

Índice

- 1 Tema 1: Internet
- 2 Tema 2: La web
- 3 Tema 3: Principios de Recuperación de Información
- 4 Tema 4: Búsqueda de información en la web
- 5 Tema 5: Búsqueda avanzada en la web
- 6 Tema 6: Integración de conocimiento lingüístico**
- 7 Tema 7: Más allá de la búsqueda textual
- 8 Pruebas de evaluación

Introducción

- Procesamiento del Lenguaje Natural (NLP, *Natural Language Processing*): tratamiento computacional del lenguaje humano
 - Objetivo: computadora comprenda el lenguaje humano
- IR como tarea de NLP: "comprender" el contenido de los documentos
- Principal problema de IR: **variación lingüística**
 - El mismo concepto puede expresarse de muy diferentes maneras
 - Impide establecer correspondencias

Introducción (cont.)

- Diferentes niveles de variación:

- Morfológica: modificaciones **flexivas** y **derivativas**
cantas / cantó cantar / cantante

- Semántica: **polisemia**
banda (de música) / banda (franja)

- Léxica: **sinonimia**
rápido / veloz

- Sintáctica: modificaciones de la **estructura sintáctica**
Juan atacó a Pepe / Pepe atacó a Juan

- Pueden darse simultáneamente: p.ej. morfo-sintáctica:
cambio climático / cambio del clima

- Solución: **técnicas de NLP**

Tratamiento de la variación lingüística

En general, dos enfoques diferenciados:

- **Normalización:** reducir las diferentes variantes de un término a una *forma canónica* común
 - Ej. sustituir una palabra por su lema (*lematización*)
- **Expansión:** añadir a la consulta variantes de sus términos originales
 - Ej. añadir sinónimos

Tratamiento de la variación morfológica: *stemming*

- Reducir de una palabra a su *stem* o raíz supuesta eliminando su terminación según una lista de sufijos
 - *Stem* o raíz contiene semántica básica

reloj
relojes
relojero

} → reloj-

- *Stemmer* de Porter
 - Demo: <http://maya.cs.depaul.edu/~classes/ds575/porter.html>
 - Snowball (descargables): <http://snowball.tartarus.org>
- Nivel de normalización
 - *Superficial*: sólo morfología flexiva simplificada; ej. sólo plurales
 - *Profundo*: flexiva y derivativa (agresivo); ej. Porter/Lovins

Tratamiento de la variación morfológica: *stemming* (cont.)

- Ventajas
 - Simplicidad
- Desventajas:
 - Problemas con idiomas de morfología compleja. Ej. español:
 - Adjetivos/nombres: +20 grupos variación género +10 grupos número
 - Verbos: 3 grupos regulares, ±40 irregulares; 118 formas flexivas cada grupo
 - Pérdida de información de cara a procesamiento futuro
 - Sobre-*stemming*: palabras no relacionadas dan igual *stem*

$$\left. \begin{array}{l} \textit{general} \\ \textit{generous} \end{array} \right\} \rightarrow \textit{gener-}$$

- Sub-*stemming*: palabras sí relacionadas dan *stems* diferentes

$$\begin{array}{l} \textit{recognize} \rightarrow \textit{recogn-} \\ \textit{recognition} \rightarrow \textit{recognit-} \end{array}$$

Tratamiento de la variación morfológica (cont.): otras aproximaciones

- Expansión de la consulta con variantes:
 - Google: busca simultáneamente el término en singular y plural
- **Lematización:** sustituir palabra por su lema
 - Mejora resultados con idiomas de morfología compleja
 - Reduce la pérdida de información

Tratamiento de la variación léxico-semántica

- Técnicas de *desambiguación del sentido de las palabras* (*WSD, Word Sense Disambiguation*)
 - Necesaria alta efectividad
- **WordNet**: base de datos léxica (*thesaurus*)
 - Sinónimos, antónimos, hipónimos, hiperónimos, etc.
 - Inglés. **EuroWordNet** para lenguas europeas
- Aproximaciones clásicas:
 - Expansión de la consulta con sinónimos, etc.
 - Google: operador ~
~simio
 - Indexación por sentidos en lugar de palabras
 - Poco efectivas salvo con consultas cortas o incompletas

Tratamiento de la variación sintáctica: introducción

- IR clásica basada en paradigma *bag-of-terms*:
 - Consultas/documentos como conjuntos de términos (desordenados)
 - Consulta y documentos relacionados si comparten términos

- Problema:

$$\left. \begin{array}{l} \text{Juan atacó a Pepe} \\ \text{Pepe atacó a Juan} \end{array} \right\} \rightarrow \{\text{atacó, Juan, Pepe}\}$$

- Solución: indexar también frases (más precisas)

Tratamiento de la variación sintáctica (cont.): identificación y extracción

- Técnicas estadísticas
 - Secuencias de palabras coocurren frecuentemente
 - Análisis estadístico (frecuencias, coocurrencias, etc.)
 - Sin base lingüística (a veces resultados extraños)
 - Mayor simplicidad
- Sintácticas
 - Secuencias de palabras satisfacen relaciones sintácticas
 - Análisis sintáctico (complejidad diversa)
 - Sí base lingüística (teóricamente superiores)
 - Mayor complejidad

Tratamiento de la variación sintáctica (cont.): representación y correspondencias

- Como conjuntos de palabras
- Almacenar árbol de análisis
 - Técnicas de comparación de árboles: gran complejidad
- Almacenar sólo las relaciones sintácticas que nos interesan
 - Sustantivo–modificador
 - Sujeto–verbo
 - Verbo–Objeto
 - ...

Tratamiento de la variación sintáctica (cont.): en buscadores

- Operadores de frase
- Operadores comodín * a nivel de palabra completa
- Aproximar sintaxis mediante distancias
 - Palabras cercanas se suponen relacionadas sintácticamente
 - Proximidad como indicador relevancia

Extracción de información (*IE, Information Extraction*)

- **Objetivo:** obtener información estructurada a partir de textos en lenguaje natural (i.e. desestructurados)

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

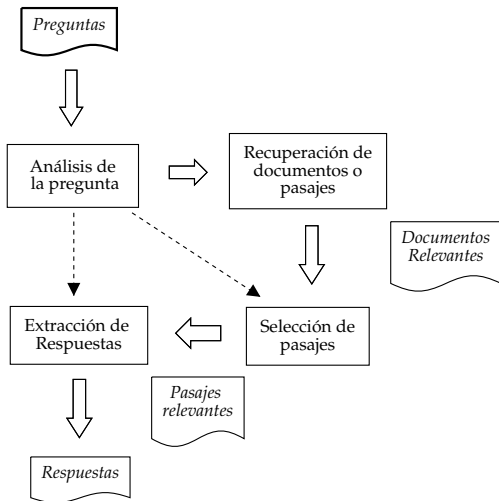
Modelo	Precio	Tel
Peugot 205	6500	981123456

- Sistemas altamente especializados de ámbito restringido
- Subtareas involucradas:
 - Reconocimiento de entidades: personas, organizaciones, lugares, expresiones temporales y numéricas, etc.
 - Resolución de coreferencias (ej. anáfora)
 - Extracción terminológica: identificación de términos relevantes
- Ej. sistema FASTUS

Búsqueda de respuestas (QA, *Question Answering*)

- **Objetivo:** dar respuestas a preguntas precisas y arbitrarias de los usuarios en base al contenido de una colección de textos en lenguaje natural
 - NO CONFUNDIR CON Google/Yahoo Answers !!! ("a mano")
 - Ejemplo: sistema START del MIT (<http://start.csail.mit.edu>)
- Combina técnicas de IR e IE:
 - IR: localiza documentos relacionados con el tema de la consulta, pero no extrae la información requerida
 - IE: extrae la información requerida, pero no permiten procesar consultas arbitrarias (sistemas muy especializados)

Búsqueda de respuestas (QA, *Question Answering*) (cont.)



Traducción automática (*MT, Machine Translation*)

- **Objetivo:** traducción de textos por ordenador
 - También sistemas semiautomáticos (interacción con el usuario)
 - SYSTRAN (<http://www.systran.co.uk>): empleado por
 - ANTERIORMENTE por *herrs. del idioma* de Google
 - *Babel Fish* de Yahoo (<http://babelfish.yahoo.com>)
 - *Babel Fish* de Altavista (<http://babelfish.altavista.com>)

Traducción automática (*MT, Machine Translation*) (cont.)

- Técnicas:

- Mediante diccionarios (*dictionary-based*): sustitución palabra por palabra empleando diccionarios bilingües
- Estadística (*statistical*): técnicas estadísticas empleando corpus paralelos bilingües
 - ACTUALMENTE traductor Google Translate (<http://translate.google.es>), empleado por herrs. del idioma de Google
- Basada en ejemplos (*example-based*): *online* por analogía empleando corpus paralelos bilingües
- Mediante interlingua (*interlingual*): gran complejidad, en 2 fases:
 - 1 Decodificar texto en lengua origen: paso a interlingua (representación independiente del idioma)
 - 2 Re-codificar en lengua destino
 - 3 SYSTRAN

Traducción automática (*MT, Machine Translation*) (cont.)

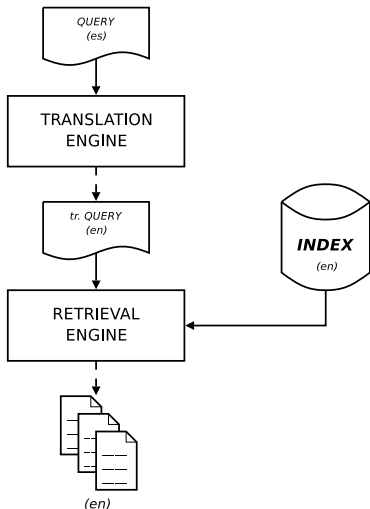
- Subtareas involucradas:
 - Reconocimiento de entidades
 - Desambiguación del sentido de las palabras (WSA)

Recuperación de Información Interlingüe (*CLIR, Cross-Language Information Retrieval*)

Caso particular de IR donde consultas y documentos en diferentes idiomas

- Aplicación de técnicas de MT
 - Restricciones menos estrictas
 - No limitados a 1 traducción: podemos combinar varias
 - No limitados por sintaxis
- Herrs. del idioma de Google (empleando Google Translate)

Recuperación de Información Interlingüe (*CLIR, Cross-Language Information Retrieval*) (cont.)



Recuperación de Información Interlingüe (*CLIR, Cross-Language Information Retrieval*) (cont.)

