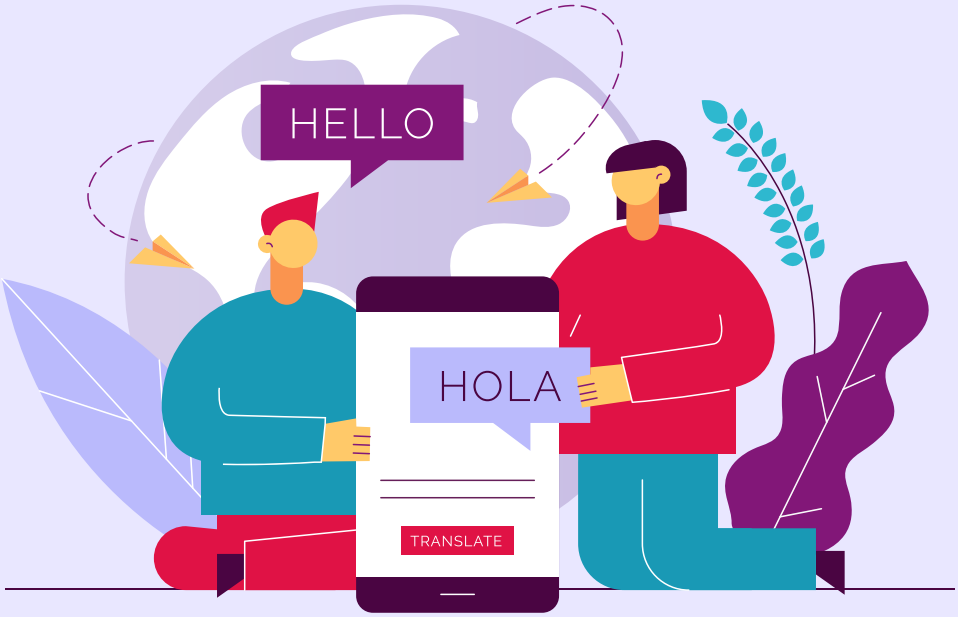
Procesamiento de Lenguaje Natural: dónde estamos y qué estamos haciendo



FELIPE BRAVO-MÁRQUEZ Profesor Asistente del Departamento de Ciencias de la Computación de la Universidad de Chile e Investigador Joven del Instituto Milenio Fundamentos de los Datos.

JOCELYN DUNSTAN Profesora Asistente de la Iniciativa de Datos e Inteligencia Artificial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile e Investigadora del Centro de Modelamiento Matemático.

El Procesamiento de Lenguaje Natural (PLN) es una rama de la Inteligencia Artificial (IA) centrada en el diseño de métodos y algoritmos que toman como entrada o producen como salida datos en la forma de lenguaje humano [1]. Esto puede venir en forma de texto o audio, y una vez que el audio es transcrito, ambos tipos de datos tienen un análisis común.

Tal como argumentan Julia Hirschberg y Chris Manning [2], tareas actuales donde el PLN entra en nuestras vidas son la traducción automática, los sistemas de pregunta-respuesta y la minería de texto en redes sociales. Ahondemos en la primera de ellas: la Web está en su mayoría en inglés, y el poder traducir páginas en forma casi instantánea es algo extraordinario. Traducir un texto no es fácil pues no hay una biyección entre palabras en ambos lenguajes, sino que una frase puede requerir menos palabras en un idioma que en otro (pensar por ejemplo traducir del español al inglés). Pero además, la traducción de una palabra requiere información del contexto en la que aparece para saber el sentido en la que se está usando. Asimismo, puede ocurrir que la palabra no tenga sentido en sí misma sino que en conjunto con la palabra que la acompaña (piense en las *phrasal verbs* del inglés). Actualmente los traductores automáticos usados por Google o DeepL están basados en sofisticadas redes neuronales.

PLN suele confundirse con otra disciplina hermana llamada Lingüística Computacional (LC). Si bien ambas están estrechamente relacionadas, tienen un foco distinto. La LC busca responder preguntas fundamentales sobre el lenguaje mediante el uso de la computación, es decir, cómo entendemos el lenguaje, cómo producimos lenguaje o cómo aprendemos lenguaje. Mientras que en PLN el foco está en resolver

problemas específicos, tales como las transcripción automática del habla, la traducción automática, la extracción de información de documentos y el análisis de opiniones en redes sociales. Es importante señalar que en PLN, el éxito de una solución se mide en base a métricas concretas (por ejemplo: qué tan similar es la traducción automática a una hecha por un humano) independientemente de si el modelo hace uso de alguna teoría lingüística.

Comprender y producir el lenguaje computacionalmente es extremadamente complejo. La tecnología más exitosa actualmente para abordar PLN es el aprendizaje automático supervisado que consiste en una familia de algoritmos que "aprenden" a construir la respuesta del problema en cuestión en base a encontrar patrones en datos de entrenamiento etiquetados.[1] Por ejemplo, si queremos tener un modelo que nos diga si un *tweet* tiene un sentimiento positivo o negativo respecto a un producto, primero necesitamos etiquetar manualmente un conjunto de *tweets* con su sentimiento asociado. Luego debemos entrenar un algoritmo de aprendizaje sobre estos datos para poder predecir de manera automática el sentimiento asociado a *tweets* desconocidos. Como se podrán imaginar, el etiquetado de datos es una parte fundamental de la solución y puede ser un proceso muy costoso, especialmente cuando se requiere conocimiento especializado para definir la etiqueta.

Los orígenes de PLN se remontan a los años cincuenta con el famoso test de Alan Turing: una máquina será considerada inteligente cuando sea capaz de conversar con una persona sin que ésta pueda determinar si está hablando con una máquina o un ser humano. A lo largo de su historia la disciplina ha tenido tres grandes periodos: 1) el racionalismo, 2) el empirismo, y 3) el aprendizaje profundo [3] que describimos a continuación.

El racionalismo abarca desde 1950 a 1990, donde las soluciones consistían en diseñar reglas manuales para incorporar mecanismos de conocimiento y razonamiento. Un ejemplo emblemático es el agente de conversación (o *chatbot*) ELIZA desarrollado por Joseph Weizenbaum que simulaba un psicoterapeuta rogeriano. Luego, a partir de la década de los noventa, el diseño de métodos estadísticos y de aprendizaje automático construidos sobre corpus llevan a PLN hacia un enfoque empirista. Las reglas ya no se construyen sino que se "aprenden" a partir de datos etiquetados. Algunos modelos representativos de esta época son los filtros de *spam* basados en modelos lineales, las cadenas de Markov ocultas para la extracción de categorías sintácticas y los modelos probabilísticos de IBM para la traducción automática. Estos modelos se caracterizaban por ser poco profundos en su estructura de parámetros y por depender de características manualmente diseñadas para representar la entrada.[2]

A partir del año 2010, las redes neuronales artificiales, que son una familia de modelos de aprendizaje automático, comienzan a mostrar resultados muy superiores en varias tareas emblemáticas de PLN [4]. La idea de estos modelos es representar la entrada (el texto) con una jerarquía de parámetros (o capas) que permiten encontrar representaciones idóneas para la tarea en cuestión, proceso al cual se refiere como "aprendizaje profundo". Estos modelos se caracterizan por tener muchos más parámetros que los modelos anteriores (superando la barrera del millón en algunos casos) y requerir grandes volúmenes de datos para su entrenamiento. Una gracia de estos modelos es que pueden ser preentrenados con texto no etiquetado como libros, Wikipedia, texto de redes sociales y de la Web para encontrar representaciones iniciales de palabras y oraciones (a lo que conocemos como *word embeddings*), las cuales pueden ser posteriormente adaptadas para la tarea objetivo donde sí se tienen datos etiquetados (proceso conocido como *transfer learning*). Aquí destacamos modelos como Word2Vec [5], BERT [6] y GPT-3 [7].

Este tipo de modelos ha ido perfeccionándose en los últimos años, llegando a obtener resultados cada vez mejores para casi todos los problemas del área [8]. Sin embargo, este progreso no ha sido libre de controversias. El aumento exponencial en la cantidad de parámetros[3] de cada nuevo modelo respecto a su predecesor, hace que los recursos computacionales y energéticos necesarios para construirlos sólo estén al alcance de unos pocos. Además, varios estudios han mostrado que estos modelos aprenden y reproducen los sesgos y prejuicios (por ejemplo: género, religión, racial) presentes en los textos a partir de los cuales se entrenan. Sin ir más lejos, la investigadora Timmnit Gebru fue despedida de Google cuando se le negó el permiso para publicar un artículo que ponía de manifiesto estos problemas [9].

1 | En PLN se le suele llamar a estos conjuntos de datos textuales (etiquetados o no etiquetados) como "corpus".

2 | La mayor parte de algoritmos de aprendizaje operan sobre vectores numéricos, donde cada columna es una característica del objeto a modelar. En PLN esas características pueden ser las palabras de una oración, las frases u otra propiedad (por ejemplo: el número de palabras con mayúsculas, la cantidad de emojis en un *tweet*, etc.).

3 | Word2Vec [5] tiene del orden de cientos de parámetros, BERT [6] tiene 335 millones de parámetros y GPT-3 [7] tiene 175 mil millones de parámetros.

Representations for Learning and Language (ReLeLa)[4] es un grupo de investigación del Departamento de Ciencias de la Computación (DCC) de la Universidad de Chile, donde también participan académicos y estudiantes de otros departamentos y centros. Sus miembros investigan varios temas en PLN: análisis de sentimiento y emociones en redes sociales, texto clínico, educación, textos legales, lenguas indígenas y el análisis de argumentos políticos.

Una línea de ReLeLa liderada por Jorge Pérez, ha sido el desarrollo de modelos preentrenados para el idioma español. Una contribución destacada ha sido BETO[5], la versión en español de BERT, que es ampliamente utilizado por investigadores y desarrolladores del mundo hispano.

En el ámbito del texto clínico, la creación de recursos para la extracción de información relevante requiere un trabajo fuertemente interdisciplinario. Recientemente fue presentado en el *workshop clínico de EMNLP*[6] el primer corpus clínico chileno etiquetado y resultados preliminares para el reconocimiento automático de entidades nombradas.

Finalmente, *The Word Embeddings Fairness Evaluation Framework* (WEFE)[7], es una herramienta de código abierto que permite medir y mitigar el sesgo de los modelos preentrenados señalados anteriormente. La principal característica de WEFE es estandarizar los esfuerzos existentes en un marco común para ser libremente utilizado.

A pesar de los grandes avances en los últimos años, aún estamos lejos de responder todas las interrogantes de PLN. En problemas como el diseño de *chatbots* las soluciones del estado del arte aún distan mucho de lo esperado y ni siquiera es claro cómo evaluarlas correctamente, luego para muchos otros problemas del mundo real simplemente no es posible obtener los recursos necesarios (datos etiquetados, hardware) para construir una solución adecuada. En RELELA confluyen visiones provenientes de la computación, las matemáticas, la lingüística y la salud para discutir esas interrogantes y sobre todo para mantenernos al día con los constantes avances del área. Todo esto ocurre en nuestros seminarios semanales donde escuchamos exposiciones de miembros del grupo o de algún charlista invitado. ■

REFERENCIAS

[1] Eisenstein, J. (2018). Natural language processing.

[2] Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261–266.

[3] Deng, L., & Liu, Y. (Eds.). (2018). *Deep learning in natural language processing*. Springer.

[4] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of machine learning research, 12(Aug):2493–2537.

[5] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13).

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.

[7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 202.

[8] NLP-progress: Repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state-of-the-art for the most common NLP tasks: http://nlpprogress.com/.

[9] Bender, Emily M., et al. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

4 | https://relela.com/.

5 | https://github.com/dccuchile/beto.

6 | https://www.aclweb.org/anthology/2020.clinicalnlp-1.32/.

7 | https://wefe.readthedocs.io/en/latest/.