Lenguas y computación

Antonio Moreno Sandoval

EDITORIAL
SINTESIS

Índice

Prólogo

> Si utilizásemos una metáfora culinaria, podríamos decir que este no es un libro de recetas de cocina, sino de consejos generales para escoger apropiadamente el menú en función de los comensales.

En 1998, la Editorial Síntesis me dio la oportunidad de publicar el libro *Lingüística computacional: una introducción a los modelos simbólicos, estadísticos y biológicos,* de cuyo prólogo están tomadas estas palabras. Dos décadas después, y dentro de una nueva colección dirigida por el profesor Juan Carlos Moreno Cabrera, se me brinda de nuevo la ocasión para hablar de lenguas y sistemas informáticos que procesan lenguas. Para mí es un tema apasionante, al que he dedicado ya unos cuantos años y me gustaría dedicar otros tantos. La disciplina cuenta con más de sesenta años de desarrollo y he tenido la oportunidad de conocer a alguno de sus grandes actores, cuyo magisterio es necesario reconocer y transmitir a las próximas generaciones. Este reconocimiento es justo y útil al mismo tiempo, pues nos permitirá entender mejor los éxitos y fracasos de las diferentes estrategias que han abordado el tratamiento automático de las lenguas naturales.

La lingüística computacional es ya una disciplina madura con un enfoque teórico (o de ciencia básica y fundamental) y otro aplicado (o de ingeniería). La rama fundamental está aportando cada vez más a la lingüística teórica, no solo en conceptos y herramientas de experimentación con teorías formales, sino también en métodos cuantitativos y estadísticos para estudiar los fenómenos de variación. La fabulosa potencia de análisis de datos que proporcionan las aplicaciones computacionales que se utilizan en la lingüística de corpus es un buen ejemplo de ello. La rama ingenieril, por su parte, está contribuyendo a la lingüística aplicada de diversas maneras: con programas que ayudan a la traducción, a los terminólogos y lexicógrafos, a los lingüistas forenses, a los sociolingüistas y a los psicolingüistas. La enseñanza de lenguas por ordenador es otro de los campos que se han beneficiado de los avances de la lingüística computacional.

El objetivo principal de este libro es proporcionar el contexto necesario para entender la relación entre todas las ramas de la lingüística. El futuro de las ciencias del lenguaje está, creo, en la apertura metodológica y conceptual a las aportaciones de diferentes especialidades. La historia de la lingüística computacional proporciona excelentes ejemplos de colaboración multidisciplinar entre lingüistas, informáticos, lógicos, psicólogos, traductores, ingenieros y matemáticos. La complementariedad permite superar los desconocimientos individuales. Y esto es necesario porque la complejidad de formalizar el lenguaje para tratarlo computacionalmente supera los conocimientos individuales, como tendremos ocasión de comprobar a lo largo del libro.

Mi libro de 1998 estaba dirigido a estudiantes de Lingüística y Computación que estuvieran interesados en una introducción al estado de la cuestión a finales de los noventa. El libro que tiene el lector en sus manos está dirigido fundamentalmente a lingüistas, traductores, psicólogos y en general cualquier humanista o científico social que quiera introducirse en los conceptos y métodos para poder interactuar con informáticos. Durante años, en mis exámenes finales de la asignatura Lingüística Computacional, formulaba la siguiente pregunta:

> ¿Qué aproximación le parece más acorde con sus intereses: formarse como lingüista y entenderse con los informáticos o formarse como lingüista computacional autónomo?

La respuesta más frecuente era la primera opción y coincido con esa apreciación. Formarse como lingüista computacional integral supone adquirir conocimientos sustanciales de diferentes especialidades: formalización lingüística en múltiples aspectos, desde el acústico hasta el interpretativo, programación informática eficiente, robusta y distribuida, conocimientos lógicos para la interpretación, así como computación estadística para modelar el aprendizaje automático. Sencillamente, el campo se ha ampliado tanto que es casi imposible abarcarlo de forma unitaria e individual. Es un signo claro de la madurez de la disciplina.

Si en los años ochenta la investigación estaba dirigida por los lingüistas gracias a las aportaciones de las gramáticas de unificación y rasgos, a partir de los noventa la disciplina ha estado dominada por los informáticos y, más en concreto, por los modelos estadísticos. Este cambio metodológico sustancial se puede comprobar en los contenidos de los manuales más conocidos: Grishman (1986) o Gazdar y Mellish (1989) frente a Jurafsky y Martin (2000; 2008). La revolución de los modelos estadísticos ha transformado la lingüística computacional y este libro persigue dotar a los lectores con conocimientos para entender sus conceptos, herramientas e implicaciones. El libro de 1998 incluía una parte dedicada a los modelos estadísticos, pero entonces estaban todavía en sus inicios y ahora presentamos una actualización que recoge la madurez de lo que se ha investigado en los últimos veinte años.

Por otra parte, lo que el lector tiene delante no es un libro de texto con ejercicios para aprender de manera pautada a elaborar gramáticas computacionales, lexicones, corpus anotados, etc. Un buen ejemplo de este tipo de manuales es el de Bird, Klein y Loper (2009). Tampoco es una enciclopedia con capítulos escritos por diferentes autores especialistas en sus dominios, para dar una visión lo más general y completa posible de todo lo que concierne a la lingüística computacional (ejemplos de esto último son los *handbook* de Jurafsky y Martin, 2008, o de Indurkhya y Damerau, eds., 2010).

Este libro pretende ser una fuente para conocer los conceptos y métodos claves en el campo. Se complementa con referencias especializadas y selectas en cada caso. Una parte significativa del libro tratará la conexión entre la lingüística computacional y otras áreas de la lingüística, tanto teórica como aplicada.

La estructura del libro se divide en dos partes. En la primera se abordan los fundamentos, que incluyen los modelos simbólicos y modelos estadísticos (dividido cada uno en dos partes para guardar un equilibrio con la extensión de los otros capítulos), así como los diferentes niveles desde la morfología a la interpretación. La segunda parte se dedica a las aplicaciones: dos capítulos se dedican a la traducción automática y a las distintas aplicaciones de la gestión de información (recuperación, extracción y análisis de opinión). En el último capítulo, de conclusiones, se repasan los hallazgos y limitaciones actuales para hacer luego una previsión de los desarrollos futuros de la disciplina.

Me he tomado la libertad, por sugerencia de la editorial, de traducir al español todas las citas de otros autores, con el objetivo de facilitar la lectura.

Finalmente, unas palabras de agradecimiento. Si bien mis maestros y compañeros en los proyectos iniciales se merecen mi sincero reconocimiento (se aprende más de los pares que de los maestros, como nos decía en clase un catedrático de Teoría de la Literatura), mis últimos veinte años como investigador responsable de los proyectos del Laboratorio de Lingüística Informática de la UAM me han hecho cambiar de perspectiva: los discípulos son los verdaderos responsables de que este libro tenga un alcance multilingüístico y multinivel. Sus tesis doctorales me han permitido trabajar (y aprender) con ellos sobre temas que se habrían escapado a mis posibilidades o intereses. Corren malos tiempos para la ciencia y debo manifestar mi convicción de que la parte más sustancial de la investigación científica que se desarrolla en los departamentos y laboratorios universitarios la realizan los becarios y estudiantes de doctorado. El relevo generacional es imprescindible. Sin ellos, no habrá progreso en el conocimiento futuro.