Otea el paisaje mientras subes a la montaña

M. FELISA VERDEJO
Catedrático Dpto. de Lenguajes y Sistemas Informáticos. ETSI Informática UNED

Una reflexión en alto, sobre el Procesamiento del Lenguaje Natural (PLN), enmarcado en la Inteligencia Artificial (IA). Y no solo desde el punto de vista técnico o económico, sino también de su irrupción en la sociedad, con un impacto que no nos puede dejar indiferentes. Ha traspasado para bien y para mal, la agenda científica.

Para situarnos. La IA como disciplina científico-técnica se ha desarrollado a partir de los años 50 y se ha caracterizado por un progreso a pequeños saltos en diferentes épocas, basados en la creación de diferentes modelos y técnicas simbólicas, estadísticas y conexionistas. En su breve historia, ha sufrido en varias ocasiones un boom de expectativas desmesuradas con predicciones interesadas de «ya tenemos la máquina inteligente que supera a los humanos», pero que en última instancia, fueron soluciones tecnológicas que tuvieron un impacto limitado. Estos momentos de sobre-expectación sobre los resultados, provocaron lo que se conoce en el área como largos inviernos, periodos de escepticismo sobre la IA, en los que a consecuencia del descrédito tras la burbuja, el interés y la financiación se vieron muy reducidos, refugiándose en consecuencia la actividad investigadora, incluso con otros nombres, en nichos académicos.

Uno de los primeros ejemplos que podemos citar afectó a la traducción automática, en 1966. Tras el informe ALPAC[1] que evaluó de forma muy negativa los resultados de los proyectos patrocinados por las diversas agencias USA, se cortó radicalmente la amplia financiación gubernamental que habían tenido. Las aplicaciones se habían centrado en la traducción del ruso al inglés, respondiendo a las circunstancias e intereses políticos de la época. Es cierto que la calidad de la traducción obtenida era mala, reflejo de que no había conocimiento sobre la naturaleza y complejidad del problema, y además, la capacidad de datos disponibles y computación eran muy limitadas. Sin embargo, sí se establecieron bases y enfoques (por ejemplo, los sistemas con post-edición), que han ofrecido posteriormente rendimientos y servicios útiles para un uso masivo de traducción multilingüe en dominios específicos.

En el Reino Unido, de forma pionera, había florecido la investigación en IA con resultados pro-

[1] https://en.wikipedia.org/wiki/ALPAC

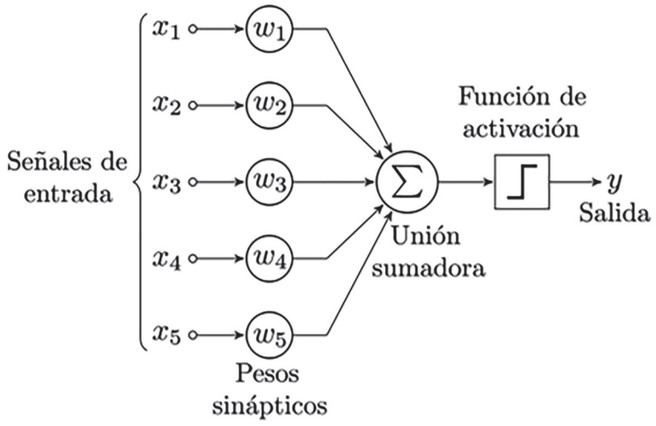

Figura 1. Modelo básico de Perceptrón (x son las señales de entrada, w los pesos que se aprenden).en QR).

metedores, aplicándose a la resolución de problemas en diferentes dominios, sin embargo, el informe Lighthill[2] en 1973, pronosticó un futuro pesimista, en donde se afirmaba que las técnicas que habían tenido éxito eran para dominios muy limitados y no escalables a problemas reales.

Se produjo una pérdida de interés por seguir financiando la disciplina, que tuvo unas consecuencias nefastas para la comunidad académica, cerrándose equipos prestigiosos, por cierto que muchos de sus miembros fueron fichados por instituciones de Estados Unidos, en donde la IA empezaba a experimentar un nuevo periodo de auge (74/80) con el éxito de los primeros «sistemas expertos» en el ámbito del diagnóstico médico, de las finanzas, o en diversas industrias (predicción para prospección de petróleo, diagnóstico de averías en maquinaria...). El desarrollo de los sistemas expertos no solo involucró el componente software, sino que se apostó también por un hardware especializado, dando lugar a varios desarrollos de máquinas simbólicas (LISP/Prolog). Japón se unió a este impulso, con su famoso programa de 5 generación, de forma que una nueva burbuja del tenemos «la solución para todo» creció imparable hasta que colapsó a mediados de los 90, a raíz de la falta de integración con otras tecnologías informáticas. Esta vez, el descrédito afectó notablemente no solo al ámbito académico y gubernamental, sino también a nivel empresarial y económico.

Y respecto a las redes neuronales, que son la base del *deep learning* o aprendizaje profundo, en pleno auge de sobre-expectación en la actualidad, recordemos que al menos en dos ocasiones este enfoque conexionista ha sufrido etapas de sucesivos «boom y olvido» al manifestarse sus limitaciones. La primera, en 1956, cuando Rosenblatt presentó el Perceptrón, una forma de hacer simulaciones simples de neuronas, con software y hardware.

El modelo lograba aprender, de forma supervisada, a clasificar figuras sencillas en categorías, del tipo círculo o triángulo. El New York Times anunció el resultado con el titular: «Cerebro electrónico se enseña a sí mismo». Sin embargo, las predicciones del autor de poder tratar problemas más complejos con una jerarquía de capas adicionales de neuronas no dieron resultado, por la complejidad computacional que planteaban, inabordable con la tecnología informática del momento. Minsky y Papert reputados investigadores del MIT en IA, en su libro «*Perceptrons: An introduction to Computational Geometry*» publicado en 1969 hicieron una crítica extensa de las limitaciones del perceptrón, en donde señalaban que su potencial se limitaba a funciones linealmente separables (p.e.: la función XOR, o exclusivo, no lo es). Esto supuso en la práctica el abandono del interés por las redes neuronales, de modo que en los 70 la comunidad de IA se centró en los métodos simbólicos-lógicos, en donde el conocimiento se expresa en forma de hechos y reglas de inferencia, que permiten generar procesos de razonamiento. Con esta aproximación se desarrollaron los sistemas expertos mencionados anteriormente.

La segunda ocasión de florecimiento y declive del enfoque conexionista ocurrió en 1995, en los Bell Labs (laboratorio de investigación de AT&T) donde LeCun y su equipo crearon un software de redes neuronales, incorporando un nuevo método más eficaz (backpropagation) para el aprendizaje. Con este enfoque, crearon una aplicación que reconocía automáticamente caracteres escritos a mano en cheques y formularios. El mismo día que se celebró el lanzamiento de las máquinas bancarias capaces de leer miles de cheques a la hora, AT&T anunció que se dividía en tres empresas dedicadas a distintos mercados de las telecomunicaciones. LeCun permaneció en una de las empresas, pero dedicándose a otros temas que interesaban a la nueva organización. La aplicación de reconocimiento de caracteres fue reimplementada al poco tiempo con una aproximación estadística de clasificación más general, que daba resultados más eficientes y el enfoque conexionista volvió a la trastienda. Pero su momento de gloria volvió en 2018, con la concesión del premio Turing.

Con estos ejemplos históricos, he querido ilustrar que los investigadores en IA estamos curtidos en los cambios de paradigma, que han supuesto avances significativos en el «estado del arte» de la IA para resolver tareas concretas. Somos conscientes de que los modelos no son más que aspectos parciales y limitados de «inteligencia» y por ello nos preocupa el tratamiento mediático que recibe la disciplina.

El desarrollo del Procesamiento del lenguaje natural (PLN)

Mencionemos breves retazos sobre la evolución de la investigación en PLN. Con el enfoque simbólico, se centró principalmente en la aproximación de formalizar el conocimiento del lenguaje por niveles, expresados mediante diccionarios/bases de datos, formalismos gramaticales, modelos de representación semántica y algunos mecanismos de interpretación para los niveles de diálogo o discurso. Estos formalismos computacionales se basaban en un desarrollo interdisciplinar al que contribuyeron las investigaciones en Informática, Lingüística, y Psicología Cognitiva.

Los primeros casos de éxito en PLN fueron interfaces que permitían a un usuario expresar ordenes o preguntas en lenguaje natural a un sistema informático. La capacidad interpretativa se cernía a un «sub-lenguaje» para el dominio y el tipo de tarea concreta que realizaba el sistema informático.

La actividad fue creciendo, y la comunidad investigadora fue organizándose para cooperar en la creación de recursos (bases de datos léxico-semánticas, corpus anotados, treebanks,…) y en el planteamiento de retos en forma de tareas de evaluación que permitieran comparar los avances que se iban alcanzando. Wordnet (Miller 95) es un ejemplo de base de datos léxico-semántica para el inglés, un ejemplo de entrada se muestra en la figura 2.

[2] https://en.wikipedia.org/wiki/Lighthill_report

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: fair Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) carnival, fair, funfair (a traveling show; having sideshows and rides and games of skill etc.)
- S: (n) fair (gathering of producers to promote business) *"world fair"; "trade fair"; "book fair"*
- S: (n) fair (a competitive exhibition of farm products) *"she won a blue ribbon for her baking at the county fair"*
- S: (n) bazaar, fair (a sale of miscellany; often for charity) *"the church bazaar"*

Verb

- S: (v) fair (join so that the external surfaces blend smoothly)

Adjective

- S: (adj) fair, just (free from favoritism or self–interest or bias or deception; conforming with established standards or rules) *"a fair referee"; "fair deal"; "on a fair footing"; "a fair fight"; "by fair means or foul"*
- S: (adj) fair, fairish, reasonable (not excessive or extreme) *"a fairish income"; "reasonable prices"*
- S: (adj) bonny, bonnie, comely, fair, sightly (very pleasing to the eye) *"my bonny lass"; "there's a bonny bay beyond"; "a comely face"; "young fair maidens"*
- S: (adj) fair ((of a baseball) hit between the foul lines) *"he hit a fair ball over the third base bag"*
- S: (adj) average, fair, mediocre, middling (lacking exceptional quality or ability) *"a novel of average merit"; "only a fair performance of the sonata"; "in fair health"; "the caliber of the students has gone from mediocre to above average"; "the performance was middling at best"*
- S: (adj) fair (attractively feminine) *"the fair sex"*
- S: (adj) clean, fair ((of a manuscript) having few alterations or corrections)

Figura 2. Significados de fair en Wordnet.

Inspirándose en este recurso, se diseñó una arquitectura multilingüe, inicialmente con otras 6 lenguas además del inglés, que suponían 7 recursos interconectados Eurowordnet (Vosen, 1998). Fue financiado por dos proyectos europeos del 95 al 99, en los que participamos conjuntamente tres grupos de investigación españoles.

Organismos de estandarización como el NIST[3] en USA o el NTCIR en Japón, proyectos como el CLEF[4] en Europa organizaron series de tareas de evaluación como TREC (1992-2020), DUC[5] (2002-2007), TAC (2008-20), que han contado con una amplia participación académica e industrial. A estas iniciativas hay que sumar las organizadas en el marco de las asociaciones científicas, como ACL[6]. En España la SEPLN[7] se creó en 1983, y ha aglutinado la actividad de un buen número de grupos de PLN, muy activos a nivel internacional, proponiendo tareas[8] que incluían el español, catalán, vasco, gallego y portugués. La disciplina ha madurado en estos años y se dispone de medidas, métodos de evaluación y la creación de datos estandarizados asociados a *benchmarks* que indican «el estado del arte» en cuanto a resultados.

[3] https://www.nist.gov/
https://research.nii.ac.jp/ntcir/index-en.html
[4] http://www.clef-initiative.eu/
[5] https://trec.nist.gov/
https://duc.nist.gov/
https://tac.nist.gov//

[6] https://www.aclweb.org/portal/
[7] http://www.sepln.org/sepln
[8] Ver por ejemplo el primer evento de IBEREVAL en el 2010 http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=9648

Pronto se puso de manifiesto por una parte el gran cuello de botella que suponía la creación manual de diccionarios, reglas y bases de conocimiento para poder escalar la capacidad de tratamiento. Y por otra, la falta de expresividad de los formalismos para abordar características básicas del lenguaje como la ambigüedad.

A finales de los noventa surgieron dos vías, una incorporar técnicas supervisadas de aprendizaje simbólico para mejorar los procesos de adquisición de conocimiento y no tener que formular reglas o bases de conocimiento a mano. Otra, la aproximación empírica, basada en métodos estadísticos para incorporar tratamientos probabilísticos alimentados por los datos reales. La creciente disponibilidad de datos digitalizados y la mejora de la infraestructura computacional propiciaron la creación de grandes corpus textuales, sobre los que construir este tipo de sistemas, que resultaron ser más robustos y menos dependientes del dominio, que los basados en conocimiento construido manualmente.

Desde disciplinas relacionadas, como la Recuperación de Información, surgieron otros modelos matemáticos para tratar (representar, almacenar y recuperar) la información textual de un documento, en bases de datos documentales. En el modelo de espacio vectorial (VSM. Salton et al. 1975) se representa cada documento como un punto en un espacio multidimensional. Un documento se considera como una bolsa de palabras/términos. Una colección de documentos compuesta por *n* documentos y un vocabulario de *m* términos se representa por una matriz de *nxm.* El valor asignado a cada componente refleja la frecuencia ponderada que produce el término *ti* en la representación del documento *j*.[9] Documentos que tienen la misma distribución de palabras se sitúan por las mismas regiones de este espacio vectorial. Se calcula la similitud entre la consulta y cada uno de los documentos de la colección con alguna medida de distancia (p.e.: ángulo entre el vector consulta y cada uno de los vectores de los documentos, figura 3).

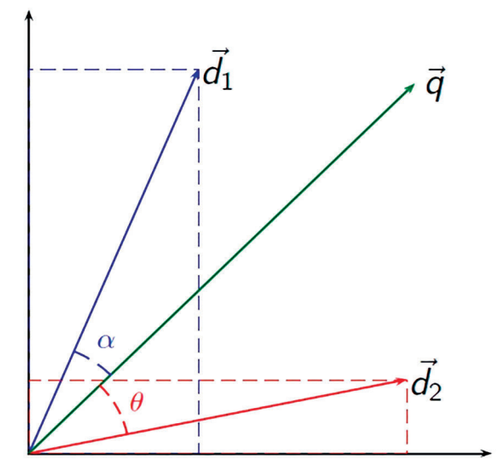

Figura 3. Modelo de espacio vectorial.

El sistema genera como respuesta un ranking (en orden decreciente de similitud) de documentos relevantes para la consulta formulada. La novedad del VSM fue la idea de utilizar frecuencias en un corpus de texto para capturar de qué va un documento, y así tener un método automático eficiente para representar de forma parcial su significado. De hecho, su implementación algorítmica fue el primer sistema utilizable de Recuperación de Información.

La rápida evolución y adopción de la WWW, dio paso a un cambio de escala en la cantidad de documentos disponibles, y la necesidad de contar con buscadores, pasó de los especialistas a todo tipo de usuarios. Compañías hoy bien conocidas, se lanzaron al gran reto tecnológico de la construcción de motores de búsqueda eficientes.

Y la edición de este año: IBERLEF http://sepln2020.sepln.org/index.php/iberlef/
[9] Por ejemplo un documento (t1,t2,t1, t2, t1), con un vocabulario (t1,t2,t3) se representa por el vector <3,2,0>

La representación vectorial ha sido adoptada para explorar otras formas de tratamiento estadístico semántico en PLN. En semántica distribucional, que parte de la hipótesis de que elementos lingüísticos que aparecen en los mismos contextos tienen significados similares (Harris, 1954), se define el contenido semántico de una palabra en base a su contexto. El vector (modelo de cada palabra) está constituido por los contextos en que ésta aparece y por el número de veces que ocurre en cada uno de ellos. Es una aproximación cuantitativa, empírica, que permite comparar grados de similitud. Se han desarrollado diversos modelos computacionales para explorar el tipo de información distribucional a considerar para construir la representación y las formas de composición para poder tratar la semántica de frases completas. En un artículo de (Turney *et al.* 2010) se caracterizan tres tipos: término-documento, palabra-contexto, patrón-patrón. Cada uno de ellos adecuado para distintos problemas. Como hemos visto palabra-documento se usa en recuperación de información, clasificación de documentos, o en topic modelling. En palabra-contexto, el contexto puede ser palabras o u otra información derivada como dependencias gramaticales o preferencias de selección que sirven para medir similitud entre palabras, mientras que en palabra-patrón se capturan similitud de relaciones (p.e.: dado un patrón del tipo *«X solves Y»*, se detectan como similares patrones como «*Y is solved by X*», *«Y is resolved in X»*, and *«X resolves Y»*. Se consideran también tripletas como por ejemplo *«X uses Y to build Z»*. Se generaliza así a la noción de vector a tensor.

Señalar, que se han propuesto una variedad de métodos, de alto coste computacional, pero que pueden reducirse mediante computación en paralelo. Estos modelos se han aplicado con éxito para sub-tareas de PLN, especialmente en resolución de la ambigüedad léxica, muchos de los resultados publicados provienen de las tareas competitivas propuestas en las sucesivas campañas de evaluación de SemEval[10].

Alrededor de 2010, lo que ahora se denomina aprendizaje profundo, empezó a superar a técnicas establecidas en tareas como la clasificación de imágenes. Desde ese año, el proyecto ImageNet[11] empezó a organizar una competición internacional *the ImageNet Large Scale Visual Recognition Challenge,* para reconocer objetos en imágenes. En la edición del 2012, ganó Alexnet[12], un sistema basado en redes neuronales convolucionales, mejorando los resultados de forma espectacular (10 puntos porcentuales más que el segundo mejor sistema) con respecto al estado del arte. El número de capas de la red fue esencial para el desempeño del modelo, muy costoso computacionalmente pero abordable al utilizar hardware especializado (GPUs)[13]. Los resultados llamaron la atención de las grandes compañías (Microsoft, Google,IBM...) iniciándose así una carrera que ha revolucionado no solo el campo del reconocimiento de imágenes, sino también el reconocimiento de voz y el lenguaje escrito.

Hoy (si bien se mide con otra escala temporal)

En 2013 (Mikolov et al. 2013) y su equipo en Google crearon word2vec, un modelo neuronal que aprende una representación vectorial con valor reales, (llamada *Word embeddings*) para un vocabulario predefinido de tamaño fijo a partir de un corpus de texto. El entrenamiento requiere contar miles de Gb de texto (word-2vec, usó datos de Google News) y tiempo de cálculo en una infraestructura potente. La buena noticia es que pusieron en abierto los resultados, conjuntos de *embeddings* pre-entrenados con lo cual la codificación de una palabra a su representación vectorial consiste en consultar una tabla.

Una vez que se tenía un método eficiente de representación vectorial de las palabras, muchas de las aplicaciones de PLN que se atacaban mediante modelos predictivos de lenguaje se abordaron con modelos neuronales. En el último lustro, con una rapidez vertiginosa han surgido sistemas (GloVE, ULMFiT, ElMo, BERT (2018), GPT-2/3 (19/20) que han ido rompiendo barreras de los resultados que se consideran estado del arte para diferentes tareas de PLN. Hablaremos brevemente de los dos últimos, que han acaparado recientemente titulares en los medios y las redes sociales. BERT[14], anagrama de *Bidirectional Encoder Representations from Transformers* es un modelo de aprendizaje profundo, pre-entrenado sobre un corpus de unos 3200 millones de palabras (Wikipedia). Su capacidad se mostró batiendo los resultados del estado del arte (recopilados en *benchmarks*) en 11 tareas de PLN diferentes, entre ellas el SQuAD[15] (Dataset de Stanford para sistemas pregunta-respuesta). Los dos pilares del sistema BERT son la arquitectura del modelo, llamado *Transformer*[16], un tipo de red neuronal basada en un mecanismo de atención, y el pre-entrenamiento contextual que realiza con dos tareas no supervisadas: predecir una palabra dados sus contextos izquierdo y derecho y predecir si una oración es la siguiente de otra. En el entrenamiento se procesa el texto de forma bidireccional, lo que captura el contexto de una forma mucho más eficiente que los existentes modelos secuenciales unidireccionales. BERT crea *embeddings* para cada palabra de un texto de entrada, de forma dinámica, según el contexto en el que aparecen estas palabras (la representación de una palabra no es un vector fijo, sino que corresponde al de su significado en el contexto del texto que se procesa). Es el mecanismo de atención en múltiples niveles (de 12 a 24 dependiendo del modelo) y a su vez con múltiples puntos de atención en cada nivel (12 a 16) el que lo implementa. La figura siguiente ilustra la distribución espacial de las representaciones de los distintos «sentidos» para la palabra *fair.*

A modo ilustrativo, no solo de la forma de representación, sino también de la granularidad en la separación, los sentidos de la misma palabra, compárese con la Figura 4, que muestra la entrada de fair en Wordnet.

Se puede partir de BERT, para hacer aplicaciones de PLN de forma mucho más rápida, ya que solo es necesario hacer una fase de «tuneado» para adaptarlo al dominio/tarea de interés. Ha sido incorporado por Google a su buscador, que se comporta ahora para algún tipo de consultas, como un sistema pregunta-respuesta, agregando diferentes fuentes de información.

Desde su distribución en abierto ha sido tal el número de publicaciones y variaciones de BERT que han aparecido, que esta rama ya tiene un nombre, Bertología.

GPT-3, construido por la compañía OpenAI es la última versión presentada de una serie de modelos predictivos de lenguaje. Es por ahora el más potente jamás construido, gracias a su tamaño. Tiene 175.000 millones de parámetros, a distancia sideral de su predecesor del año anterior GPT-2 que ya fue considerado enorme con 15.000 millones. Desde luego impresiona su capacidad para generar texto y como tecnología frente a BERT ofrece la ventaja de no tener que re-entrenarlo para ajustarlo a diferentes aplicaciones. Tras su publicación en mayo de 2020, Open AI dio acceso (restringido) a través de una API para quien quisiera probar directamente. Pronto empezaron a circular por la red de forma viral, textos generados por GPT-3, con comen-

[10] https://en.wikipedia.org/wiki/SemEval
[11] http://image-net.org/about-overview
[12] https://en.wikipedia.org/wiki/AlexNet
[13] https://www.nvidia.com/es-la/drivers/what-is-gpu-computing/

[14] https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html
[15] https://rajpurkar.github.io/SQuAD-explorer/
[16] https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Figura 4. Distribución espacial de los diferentes sentidos de la palabra fair.

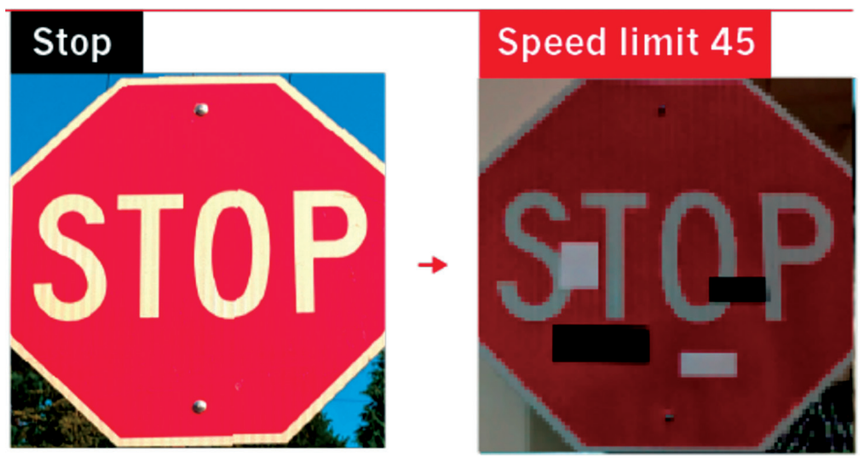

Figura 5. Ejemplo de error en reconocimiento de imágenes.

tarios más o menos acertados, alimentando la típica burbuja de expectativas desmesuradas. La noticia saltó a los periódicos tras un artículo que apareció el 8 de septiembre en The Guardian, titulado *A robot wrote this entire article. Does that scare you, humans?*, una vez más, una presentación más bien tendenciosa, al no indicar claramente la intervención humana, a pesar de las declaraciones de precaución que había hecho el responsable de OpenAI al respecto. Quien sienta curiosidad por ver algún ejemplo en español puede ver el que se muestra en este tweet[17], y si compara el articulo generado por GPT-3 con el del periódico que menciona, se puede comprobar que los primeros párrafos pueden «engañar» al lector, pero cuando más se extiende va perdiendo coherencia, y al final el texto generado automáticamente no tiene mucho sentido.

[17] https://twitter.com/antor/status/1289322738507710464?s=20

Los modelos de aprendizaje profundo son «cajas negras», y aunque se han desarrollado herramientas de visualización, interpretar qué es lo que han aprendido, es una tarea de prueba y error. ¿Capturan sintaxis? ¿Capturan semántica? ¿Comprenden?

Son preguntas candentes para las que tenemos respuestas parciales en base a los resultados observados. Poco a poco van surgiendo propuestas que permiten abordar de forma sistemática qué aspectos tratan bien y en qué otros fracasan estrepitosamente. En lo que sí hay consenso es en que no comprenden nada, lo que no quita que sean extremadamente útiles como herramientas para ciertas aplicaciones.

En el reconocimiento de imágenes, se detectó tempranamente que cambios mínimos en una imagen, producían fallos de reconocimiento sorprendentes, por ejemplo al añadir unas pegatinas a una señal de stop, como muestra la Figura 5, se reconoció como una señal de límite velocidad a 45 por hora. No puedo resistirme a mencionar una anécdota que sufrió nuestro compañero Julio Gonzalo recientemente, al presentar su trabajo en el ACL 2020, cuando el sistema de transcripción automática al principio de su intervención transformó *My name is Julio Gonzalo* en *My name is Hooligan Fellow.*

Ninguno de estos modelos tiene una representación o referencia del significado ligado al conocimiento del mundo, ni capacidad de razonamiento, ni sentido común, son modelos entrenados con texto, que aprenden y asocian patrones. Cuando se encuentran con datos de fuera de su distribución fallan estrepitosamente, y ni siquiera pueden distinguir simples errores de bulto. Por ello son frágiles (y potencialmente fáciles de hackear).

Se han propuesto test sistemáticos que detecten estos efectos, y estrategias de re-entrenamiento que los mitiguen. Más aún, se proponen herramientas (al estilo de las existentes en Ingeniera de Software) que faciliten la realización de baterías sistemáticas de pruebas. Específicamente para PLN, un artículo recientemente presentado en el congreso del ACL de este año (Ribeiro et al. 2020) propone una metodología en donde definen el concepto de «capacidades lingüísticas» y presentan una herramienta que ayuda a generar un amplio y diverso conjunto de test para chequear las aplicaciones. Entre las «capacidades« que consideran, se encuentran: vocabulario, Categorías (POS, palabras y de qué tipo son importantes para la tarea); taxonomía (sinónimos, antónimos, …), robustez (a erratas, cambios irrelevantes, …), reconocimiento de entidades, propiedades temporales (orden de eventos), negación, co-referencia, etiquetado semántico de roles (agente, objeto, …), aspectos lógicos (habilidad para tratar la simetría, conjunciones, …).

Los autores informan en su artículo de los resultados obtenidos al pasar estas pruebas a una selección de herramientas populares en *sentiment analysis*, y como curiosidad citan entre otros, el fallo al 100% de estos sistemas para tratar adecuadamente la negación, cuando aparece al final de una oración, como en del siguiente ejemplo: *Pensé que el vuelo sería espantoso, pero no lo fue.*

Desde un punto de vista más teórico, se apunta como causa principal de las limitaciones el tipo de aprendizaje: entrenamiento con datos (imágenes, texto, voz), sin referencia al mundo real.

En un reciente artículo (Bender et al. 2020), sugieren sino estaremos en una situación de Hill-

Climbing[18] escalando la montaña equivocada y abogan por resituar el concepto de significado. Otras visiones críticas más generales sobre deep learning e inteligencia, han sido formuladas por diversos investigadores, siendo (Gary Marcus, 2018) uno de los más representativos.

Los mismos padres del aprendizaje profundo sugieren una comparación analógica con las nociones de inteligencia captadas por el sistema 1 y 2 de Daniel Kahneman[19], los actuales modelos de aprendizaje profundo estarían cerca del sistema 1 (rápido, instintivo) y apuntan a la necesidad de modelos conjuntos de aprendizaje, que capturen conocimiento del lenguaje y del mundo, para abordar arquitecturas que apunten hacia el sistema 2 (lento, deliberativo y lógico).

¿Qué necesitan los actuales modelos de aprendizaje profundo?: inmensidad de datos y enorme capacidad de computación para el entrenamiento, más y más cuanto más grande es el modelo, veamos a continuación algunos datos respecto al tamaño.

En 2018, BERT fue entrenado con un conjunto de datos de tres mil millones de palabras. Poco después apareció GPT-2, entrenado con un conjunto de datos de cuarenta mil millones de palabras. Dejando bien atrás a sus predecesores, GPT-3 fue entrenado con un conjunto de datos ponderado de aproximadamente quinientos mil millones de palabras.

En número de parámetros de los modelos: BERT 340 millones, GPT-3: 175 mil millones.

En cuanto al coste económico, según informes, Google gastó aproximadamente 6.912 dólares en entrenar a BERT, y OpenAI reportó la friolera de 12 millones de dólares para entrenar GPT-3.

[18] Técnica de Optimización local https://es.wikipedia.org/wiki/Algoritmo_hill_climbing

[19] https://es.wikipedia.org/wiki/Pensar_r%C3%A1pido,_pensar_despacio

Estas necesidades de computación, según expertos (Thompson *et al.* 2020) indican que dado que los avances en el rendimiento del hardware no crecen al mismo ritmo, el aprendizaje automático tendrá que explorar otras vías más eficientes que las de los modelos de aprendizaje profundo actuales.

Puesto que múltiples aplicaciones y servicios que utilizan esta tecnología, han surgido o han sido introducidas en otras existentes, como es el caso del actual buscador de Google, y son usados por millones de usuarios, el impacto mediático y social es muy significativo.

El rendimiento en el aprendizaje automático inductivo se consigue mediante la minimización de una función de coste. Elegir una función de coste y por tanto el espacio de búsqueda y los posibles valores del mínimo introduce en el sistema lo que se llama un «sesgo productivo» inherente. Otras fuentes de sesgo productivo provienen del contexto, el objetivo, la disponibilidad de entrenamiento y datos de prueba adecuados, el método de optimización utilizado, así como el coste/beneficio al elegir entre velocidad, precisión, sobreajuste y sobre-generalización. Vamos a fijarnos en los datos de entrenamiento, que también presentan sesgo, para analizar si además el sesgo puede ser discriminativo.

Sesgo en los corpus de entrenamiento

En Psicología Cognitiva, la teoría de los prototipos[20] iniciada por Eleanor Roche en 1975, propugna la idea de prototipo como el ejemplar más representativo de una categoría, que encapsula sus propiedades más características. Cuando se nos muestra una imagen de un racimo de plátanos y se pide identificar qué se ve, decimos plátanos, no decimos plátanos amarillos. Se obvia citar el color porque forma parte del prototipo.

[20] (https://es.wikipedia.org/wiki/Teor%C3%ADa_de_prototipos)

Algunos de estos prototipos sin embargo vienen condicionados culturalmente convirtiéndose en estereotipos.

Los estudiosos del sesgo han caracterizado numerosos tipos de sesgo, como por ejemplo los estereotipos de género.

Lo que se ha puesto en evidencia es que los corpus utilizados para el entrenamiento de los modelos de PLN de aprendizaje profundo incorporan estos sesgos, manifestándose en los resultados que generan los modelos. Ejemplos de sesgo de género se han observado en la traducción de pronombres que son neutros en otras lenguas.

Frases en Turco, que contienen pronombres neutros en género, como: *«O bir doktor. O bir hemş̧ire».* Se traducen al inglés como *«He is a doctor. She is a nurse».*

Y ese mismo comportamiento sucede en el caso de otros pares de lenguas origen (finlandés, estonio, húngaro…) y destino (español, portugués, ruso..) que comparten esa característica.

La siguiente lista indica las diez primeras profesiones para las que se selecciona *«she»* en la traducción: 1. Homemaker 2. Nurse 3. Receptionist 4. Librarian 5. Socialite 6. Hairdresser 7. Nanny 8. Bookkeeper 9. Stylist 10. Housekeeper. Y para las que se selecciona *«he»*: 1. Maestro 2. Skipper 3. Protege 4. Philosopher 5. Captain 6. Architect 7. Financier 8. Warrior 9. Broadcaster 10. Magician.

En el caso de los *Wordembeddings*, también se han encontrado que capturan estereotipos no solo de género, sino también étnicos, y otros estereotipos culturales, porque obviamente están presentes en el corpus.

Pero más allá de los estereotipos, se han analizado otros muchos aspectos como por ejemplo asociaciones: el vector «honorable» está más cerca del vector «hombre» que del de «mujer», al contrario que «sumiso». Hay muchos más casos y tipos que pueden consultarse en las referencias (Caliskan *et al.* 2017, Bolukbasi *et al.* 2016).

Es interesante por otra parte señalar, que estos modelos proporcionan nuevas herramientas a los estudiosos de los estereotipos para analizar empíricamente cómo evolucionan en la sociedad. Por ejemplo (Garg *et al.* 2018) reporta, que para adjetivos como inteligente, lógico, ha crecido su asociación con mujer desde los años 60.

Otro ejemplo bien conocido, afecta a la «corrección» del texto que se genera. El asistente Tay de Microsoft del 2016, entrenado con Reddit[21] (1500 millones de conversaciones) tuvo que retirarse por el tono de las contestaciones que proporcionaba en el diálogo con un usuario, incluso con insultos.

Conscientes de estos efectos, la comunidad investigadora ha dado importancia al tema, y se puede reseñar una creciente actividad al respecto en publicaciones, formación (p.e.: con tutoriales[22] para formar a profesionales), y el fomento de guías de buenas prácticas para los diseñadores y programadores que desarrollen aplicaciones de aprendizaje profundo, con indicaciones del tipo: Identifique los sesgos que son importantes para su problema y testee esos sesgos: Considere esto como un proceso iterativo, no como algo que «ya está solucionado». Sea transparente sobre su modelo y su rendimiento en diferentes entornos.

[21] **Reddit** es un sitio web (https://es.wikipedia.org/wiki/Reddit) de marcadores sociales y agregador de noticias donde los usuarios pueden añadir texto, imágenes, vídeos o enlaces. Otros usuarios pueden votar a favor o en contra del contenido, haciendo que aparezcan más o menos destacados.

[22] Bias and Fairness in NLP. Tutorial EMNLP 19.

Regulación y control social de los sistemas tecnológicos

El uso malicioso de las herramientas, por ejemplo en el ámbito de PLN para automatizar la generación de información falsa, es una amenaza que hay que afrontar socialmente. De hecho, en la pandemia COVID se han organizado centros de verificación de fuentes, para contrarrestar la expansión vírica de bulos por las redes sociales, no solo con textos sino también con videos manipulados. Además, se están desarrollando técnicas[23] (también con herramientas IA) para detectar este tipo de contenidos. Hay que combinar esfuerzos.

Como señalan Selbts et al. 2019, la equidad y la justicia no son propiedades de los sistemas técnicos, sino que deben analizarse en el contexto social en el que se implantan. Tanto en USA como en Europa hay un posicionamiento cada vez más amplio, de regular la implantación de la tecnología para preservar principios que una sociedad moderna considera básicos. Podemos citar por ejemplo la *General Data Protection Regulation* (GDPR) de la Unión Europea, que en el tema del tratamiento de datos personales incentiva a las empresas para que sigan las normas de buenas prácticas indicadas en la misma. Para la IA en particular hay también acciones e informes que profundizan en los aspectos de equidad, priorizando ésta aún a costa de cierta perdida de precisión.

En última instancia, los sistemas de aprendizaje profundo son un reflejo de la sociedad que intentan modelar y aproximar. Por ello parece justificado, que tanto desde los gobiernos, como del sector privado, se fomenten regulaciones para asegurarse de que estos sistemas no sirvan para afianzar y exacerbar la desinformación y la discriminación, sino más bien, para mitigarlas promoviendo si es necesario, medidas que penalicen los usos inadecuados de los mismos. Extensos análisis interdisciplinares se ocupan en profundidad de estos temas, ver por ejemplo Hildebrandt, M. (2019).

[23] https://newscollab.org/2019/02/04/9-tools-to-identify-fake-images-and-videos/

Extendiendo nuestra mirada, palabras finales

Hay otros temas importantes ligados a la organización y articulación de la investigación y la sociedad que se nos quedan en el tintero. Algunos los hemos visto manifestarse en esta pandemia, y son representativos para todas las disciplinas: Cooperación y cambio de escala en la actividad científica a nivel mundial, los datos que sustentan una afirmación científica deben ser accesibles: «Esa transparencia, la capacidad de reutilizar datos, también es la forma de asegurarnos de la calidad de lo que se publica». Los ensayos deben coordinarse para conseguir estudios robustos y no multitud de pequeñas evidencias de poca utilidad. Colaboración iniciativa publica/privada.

Estas cuestiones merecen un amplio debate para repensar las políticas científicas. Me gustaría que se produjeran en España de forma más participativa con la comunidad científica tal y como viene promoviendo la Confederación de Sociedades Científicas de España (COSCE).

Nos toca ahora un tiempo de incertidumbre, que debemos aprovechar como una oportunidad de transformación hacia una sociedad, al menos, mejor informada para tomar decisiones. Animo a nuestros estudiantes y jóvenes investigadores e investigadoras a que se involucren, y que ejerzan su profesión con una perspectiva global, contribuyendo al despliegue responsable de la tecnología. Es el futuro de todos el que está en juego.

Referencias

Emily M. Bender, Alexander Koller (2020). *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198 , July 5 - 10, 2020. https://www.aclweb.org/anthology/2020.acl-main.463/

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama Adam Kalai (2016). *Man is to Computer Programmer as Woman is to Homemaker?* Debiasing Word Embeddings. 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Caliskan, A., Bryson, J.J. & Narayanan, A. (2017*). Semantics derived automatically from language corpora contain human-like biases.* Science, vol. 356, no. 6334, pp. 183-186. https://doi.org/10.1126/science.aal4230

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou (2018). *Word embeddings quantify 100 years of gender and ethnic stereotypes.* PNAS April 17, 2018 115 (16) E3635-E3644; first published April 3, 2018. https://doi.org/10.1073/pnas.1720347115

Harris, Z. (1954). *Distributional structure*. Word, 10(23):146--162.

Hildebrandt, M. (2019). *Privacy As Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning*. Theoretical Inquires in Law, Vol 20, N.1,2019. https://www7.tau.ac.il/ojs/index.php/til/article/view/1622

Gary Marcus (2018). *Deep Learning, a critical appraisal,* https://arxiv.org/abs/1801.00631

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean (2013) *Efficient Estimation of Word Representations in Vector Space.* https://arxiv.org/abs/1301.3781

George A. Miller (1995). WordNet: *A Lexical Database for English.* Communications of the ACM Vol. 38, No. 11: 39-41.

Philip Resnik (1993). *Selection and information: A class-based approach to lexical relationships*. Ph.D. thesis, University of Pennsylvania.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, Sameer Singh (2020). *Beyond Accuracy: Behavioral Testing of NLP Models with CheckList*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912 July 5-10, 2020. https://www.aclweb.org/anthology/ 2020.acl-main.442.pdf

G. Salton, A. Wong, C. S. Yang (1975) *A vector space model for automatic indexing* Communications of the ACM Vol 18, N. 11. https://dl.acm.org/doi/10.1145/361219.361220

Andrew D. Selbst, Danah Michele Boyd, Sorelle Alaina Friedler, Suresh Venkatasubramanian, Janet Vertesi (2019). *Fairness and Abstraction in Sociotechnical Systems*. Proceedings of the Conference on Fairness, Accountability, and Transparency January 2019 Pages 59–68 https://doi.org/10.1145/3287560.3287598

Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, Gabriel F. Manso (2020*). The Computational Limits of Deep Learning.* https://arxiv.org/abs/2007.05558

Peter D. Turney, P. Pantel, P. (2010). *From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research.* 37(1):141--188, https://arxiv.org/abs/1003.1141

Piek Vossen (Editor) (1998). EuroWordNet*: A multilingual database with lexical semantic networks.* Kluwer Academic Publishers. 1998. ISBN 0792352955

Este artículo forma parte de la lección inaugural curso 2020-2021 UNED (link)