Proyecto Espacio de Observación de Inteligencia Artificial en Español. Resumen Ejecutivo Año 2.

Enrique Amigó[1], Jorge Carrillo-de-Albornoz[1], Andrés Fernández[1], Julio Gonzalo[1], Miguel Lucas[2], Guillermo Marco[1], Roser Morante[1], Jacobo Pedrosa[1], Laura Plaza[1], Eva Sánchez[1], Augusto Villa[2]
[1] Natural Language Processing and Information Retrieval Group, UNED
[2] LLorente & Cuenca Madrid, S.L.
Autor de contacto: Julio Gonzalo - julio@lsi.uned.es

Resumen ejecutivo

En el marco del proyecto ODESIA (espacio de observación para la Inteligencia Artificial en español, fruto de un convenio entre Red.es y UNED financiado por la Estrategia Nacional de Inteligencia Artificial), se ha realizado una estimación de la brecha de desarrollo de la Inteligencia Artificial en inglés y en español para el Año 2 del proyecto. Como en el Año 1, esta brecha se ha medido en cuatro ámbitos: (1) estado del arte de las tecnologías del lenguaje; (ii) soluciones de mercado; (iii) nivel de adopción de la tecnología; y (iv) experiencia de uso.

En la primera iteración del proyecto se realizó un estudio en profundidad de los dominios y tipos de problemas a nivel abstracto (clasificación, etiquetado, ranking, etc.) en las tecnologías del lenguaje con el fin de asegurar una buena cobertura en el estudio de la brecha lingüística. A lo largo del último año las tecnologías basadas en grandes modelos de lenguaje ha revolucionado la capacidad de los sistemas de resolver problemas diversos. Por ello, se han revisados las tipologías de tareas y las dimensiones de evaluación de sistemas inteligentes en el contexto de las tecnologías del lenguaje, incluyendo aspectos como los sesgos en las respuestas, contenidos no informativos o engañosos, competencias cognitivas de los sistemas, etc. El análisis de estas dimensiones se ha reflejado tanto en el diseño de nuevos datasets como en la elaboración de encuestas.

Los resultados, que pueden verse como tabla resumida en la Figura 1, son los siguientes:

- **Ámbito 1 (Estado del arte): brecha global del 66 %**. La brecha promedio sobre todos los aspectos medidos es similar a la del año 1. Dentro del estado del arte, en cuanto a diseminación y recursos, se mantiene la tendencia observada en la primera iteración del proyecto, con algunas diferencias. Al igual que en la iteración anterior, el factor más desfavorable es la diseminación, con una **brecha en publicaciones y proyectos subvencionados del 98 % y 96 % respectivamente**. En concreto, la brecha en proyectos subvencionados ha ascendido del 88 % al 96 % respecto del año anterior. En cuanto a recursos, **la disponibilidad de textos en internet (R.0) se mantiene estable**, como era de esperar dado que no es un indicador susceptible de cambios bruscos. **La brecha en disponibilidad de modelos de lenguaje se mantiene también muy similar (R.1)**. Se observa un **aumento importante de la brecha en disponibilidad de datos anotados en repositorios (R.2)**, sobre todo debido al incremento de datos para el inglés en Hugging Face. La presencia de datos de campañas de evaluación, sin embargo, permanece bastante constante en las fuentes consideradas, si tenemos en cuenta el reducido número de muestras y el consiguiente efecto en la volatilidad del indicador R.2.b. El mayor esfuerzo de medición en este ámbito ha sido para calcular la brecha de efectividad de los modelos de lenguaje:
    - En el segundo año, **hemos pasado de 6 a 10 tareas discriminativas en el leaderboard ODESIA CORE** (con datos generados en el proyecto), para un total de 15 tareas discriminativas

Figura 1: Estimación ODESIA de la brecha entre español e inglés, Año 2

en el leaderboard ODESIA EXTENDED (que incluye 5 datasets más de dominio público). En cuanto a tareas abstractas se refiere, para la estimación de la brecha, se han cubierto la clasificación binaria (EXIST 2022 tarea 1, EXIST 2023 tarea 1, DIPROMATS 2023 tarea 1), la clasificación multiclase, jerárquica y/o multilabel (EXIST 2022 tarea 2, EXIST 2023 tareas 2 y 3, DIPROMATS 2023 tareas 2 y 3), la evaluación en modo *learning with disagreement* (EXIST 2023, tareas 1,2 y 3), la regresión (STS 2017), y el etiquetado de secuencias (DIANN 1 y 2). Dentro de lo que se consideran problemas dinámicos, en la estimación de la brecha consideramos el *question answering* con anotación de secuencias (SQUAD/SQAC 2024). Además, hemos completado dos datasets adicionales para medir la efectividad de modelos generativos: UNED-ACCESO (de exámenes tipo test de once asignaturas de acceso a la universidad) y CURIA (de resúmenes en lenguaje claro de textos jurídicos).

- En conjunto, los dominios y áreas de aplicación cubiertos en el segundo año en el cálculo de la brecha en efectividad incluyen: geopolítica y desinformación (DIPROMATS), biomedicina y extracción de información (DIANN), publicaciones académicas y machine reading (SQUAD/SQAC), redes sociales y contenidos tóxicos (EXIST), noticias (MLDOC), conocimiento enciclopédico y consultas en buscadores (MULTICONER), resolución de similitud textual (STS), información jurídica y resúmenes en lenguaje claro (CURIA), y exámenes de conocimiento general (UNED-ACCESO).
- **Sobre las tareas discriminativas se ha medido en el leaderboard ODESIA EXTENDED una brecha promedio del** 20 ± 06 **%, consistente con la medición del año anterior**. Hay que destacar que la brecha es positiva en todas las tareas discriminativas evaluadas, excepto en una. Es decir, independientemente del problema abordado, los modelos tienen una efectividad menor en español que en ingles para tareas equivalentes. Otro resultado destacable de este estudio es que **los modelos de lenguaje en español no obtienen mejores resultados que los modelos multilingües equivalentes en español**.
- Más allá de las actividades previstas en el convenio, ante la irrupción de la IA generativa se ha comenzado a evaluar en ODESIA la brecha de rendimiento de modelos generativos. Se han realizado dos experimentos: (i) se ha evaluado GPT-4 en modo zero-shot sobre tres tareas discriminativas del leaderboard, en las que se ha obtenido una **brecha promedio del 18 % en el rendimiento de GPT-4 en inglés y español**. Esta cifra es compatible con la brecha medida para los modelos discriminativos, aunque hay que ampliar la experimentación para consolidarla e incluirla en la medición de la brecha; y (ii) se han evaluado seis modelos generativos (GPT-4, Claude 3 Opus, GPT-3.5, Llama-2, Mistral y Gemma) sobre el dataset de exámenes UNED-ACCESO desarrollado dentro del proyecto. En este caso se ha observado una **brecha promedio del 12 % en los modelos abiertos** y ligeramente negativa en los propietarios (-1 %). Esta estimación de la brecha tiene seguramente un sesgo derivado de posible contaminación, ya que las preguntas originales están en español y se han traducido dentro del proyecto. Es decir, es probable que, al menos los modelos propietarios, hayan visto las soluciones a las preguntas de examen en su formato original en español. En conjunto, se requiere más experimentación para medir con fidelidad la brecha de los modelos generativos.

- **Ámbito 2 (Soluciones de mercado): 8 %**. Esta brecha corresponde a la brecha de funcionalidades en productos comerciales disponibles en ambos idiomas y ha descendido un punto respecto al año anterior.

- **Ámbito 3 (Nivel de adopción): 61 %**. En cuanto a la brecha en nivel de adopción, aparecen variaciones a nivel de indicador específico, aunque el promedio se mantiene prácticamente constante. En concreto, ascienden las menciones de productos en medios (I.A.3) de un 76 % a un 83 %, las menciones de tecnologías de la lengua en medios de un 49 % a un 70 % (I.A.4) y la brecha en adopción para uso personal (I.A.7) que asciende de un 33 % a un 36 %. Sin embargo, descienden la brecha en menciones de productos informes (I.A.1) de un 95 % a un 93 %, las menciones de

tecnologías en informes (I.A.2) de un 56 % a un 53 %, el impacto de las tecnologías en la empresa (I.A.5) de un 66 % a un 60 % y la adopción para uso profesional (I.A.7) de un 46 % a un 33 %.

- **Ámbito 4 (Experiencia de uso): -6 %**. En cuanto al ámbito de experiencia de usuario, también se mantiene constante en promedio, aunque hay variaciones a nivel de indicador específico. Se ha obtenido el mismo patrón que el año anterior. Los indicadores de polaridad reputacional (I.E.1) y curvas de valor (I.E.2) donde los indicadores se estiman a partir de opiniones en la web, aparece una brecha negativa en favor del español. Lo mismo ocurre en el indicador de limitaciones (I.E.4) en donde se encuesta a individuos sobre las deficiencias específicas de los productos analizados. Sin embargo, en el caso de las encuestas de satisfacción (I.E.3) los usuarios de tecnologías en inglés se sienten más satisfechos, efecto que ha crecido en los resultados de este año (12 % frente al 2 % obtenido en el año anterior). Esto se compensa con la reducción de brecha en favor del español en los otros tres indicadores (-9 % frente a 2 %, -9 % frente a -4 % y 16 % frente a 25 %). De manera adicional, se han ampliado las encuestas sobre limitaciones para identificar aspectos de la calidad de las tecnologías definidos en la sección **??** de este documento.

En conjunto, hemos medido de nuevo una brecha significativa en casi todos los ámbitos estudiados, lo que confirma la necesidad de impulsar la IA en español como parte de cualquier estrategia nacional de desarrollo tecnológico.

Agradecimientos

Este trabajo ha sido financiado por la Unión Europea - NextGenerationEU a través del “Plan de Recuperación, Transformación y Resiliencia”, por el Ministerio de Asuntos Económicos y Transformación Digital y por la UNED. Sin embargo, los puntos de vista y las opiniones expresadas son únicamente los del autor o autores y no reflejan necesariamente los de la Unión Europea o la Comisión Europea. Ni la Unión Europea ni la Comisión Europea pueden ser consideradas responsables de los mismos.