

# Grado en Ingeniería Informática

## Explotación de la Información Módulo 3. Extracción de Información

Antonio Ferrández Rodríguez




UNIVERSIDAD DE ALICANTE



Grupo de Procesamiento  
del Lenguaje y Sistemas  
de Información

## Índice


1. Introducción a los sistemas de EI
2. Arquitectura de los sistemas de EI
3. Módulo de análisis léxico
4. Módulo de análisis sintáctico
5. Módulo de reconocimiento de entidades
6. Módulo de análisis semántico
7. Módulo de resolución de correferencias
8. Módulo de análisis contextual
9. Módulo de extracción, rellenado y almacenamiento de plantillas
10. Ejemplos de sistemas de EI



# 1. Introducción a los sistemas de Extracción de Información

## # Definiciones de Extracción de Información (*Information Extraction*):

- Cowie y Lehnert (1996). “Técnica que proporciona determinada información denominada relevante de un conjunto de textos todos ellos relevantes”
- Gaizauskas y Wilks (1998). “Es la actividad de extraer automáticamente un tipo de información pre-especificada desde textos”

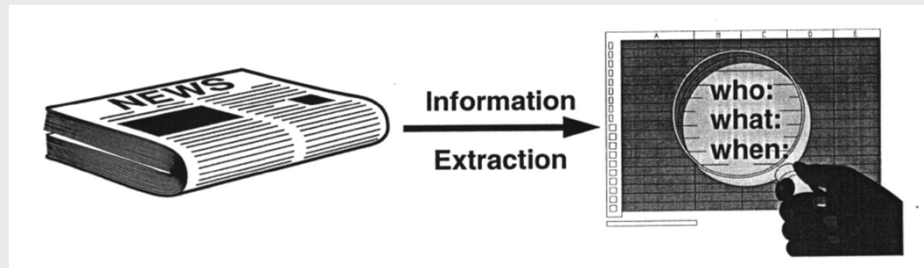


# 1. Introducción a los sistemas de Extracción de Información

## # Objetivo:

- Encontrar y relacionar información relevante mientras ignoran otras informaciones NO relevantes
- La relevancia se determina a partir de guías predefinidas: *plantillas*
  - Deben especificar con la mayor exactitud posible el tipo de información a extraer
- Desde la perspectiva del Procesamiento del Lenguaje Natural, los sistemas de EI deben trabajar a distintos niveles:
  - Desde el reconocimiento de palabras hasta el análisis de frases y desde el entendimiento a nivel de frase hasta el texto completo
  - Entrada: texto no estructurado
  - Salida: texto estructurado en forma de plantillas

# 1. Introducción a los sistemas de Extracción de Información



# 1. Introducción a los sistemas de Extracción de Información

## # Ejemplo de extracción de información:

**Hadson Corp.** said **it** expects to report a **third quarter net loss** of \$ 17 million to \$ 19 million because of special reserves and continued low natural gas prices. **The Oklahoma City energy and defense concern** said **it** will record a \$ 7. 5 million reserve for **its** defense group, including a \$ 4. 7 million charge related to problems under a fixed price development contract and \$ 2. 8 million in overhead costs that won't be reimbursed. In addition, **Hadson** said **it** will write off about \$ 3. 5 million in costs related to international exploration leases where exploration efforts have been unsuccessful. **The company** also cited interest costs and amortization of goodwill as factors in **the loss** . A year earlier, net income was \$ 2. 1 million, or six cents a share, on revenue of \$ 169. 9 million

Company Losses

| company name | company description                          | loss description         | amount                         | link to text           |
|--------------|--|--------------------------|--------------------------------|------------------------|
| Hadson Corp. | The Oklahoma City energy and defense concern | a third quarter net loss | \$ 17 million to \$ 19 million | <a href="#">source</a> |

# 1. Introducción a los sistemas de Extracción de Información

## # Ejemplo de EI (LabTL-INAOE México):

El senador liberal Federico Estrada Vélez fue secuestrado el tres de abril en la esquina de las calles 60 y 48 oeste en Medellín... Horas después, por medio de una llamada anónima a la policía metropolitana y a los medios, los Extraditables se atribuyeron la responsabilidad del secuestro... La semana pasada Federico Estrada Vélez había rechazado pláticas entre el gobierno y traficantes de drogas.

### INFORMACIÓN DEL INCIDENTE

|                 |                        |
|-----------------|------------------------|
| CATEGORÍA       | Ataque terrorista      |
| TIPO            | Secuestro              |
| FECHA           | 03 de abril (de 1990)  |
| LUGAR           | Medellín (Colombia)    |
| GRUPO AUTOR     | Los Extraditables      |
| OBJETIVO HUMANO | Federico Estrada Vélez |

7

# 1. Introducción a los sistemas de Extracción de Información

## # Ejercicio 1:

- Explica las diferencias que encuentras entre las tareas de recuperación de información y la de extracción de información

8

Explotación de la información. Extracción de Información

## 2. Arquitectura de los sistemas de Extracción de Información

# **Módulos:**

- Análisis léxico
- Análisis sintáctico
- Reconocimiento de entidades
- Análisis semántico
- Resolución de correferencias
- Análisis contextual o pragmático
- Patrones de extracción
- Rellenado y almacenamiento de plantillas

9

Explotación de la información. Extracción de Información

## 2. Arquitectura de los sistemas de Extracción de Información

# Nivel fonológico → sonido

# Nivel morfo-léxico → palabra

# Nivel sintáctico → sintagma

# Nivel semántico → significado

# Nivel pragmático → texto

Niveles de Representación Lingüística

Explotación de la información. Extracción de Información

### 3. Módulo de análisis léxico

#### # Ambigüedad léxica

- Se sentó en el banco.
- Entró en el banco y fue a la ventanilla.
- Juan se dejó el periódico en el banco.
- El avión localizó el banco y comunicó su situación.

Explotación de la información. Extracción de Información

### 3. Módulo de análisis léxico

#### # Análisis léxico:

- Proceso que transforma el texto de entrada (caracteres) en una secuencia de unidades significativas (unidades léxicas) con información asociada

#### # Tareas:

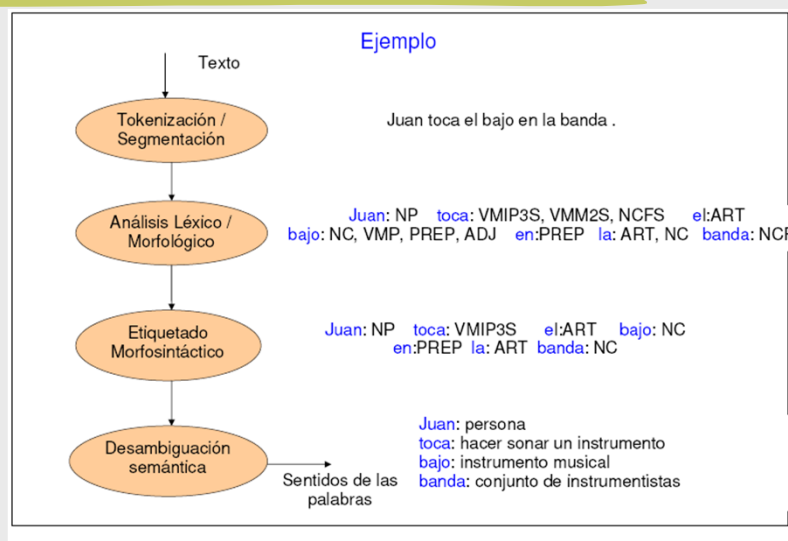
- Segmentación
- Análisis morfológico
- Etiquetado morfosintáctico
- Desambiguación del sentido de las palabras

### 3. Módulo de análisis léxico

#### # Características léxicas

- Características morfológicas:
  - Raíz, lema, género, número, persona, tiempo, modo, etc.
- Categoría morfosintáctica o gramatical:
  - Nombre común, nombre propio, pronombre (él, ella), verbo, adjetivo (determinado: el, numeral: 10, demostrativo: este), adverbio (allí), preposición (por), conjunción (y).
- Información semántica
  - Sentido, glosa
- Pronunciación de la palabra

### 3. Módulo de análisis léxico



### 3. Módulo de análisis léxico

**ENTRADA:** Roberto González Cepeda que nació el 12 de enero de 1900, era un gran hombre de letras que por lo menos publicó 100 libros.

Roberto\_González\_Cepeda roberto\_gonzález\_cepeda 5 NP00000  
 que que CS00 que PR3CN000  
 nació nacer VMIS3S0  
 el el TDMS0  
 12\_de\_enero\_de\_1900 12/1/1900 F W  
 , , E2 Fc  
 era era NCFS000 erar VMMP2S0 erar VMIP3S0 ser VAIH3S0 ser VAIH3S0  
 un un TIMS0 un MCMS00  
 gran gran AQ0CS00  
 hombre hombre NCMS000  
 de de SPS00 de NCFS000  
 letras letra NCFP000  
 que que CS00 que PR3CN000  
 por\_lo\_menos por\_lo\_menos RG000  
 publicó publicar VMIS3S0  
 100 100 Z  
 libros libro NCMP000

Tokenización

+

Análisis Morfológico

15

### 3. Módulo de análisis léxico

#### # Desambiguación del sentido de las palabras (*Word Sense Disambiguation, WSD*):

- Asignar el sentido correcto a las palabras
  - “Te voy a firmar la cara con la planta de mi pié”

#### WordNet 1.5

1. **planta, piso** -- a room or set of rooms comprising a single level of a multi-level building
2. **planta, flora** -- a living organism lacking the power of locomotion
3. **planta** -- the underside of the foot
4. **planta, fábrica** -- buildings for carrying on industrial labor
5. **planta, distribución** -- a floor plan for the ground level of a building

16



## 10. Ejemplos de sistemas de EI

# Ejercicio 2:

- Tarea de EI del *MUC-4 Terrorism Task*. ¿Qué información léxica sería útil para extraer: fechas, personas y localizaciones?
  - <http://www.opener-project.eu/webservices/entrance.html>
  - <http://demo2-opener.rhcloud.com/welcome.action>
  - [http://cogcomp.cs.illinois.edu/page/demo\\_view/POS](http://cogcomp.cs.illinois.edu/page/demo_view/POS)

SANTIAGO, 10 JAN 90 -- [TEXT] POLICE ARE CARRYING OUT INTENSIVE OPERATIONS IN THE TOWN OF MOLINA IN THE SEVENTH REGION IN SEARCH OF A GANG OF ALLEGED EXTREMISTS WHO COULD BE LINKED TO A RECENTLY DISCOVERED ARSENAL. IT HAS BEEN REPORTED THAT CARABINEROS IN MOLINA RAIDED THE HOUSE OF 25-YEAR-OLD WORKER MARIO MUNOZ PARDO, WHERE THEY FOUND A FAL RIFLE, AMMUNITION CLIPS FOR VARIOUS WEAPONS, DETONATORS, AND MATERIAL FOR MAKING EXPLOSIVES.


IT SHOULD BE RECALLED THAT A GROUP OF ARMED INDIVIDUALS WEARING SKI MASKS ROBBED A BUSINESSMAN ON A RURAL ROAD NEAR MOLINA ON 7 JANUARY. THE BUSINESSMAN, ENRIQUE ORMAZABAL ORMAZABAL, TRIED TO RESIST; THE MEN SHOT HIM AND LEFT HIM SERIOUSLY WOUNDED. HE WAS LATER HOSPITALIZED IN CURICO. CARABINEROS CARRIED OUT SEVERAL OPERATIONS, INCLUDING THE RAID ON MUNOZ' HOME. THE POLICE ARE CONTINUING TO PATROL THE AREA IN SEARCH OF THE ALLEGED TERRORIST COMMAND.

17

## 4. Módulo de análisis sintáctico


# Ambigüedad sintáctica

- La vendedora de periódicos del barrio.
- Juan vio al ladrón con los prismáticos
- Pedro vio a Juan en lo alto de la montaña con los prismáticos



## 4. Módulo de análisis sintáctico

- # **Las palabras se combinan formando constituyentes a un nivel sintáctico superior**
- # **Tipos de constituyentes:**
  - Básicos (símbolos terminales)
  - Superiores (símbolos no terminales)



## 4. Módulo de análisis sintáctico

- # **Símbolos terminales. Tipos:**
  - **Clases abiertas:**
    - Regularmente se van introduciendo nuevas palabras pertenecientes a estas categorías (nombres, adjetivos, etc.)
  - **Clases cerradas:**
    - Raramente se introducen nuevas palabras en estas clases (artículos, preposiciones, pronombres, etc.)

Explotación de la información. Extracción de Información en AN

4. Módulo de análisis sintáctico

# Símbolos no terminales. Ejemplos:

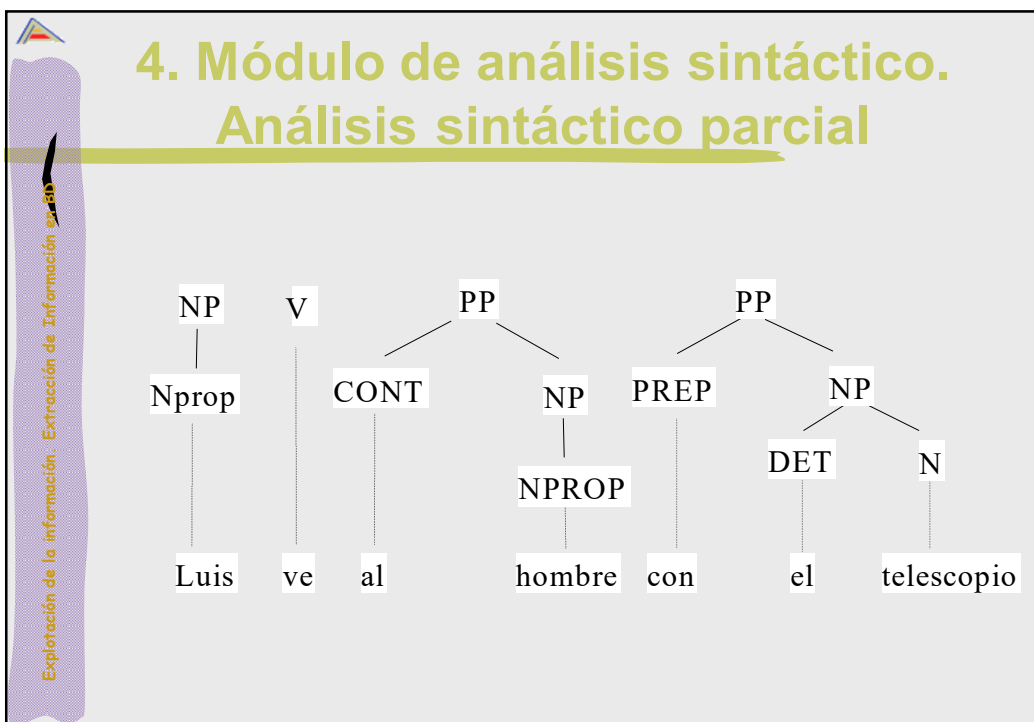
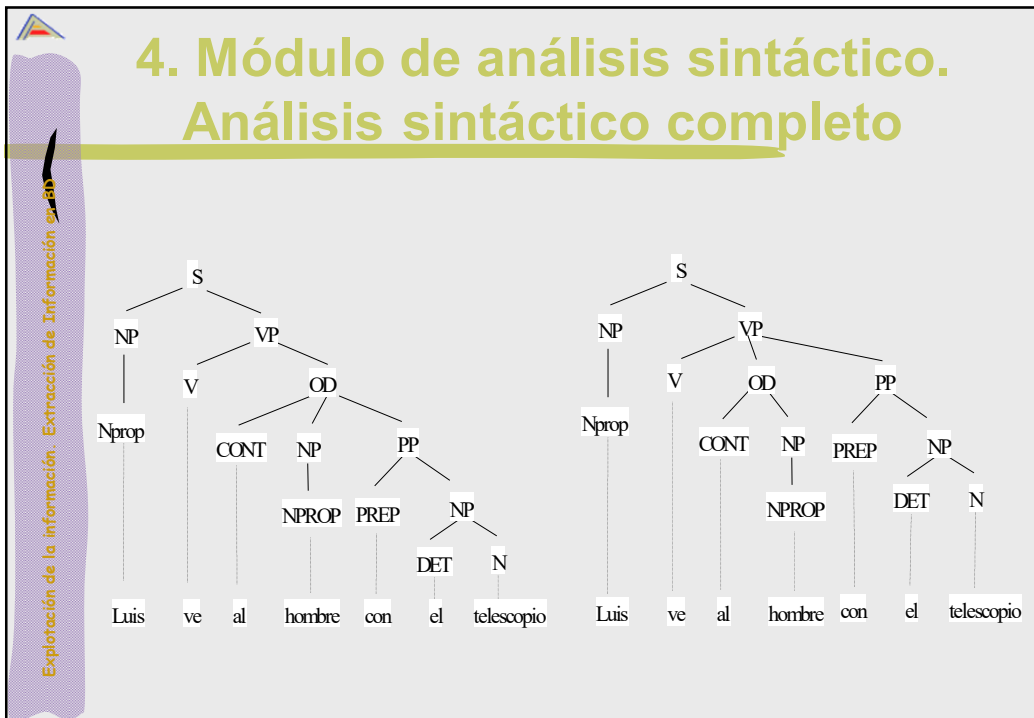
- Los sintagmas nominales (np)
  - Se utilizan para referirse a objetos, lugares, conceptos, cualidades, etc.
  - Distintas composiciones:
    - Un pronombre personal
    - o un nombre propio
    - o cualquier otra combinación de palabras cuyo núcleo es un nombre.
- Sintagmas preposicionales (pp), oraciones, cláusulas de relativo, etc.


Explotación de la información. Extracción de Información en AN

4. Módulo de análisis sintáctico

# Tipos de análisis sintáctico:

- Completo:
  - Determinar si una frase es gramaticalmente correcta.
  - Proporcionar una estructura asociada a la frase que refleje sus relaciones sintácticas.
- Parcial:
  - Extraer determinados constituyentes, saltándose los no conocidos.






## 4. Módulo de análisis sintáctico

### # Analizador sintáctico-semántico:

- Hace corresponder una oración con su estructura sintáctica y su forma lógica.
- Usa el conocimiento sobre palabras, su significado y un conjunto de reglas que definen las estructuras correctas del lenguaje (gramática).
- El objetivo de realizar estos dos procesos conjuntamente es para reducir el número de posibles interpretaciones.



## 5. Módulo de reconocimiento de entidades

### # Reconocimiento de entidades:

- Identificación:
  - Normalmente sintagmas nominales
  - Necesita un tratamiento adecuado de la coordinación

### # Clasificación de entidades:

- Persona
- Organización
- Lugar
- Fecha
- Tiempo
- Moneda
- Porcentaje

## 5. Módulo de reconocimiento de entidades

### # Técnicas de clasificación de entidades:

#### ■ Basados en conocimiento:

##### ■ Utilización de diccionarios:

- # Diccionarios específicos de entidades (del dominio o de uso general como WordNet)
- # Diccionarios de palabras comunes que inician una frase y aparecen en mayúsculas (p.ej. La)
- # Diccionarios de números escritos como letras

##### ■ Utilización de reglas:

- # entidad\_fecha → mes | mes conector\_fecha número | número conector\_fecha entidad\_fecha
- # entidad\_nombre → nombre | nombre conector\_nombre entidad\_nombre
- # entidad\_cantidad → número([“.” | “,”]número)? | número([“.” | “,”]número)?
- # entidad\_cantidad
- # mes → enero | ... | diciembre
- # conector\_fecha → de | - | ... | e
- # nombre → [A-Z][A-Za-z]\*
- # conector\_nombre → de | la | ... | e
- # número → [0-9]+

#### ■ Basados en aprendizaje (estadísticos):

- Supervisados
- No supervisados

## 10. Ejemplos de sistemas de EI


### # Ejercicio 2 (cont.):

- Tarea de EI del *MUC-4 Terrorism Task*. ¿Qué información sintáctica sería útil para extraer: fechas, personas y localizaciones? Etiqueta las entidades

- <http://demo2-opener.rhcloud.com/welcome.action>
- [http://cogcomp.cs.illinois.edu/page/demo\\_view/ShallowParse](http://cogcomp.cs.illinois.edu/page/demo_view/ShallowParse)
- [http://cogcomp.cs.illinois.edu/page/demo\\_view/NER](http://cogcomp.cs.illinois.edu/page/demo_view/NER)

SANTIAGO, 10 JAN 90 -- [TEXT] POLICE ARE CARRYING OUT INTENSIVE OPERATIONS IN THE TOWN OF MOLINA IN THE SEVENTH REGION IN SEARCH OF A GANG OF ALLEGED EXTREMISTS WHO COULD BE LINKED TO A RECENTLY DISCOVERED ARSENAL. IT HAS BEEN REPORTED THAT CARABINEROS IN MOLINA RAIDED THE HOUSE OF 25-YEAR-OLD WORKER MARIO MUNOZ PARDO, WHERE THEY FOUND A FAL RIFLE, AMMUNITION CLIPS FOR VARIOUS WEAPONS, DETONATORS, AND MATERIAL FOR MAKING EXPLOSIVES.

IT SHOULD BE RECALLED THAT A GROUP OF ARMED INDIVIDUALS WEARING SKI MASKS ROBBED A BUSINESSMAN ON A RURAL ROAD NEAR MOLINA ON 7 JANUARY. THE BUSINESSMAN, ENRIQUE ORMAZABAL ORMAZABAL, TRIED TO RESIST; THE MEN SHOT HIM AND LEFT HIM SERIOUSLY WOUNDED. HE WAS LATER HOSPITALIZED IN CURICO. CARABINEROS CARRIED OUT SEVERAL OPERATIONS, INCLUDING THE RAID ON MUNOZ' HOME. THE POLICE ARE CONTINUING TO PATROL THE AREA IN SEARCH OF THE ALLEGED TERRORIST COMMAND.




Explotación de la información. Extracción de Información en IA

## 6. Módulo de análisis semántico

### # Ambigüedad semántica

- Juan dio un pastel a los niños
  - Uno para todos?
  - Uno para cada uno?

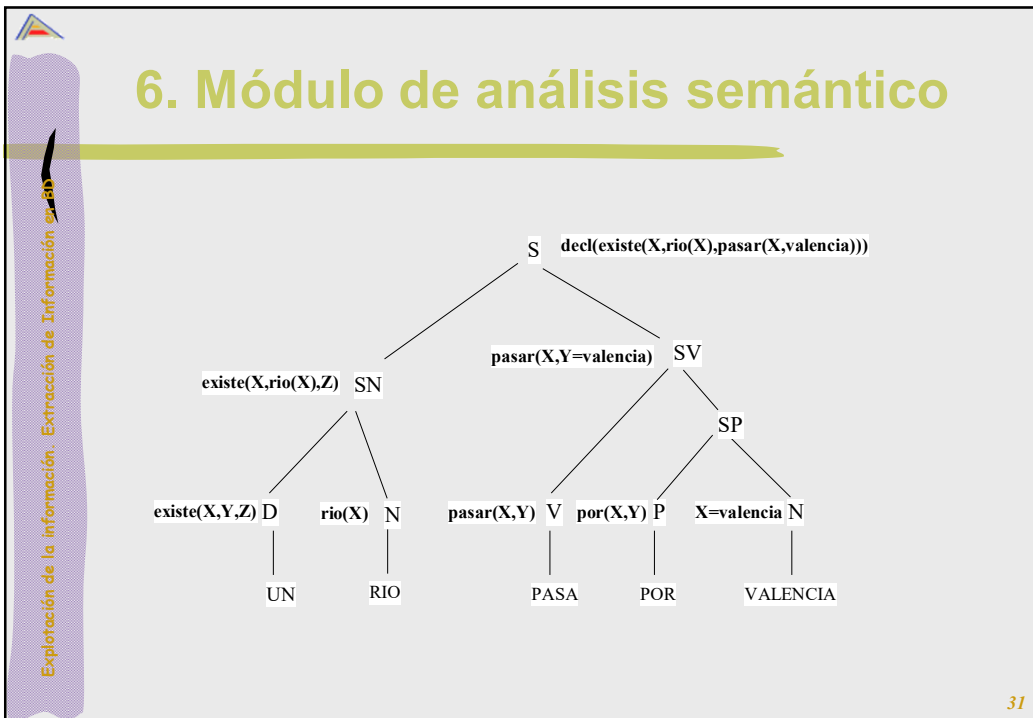


Explotación de la información. Extracción de Información en IA

## 6. Módulo de análisis semántico

### # Análisis semántico:

- Obtención de la *Forma Lógica*:
  - Representación del significado de una oración que es independiente del contexto
  - Un único significado que puede ser utilizado con propósitos diferentes
- Ejemplos:
  - Dime los ríos que nacen en Madrid y desembocan en Valencia:
    - $\# \text{preg}(X, \text{río}(X) \ \& \ \text{nacer}(X, \text{madrid}) \ \& \ \text{desembocar}(X, \text{valencia}))$
  - Todos los ríos pasan por Valencia:
    - $\# (\forall x) R(x) \rightarrow P(x, \text{valencia})$
  - Un río grande pasa por Valencia:
    - $\# (\exists x) R(x) \wedge G(x) \wedge P(x, \text{valencia})$



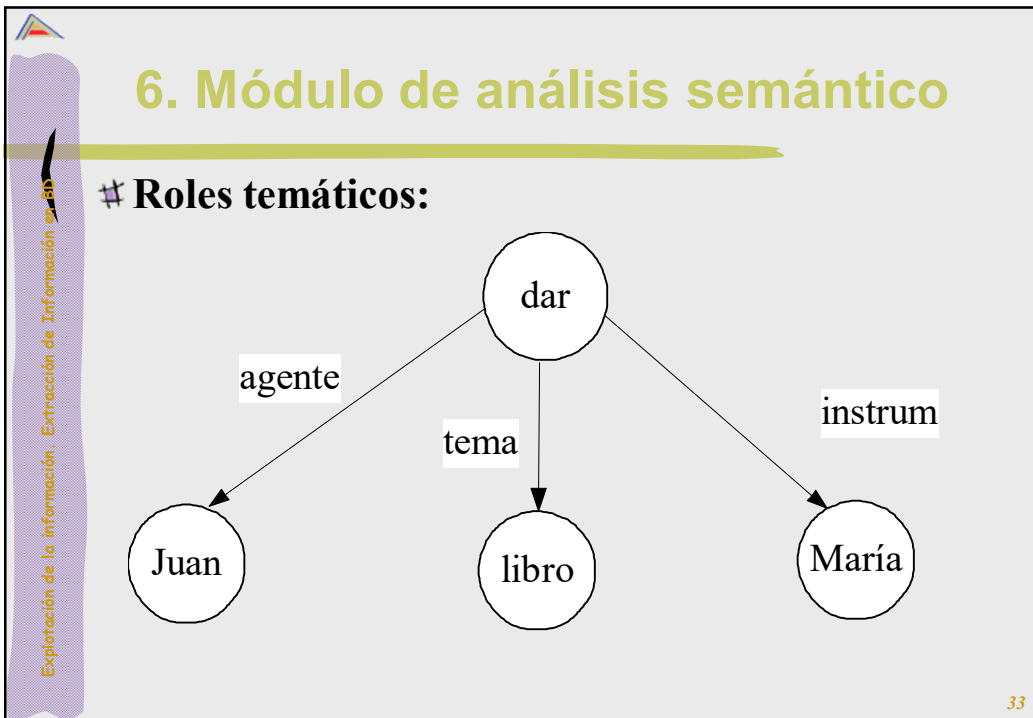
6. Módulo de análisis semántico

# Un plural *distributivo* en el que la propiedad debe distribuirse entre todas las combinaciones posibles, para cada una de las entidades afectadas:

- “Juan y María saben Latín y Griego”
  - $saber(Juan, Latín) \wedge saber(María, Latín) \wedge$   
 $saber(Juan, Griego) \wedge saber(María, Griego)$
- “Juan y María saben Latín y Griego respectivamente”
  - $saber(Juan, Latín) \wedge saber(María, Griego)$

32





Explotación de la información. Extracción de Información

## 6. Módulo de análisis semántico

### # Actos del habla:

- Oraciones declarativas o aserciones: DECL
  - El Turia pasa por Valencia → decl(pasar(turia,valencia))
- Oraciones interrogativas de cierto-falso: SÍ\_NO
  - ¿El Turia pasa por Valencia? → sino(pasar(turia,valencia))
- Oraciones interrogativas de tipo cantidad : CANT
  - ¿Cuántos ríos pasan por Valencia? → cant(X, rio(X) & pasar(X,valencia))
- Oraciones interrogativas de tipo general: PREG
  - ¿Qué ríos pasan por Valencia? → preg(X, rio(X) & pasar(X,valencia))
- Oraciones imperativas u órdenes: PREG
  - Dime los ríos que pasan por Valencia → preg(X, rio(X) & pasar(X,valencia))

34

Explotación de la información. Extracción de Información

## 7. Módulo de resolución de correferencias

### # Ambigüedad referencial

- él le dijo, después, que lo pusiera encima
  - Quién dijo?
  - A quién?
  - Cuándo, después de qué?
  - Que pusiera qué?
  - Encima de dónde?

Explotación de la información. Extracción de Información

## 7. Módulo de resolución de correferencias

### # Definición de anáfora:

- (Hirst, 81) “La anáfora es el mecanismo que nos permite hacer en un discurso una referencia abreviada a alguna entidad o entidades, con la confianza de que el receptor del discurso sea capaz de interpretar la referencia y por consiguiente determinar la entidad a la que se alude”
  - A la referencia abreviada se la llama expresión o elemento anafórico
  - A la entidad referenciada se la denomina antecedente o referente

36

Explotación de la información. Extracción de Información

## 7. Módulo de resolución de correferencias

# **Noción de correferencialidad:**

- Una expresión anafórica no es que se refiera a su antecedente, sino al referente de la expresión que sirve de antecedente
- Por ello ha de hablarse de correferencialidad entre expresión anafórica y antecedente
  - Pedro<sub>i</sub> entró en la tienda... Él<sub>i</sub> buscaba un regalo...

37

Explotación de la información. Extracción de Información

## 7. Módulo de resolución de correferencias


# **Distinción entre antecedente y referente (Brown y Yule, 83):**

- Referente: constituiría la representación mental de los objetos evocados por el texto
- Antecedente: sería la representación lingüística que estos toman en el mismo


Expresión anafórica (él)

Antecedente (Pedro)

Referente




38



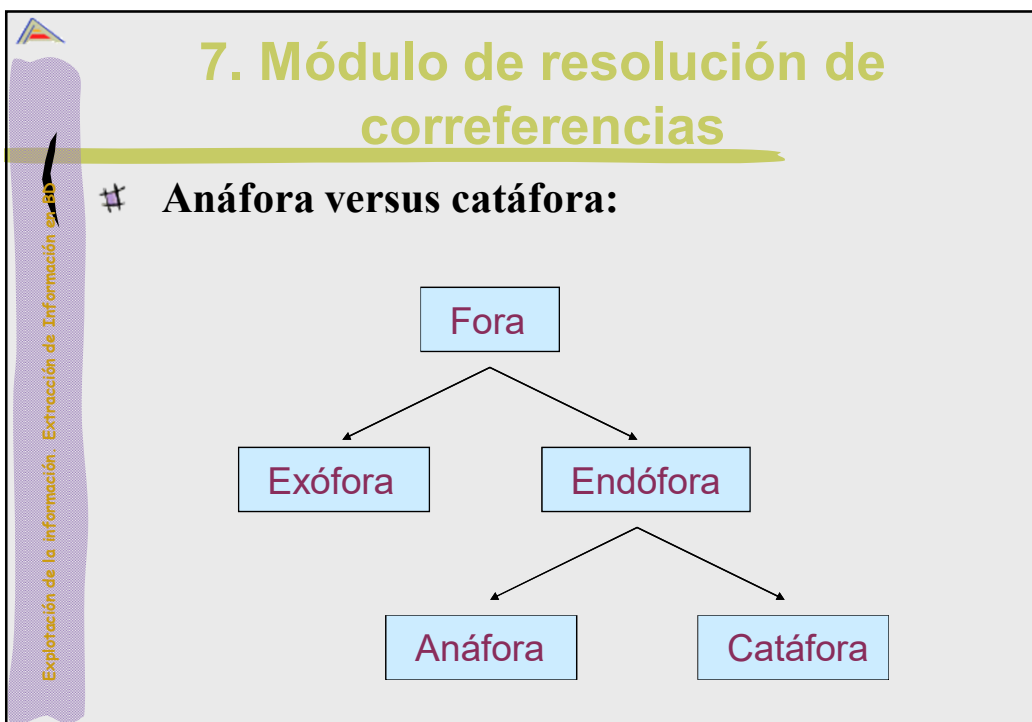
## 7. Módulo de resolución de correferencias

- # **Resolución de la anáfora generada por el usuario.**
  - Debe ser interpretada por el sistema
- # **Generación de la anáfora por el sistema.**
  - Proporciona naturalidad
  - Topicaliza
  - Remarca la estructura del diálogo



## 7. Módulo de resolución de correferencias

- # **Contextos en los que se desarrolla la anáfora:**
  - Convencional: darla con queso, arreglárselas o pasarlo bien
  - Situacional: Dame éste
  - Lingüístico: Juan enjabona al bebé<sub>i</sub> y María lo<sub>i</sub> seca



7. Módulo de resolución de correferencias

# Catáfora:

- Casos en que la expresión anafórica aparece antes que el antecedente al cual se refiere:
  - Cerca de él<sub>i</sub>, Juan<sub>j</sub> vio una serpiente.
- Elipsis catafórica:
  - Si  $\emptyset$ <sub>i</sub> gana en la lotería, Juan<sub>j</sub> se compra un piano.

Explotación de la información. Extracción de Información

## 7. Módulo de resolución de correferencias

# **Clasificación de la anáfora según la accesibilidad del antecedente:**

1. Morfosintáctica:
  - Nominal: Tu hijo<sub>i</sub> rompió el cristal. Yo le<sub>i</sub> vi
  - Verbal: Pedro [jugó muy bien al tenis ayer]<sub>i</sub>, pero Juan [lo hizo]<sub>i</sub> muy mal
  - Oracional: [No deberíamos salir esta noche]<sub>i</sub>. Yo no opino eso<sub>i</sub>
2. Semántica:
  - Sinonimia: Pedro se quitó sus gafas<sub>i</sub>... Estas lentes<sub>i</sub>...
  - Hiperonimia: No sabía que ese coche<sub>i</sub> es tuyo. Opino que es un buen vehículo<sub>i</sub>
  - Contextual: Él se limpió las gafas<sub>i</sub> y se las ajustó a la nariz. Su montura<sub>i</sub> y cristales<sub>i</sub> estaban húmedas
3. Pragmática: ...”La isla del tesoro”<sub>i</sub>... En ese libro<sub>i</sub>...


Explotación de la información. Extracción de Información

## 7. Módulo de resolución de correferencias

# **Clasificación según el tipo de expresión anafórica:**

- Pronominal: Tu hijo<sub>i</sub> rompió el cristal. Yo le<sub>i</sub> vi
- Descripciones definidas: Juan<sub>i</sub> perdió el dinero, [el pobre chico]<sub>i</sub> está hundido
- Tipo “one”: Peter bought [a blue pen]<sub>i</sub> yesterday. He has bought [another one]<sub>i</sub> today
- Adjetiva: Compré [una pera verde]<sub>i</sub> y [otra roja]<sub>i</sub>. Yo prefiero [la verde]<sub>i</sub>
- Superficial numérica: Pedro miró [al perro<sub>i</sub> y al gato<sub>i</sub>], pero finalmente eligió el primero<sub>i</sub>
- Verbal: Pedro [besó a su mujer]<sub>i</sub>. Juan también [lo hizo]<sub>i</sub>
- Adverbios: La iglesia estaba [detras de la librería]<sub>i</sub>. Luis fue ahí<sub>i</sub> después del almuerzo
- Complementos circunstanciales: El despertador suena a [las 6 de la mañana]<sub>i</sub>. [Las siguientes dos horas]<sub>i</sub>...


44



Explotación de la información. Extracción de Información

## 7. Módulo de resolución de correferencias

- # **Estrategias basadas en conocimiento lingüístico**
  - Imitan fuentes de conocimiento humano
  - Consultivos
    - una única fuente de información
  - Democráticos
    - combinan varias fuentes de información
    - mecanismos de restricciones y preferencias
      - # reglas para descartar candidatos
      - # reglas para ordenar los candidatos
- # **Estrategias basadas en corpus**
  - Estudian corpus a través de herramientas estadísticas
  - Proponen modelos probabilísticos



Explotación de la información. Extracción de Información


## 10. Ejemplos de sistemas de EI

- # **Ejercicio 2 (cont.):**
  - Tarea de EI del *MUC-4 Terrorism Task*. Etiqueta anáforas y sus soluciones
    - [http://cogcomp.cs.illinois.edu/page/demo\\_view/Corf](http://cogcomp.cs.illinois.edu/page/demo_view/Corf)

SANTIAGO, 10 JAN 90 -- [TEXT] POLICE ARE CARRYING OUT INTENSIVE OPERATIONS IN THE TOWN OF MOLINA IN THE SEVENTH REGION IN SEARCH OF A GANG OF ALLEGED EXTREMISTS WHO COULD BE LINKED TO A RECENTLY DISCOVERED ARSENAL. IT HAS BEEN REPORTED THAT CARABINEROS IN MOLINA RAIDED THE HOUSE OF 25-YEAR-OLD WORKER MARIO MUNOZ PARDO, WHERE THEY FOUND A FAL RIFLE, AMMUNITION CLIPS FOR VARIOUS WEAPONS, DETONATORS, AND MATERIAL FOR MAKING EXPLOSIVES.


IT SHOULD BE RECALLED THAT A GROUP OF ARMED INDIVIDUALS WEARING SKI MASKS ROBBED A BUSINESSMAN ON A RURAL ROAD NEAR MOLINA ON 7 JANUARY. THE BUSINESSMAN, ENRIQUE ORMAZABAL ORMAZABAL, TRIED TO RESIST; THE MEN SHOT HIM AND LEFT HIM SERIOUSLY WOUNDED. HE WAS LATER HOSPITALIZED IN CURICO. CARABINEROS CARRIED OUT SEVERAL OPERATIONS, INCLUDING THE RAID ON MUNOZ' HOME. THE POLICE ARE CONTINUING TO PATROL THE AREA IN SEARCH OF THE ALLEGED TERRORIST COMMAND.

46



## 8. Módulo de análisis contextual

- # **Interpretación Contextual:**
  - Proceso de emparejar una forma lógica al lenguaje de representación final del conocimiento/significado
- # **Incluye diversos mecanismos para cubrir aspectos tales como:**
  - Identificación de objetos referenciados por un SN
  - Análisis de aspectos temporales
  - Identificación de la intención del hablante
  - Proceso inferencial requerido para interpretar la oración dentro del dominio de aplicación



## 9. Módulo de extracción, rellenado y almacenamiento de plantillas

- # **Algoritmo:**
  - Obtención de la información de cada módulo
  - Marcado de la información clave del documento:
    - Entidades, fechas, lugares, cantidades, relaciones entre ellas, etc.
  - Aplicación de patrones de extracción:
    - Los patrones se diseñan ad-hoc para el dominio o bien se aplica un proceso de aprendizaje automático a partir de texto previamente etiquetado
  - Rellenado de las plantillas

48



## 9. Módulo de extracción, rellenado y almacenamiento de plantillas

### # Representación de Documentos mediante Grafos de Relaciones:

- Tras la salida de los módulos
- Se estructura en forma de grafos:
  - Cada nodo es una entidad
  - Cada arco es una relación entre entidades
  - Se “colapsa” para reflejar las relaciones de coreferencia

49

## 9. Módulo de extracción, rellenado y almacenamiento de plantillas

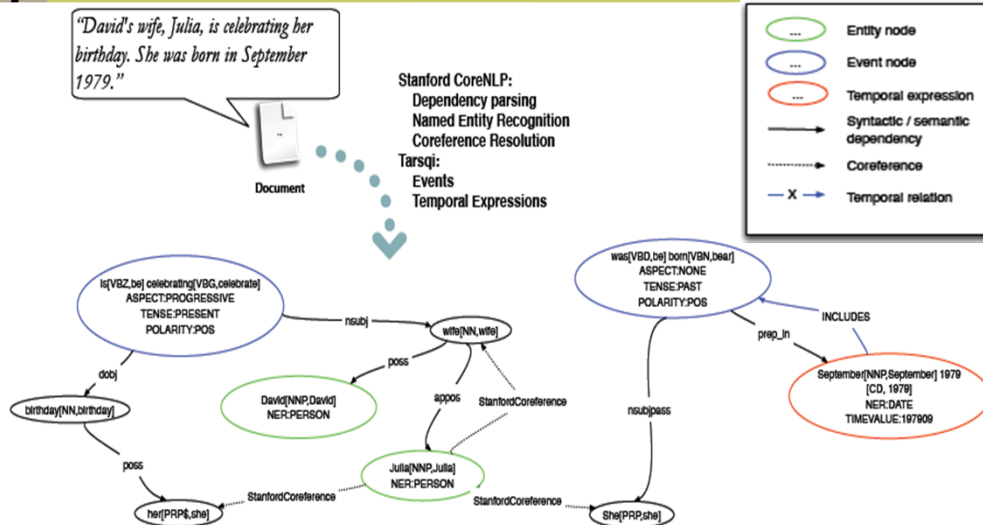
*“David’s wife, Julia, is celebrating her birthday. She was born in September 1979.”*



Document

Stanford CoreNLP:  
Dependency parsing  
Named Entity Recognition  
Coreference Resolution

Tarsqi:  
Events  
Temporal Expressions



50

## 9. Módulo de extracción, relleno y almacenamiento de plantillas

Collapse referents of discourse

Graph normalization

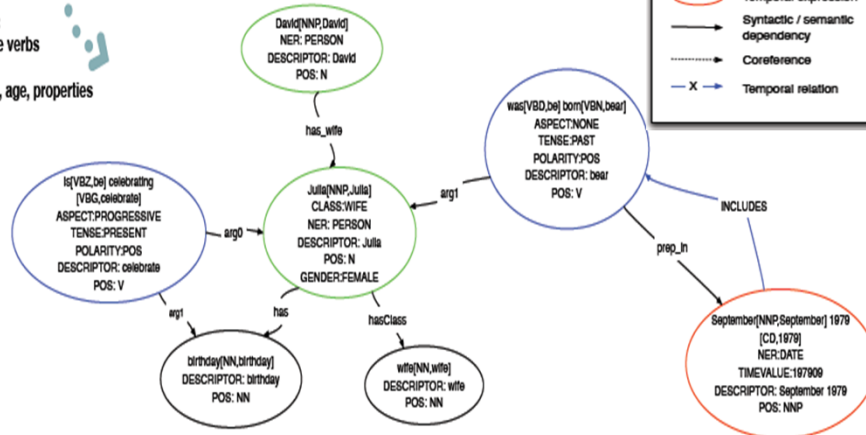
Semantic information:

Normalize copulative verbs

Normalize genitives

Infer semantic class, age, properties

Assign a gender



51

## 10. Ejemplos de sistemas de EI

### # Ejercicio 2 (cont.): EJERCICIO A ENVIAR COMO TUTORÍA CV

- Aplica los módulos de EI, detallando la salida de cada uno, para realizar la siguiente tarea de EI del MUC-4 Terrorism Task, sobre el documento:

- POS tagger inglés: [http://cogcomp.cs.illinois.edu/page/demo\\_view/POS](http://cogcomp.cs.illinois.edu/page/demo_view/POS)

- POS tagger español: [http://www.mystilus.com/Morphosyntactic\\_analyzer#](http://www.mystilus.com/Morphosyntactic_analyzer#)

- WordNet: [http://adimen.si.edu/cgi-bin/wei/public/wei.consult.perl?item=sales&button1=Look\\_up&metode=Word&pos=Nouns&llengua=English\\_3.0&search=has\\_hyperonym&estructura=English\\_3.0&glos=Gloss&levin=1&eng-30=English\\_3.0](http://adimen.si.edu/cgi-bin/wei/public/wei.consult.perl?item=sales&button1=Look_up&metode=Word&pos=Nouns&llengua=English_3.0&search=has_hyperonym&estructura=English_3.0&glos=Gloss&levin=1&eng-30=English_3.0)

SANTIAGO, 10 JAN 90 -- [TEXT] POLICE ARE CARRYING OUT INTENSIVE OPERATIONS IN THE TOWN OF MOLINA IN THE SEVENTH REGION IN SEARCH OF A GANG OF ALLEGED EXTREMISTS WHO COULD BE LINKED TO A RECENTLY DISCOVERED ARSENAL. IT HAS BEEN REPORTED THAT CARABINEROS IN MOLINA RAIDED THE HOUSE OF 25-YEAR-OLD WORKER MARIO MUNOZ PARDO, WHERE THEY FOUND A FAL RIFLE, AMMUNITION CLIPS FOR VARIOUS WEAPONS, DETONATORS, AND MATERIAL FOR MAKING EXPLOSIVES.

IT SHOULD BE RECALLED THAT A GROUP OF ARMED INDIVIDUALS WEARING SKI MASKS ROBBED A BUSINESSMAN ON A RURAL ROAD NEAR MOLINA ON 7 JANUARY. THE BUSINESSMAN, ENRIQUE ORMAZABAL ORMAZABAL, TRIED TO RESIST; THE MEN SHOT HIM AND LEFT HIM SERIOUSLY WOUNDED. HE WAS LATER HOSPITALIZED IN CURICO. CARABINEROS CARRIED OUT SEVERAL OPERATIONS, INCLUDING THE RAID ON MUNOZ' HOME. THE POLICE ARE CONTINUING TO PATROL THE AREA IN SEARCH OF THE ALLEGED TERRORIST COMMAND.

52

## 10. Ejemplos de sistemas de EI

Explotación de la información. Extracción de Información en 3D

### # Ejercicio 2 (cont.):

- Rellenando la siguiente plantilla:
  - INCIDENT: DATE
  - INCIDENT: LOCATION
  - INCIDENT: TYPE
  - INCIDENT: STAGE OF EXECUTION
  - INCIDENT: INSTRUMENT TYPE
  - HUM TGT: NAME
  - HUM TGT: DESCRIPTION
  - HUM TGT: TYPE

53


## 10. Ejemplos de sistemas de EI

Explotación de la información. Extracción de Información en 3D

### # LabTL-INAOE México:

- Escenario
  - Noticias en Español
    - # Desastres Naturales
      - Forestal, Huracán, Inundación, Sequía, Sismo
- Plantilla de extracción
  - Información del evento
    - Fecha, Lugar, Magnitud
  - Información de personas
    - Muertos, Heridos, Desaparecidos, Damnificados, Afectados
  - Información de viviendas e infraestructura
    - Destruídas, Afectadas, Hectáreas, Pérdida económica
- Técnica: aprendizaje automático supervisado

54




Explotación de la información. Extracción de Información en EII

Explotación de la información. Extracción de Información en EII

## 10. Ejemplos de sistemas de EI

|              | Filtrado de documentos  | Extracción de información  |
|--------------|---|--|
| Relevantes   | 439   | 2025 (20%)   |
| Irrelevantes | 229 (34%)   | 7926   |
|              | Noticia   | Entidad  |
| Relevante    | El <u>huracán</u> Isidore dejó en la península de Yucatán 300 mil personas <u>damnificadas</u> y el <u>deceso</u> de una persona              | En el peor temblor del siglo en Puebla, <u>11</u> muertos  |
| Irrelevante  | Cuando Beijing estaba en el ojo del <u>huracán</u> de la neumonía atípica, dejó 2 mil 561 <u>enfermos</u> , de los cuales 192 <u>murieron</u> | El <u>palacio</u> municipal, construido en <u>1536</u> , fue el monumento que presentó los daños más severos |



Explotación de la información. Extracción de Información en EII

Explotación de la información. Extracción de Información en EII

## 10. Ejemplos de sistemas de EI

|                   | REC | PRE |
|-------------------|-----|-----|
| EVE_FECHA         | 99  | 95  |
| EVE_LUGAR         | 60  | 50  |
| EVE_MAGNITUD      | 96  | 73  |
| PER_MUERTAS       | 73  | 66  |
| PER_HERIDAS       | 84  | 91  |
| PER_DESAPARECIDAS | 93  | 78  |
| PER_DAMNIFICADAS  | 84  | 62  |
| PER_AFECTADAS     | 82  | 60  |
| VIV_DESTRUIDAS    | 83  | 70  |
| VIV_AFECTADAS     | 83  | 72  |
| INF_HECTAREAS     | 92  | 66  |
| INF_ECONOMICA     | 95  | 49  |

Explotación de la información. Extracción de Información

## 10. Ejemplos de sistemas de EI

# **Empirical methods in Information Extraction**

- Claire Cardie. Department of Computer Science. Cornell University
- <http://www.cs.cornell.edu/home/cardie/papers/ai-mag.pdf>

57

Explotación de la información. Extracción de Información

## 10. Ejemplos de sistemas de EI. Cornell University

```

    graph TD
      A[Free Text: 4 Apr Dallas - Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister occurred without warning at approximately 7:15 p.m. and destroyed two mobile homes. The Texaco station, at 102 Main Street, Farmers Branch, TX, was also severely damaged, but no injuries were reported. Total property damages are estimated to be $350,000.] --> B[Information Extraction System]
      B --> C[Output Template]
  
```

**Free Text**

4 Apr Dallas - Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister occurred without warning at approximately 7:15 p.m. and destroyed two mobile homes. The Texaco station, at 102 Main Street, Farmers Branch, TX, was also severely damaged, but no injuries were reported. Total property damages are estimated to be \$350,000.

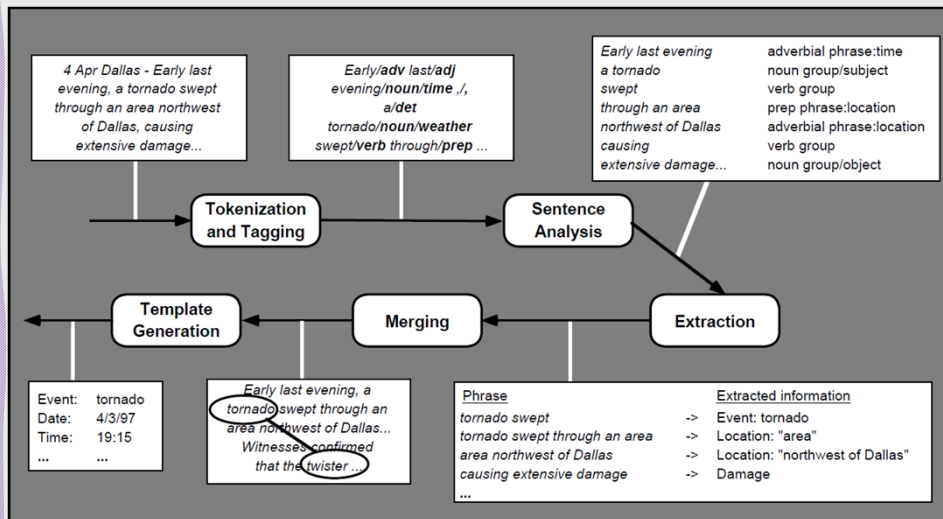
**Information Extraction System**

**Output Template**

|                   |   |
|-------------------|---|
| Event:            | tornado   |
| Date:             | 4/3/97  |
| Time:             | 19:15   |
| Location:         | Farmers Branch : "northwest of Dallas" : TX : USA |
| Damage:           | "mobile homes" (2)<br>"Texaco station" (1)        |
| Estimated Losses: | \$350,000   |
| Injuries:         | none  |

58

## 10. Ejemplos de sistemas de EI. Cornell University



59

## 10. Ejemplos de sistemas de EI

### # Sistema Annie (GATE):

📄 <http://services.gate.ac.uk/annie/>

Enter a URL:

- ☒ Person
- ☒ Location
- ☒ Organization
- ☒ Date
- ☒ Address
- ☒ Money
- ☒ Percent

📄 Salida: [annie.jsp.htm](http://annie.jsp.htm)

60

Explotación de la información. Extracción de Información en 2010

10. Ejemplos de sistemas de EI

# Ejercicio 3:

- Aplica los módulos de EI, detallando la salida de cada uno, para realizar la siguiente tarea de EI:
  - Sucesión de 6 eventos de la plantilla:
    - <SUCCESSION-1>
      - # ORGANIZATION :
      - # POST :
      - # WHO\_IS\_IN :
      - # WHO\_IS\_OUT :
  - Sobre el siguiente documento:

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent

61

Explotación de la información. Extracción de Información en 2010

10. Ejemplos de sistemas de EI

# Utilidad de este tipo de sistemas:

- <http://www.agendadigital.gob.es/planes-actuaciones/Paginas/plan-impulso-tecnologias-lenguaje.aspx>
- Oferta trabajo Xerox

62