Extracción de información

Durante los últimos treinta años, la extracción de información, se ha convertido en una de las principales áreas de investigación en el área de la gestión de la información. A diferencia de la recuperación de información tradicional, en la que el objetivo es recuperar un documento que sea relevante a la respuesta de una pregunta dada. La extracción de información trata de obtener la respuesta directamente de los documentos. Los datos no tienen por qué estar escritos, solamente, en lenguaje natural, sino que podrían estar en metadatos, tablas u otro soporte. Puesto como ejemplo, a la pregunta quién escribió El Quijote, un sistema tradicional devolvería documentos que son relevantes para saber la autoría de la obra, un sistema de extracción debería responder solamente "Miguel de Cervantes". Y asi, una tarea típica, sería rellenar un formulario asociado a un documento, cuyos campos puedan ser, por ejemplo, autor, título y fecha de edición.

Dado el bajo rendimiento que tenían estos sistemas en los primeros años, se decidió reducir el tipo de consultas a las que se podía dar respuesta. Básicamente, estos tipos básicos fueron responder al Quién (quién es el autor o entidad responsable del dato o qué personas o entidades se mencionan en relación a un evento), Dónde (qué lugares o emplazamientos son mencionados en relación a un evento o un dato) o Cuando (qué fechas o marcadores temporales se mencionan). Además, se incluyen algunos Qué, en relación a qué tema o artefacto se menciona en cierto tipo de documentos. Para todas estas tareas reducir la ambigüedad propia del lenguaje natural es un objetivo.

Los elementos por los que se recupera suelen estar representados por una denominación con un bajo grado de ambigüedad. A estos elementos se les ha denominado "nombre de entidades", "entidades nombradas" o "named entities". Estas entidades son habitualmente nombres propios, es decir la designación no ambigua de una persona en el lenguaje natural suele hacerse con su nombre y apellidos, ya que utilizar en una conversación habitual el número de identificador no es práctico. Un problema asociado sería la designación múltiple de la misma entidad: "M. de Cervantes", "Cervantes" o "Cervantes Saavedra, M."

Una vez se identifican estas entidades, la siguiente tarea es encontrar relaciones entre las mismas. Por ejemplo, <entidad: Miguel de Cervantes><fue_el_autor_de_la_novela><entidad: El Quijote>. Asi entre las dos entidades podemos determinar que existe una relación de autoría.

Por supuesto el lenguaje tiene problemas adicionales debido a su ambigüedad, por ejemplo, uso de sinónimos, términos polisémicos, pronombres o anáforas de todo tipo. Este es uno de los motivos por los que estos sistemas utilizan tanto análisis del lenguaje (PLN) como tesauros y ontologías. La técnica fundamental para extraer la información es detectar patrones. Esto es identificar una secuencia de términos o categorías gramaticales que asocian entidades, p.e. <fue_el_autor_de_la_novela> se podría interpretar como una asociación de tipo "autoría". Puede haber elementos opcionales, p.e. <fue_el_autor_de > conduce a la misma relación.

Para estimar la relevancia de estos patrones se suele utilizar una carga variable de etiquetado a mano, con el fin de aprender. Y otra carga variable de aprendizaje estadístico, con el fin de detectar los patrones relevantes. Sobre el resultado se estimará el ruido y silencio, vistos en anteriores temas.

Como sistema para comparar la eficacia de los sistemas se establecieron competiciones, esto es, foros en los que se proponen un conjunto de tareas para ejecutar sobre un conjunto de documentos y comparar la eficacia con unas métricas predefinidas.