Extracción de información

- Área del Procesamiento del Lenguaje Natural
- Objetivo:
  - Convertir la información no estructurada contenida en los textos en datos estructurados:
    - Detectar y extraer automáticamente ciertas partes relevantes de un texto
    - Combinar la información encontrada en base a sus relaciones

Extracción de información

Ejemplo:

Paciente varón de 86 años con antecedentes de TBC pulmonar, refiere F no cuantificada desde hace dos semanas, así como tos esporádica productiva y perdida de peso.

Documento: *Informe de urgencias*

- Sexo: *hombre*
- Edad: *86*
- Antecedentes:
  Tuberculosis pulmonar
- Síntomas:
  - *Fiebre no cuantificada*
  - *Tos esporádica productiva*
  - *Perdida de peso*
- Tiempo: *dos semanas*

EI: tareas y subtareas

- Named Entity Recognition (NER): identificación de entidades
  - Detección:
    - *Spokesman Tim Wagner said* → [Tim Wagner]: ENTIDAD
  - Clasificación:
    - *Spokesman Tim Wagner said* → [Tim Wagner]: PERSONA

EI: tareas y subtareas

- Extracción de relaciones:

Relación	Tipo	Ejemplo
Physical-Located	PER-GPE	He was in Tennessee
Part-Whole-Subsidiary	ORG-ORG	XYZ, the parent company of ABC
Person-Social-Family	PER-PER	Yoko's husband John
Org-AFF-Founder	PER-ORG	Steve Jobs, co-founder of Apple

EI: tareas y subtareas

- Reconocimiento de expresiones temporales
- Extracción de sucesos
- Rellenado de plantillas

Reconocimiento de entidades

- Definiciones y objetivo
- Métodos:
  - Diccionarios
  - Reglas
  - Aprendizaje automático (Machine learning)

Reconocimiento de entidades(NER)

Definiciones y objetivo:

- Entidad nombrada (named entity): nombres propios (personas, organizaciones, lugares, etc.)
- Por extensión expresiones que identifican conceptos de una clase: enfermedades, medicamentos, proteínas, etc.
- Reconocimiento de entidades: identificar el segmento de texto que constituye la entidad y asignarle una clase.

NER: ejemplo (MUC-7)

- Ejemplo

<ENAMEX TYPE=„LOCATION“>Italy</ENAMEX>‘s business world was rocked by the announcement <TIMEX TYPE=„DATE“>last Thursday</TIMEX> that Mr. <ENAMEX TYPE=„PERSON“>Verdi</ENAMEX> would leave his job as vice-president of <ENAMEX TYPE=„ORGANIZATION“>Music Masters of Milan, Inc</ENAMEX> to become operations director of <ENAMEX TYPE=„ORGANIZATION“>Arthur Andersen</ENAMEX>.

Se delimita con etiquetas de la categoría de NER:

Milan es parte de una Organizacion

Arthur Andersen es una compañía

NER: Dificultades

- En muchos casos demasiado numerosas para recogerlas todas en diccionarios
- Cambian frecuentemente
- Aparecen en formas variadas
- A veces aparecen abreviadas

NER: fuentes de información

- Identificar un nombre propio (o concepto) y su clase depende de:
  - La estructura interna
    - Ej: Sr. Alvarez
  - El contexto:
    - His job as vice-president of Music Masters of Milan
      - Indicio de que Music Masters of Milan es una compañía

NER: Diccionarios

- Diccionarios o gazeteers: Un gazeteer es una lista de nombres, generalmente de lugares.
- Incluyen las distintas formas de nombrar una localización, junto con información política, geográfica y geológica detallada.
- Ejemplos:
  - https://gate.ac.uk/sale/tao/splitch13.html

NER: reglas

- Un método para NER es usar reglas o patrones que se ajustan al tipo de entidades buscadas.
- Suelen utilizarse expresiones regulares, que en muchos casos se construyen manualmente.
- Funciona bien cuando existen un patrón común a las entidades buscadas

NER: Expresiones regulares

- La expresiones regulares (regex) son un lenguaje para construir patrones a partir de otros más simples:
  - Un carácter es una regex
  - Regex como unión de dos regex: (e_1 | e_2)
  - Concatenación: e_1e_2
  - Clausura: e_1* (secuencia de 0 o más cadenas que encajan con e_1)
  - Etc.

NER: reglas

- Hay términos especiales para conjuntos de caracteres comunes: alfanuméricos, dígitos, etc.
- Hay operadores para representar operaciones frecuentes.

NER: expresiones regulares

- . Wildcard, matches any character
- ^abc Matches some pattern abc at the start of a string
- abc$ Matches some pattern abc at the end of a string
- [abc] Matches one of a set of characters
- [A-Z0-9] Matches one of a range of characters
- ed|ing|s Matches one of the specified strings (disjunction)
- * Zero or more of previous item, e.g. a*, [a-z]* (also known as Kleene Closure)
- + One or more of previous item, e.g. a+, [a-z]+
- ? Zero or one of the previous item (i.e. optional), e.g. a?, [a-z]?
- {n} Exactly n repeats where n is a non-negative integer
- {n,} At least n repeats
- {,n} No more than n repeats
- {m,n} At least m and no more than n repeats
- a(b|c)+ Parentheses that indicate the scope of the operators
- VER REFERENCIAS para más casos!

NER: Reglas

- Ejemplos de patrones que podrían capturar reglas formuladas con exp. regulares:
- Rasgo Ejemplo Idea

  núm. 4 dígitos 2019 año de 4 dig.

  Todo mayúsculas ONU organización

  Mayúscula . Inicial persona L.

  Dígitos y letras X2345 código de producto

  etc..

NER: Aprendizaje automático (ML)

- Entrenamiento
  - Recoger un conjunto de documentos de entrenamiento representativo
  - Etiquetar cada token (palabra/signo) con la clase de su entidad o con *other* (O)
  - Diseñar los rasgos apropiados para caracterizar el texto y las clases
  - Entrenar un clasificador secuencial para predecir las etiquetas de los datos

NER: Aprendizaje automático (ML)

- Predicción/evaluación
  - Recibir un conjunto de documentos a etiquetar
  - Ejecutar el sistema de ML entrenado en la fase anterior para predecir la etiqueta de cada token
  - Producir una salida con las entidades reconocidas en un formato legible

NER: codificación de las clases

- Sistemas de etiquetado:
  - IOB:
    - I: Inside
    - B: Begin
    - O:other
  - IO:
    - I: Inside:
    - O: Other
  - Existen otros: BILOU, etc.

NER: codificación de las clases

Words	IOB Label	IO Label
American	B-ORG	I-ORG
Airlines	I-ORG	I-ORG
,	O	O
a	O	O
unit	O	O
of	O	O
AMR	B-ORG	I-ORG
Corp.	I-ORG	I-ORG
,	O	O
immediately	O	O
matched	O	O
the	O	O
move	O	O
,	O	O
spokesman	O	O
Tim	B-PER	I-PER
Wagner	I-PER	I-PER
said	O	O
.	O	O

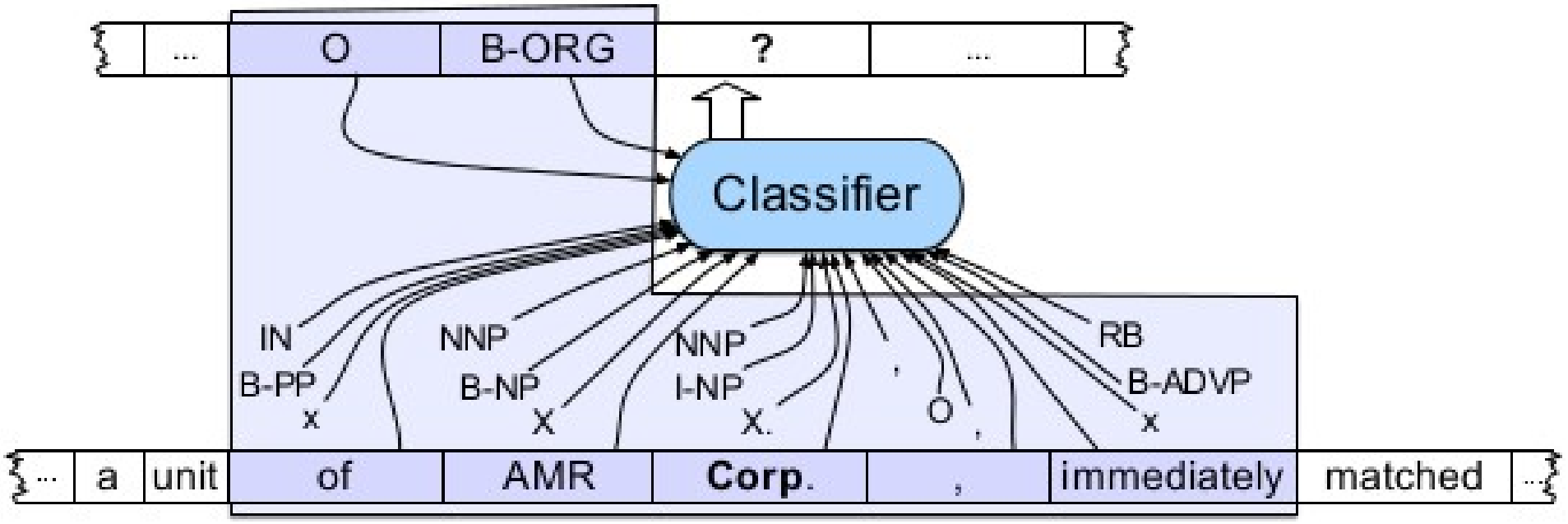
NER: rasgos típicos clasificador

- Palabra w_i
- Palabras vecinas
- POS tag de w_i
- POS tags de las palabras vecinas
- Contiene un determinado prefijo
- Contiene un determinado subfijo
- Empieza por mayúscula
- Etc.

NER: ejemplo

Word	POS	Chunk	Short shape	Label
American	NNP	B-NP	Xx	B-ORG
Airlines	NNPS	I-NP	Xx	I-ORG
,	,	O	,	O
a	DT	B-NP	x	O
unit	NN	I-NP	x	O
of	IN	B-PP	x	O
AMR	NNP	B-NP	X	B-ORG
Corp.	NNP	I-NP	Xx.	I-ORG
,	,	O	,	O
immediately	RB	B-ADVP	x	O
matched	VBD	B-VP	x	O
the	DT	B-NP	x	O
move	NN	I-NP	x	O
,	,	O	,	O
spokesman	NN	B-NP	x	O
Tim	NNP	I-NP	Xx	B-PER
Wagner	NNP	I-NP	Xx	I-PER
said	VBD	B-VP	x	O
.	,	O	.	O

NER: ejemplo


...
O
B-ORG
?
...
Classifier
IN
B-PP
x
NNP
B-NP
X
NNP
I-NP
X.
,
O
,
RB
B-ADVP
x
...
a
unit
of
AMR
Corp.
,
immediately
matched
...

NER in nltk

- NLTK proporciona un clasificador que ya ha sido entrenado para reconocer nombres propios. Se accede con la función nltk.ne_chunk().
- Si fijamos el parámetro binary=True, entonces las entidades se etiquetan como NE; en otro caso el clasificador añade la etiqueta de categoría como por ejemplo PERSON, ORGANIZATION, and GPE.

NER en nltk

```
doc = '''Andrew Yan-Tak Ng is a Chinese American computer scientist. He is the former chief scientist at Baidu,  where he led the company's Artificial Intelligence Group. He is an adjunct professor (formerly associate professor) at Stanford University. Ng is also the co-founder and chairman at Coursera, an online education platform. Andrew was born in the UK in 1976. His parents were both from Hong Kong.'''

# sentence tokenizer

nltk.download('punkt')

tokenized_doc = nltk.word_tokenize(doc)


nltk.download('averaged_perceptron_tagger')

nltk.download('maxent_ne_chunker')

nltk.download('words')

# tag sentences and use nltk's Named Entity Chunker

tagged_sentences = nltk.pos_tag(tokenized_doc)

ne_chunked_sents = nltk.ne_chunk(tagged_sentences)
```

NER en nltk

```
# extract all named entities
named_entities = []
for tagged_tree in ne_chunked_sents:
    if hasattr(tagged_tree, 'label'):
        entity_name = ' '.join(c[0] for c in tagged_tree.leaves()) #
        entity_type = tagged_tree.label() # get NE category
        named_entities.append((entity_name, entity_type))
print(named_entities)
```

NER en Spacy

```
modeloEs = spacy.load('es')
doc = “…..”
docEsAnotado = modeloEs(doc)
# imprimer las NER encontradas
for ent in docEsAnotado.ents:
      print(ent.text,ent.label)
# imprime informacion de los tokens relativa a NER
for token in docEsAnotado:
   print (token.text, " ", token.pos_, " ",
   token.ent_iob_+"-"+token.ent_type_ )
```

NER en Spacy

Información sobre entidades en los tokens (https://spacy.io/api/token):

- ent_type_ : nombre del tipo de la entidad: ORG, PER, MISC, LOC
- ent_iob_: código IOB de la entidad. “B” significa token inicial de la NER, “I” significa que están dentro de la NER, “O” significa que está fuera, y "" significa que no se ha asignado etiqueta de entidad (ent_type).

NER: evaluación

- Competiciones que han tratado NER:
  - Conll 2002 (Spanish, Dutch)

    Language-Independent Named Entity Recognition (I)

    https://www.clips.uantwerpen.be/conll2002/ner/
  - Conll 2003 (English, German)

    Language-Independent Named Entity Recognition (II)

    https://www.clips.uantwerpen.be/conll2003/ner/

NER: Evaluación

- Se evalua sobre un conjunto de datos (test) no usados en el desarrollo del sistema (entrenamiento).
- Medida por cada documento de test:
  - Número total de entidades en el documento de referencia:N
  - Número total de entidades extraidas por el sistema: E
  - Número de entidades extraídas que son correctas: *C*
- Cálculo de a métricas más usuales:
  - Cobertura (Recall) = *C*/*N*
  - Precision = *C*/*E*
  - *Medida-F (*F-Measure: media armónica de Prec. y recall):

    2 Precision x Cobertura /(Precision + Cobertura)

NER: Evaluación

- Consideraciones:
  - La unidad de evaluación es la entidad, no la palabra: “Tim Wagner” cuenta como una única respuesta.
  - Este componente de segmentación que no está presente, por ejemplo en el POS tagging causa algunos problemas en la evaluación.
  - Por ejemplo, un sistema que etiquetara a *American* pero no a *American Airlines* como una organización causaría dos errores, un falso positivo para O y un falso negativo para I-ORG.
  - Además, si se usan entidades como unidad de respuesta pero de palabras como unidad de entrenamiento se produce un desajuste entre las condiciones de entrenamiento y de test.

NER: Evaluación

- Validación cruzada (10-fold cross-validation): resultados promediados sobre 10 pruebas con distintas selecciones del conjunto de entrenamiento y test (aunque siempre disjuntos).
- Los parámetros del modelo pueden favorecer la cobertura o la precisión.

NER: Deep learning

- Suelen utilizar embeddings de palabras para representar la entrada:
- Word embeddings es el nombre de un conjunto de técnicas para construir representaciones vectoriales del significado de las palabras de un idioma, de forma que la dimensión de los vectores es mucho menor que la cantidad de palabras del idioma.
- Su generación se basa en la hipótesis distribucional: las palabras que aparecen en contextos similares tienden a tener significados similares.

NER: Deep learning

- Los vectores resultantes tienen representaciones cercanas en el espacio → Se puede utilizar el álgebra de vectores para encontrar propiedades semánticas:

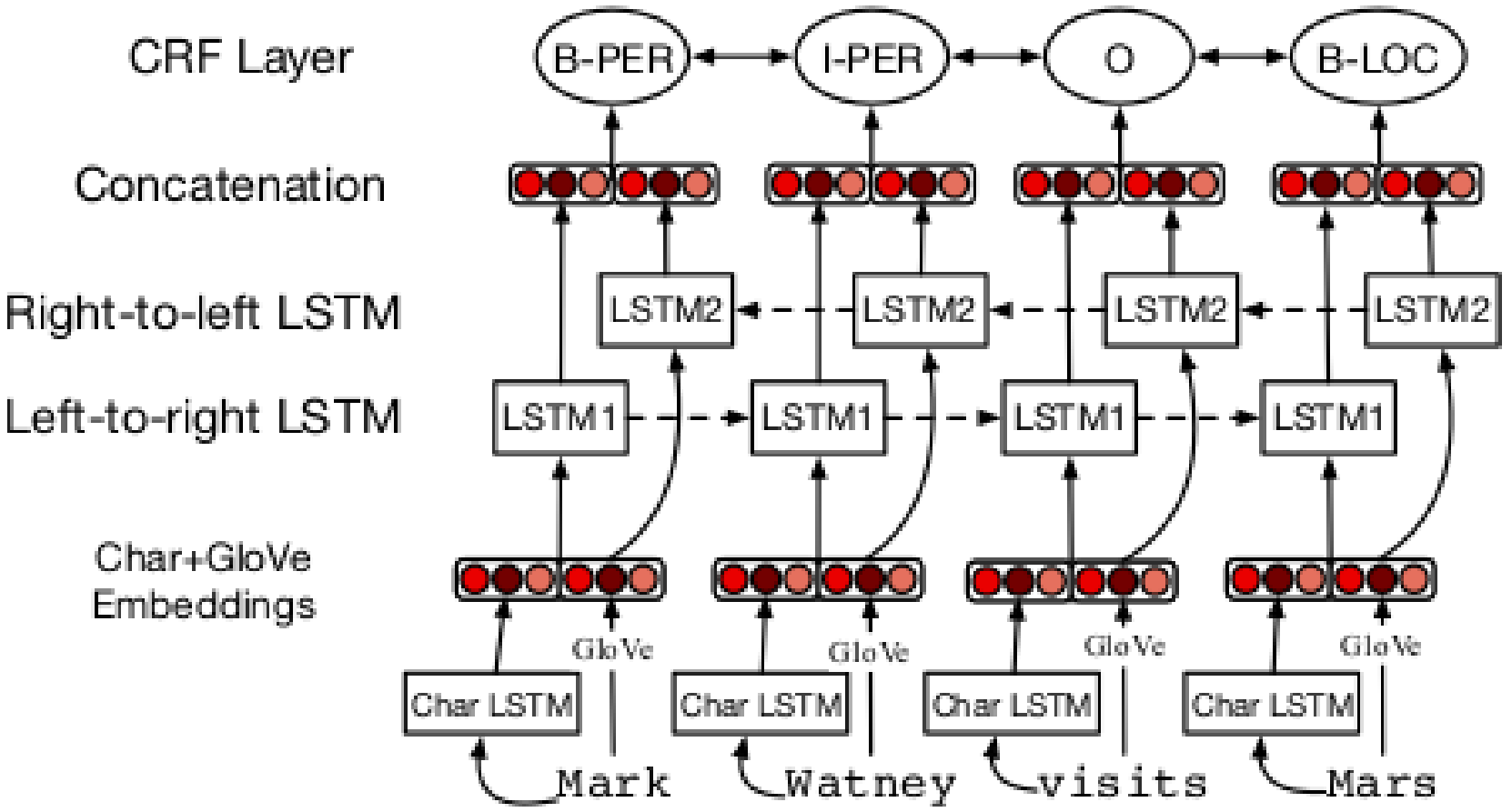
  Ej: rey + mujer – hombre = reina

- Se construyen con técnicas basadas en contadores de palabras o con redes neuronales (word2vec, Glove)

NER: Deep learning

- Los sistemas DL de detección de entidades usan típicamente Bi-LSTM (Long Short Term Memory).
- La entrada son los embedding de palabras y caracteres de las palabras de entrada.
- Producen una probabilidad i para cada posible etiqueta asignable a cada palabra.
- Pueden mejorarse concatenando layers que mejoren la salida.

NER: Deep learning



NER: Deep learning

- Código ejemplo Github:
- POS tagging: https://github.com/UKPLab/deeplearning4nlp-tutorial/tree/master/2016-11_Seminar/Session%201%20-%20SENNA/code%20for%20POS
- NER: https://github.com/UKPLab/deeplearning4nlp-tutorial/tree/master/2016-11_Seminar/Session%201%20-%20SENNA/code%20for%20NER

Reconocimiento de entidades

- Algunos corpus con entidades anotadas:
  - The CoNLL 2003 corpus (inglés)

    https://www.clips.uantwerpen.be/conll2003/ner/ (annotations) and NIST (text).
  - CoNLL 2002 a traves de python NLTK (español).
  - Corpus del dominio biomédico:
    - ADE: reacciones adversas a medicamentos https://github.com/trunghlt/AdverseDrugReaction/tree/master/ADE-Corpus-V2

Extracción de relaciones

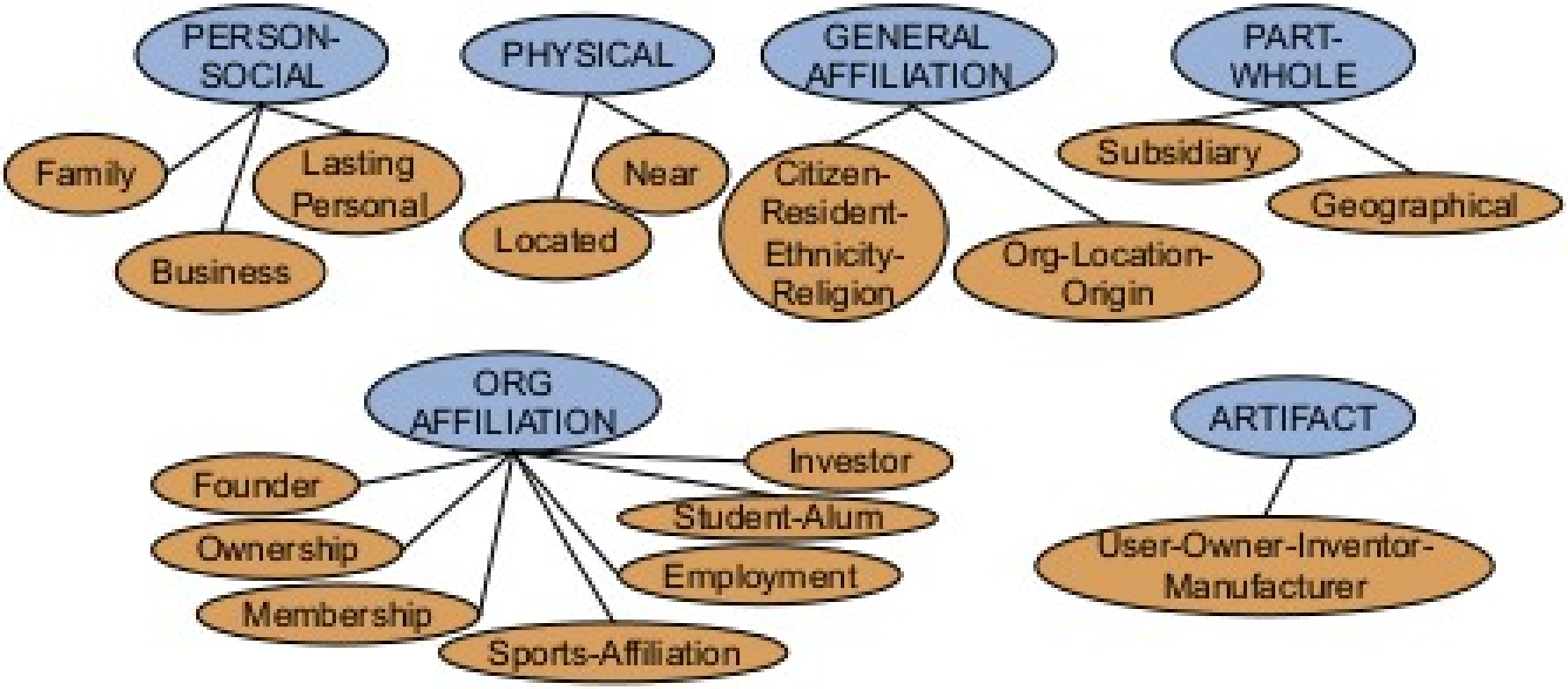
- Objetivo: Identificar la existencia de relaciones entre ciertos tipos de entidades:
  - Parte de (part-of)
  - PERSONA Presidente-de ORGANIZACIÓN
  - MEDICAMENTO Cura ENFERMEDAD
  - MEDICAMENTO provoca EFECTO ADVERSO (ENFERMEDAD)
- Generalmente una relación es un conjunto ordenado de tuplas sobre los elementos de un dominio.

Extracción de relaciones

- “La UNED es la mayor universidad de España, con sus más de 250.000 estudiantes que cursan sus titulaciones oficiales (27 grados, 65 másteres universitarios, ...”
- Relaciones:
  - UNED ↔ Universidad (relación “es”)
  - UNED ↔ España (relación LOCALIZACIÓN)
  - UNED ↔ > 250.000 (relación NUM ESTUDIANTES)
  - ...

Extracción de relaciones

- Relaciones de la competición *ACE relation extraction task*:



Extracción de relaciones

- Dominio biomédico: UMLS (Unified Medical Language System) de la Biblioteca Nacional de USA, red con 134 categorías mayores, y tipos de entidades y 54 relaciones, como:
  - Entidad Relación Entidad
  - Injury disrupts Physiological Function
  - Bodily Location location-of Biologic Function
  - Anatomical Structure part-of Organism
  - Pharmacologic Substance causes Pathological Function
  - Pharmacologic Substance treats Pathologic Function

Extracción de relaciones

- Infoboxes de Wikipedia: tablas estructuradas asociadas con los artículos de Wikipedia que proporcionan una gran cantidad de relaciones.
  - La infobox para Stanford incluye state=”california”, etc.
  - Los hechos se pueden convertir en relaciones:

    president-of
  - O en relaciones del metalenguaje RDF (Resourde Descripcion Framework): una tripleta RDF es una tupla entidad-relación-entidad.
- Otras fuentes de relaciones son las ontologías. Por ejemplo Wordnet tiene la relación is-a o hiperonimia entre clases.

Extracción de relaciones

- Ejemplo: Corpus ADE (Medicamentos- Efectos adversos)
- Formato:
  - Columna-1: Identificador del art. científico (PubMed-ID)
  - Columna-2: Oración
  - Columna-3: Efecto adverso
  - Columna-4: Begin offset of Adverse-Effect at 'document level'
  - Columna-5: End offset of Adverse-Effect at 'document level'
  - Columna-6: Medicamento
  - Columna-7: Begin offset of Drug at 'document level'
  - Columna-8: End offset of Drug at 'document level'

Corpus ADE

ID medline

Oración

9934637|In this article, we present the case of a vitiligo patient who was admitted to our facility with an intense burn after the topical use of 8-methoxypsoralen solution as a suntanning agent.|intense burn|771|783|8-methoxypsoralen|809|826

Medicamento

Reacción adversa

Extracción de relaciones

- Ejemplo: Corpus ADE (Medicamentos- Efectos adversos)
- 10030778|Intravenous azithromycin-induced ototoxicity.|ototoxicity|43|54| azithromycin|22|34
- 10048291|Immobilization, while Paget's bone disease was present, and perhaps enhanced activation of dihydrotachysterol by rifampicin, could have led to increased calcium-release into the circulation.|increased calcium-release|960| 985|dihydrotachysterol|908|926
- 10048291|Unaccountable severe hypercalcemia in a patient treated for hypoparathyroidism with dihydrotachysterol.|hypercalcemia|31|44| dihydrotachysterol|94|112
- 10082597|METHODS: We report two cases of pseudoporphyria caused by naproxen and oxaprozin.|pseudoporphyria|620|635|naproxen|646|654
- 10082597|METHODS: We report two cases of pseudoporphyria caused by naproxen and oxaprozin.|pseudoporphyria|620|635|oxaprozin|659|668

Extracción de relaciones

- Corpus de datos anotados con relaciones permiten:
  - Desarrollar sistemas (entrenamiento, etc.)
  - Evaluarlos
  - Compararlos

Extracción de relaciones

- Medidas de evaluación:
  - Precisión
  - Cobertura
  - Medida-F

Principales técnicas

- Patrones
- Aprendizaje automático supervisado
- Técnicas semisupervisadas
- Técnicas no supervisadas

Patrones lexico-sintácticos

- Ejemplo (Hearst 1992: Automatic Acquisition of Hyponyms (IS-A):
  - Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use
  - X → Gelidium (sub-type)
  - Y → red algae (super-type)
  - X → IS-A → Y
  - Ejemplos de patrones:

    “Y such as X”

    “Y, such as X”

    “X or other Y”

    “X and other Y”

    “Y including X”

Patrones lexico-sintácticos

Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	Bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
Such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y , especially X	European countries, especially France, England, and Spain...

Patrones lexico-sintácticos

- Construcción manual de patrones
  - Ventajas
    - Suelen tener una precisión alta
    - Pueden adaptarse a dominio específicos
  - Desventajas
    - Baja cobertura (difícil pensar en todos los patrones posibles)
    - Alto esfuerzo

Patrones lexico-sintácticos

- Aproximaciones basadas en aprendizaje
  - Patrones aprendidos automáticamente
  - Modelos estadísticos de identificación de patrones
  - Pueden aplicarse a nuevos dominios
  - Necesitan ejemplos de entrenamiento
  - Evitan el esfuerzo humano del diseño de patrones a costa de necesitar ejemplos anotados

Extracción de relaciones supervisada

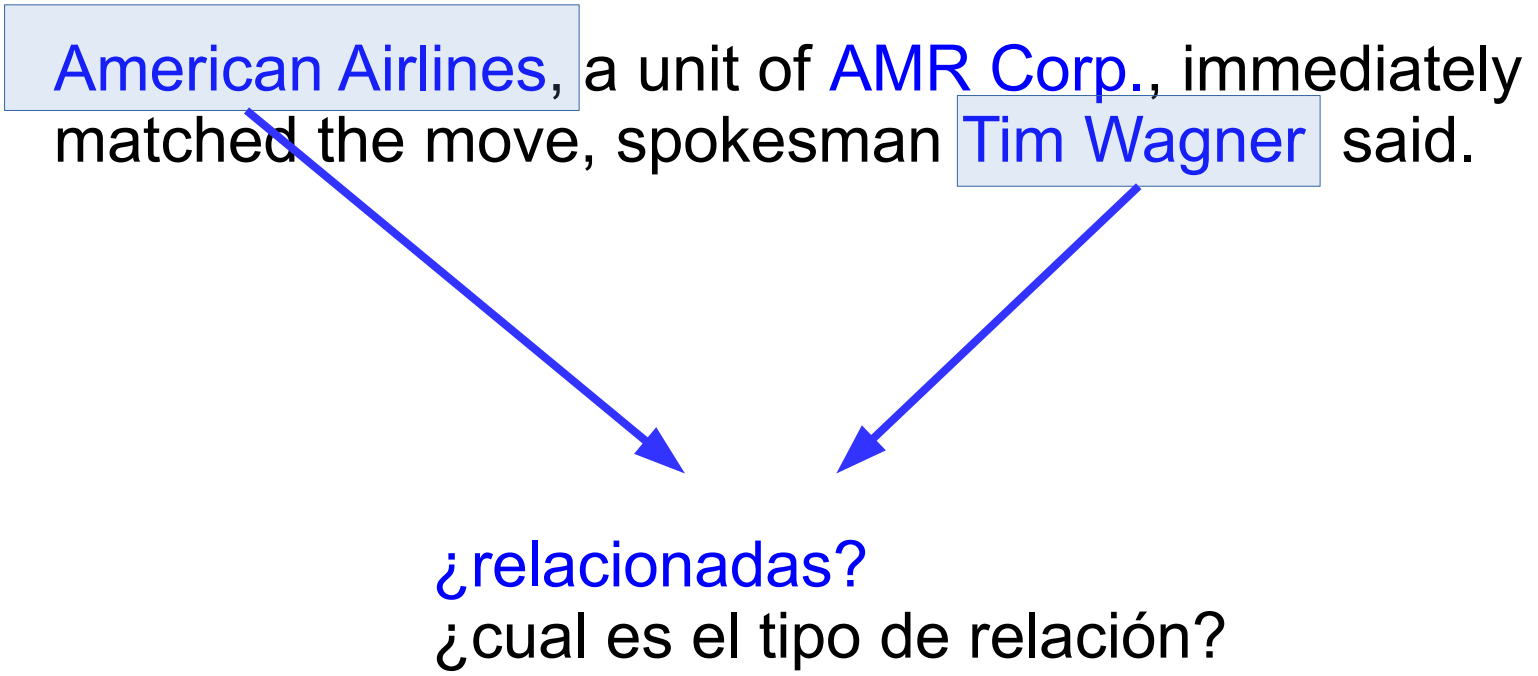
Pasos:

- Diseño de un marco de aprendizaje
- Decidir cuales son las relaciones de interés
- Seleccionar las entidades relevantes
- Encontrar o crear datos anotados
  - Corpus representativo
  - Entidades identificadas en el corpus (técnicas de NER)
  - Relaciones anotadas entre las entidades
  - Partición del corpus: entrenamiento/desarrollo/test o diseño de validación cruzada
- Entrenar, ajustar parámetros y evaluar

Extracción de relaciones supervisada

- Posibles enfoques a la extracción de relaciones como un problema de clasificación:
  - Decidir si dos entidades están relacionadas
  - Decidir la clase de un par de entidades relacionadas

Extracción de relaciones supervisada



Extracción de relaciones supervisada

- Rasgos típicos para el clasificador:
  - Núcleos de las entidades consideradas M1 y M2:
    - Airlines, Wagner
  - Bolsas de palabras y N-gramas de N1 y N2:
    - American, Airlines, Tim, Wagner, American Airlines, Tim Wagner
  - Palabras o bigramas de determinadas posiciones:
    - M2 – 1: spokesman
    - M2 + 1: said
  - Bolsas de palabras o bigramas entre:
    - a, AMR, of, immediately, matched, move, spokesman, the, unit
  - Versión con stems de los rasgos anteriores

Métodos basados en clasificadores secuenciales

- El problema se transforma en una clasificación de secuencias.
- El clasificador lee los datos secuencialmente de izquierda a derecha.
- Los datos se anotan en forma de secuencia con las etiquetas de salida consideradas

Extracción de relaciones semisupervisada: bootstrap

- Si hay suficientes datos de entrenamiento y el conjunto de test es similar al de entrenamiento, los métodos supervisados identifican relaciones de forma muy precisa.
- Sin embargo, el etiquetado manual de un conjunto de entrenamiento extenso es muy costoso.
- Por ello se recurre a otras técnicas que requieren menos datos como las semi-supervisadas

Extracción de relaciones semisupervisada: bootstrap

- Se parte de un conjunto de patrones semilla muy precisos o de un pequeño conjunto semilla de tuplas de entidades relacionadas.
- Se utilizan para hacer bootstrapping de un clasificador:
  - Se toman las entidades de los pares de la semilla
  - Se buscan oraciones en nuevos documentos (web, otros conjuntos de datos) que contengan a ambas entidades
  - Se extraen y generalizan los contextos de esas oraciones para aprender nuevos patrones.
  - Los nuevos patrones se pueden utilizar para recoger nuevas tuplas

Extracción de relaciones semisupervisada: bootstrap

- <Mark Twain, Elmira> Tupla semilla
- Buscar(google) los contextos de la tupla semilla:
  - “Mark Twain is buried in Elmira, NY.”
    - X is buried in Y
  - “The grave of Mark Twain is in Elmira”
    - The grave of X is in Y
  - “Elmira is Mark Twain’s final resting place”
    - Y is X’s final resting place.
- Usar esos patrones para buscar nuevas tuplas
- Iterar

Extracción de relaciones semisupervisada: bootstrap

- Los sistemas de bootstrapping asignan valores de confianza a las nuevas tuplas para evitar la deriva semántica (semantic drift):
  - Un patrón erróneo lleva a la introducción de tuplas erróneas, que a su vez lleva a patrones erróneos.
- Los valores de confianza de los patrones se basan en el equilibrio de dos factores:
  - el rendimiento del patrón con respecto al conjunto actual de tuplas (con cuantas encaja) y,
  - la productividad del patrón en términos del número de coincidencias que produce en toda la colección de documentos.

Extracción de relaciones semisupervisada: bootstrap

- Establecer umbrales de confianza conservadores para la aceptación de nuevos patrones y tuplas durante el proceso de bootstrapping ayuda a evitar que el sistema se aleje de la relación objetivo.

Extracción de relaciones con supervisión a distancia

- En lugar de usar un pequeño conjunto de semillas, utiliza una base de datos para obtener un enorme conjunto de semillas (no validadas)
- Extrae numerosos rasgos como patrones de todos estos ejemplos
- Los combina en un clasificador

Extracción de relaciones con supervisión a distancia

- Ejemplo: se busca la relación *place-of-birth*
- Se buscan en Dbpedia o Freebase muchos (10000 por ejemplo) casos de esta relación:
  - <Edwin Hubble, Marshfield>, <Albert Einstein, Ulm>, etc.
- Se usa un etiquetador de NER sobre una gran colección de textos para extraer todas las oraciones que contienen algunos de los pares de entidades seleccionados:
  - Hubble was born in Marshfield
  - Einstein, born (1879), Ulm
  - Hubble's birthplace in Marshfield, etc.

Extracción de relaciones con supervisión a distancia

- Se extraen rasgos frecuentes (análisis sintáctico, palabras, etc.)
- Se entrena un clasificador supervisado usando miles de patrones.

Extracción no supervisada de relaciones

- Open Information Extraction (Open IE)
- Sin utilizar datos de entrenamiento, ni listas de relaciones
- Se basan en modelos estadísticos.
- Algunos se basan en extraer ciertas relaciones entre determinadas partes de la oración (SN y verbos) que cumplen determinadas restricciones léxicas:
  - Se eliminan las relaciones que no aparecen con suficiente frecuencia.

Evaluación de la extracción de relaciones semi o no supervisada

- Relaciones nuevas → No hay un gold estándar con el que comparar.
  - No se puede calcular la precisión: no se sabe cuales son correctas
  - No se puede calcular la cobertura: no se sabe cuales han escapado a la detección
- Posible aproximación a la precisión:
  - Elegir una muestra aleatoria de la salida y comprobar la precisión manualmente.

Evaluación de la extracción de relaciones semi o no supervisada

- También puede calcularse la precisión a distintos niveles de cobertura:
  - Precisión para las 1000 relaciones de más peso, para las 10000 de más peso, para las 10000, etc.
  - En cada caso se toma una muestra aleatoria del conjunto
- Pero no se puede evaluar la cobertura.

Expresiones temporales

- Reconocimiento y normalización de expresiones temporales (horas, fechas):
  - 3 de la tarde
  - Mañana
  - Desde ayer
  - Dos veces al mes
- De particular importancia en tareas como la búsqueda de respuestas

Extracción de expresiones temporales

- Pueden referirse a momentos en el tiempo, a momentos relativos, a duraciones, o a combinaciones de los anteriores.
- Expresiones temporales absolutas: se corresponden con una fecha u hora concretas:
  - April 24, 1916
  - The summer of ’77
  - 10:15 AM
  - The 3rd quarter of 2006

Expresiones temporales

- Relativas: hacen referencia a otro punto del tiempo
  - yesterday
  - next semester
  - two weeks from yesterday
  - last quarter
- Duraciones: periodos de tiempo de distinta granularidad
  - four hours
  - three weeks
  - six days
  - the last three quarters

Expresiones temporales

- Son construcciones gramaticales que tienen un disparador léxico como núcleo
- Los disparadores pueden ser nombres, comunes y propios, adjetivos y adverbios.
- Ejemplos de disparadores (inglés):
  - Nombres: morning, noon, night, winter, dusk, dawn, ...
  - Nombres propios: January, Monday, Ides, Easter, Ramadan,...
  - Adjetivos: recent, past, annual, former
  - Adverbios: hourly, daily, monthly, yearly

Extracción de expresiones temporales

- Esquema de anotación TimeML
  - Las expresiones temporales se anotan con la etiqueta <TIMEX3> y varios atributos para ella:
  - Existen corpus anotados con este esquema como

    TimeBank 1.2
  - Ejemplo:

    A fare increase initiated <TIMEX3>last week</TIMEX3> by UAL Corp's United Airlines was matched by competitors over<TIMEX3>the weekend</TIMEX3>, marking the second successful fare increase in <TIMEX3>two weeks</TIMEX3>

Extracción de expresiones temporales

- Aprendizaje automático: Anotación para aproximaciones de etiquetado secuencial:
  - Esquema IOB (inside, outside, begin) para las expresiones delimitadas por TIMEX3:
- Ejemplo

A	fare	increase	initiated	last	week	by	UAL	Corp's…
O	O	O	O	B	I	O	O	O

Extracción de expresiones temporales

- Aprendizaje automático-Rasgos típicos:
  - Token: token objetivo a etiquetar
  - Tokens en la ventana: bolsa de tokens en la ventana alrededor del objetivo
  - Forma: rasgos de los caracteres: mayúsculas, longitud, etc.
  - POS: tag del objetivo y de la ventana de palabras
  - Etiquetas del chunk: tag de la frase del objetivo y de la ventana de palabras
  - Disparadores léxicos: presencia en la lista de disparadores temporales.

Normalización temporal

- Proceso de correspondencia de una expresión temporal con un punto específico del tiempo o con una duración.
- Puntos del tiempo pueden ser días del calendario y/o horas del día.
- Duraciones son longitudes de tiempo y también puntos de comienzo y fin.
- Los tiempos normalizados se representan con un atributo VALUE de la norma estándar ISO 8601

Normalización temporal

- <TIMEX3 id = ’’t1’’ type=”DATE” value =”2007-07-02” functionInDocument=”CREATION_TIME”> July 2, 2007 </TIMEX3> A fare increase initiated <TIMEX3 id=”t2” type=”DATE” value =”2007-W26” anchorTimeID=”t1”>last week</TIMEX3> by United Airlines was matched by competitors over <TIMEX3 id =”t3” type=”DURATION” value =”P1WE” anchorTimeID=”t1”> the weekend </TIMEX3>, marking the second successful fare increase in <TIMEX3 id=”t4” type =”DURATION” value =”P2W” anchorTimeID=”t1”> two weeks </TIMEX3>.
- Fecha del documento: 2 de julio de 2007 (en la representación ISO, YYYY-MM-DD: 2007-07-02)
- Esta es la referencia para el resto de las expresiones temporales del documento

Normalización temporal

- Las exp. temporales totalmente calificadas contienen año, mes y día en alguna forma convencional.
- Son infrecuentes en los textos reales.
- La mayor parte de las expresiones y suelen hacer referencia implícita a un punto del tiempo: el ancla temporal (temporal anchor) del documento.

Extracción de sucesos

- Suceso (event): suceso de la vida real que ocurre en un punto del tiempo y del espacio.
- Extracción de sucesos: nombre, tipo de evento, agente, tiempo y lugar.
- Ejemplo:

  [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.

Extracción de sucesos

- Los sucesos se clasifican en acciones, estados, de notificación (dice, explica, etc.), percepción, etc.
- La extracción de sucesos suele abordarse con aprendizaje supervisado (IO tagging).
- Rasgos: prefijos y sufijos de la palabra, POS tag, tipo de verbo, etc.

Extracción de sucesos

- Ordenación temporal de sucesos: ordenar los sucesos y expresiones temporales en una línea de tiempo.
- Útil en sistemas de búsqueda de respuesta y generación de resúmenes.
- Una tarea más simple es la ordenación parcial (relaciones de Allen): A before B, B after A, A overlaps B, A meets B, A equal B, A starts B, A finishes B, A during B, etc.
- El corpus TimeBank contiene la mayor parte de la información mencionada en esta sección.

Rellenado de plantillas (templates)

- Organizar la información de los textos en secuencias estructuradas de información (scripts) sobre sucesos puede ser muy útil para realizar inferencias (razonamiento).
- Las plantillas son un caso simple de script con un conjunto fijo de slots a rellenar.
- Ejemplo:

FARE-RAISE-ATTEMP: LEAD AIRLINE: UNITED AIRLINES

AMOUNT: $6

EFFECTIVE DATE: 2006-10-26

FOLLOWER: AMERICAN AIRLINES

Rellenado de plantillas (templates)

- La tarea suele modelarse con dos sistemas supervisados separados:
  - Reconocimiento de plantillas
  - Extracción de rellenos de rol (role-filler extraction): detección del rol de la entidad (LEAD-AIRLINE, AMOUNT, etc.)

Referencias

- Speech and Language Processing (3rd ed. draft)

  Dan Jurafsky and James H. Martin (Draft chapters in progress, Sep 23, 2018)

  - Tema 17: Information Extraction

Referencias

- Natural Language Processing with Python

  Steven Bird, Ewan Klein, and Edward Loper

  O'Reilly Media Inc. *https://www.nltk.org/book/*

  - 7. Extracting Information from Text