Índice

Datos Estructurados vs. No Estructurados

- **No estructurados**: aquéllos donde la información está codificada de forma que no permite su procesamiento automático inmediato
  - i.e. **en lenguaje natural**

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.

- **Estructurados**: aquéllos de semántica definida y susceptibles de ser procesados automáticamente por el ordenador
  - Bases de datos, hojas de cálculo, registros, etc.

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

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 no estructuradas (textos), para así permitir su posterior tratamiento automático en tareas de procesamiento de la información.
- **Objetivo:** dada una colección de documentos (texto **no estructurado**), identificar y extraer de los mismos aquellos hechos y relaciones relevantes para un dominio particular (*dominio de extracción*), ignorando la información extraña e irrelevante (i.e. **obtener información estructurada a partir de docs. en lenguaje natural**)
  - La información obtenida se devuelve de forma **estructurada**.
  - Ha de **establecerse a priori** qué constituye un hecho/rela. relevante.
  - Sistemas muy 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 consultora financiera abonada a un proveedor de noticias económicas quiere realizar un seguimiento a nivel mundial de las fusiones, opas y quiebras de empresas en bolsa. Dicha información será organizada cronológicamente y por compañía.
- Una agencia de seguridad desea hacer un seguimiento del tráfico de *email* en busca de indicios de actividades delictivas.
- Un empresa de investigación biotecnológica quiere analizar toda la literatura disponible para conocer todas las interacciones de un cierto grupo de proteínas con cualquier otra proteína.

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
  - Aplicaremos **patrones** y otras técnicas afines sobre el texto para identificar y extraer la información relevante

Especificidad de la Información

A tres niveles:

(1) **Tipo de información (semántica) a extraer**: especificada a priori
Ej. fusiones empresariales

- Las formas de expresar un evento/información son limitadas
Ej. concepto de *"fusión"*
- Consecuentemente, se puede diseñar un método para identificarlos

(2) **Unidad de extracción**: no se devuelve el documento completo, sino frases simples (gen. *chunks*) 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

Clasificación y Estructuración (I): Clasificación

Objetivo: convertir la información no estructurada inicial en información estructurada lista para ser procesada

I. 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

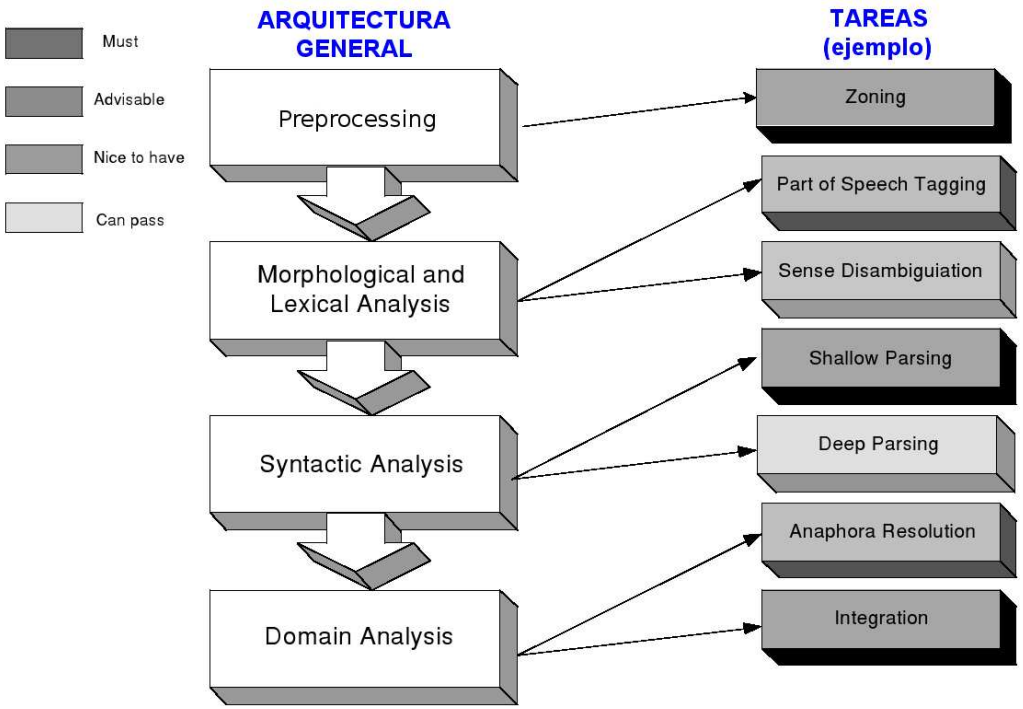
II. Estructuración

- 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/relación
- Objetivo: ir rellenando la plantilla mapeando en los diferentes slots la información contenida en el texto procesado

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

Arquitectura General

- Consenso en la **estructura general** de un sistema de IE ...
- ... pero no en las **tareas particulares** involucradas, muy variables entre sistemas
  - Opcionales vs. obligatorias
  - Fusionables: Ej. detectar eventos relevantes y a la vez generar su plantilla



- A grosso modo es una cascada de módulos que en cada paso ...
  i. agregan estructura al documento
  ii. filtran la información relevante

Preprocesamiento

- **Delimitador** (*text zoner*): dividir un texto en segmentos de texto (ej. en 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 (cuidado!!!, pérdida de información)
- **Desambiguación del sentido de la palabra** (*Word Sense Disambiguation, WSD*): en el caso de palabras polisémicas, identificar el significado/sentido concreto en ese contexto
- **Detección y análisis de entidades** (*entity recognition*):
  i. Entidades "con nombre" (*named entities*): nombres de personas, organizaciones, lugares, etc.
  ii. Expresiones temporales y numéricas

Análisis Sintáctico

- Simplifica las fases de extracción posteriores:
  - Los argumentos a extraer suelen corresponderse con los NPs.
  - Las relaciones entre argumentos a extraer suelen correponderse con las relaciones gramaticales funcionales existentes entre ellos.
- En ocasiones pueden aplicarse restricciones/información semántica propios del dominio durante el proceso de análisis para mejorar el análisis (ej. *adjunción PPs*), pero a expensas de perder generalidad

  `<POSITION> of <COMPANY>` $\Rightarrow$ $[_{NP}$ *vice president of Hupplewhite Inc.]*

Análisis Sintáctico (cont.)

Aproximaciones posibles:

(1) **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

(2) **Análisis sintáctico superficial** (*shallow parsing*; a.k.a. **chunking**, *partial parsing*):

- Devuelve una representación *"superficial"* (i.e. aproximativa, incompleta) de la estructura sintáctica del texto:
  - Opera en base a **grupos de palabras** o **chunks**
  - Plana, i.e. no contempla estructuras arborescentes
- Requerimientos menores
- Mayor robustez
- Bajo coste

Análisis del Dominio

- **Resolución de co-referencias**: identificar y resolver las expresiones que referencian al mismo objeto: anáforas, pronombres, etc.
  - Ej. *Barack Obama*, *Obama*, *el presidente Obama*, *el presidente norteamericano*, *el presidente de EE.UU.*, *el presidente* ...
  - *Sara ha comprado un piso. Ahora está reformándolo.*
- **Tratamiento de la elipsis** (i.e. omitir una o más palabras)
  - Ej. *Sara ha comprado un piso. Ahora [Sara] está reformándolo.*
- **Detección y análisis de relaciones/eventos:** identificar y clasificar los eventos y relaciones relevantes para el dominio presentes en el texto.
- **Generación de plantillas de salida**: volcar los elementos de información extraídos del texto al formato de salida deseado (infor. estructurada)
- **Combinación de resultados parciales**: diferentes oraciones/documentos pueden hablar del mismo suceso

Reconocimiento de Entidades

- a.k.a *entity recognition*
- **Objetivo:** identificar aquellas expresiones del texto (i.e. 1+ palabras) correspondientes a:
  - Entidades "con nombre" (*named entities*); i.e. **nombres propios** denotando personas, lugares, organizaciones, etc. (**named entity recognition**, *NER*)
  - Expresiones temporales
  - Expresiones numéricas (i.e. cantidades)

Reconocimiento de Entidades (cont.)

Citing high fuel prices, [$_{ORG}$ United Airlines] said [$_{TIME}$ Friday] it has increased fares by [$_{MONEY}$ $6] per round trip on flights to some cities also served by lower-cost carriers. [$_{ORG}$ American Airlines], a unit of [$_{ORG}$ AMR Corp.], immediately matched the move, spokesman [$_{PER}$ Tim Wagner] said. [$_{ORG}$ United], a unit of [$_{ORG}$ UAL Corp.], said the increase took effect [$_{TIME}$ Thursday] and applies to most routes where it competes against discount carriers, such as [$_{LOC}$ Chicago] to [$_{LOC}$ Dallas] and [$_{LOC}$ Denver] to [$_{LOC}$ San Francisco]

- Identificables por su forma de expresión, diferente de la del resto del texto (uso de abreviaturas de introducción, mayúsculas, etc.)
- El proceso consta de dos fases (como en *chunking*):
  (1) Detección
  (2) Clasificación

Esquemas de Anotación

- Definidos para las *Message Understanding Conferences (MUC)* y posteriormente adoptados por sistemas comerciales
- Emplean XML:
  - Un elemento por clase
  - Subcategorización en base al valor del atributo `TYPE`
- Entidades "con nombre": elemento `<ENAMEX>`; clasificadas en `TYPE`$=\{$`ORGANIZATION`, `PERSON`, `LOCATION`$\}$

```
<ENAMEX TYPE="PERSON">Clinton</ENAMEX> government
<ENAMEX TYPE="ORGANIZATION">Bridgestone Sports Co.</ENAMEX>
<ENAMEX TYPE="ORGANIZATION">European Community</ENAMEX>
<ENAMEX TYPE="ORGANIZATION">University of California</ENAMEX>
<ENAMEX TYPE="LOCATION">Los Angeles</ENAMEX>
```

NER: Esquemas de Anotación (cont.)

- Expresiones temporales: elemento <TIMEX>; clasificadas en TYPE={DATE, TIME}

```
<TIMEX TYPE="TIME">twelve o'clock noon</TIMEX>
<TIMEX TYPE="TIME">5 p.m.  EST</TIMEX>
<TIMEX TYPE="DATE">January 1990</TIMEX>
```

- Expresiones cuantitativas: elemento <NUMEX>; clasificadas en TYPE={MONEY, PERCENT}

```
<NUMEX TYPE="MONEY">20 million New Pesos</NUMEX>
<NUMEX TYPE="MONEY">$42.1 million</NUMEX>
<NUMEX TYPE="MONEY">million-dollar</NUMEX> conferences
<NUMEX TYPE="PERCENT">15 pct</NUMEX>
```

Named Entity Recognition (NER): Conceptos Básicos

- **Objetivo:** Identificar en el texto expresiones correspondientes a **entidades "con nombre"** (*named entities*); i.e. **nombres propios** denotando personas, lugares, organizaciones, etc.

Tipo	**Tag**	**Ejemplos**
Personas	PER	individuos, personajes, pequeños grupos
Organizaciones	ORG	compañías, agencias, partidos políticos, grupos religiosos, equipos deportivos
Lugares	LOC	montañas, lagos, mares
Entidades geopolíticas	GPE	países, estados, provincias, ciudades

- **Problema:** un mismo nombre puede referirse a entidades diferentes. Ejemplo: "JFK"
  - de igual tipo: presidente estadounidense, su hijo (problema de referencia)
  - de distinto tipo: persona (anteriores), aeropuerto NY, colegios, calles, etc.

NER: Procesamiento

Su procesamiento es muy similar al *chunking*:

- 2 fases:
  (1) Detección: delimitar el grupo de palabras que denotan la entidad
  (2) Clasificación
- Aproximaciones similares:
  I. Mediante aprendizaje automático (i.e. estadísticas)
  II. Mediante patrones (y heurísticas)

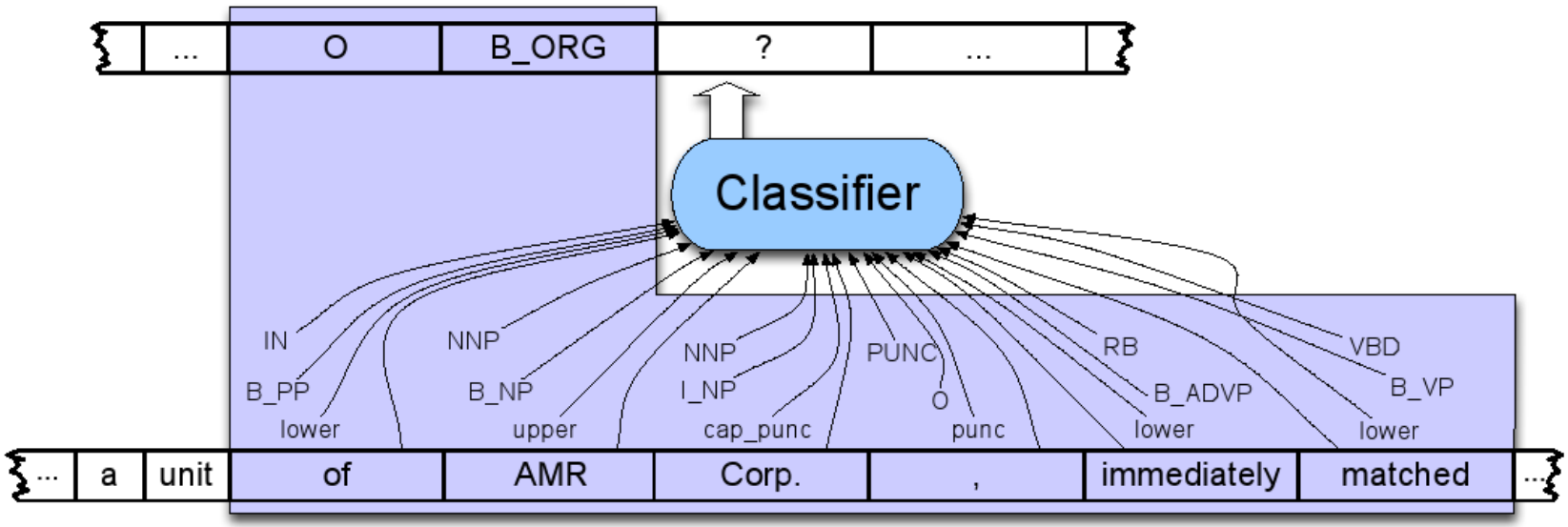
NER: Procesamiento (cont.)

En base a diversas características:

- La propia palabra
- Su stem, raíz o lema
- La grafía/forma de la palabra:
  - Sí/no empieza en mayúsculas: *"George"*
  - Todo mayúsculas: *"UGT"*
  - Alternancia de mayús. y minús: *"eBay"*
  - Inicial (mayúscula seguida por punto): *"H."*
  - Terminar en dígito: *"(DIN) A4"*
  - Contener guión: *"AP-9"*
- Etiqueta morfosintáctica
- Tipo del *chunk* (suelen corresponderse con NPs)
- Ocurrencia dentro de un *gazetteer* (listas/diccionarios especializados de nombres propios de personas, compañías, lugares, etc.)
- Presencia de palabras indicativas del tipo de entidad (*"empresa"*), cargo (*"director"*), título (*"Sr."*), abreviaturas comerciales (*"S.A."*), etc.
- Palabras del contexto

NER: Mediante Aprendizaje Automático

- IOB tagging con clasificadores secuenciales (p.ej. HMM): decidir si una palabra pertenece o no a la secuencia de palabras que denotan la entidad y, en caso afirmativo, su tipo:



NER: Mediante Aprendizaje Automático (cont.): Ejemplo

Features				Label
American	NNP	B_{NP}	cap	B_{ORG}
Airlines	NNPS	I_{NP}	cap	I_{ORG}
,	PUNC	O	punc	O
a	DT	B_{NP}	lower	O
unit	NN	I_{NP}	lower	O
of	IN	B_{PP}	lower	O
AMR	NNP	B_{NP}	upper	B_{ORG}
Corp.	NNP	I_{NP}	cap_punc	I_{ORG}
,	PUNC	O	punc	O
immediately	RB	B_{ADVP}	lower	O
matched	VBD	B_{VP}	lower	O
the	DT	B_{NP}	lower	O
move	NN	I_{NP}	lower	O
,	PUNC	O	punc	O
spokesman	NN	B_{NP}	lower	O
Tim	NNP	I_{NP}	cap	B_{PER}
Wagner	NNP	I_{NP}	cap	I_{PER}
said	VBD	B_{VP}	lower	O
.	PUNC	O	punc	O

NER: Mediante Patrones y Heurísticas (Mikheev et al., 1998)

En la práctica combina diversas técnicas (reglas, listas, téc. estadísticas, etc.) aplicadas en un determinado orden:

(1) **Aplicación de reglas I (seguras):** se aplican heurísticas basadas en el contexto y de altísima fiabilidad. Ejemplo:

$$\texttt{<SECUENCIA\_EN\_MAYÚSCULA>}_{\$1} \texttt{ , (<CARGO> | <PROFESIÓN>)} \Rightarrow [_{PERsona}\ \$1]$$

ej. $[_{PER}$ *John Smith*$]$, *director*

$$\texttt{<CARGO> of <SECUENCIA\_EN\_MAYÚSCULA>}_{\$1} \Rightarrow [_{ORGanizacion}\ \$1]$$

ej. *president* of $[_{ORG}$ *Microsoft Corporation*$]$

(2) **Gazetteers**: comprobamos candidatos en los *gazetteers* del sistema.

- Sólo se acepta si el contexto concuerda con el tipo propuesto. Ej.:

  . . . in the Washington area . . . → lugar, persona

  . . . Washington was born in . . . → lugar, persona

NER: Mediante Patrones y Heurísticas (cont.)

(3) **Correspondencias parciales I**: en 2 fases:

i. Derivamos patrones parciales a partir de las entidades ya reconocidas hasta el momento y buscamos ocurrencias de los mismos, marcándolas como candidatas. Ejemplo:

"Lockheed Martin Production" $\Rightarrow$ { *"Lockheed Martin Production"*, *"Lockheed Martin Production"* (en otras posiciones ambiguas), *"Lockheed Martin"*,*"Lockheed Production"*, *"Martin Production"*, *"Lockheed"*, *"Martin"* }

ii. Cada ocurrencia candidata es chequeda contra un clasificador estadístico, si éste la acepta, la confirmamos como entidad válida.

(4) **Aplicación de reglas II**: similar a (1), pero las restricciones han sido relajadas para dejar de considerar posibles ambigüedades que ya habrían logrado resolverse en pasos anteriores. Ejemplo (supuesto):

- Ocurrencia de *"Philip Morris"* en una posición inicialmente ambigua que permitía que fuese tanto *persona* como *organización*.
- Sólo aparece una vez en el texto, luego no hemos podido aplicar (3)
- Sin embargo, tal donde está, resultaría que si hubiese sido una *organización*, ya hubiera sido identificada como tal en los pasos anteriores.
- Eso implica que, por eliminación, sólo puede ser una *persona*.

NER: Mediante Patrones y Heurísticas (cont.)

(5) **Correspondencias parciales II**: similar a (3), pero partiendo de las nuevas entidades que hayamos reconocido desde entonces.

(6) **Procesamiento de títulos**: similar a (3) y (5), pero actuando únicamente sobre el título, que al estar todo en mayúsculas debe ser procesado sin diferenciar entre mayús. y minús.

NER: Ambigüedad en Palabras en Mayúscula

- **Problema:** posiciones en las que se emplean mayúsculas: comienzo de oración, listas enumeradas, etc.

"(...). Bush said (...)" { nombre propio / arbusto } *"(...). Rosa dijo (...)"* { nombre propio / flor, color (nombre) / color (adjetivo) }

- **Soluciones:**
  - Emplear un etiquetador (PoS tagger) para filtrar (reduce error ∼2%):
    - Problema: no resuelve coincidencias de nombres propios y comunes:

"(...). Bush said (...)" { nombre propio / arbusto } *"(...). Rosa dijo (...)"* { nombre propio / flor, color (nombre) }

  - Emplear co-referencias: un nombre ambiguo probablemente haya sido usado anteriormente en el texto de forma no ambigua.

The former president Bush (...). Bush said (...)

  - Ídem, pero buscando subcadenas de la entidad original (*"sequence strategy"*):

(...) by Rocket Systems Co. (...). Rocket Co. (...)

Expresiones Cuantitativas y Temporales

```
<NUMEX TYPE="MONEY">20 million New Pesos</NUMEX>
<NUMEX TYPE="MONEY">$42.1 million</NUMEX>
<NUMEX TYPE="MONEY">million-dollar</NUMEX> conferences
<NUMEX TYPE="PERCENT">15 pct</NUMEX>

<TIMEX TYPE="TIME">twelve o'clock noon</TIMEX>
<TIMEX TYPE="TIME">5 p.m.  EST</TIMEX>
<TIMEX TYPE="DATE">January 1990</TIMEX>
```

- Aproximaciones similares a NER (patrones+heurísticas o aprend. máquina) en base a:
  - La propia palabra
  - Las palabras contiguas
  - Grafía: p.ej. presencia de símbolos (*"$"*, *"%"*) o dígitos (*"2009"*)
  - Etiqueta morfosintáctica de la palabra y sus contiguas
  - Chunk-tag de las mismas
  - Presencia de indicadores léxicos (i.e. términos temporales/cuantitativos): *"euros"*, *"millón"*, *"junio"*, *"lunes"*, *"p.m."*, *"o'clock"*, etc.
- Su procesamiento completo (opcional) requeriría su **normalización**:

  mil cuatro euros → 1004 EUR *once de la mañana* → 11:00:00

Expresiones Temporales

- 3 tipos:
  i. Absolutas: indican un instante del tiempo de forma explícita
     1 de enero de 2010, verano del 77, 10:15 am, dos de la tarde
  ii. Relativas: indican un instante del tiempo en relación a otro
     ayer, la próxima semana, hace tres días, dentro de 15 min.
  iii. Duración: indican un intervalo de tiempo (granularidad variable)
     tres horas, 2 semanas, 10 min.

- Fácilmente identificables:
  - Formas de expresión más o menos acotadas
  - Presencia de indicadores léxicos (i.e. términos temporales):

sustantivos:	[*"enero"*, *"lunes"*], *"mañana"*, *"mediodía"*, *"víspera"*
propios:	[*"January"*, *"Monday"*], *"Semana Santa"*, *"Navidad"*
adjetivos:	*"anual"*, *"pasado"*, *"actual"*
adverbios:	*"anualmente"*, *"diariamente"*, *"hoy"*, *"ayer"*

Expresiones Temporales: Normalización

- Def.: mapear la expresión temporal original a:
  i. Un punto específico del tiempo (fecha y/o hora).
  ii. Duración (pudiendo incluir instantes de inicio y fin).
- Se emplea el estándar ISO 8601:

Unidad	**Formato**	**Ejemlo**
Fechas coml. especificadas	YYYY-MM-DD	2009-11-18
Hora (24h.)	HH:MM:SS	14:37:45
Fecha y hora	YYYY-MM-DDTHH:MM:SS	2009-11-18T14:37:45

- Mediante patrones para identificar cada componente de la entrada y mapearlo a su componente de salida:

  `FQTE` $\longrightarrow$ `<DIA>` de `<MES>` de `<AÑO>` $\quad \{$`AÑO.val` $-$ `MES.val` $-$ `DIA.val`$\}$

  - En **expresiones relativas**, se necesita el punto de **referencia temporal** (*temporal anchor*) para calcular la fecha/hora absoluta en base a él:

    expr. relativa: *"ayer"*
    anchor (fecha publicación): 18-11-2009
    $\Big\}$ $\longrightarrow$ fecha absoluta: 17-11-2009

Detección y Clasificación de Relaciones: Intro.

Doble objetivo:

(1) **Identificar las relaciones relevantes para el dominio** existentes entre las entidades contenidas en el texto.

- Deben haber sido **especificadas a priori**
- Son dependientes del dominio
- 2 subtareas:
  i. Detección
  ii. Clasificación

Relaciones		**Ejemplos**	**Tipo**
AFILIACIÓN:	Personal	*casado con*, *madre de*	PER → PER
	Organizacional	*portavoz de*, *presidente de*	PER → ORG
	"Artefactual"	*propietario de*, *fabricante de*	(PER|ORG) → ART
GEOESPACIAL:	Direccional	*al noroeste de*	LOC → LOC
PARTE-DE:	Organizacional	*sucursal de*, *matriz de*	ORG → ORG

(2) Generar su **representación** (p.ej. tuplas, proposiciones lógicas)

$$\text{FABRICANTE\_DE} \begin{bmatrix} \text{FABRICANTE:} & \textit{Nintendo} \\ \text{PRODUCTO:} & \textit{Wii} \end{bmatrix} \qquad \texttt{FABRICANTE\_DE}\ (\textit{Nintendo}, \textit{Wii})$$

Detección y Clasificación de Relaciones: Intro. (cont.)

- **Aproximaciones:**
  I. Mediante aprendizaje automático (i.e. estadísticas)
  II. Mediante patrones (y heurísticas)

Mediante Aprendizaje Automático

- Preciso **corpus de entrenamiento** anotado a mano indicando:
  (1) Argumentos (entidades) relacionados entre sí
  (2) Tipo de la relación
  (3) Rol (semántico) de cada uno dentro de la relación

- El proceso involucra 2 subtareas:
  (1) Detectar la existencia de la relación
  (2) Clasificarla (i.e. identificar su tipo)

Mediante Aprendizaje Automático: Detección

- Mediante **clasificadores binarios** que deciden si 2 entidades del texto están relacionadas
- Entrenados sobre el corpus de entrenamiento:
  - Ejemplos positivos: los marcados en el corpus. Ejemplo:

    En 2006 Nintendo lanzó la Wii, y Sony la PlayStation 3.
  - Ejemplos negativos: pares de entidades del corpus que están dentro de la misma oración pero que NO están relacionados. Ejemplo:

    En 2006 Nintendo lanzó la Wii, y Sony la PlayStation 3.

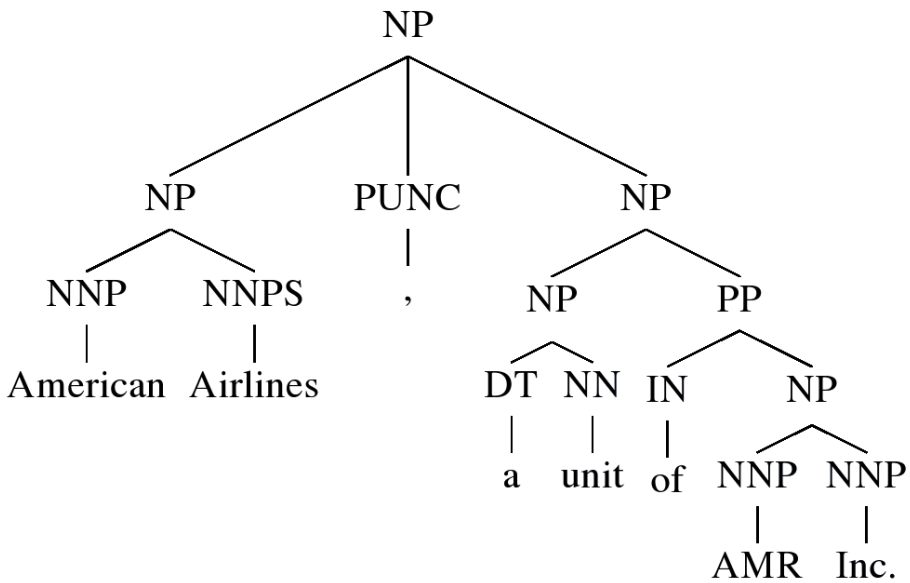
Mediante Aprendizaje Automático: Clasificación

- **Objetivo: identificar el tipo de la relación** detectada en la fase anterior.
- Mediante **algoritmos de clasificación** (árboles de decisión, bayesianos, de máxima entropía, ...)
- Factores/características en base a los cuales hacer la clasificación. 3 fuentes:
  1. **Los argumentos (entidades) relacionados**:
     - Su tipo, de cada una y en conjunto. Ej. una relación `president_of` se puede establecer entre PER→ORG, pero no al revés (ORG→PER), ni tampoco entre LOC→ORG.
     - Sus núcleos.
     - *Bag_of_words* (i.e. conjunto de palabras) que los forman
  2. **El contexto**:
     - Palabras, *stems* y/o lemas entre/antes/después de las entidades (ventana)
     - *Bag_of_words* y *bag_of_bigrams* entre ellas (y/o de sus *stems* y/o lemas)
     - Distancia entre las dos entidades
     - Número de (otras) entidades entre ellas

Mediante Aprendizaje Automático: Clasificación (cont.)

III. **La estructura sintáctica** (la naturaleza de dicha información sintáctica variará según el tipo de *parsing* empleado): las relaciones suelen corresponderse con las **relaciones gramaticales** funcionales ya existentes entre las entidades:

- Presencia de determinadas construcciones y relaciones gramaticales
  - Id. qué tipo de construcciones sintácticas se corresponden con determinadas relaciones: empleando "detectores" construidos manualmente o aprendidos automáticamente:



Ej. construcción apositiva asociada a una relación `part_of`

Mediante Patrones

- Patrones (ej. expresiones regulares extendidas) correspondientes a las relaciones relevantes del dominio que queremos capturar e incluyendo:
  i. Las entidades relacionadas
  ii. Su contexto (léxico, sintáctico o semántico)

- Ejemplo: Identificar *hubs* de aerolíneas:

  `/* has a hub at */` $=\sim$ *[Delta]* `has a hub at` *[LaGuardia]*
  [Bulgaria Air] `has a hub at` *[Sofia Airport]*
  [American Airlines] `has a hub at` *[the San Juan airport]*

- **Posibles problemas** con los patrones:
  i. Falta de precisión
  ii. Falta de cobertura

Mediante Patrones: Falta de Precisión

- Ejemplo anterior:

`/* has a hub at */` $=\sim$ *[The catheter]* `has a hub at` *[the proximal end]**
Many times, [a star topology] `has a hub at` *[its center]**

- **Solución:** hacer el patrón más específico añadiendo restricciones; ej. sobre el tipo de entidades a relacionar

`/<ORG> has a hub at <LOC>/` $=\sim$ $[_{ORG}$ *Delta]* `has a hub at` $[_{LOC}$ *LaGuardia]*
$[_{ORG}$ *Bulgaria Air]* `has a hub at` $[_{LOC}$ *Sofia Airport]*
$[_{ORG}$ *American Airlines]* `has a hub at` $[_{LOC}$ *the San Juan*

$!\sim$ *The catheter has a hub at the proximal end*
Many times, a star topology has a hub at its center

Mediante Patrones: Falta de Cobertura

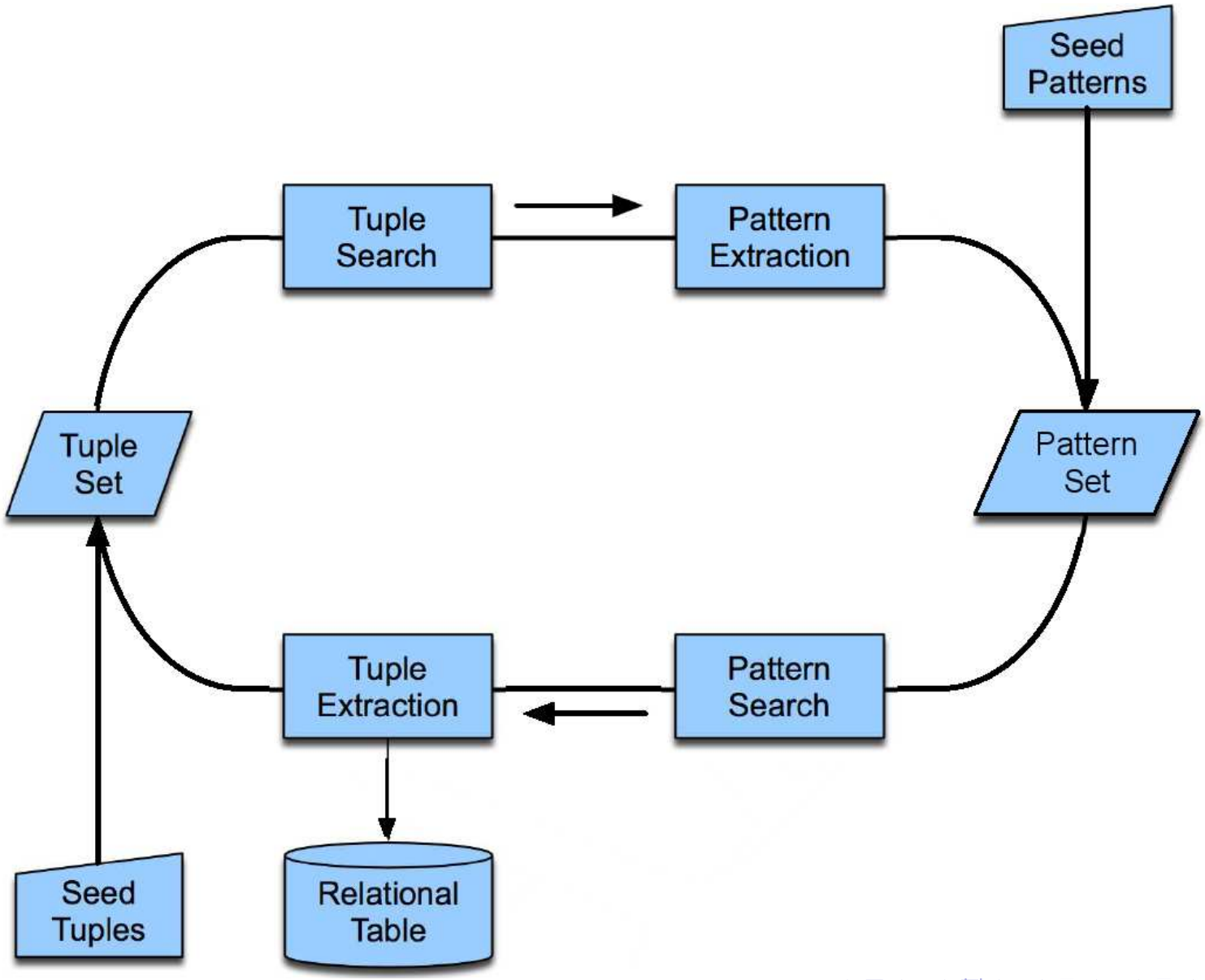
- Causa: **variación lingüística**

`/<ORG> has a hub at <LOC>/` !∼ *EasyJet, which has established a hub at Liverpool*
Ryanair also has a continental hub at Charleroi airbort

- Solución: incrementar la cobertura de los patrones
  - Relajando el patrón permitiendo *matchings* intermedios
    - Riesgo de introducir ruido

      `/<ORG> has a * hub at <LOC>/` =∼
      [ORG Dow Chemical] `has a chemical hub at` *[LOC West Bengal]**
  - Expandiendo el conjunto de patrones generando variantes de ellos ...
    I. Manualmente: costoso
    II. Automáticamente mediante *bootstrapping*

Mediante Patrones: Bootstrapping



Mediante Patrones: Bootstrapping

(1) Tomamos un par de entidades (*tupla*) que ya sepamos están relacionadas (*seed tuplas*)
- Ejemplo: sabemos que Ryanair tiene un *hub* en Charleroi:

$$\texttt{has\_hub\_at}(\textit{Ryanair}, \textit{Charleroi})$$

(2) Localizamos documentos (ej. con Google) que contengan, cerca unos de otros, los términos involucrados (*"Ryanair"*, *"Charleroi"* y *"hub"*), y buscamos oraciones que contengan la relación deseada:
- (a) *A Budget airline Ryanair, which uses Charleroi as a hub, scrapped all weekend flights out of the airport.*
- (b) *All flights in and out of Ryanair's Belgian hub at Charleroi airport were grounded on Friday.*
- (c) *A spokesman at Charleroi, a main hub for Ryanair, estimated that 8000 passengers had already been affected.*

(3) En base a esos textos generamos nuevos patrones que también capturan dicha relación
- (a) `/<ORG>, which uses <LOC> as a hub/`
- (b) `/<ORG>'s hub at <LOC>/`
- (c) `/<ORG> a main hub for <LOC>/`

Mediante Patrones: Bootstrapping (cont.)

(4) Mediante esos nuevos patrones generamos nuevas tuplas con entidades relacionadas

(a) $[_{ORG}$ *US Airways*$]$ `, which uses` $[_{LOC}$ *Pittsburgh*$]$ `as a hub` $\Rightarrow$
`has_hub_at`(*US Airways*, *Pittsburgh*)

(b) $[_{ORG}$ *Continental*$]$`'s hub at` $[_{LOC}$ *Cleveland*$]$ $\Rightarrow$
`has_hub_at`(*Continental*, *Cleveland*)

(c) $[_{LOC}$ *Minneapolis/St. Paul*$]$ `is a main hub for` $[_{ORG}$ *Northwest Airlines*$]$ $\Rightarrow$
`has_hub_at`(*Northwest Airlines*, *Minneapolis/St. Paul*)

(5) Volvemos a (1) tomando esta vez como entrada las nuevas tuplas generadas.

- El proceso puede también iniciarse en (3) si partimos de un patrón(es) inicial(es) (*seed patterns*) en lugar de tuplas

Mediante Patrones: Bootstrapping (cont.)

- **Problema: deriva semántica** (*semantic drift*): generar patrones erróneos que no correspondan a la relación deseada, lo cual generará tuplas erróneas, las cuales generarán más patrones erróneos, etc. (y viceversa):

  Sydney has a ferry hub at Circular Quay* $\Rightarrow$
  `/<LOC> has a ferry hub at <LOC>/` $\Rightarrow$
  [$_{LOC}$ Hamburg] has a ferry hub at *[$_{LOC}$ Landungsbrucken]* $\Rightarrow$
  `has_hub_at`(*Hamburg*, *Landungsbrucken*)

  - Necesario introducir mecanismos de comprobación de la fiabilidad de los patrones y tuplas generadas

Detección y Clasificación de Eventos: Intro.

Objetivo: Identificar los eventos relevantes para el dominio presentes en el texto

- Deben haber sido **especificados a priori**
- Son dependientes del dominio
- Indican estados y transiciones entre estados que pueden ser asignados a un determinado punto/intervalo de tiempo; ej. *"(...) la cotización de X subió un 25% (...)"*
- Por lo general **se corresponden con**:
  i. Verbos: ej. *"(...) increased (...)"*
  ii. Nombres de acción verbal: ej. *"(...) the increase of (...)"*
- 2 subtareas:
  (1) Detección
  (2) Clasificación

Detección y Clasificación de Eventos: Intro. (cont.)

[$_{EVENT}$ Citing] high fuel prices, United Airlines [$_{EVENT}$ said] Friday it has [$_{EVENT}$ increased] fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately [$_{EVENT}$ matched] [$_{EVENT}$ the move], spokesman Tim Wagner [$_{EVENT}$ said]. United, a unit of UAL Corp., [$_{EVENT}$ said] [$_{EVENT}$ the increase] took effect Thursday and [$_{EVENT}$ applies] to most routes where it [$_{EVENT}$ competes] against discount carriers, such as Chicago to Dallas and Denver to San Francisco

- **Aproximaciones:**
  I. Mediante aprendizaje automático (estadística)
  II. Mediante patrones

Mediante Aprendizaje Automático

- Preciso **corpus de entrenamiento**
- En base a diversas características:
  - La propia palabra
  - Su *stem*, raíz y/o lema
  - Prefijos y sufijos (ej. sufijos de nominalización: *"-ción"*)
  - Etiqueta morfosintáctica
  - Información semántica (WordNet): ej. hiperónimos
  - Tipo del sujeto

Mediante Patrones

- Patrones en base a:
  - Los tipos de las entidades implicadas
  - Los núcleos de los *chunks* implicados
  - Características anteriores (usadas para aprendizaje automático)

Mediante Patrones (cont.): Ejemplo (Hobbs et al., 1997)

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 clups 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 with production of 20,000 iron and "metal wood" clubs a month.

/<$_{NP}$ COMPANY(IES)>$_{\$1}$ <$_{VP}$ FORM>$_{\$2}$ <$_{NP}$ JOINT-VENTURE>$_{\$3}$ with <$_{NP}$ COMPANY(IES)>$_{\$4}$/ =~

[$_{NP}$ Bridgestone Sports Co.]$_{\$1}$ said Friday it [$_{VP}$ has set up]$_{\$2}$ [$_{NP}$ a joint venture]$_{\$3}$ in Taiwan with *[$_{NP}$ a local concern]$_{\$4.1}$ and [a Japanese trading house]$_{\$4.2}$ to produce golf clups to be shipped to Japan.*

/<$_{NP}$ COMPANY(IES)>$_{\$1}$ capitalized at <$_{NUMEX}$ CURRENCY>$_{\$2}$/ =~

The joint venture, [$_{NP}$ Bridgestone Sports Taiwan Co.]$_{\$1}$,* capitalized at *[$_{NUMEX}$ 20 million new Taiwan dollars]$_{\$2}$, will start production in (...)

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

Relationship:	**TIE-UP**
Entities:	**--**
JV Company:	**"Bridgestone Sports Taiwan Co."**
Capitalization:	**20000000 TWD**

Generación de Plantillas

- **Objetivo:** volcar los eventos y relaciones relevantes extraídos del texto al formato de salida deseado
  - Plantillas (*templates*) tipo *frame*.
- Puede ser necesario adaptar el elemento extraído al registro destino:
  Ej., registros con conjunto de valores predefinido, normalización de fechas/cantidades, etc.
- Umbral mínimo de "interés" del evento/relación: desechar eventos/relaciones incumplen
  (Ej. determinados campos sin rellenar)
- Aproximaciones:
  - Mediante aprendizaje automático (estadística)
  - Mediante patrones y heurísticas: ej. FASTUS (simultáneamente a la detección de relaciones/eventos)

Combinación de Resultados Parciales

- Diferentes textos (oraciones, documentos, ...) pueden referenciar el mismo evento/relación.
- Por lo tanto la información referida a él está distribuida entre ellos.
- Combinando dichas fuentes se obtiene una información más completa.
- ¿Sí/No combinar y cuáles? Decidimos en base a:
  - Estructura interna de los términos potencialmente relacionados
  - Proximidad
  - Compatibilidad y consistencia de ambas fuentes de información
- ¿Cuándo realizarla?

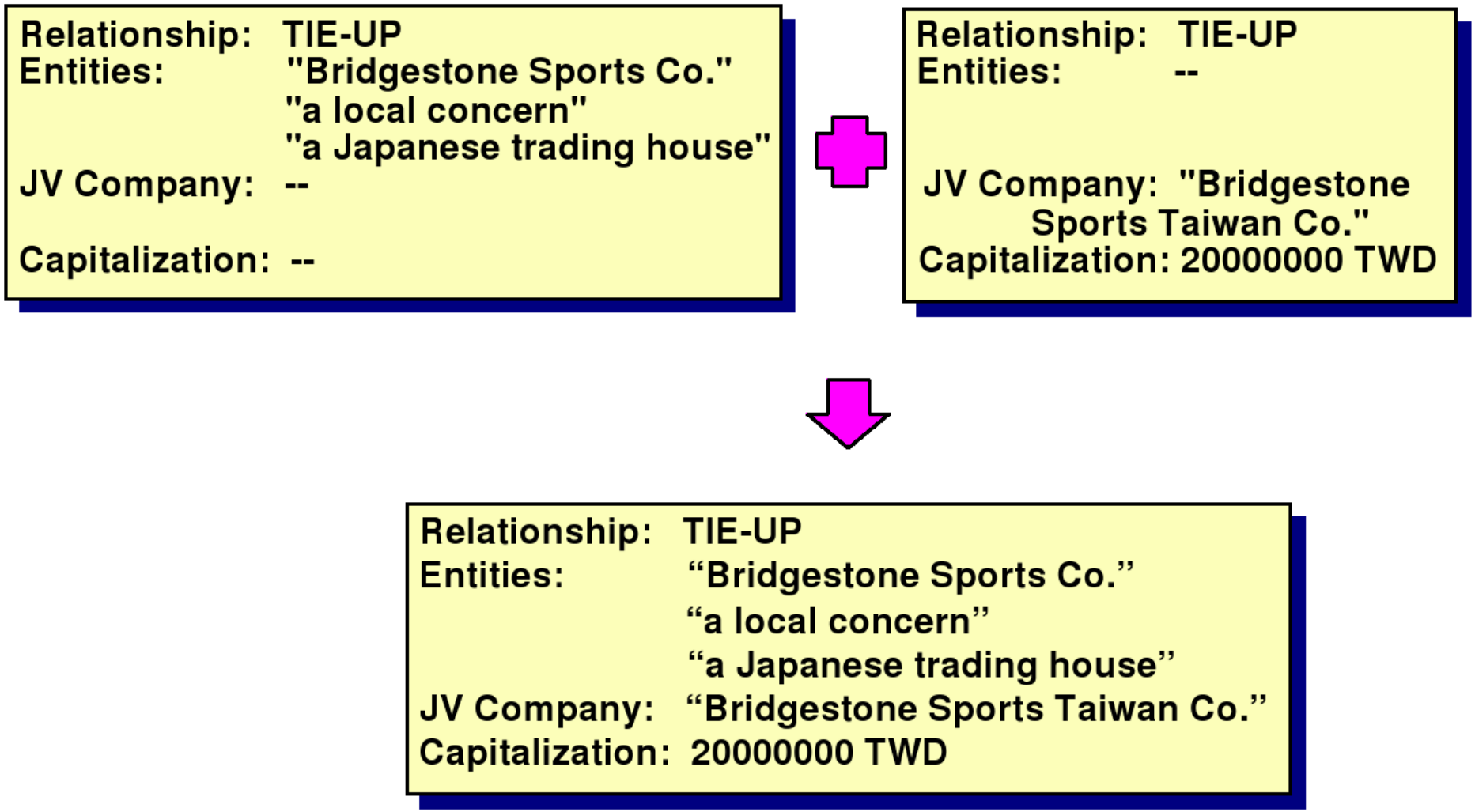
  I. Antes de generar la plantilla; ej. empleando reglas de producción

$$\texttt{start\_job}(\texttt{person}_X, \texttt{job}_Y) \ \& \ \texttt{succeed}(\texttt{person}_X, \texttt{person}_Z) \Rightarrow \texttt{leave\_job}(\texttt{person}_Z, \texttt{job}_Y)$$

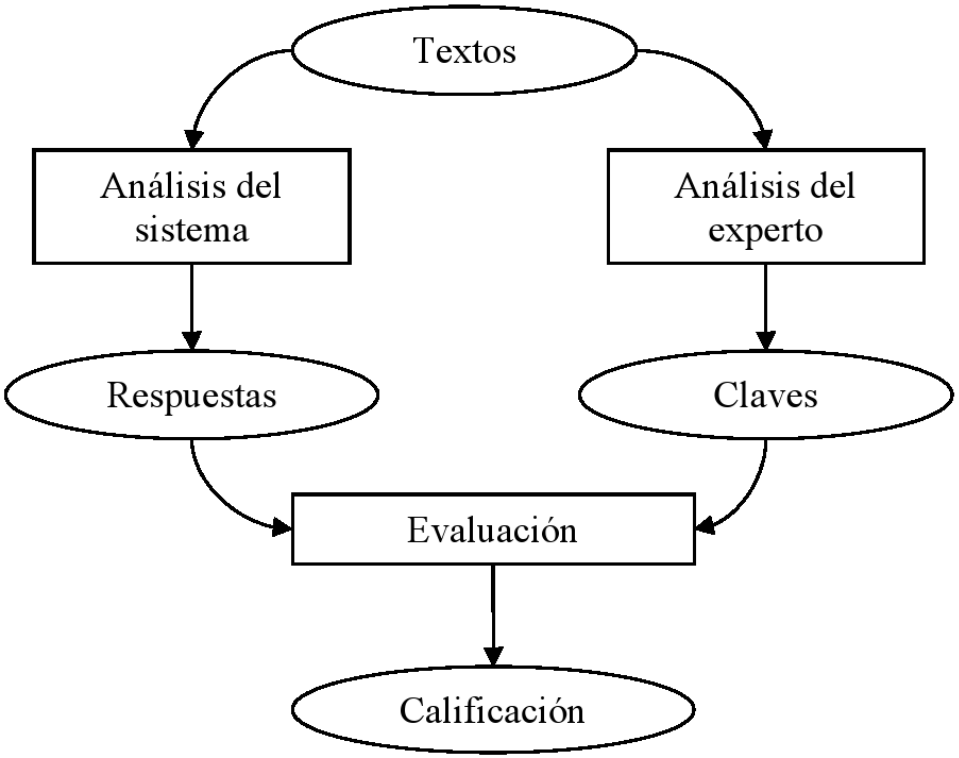
  Ej. *"Juan es ahora director. Sustituye a Pepe."* $\Rightarrow$ Pepe ha dejado de ser director.

  II. Después, combinando plantillas parciales

Comb. de Res. Parciales (cont.): Ejemplo Plantillas Parciales (Hobbs et al., 1997)



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
perdida: NO respuesta, SÍ clave
espúrea: SÍ respuesta, NO clave

Métricas de Evaluación (I): Basadas en Error

#claves=#correctas+#incorrectas+#parciales+#perdidas

#respuestas=#correctas+#incorrectas+#parciales+#espúreas

- **Error en respuestas (error per response fill)**: error "global" **(oficial)**

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

- **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 devueltas "cambiadas"

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

Métricas de Evaluación (II): "Clásicas"

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

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

Capacidad para extraer sólo registros correctos.

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

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

Capacidad para extraer todos los registros correctos.

- **Medida-F (F-measure):** pondera ambas conforme a un parámetro $\beta \in [0, \infty)$

$$F = \frac{(\beta^2 + 1) \times Re \times Pr}{Re + \beta^2\ Pr} \quad \text{con} \quad \beta^2 = \frac{1 - \alpha}{\alpha} \quad \text{y} \quad \alpha \in [0, 1]$$

Si β=1 (F_1) ambas se ponderan igual

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

MUC

- *Message Understanding Conference (MUC)* (`http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/muc.htm`)
  - *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**
- **Evolución y dominio de trabajo**:
  - MUCK-I..II: experimentos iniciales, muy limitados. Comunicaciones militares navales.
  - MUC-3..4: ya se requiere filtrado de información (infor. s/n relevante). Ataques terroristas.
  - MUC-5: disponibilidad de *gazetteers*. Métricas más completas. Fusiones de empresas y anuncios de productos de microelectrónica.
  - MUC-6: Nuevas tareas. Sucesiones en la dirección de empresas.
  - MUC-7: Nuevas tareas. Accidentes de avión; lanzamientos de cohetes/misiles.

MUC (cont.): Tareas de evaluación

- **"Extracción de información"**: proceso clásico-completo (tarea original)
- **Reconocimiento de entidades (entity recognition)***: encontrar y clasificar las entidades del texto
  - *Multilingual Entity Task (MET)*: ampliación a multilingüe: español, japonés y chino (`http://www-nlpir.nist.gov/related_projects/tipster/met.htm`)
- **Resolución de correferencias (co-reference)***: identificar las expresiones en el texto que hacen referencia al mismo objeto
- **Plantillas de escenario (scenario templates)***: readaptar tu sistema de IE a un nuevo dominio en 1 mes. Testea la flexibilidad y portabilidad del sistema.
- **Relación de plantillas (template relations)***: identificar relaciones entre plantillas. Ej. `empleado_de`, `localizado_en`, `producto_de`, etc.

(*) sólo en las últimas ediciones

FASTUS

- *Finite State Automata-Based Text Understanding System (FASTUS)* (`http://www.ai.sri.com/natural-language/projects/fastus.html`)
- Sistema clásico "de referencia" en EI
- **Cascada de traductures finitos no deterministas:** 5 etapas/niveles
  (1) **Términos complejos:** reconocimiento de expresiones multipalabra y propios
  (2) **Frases básicas:** reconocimiento de grupos nominales y verbales simples, y ciertas partículas de interés (ej. preposiciones)
  (3) **Frases complejas:** reconocimiento de grupos nominales y verbales complejos (ej. adjunción de PPs)
  (4) **Eventos/relas. del dominio:** búsqueda de correspondencias con patrones de eventos/relas. de interés y generación de su plantilla
  (5) **Combinación de estructuras:** combinación de la información sobre un mismo evento/rela. repartida entre diferentes plantillas

Referencias

- [FASTUS, n.d.] Finite State Automata-based Text Understanding System (FASTUS). Site: `http://www.ai.sri.com/natural-language/projects/fastus.html`
- [MUC, n.d.] Message Understanding Conference (MUC). Site: `http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/muc.htm`
- [Chinchor & Sundheim, 1993] Chinchor, N. & Sundheim, B. (1993). MUC-5 evaluation metrics. In *Proc. of the 5th Message Understanding Conference (MUC-5)*, pp. 69-78.
- [Grisham, 1997] Grishman, R. (1997). Information Extraction: Techniques and Challenges. In *Lecture Notes in Computer Science*, 1299:10–27. Springer-Verlag.
- [Hobbs, 1993] Hobbs, J.R. (1993). The Generic Information Extraction System. In *Proc. of the 5th Message Understanding Conference (MUC-5)*, pp. 87-91.
- [Hobbs et al., 1997] Hobbs, J.R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M. & Tyson, M. (1997). Chapter 13: FASTUS - A Cascaded Finite-State Transducer for Extracting Information from Natural-Language text. In *Finite-State Language Processing*. MIT Press. Available in (FASTUS, n.d.).

Referencias (cont.)

- [Jackson & Moulinier, 2007] Jackson, P. & Moulinier, I. (2007). Chapter 3: Information extraction & Chapter 5: Text mining. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization (2nd Revised Ed.)*. John Benjamins Publishing.

- [Jurafsky & Martin, 2009] Jurafsky, D. & Martin, J.H. (2009). Chapter 22: Information Extraction. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd ed.)*. Pearson–Prentice Hall.

- [Mikheev et al., 1998] Mikheev, A., Grover, C. & Moens, M. (1998). Description of the LTG system used for MUC-7. In *Proceedings of 7th Message Understanding Conference (MUC-7)*.

- [Moens, 2006] Moens, M.-F. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Springer.

- [Nugues, 2006] Nugues, P.M. (2006). Chapter 9: Partial Parsing. *An Introduction to Language Processing with Perl and Prolog*. Springer.