

Evaluación del Módulo 5.1

Laura Rodríguez Navas
rodrigueznava@posgrado.uimp.es

Marzo 2020

1. Defina la tarea de Extracción de Información (2 puntos).

La Extracción de Información (EI) es un tipo de recuperación de la información cuyo objetivo es extraer automáticamente fragmentos de texto relevantes, los cuales pueden ser un nombre, un evento o una acción, en definitiva, una entidad preespecificada, y almacenar las entidades extraídas en una base de datos. En resumen, convertir información no estructurada en información estructurada.

Típicas tareas de la extracción de información:

- Reconocimiento de nombres de entidades. Esta tarea se basa en buscar, localizar y clasificar elementos atómicos en texto (nombres) sobre categorías predefinidas (entidades).
- Resolución de la correferencia. Esta tarea está restringida a encontrar vínculos entre las entidades de nombres que se han extraído previamente.
- Extracción de terminología. Esta tarea se basa identificar y extraer candidatos a términos de los textos explorados.
- Extracción de relaciones. Esta tarea se basa en detectar y clasificar menciones a relaciones semánticas.

La mayor aplicación de la extracción de información es la construcción de sistemas que encuentren y relacionen información relevante mientras ignoran otras informaciones no relevantes. Estos sistemas deben trabajar desde el reconocimiento de palabras hasta el análisis de frases y desde el entendimiento a nivel de frase hasta el texto completo.

2. Atendiendo a la lección ¿es cierto que la extracción de información no requiere de la definición de la información relevante a extraer? Razone su respuesta (2 puntos).

No es cierto.

Como se ha comentado anteriormente, la mayor aplicación de la extracción de información es la construcción de sistemas que encuentren y relacionen información relevante. Para determinar si esta información es relevante, se hace a partir de las guías predefinidas de dominio, las cuales deben especificar con la mayor exactitud posible el tipo de información a extraer. Así que, la extracción de información requiere de la definición de la información a extraer para especificar con la mayor exactitud las guías predefinidas de dominio necesarias para la construcción de los sistemas.

3. Tras leer el artículo **Open Information Extraction from the Web** , indique las diferencias existentes con la tarea de **Extracción de Información**. (4 puntos).
4. Atendiendo a la lección, ¿podría indicar cuáles han sido las principales conferencias relacionadas con la **Extracción de Información**? (2 puntos)

Se desarrollaron 7 conferencias para la comprensión de mensajes financiadas por DARPA , la Agencia de Investigación Avanzada de la Defensa Proyectos, para fomentar el desarrollo de nuevos y mejores métodos de extracción de información. Estas conferencias tuvieron como objetivo el determinar un régimen de evaluación cuantitativo para los sistemas de extracción de información, estableciendo las tareas a llevar a cabo y el conjunto de textos sobre los que se deberían realizar dichas tareas.

Concretamente las conferencias fueron:

- MUC-1 (Mayo 1987) sobre operaciones navales.
Participaron 6 sistemas.
No hubo definición de tareas ni se establecieron medidas de evaluación.
- MUC-2 (Mayo 1989) sobre operaciones navales.
Participaron 8 sistemas.
Se definió una plantilla y las reglas para el relleno de los diferentes atributos.
Se definieron unos criterios de evaluación que se desecharon al no ser óptimos.
- MUC-3 (Mayo 1991) sobre atentados terroristas en América Latina.
Participaron 15 sistemas.
Se adaptaron los criterios de evaluación que se utilizaban en Recuperación de Información, las medidas Precision y Recall.
- MUC-4 (Junio 1992) sobre atentados terroristas en América Latina.
Participaron 17 sistemas.
Se definieron criterios de evaluación independientes de los textos, la medida F-measure (combinación de las medidas Precision y Recall).
- MUC-5 (Agosto 1993) sobre fusiones de empresas y anuncios de productos microelectrónicos.
Participaron 17 sistemas.
Se definieron dos idiomas, el inglés y el japonés.
- MUC-6 (Noviembre 1995) sobre fusiones de empresas y anuncios de productos micro-electrónicos.
Participaron 17 sistemas.
El objetivo fue la modularidad y la portabilidad de los sistemas.
Se recuperaron las medidas de evaluación: Precision y Recall.
Se definieron las tareas:
 - Reconocimiento de entidades (RE).
 - Resolución de correferencias (CO).
 - Plantillas de elementos (PE).
 - Plantillas de escenarios (ES).
- MUC-7 (Abril 1998) sobre accidentes de aviones y lanzamientos de misiles.
Participaron 17 sistemas.
Se definió una nueva tarea, la Relación de plantillas (RP).