Número 3

LOROS

Sobre los loros que hablan español

Gimena del Rio Riande

agosto 2022

doi:10.5281/zenodo.6567850 (English)
doi:10.5281/zenodo.6567966 (Español)

1. Una lora llamada MarIA

El español es el segundo idioma más hablado en el mundo como lengua materna.[1] Informes oficiales, estudios basados en encuestas y Wikipedia lo confirman. Y Google lo puede predecir. Los datos hablan por sí solos: más de veinte países cuya lengua oficial es el español, un gran crecimiento en su enseñanza como segunda lengua en universidades norteamericanas y veintitrés academias de la lengua española repartidas por el mundo.[2] No obstante, el español naufraga desde el siglo pasado ante el inglés en la batalla geopolítica y simbólica de las lenguas por la publicación y comunicación científica, sin reclamar demasiados avances tecnológicos para revertir esta situación.[3]

En los últimos años las investigaciones a nivel global han puesto su lupa en la inteligencia artificial (IA) y el machine learning, otorgando a la lingüística computacional y al procesamiento del lenguaje natural (PLN) un lugar de relevancia en la currícula académica de diferentes disciplinas y campos, como, por ejemplo, las Humanidades Digitales. Sin embargo, el desbalance Norte-Sur en este campo es evidente. La mayoría de los recursos PLN todavía se desarrollan principalmente para el idioma inglés. Hacer que estos recursos estén disponibles en español toma bastante tiempo, e incluso cuando esto sucede es generalmente a través de versiones multilingües que no son tan eficientes como el recurso original.

> *Los países del indefinido Sur Global los que miran de reojo la tecnología, desde ese lugar de consumidor incómodo de desarrollos ajenos y receptor de críticas de las grandes naciones.*

Si nos recortáramos únicamente sobre los países hispanohablantes, ha de decirse que España viene tomando la delantera en la investigación sobre IA, PLN y lengua española, con proyectos oficiales como, por ejemplo, *Aporta*, donde se lleva a cabo la actividad "Tecnologías emergentes y datos abiertos: Inteligencia Artificial,"[4] anunciada desde el portal del Ministerio de Asuntos Económicos y Transformación Digital.[5] Si bien la iniciativa española es necesaria y da cuenta de un trabajo sostenido desde lo humano y lo tecnológico, se trata de una propuesta que adopta un enfoque absolutamente tecnopositivista, donde no hay demasiada reflexión crítica sobre los peligros y dificultades de la IA a la hora de trabajar con datos humanos para desarrollar modelos de lenguas. Tampoco encontramos mención alguna, por ejemplo, acerca de las tecnologías del silicio, imprescindibles para el desarrollo de infraestructuras para la IA, que son parte de una

industria estratégica y un activo geopolítico que hoy muchos países producen pero solo aprovechan realmente unos pocos.

Hacia fines del mes de julio de este año desde el país ibérico se dio a conocer el primer gran proyecto sobre lengua e IA desarrollado por la Biblioteca Nacional de España (BNE). MarIA, tal es su nombre, es el primer modelo de IA masivo de la lengua española y nació de una gran cantidad de datos que la BNE ingestó en el supercomputador MareNostrum del Barcelona Supercomputing Cente. Estos datos no están en dominio público ni accesibles en internet, sino que son los archivos en formato WARC resultantes del rastreo y archivado de la web española, que, por ley la BNE *scrapea* y conserva.[6]

Vayamos a los números. Para crear este enorme corpus se utilizaron 59 terabytes del archivo web de la BNE, y para dar este primer paso de creación, curación y compilación se necesitaron 6.910.000 horas del superordenador MareNostrum. Como resultado, se obtuvieron 201.080.084 documentos limpios y sin duplicidades que ocupan un total de 570 gigabytes. El segundo paso del entrenamiento, basado en la tecnología de redes neuronales, necesitó de 184.000 horas de procesador y más de 18.000 horas de GPU.[7] De acuerdo al *paper* publicado hace unos meses, los modelos liberados, que tienen entre 125 millones y 355 millones de parámetros respectivamente, seguirán ampliándose con nuevas y diferentes fuentes de archivos, como por ejemplo publicaciones científicas del Consejo Superior de Investigaciones Científicas (CSIC) de España y la Wikipedia en español.[8]. También se prevé comenzar a entrenar modelos para otras lenguas romances como el catalán, gallego, euskera, portugués y para variantes mucho más complejas o tal vez inexistentes, como eso que se suele llamarse español de América.

Como viene sucediendo en los últimos tiempos, son los países del indefinido Sur Global los que miran de reojo la tecnología, desde ese lugar de consumidor incómodo de desarrollos ajenos y receptor de críticas de las grandes naciones.[9] Un claro ejemplo es la acusación a ciertos países del sur sobre sus industrias contaminantes que convive con un total olvido acerca de cuánto contamina — principalmente al sur — la nueva industria del bitcoin, blockchain y de la IA de los países ricos.[10]

Podría decir que mi perspectiva acerca de los modelos de lengua y la IA es la de una investigadora del sur que sospecha bastante de la representatividad del español que hoy habla MarIA y de sus capacidades, por ejemplo, para adaptarse a la llamada *norma rioplatense*[11] en aplicaciones de traducción, subtitulación, predicción o corrección automática de lengua, o los populares chatbots. Si bien, como leíamos más arriba, es probable que el modelo siga creciendo y mejorándose, quisiera recordar, siguiendo a Mboa Nkoudou, que tecnología y colonialismo pueden convertirse en sinónimos.[12] El investigador camerunés acuña el neologismo tecnocolonialidad y lo describe en diferentes dimensiones que nos ayudan a comprender por qué es importante pensar la tecnología antes de usarla: la primera dimensión es la del discurso tecno-utópico que acuñan muchos proyectos y desarrolladores de tecnología, que esconde el modo en el que la tecnología puede convertirse en un elemento de dominación o eliminación de una cultura; la segunda dimensión es la relacionada con la facilidad de transferencia de la tecnología, lo que la vuelve invisible; la tercera es la del peligro de colonizar a través del conocimiento producido tecnológicamente e incorporarse a cualquier práctica (neo)capitalista.

Que no se equivoque el lector: estoy convencida de que necesitamos muchas y más MarIAs. Así y todo, creo que estas MarIAs deberían explicarse en un contexto de investigación abierta, si de verdad quieren superar los debates acerca del colonialismo tecnológico y convertirse en modelos reusables y reproducibles en los más de veinte países que hablan español en nuestro

planeta y que, en su gran mayoría, ocupan el conglomerado comúnmente llamado América Latina. No conozco estudios sobre el modelo de la BNE en cuanto a sesgo lingüístico, geográfico, o racial. Tratándose de la liberación de un modelo que quiere convertirse en oficial o pionero para el español, algo más de tiempo y dinero puede invertirse para que la BNE trabaje junto con las academias de la lengua española de los diferentes países antes de liberar su modelo para adopción masiva o, al menos, para documentar los problemas que, a día de hoy, este modelo podría tener.

2. Un loro llamado BERTIN

La misma semana en la que la llegada de MarIA era ampliamente anunciada en los medios nació BERTIN. Un proyecto cuyo propósito era pre-entrenar un modelo de RoBERTa desde cero, durante el Flax/JAX Community Event llevado a cabo entre la semana del 7 al 14 de julio de 2021.[13] Quedé muy impresionada al ver que, de un momento a otro, había dos modelos de RoBERTa disponibles en español, así que decidí entrevistar a uno de sus mentores, el Doctor Javier de la Rosa, un joven y brillante investigador español, humanista y lingüista computacional, además de experto en PLN. En la siguiente sección retomo unas pocas preguntas de nuestro diálogo, ya que considero ayudan a pensar por qué, cómo y para qué necesitamos modelos de lengua para el español y cómo las Humanidades y las Humanidades Digitales podrían participar en este tipo de iniciativas.

2.1. Javier, BERTIN y Gimena

Gimena (G): *BERTIN se propone como un proyecto colaborativo ¿Qué papel juegan o podrían jugar los humanistas o lingüistas en la creación de datasets y en adjudicación de significado de los outputs?*

Javier (J): La construcción de BERTIN ha estado desde el inicio orientada a la comunidad. Todo en abierto por y para la comunidad. Mucha gente mostró su interés en participar desde que comenzamos la propuesta: programadores, ingenieros, investigadores en IA, humanistas digitales, lingüistas computacionales. Por desgracia, por una cuestión meramente práctica (sólo disponíamos de recursos durante los 10 días del Flax/JAX Community Event esponsoreado por Google Cloud), no todos pudieron participar. Sin embargo, mantuvimos conversaciones muy interesantes y enriquecedoras acerca de la orientación y las metas del proyecto.

Creo que uno de los aspectos al que podrían contribuir tanto los humanistas como los lingüistas en este tipo de proyecto es a la caracterización del sesgo. En nuestro caso, este tema surgió como un *afterthought* pero poco a poco, según veíamos que el modelo, por ejemplo, solía preferir palabras del castellano peninsular (“coche” versus “auto” versus “carro”), nos dimos cuenta de que este podía padecer de cierto sesgo geográfico que había que resaltar e intentar combatir. Lamentablemente, no es fácil disminuir el sesgo antes de que el modelo esté entrenado. Pero miembros del equipo con *background* lingüístico hicieron un muy buen trabajo documentándolo.

No todos los sesgos son malos. Por ejemplo, una forma de tener un modelo de lengua que sea capaz de comprender bien otras variantes del español es sesgando el dataset para sobrerrepresentar textos de una variante menos extendida y producir un modelo capaz de adaptarse mejor a diversas variedades.

G: *¿Cómo deberíamos crear datos o datasets lingüísticos de manera que estos no perpetúen los prejuicios y las normas hegemónicas?*

J: En estos momentos, la IA, en general, y los modelos de lengua, en particular, son un campo muy activo. Empiezan a aparecer *toolkits* para analizar los datos antes de que comience el entrenamiento para, por ejemplo, disminuir ciertos sesgos conocidos o encontrar patrones que podrían llegar a ser problemáticos de cara a la explotación de los modelos. Pero he de decir que no todos los sesgos son malos. Por ejemplo, una forma de tener un modelo de lengua que sea capaz de comprender bien otras variantes del español es sesgando el dataset para sobrerrepresentar textos de una variante menos extendida y producir un modelo capaz de adaptarse mejor a diversas variedades.

G: *Si tuvieras que dar una asignatura sobre curaduría de datos con perspectiva humanística, ¿qué métodos o prácticas enseñarías? ¿Podemos imaginar un futuro en el que las humanidades se vuelvan tan importantes para los profesionales de datos?*

Es cierto que los humanistas llevamos curando datos desde el inicio de los tiempos, pero estamos muy por detrás y nos movemos muy lentamente en comparación con el avance de la tecnología y la IA. Dicho lo cual, existen iniciativas para equiparar un poco los dos mundos en cuanto a datos se refieren. Por ejemplo, hace unos daños se inició el proyecto *Collections as Data*, de Thomas Padilla, que distintas instituciones, sobre todo bibliotecas y archivos, adoptaron, y que de una forma u otra han cristalizado en modelos de lengua y otros artefactos de la IA como el de la Biblioteca Nacional de Noruega, que usó su fondo digital para construir un modelo para el noruego,[14] o el de la BNE. Así que supongo que si tuviera que enseñar curaduría de datos lo haría desde esta perspectiva híbrida, señalando la importancia de la preservación para el futuro y la necesidad de poder disponer de los datos en formatos no sólo interoperables sino reutilizables por quiénes lo necesiten.

G: *¿Ves consecuencias negativas sobre el hecho de que los humanistas dependan de Google para usar o poner a prueba sus modelos? ¿Hay posibilidad de que podamos desarrollar recursos de código abierto para este tipo de investigación o deberíamos seguir confiando en las grandes empresas de tecnología con prácticas éticas cuestionables, o tal vez empezar a desarrollar modelos de lengua más pequeños?*

Las grandes empresas no tienen su negocio en el software, sino en el hardware y la infraestructura. La mayoría de los avances en IA se hacen en abierto y ese es el motivo de que se haya avanzado tan rápido en los últimos 10 o 15 años. Aun así, también he de añadir que los servicios de infraestructura son sólo una parte de la recaudación en las cuentas anuales de estas empresas, pues por ejemplo, en el caso de Google, la mayor partida sigue viniendo de la publicidad. Sin embargo, la adopción del software libre y la liberación tanto de modelos como de librerías de código abierto es una forma de presionar para que los desarrollos se hagan en sus

doi:10.5281/zenodo.6567966

plataformas. Por ejemplo, si quieres entrenar un modelo para el vasco, puedes hacerlo en Google Cloud y luego marcharte. Pero si Google te da el código, la integración con la plataforma, y te enseña cómo hacerlo para otros 30 idiomas, pues te ahorra buena parte del trabajo. Por supuesto, también puedes intentar entrenar tu modelo en un superordenador como el MareNostrum, que se ha usado para el modelo de la BNE, pero el acceso a este tipo de recursos no es fácil ni directo. O incluso podrías comprar unas 120 tarjetas gráficas NVIDIA (que no son baratas ni fáciles de conseguir) y ponerlas a entrenar, pero es una inversión en fungibles difícil de justificar cuando existen soluciones *on demand* que prometen un uso eficiente del dinero. Lo que nosotros hemos intentado con BERTIN es que se puedan reducir los costes al entrenar un modelo. Aún estamos muy lejos de poder entrenar estos modelos masivos en nuestros ordenadores personales, pero no por ello debemos dejar de avanzar en este aspecto.

3. Algunas conclusiones

Creo que una conclusión importante que surge del modelo de BERTIN es que demuestra que, aunque no podemos escapar de la necesidad de grandes datos y grandes máquinas para producir modelos de lengua con buen rendimiento, puede reducirse un poco el gasto que supone entrenarlos, disminuyendo el tiempo y los datos necesarios para ello. Esto posibilita que equipos más pequeños puedan acercarse a este mundo que de momento parece sólo reservado a las grandes compañías e instituciones del norte global . También supone una propuesta más alineada con los ideales de los movimientos de ciencia abierta y hasta con los activistas ambientalistas, más conscientes del uso de la tecnología, que nos evita caer en un tecnopositivismo que ve en el desarrollo y el aporte de las nuevas tecnologías el sistemático reemplazo de los sistemas productivos existentes o el simple avance de la ciencia y la civilización.

Los modelos de lengua . . . pueden generar resúmenes de textos, encontrar similitudes en colecciones textuales, clasificar obras en base a tema, género u otros aspectos del contenido, y enlazar datos a partir únicamente de información contenida en el texto.

Si bien el campo de la IA supera el horizonte de trabajo de los humanistas, es evidente que los modelos de lengua tienen muchas aplicaciones con un impacto inmediato en las humanidades y las humanidades digitales: pueden optimizar las salidas de los sistemas para el reconocimiento óptico de caracteres (OCR), mejorar técnicas de estilometría y atribución de autoría, generar resúmenes de textos, encontrar similitudes en colecciones textuales, clasificar obras en base a tema, género u otros aspectos del contenido, enlazar datos a partir únicamente de información contenida en el texto, etc. Pero también es importante señalar que las salidas de estos sistemas también deben ser documentadas y sometidas al estudio y a la crítica. Y aquí es donde nuevamente este tipo de trabajo bien puede alinearse con las propuestas de la ciencia abierta. Volviendo a los proyectos de Javier, puedo traer aquí otra de sus iniciativas, ALBERTI, donde estos modelos se aplican a proyectos de humanidades digitales, completando poemas de forma automática para que, en un segundo paso, evaluadores humanos determinen qué sustituciones de ALBERTI resultan “más poéticas.”[15] Un buen ejemplo que, desde las humanidades, demuestra que todos los sistemas deben ser evaluados y limitados por los humanos, ya que no sólo en las humanidades sino en la vida cotidiana necesitamos de la reflexión sopesada para tomar decisiones acertadas (pensemos en los

problemas que hasta ahora han traído la revisión no supervisada de curriculum vitae o de antecedentes para pedir hipotecas, etc.). “A mucho hablar, mucho errar”, dice el refrán. Y parece cierto para los humanos y las máquinas.

doi:10.5281/zenodo.6567966

1. El título y contenido de este ensayo juega con el del artículo “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", de Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* 610-623. https://doi.org/10.1145/3442188.3445922. También, quiero destacar que este ensayo y “On Spanish Speaking Parrots” no tienen una relación de traducción directa. En un principio, al ser invitada a participar de la mesa redonda virtual “Machine Predictions and Synthetic Text”, que se celebró el 26 de octubre de 2021, y de la cual participé junto a destacadísimos especialistas como Lauren Klein, Ted Underwood, Toma Tasovac y dos de las autoras del mencionado artículo (Angelina McMillan-Major y Margaret Mitchell), escribí el texto en español, con el fin de ordenar mis ideas. No obstante, a medida que se acercaba el evento, reescribí el texto en inglés, intentando huir (con mediana o poca suerte) del insalvable problema del traslado de estructuras gramaticales y expresiones del español al inglés. Agradezco a mi querida hermana, María Cruz del Rio, la corrección del primer borrador en inglés y a Grant Wythoff la revisión final. ↩

2. Según el Instituto Cervantes, 528 millones de personas en el mundo usan el español como lengua nativa, segunda lengua o lengua extranjera. El español, una lengua que hablan 580 millones de personas, 483 millones de ellos nativos. (2019). *Instituto Cervantes*. https://www.cervantes.es/sobre_instituto_cervantes/prensa/2019/noticias/presentacion_anuario_madrid.htm. ↩

3. A este respecto, una herramienta multilingüe que podría adaptarse a diferentes disciplinas, creada por un biólogo argentino y un bioinformático estadounidense, es PanLingua. PanLingua permite hacer búsquedas en la lengua del usuario sobre la base de datos de bioRxiv.org utilizando Google Translate para proporcionar traducciones automáticas del término de consulta a diferentes lenguas. Accesible desde: https://panlingua.rxivist.org/. ↩

4. Accesible desde: https://datos.gob.es/es/documentacion/tecnologias-emergentes-y-datos-abiertos-inteligencia-artificial. ↩

5. Accesible desde: https://datos.gob.es/es/acerca-de-la-iniciativa-aporta ↩

6. Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., & Villegas, M. (2021). Spanish language models. arXiv:2107.07253 [cs]. http://arxiv.org/abs/2107.07253. Los autores del artículo, además desarrolladores del proyecto, no mencionan si, para alimentar a MarIA se utilizó la gran cantidad de texto literario a la que sólo la BNE, como biblioteca nacional que ha digitalizado la mayor parte de su acervo, tiene acceso. ↩

7. Gutiérrez-Fandiño et al., http://arxiv.org/abs/2107.07253. ↩

8. Tampoco se precisa si se trata de todas las Wikipedias que se escriben en español o la Wikipedia española. ↩

9. Destaco esta interesante iniciativa sobre IA surgida desde espacios tradicionalmente considerados como periféricos con relación al desarrollo de este tipo de investigaciones: https://points.datasociety.net/ai-in-the-global-south-sites-and-vocabularies-e3b67d631508 ↩

10. Howson, P. (2020). Climate crises and crypto-colonialism: Conjuring value on the blockchain frontiers of the global south. *Frontiers in Blockchain,* 3 (22). https://doi.org/10.3389/fbloc.2020.00022. También, muy interesante esta nota sobre las granjas ilegales de bitcoin y su relación con los cortes de luz en Argentina: https://www.iproup.com/finanzas/28621-cortes-de-luz-el-gobierno-busca-granjas-de-minado-de-bitcoin. ↩

11. Ríos de tinta han corrido acerca del español hablado en la zona del Río de la Plata, el origen del *voseo*, etc. Para quien no está demasiado familiarizado con él, puede empezar leyendo la entrada de Wikipedia: https://es.wikipedia.org/wiki/Espa%C3%B1ol_rioplatense. ↩

12. Mboa Nkoudou, Thomas (2020). Les makerspaces en Afrique francophone, entre développement local durable et technocolonialité: trois études de cas au Burkina Faso, au Cameroun et au Sénégal. Tesis de doctorado defendida en la Université Laval, Canadá. https://corpus.ulaval.ca/jspui/handle/20.500.11794/67577. ↩

13. Puede leerse más sobre BERTIN en: https://huggingface.co/bertin-project/bertin-roberta-base-spanish. ↩

14. Nota de la autora: Se refiere al trabajo que de describe en Kummervold, P., de la Rosa, J., Wetjen, F., & Brygfjeld, S.A. (2021). Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. http://arxiv.org/abs/2104.09617. ↩

15. Para saber más sobre ALBERTI, véase: https://huggingface.co/spaces/flax-community/alberti ↩