Técnicas básicas en el tratamiento informático de la lengua

Basic techniques in computerized language processing

Horacio Rodríguez

Los procesadores lingüísticos son parte necesaria en la mayoría de los sistemas que incluyen alguna forma de tratamiento de la lengua. El autor describe los procesos iniciales que se desarrollan en el tratamiento de textos: nivel superficial, morfológico y presintáctico. Son los pasos previos a la interpretación semántica limitados al ámbito del tratamiento de textos escritos.

Linguistic processors are an essential part of most systems that include some sort of language processing. The author describes the initial processes being developed in text treatment: superficial, morphological and pre-syntactic levels. These are the preliminary steps to semantic interpretation limited to the treatment of written texts.

La ingeniería lingüística es la aplicación del conocimiento de la lengua al desarrollo de sistemas informáticos capaces de reconocer, comprender, interpretar y generar lenguaje humano en todas sus formas. La ingeniería lingüística comprende una amplia gama de métodos, técnicas y herramientas; hace uso de una gran cantidad de recursos, lingüísticos y de otros tipos, y se utiliza de forma creciente en un gran número de aplicaciones.

A un nivel muy general podemos considerar que una aplicación de ingeniería lingüística incluye alguna forma de comprensión del lenguaje humano, algún tipo de proceso de la información adquirida y eventualmente una comunicación del resultado del proceso, a veces también utilizando alguna forma de lenguaje humano. Un ejemplo claro es el de un *sistema de traducción automática*, donde un texto en una lengua fuente es procesado hasta lograr un nivel de comprensión suficiente. La información extraída, expresada en algún lenguaje de representación adecuado es transferida al sistema de representación equivalente en la lengua objetivo y desde aquí un proceso de generación logrará la traducción final en dicha lengua objetivo.

Por supuesto, no todas las aplicaciones de la ingeniería lingüística incluyen estos elementos. Así, un sistema de generación de cartas personalizado no precisa ningún tratamiento de comprensión, o un sistema de identificación de la lengua o un detector de errores ortográficos no necesitan generar lenguaje humano. La mayoría de las aplicaciones incluyen, sin embargo, alguna forma más o menos precisa de comprensión: así, un sistema de consulta en lenguaje humano a una base de datos precisará un nivel muy alto de comprensión de las expresiones del interlocutor humano para que la respuesta del sistema sea de utilidad, en cambio en un sistema de traducción o de resumen automáticos se pueden lograr niveles de corrección muy notables con niveles de comprensión bajos. Es decir, no es preciso comprender totalmente una oración para ser capaz de traducirla correctamente.

La inmensa mayoría de los sistemas que incluyen alguna forma de tratamiento de la lengua requieren la actuación de una serie de procesadores lingüísticos. La descripción lingüística del material a tratar, sea textual u oral, suele organizarse en forma estratificada: nivel fonológico, nivel fonémico, nivel textual, nivel morfológico, nivel léxico, nivel sintáctico, nivel lógico, nivel semántico, nivel pragmático, etc., y a cada uno de ellos se le suele asociar una colección de tales procesadores que deben resolver, en forma aislada o colaborativa, los problemas de tratamiento de la información propia del nivel. Así, en un sistema de comprensión de un documento escrito, el nivel textual del documento estará constituido por los caracteres que aparecen en el mismo y las tareas a llevar a cabo incluirán la separación del texto de la metainformación eventualmente presente, la segmentación del texto en unidades a tratar, párrafos, oraciones, etc., la localización de las palabras ortográficas, etc. En el nivel sintáctico, los elementos a tratar son palabras dotadas ya de información morfosintáctica (categoría, propiedades morfosintácticas) y lo que se pretende es explicitar las relaciones sintácticas subyacentes.

Aunque, como se ha indicado antes, la complejidad del tratamiento, y por lo tanto la de los procesadores implicados, depende de la aplicación concreta, es evidente que los procesos iniciales: tratamiento textual, consulta en diccionarios, análisis y desambiguación morfosintáctica, etc., aparecen en prácticamente todos los sistemas. De estos procesos iniciales del tratamiento de los textos, previos al análisis sintáctico o a la interpretación semántica, nos ocupamos en este trabajo que además queda limitado al ámbito del tratamiento de textos escritos.

Tradicionalmente, las aproximaciones usadas para el tratamiento de la lengua se basaban en el

conocimiento lingüístico. Recientemente, sin embargo, han resurgido con éxito los sistemas empíricos, basados no en el conocimiento lingüístico sino en la modelización del comportamiento lingüístico a partir de su constatación empírica. Este resurgir es debido a factores como la disponibilidad de recursos como los corpus lingüísticos, la capacidad de proceso de los sistemas informáticos y la necesidad de aplicaciones de tratamiento de textos no restringidos. La influencia del desarrollo de Internet ha sido decisiva a este respecto. Dentro de los métodos empíricos cabe mencionar los de base estadística y últimamente los basados en técnicas de aprendizaje automático (*Machine Learning*). Los sistemas híbridos pretenden recoger lo mejor de estas dos tradiciones. De todo ello nos ocuparemos en los siguientes apartados de este artículo.

Análisis textual

La primera etapa de cualquier sistema de procesamiento de la lengua tiene lugar a nivel textual. En este nivel, el texto puede ser considerado como una simple secuencia de caracteres. Las tareas básicas que deben abordarse a este nivel son: *a)* la segmentación del texto, *b)* el filtrado de información no relevante y *c)* la localización de unidades tratables. A pesar de su aparente simplicidad estas tareas presentan notables complicaciones que describiremos someramente:

Segmentación del texto. El texto debe ser segmentado en fragmentos que puedan tratarse de forma hasta cierto punto independiente. La dificultad de la tarea depende tanto de las características de los fragmentos a obtener (párrafos, oraciones, intervenciones de diversos interlocutores, etc.), como de la fuente de la cual se obtienen (texto marcado, texto plano, resultado de una transcripción a partir de voz, etc.). Si deseamos segmentar un texto en párrafos y disponemos de signos de puntuación y somos capaces de distinguir mayúsculas y minúsculas, entonces la tarea es relativamente sencilla (aun cuando un signo de puntuación puede cumplir funciones diferentes de la de separación; por ejemplo, un punto puede formar parte de un nombre propio, de una fórmula o de un acrónimo). Si deseamos segmentar un texto en oraciones, la tarea es obviamente más difícil y será necesario algún tipo de conocimiento lingüístico, lo mismo que si la fuente a tratar es producto de una transcripción y carece de signos de puntuación y de distinción entre mayúsculas y minúsculas.

Filtrado de información no relevante. Los textos a tratar vienen a menudo acompañados de otros materiales que deben ser eliminados o extraídos (a veces el uso de esta metainformación ejerce un papel relevante en el procesado de los textos) para facilitar el tratamiento. Así, si la fuente de información es una página de Internet, junto a los fragmentos de texto tratables, nos aparecen diferentes tipos de marcas que definen las características de visualización de la página, enlaces con otras páginas o dentro de la misma página, objetos (o apuntadores a los mismos) no textuales, como por ejemplo imágenes, animaciones, tablas, programas, etc. Si el texto está marcado de forma consistente, el filtrado es relativamente sencillo, pero, a menudo, ello no es así y la dificultad de la tarea aumenta. Podemos situar aquí también la tarea de la identificación de la lengua en los casos en que se desconozca *a priori* la lengua en que un texto está escrito, por ejemplo si el texto ha sido recuperado a través de Internet, o en los casos en que un texto contenga fragmentos correspondientes a lenguas diferentes.

Localización de unidades tratables. Las unidades básicas de tratamiento son las palabras. Localizar las palabras ortográficas es sencillo si el espacio o los signos de puntuación actúan como separadores. En las lenguas en que ello no es así, o en los casos en que la puntuación no existe el problema es mayor. Pero aún en los casos en que se han localizado de forma precisa las palabras ortográficas existen casos problemáticos. Podemos citar algunos:

• Distinción entre palabras ortográficas y palabras gramaticales. Casos como la conjunción «sin embargo» en que una palabra gramatical corresponde a dos ortográficas o «dímelo» en que una palabra ortográfica corresponde a tres palabras gramaticales, o contracciones como «del» o «al».
• Términos multipalabra, locuciones, lexías, frases hechas, etc.
• Fechas, fórmulas, siglas, jerga, abreviaturas, etc.
• Nombres propios (de persona, geográficos, etc.).

• Palabras desconocidas. Es decir palabras que no figuran en los diccionarios de que se dispone. Pueden tratarse de barbarismos, de neologismos o simplemente de errores.

El uso de lexicones específicos (frasales o terminológicos) que complementen a los lexicones generales, de procesadores especializados para tratar las unidades no estándar (por ejemplo extractores de fechas o identificadores de nombres propios), de autómatas o transductores de estados finitos y de técnicas estadísticas y de aprendizaje automático contribuye a solucionar todos estos problemas.

Análisis morfosintáctico

El siguiente paso en el tratamiento de la lengua consiste en la asignación de información a las unidades elementales que se han localizado en la etapa anterior. El tipo de información suele consistir en: *a)* las categorías morfosintácticas válidas para la unidad, *b)* otros tipos de información morfosintáctica, como los rasgos morfológicos: género, número, persona, etc., *c)* las categorías semánticas, por ejemplo las acepciones, etiquetas de dominio u otras clases semánticas, *d)* otras informaciones semánticas, como las preferencias semánticas sobre los argumentos o los patrones de subcategorización, y *e)* información de tipo empírico como las probabilidades léxicas.
La manera más habitual de obtener esta información es sencillamente a través de consultas a uno o varios lexicones y diccionarios (es frecuente disponer de lexicones morfosintácticos generales y de lexicones semánticos específicos de un dominio). A veces es necesario, sin embargo, recurrir a algún tipo de análisis morfológico.
El papel de un analizador morfológico es el de recuperar la morfología de las palabras, es decir recorrer en sentido inverso el mecanismo de formación de la palabra. Generalmente, las palabras se forman a través de mecanismos de flexión, derivación o composición a partir de unas formas básicas. Así, a partir de la raíz «tembl» (o del lema «temblar») se pueden construir las formas flexivas del verbo añadiendo sufijos flexivos («temblaba», «temblado», etc.) o las formas derivadas («temblorosos»).
El análisis morfológico consiste en la descomposición de la palabra a analizar en una serie de fragmentos: raíz, prefijos, sufijos, en algunas lenguas infijos, pertenecientes a conjuntos finitos y siguiendo determinadas reglas de combinación. La palabra «temblorosamente» se puede descomponer en «tembl + oros + a + mente», de forma que «tembl» perteneciera al conjunto de raíces, y «oros», «a» y «mente» a conjuntos de sufijos de diversos tipos. Una determinada regla autorizaría la combinación.
El análisis morfológico de las formas flexivas es relativamente sencillo, ya que la flexión responde a patrones bastante regulares y, además, la categoría gramatical y el significado no cambian (aunque a veces se refinan) al añadirse los sufijos. Para el castellano o el catalán, el número de sufijos flexivos es de unos 200 y el número de reglas de combinación de unas 500. En cambio, la derivación o la composición son más complicadas y suelen venir combinadas con la flexión.
A menudo, sobre todo para lenguas con poca complejidad morfológica, como la inglesa, el analizador morfológico se reduce a un *formario*, es decir a un diccionario de formas completas. Si el formario está implementado razonablemente, la eficiencia del proceso de análisis es alta. Por otra parte, los formarios son fácilmente extensibles, soportan entradas multipalabra y es posible su construcción a partir de generadores morfológicos.
En otras ocasiones, la representación explícita de todas las formas no es conveniente y se debe recurrir a un proceso de análisis morfológico realizado no *a priori*, sino en el momento de llevar a cabo el tratamiento de la oración. Consideremos el caso del castellano. La capacidad flexiva de tres de las cuatro categorías principales (nombre, adjetivo y adverbio) es muy reducida. En cambio, el verbo tiene una alta capacidad flexiva (unas 40 formas por lema). Si tenemos en cuenta que existen unos 5000 verbos que se usan con cierta frecuencia el número de formas verbales que debiera incluir un formario del castellano sería de unas 200 000. Esta cifra puede ser aceptable para algunas aplicaciones y no serlo para otras. Otros fenómenos morfológicos difícilmente reducibles a colecciones de formas son la derivación, las formas verbales que incluyen pronombres enclíticos («dímelo», «diríaselo»), los adverbios de modo acabados en «mente», derivados de adjetivos femeninos, etc. En estos casos, un analizador morfológico que actúe en el momento del tratamiento debe sustituir o completar al formario.
Existen dos grandes familias de analizadores morfológicos: los analizadores de un nivel y los analizadores de dos niveles. Los primeros consideran un solo nivel de representación: el superficial (el de las palabras),

mientras que los segundos distinguen un nivel léxico (el de los lexemas) y otro superficial. Para los analizadores de un solo nivel, el análisis consiste simplemente en determinar las posibles particiones de una palabra en secuencias válidas de caracteres tal como indicamos anteriormente. Es habitual el uso de autómatas de estados finitos para implementar de forma eficiente este tipo de analizadores. El problema que presentan es la dificultad de hacer frente a las variaciones fonológicas, es decir los cambios en los fragmentos cuando se producen determinadas combinaciones. Estos problemas han dado lugar a los sistemas de dos niveles en los cuales el nivel léxico establece las combinaciones válidas y el nivel superficial su realización en forma de palabras. Un conjunto de restricciones establece las correspondencias entre estos dos niveles. En este caso se suelen utilizar los denominados transductores de estados finitos.

Desambiguación morfosintáctica (*pos tagging*)

El resultado del análisis morfológico es el de asignar a cada una de las unidades léxicas presentes el conjunto de sus categorías gramaticales posibles. El problema es que las palabras tomadas en forma aislada son ambiguas respecto a su categoría. Consideremos el siguiente ejemplo: «Yo bajo con el hombre bajo a tocar el bajo bajo la escalera». La palabra «bajo» puede tener, dependiendo del conjunto de etiquetas que se manejen, un mínimo de cuatro categorías diferentes: verbo, adjetivo, nombre y preposición. El analizador morfológico devolverá toda ellas para cada una de las apariciones de la forma «bajo» en la oración. Afortunadamente la categoría de la mayoría de las palabras no es ambigua dentro de un contexto. Es relativamente simple para una persona eliminar la ambigüedad en la categorización para establecer que las apariciones de «bajo» corresponden respectivamente a un verbo, un adjetivo, un nombre y una preposición. La misión de los desambiguadores morfosintácticos (*pos taggers*) es la de realizar automáticamente esta tarea.
El objetivo de un desambiguador (también llamado *etiquetador morfosintáctico*) es, pues, el de asignar a cada palabra la categoría más *apropiada*, dentro de un contexto. Es decir, dada una secuencia de palabras, dotada cada una del conjunto de etiquetas posibles, el desambiguador deberá devolver una secuencia de etiquetas que sea la más *verosímil* dado el contexto. Por supuesto, la calidad del desambiguador dependerá del grado de precisión (la *granularidad*) del etiquetado, del contexto considerado y de la información de que disponga el desambiguador para considerar *apropiada* una etiqueta o *verosímil* una secuencia de etiquetas. A veces, los desambiguadores no resuelven totalmente el problema de la ambigüedad gramatical y se limitan a eliminar las opciones imposibles o menos probables. Es el caso de los denominados *desambiguadores reduccionistas*.
Existen tres grandes familias de etiquetadores: los basados en reglas, los estadísticos y los híbridos. Los *etiquetadores basados en reglas* utilizan conocimiento lingüístico (*knowledge-driven taggers*), generalmente expresado en forma de reglas o restricciones para establecer las combinaciones de etiquetas aceptables o prohibidas. Las reglas suelen construirse manualmente, responden a criterios lingüísticos y se representan en forma explícita. Se trata de sistemas de muy alta precisión (por ejemplo, el EngCG de Karlsson, que implementa las denominadas *Constraint Grammars* para el inglés, comprende un millar de reglas y supera el 99,5 % de corrección). El coste de desarrollo es alto y también lo es el coste de adaptación a otros dominios.
Los *desambiguadores estadísticos* obtienen los modelos del lenguaje y generalizaciones en que basan su actuación automáticamente a partir de la evidencia empírica obtenida de corpus lingüísticos voluminosos (*data-driven taggers*). El coste es por ello mucho menor aunque también es menor su grado de precisión, superior en cualquier caso al 97 %, suficiente en algunas aplicaciones. Los sistemas son independientes de la lengua y fácilmente transportables a lenguas y dominios diversos con un coste limitado. El problema de estos sistemas reside en el aprendizaje (estimación) de los parámetros del modelo estadístico utilizado. En este sentido es notable, y creciente, el uso de técnicas de aprendizaje automático (*machine learning*). Se han utilizado técnicas de aprendizaje supervisado partiendo de corpus desambiguados manualmente y técnicas de aprendizaje no supervisado en las que no es precisa (o está limitada) esa intervención manual. Se han utilizado modelos muy sencillos del lenguaje, como por ejemplo los bigramas en los que la probabilidad de una etiqueta se estima simplemente con el contexto de la etiqueta anterior, o los trigramas en los que el contexto abarca a las dos etiquetas precedentes; y también modelos más complicados como el utilizado por E. Brill, basado en el aprendizaje de las transformaciones que deben corregir los errores de un desambiguador simple, el uso de árboles de decisión por parte de L. Márquez o el de aprendizaje basado en

casos por parte de W. Daelemans.
En tercer lugar, los *sistemas híbridos* combinan varias fuentes de conocimiento, estadísticas y lingüísticas para intentar recoger los aspectos positivos de cada una de ellas y evitar sus limitaciones. El uso de las técnicas de máxima entropía por A. Rapnaparkhi o la optimización por relajación por L. Padró son ejemplos notables.
Recientemente han comenzado a utilizarse sistemas de desambiguación por combinación. Se trata de combinación de diferentes modelos del lenguaje en un único desambiguador, de combinación de desambiguadores mediante votación u otros procedimientos más sofisticados (*bagging, boosting*), de aprendizaje iterativo (*bootstrapping*) y otros muchos.

Agrupación sintáctica (*chunking*)

Una vez analizado morfológicamente y desambiguado total o parcialmente el texto, podemos plantearnos un análisis sintáctico en profundidad. No obstante hay ocasiones en que este análisis no es posible o no es conveniente y se nos presenta la necesidad de realizar un análisis superficial o fragmental en vez de o como paso previo al análisis sintáctico en profundidad.
Se ha señalado la insuficiencia de los métodos convencionales de análisis sintáctico para tratar textos no restringidos. Problemas como la dificultad de una segmentación adecuada, la obtención de no uno sino varios (a menudo muchos) árboles de análisis o la necesidad de ampliar la cobertura del analizador al tratamiento de oraciones no gramaticales o que incluyan palabras desconocidas tienen difícil solución en el marco tradicional de un analizador sintáctico que trate de obtener un árbol de análisis completo del texto basándose en una gramática de amplia cobertura. Ante ello caben dos aproximaciones: analizar en profundidad pero no todo, caso del *análisis fragmental*, y analizar todo pero no en profundidad, caso del *análisis superficial*.
El *análisis superficial* se aparta del contenido de este artículo y entra más en el correspondiente al análisis sintáctico. Las técnicas de *análisis fragmental* son, en cambio, más próximas a las de desambiguación y a ellas nos referiremos a continuación.
El objetivo de los analizadores fragmentales, también denominados agrupadores sintácticos o *chunkers*, es la detección de frases nominales, verbales, adjetivas, adverbiales básicas (sin recursión). A veces se trata simplemente de detectar el segmento (es lo que se denomina *parentizado* o *bracketting*), mientras que en otras ocasiones se desea obtener el etiquetado correcto y la estructura sintáctica del segmento. Es frecuente el uso de técnicas de estados finitos y la actuación de transductores en cascada. También se han utilizado técnicas de aprendizaje automático. A menudo el siguiente paso es la obtención de dependencias y de relaciones sintácticas entre los segmentos.

Conclusiones

Se ha presentado una serie de técnicas para llevar a cabo las primeras etapas del proceso de textos no restringidos en lenguaje natural. Estas técnicas cubren los niveles superficial, morfológico y presintáctico y su utilización es necesaria en prácticamente todas las aplicaciones de tratamiento de la lengua. Las exigencias de calidad dependen de la aplicación aunque, al tratarse de las etapas iniciales del proceso y debido a la propagación de los posibles errores, suelen ser altas. Aun teniendo en cuenta la aparente sencillez de las tareas existen múltiples técnicas diferentes con niveles de conocimiento lingüístico distintos. Es notable la introducción creciente de las técnicas con base estadística y las que implican aprendizaje automático. Las diferentes prestaciones, grado de corrección de los resultados, eficiencia, recursos, coste de desarrollo, aprendizaje, etc. parecen favorecer formas de combinación entre los métodos y técnicas empleados.
Se trata de un campo de la tecnología lingüística bastante maduro y consolidado. Sin embargo, existen temas abiertos en los que la investigación es muy activa y en los que aún hay margen para mejoras apreciables. Entre ellos podemos citar: la fragmentación del texto en unidades, el reconocimiento de nombres propios, las hipótesis sobre palabras desconocidas, el reconocimiento de expresiones multipalabra, el transporte y, por tanto, afinado de recursos, y la extracción de relaciones sintácticas, entre otros.

Bibliografía

Algunas referencias en castellano o catalán que cubren parcialmente lo aquí presentado se pueden encontrar en:

Badia, T.: «Tècniques de processament del llenguatge» (cap. 7). En: M.A. Martí (ed.): *Les tecnologies del llenguatge*, Barcelona, UOC, Filologia Catalana, 2000.
Rodríguez, H.: «Tratamiento del lenguaje natural» (cap. 10). En: Cortés *et al.* (eds.): *Inteligencia artificial*, Barcelona, Politext Ediciones UPC, 1993.
Rodríguez, H.: «Técnicas estadísticas para el tratamiento del lenguaje natural» (cap. 4). En: J.M. Blecua, G. Clavería, C. Sánchez, J. Torruella (eds): *Filología e informática: nuevas tecnologías en los estudios filológicos,* Seminario de Filología e Informática, UAB, Barcelona, 1999.

Sobre las técnicas estadísticas para el tratamiento de la lengua:
Manning, Ch.D.; Schütze, H.: *Foundations of statistical natural language processing,* Cambridge, Massachusetts, MIT Press, 1999.
Jelinek, F.: *Statistical methods for speech recognition*, Cambridge, Massachusetts, MIT Press, 1998.
Revista *Machine Learning* (vol. 34, num. 1/2/3 February 1999).

Para los temas de análisis morfológico:
Sproat, R: *Morphology and computation*, Cambridge, Massachusetts, MIT Press, 1993.
Martí, M.A.: «Processament informàtic del llenguatge natural: un sistema d'anàlisi morfològica per ordinador, Tesis doctoral, Universidad de Barcelona, 1988.

Para profundizar en las técnicas de estados finitos, tanto autómatas como transductores:
Roche, E., Schabes, Y.: *Finite State language processing*, Cambridge, Massachusetts, MIT Press, 1997.

Sobre la desambiguación morfosintáctica:
Màrquez, L.: «Part-of-speech Tagging: A Machine Learning Approach based on Decision Trees», Tesis doctoral, Departamento de Lenguajes y Sistemas Informáticos, UPC, Barcelona, julio 1999.
Padró, L.: «A hybrid environment for syntax-semantic tagging», Tesis doctoral, Departamento de Lenguajes y Sistemas Informáticos, UPC, Barcelona, 1997.
Karlsson, F.; Voutilainen, A.; Heikkilä, J; Anttila, A. (eds): *Constraint grammar: a language-independent system for parsing unrestricted text,* Mouton de Gruyter, 1993.

Sobre agregación sintáctica:
Young, S.; Bloothooft, G. (eds.): *Corpus-based methods in language and speech processing*, Dordrecht, Kluwer Academic Publishers, 1997.

Una bibliografía más amplia se puede encontrar en http://www.lsi.upc.es/~horacio/

Horacio Rodríguez Hontoria
Ingeniero industrial, licenciado en ciencias físicas y doctor en informática por la Universitat Politècnica de Catalunya (UPC). Desde 1978 es profesor del Departamento de Lenguaje y sistemas informáticos (LSI) de la UPC. Su investigación se ha llevado a cabo en el campo del procesamiento de la lengua, tanto en cuanto a la adquisición de recursos como al desarrollo de técnicas y herramientas. Ha dirigido y participado en varios proyectos de investigación tanto nacionales como internacionales.
horacio@lsi.upc.es
http://www.lsi.upc.es/~horacio/