EL DIARIO VASCO

Latxa

La tecnología vasca que supera a ChatGPT

El centro Hitz, de la UPV/EHU, ha informado que con las últimas mejoras implantadas el modelo en euskera ha dado «resultados esperanzadores»

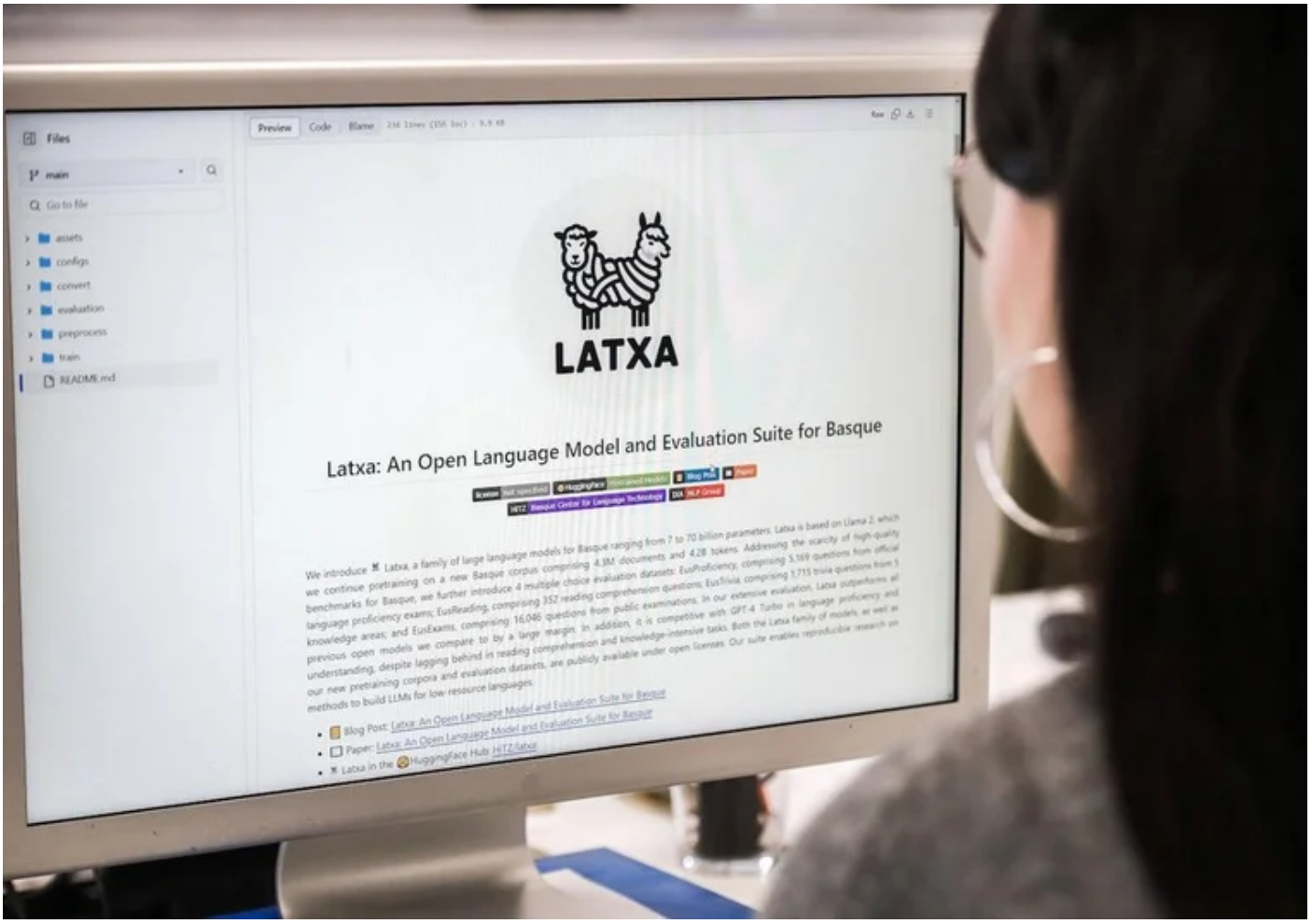

Imagen de la web del modelo Latxa elaborado por Hitz zentroa. Borja Luna

Jon Agirre
Donostia

Seguir

Lunes, 29 de abril 2024, 17:30

Comenta

l modelo de lenguaje Latxa, elaborado por el centro Hitz de la UPV/EHU ha superado a la versión de ChatGPT original lanzado hace un año (ahora conocido como GPT 3.5), y por primera vez para un modelo abierto de un idioma de pocos recursos digitales, supera la última versión (GPT-4) en competencia lingüística. Así lo comunicó ayer el centro, que señala que con las últimas mejoras implementadas en el modelo en euskera este ha dado «resultados esperanzadores».

El paso logrado por Hitz con el modelo Latxa quizá sea difícil de medir para el usuario medio, ya que estos modelos no están pensados para que el público general los utilice directamente, pero es considerable y abre la puerta a nuevos avances y a desarrollar herramientas exitosas que utilicen tecnología lingüística para el euskera. La irrupción de la Inteligencia Artificial ha cambiado de raíz el sector tecnológico y esta vía de trabajo de la UPV/EHU busca minimizar la brecha existente entre los idiomas hegemónicos como el ingles y el euskera, en situación minorizada, con menos recursos digitales para desarrollar modelos o tecnologías similares.

No es la primera vez que desde Hitz insisten en la brecha entre idiomas. El director del centro, Eneko Agirre, ya subrayó que los resultados logrados, en general, suelen ser «bastante inferiores» si se comparan con el mencionado ChatGPT en inglés o el Gemini en castellano. En esa línea de trabajo, la primera versión de Latxa, que fue presentada en enero de 2024, ha sido entrenada «sobre el mayor corpus público en euskera», que extiende el ya existente EusCrawl, y que también se distribuye junto con los modelos, han explicado los investigadores.

La nota de Hitz detalla que en total son más de 4 millones de documentos y 1200 millones de palabras, doblando en tamaño de corpus de textos de licencia libre a sus predecesores hasta el momento, y que los resultados también serán «mejores» a medida que aumenta el modelo «en tamaño». En esa línea, con el fin de evaluar la calidad de los modelos, se han construido varios bancos de prueba sobre competencia lingüística, comprensión lectora, cultura general y exámenes profesionales.

Software libre

Latxa está basado en los modelos Llama de Meta, denominados LLM, y reúne entre 7 y 70 mil millones de parámetros. Uno de los tres autores principales, Julen

Etxaniz ha subrayado el mérito de «haber conseguido resultados tan buenos con un corpus relativamente tan pequeño, porque abre la puerta a más mejoras según investiguemos en nuevas técnicas. Es un resultado muy prometedor, no solamente para el euskera sino para el resto de lenguas con pocos recursos digitales».

Por su parte, Pérez destaca que «el euskera se encuentra en la posición 50 entre los idiomas del mundo según la cantidad de texto en Internet, y hay decenas de otras lenguas que tienen cantidades similares de texto, con lo que las técnicas aplicadas al euskera pueden también aplicarse a esos idiomas con resultados previsiblemente similares».

Finalmente, Sainz subraya que «se ha superado a GPT-4 Turbo en competencia lingüística pero no en el resto de pruebas, lo que sugiere que las capacidades de los modelos de lenguaje de gran tamaño en un idioma dado no están determinadas solamente por su capacidad lingüística. Esto abre la puerta a mejorar los resultados en idiomas con pocos recursos digitales como el euskera, según se desarrollen mejores modelos abiertos para el inglés. Con la creciente competencia en esta área, bastaría con esperar a modelos abiertos mejores que Llama y entonces podríamos entrenar un Latxa que supere a GPT-4».

Modelo exportable

Los resultados obtenidos con el modelo Latxa pueden ser de utilizada para el resto de lenguas minoritarias, según avanzaron los investigadores. «Las técnicas aplicadas al euskera se pueden aplicar a otras lenguas y pensar que los resultados serán similares». Cabe recordar que desde el pasado noviembre Hitz está adherida a las infraestructuras tecnológicas punteras europeas. Por un lado, el actual CLARIN-ERIC europeo, orientado a los datos y procesos digitales relacionados con la lengua; y, por otro, DARIAH-ERIC, orientado a los datos y procesos digitales relacionados con las ciencias humanas digitales.

Comenta

Reporta un error