ODESIA

Informe del estado del arte

1. INTRODUCCIÓN

El objetivo del presente informe es presentar, de forma breve y con fines divulgativos, la metodología y resultados del estudio realizado para medir la brecha o "gap" existente entre el inglés y el español con relación al estado del arte de las tecnologías de la lengua. Este gap, a su vez, se utilizará para el cálculo de una métrica agregada que medirá la brecha en tecnologías de la lengua entre el español y el inglés.

El estado de las tecnologías de la lengua puede estudiarse desde diferentes ámbitos o perspectivas complementarias. En concreto, en el marco del proyecto del Espacio de Observación de Inteligencia Artificial en Español, este estudio se realiza desde los siguientes ámbitos: (1) estado del arte, cercano al mundo académico; (2) soluciones de mercado, referido a la disponibilidad de dispositivos de consumo que utilizan tecnologías de Procesamiento del Lenguaje Natural (PLN), sus funcionalidades y características; (3) nivel de adopción, que mide la incorporación de tecnologías de la lengua en el entorno industrial; y (4) experiencia de usuario, que analiza el grado de satisfacción por parte de usuarios finales.

El presente informe aborda el primero de los ámbitos, estado del arte, y considera, a su vez, los distintos componentes o elementos que contribuyen al cálculo de la brecha entre el inglés y el español en este ámbito: diseminación científica y divulgación de resultados, financiación de la investigación, disponibilidad de recursos lingüísticos y datos anotados, y efectividad de las tecnologías y modelos. Como resultado, se ofrece un estimador del gap existente en cada uno de estos componentes, así como del gap agregado que representa la distancia entre ambos idiomas en el estado del arte de las tecnologías de la lengua.

2. DISEMINACIÓN CIENTÍFICA, FINANCIACIÓN Y DIVULGACIÓN DE RESULTADOS

Los indicadores de diseminación (referidos en nuestra documentación con la letra D) reflejarán el estado de las tecnologías en español e inglés en cuanto a la divulgación de resultados en medios científicos. Para el cálculo de estos indicadores se ha tenido en cuenta dos elementos:

- Indicador D.1: Artículos científicos publicados.
- Indicador D.2: Proyectos subvencionados.

Para el cálculo del indicador D.1, se consideran congresos internacionales de primer nivel de índole internacional celebrados en el intervalo temporal considerado (en este primer informe, 2018-2022). Se realizan búsquedas manuales y mediante palabras clave para identificar (1) artículos que describen investigaciones realizadas sobre datos en inglés y (2) artículos que

Financiado por la Unión Europea NextGenerationEU

artículos que describen investigaciones realizadas sobre datos en español. El indicador se calcula como la diferencia entre ambos valores dividido por la suma de ambos.

Para el cálculo del indicador D.2, se realizan búsquedas en bases de datos públicas de proyectos de investigación subvencionados. En las búsquedas, se usan los filtros disponibles para seleccionar proyectos de PLN, se filtran los proyectos obtenidos en la primera búsqueda conforme al intervalo temporal considerado (2018-2022), y se revisa manualmente el título y la descripción de los proyectos, para hallar los que experimentan sobre datos en español y los que experimentan sobre datos en inglés. El indicador se calculará como la diferencia entre proyectos que trabajan con datos textuales en inglés y proyectos que trabajan con texto en español, dividido por la suma de ambos.

Como resultado, se obtienen los siguientes valores que cuantifican la brecha existente entre ambos idiomas para los dos conceptos analizados:

INDICADOR D.1: BRECHA EN PUBLICACIONES CIENTÍFICAS

98%

INDICADOR D.2: BRECHA EN PROYECTOS SUBVENCIONADOS

88%

3. RECURSOS LINGUÍSTICOS Y DATOS ANOTADOS DISPONIBLES

Consideraremos como recursos en tecnologías de la lengua los siguientes: cantidad de texto disponible en Internet, modelos de lenguaje y datos anotados. Para medir la brecha en la disponibilidad de recursos y datos entre las dos lenguas, calculamos los tres indicadores siguientes:

- Indicador R.0: Texto disponible en Internet
- Indicador R.1: Modelos de lenguaje pre-entrenados
- Indicador R.2: Datos anotados

El **indicador R.0** mide la brecha entre el español y el inglés en cuanto a disponibilidad de texto en Internet para estas lenguas. Para su cálculo, se utiliza la siguiente información: número de artículos en Wikipedia, porcentaje de páginas en internet, número de textos en Internet Archive, número de textos en PubMed y porcentaje de páginas en el último crawl de Common Crawl.

INDICADOR R.0: TEXTO EN INTERNET

90%

El **indicador R.1** mide la brecha entre el español y el inglés en cuanto a disponibilidad de modelos de lenguaje pre-entrenados. Los últimos avances en tecnología de la lengua muestran sistemáticamente que disponer de modelos pre-entrenados se traduce en una mejora significativa en términos de efectividad en la gran mayoría de problemas. Para el cálculo del indicador se utiliza la información disponible en Hugging Face[1], y se computa como la diferencia entre el número de modelos entrenados en inglés y el número de modelos disponibles en español dividido por la suma de ambos.

INDICADOR R.1: MODELOS DE LENGUAJE

76%

El **indicador R.2** mide la brecha entre el español y el inglés en cuanto a disponibilidad de datos anotados. El éxito de los sistemas de PLN está, en muchos casos, condicionado por la disponibilidad de datos de entrenamiento. En esta primera anualidad, se considera la presencia de datos anotados en ambas lenguas en las principales campañas de evaluación y repositorios a nivel nacional, europeo e internacional. El indicador se calcula como el número de recursos encontrados para el inglés menos los encontrados para el castellano dividido por la suma de ambos.

[1] https://huggingface.co/

INDICADOR R.2: DATOS ANOTADOS

54%

4. EFECTIVIDAD

Los indicadores de efectividad evalúan el desempeño de las aplicaciones en tecnologías de la lengua en base a criterios extrínsecos, es decir, en base a la eficacia de estas en tareas específicas. Para calcular la brecha o diferencia en la efectividad de los sistemas de PLN entre el inglés y el español, se han realizado experimentos en laboratorio dentro del contexto del proyecto en el que se compararán, para ambos idiomas, sistemas base carentes de tecnología lingüística con modelos pre-entrenados y re-entrenados sobre datos anotados.

En concreto, se han realizado experimentos sobre once tareas, de las cuales siete son de clasificación (de múltiples tipos), tres de etiquetado (una de ellas de answer extraction) y una de regresión (similitud entre oraciones). La brecha es muy variable entre tareas, probablemente como reflejo de la dificultad relativa que tienen; en las tareas más sencillas, incluso las aproximaciones no lingüísticas obtienen resultados competitivos, y la incorporación de modelos de lenguaje no provoca diferencias apreciables entre idiomas. Esta puede ser la razón por la que, en las escasas referencias de la literatura que permiten una comparación directa, no se habían observado diferencias sustanciales entre ambos idiomas. En nuestra evaluación, sin embargo, incluimos tareas difíciles de acometer con aproximaciones no lingüísticas, y el gap promedio que hemos encontrado es del 18 %.

INDICADOR E.1: EFECTIVIDAD

18%

5. CONCLUSIONES

A partir de los indicadores anteriores, es posible calcular un indicador de la brecha en el Estado del Arte de nuestro idioma con respecto a la lengua inglesa. La agregación se ha realizado como una estimación del esfuerzo adicional necesario para construir una aplicación de Procesamiento del Lenguaje Natural (el campo de la Inteligencia Artificial que nos ocupa) a un nivel similar de eficiencia en español, respecto del inglés. Para ello hemos realizado un promedio de los siguientes indicadores de brecha: D2 (proyectos subvencionados), R.0 (corpora disponible en

Internet para desarrollar modelos de lenguaje preentrenados), R.1 (modelos de lenguaje disponibles), R.2 (datos anotados disponibles), y E.1 (efectividad de los modelos de lenguaje en tareas de PLN usando datos comparables para el fine-tuning). Nótese que para hacer este promedio hemos descartado el indicador de brecha de publicaciones científicas, ya que no podemos establecer un impacto directo en el coste de desarrollar aplicaciones eficaces y eficientes, dado que la algorítmica subyacente en el estado del arte es independiente de la lengua. Aun así, sigue siendo un indicador útil en sí mismo para estimar la intensidad del esfuerzo de investigación en uno y otro idioma.

La estimación de la brecha en este ámbito es, finalmente del 65 %. El factor más desfavorable es el de texto en Internet (90 %), seguido de proyectos subvencionados (88 %), modelos de lenguaje (76 %), datos anotados (54 %) y efectividad de los modelos (18 %).

BRECHA EN EL ESTADO DEL ARTE

65%