Creado el 30/01/2014
Última versión: 30/01/2014
Carlos Gómez Rodríguez, Universidade da Coruña
Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

Enfoques y recursos básicos/NIVELES DE ANÁLISIS DEL LENGUAJE

Análisis sintáctico

Hola, soy Carlos Gómez Rodríguez, profesor de la Universidade da Coruña y miembro del grupo de investigación en Lengua y Sociedad de la Información de esta universidad. Bienvenidos a la lección sobre análisis sintáctico.

En esta lección vamos a aprender tres cosas: primero veremos qué es el análisis sintáctico, a continuación para qué sirve, y por último, cómo se puede llevar a cabo automáticamente en un ordenador.

Empezamos, pues, viendo qué es el análisis sintáctico. Se trata de una tarea que consiste en determinar la estructura de una oración en lenguaje natural. Esta estructura se puede representar de diferentes maneras, dependiendo de la teoría lingüística que queramos usar. Pero hay dos maneras de hacerlo especialmente populares, que son el análisis sintáctico de constituyentes y el análisis sintáctico de dependencias.

El análisis de constituyentes se basa en dividir la oración en las partes que la componen, que se llaman constituyentes, de forma que cada una de esas partes se va dividiendo a su vez en partes más pequeñas hasta llegar a las palabras. Por otro lado, el análisis de dependencias se basa en buscar cuáles son las relaciones entre las distintas palabras de la oración. Pero veremos mejor todo esto con un ejemplo.

Supongamos que queremos analizar la oración "Juan comió la manzana".

Antes de proceder al análisis sintáctico propiamente dicho, normalmente haremos una fase previa de análisis léxico y morfológico, que habéis visto en las lecciones anteriores. Esto nos proporciona información sobre cada una de las palabras de la oración individualmente, cosa que nos ayudará a ver cómo se agrupan o se relacionan en el análisis sintáctico. Por ejemplo, aquí el análisis léxico y morfológico nos estaría diciendo que "comió" es un verbo en tercera persona del singular y está en pasado, o que "manzana" es un nombre común, femenino singular.

Toda esa información es la que serviría de entrada a un analizador sintáctico, que nos podría dar un análisis como éste. Podéis ver que se trata de un análisis de constituyentes porque lo que hace es dividir la oración en partes: nos está diciendo que la oración se compone de un sintagma nominal (Juan) y un sintagma verbal (comió la manzana). El sintagma nominal sólo tiene dentro un sustantivo que es Juan, mientras que el sintagma verbal tiene un núcleo verbal

(comió) del que "cuelga" otro sintagma nominal (la manzana).

[transparencia 7]

Por otra parte, si en vez de un analizador de constituyentes utilizáramos uno de dependencias, obtendríamos algo como lo que vemos aquí. En este análisis, estamos relacionando las palabras entre sí mediante flechas, que es lo que llamamos una relación de dependencia. Por ejemplo, decimos que la palabra Juan depende de comió, y la etiqueta encima de la dependencia nos aclara cómo: Juan es el sujeto de comió. De la misma forma, este análisis nos da la información de cuál es el objeto de comió: la manzana.

¿Cuál es la utilidad de un análisis sintáctico como éste?

El análisis sintáctico nos proporciona información útil para extraer significado (semántica) de la oración. Por ejemplo, de este análisis de dependencias podemos obtener la información de que la oración describe una acción, comer, que se realizó en el pasado. Para saber quién fue la persona que comió miraremos el sujeto (que es Juan), y para saber qué fue lo que se comió, miraremos el objeto (que es la manzana). Por lo tanto, el análisis sintáctico es lo que nos permite ver cómo los significados individuales de cada palabra se relacionan para formar un todo con sentido, y así enterarnos de qué dice realmente la oración.

Esto va a ser útil en gran cantidad de aplicaciones prácticas de procesamiento del lenguaje natural, dado que nos permite interpretar significados en las oraciones que nunca podríamos extraer viéndolas como una secuencia de palabras sin estructura interna. Por ejemplo, supongamos un sistema automático de minería de opiniones, que busca determinar si una opinión sobre un producto es positiva o negativa. Las opiniones "es bueno y no es caro" y "no es bueno y es caro" tienen exactamente las mismas palabras, con lo cual si extraemos la opinión metiendo todas las palabras en una bolsa sin estructura (lo que se llama modelo bag-of-words) ambas oraciones darán el mismo resultado. Sin embargo, vemos que en realidad expresan opiniones contrarias. La diferencia no nos las da el análisis individual de las palabras, sino el análisis sintáctico, que va a ser lo que nos diga que la palabra "no" en la primera opinión está negando "es caro", mientras que en la segunda niega "es bueno", cambiando totalmente el sentido.

De la misma manera que en minería de opiniones, detectar matices como éste es imprescindible en otras muchas aplicaciones de procesamiento de lenguaje natural, todas en las cuales haga falta extraer alguna semántica de las oraciones. Por ejemplo, en recuperación de información, búsqueda de respuestas, extracción de información o traducción automática.

Ahora que espero haberos convencido de que el análisis sintáctico es muy útil, nos podemos preguntar ¿cómo podemos llevarlo a cabo automáticamente con un ordenador?

Para esto, primero es importante observar que para analizar sintácticamente frases en un idioma, hace falta conocer dicho idioma. Por ejemplo, si leemos la oración "Juan comió la manzana", nosotros sabemos que la manzana es el objeto porque en español las oraciones tienden a seguir un orden "sujeto-verbo-objeto". Pero esto es diferente por ejemplo en alemán, donde el verbo se pone al final.

Por lo tanto, para crear un analizador sintáctico tendremos que enseñarle de alguna manera al

ordenador cuáles son las normas de funcionamiento del idioma (las reglas de la sintaxis).

Esto se puede hacer al menos de dos maneras. La más tradicional es escribir esas reglas explícitamente, lo que se denomina una *gramática*, y dárselas al programa. Entonces el analizador sintáctico podrá construir una estructura de la oración viendo cómo encajan las reglas gramaticales con las palabras de la misma. Son los modelos dirigidos por gramática.

Por ejemplo, aquí vemos cómo un algoritmo concreto dirigido por la gramática, llamado CYK, construiría el árbol de constituyentes del ejemplo. A la izquierda tenemos una serie de reglas gramaticales. Por ejemplo, una de ellas [transparencia 13] nos dice que un sintagma nominal puede componerse de un nombre, y podemos "encajarla" con el nombre Juan. Otra [transparencia 14] nos dice que un sintagma nominal también puede tener un determinante y un nombre, lo cual encajamos con "la manzana". A su vez, [transparencia 15] podemos enlazar "comió" con "la manzana", mediante otra regla que nos define un sintagma verbal, y por último [transparencia 16] juntar el nominal con el verbal para obtener [transparencia 17] el árbol completo de constituyentes.

Pero como escribir una buena gramática es muy complicado y costoso, otra manera muy en auge es hacer uso del aprendizaje automático: usar técnicas estadísticas para mostrarle al programa muchas oraciones analizadas (un corpus), y que a partir de ellas *aprenda* automáticamente a analizar oraciones nuevas. Son los modelos dirigidos por los datos.

También existen aproximaciones híbridas entre ambas, como por ejemplo aprender reglas gramaticales mediante aprendizaje automático, o partir de una gramática pero usar técnicas de aprendizaje para decidir qué regla usar en un momento dado.

Pero sea cual sea la aproximación que utilicemos, los desafíos que nos vamos a encontrar vienen de la mano de las características de los lenguajes humanos.

Por un lado, la expresividad y riqueza de los lenguajes naturales hace imposible que podamos tener una gramática o un corpus exhaustivos, que contengan todos los fenómenos sintácticos que pueden aparecer en el idioma, y más aún teniendo en cuenta que las lenguas no se mantienen fijas en el tiempo sino que son algo dinámico.

Por otro lado, los lenguajes naturales son ambiguos, y a nivel sintáctico esto se manifiesta en que la misma oración puede tener dos análisis diferentes, correspondiente a distintas interpretaciones. [transparencia 20] Esto sucede, por ejemplo, con la frase "Juan vio a un hombre con un telescopio": podríamos referirnos a que Juan utilizó un telescopio para ver al hombre, o bien a que Juan vio a un hombre que llevaba un telescopio, y en cada caso tendremos un análisis sintáctico distinto. Para decidir cuál es el que procede, necesitaríamos conocer el contexto de la oración.

De ahí que los analizadores sintácticos automáticos; aunque ya presentan una precisión suficientemente buena para ser útiles en muchas tareas de procesamiento del lenguaje natural, son componentes que están en constante mejora y que plantean todavía muchos desafíos de investigación.

Sin más, me despido recordándoos que en el documento asociado a esta lección tenéis

actividades y referencias bibliográficas que amplían esta información.