

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

1


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

Explotación de la Información. Extracción de Información

2

2




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”

3



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

4

1. Introducción a los sistemas de Extracción de Información

Explotación de la información. Extracción de Información

5

1. Introducción a los sistemas de Extracción de Información

Explotación de la información. 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	source

6

1. Introducción a los sistemas de Extracción de Información

Ejemplo de EI (LabTL-INAOE México):

Explotación de la información. Extracción de Información

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

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

Explotación de la información. Extracción de Información

8

8

Explotación de la información. Extracción de Información

1. Introducción a los sistemas de Extracción de Información

Importancia del procesamiento del lenguaje natural (PLN o *Natural Language Processing, NLP*):

- [Ex.Inf. Modulo 3. Las tecnologías lingüísticas.pdf](#)
- <https://www.red.es/redes/es/que-hacemos/tecnolog%C3%ADas-del-lenguaje>
 - <https://www.youtube.com/watch?v=3ixz-0SrXMw>
 - [Ex.Inf. Modulo 3. Plan de Impulso de las Tecnologías del Lenguaje - Agenda Digital.pdf](#)
 - [Ex.Inf. Modulo 3. I Hackaton en TLH.pdf](#)
- [Oferta trabajo Xerox: Ex.Inf. Modulo 3. Oferta trabajo Xerox.pdf](#)
- [Google launches new API to help you parse natural language](#)
 - [Ex.Inf. Modulo 3. Google launches new API to help you parse natural language.pdf](#)
 - <https://techcrunch.com/2016/07/20/google-launches-new-api-to-help-you-parse-natural-language/>
 - <https://cloud.google.com/natural-language/>
- [Ex.Inf. Modulo 3. The Stanford Natural Language Processing Software.pdf](#)
 - <https://nlp.stanford.edu/software/>
- [Ex.Inf. Modulo 3. Qué -o quién- es IBM Watson_ esta es su historia - MediaTrends.pdf](#)

9

9

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

10

10

2. Arquitectura de los sistemas de EI

Arquitectura típica de PLN en pipeline (ACM SIGAI Learning Webinar, "On the Evolution of NLP, QA, and IE, and Current Research and Commercial Trends", Dan Moldovan):

- [Ex.Inf. Modulo 3. Extraccion de Informacion. Dan Moldovan.pdf](#)

```

graph TD
    Text --> DP[Document Preprocessing]
    DP --> TS[Text Segmentation]
    TS --> T[Tokenization]
    T --> P[Part-of-speech Tagging]
    P --> NEE[Named Entity Extraction]
    NEE --> EE[Event Extraction]
    EE --> WSD[Word Sense Disambiguation]
    WSD --> SP[Syntactic Parsing]
    SP --> S[Semantic Parsing]
    S --> CR[Coreference Resolution]
    CR --> TD[Topic Detection]
    TD --> RT[RDF/TriX Representation]
    RT --> ST[Semantic Triples]
  
```

11

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

12




2. Arquitectura de los sistemas de Extracción de Información

Multidisciplinariedad del PLN:

- *Linguistics*: cómo se forman palabras, sintagmas y oraciones
- *Psycholinguistics*: cómo la gente se entiende y comunica utilizando el lenguaje humano
- *Computational Linguistics*: trata con los modelos y aspectos computacionales del lenguaje natural (algoritmos)
- *Philosophy*: semántica del lenguaje, nociones del significado, cómo las palabras identifican objetos

13




2. Arquitectura de los sistemas de Extracción de Información

Multidisciplinariedad del PLN (cont.):

- *Computer Science*: formulación de modelos y su implementación
- *Artificial Intelligence*: temas relacionados con la representación del conocimiento y su razonamiento
- *Statistics*: muchos problemas de PLN se resuelven utilizando modelos probabilísticos
- *Machine Learning*: aprendizaje automático de reglas y procedimientos basándose en características léxicas, sintácticas y semánticas
- *NL Engineering*: implementación de sistemas informáticos que procesen el lenguaje natural

14




2. Arquitectura de los sistemas de Extracción de Información

Aplicaciones del PLN:

- *Sentiment Analysis*
- *Text Summarization*
- *Textual Entailment*
- *Information Extraction*
- *Topic Segmentation*
- *Question Answering*
- *Semantic comparison of two documents*
- *Ontology Building*
- *Event Detection and Reasoning*
- *Regulatory Compliance*

15



2. Arquitectura de los sistemas de Extracción de Información

Aplicaciones del PLN (cont.):

- *Trend Analysis and Prediction*
- *Risk Management; Risk Assessment*
- *Decision Making and Evidence Support*
- *Customer Profiles and Business Intelligence*
- *CRM (Customer Relationship Management), call centers using text messages*
- *Translation*
- *Argumentation Mining*

16

Explotación de la información. Extracción de Información

2. Arquitectura de los sistemas de Extracción de Información

Aplicaciones del PLN (cont.):

- *Systems for processing Big Data*
- *IE: converting unstructured text to actionable knowledge (structured data: RDF, XML, JSON)*
- *Creating/refining domain ontologies, Increased use of ontologies*
- *Enterprise Knowledge Graphs, RDF stores, graph databases*
- *Conversational agents: chatbots and dialog systems*
- *Human to Machine communication in addition to M2M*
- *QA: Information retrieval systems with natural language query*

17

Explotación de la información. Extracción de Información

2. Arquitectura de los sistemas de Extracción de Información

Aplicaciones del PLN (cont.):


- *Tools for rapid domain customization*
- *Decision and predictions as a service*
- *Document understanding combining text, tables, figures, images, drawings and graphs.*
- *Applications get smarter; NLP – enabled AI in areas such as contract processing, decision making, argumentation reasoning, etc.*

Paquetes de software:

- <http://nlp.lsi.upc.edu/freeling/node/1>
- <https://nlp.stanford.edu/software/>
- <https://github.com/flairNLP/flair>

18

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.
 - ¿Banco de sentarse o entidad financiera?
- El avión localizó el banco y comunicó su situación.
 - ¿Cuántos significados diferentes tiene “banco” en estas frases?
 - ¿Se te ocurre algún significado más?

19

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

20

Explotación de la información. Extracción de Información

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

21

Explotación de la información. Extracción de Información

3. Módulo de análisis léxico

Texto

↓

Tokenización / Segmentación

↓

Análisis Léxico / Morfológico

↓

Etiquetado Morfosintáctico

↓

Desambiguación semántica

→

Sentidos de las palabras

Ejemplo

Juan toca el bajo en la banda .


Juan: NP toca: VMIP3S, VMM2S, NCFS el: ART
 bajo: NC, VMP, PREP, ADJ en: PREP la: ART, NC banda: NC

Juan: NP toca: VMIP3S el: ART bajo: NC
 en: PREP la: ART banda: NC

Juan: persona
 toca: hacer sonar un instrumento
 bajo: instrumento musical
 banda: conjunto de instrumentistas

22

22



Explotación de la información. Extracción de Información

3. Módulo de análisis léxico

POS tagger (*Part-of-speech*):

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 TDM50
12_de_enero_de_1900 12/1/1900 F W
,, E2 Fc
era era NCFS000 erar VMMP2S0 erar VMIP3S0 ser VAI1S0 ser VAI13S0
un un TIMS0 un MCMS00
gran gran AQOC500
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

23



Explotación de la información. Extracción de Información

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

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

24

Explotación de la información. Extracción de Información

3. Módulo de análisis léxico

Explotación de la información. Extracción de Información

Ejercicio 2:

- Con el objetivo del rellenando de la siguiente plantilla:
 - INCIDENT: DATE
 - INCIDENT: LOCATION
 - INCIDENT: TYPE
 - INCIDENT: STAGE OF EXECUTION
 - INCIDENT: INSTRUMENT TYPE
 - HUM TARGET: NAME
 - HUM TARGET: DESCRIPTION
 - HUM TARGET: TYPE

25

25

Explotación de la información. Extracción de Información

3. Módulo de análisis léxico

Explotación de la información. Extracción de Información

Ejercicio 2 (cont.):


- Tarea de EI del *MUC-4 Terrorism Task*. ¿Qué información léxica sería útil para rellenar esa plantilla (p.ej. fechas, personas y localizaciones)? Consultad la salida del análisis léxico en [Ex.Inf. Modulo 3. Ejercicio 2.pdf](#)
 - <https://corenlp.run/>
 - <http://nlp.lsi.upc.edu/freeling/demo/demo.php>
 - <http://services.gate.ac.uk/annie/>
 - <http://www.opener-project.eu/webservices/entrance.html>
 - 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.

26

26




Explotación de la información. Extracción de Información

4. Módulo de análisis sintáctico

Ambigüedad sintáctica

- La vendedora de periódicos del barrio.
 - ¿Quién es del barrio: la vendedora o los periódicos?
- Juan vio al ladrón con los prismáticos
 - ¿Quién tenía los prismáticos: Juan o el ladrón?
- Pedro vio a Juan en lo alto de la montaña con los prismáticos
 - ¿Quién tenía los prismáticos: Pedro o Juan?

27



Explotación de la información. Extracción de Información


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)

28




Explotación de la información. Extracción de Información

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.)

29



Explotación de la información. Extracción de Información

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.

30

Explotación de la información. Extracción de Información

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.

31

Explotación de la información. Extracción de Información

4. Módulo de análisis sintáctico

#

Gramáticas de Cláusulas Definidas (DCG):

S --> NP, VP. %**Símbolos no terminales**

NP --> det, n.

NP --> nprop.

VP --> v, NP.

det --> [el]. % **Símbolos terminales.**

det --> [las]. **Diccionario**

det --> [una].

n --> [perro].

n --> [hueso].

n --> [orejas].

nprop --> [Pepe].


v --> [come].

v --> [movía].

```

graph TD
    S --> NP1[NP]
    S --> VP[VP]
    NP1 --> nprop[nprop]
    nprop --> Pepe[Pepe]
    VP --> v[v]
    v --> come[come]
    VP --> NP2[NP]
    NP2 --> det[det]
    det --> una[una]
    NP2 --> n[n]
    n --> manzana[manzana]
  
```

32



Explotación de la información. Extracción de Información

4. Módulo de análisis sintáctico.

Análisis sintáctico completo

¿Quién tiene el telescopio en cada árbol sintáctico?

S

NP

Nprop

Luis

VP

V

ve

OD

CONT

al

NP

NPROP

hombre

PP

PREP

con

NP

DET

el

N

telescopio

S

NP

Nprop

Luis

VP

V

ve

OD

CONT

al

NP

NPROP

hombre

PP

PREP

con

NP


DET

el

N

telescopio

33



Explotación de la información. Extracción de Información

4. Módulo de análisis sintáctico.

Análisis sintáctico parcial

NP

Nprop

Luis

V

ve

PP

CONT

al

NP

NPROP

hombre

PP

PREP

con

NP


DET

el

N

telescopio

34




Explotación de la información. Extracción de Información

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

35



Explotación de la información. Extracción de Información

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

36

Explotación de la información. Extracción de Información

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:
 - Supervisados
 - No supervisados

37

Explotación de la información. Extracción de Información

5. Módulo de reconocimiento de entidades

Ejercicio 2 (cont.):


- Tarea de EI del *MUC-4 Terrorism Task*. ¿Qué información **SINTÁCTICA (REGLAS SINTÁCTICAS)** sería útil para rellenar esa plantilla (p.ej. fechas, personas y localizaciones)? Consultad la salida [Ex.Inf. Modulo 3. Ejercicio 2.pdf](#)
 - <https://corenlp.run/>
 - <http://nlp.lsi.upc.edu/freeling/demo/demo.php>
 - <http://www.opener-project.eu/webservices/entrance.html>
 - http://cogcomp.cs.illinois.edu/page/demo_view/POS
 - http://cogcomp.cs.illinois.edu/page/demo_view/ShallowParse
 - 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.

38

38




Explotación de la información. Extracción de Información

6. Módulo de análisis semántico

Ambigüedad semántica

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

39



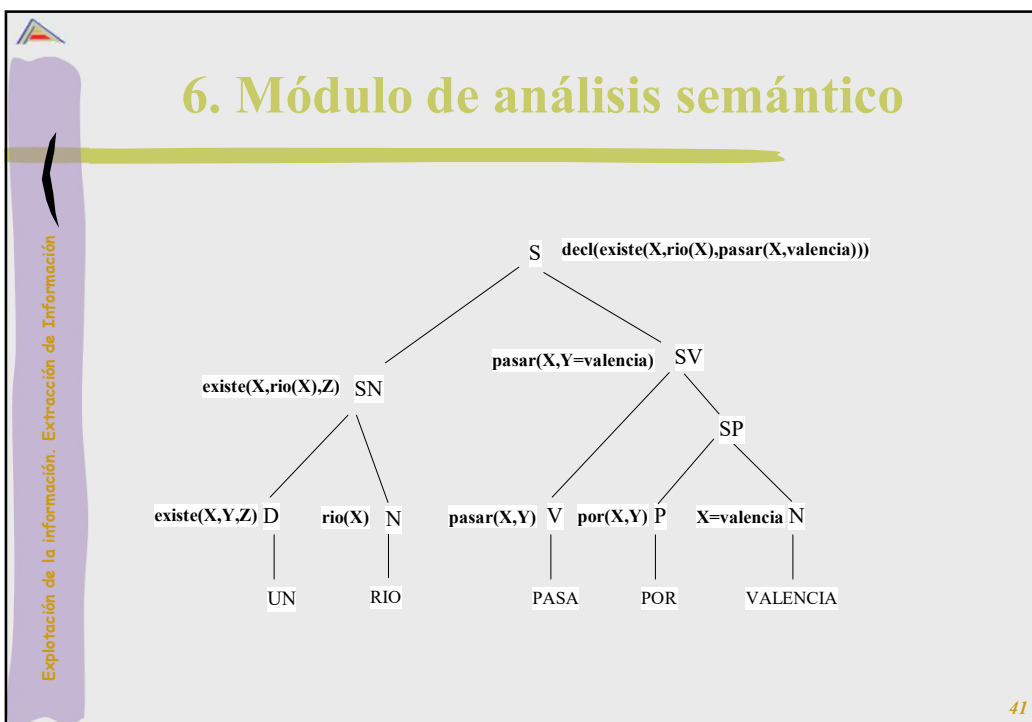
Explotación de la información. Extracción de Información

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:
 - $\# (\exists x) R(x) \wedge G(x) \wedge P(x, \text{valencia})$

40



41

6. Módulo de análisis semántico

Explotación de la información. Extracción de Información

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”
 - $saber(Juan, Latín) \wedge saber(María, Griego)$

42

42

Explotación de la información. Extracción de Información

6. Módulo de análisis semántico

Roles temáticos:

```

graph TD
    dar((dar)) -- agente --> Juan((Juan))
    dar -- tema --> libro((libro))
    dar -- instrum --> Maria((María))
        
```

43

43

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))

44

44

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 dijo?
 - ¿Cuándo dijo?
 - ¿Que pusiera qué?
 - ¿Que pusiera encima de dónde?

45

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

46

46



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_i entró en la tienda... Él_i buscaba un regalo...

47

47



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


48

48




Explotación de la información. Extracción de Información

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

49




Explotación de la información. Extracción de Información

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_i y María lo_i seca

50



Explotación de la información. Extracción de Información

7. Módulo de resolución de correferencias

Anáfora versus catáfora:

Fora


Exófora

Endófora

Anáfora

Catáfora

51



Explotación de la información. Extracción de Información

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_i, Juan_i vio una serpiente.
- Elipsis catafórica:
 - Si Ø_i gana en la lotería, Juan_i se compra un piano.

52

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_i rompió el cristal. Yo le_i vi
 - **Verbal:** Pedro [jugó muy bien al tenis ayer]₁, pero Juan [lo hizo]₁ muy mal
 - **Oracional:** [No deberíamos salir esta noche]_i. Yo no opino eso_i
2. **Semántica:**
 - **Sinonimia:** Pedro se quitó sus gafas_i... Estas lentes_i...
 - **Hiperonimia:** No sabía que ese coche_i es tuyo. Opino que es un buen vehículo_i
 - **Contextual:** Él se limpió las gafas_i y se las ajustó a la nariz. Su montura_i y cristales_i estaban húmedas
3. **Pragmática:** ...”La isla del tesoro”_i... En ese libro_i...

53

Explotación de la información. Extracción de Información

7. Módulo de resolución de correferencias

Clasificación según el tipo de expresión anafórica:

- **Pronominal:** Tu hijo_i rompió el cristal. Yo le_i vi
- **Descripciones definidas:** Juan_i perdió el dinero, [el pobre chico]_i está hundido
- **Tipo “one”:** Peter bought [a blue pen]₁ yesterday. He has bought [another one]₁ today
- **Adjetiva:** Compré [una pera verde]_i y [otra roja]_j. Yo prefiero [la verde]_i
- **Superficial numérica:** Pedro miró [al perro_i y al gato_j], pero finalmente eligió el primero_i
- **Verbal:** Pedro [besó a su mujer]₁. Juan también [lo hizo]₁
- **Adverbios:** La iglesia estaba [detras de la librería]_i. Luis fue ahí_i después del almuerzo
- **Complementos circunstanciales:** El despertador suena a [las 6 de la mañana]₁. [Las siguientes dos horas]₁...

54

54

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

55

Explotación de la información. Extracción de Información

7. Módulo de resolución de correferencias

- # **Ejercicio 2 (cont.):**
 - Tarea de EI del *MUC-4 Terrorism Task*. **Etiqueta anáforas y sus soluciones**
 Consultad la salida [Ex.Inf. Modulo 3. Ejercicio 2.pdf](#)
 - http://cogcomp.cs.illinois.edu/page/demo_view/Coref

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.

56

56

Explotación de la información. Extracción de Información

8. Módulo de análisis contextual

- # **“La Isla de la Calavera me ha gustado mucho”:**
 - ¿Libro o lugar?
- # **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 (p.ej. fecha noticia periódico)
 - Identificación de la intención del hablante (fundamental en un interfaz en LN a BD)
 - Proceso inferencial requerido para interpretar la oración dentro del dominio de aplicación

57

Explotación de la información. Extracción de Información

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

58

58

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 correferencia

59

59

9. Módulo de extracción, rellenado y almacenamiento de plantillas

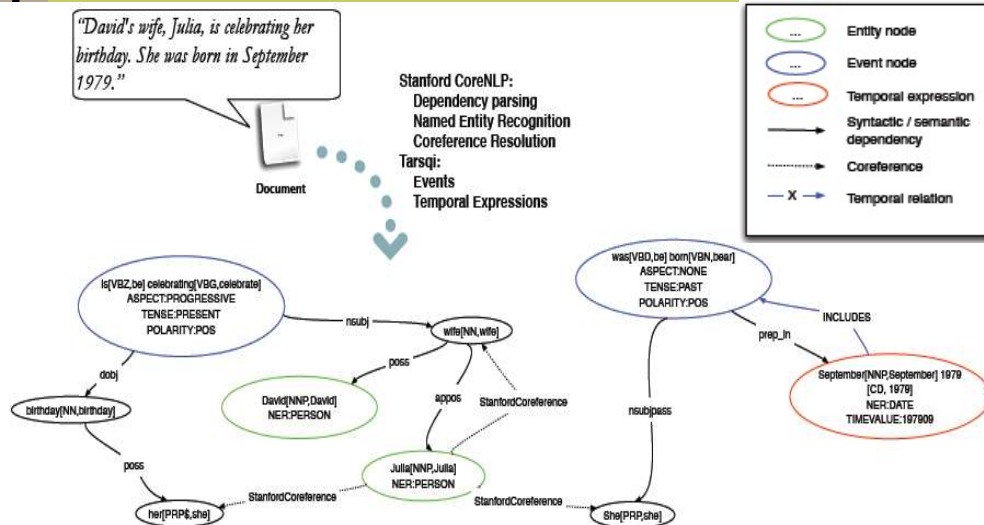
"David's wife, Julia, is celebrating her birthday. She was born in September 1979."



Document

Stanford CoreNLP:
Dependency parsing
Named Entity Recognition
Coreference Resolution

Tarsqi:
Events
Temporal Expressions



60

60

9. Módulo de extracción, relleno y almacenamiento de plantillas

Collapse referents of discourse

Graph normalization

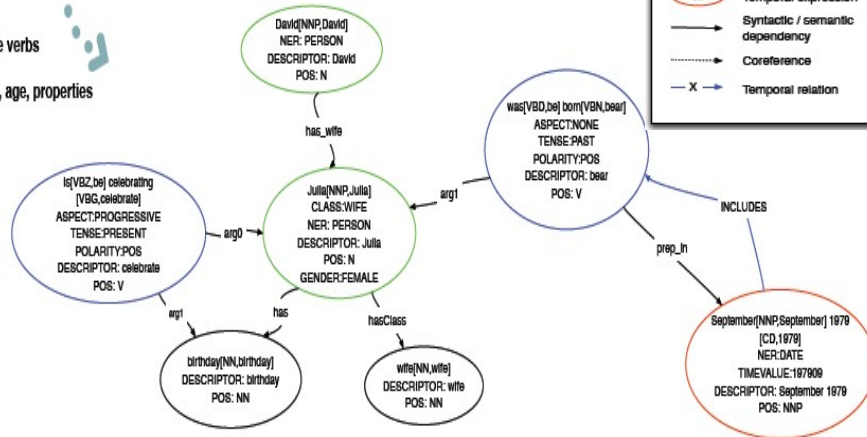
Semantic information:

Normalize copulative verbs

Normalize genitives

Infer semantic class, age, properties

Assign a gender



61

61

9. Módulo de extracción, relleno y almacenamiento de plantillas

✦ Ejercicio 2 (cont.):

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

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.

62

62

Explotación de la información. Extracción de Información

9. Módulo de extracción, relleno y almacenamiento de plantillas

Ejercicio 2 (cont.):

- Indicando un algoritmo para el relleno de la siguiente ÚNICA plantilla (aplica dicho algoritmo y muestra los resultados que saldrían de su aplicación, analizando su precisión y taxonomías utilizadas):
 2. INCIDENT: DATE
 3. INCIDENT: LOCATION
 4. INCIDENT: TYPE
 5. INCIDENT: STAGE OF EXECUTION
 7. INCIDENT: INSTRUMENT TYPE
 18. HUM TARGET: NAME
 19. HUM TARGET: DESCRIPTION
 20. HUM TARGET: TYPE

63

63

Explotación de la información. Extracción de Información


10. Ejemplos de sistemas de EI

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

64


64



Explotación de la información. Extracción de Información

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

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>

67

67



Explotación de la información. Extracción de Información

10. Ejemplos de sistemas de EI.

Cornell University

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

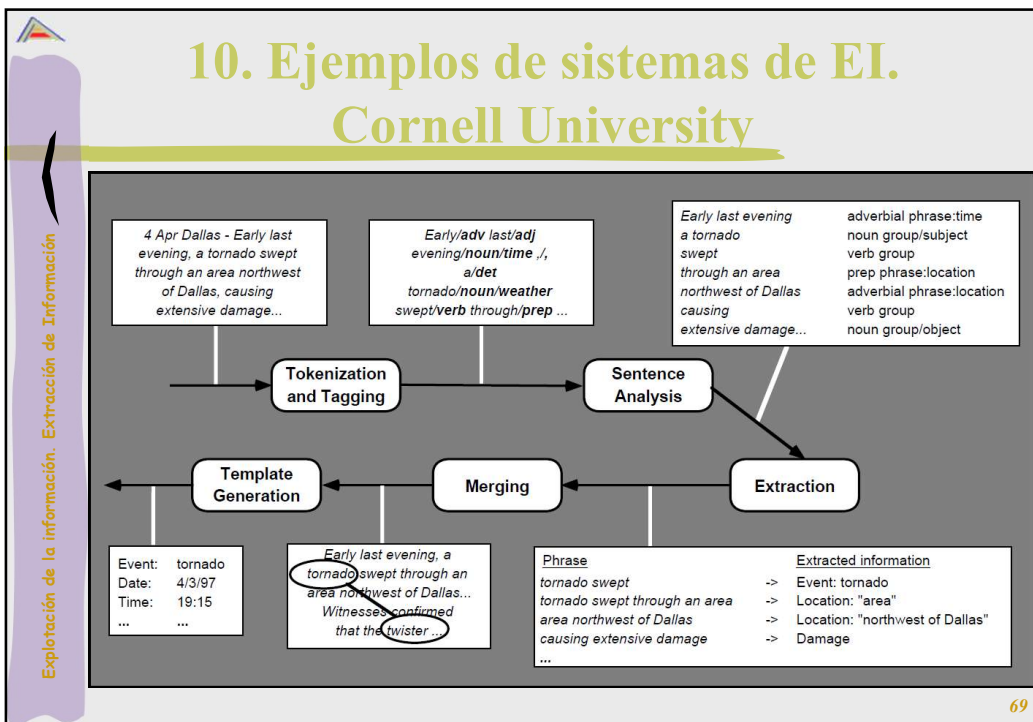
Output Template

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

68

68

34



69

Explotación de la información. Extracción de Información

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

70

Explotación de la información. Extracción de Información

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

71

71

Explotación de la información. Extracción de Información

10. Ejemplos de sistemas de EI

Ex.Inf. Modulo 3. Esta IA ha leído 3,3 millones de estudios y ha descubierto datos que nadie había visto antes.pdf

Ahora un grupo de investigadores del Lawrence Berkeley National Laboratory (California, Estados Unidos) ha utilizado esta tecnología para analizar 3,3 millones de estudios científicos, según publica *MIT Technology Review*. Se trata de investigaciones publicadas entre 1922 y 2018 en revistas que hablan sobre la ciencia de los materiales. Analizando la relación entre palabras, la inteligencia artificial fue capaz de capturar el conocimiento fundamental dentro del campo, incluida la tabla periódica y la forma en que las estructuras de los químicos se relacionan con sus propiedades.

72

72

10. Ejemplos de sistemas de EI

Ejercicio 4: **EJERCICIO A ENVIAR COMO TUTORÍA CV (evaluación parte teórica): (cont.)**

- ¿Cómo mejoraríais el modelo clásico de RI visto en el módulo 2 introduciendo técnicas de PLN para alcanzar una mejor comprensión del significado de los documentos?
- ¿Cómo se incorporaría en el modelo del coseno con pesos?

$$\text{sim}(Q, D) = \frac{\sum_{i=1}^k q_i * d_i}{\|Q\| * \|D\|}, \quad \|Q\| = \sqrt{\sum_{i=1}^k q_i * q_i}, \quad q_i = \text{ft}_{Q,i} * \log_e\left(\frac{N}{\text{fd}_i}\right)$$

$$\|D\| = \sqrt{\sum_{i=1}^k d_i * d_i}, \quad d_i = \text{ft}_{D,i} * \log_e\left(\frac{N}{\text{fd}_i}\right)$$

- Medida del coseno según (Kaszkiet et al., 1999):

$$q_i = \log_e(\text{ft}_{q,i} + 1) * \log_e\left(\frac{N}{\text{fd}_i} + 1\right)$$

$$d_i = \log_e(\text{ft}_{d,i} + 1)$$

73

73

10. Ejemplos de sistemas de EI

Ejercicio 4 (cont.): **EJERCICIO A ENVIAR COMO TUTORÍA CV (evaluación parte teórica)**

- Para ello se aconseja buscar ejemplos y proponer soluciones con PLN para resolver problemas de comprensión de la RI tradicional, dando ejemplos de cuándo funciona bien, y cuándo mal. Por ejemplo:
 - Variaciones léxicas:
 - Valorar más los términos de la query de tipo nombre propio (NP):
 - Funciona bien: query “Comida rápida en Japón” si el doc. tiene “Japón”
 - Funciona mal: si el doc. tiene “Comida rápida en Asia” “Comida nipona”
 - Tener en cuenta conversiones léxicas:
 - (n-v) Recorte de los gastos → recortar los gastos.
 - (n-adj) Cambio del clima → cambio climático.

74

74

Explotación de la información. Extracción de Información

10. Ejemplos de sistemas de EI

Ejercicio 4 (cont.): **EJERCICIO A ENVIAR COMO TUTORÍA CV (evaluación parte teórica)**

- Más ejemplos (cont.):
 - Variaciones sintácticas:
 - # Términos dentro de sintagmas nominales de la query que aparecen separados en los documentos
 - Funciona bien query: “Arabella Kiesbauer” (NP+NP) “siete maravillas” (det+n)
 - Funciona mal query: “Reichstag alemán” (N-Adj) “países europeos” (N-N)
 - # ¿Diferentes estructuras sintácticas para el mismo significado?
 - Funciona bien query: “Comida rápida en Japón” (SN+SP) doc: “Comida rápida nipona”
 - Funciona mal query: “diamantes de Sierra Leona” (SN+SP) doc: “Sierra Leona ... diamantes”
 - # ¿Cómo valorar los modificadores de más que aparecen en un sintagma nominal? ¿Y los que no aparecen?

75

75

Explotación de la información. Extracción de Información


10. Ejemplos de sistemas de EI

Ejercicio 4 (cont.): **EJERCICIO A ENVIAR COMO TUTORÍA CV (evaluación parte teórica)**

- Más ejemplos (cont.):
 - Variaciones semánticas:
 - # Sinonimia, polisemia, hiperonimia, meronimia, ...
 - Resolución de elipsis o anáfora
 - Tesis Doctoral:
 - # Karam Abdulahhad. Information Retrieval (IR) Modeling by Logic and Lattice. Application to Conceptual IR. Information Retrieval [cs.IR]. Université de Grenoble, 2014.
 - # <https://tel.archives-ouvertes.fr/tel-00991669/document>

76

76



Explotación de la información. Extracción de Información


10. Ejemplos de sistemas de EI

Ejercicio 4 (cont.): **EJERCICIO A ENVIAR COMO TUTORÍA CV (evaluación parte teórica)**

- Objetivos a evaluar:
 - Nivel de detalle de los ejemplos propuestos
 - Nivel de detalle de la solución propuesta a esos ejemplos
 - Nivel de detalle de la incorporación de esa solución al modelo del coseno con pesos
 - Nivel de detalle de la memoria presentada

77

77



Explotación de la información. Extracción de Información

10. Ejemplos de sistemas de EI

Ejercicio 4 (cont.): **EJERCICIO A ENVIAR COMO TUTORÍA CV (evaluación parte teórica)**

- [Ex.Inf. Modulo 3. Extraccion de Informacion. RI con PLN Ejemplo1 pregunta 146.pdf](#)
- [Ex.Inf. Modulo 3. Extraccion de Informacion. RI con PLN Ejemplo1 pregunta 189.pdf](#)

78

78

Explotación de la información. Extracción de Información

79

10. Ejemplos de sistemas de EI

EJERCICIO OPCIONAL A ENVIAR COMO TUTORÍA CV

- Sobre vuestra tercera práctica de búsqueda, implementar las soluciones propuestas en el ejercicio anterior
- Compararlo con los resultados de dicha práctica: calcular la gráfica comparativa
- Los que quieran implementar soluciones a las variaciones sintáctico-semánticas, que me pidan por tutoría de CV la versión del corpus con dicha información
- Objetivos a evaluar:
 - Mejoras conseguidas en la precisión y cobertura
 - Nivel de detalle en la solución propuesta
 - Nivel de detalle de la memoria presentada