Evaluación del Módulo 5.1

Laura Rodríguez Navas
rodrigueznavas@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).

Tradicionalmente, la extracción de información (IE) se ha centrado en satisfacer la precisión, preespecificando pequeños corpus homogéneos. Cambiar a un nuevo dominio requiere que el usuario nombre las relaciones y manualmente cree nuevas tareas de extracción. Este artículo presenta un nuevo modelo de extracción donde el sistema solo hace un recorrido por los datos del corpus y extrae un gran conjunto de tuplas sin ninguna interacción humana.

TODO

Este documento presenta la extracción de información abierta (OIE) - Un nuevo paradigma de extracción que facilita el dominio independiente descubrimiento de relaciones extraídas de texto y se adapta fácilmente a la diversidad y el tamaño del corpus web. La única entrada a un sistema OIE es un corpus, y su salida Es un conjunto de relaciones extraídas. Un sistema de la OIE hace un solo pasar sobre su cuerpo garantizando la escalabilidad con el tamaño de el corpus

La extracción de información (IE) se ha basado tradicionalmente en participación humana en forma de extracción artesanal reglas o ejemplos de entrenamiento etiquetados a mano. Además, el usuario debe especificar previamente de forma explícita cada relación de interesar. Mientras que IE se ha vuelto cada vez más automatizado tiempo, enumerando todas las posibles relaciones de interés para la extracción por un sistema IE es muy problemático para corpus como grande y variada como la web. Para que los usuarios puedan emitir consultas diversas sobre corpus heterogéneos, sistemas IE debe alejarse de las arquitecturas que requieren relaciones con ser especificado antes del tiempo de consulta a favor de aquellos que apuntan a Descubre todas las relaciones posibles en el texto. En el pasado, IE se ha utilizado en cuerpos pequeños y homogéneos. como historias de noticias o anuncios de seminarios. Como Como resultado, los sistemas IE tradicionales pueden confiar en la lingüística "pesada" tecnologías sintonizadas con el dominio de interés, como analizadores de dependencia y reconocedores de entidades con nombre (NER). Estos sistemas no fueron diseñados para escalar en relación con el tamaño del corpus o el número de relaciones extraídas, ya que ambas Los parámetros fueron fijos y pequeños. El problema de extraer información de la Web viola Todos estos supuestos. Los cuerpos son masivos y heterogéneos, las relaciones de interés no se anticipan, y Su número puede ser grande. A continuación, consideramos estos desafíos. con más detalle.

Este documento presenta Open IE desde la Web, un sitio sin supervisión paradigma de extracción que evita la extracción específica de relación a favor de una sola pasada de extracción sobre el cuerpo durante el cual las relaciones de interés se descubren automáticamente y eficientemente almacenado. A diferencia de los sistemas IE tradicionales que incurrir repetidamente en el costo del análisis de corpus con el nombramiento de cada nueva relación, el descubrimiento de relación única de Open IE procedimiento permite a un usuario nombrar y explorar relaciones en Velocidades interactivas. El documento también presenta TEXTRUNNER, una aplicación totalmente implementada. Sistema IE abierto, y demuestra su capacidad para extraer cantidades masivas de información de alta calidad de un corpus de nueve millones de páginas web. Hemos demostrado que TEXTRUNNER puede igualar el retiro del KNOWITALL Sistema IE web de última generación, al tiempo que se logra una mayor precisión.

4. Atendiendo a la lección, ¿podría indicar cuáles han sido las principales conferencias relacionadas con la Extracción de Información? (2 puntos).

Se desarrollaron 7 conferencisa 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 Precison 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-mesure (combinación de las medidas Precison 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).