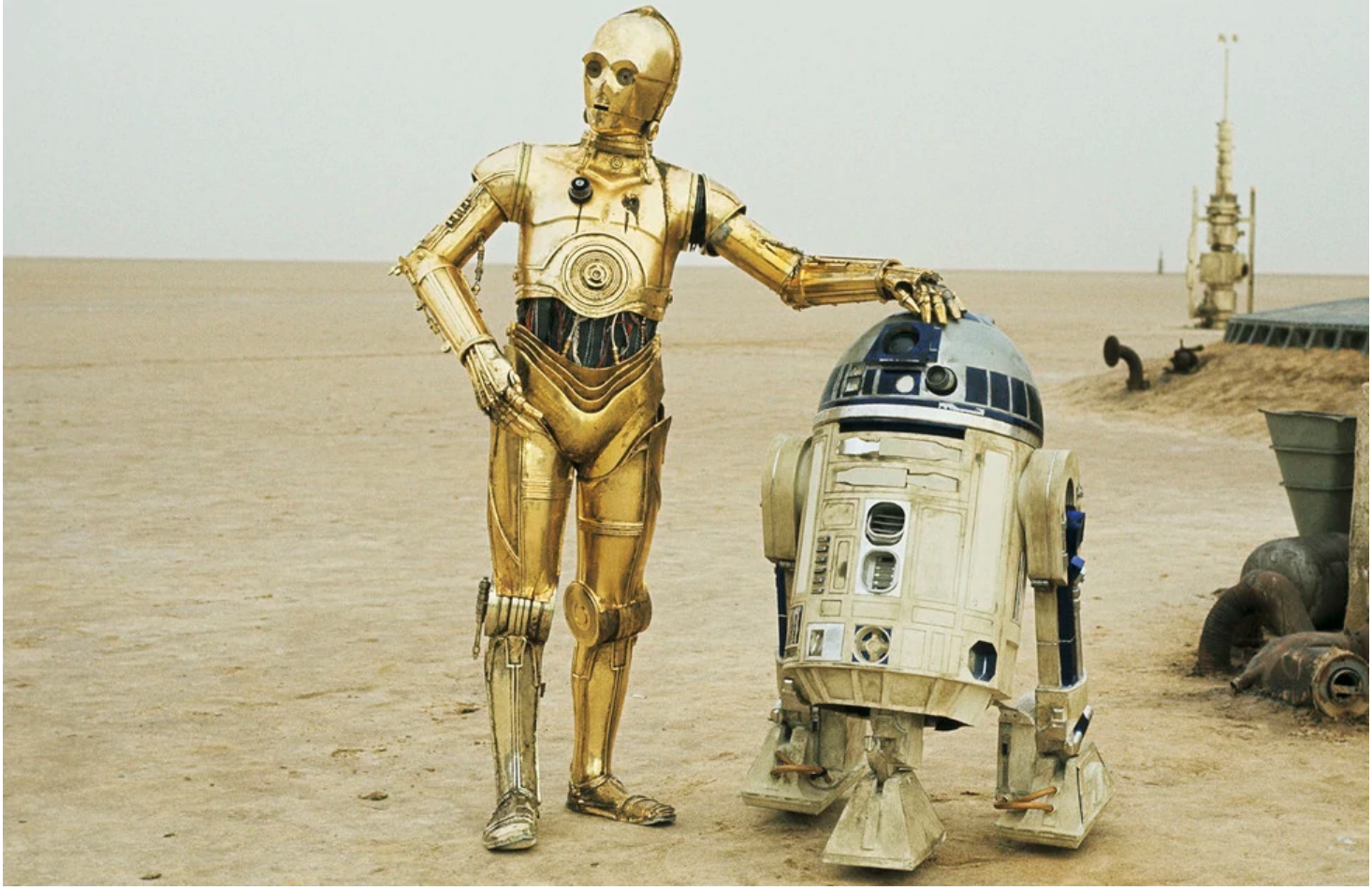

El procesamiento del lenguaje natural

FREDY NÚÑEZ
Y
CARLOS
GONZÁLEZ[1]

Este texto girará en torno a tres grandes preguntas. En primer lugar, ¿qué es el procesamiento del lenguaje natural y qué problemas aborda?; en segundo lugar, ¿cuáles han sido los aportes de los estudios del lenguaje a este campo y qué puede aportar en el futuro? Finalmente, ¿cómo puede contribuir el procesamiento del lenguaje natural a resolver problemas sociales y, en consecuencia, a propiciar un mejor futuro para la humanidad?

El procesamiento del lenguaje natural y sus desafíos

El procesamiento del lenguaje natural tiene como fin estudiar, diseñar y aplicar sistemas informáticos que faciliten la comunicación entre personas y entre personas y máquinas, logrando que esta comunicación sea flexible, eficiente y fluida. Su objetivo es imitar artificialmente algunos de los aspectos de la capacidad humana para el lenguaje, lo que se traduce en procesos de producción y comprensión. Algunos de sus productos podemos verlos actualmente en sistemas de traducción automática (Google Translate[2],

[1] Pontificia Universidad Católica de Chile.

[2] https://translate.google.com

DeepL[3]), o asistentes virtuales, como *Siri* de Apple, *Cortana* de Google o *Alexa* de Amazon. Un debate tradicional, pero que continúa hasta nuestros días, es a qué ámbito de estudios pertenece el procesamiento de lenguaje natural: a la informática o a la lingüística, lo que ha tenido consecuencias en la manera en que estas dos diferentes perspectivas han abordado sus problemas. En los años 50 del siglo XX surge el primer interés por desarrollar sistemas de traducción automática, los que fueron abordados inicialmente desde una perspectiva únicamente informática, sin tomar en cuenta aportes desde la lingüística, lo que no tuvo resultados eficientes. Esto llevó a la necesidad de integrar modelos lingüísticos que permitieran resolver estos problemas. Otros hitos que sucedieron a este fueron el desarrollo de sistemas de diálogo interactivos, métodos de aprendizaje automático y sistemas capaces de inferir conocimiento lingüístico a partir de algoritmos. Actualmente, los problemas que han sido abordados exitosamente son, entre otros, la detección de correo electrónico no deseado (*spam detection*), el reconocimiento automático de partes de la oración (*part-of-speech tagging*) y el reconocimiento de entidades nombradas (*named entity recognition*). Algunos ejemplos de las tareas que actualmente se encuentran en desarrollo son el análisis de sentimientos (*sentiment analysis*), la desambiguación léxica automática (*word sense disambiguation*) y la traducción automática (*machine translation*). Finalmente, los mayores desafíos que aún no han sido resueltos son el desarrollo de sistemas de respuesta automatizados (*automatic question answering*), mecanismos generadores de resúmenes (*automatic paraphrase-summarization*) y el desarrollo de agentes conversacionales (*conversational agents*).

El aporte de las ciencias del lenguaje al procesamiento del lenguaje natural

Desde los inicios de los estudios en inteligencia artificial ha existido una tensión en la integración de las disciplinas de la informática y la lingüística. La informática ha propiciado esencialmente el desarrollo de modelos estocásticos, que se caracterizan por la aplicación de técnicas matemáticas sobre un gran volumen de datos textuales con el objetivo de inferir conocimiento lingüístico. Un ejemplo de esta perspectiva son las aplicaciones de aprendizaje automático (*machine learning*). Las ciencias del lenguaje, por su parte, han favorecido el desarrollo de modelos simbólicos, que –a diferencia de los estocásticos– almacenan conocimiento lingüístico en forma de esquemas de representación del conocimiento. Un producto de esta última perspectiva ha sido el desarrollo de lexicones automatizados y ontologías. Si bien los métodos estocásticos han demostrado ser altamente eficientes, no son consecuentes con el mayor objetivo de la inteligencia artificial, que es imitar el lenguaje entendido como una capacidad cognitiva, lo que se hace evidente al tratar de mantener un diálogo fluido y espontáneo con un asistente virtual (como *Siri* o *Alexa*). Desde esta perspectiva, la mayor crítica que se puede hacer a los métodos estocásticos es que reducen las capacidades lingüísticas a decisiones probabilísticas basadas en

[3] https://www.deepl.com/es/translator

algoritmos, lo que no parece coincidir con la manera en que funciona el lenguaje humano. Las ciencias del lenguaje indican que existen diferentes habilidades lingüísticas que una máquina debería desplegar en la interacción entre humanos y sistemas artificiales: ser, por ejemplo, capaz de establecer relaciones entre el contenido conceptual que se encuentra en las diferentes palabras, de inferir información a partir de esas relaciones, y de actuar en diferentes entornos y con múltiples propósitos. Para abordar estos desafíos se hace necesario un nuevo especialista que pueda proveer tanto el conocimiento del funcionamiento del lenguaje natural como los modelos informáticos necesarios para su implementación en sistemas artificiales. Esta nueva especialidad ha recibido el nombre de ingeniería del conocimiento. Un producto de este nuevo abordaje puede verse en el desarrollo de proyectos como WordNet[4] y FrameNet[5] y, especialmente, en la creación de bases de conocimiento que integren diferentes aspectos del conocimiento del lenguaje (gramatical, enciclopédico, procedimental), como FunGramKB[6].

La contribución del PLN a la solución de problemas sociales

Los productos derivados de la aplicación e integración de estas perspectivas se pueden poner al servicio de la sociedad. Algunos ejemplos son los siguientes: (a) Detección de ciberacoso (*cyberbullying*) en redes sociales[7]. Este es un sistema que analiza tweets para encontrar patrones lingüísticos que pueden ser casos de ciberacoso, para lo cual se fundamenta en las características lingüísticas del acoso por redes sociales, que son típicamente diferentes de las que se dan en el acoso cara a cara. (b) Sensores sociales (*social sensors*) para el desarrollo de respuesta ante emergencias[8]. Este sistema es capaz de detectar automáticamente el nivel de riesgo de un desastre natural (terremotos, inundaciones, tornados, etc.) por medio del análisis de los comentarios que las personas hacen en redes sociales y las percepciones que pueden inferirse de esos comentarios.

Dos conclusiones relevantes pueden alcanzarse de estas reflexiones. La primera es que la integración de campos disciplinares como la informática y la lingüística hace posible un abordaje más profundo de problemas complejos, como es la naturaleza del conocimiento lingüístico, lo que muestra que nuestra sociedad actual requiere de generosidad intelectual para resolver los desafíos que propone la sociedad de la información. La segunda es que el desarrollo de aplicaciones como las anteriormente mencionadas pueden ponerse al servicio de la comunidad para resolver problemas sociales como el ciberacoso o la alerta temprana de desastres naturales.

***

[4] https://wordnet.princeton.edu
[5] https://framenet.icsi.berkeley.edu/fndrupal/
[6] http://www.fungramkb.com
[7] https://polipapers.upv.es/index.php/jclr/article/view/11013
[8] https://ebooks.lospress.nl/volumenarticle/47213