Proyecto Espacio de Observación de Inteligencia Artificial en Español. Ámbito 1. Estado del Arte. Brecha español/ingles en tecnologías de la lengua. Informe 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

1. Introducción

En este documento se presenta la medición del cálculo de la brecha entre español e inglés en cuanto al estado del arte de tecnologías de la lengua, en el marco del proyecto del Espacio de Observación de Inteligencia Artificial en Español, que se lleva a cabo mediante un convenio entre Red.es y la UNED. Dentro del *Ámbito 1: estado del arte*, cercano al mundo académico, estudiaremos para ambas lenguas: el grado de diseminación y proyectos subvencionados, los recursos existentes (corpus de texto, modelos de lenguaje, datos anotados, herramientas de procesamiento lingüístico) y la efectividad de los modelos. Este documento se corresponde con la segunda iteración del proyecto, realizada desde marzo de 2023 hasta marzo de 2024.

A continuación se presenta la descripción de la metodología aplicada para definir y calcular los indicadores en el Ámbito del Estado del Arte, así como los resultados obtenidos al aplicarlos. Se formalizan los indicadores y se define la metodología general aplicada para recopilar la información necesaria para estimar cada uno de los indicadores. Además, se presentan los resultados obtenidos durante el segundo año de proyecto y se comparan con los obtenidos en el primer año.

2. Definición de indicadores

En esta segunda iteración del proyecto se mantiene la misma definición de indicadores que en el primer año. Incluimos en esta sección la definición de estos indicadores con el fin de que el documento sea autocontenido.

2.1. Indicadores Ámbito 1: Estado del Arte

2.1.1. Indicadores de diseminación

Los indicadores de diseminación en ODESIA miden 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 tienen en cuenta los artículos publicados y los proyectos subvencionados. En cuanto a los artículos publicados (Indicador D.1), las publicaciones se buscan en los proceedings de congresos recientes de PLN de índole internacional. Para determinar si un artículo o proyecto está orientado al español, inglés o a ambos se determina si la experimentación se ha realizado sobre datos en español, inglés o en ambas lenguas. La identificación de estos artículos se realiza mediante una combinación de búsquedas por palabras clave y revisión manual (en el primer año, se realizó un estudio en profundidad de 50 artículos para comprobar el margen de error del proceso). En cuanto a proyectos subvencionados, se analizan proyectos a nivel europeo y de Estados Unidos. En la secciones correspondientes a resultados se proporcionan más detalles sobre el proceso de recopilación de datos en esta segunda iteración.

El indicador de publicaciones se calcula como la diferencia entre publicaciones que trabajan con textos en inglés y publicaciones que trabajan con textos en español dividido por la suma de ambos. La brecha

obtenida sería nula solo en el caso de que el número de publicaciones en inglés sea el mismo que el número de publicaciones que cubran además en español. La brecha sería negativa y del 100 % si todos los artículos cubrieran el español y ninguno el inglés. Sería positiva del 100 % si ningún artículo cubriera el español y todos ellos cubrieran el inglés.

Indicador D.1: Brecha en publicaciones

Representa la diferencia porcentual en cuanto a publicaciones en tecnologías de la lengua en castellano. Se obtendrá de la siguiente forma:

$$\text{D. } 1 = \frac{|P_I| - |P_E|}{|P_E \cup P_I|} \cdot 100$$

donde $|P_E|$ y $|P_I|$ representan el número de publicaciones de tecnologías que cubren el español y que cubren exclusivamente el inglés.

En relación al indicador D.2 (Brecha en proyectos subvencionados), se realizan búsquedas en bases de datos públicas de proyectos de investigación subvencionados. En las búsquedas se usan filtros disponibles para seleccionar proyectos de PLN. Posteriormente, se filtran los proyectos obtenidos en la primera búsqueda conforme al intervalo temporal considerado. A continuación 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 exclusivamente en inglés y proyectos que además cubran el español dividido por la suma de ambos. La brecha obtenida sería nula solo en el caso de que el número de proyectos exclusivamente en inglés sea el mismo que el número de proyectos que cubran además el español. La brecha sería negativa y del 100 % si todos los proyectos cubrieran además el español. Sería positiva del 100 % si ningún proyecto cubriera el español.

Indicador D.2: Brecha en proyectos subvencionados.

Representa la diferencia porcentual en cuanto a proyectos subvencionados en tecnologías de la lengua en castellano. Se obtendrá de la siguiente forma:

$$\text{D. } 2 = \frac{|P_I| - |P_E|}{|P_E \cup P_I|} \cdot 100$$

donde $|P_E|$ representa proyectos donde se cubre el español y $|P_I|$ representan el número de proyectos de tecnologías exclusivamente en inglés.

2.1.2. Indicadores de recursos

Consideraremos como *recursos* en tecnologías de la lengua los corpora de texto disponible en ambos idiomas, y los modelos de lenguaje que se usen en el desarrollo de tecnologías de PLN.

2.1.3. Indicador de texto disponible en internet

Los modelos de lenguaje generativos que se han desarrollado en los últimos años y que lideran los avances en la inteligencia artificial relacionada con el lenguaje están entrenados con grandes masas de texto disponibles en Internet. Este indicador medirá la brecha entre el español y el inglés en cuanto a a disponibilidad de texto en Internet para estas lenguas. Para calcularlo se emplearán la información para inglés y español:

- Número de artículos en Wikipedia.
- Porcentaje de páginas en Internet (tomando como referencia el número de páginas para las que se conoce la lengua).
- Número de textos en Internet Archive.
- Número de textos en PubMed.
- Porcentaje de páginas en el último crawl de Common Crawl.

Indicador R.0: Brecha en masa de texto disponible en internet

Representa la diferencia porcentual en cuanto a disponibilidad de corpus de texto en las respectivas lenguas. Sea F el conjunto de fuentes (wikipedia, etc.), y sea P_f el peso asignado a fuente:

$$\text{R. } 0 = \sum_{f \in F} P_f \frac{T_I^f - T_E^f}{T_I^f + T_E^f} \cdot 100$$

donde T_I^f y $\frac{f}{E}$ representan el volumen de texto en la fuente f en inglés y en español respectivamente.

2.1.4. Indicadores 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 ello, explotamos la información disponible en Hugging Face,[1] la biblioteca de código abierto más popular actualmente, que permite a los desarrolladores utilizar modelos pre-entrenados para tareas como clasificación de texto, traducción automática y generación de texto, entre otros. Los modelos se organizan en categorías basadas en el tipo de tarea de procesamiento que abordan, como la clasificación de texto, la generación de texto y la traducción automática. Estas tareas se corresponden de forma aproximada con la categorización de aplicaciones desde un punto de vista técnico descrita en la sección **??**. Además, Hugging Face permite filtrar por idioma, arquitectura de modelo y otros criterios relevantes.

Indicador R.1: Brecha en modelos de lenguaje

Representa la diferencia porcentual en cuanto a disponibilidad de modelos de lenguaje. Sea D el conjunto de tareas o dominios considerados, y sea P_d el peso asignado a cada tarea o dominio:

$$\text{R. } 1 = \sum_{d \in D} P_d \frac{|M_I| - |M_E|}{|M_I \cup M_E|} \cdot 100$$

donde M_I y M_E representan los conjuntos de modelos entrenados sobre texto en inglés y en español respectivamente. Los modelos multilingües seran considerados como pertenecientes a ambos conjuntos.

2.1.5. Indicadores de datos anotados

En muchos escenarios, el éxito de los sistemas está condicionado por la disponibilidad de datos de entrenamiento, especialmente en cierto tipo de tareas de minería de textos (ver Tabla **??**) en donde la información implícita que debe ser extraída de los textos depende del dominio de aplicación y del problema concreto. Algunos ejemplos son la clasificación temática de textos en dominios específicos, el análisis de sentimientos, o tareas de extracción de información. En estos casos no es suficiente pre-entrenar los modelos directamente sobre grandes colecciones de documentos, sino que es necesario disponer de datos anotados manualmente específicos para cada problema.

Podemos distinguir entre tres tipos de datos anotados. En primer lugar, aquellos que son anotados por expertos en un laboratorio. El número de datos anotados es limitado dado que la contratación de expertos es cara. En segundo lugar, estos costes pueden reducirse mediante el uso de servicios de anotación on-line de no expertos. La calidad de la anotación es menor, pero puede compensarse con el análisis de los datos, descartando anotaciones incoherentes o incluyendo pruebas de comprobación en los propios datos a anotar. Una tercera vía consiste en explorar información colaborativa, es decir, información anotada por los propios usuarios durante el uso de sistemas. Por ejemplo, los *logs* de navegación son una herramienta clave para entrenar motores de búsqueda. Es posible también emular datos anotados de análisis de sentimiento a partir de emoticonos y feedback de los usuarios.

Para la recopilación de datos se consideran campañas de evaluación competitivas y repositorios de recursos lingüísticos. Concretamente, se estudiará la presencia de datos anotados en ambas lenguas en las principales campañas de evaluación y repositorios a nivel nacional, europeo e internacional. Generalmente, las campañas de evaluación como CLEF, IBERLEF o SEMEVAL aglutinan una serie de tareas en las que

[1] https://huggingface.co/

grupos de investigación presentan sus aproximaciones a problemas específicos que son evaluadas sobre un conjunto de datos anotados común para cada tarea.

La brecha se computa como los recursos encontrados para el inglés menos los encontrados para el castellano dividido por la suma de ambos. La brecha será nula si los recursos en ambas lenguas son igualmente numerosos, obteniendo una brecha positiva o negativa del 100 % si solo hubiera recursos en uno de los idiomas. Los criterios de ponderación se especificarán en el informe de resultados de cada iteración.

Indicador R.2.a: Presencia de datos anotados en repositorios

Representa la diferencia porcentual en cuanto a disponibilidad de corpus anotado para entrenamiento y test en foros internacionales (campañas de evaluación competitiva y repositorios). Sea R el conjunto de repositorios y P_r el peso asignado a cada repositorio:

$$\text{R. 2. a} = \sum_{r \in \mathcal{R}} P_r \frac{|D_I| - |D_E|}{|D_I \bigcup D_E|} \cdot 100$$

donde D_I y D_E representan los conjuntos de datos anotados en inglés y castellano respectivamente en repositorios públicos.

Indicador R.2.b: Presencia de datos anotados en campañas de evaluación

Representa la diferencia porcentual en cuanto a disponibilidad de corpus anotado para entrenamiento y test campañas de evaluación competitiva. Sea $\mathcal{C}$ el conjunto de campañas de evaluación y P_c el peso asignado a cada campaña:

$$\text{R. 2. b} = \sum_{c \in \mathcal{C}} P_c \frac{|T_I| - |T_E|}{|T_I \bigcup T_E|} \cdot 100$$

donde T_I y T_E representan el conjunto de tareas competitivas en las campañas con datos anotados en inglés y castellano respectivamente.

El tercer indicador de recursos anotados tiene como objetivo obtener una visión general de aquellos dominios y tareas en los que se dispone de datos anotados para el español, identificando a su vez escenarios en los que existe aún un vacío. Además, se considerarán las tareas de procesamiento de lenguaje (lematización, análisis de dependencias, reconocimiento de entidades nombradas etc.) como un dominio adicional.

Indicador R.2.c: Brecha en diversidad de datos anotados

Representa la diferencia porcentual en cuanto a disponibilidad de corpus anotado para entrenamiento y test en diferentes aplicaciones. Sea $\mathcal{D}$ el conjunto de dominios, P_d el peso asignado a cada dominio, y $\mathcal{H}_d$ el conjunto de familias de aplicaciones identificadas para dicho dominio, se calcula:

$$\text{R. 2. c} = \sum_{d \in \mathcal{D}} P_d \sum_{h \in \mathcal{H}_d} \frac{C_I^h - C_E^h}{C_{I \cup E}^h}$$

donde C_I^h y C_E^h toman el valor uno o cero dependiendo de si existen datos anotados para la familia de aplicaciones h en el idioma correspondiente. $C_{I \cup E}^h$ toma valor 1 si existen datos anotados en cualquiera de los dos idiomas. Solo se considerarán familias de aplicaciones con datos en al menos uno de los dos idiomas.

2.1.6. Indicadores de efectividad

Comúnmente, las aplicaciones en tecnologías de la lengua se evalúan en base a criterios extrínsecos, es decir, en base a la eficacia del sistema en tareas específicas. Estas metodologías están relacionadas con la usabilidad y captan la capacidad de aprender y generalizar en el contexto de problemas específicos. Estos marcos de evaluación pretenden capturar escenarios de uso como la traducción automática, la respuesta a preguntas, el resumen automático, la minería de opiniones, etc. Idealmente, los datos de prueba deberían ser una muestra representativa del escenario de uso del sistema. La fortaleza de estos marcos de evaluación radica en que proporcionan información directa sobre la usabilidad del sistema. El punto débil es el efecto del sobre-ajuste. No siempre es fácil recoger los datos de entrada-salida en un escenario real. Por ejemplo, no hay demasiados datos de entrenamiento disponibles para las lenguas poco comunes en la traducción automática o para los temas dominios no populares en sistemas de diálogo.

Definir un indicador de brecha en efectividad entre lenguas requiere tener en cuenta muchas variables, que en muchos casos dependen también del problema y de los datos disponibles para la evaluación. El primer problema a resolver es que no todas las métricas de efectividad tienen las mismas propiedades de

escala. La mayoría de las métricas, como por ejemplo la tasa de aciertos, están acotadas entre cero y uno. Otras métricas no tienen una cota superior. Por tanto, no se pueden equiparar las diferencias obtenidas para varios problemas en los que se emplean diferentes métricas. Es decir, un intervalo en una métrica puede tener una relevancia completamente distinta del mismo intervalo en otra métrica. Por tanto, es necesario establecer un intervalo-unidad o intervalo de referencia.

El segundo gran problema es que la efectividad obtenida puede ser sensible a la dificultad intrínseca de los datos de evaluación. Por ejemplo, sobre un tamaño de datos de entrenamiento menor, es normal obtener valores de efectividad menores. Esto no significa que exista una brecha intrínseca en cuanto a efectividad de los sistemas en cada uno de los idiomas. Para controlar este aspecto, será necesario tomar como referencia un sistema base que cumpla ciertas características. El sistema base no debe incluir ningún tipo de tecnología de la lengua específica de idioma. Es decir, debe de ser un sistema no pre-entrenado para ningún idioma, y que además no emplee herramientas de procesamiento lingüístico. Por ejemplo, emplear un clasificador SVM sobre conjuntos de *tokens* para clasificar textos no supone pre-entrenamiento ni pre-procesamiento lingüístico. Por tanto, las diferencias de efectividad entre idiomas vendrán determinadas únicamente por la dificultad del conjunto de datos de evaluación.

Teniendo en cuenta estos dos factores, tomaremos como intervalo de referencia en cada idioma la distancia entre la efectividad del sistema base que denotaremos como b, y un punto de referencia en la escala de la métrica que denotaremos como r. Es decir, tomaremos como intervalo unidad $|b - r|$.

El punto de referencia r también requiere un análisis previo. En casos en los que la métrica no esté acotada superiormente o en casos en los que la efectividad de los sistemas sea muy baja, este punto debería ser la cota inferior. Por otro lado, en casos en los que exista una cota superior y la efectividad sea alta debería tomarse como punto de referencia el valor superior en la escala de la métrica. Por ejemplo, dado un sistema base con efectividad cercana al uno en una métrica acotada superiormente, por ejemplo, 82 % de acierto, y tomando como punto de referencia el 100 % de acierto, el intervalo unidad debería ser del 18 %.

Una vez definido este intervalo unidad $|b-r|$ la aportación lingüística efectiva en un idioma se calculará como el ratio de la diferencia entre efectividad del sistema evaluado y el sistema base respecto al intervalo unidad:

$$\Delta = \frac{s - b}{|b - r|} \cdot 100$$

La aportación lingüística en cada idioma cumple las siguientes propiedades. En primer lugar, la aportación es nula cuando el mejor sistema se comportan igual que el sistema base ausente de tecnología lingüística:

$$s = b \Longrightarrow \Delta = 0$$

En segundo lugar, dada una efectividad fija por parte del sistema base, la aportación es proporcional a la diferencia entre efectividad del sistema evaluado y la del sistema base:

$$b = k \Longrightarrow \Delta \propto s - b$$

En tercer lugar, dado una diferencia fija entre el sistema y el sistema base, la contribución será proporcional a la inversa del intervalo unidad:

$$s - b = k \Longrightarrow \Delta \propto \frac{1}{|b - r|}$$

Esto quiere decir que, tomando como referencia la máxima puntuación ($r = 1$) a medida que la efectividad del sistema y del sistema base se aproximen al punto máximo, la contribución será mayor. Por ejemplo, una mejora de 0.97 a 0.98 es más importante que una mejora de 0.67 a 0.68. De la misma forma, a valores bajos de efectividad, tomando como punto de referencia ($r = 1$), una mejora de 0.1 a 0.2 será más significativa que una mejora de 0.3 a 0.4.

El indicador de la brecha de efectividad entre idiomas inglés (I) y español (E) se calculará mediante el indicador:

$$Ind(I, E) = \Delta_I - \Delta_E = \frac{s_I - b_I}{|b_I - r|} - \frac{s_E - b_E}{|b_E - r|}$$

Este indicador cumple las siguientes propiedades. En primer lugar, es simétrico respecto a los idiomas:

$$Ind(I, E) = -Ind(E, I)$$

En segundo lugar, un comportamiento idéntico en ambos idiomas se corresponde con una brecha cero:

$$Ind(I, I) = Ind(E, E) = 0$$

Tomando como punto de referencia $r = 0$, se obtiene una brecha nula cuando ambos presentan la misma diferencia porcentual respecto al baseline.

$$\left.\begin{array}{l} r = 0 \\ \frac{s_I - b_I}{b_I} = \frac{s_E - b_E}{b_E} \end{array}\right\} \Rightarrow Ind(I, E) = 0$$

que es lo mismo que decir que ambos tienen la misma proporción de efectividad respecto a sus sistemas base.

$$\left.\begin{array}{l} r = 0 \\ \frac{s_I}{b_I} = \frac{s}{b_E} \end{array}\right\} \Rightarrow Ind(I, E) = 0$$

Tomando como punto de referencia $r = 0$, se obtiene una brecha del 100 % cuando la efectividad del sistema en ingles consigue la diferencia porcentual conseguida por el sistema en español (brecha cero) más la efectividad del sistema base en inglés.

$$\left.\begin{array}{l} r = 0 \\ s_I = b_I \cdot \frac{s_E}{b_E} + b_I \end{array}\right\} \Rightarrow Ind(I, E) = 100\,\%$$

Tomando como referencia $r = 1$ (cota máxima), la brecha es nula cuando la efectividad del sistema en ambas lenguas es proporcional la diferencia entre los sistemas base y la cota superior:

$$\left.\begin{array}{l} r = 0 \\ \frac{s_I}{1 - b_I} = \frac{s}{1 - b_E} \end{array}\right\} \Rightarrow Ind(I, E) = 0$$

Tomando como referencia $r = 1$, hay una brecha del 100 % cuando el sistema en español no supera al sistema base, mientras que el sistema en inglés obtiene la máxima puntuación.

$$\left.\begin{array}{l} r = 1 \\ s_I = 1 \\ s_E = b_E \end{array}\right\} \Rightarrow Ind(I, E) = 100\,\%$$

Para este indicador se emplearán dos fuentes de datos. Por un lado, experimentos en laboratorio dentro del contexto del proyecto en el que se compararán para ambos idiomas sistemas base carentes de tecnología linguística con modelos pre-entrenados y re-entrenados sobre datos anotados. Estas fuentes tienen la ventaja de ser experimentación controlada orientada a la brecha lingüística. La limitación es que no será posible cubrir muchas de las familias de aplicaciones. Estos datos se complementarán con el análisis de contribuciones en la literatura donde se aporten resultados de experimentos en distintas lenguas sobre sistemas base y datos anotados comparables.

Indicador E.1: Brecha en efectividad

Representa la diferencia entre idiomas entre mejoras porcentuales sobre un sistema base no lingüístico. Sea $\mathcal{D}$ el conjunto de dominios, P_d el peso asignado a cada dominio, y $\mathcal{H}_d$ el conjunto de aplicaciones seleccionadas para dicho dominio, se calcula:

$$\text{E. } 1 = \sum_{d \in \mathcal{D}} P_d \sum_{h \in \mathcal{H}_d} \frac{s_I^h - b_I^h}{|b_I^h - r^h|} - \frac{s_E^h - b_E^h}{|b_E^h - r_0^h|}$$

donde s_I^h, b_I^h y r^h representan la efectividad de la herramienta h, del sistema base y el punto de referencia en inglés. La notación para el español es análoga.

3. Cálculo de indicadores: Ámbito 1 - Estado del arte

En esta sección se presentan los resultados obtenidos dentro del Ámbito 1, Estado del Arte.

3.1. Cálculo de indicadores de diseminación

3.1.1. D.1: Brecha en publicaciones científicas [D.1: 98 %]

A continuación se detalla la metodología adoptada para obtener datos para el cálculo del indicador D.1 (brecha en publicaciones) correspondiente al segundo año.

- **Fuente de datos:** Se han considerado las publicaciones en los dos congresos de mayor prestigio internacional en las áreas de PLN (ACL, Association for Computational Linguistics)[2] y de la recuperación de información (SIGIR, Special Interest Group on Information Retrieval),[3] respectivamente. Ambas conferencias están indexadas en la base de datos SCIE, en la categoría 1, y en la base de datos CORE, en la categoría A*. Además, se han considerado las conferencias de índole internacional CONLL (Computational Natural Language Learning)[4], EACL (European Chapter of the Association for Computational Linguistics)[5], y NAACL (North American Chapter of the Association for Computational Linguistics). Además de las conferencias anteriores ya monitorizadas en el año anterior, se ha incorporado en esta iteración el congreso EMNLP, que el segundo congreso internacional de mayor importancia tras el ACL dentro del área de procesamiento de lenguaje natural. No hemos añadido en este indicador las publicaciones asociadas a las campañas de evaluación en donde se comparten datos de entrenamiento y test (CLEF, IBERLEF, SEMEVAL, etc.), pues estos datos se tendrán en cuenta en el indicador de brecha en datos anotados de entrenamiento.

- **Intervalo temporal:** En este informe, se ha considerado para el cálculo del indicador las publicaciones referidas a los últimos 5 años, de 2019 a 2023. Respecto al informe del año anterior, desplazamos un año la ventada temporal (2018-2022) con el fin de estudiar la evolución de los indicadores bajo las mismas condiciones.

- **Extracción de datos:** Se ha realizado una búsqueda semi automática de la palabra "Spanish" en el título y resumen (*abstract*) de los artículos publicados, para localizar aquellos que experimentaron sobre datos en español. En el caso de las campañas de evaluación se ha considerado el artículo principal publicado por los organizadores. Se han revisado manualmente los resultados obtenidos para validarlos y se ha comprobado si se trata de trabajos multi-lingües inglés-español. Para obtener las publicaciones que han experimentado sobre datos en inglés, se ha considerado como tales todos aquellos artículos que no mencionan el nombre de los idiomas más representados en la conferencia ("Spanish", "French", "German","Italian", "Chinese").

- **Fecha de la búsqueda**: 31 de diciembre de 2023.

- **Cálculo del indicador:** En este caso se ha calculado el Indicador D.1 sobre el total de publicaciones en cada congreso a lo largo de los años para cada uno de los idiomas.

La Tabla 1 muestra los resultados para congresos internacionales y las cifras en las que se ha basado el cómputo.

En base a estos datos, en el congreso SIGIR de recuperación de información hay una brecha del 100 %, en el congreso ACL de PLN una brecha del 99 % y en el congreso CoNLL centrado en lingüística computacional, una brecha del 99 %, siendo sensiblemente mayor que el calculado en el informe anterior. Al igual que en el informe del año anterior, la predominancia del inglés respecto a otros idiomas parece ser un fenómeno bastante general.

[2] https://aclanthology.org/venues/acl/
[3] https://sigir.org/
[4] https://www.conll.org/
[5] https://eacl.org/

La brecha entre el español y el inglés desciende sensiblemente en el caso de los congresos internacionales de índole europeo y americano como EACL o NAACL (97 % en ambos casos).

Con respecto al informe del año anterior, se se ha ampliado el espacio de búsqueda a nuevos congresos, en concreto los congresos EMNLP o NAACL, obteniéndose resultados similares a los del congreso homólogo ACL.

Como criterio de agregación consideramos la suma de publicaciones en todos los congresos (en los dos últimos años). La última parte de la tabla muestra los resultados agregados. **En global, el español presenta una brecha similar a otras lenguas europeas como el francés o el alemán, en torno al 98 %.** La brecha desciende en el caso del chino (93 %).

De nuevo, no se observa una evolución de reducción de brecha a lo largo de los años. De hecho, la brecha global aumenta sensiblemente, por lo que no parece que existe ninguna tendencia beneficiosa para el español en este sentido. Por otro lado, se planteó en el informe del año anterior desgranar los indicadores en foros internacionales y nacionales además de un análisis por áreas de aplicaciones. Sin embargo, la disponibilidad de muestras para el español no ha permitido realizar dicho desgranamiento de indicadores.

3.1.2. D.2: Brecha en proyectos subvencionados [D.2: 96 %]

Al igual que en el año anterior, para la elaboración del indicador de brecha en proyectos subvencionados, hemos accedido a dos fuentes de datos que cubren proyectos europeos y estadounidenses respectivamente. En el ámbito europeo, accedemos a CORDIS (Community Research and Development Information Service[6]), portal de la Comisión Europea que recoge los proyectos financiados por los programas marco de investigación e innovación de la Unión Europea. En el ámbito estadounidense accedemos al portal de NSF (National Science Foundation)[7], una agencia independiente federal fundada en 1950 para la promoción de la ciencia en E.E.U.U. La consulta se ha realizado el 31 de diciembre de 2023.

En ambos casos se han considerado proyectos iniciados entre el 1 de enero de 2019 hasta el 31 de diciembre de 2023. En el caso de CORDIS, se ha realizado la siguiente búsqueda "Natural Language Processing.ªnd "Spanish", para recuperar posibles proyectos sobre procesamiento del lenguaje en nuestro idioma, y la búsqueda "Natural Language Processing"para recuperar posibles proyectos sobre procesamiento del lenguaje en inglés. Al resultado de la búsqueda en inglés, se han restado aquellos que proyectos en los que aparece mencionado alguno de los siguiente idiomas: "German", "French", "Spanish.ªnd Çhinese". Como resultado, en CORDIS se han encontrado 9 proyectos con experimentación en español frente a 226 proyectos con experimentación en inglés. Esto supone una brecha del 93 % según el indicador D.2.

En el caso de NSF, la búsqueda se ha realizado considerando los projectos asignados al área textit-Division of Information & Intelligent Systems en el caso de NSF. Se han encontrado 0 proyectos con experimentación en español frente a 126 proyectos con experimentación en inglés. lo que supone una brecha del 100 %. **En promedio, podemos considerar una brecha del 96 % para el indicador D.2.**. De nuevo, respecto al año anterior no hay un decrecimiento de la brecha sino más bien un aumento sensible.

Siguiendo las sugerencias en el informe anterior se ha estudiado las temáticas de investigación en español frente a inglés. No se ha encontrado ningún patrón característico. Se ha identificado un único proyecto orientado excusivamente al español denominado *Making contracts more human with state-of-the-art natural language processing techniques in Spanish.*

[6] https://cordis.europa.eu/search/es
[7] https://www.nsf.gov/awardsearch/advancedSearch.jsp

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE TRANSFORMACIÓN DIGITAL
red.es
Plan de Recuperación, Transformación y Resiliencia
UNED

Tabla 1: Número de publicaciones en conferencias internacionales, por lengua y año, que describen experimentación para las lenguas consideradas, desde 2019 a 2023.

	2019	2020	2021	2022	2023	Brecha
			ACL			
Spanish	4	3	0	8	8	99 %
French	3	2	0	2	11	99 %
German	16	5	6	5	30	98 %
Chinese	19	18	36	25	36	93 %
English	618	751	668	662	825	
			SIGIR			
Spanish	0	0	0	0	0	100 %
French	0	0	0	0	0	100 %
German	0	0	0	0	0	100 %
Chinese	0	10	2	6	0	98 %
English	224	330	380	334	165	
			CONLL			
Spanish	0	2	0	-	0	99 %
German	6	1	3	-	0	94 %
French	1	1	2	-	0	97 %
Chinese	3	0	3	-	1	97 %
English	51	88	50	45	-	74
			EACL			
Spanish	8	-	-	-	4	96 %
German	19	-	-	-	4	92 %
French	8	-	-	-	6	95 %
Chinese	7	-	-	-	3	96 %
English	285	-	-	-	266	
			NAACL			
Spanish	15	-	21	3	-	94 %
German	19	-	17	6	-	93 %
French	12	-	10	4	-	97 %
Chinese	30	-	16	13	-	90 %
English	348	-	414	419	-	
			EMNLP			
Spanish	9	8	7	2	9	98 %
German	24	19	21	3	10	96 %
French	10	11	14	2	10	98 %
Chinese	33	26	27	27	30	93 %
English	606	688	779	793	989	
			Aggregated			
Spanish	43	13	29	16	21	0.98 %
French	82	22	47	17	25	0.97 %
German	52	17	36	15	46	0.97 %
Chinese	112	54	96	84	70	0.93 %
English	2513	1857	2746	2670	2245	

3.2. Cálculo de indicadores de recursos

3.2.1. R.0: Indicador de texto disponible en internet [83 %]

La información que se ha usado para calcular este indicador en esta segunda iteración se muestra en la Tabla 2 y se corresponde con as fuentes empleadas en el año anterior. Se ha recabado información sobre las siguientes colecciones de textos: Wikipedia, Internet, Internet Archive, PubMed y CommonCrawl. La consulta de información se realizó el 19 de febrero de 2024. Algunas fuentes proporcionan información en números absolutos y otras en términos de porcentajes.

Tabla 2: Datos sobre volumen de texto disponibles en Internet en inglés y en español.

Colección	Medida	Búsqueda Esp.	Esp.	Búsqueda Ing.	Ing.	Fuente	R.0
Wikipedia	# artículos	https://es.wikipedia.org/wiki/Wikipedia_en_espa%C3%B1ol	1932427	https://es.wikipedia.org/wiki/Wikipedia_en_espa%C3%B1ol	6785774		56 %
Internet	% páginas		5.6		51.20	https://w3techs.com/technologies/history_overview/content_language	80 %
Internet Archive	# textos	https://archive.org/search?query=Language%3A%28spanish%29	49017	https://archive.org/search?query=Language%3A%28english%29	8234197	https://archive.org/	99 %
PubMed	# textos	https://pubmed.ncbi.nlm.nih.gov/?term=Spanish%5BLanguage%5D&sort=	385212	https://pubmed.ncbi.nlm.nih.gov/?term=English%5BLanguage%5D	32013562	https://pubmed.ncbi.nlm.nih.gov/	98 %
Common Crawl: CC-MAIN-2023-50	% páginas		4.5391		44.4285	https://commoncrawl.github.io/cc-crawl-statistics/plots/languages	81 %
promedio							**83 %**

Para el cálculo del indicador, se asigna el mismo peso a todas las fuentes. De este modo, **obtenemos una brecha promedio R.0 del 83 %**. Ésta es inferior a la obtenida el año anterior (84 %).

3.2.2. R.1: Indicador de modelos pre-entrenados [76 %]

En esta iteración se ha considerado de nuevo el repositorio Hugging Face[8], que contiene gran cantidad de modelos de lenguaje pre-entranados. La interfaz de búsqueda permite filtrar modelos pre-entrenados por lenguas y por categorías de tareas de PLN. La Tabla 3 muestra los resultados obtenidos, según la búsqueda por lenguas y por tareas, realizada el 19 de febrero de 2024. Los modelos bilingües (inglés/español) no están incluidos en el recuento de modelos por idioma individual. Por ello, las variables del indicador se han fijado de la siguiente forma: $|M_I|$ y $|M_E|$ se han calculado como la primera o la segunda columna más la tercera respectivamente para cada idioma. La variable $|M_I \bigcup M_E|$ se ha estimado como la suma de las dos primeras columnas mas la tercera.

Como muestran los resultados, teniendo en cuenta el total de modelos se obtiene una brecha del 84 %, frente a un 83 % en el año anterior. Si se promedia las brechas obtenidas por tarea de PLN, se obtiene un 76 %, frente a un 75 % obtenido el año anterior. Sin embargo, como muestra la figura 1, las diferencias a escala de tarea son muy variables, aumentando o disminuyendo la brecha dependiendo de la tarea. Al igual que el año anterior, la menor brecha se obtiene en el caso de la traducción automática, que siempre implica varias lenguas (38 %). Si descartamos la traducción automática, obtenemos un indicador promedio por tareas de 79 %. Es decir, asignamos el mismo peso P_d a todas las categorías de tareas en Hugging Face. **Por consistencia con el informe del año anterior, como valor final del indicador R.1 consideramos el promedio por tareas, obteniendo una brecha del 76 %.**

[8] https://huggingface.co/

Tabla 3: R.1: Resultados del indicador de modelos pre-entrenados basado en datos de Hugging Face.

	Inglés	Español	Ing./Esp.	R.1
Total	34142	2495	1227	84 %
Por categorías de PLN				
Generación de textos	13785	465	368	91 %
Clasificación de textos	3506	266	103	84 %
Generación de textos Text2Text	2087	190	83	80 %
Traducción automática	925	370	150	38 %
Enmascaramiento de palabras	730	108	53	70 %
Clasificación de tokens	983	216	52	61 %
Búsqueda de respuestas	569	44	8	85 %
Resumen automático	448	36	16	82 %
Similitud entre frases	356	56	40	66 %
Sistemas conversacionales	64	0	0	100 %
Clasificación Zero-shot	107	27	18	53 %
Búsqueda de respuestas sobre tablas	49	1	0	96 %
Promedio por tareas				**76 %**

3.2.3. R.2:Indicadores de datos anotados [R.2: 55 %; R.2.a: 81 %; R.2.b: 29 %]

En esta segunda iteración se han estimado los indicadores R.2.a y R.2.b siguiendo lo mismos criterios que en el primer año de proyecto. A continuación se detalla la metodología adoptada:

- **Selección de repositorios:** Se han considerado los repositorios LRE Map,[9] repositorio de recursos lingüísticos promovido y mantenido por la European Language Resources Association (ELRA), y el catálogo LDC,[10] Linguistic Data Consortium, organización con sede en la Universidad de Pennsilvania, que crea y distribuye recursos lingüísticos. Además, se ha considerado el repositorio Hugging Face, cuya interfaz de búsqueda permite filtrar conjuntos de datos anotados por tarea e idioma. Los tres repositorios son repositorios de referencia en el área del PLN y el aprendizaje automático.

- **Selección de campañas de evaluación:** Se han considerado los datasets generados en las campañas de evaluación más importantes en tareas de PLN, IberEVAL y SEMEVAL, que son de ámbito nacional e internacional respectivamente. Además se han considerado los datasets generados en los foros de evaluación de sistemas de acceso a la información CLEF y TREC, de ámbito europeo e internacional respectivamente.

- **Intervalo temporal:** Se han considerado para el cálculo del indicador los recursos liberados o publicados durante los últimos 5 años, de 2019 a 2023.

- **Fecha de búsqueda**: 31 de diciembre de 2023.

- **Extracción de datos:** Se ha realizado de forma manual, mediante búsquedas en los repositorios indicados y en los *proceedings* y *working notes* de las campañas de evaluación. Se han clasificado los distintos conjuntos de datos según su idioma (inglés y/o español) y dominio.

[9] https://lremap.elra.info/
[10] https://catalog.ldc.upenn.edu/

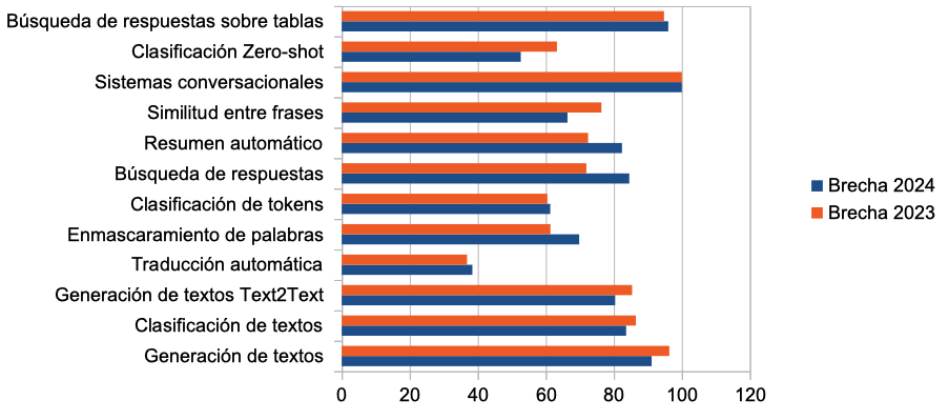

Figura 1: Comparativa de brechas en cantidad de modelos entrenados respecto al informe anterior.

En cuanto al indicador R.2.a (disponibilidad de corpus anotado para entrenamiento y test en foros internacionales), las búsquedas por idioma en el repositorio ELRA se han realizado sobre la categoría de recursos *"Dataset & evaluation data"* e introduciendo los filtros *Language="Spanish"* y *Language="English"* respectivamente. Sin considerar fecha, encontramos 213 recursos en español frente a 1773 recursos en inglés, lo que supone una brecha del 79 %. Asumimos que los recursos son de idioma único, por lo que el denominador del indicador es la suma de ambos números.

En cuanto al catálogo LCD, entre 2019 y 2023, encontramos 10, 19, 12, 7 y 7 datasets en inglés, frente a 3, 2, 4, 0 y 1 datasets en español. Esto supone una brecha promedio del 74 %. Si no se tiene en cuenta los años, encontramos 40 recursos en español frente a 410 en inglés. Esto supone una brecha del 82 %. Por coherencia con el repositorio ELRA y para evitar el efecto de la variabilidad entre años, tomamos este indicador.

En cuanto a Hugging Face, considerando el total de conjuntos de datos anotados, encontramos 5,879 en inglés frente a 555 en español, lo que supone una brecha del 83 %, que al igual que se mantiene cuando promediamos por tareas. Consideramos esté último indicador por el mismo motivo ya descrito en la sección anterior. Aplicamos el mismo peso P_r a todos los repositorios. **Calculando el promedio sobre los tres repositorios, obtenemos una brecha R.2.a del 81.17 %**. Esta brecha supera sustancialmente a la brecha obtenida en la iteración en el año anterior. Esto se debe a una explosión de recursos anotados en inglés. Como muestra la figura 2, la brecha en recursos anotados en Hugging face asciende considerablemente en todas las tareas.

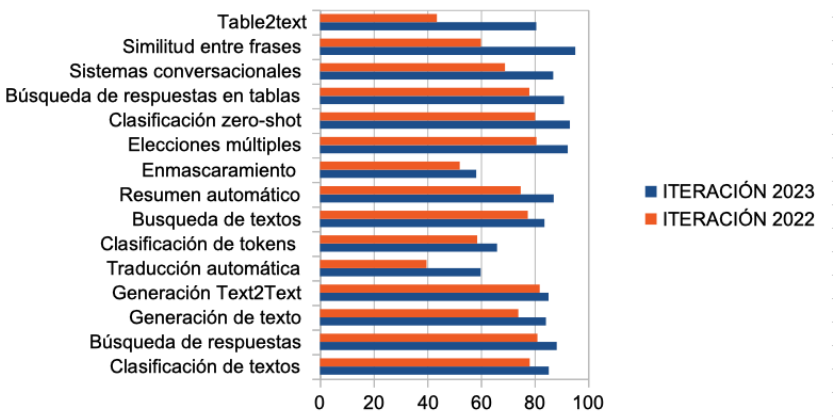

Figura 2: Comparativa de brechas en cantidad de recursos anotados en Hugging Face respecto al informe anterior.

Tabla 4: R.2.a: Indicador de conjuntos de datos anotados procedentes de repositorios.

	Inglés	Español	R.2.a
LRE			
Sin fecha	1773	213	**79 %**
LCD			
Sin fecha	410	40	**82 %**
2019	10	3	54 %
2020	19	2	81 %
2021	12	4	50 %
2022	7	0	100 %
2023	12	1	85 %
Promedio por años			74 %
Hugging Face			
Sin fecha	5,879	555	83 %
By NLP Task:			
Clasificación de textos	1310	104	85 %
Búsqueda de respuestas	802	50	88 %
Generación de texto	1157	99	84 %
Generación Text2Text	363	29	85 %
Traducción automática	343	86	60 %
Clasificación de tokens	259	53	66 %
Busqueda de textos	169	15	84 %
Resumen automático	480	33	87 %
Enmascaramiento	163	43	58 %
Elecciones múltiples	126	5	92 %
Clasificación zero-shot	141	5	93 %
Búsqueda de respuestas en tablas	127	6	91 %
Sistemas conversacionales	358	25	87 %
Similitud entre frases	81	2	95 %
Table2text	28	3	81 %
Promedio por tareas			**82 %**
Promedio por repositorios			**81 %**

En cuanto a recursos anotados en campañas de evaluación (indicador R.2.b), como muestra la Tabla 5, la brecha entre español e inglés en el caso de las campañas de evaluación en SEMEVAL, orientadas a PLN y de índole internacional, es del 78.95 %. En el caso de las campañas CLEF, orientadas a recuperación de información y de índole europeo, la brecha desciende hasta el 71.43 %. Al igual que en la iteración anterior, en ambos casos la brecha parece ser inferior o semjante a la de otros idiomas como el francés, alemán o chino. Por supuesto, en IBERLEF, conferencia que aglutina campañas de evaluación en el ámbito ibero-americano, la brecha se invierte, con 49 campañas sobre datos en español frente a 6 con datos en inglés. La brecha promedio para el español es del 24.06 %, promediando por campañas, es decir, asignando un p_c fijo para todas las campañas. Sin embargo, con el fin de tener en cuenta el tamaño de éstas (número de tareas), tomaremos como peso P_c de cada campaña su número de tareas, siendo esto equivalente a aplicar el indicador sobre la suma de las tareas en todas las campañas (internacional, europeo y nacional), **se obtiene una brecha del 28.57 % para el indicador R.2.b.**, ligeramente inferior a la brecha del 33 % obtenida en la iteración anterior. Esto se debe a una reducción sensible de la brecha en campañas

Tabla 5: Datos anotados en diferentes idiomas en las campañas de evaluación SEMEVAL, CLEF e IBERLEF.

	2019	2020	2021	2022	2023	R.2.c
	SEMEVAL-Workshops					
Español	1	1	0	1	3	78.95
Aleman	1	1	0	1	2	82.14
Francés	1	0	1	2	3	75.86
Chino	0	0	1	1	2	85.45
Inglés	8	10	9	12	12	
	CLEF					
Español	0	2	3	2	3	71.43
Francés	0	1	0	1	1	90.48
Aleman	1	1	1	2	0	84.62
Chino	0	0	0	0	0	100
Inglés	9	12	12	14	13	
	IBERLEF					
Español	8	7	11	9	14	-78.18
Inglés	0	0	1	1	4	
	Total					
Español	9	10	14	12	20	28.57
Inglés	17	22	22	27	29	
Brecha promedio inglés/español por campañas						24.06 %

de índole internacional (SEMEVAL) y europea (CLEF).

El promedio de R.2.a y R.2.b nos proporciona una **brecha en datos anotados R.2 del 55 %**, muy similar a la del año anterior (54 %), ya que el aumento de brecha en datos anotados en repositorios se compensa con la reducción de brecha en datos de campañas de evaluación.

3.3. Cálculo de indicadores de efectividad basados en experimentos [E1: 20±06 %]

3.3.1. Criterios de selección de datasets y tareas

- En cada dataset, los datos en inglés y en español se deben haber recolectado y anotado siguiendo la misma metodología, y el volumen de datos debe ser comparable entre ambos idiomas. No consideramos que cumplan este requisito los datasets que son traducción el uno del otro, ya sea automática o manual. En ambos casos pueden producirse sesgos en la evaluación (por supuesto, el supuesto de traducción automática introduce sesgos más acusados).

- En cada dataset, el subconjunto de test (sobre el que se evalúan los sistemas) no debe haber sido distribuido públicamente. De esta manera se evita el sobreajuste de datos y se minimiza la posibilidad de contaminación en los modelos (es decir, que el modelo haya visto las anotaciones manuales en la fase de preentrenamiento). La contaminación provoca una sobreestimación del rendimiento de un modelo contaminado con respecto a sus homólogos no contaminados. Desde un punto de vista científico las consecuencias son muy perjudiciales, ya que se publican conclusiones científicas erróneas.

- Para los datasets existentes, debe ser posible crear un subconjunto adicional de test privado utilizando la misma metodología con la que se construyó el dataset original. En este sentido, es preferible

adaptar datasets en los que podamos contar con el equipo original que los desarrolló y los anotó.

Otros aspectos que hemos considerado son los siguientes:

- La selección de tareas debe estar más orientada hacia las aplicaciones de la tecnología que hacia evaluaciones puramente lingüísticas. Las tareas más lingüísticas son adecuadas para evaluar el grado de conocimiento que tienen los modelos sobre la lengua, pero tienen una relación menos directa con el comportamiento de las aplicaciones de IA usadas por ciudadanos y organizaciones. Nuestro interés es realizar mediciones que tengan algún tipo de correspondencia con el uso práctico de estas tecnologías, así que queremos priorizar las tareas más cercanas al mercado de aplicaciones.

- Las tareas deben tener un grado de dificultad realista. Por un lado, hay datasets que pueden resolverse con un grado muy alto de acierto, pero no porque las tareas sean realmente sencillas, sino porque los sistemas aprenden los sesgos del dataset. En esos casos, el comportamiento de los sistemas fuera del laboratorio (es decir, sobre conjuntos de datos que no tienen los sesgos del conjunto de entrenamiento) es mucho peor que en las evaluaciones de laboratorio. En el otro extremo se encuentran los "diagnostic datasets", en los que se escogen ejemplos para el conjunto de test que requieran un conocimiento profundo del lenguaje para poder resolverse. Este tipo de datasets son muy útiles para identificar los puntos débiles de los modelos de lenguaje y para mejorarlos, pero son más difíciles que las situaciones promedio que nos encontramos en entornos de aplicación, de forma que tampoco ofrecen una imagen realista del rendimiento de las aplicaciones de Procesamiento del Lenguaje Natural. Nos interesa utilizar datasets en los que el rendimiento de los modelos de lenguaje se acerque al que encontraríamos en entornos reales de aplicación.

- Dificultad similar entre idiomas. Hemos establecido mecanismos para calibrar el leaderboard en función de la dificultad relativa intrínseca de los datasets en inglés y español, de manera que si para una tarea determinada hay una dificultad intrínseca (independiente del conocimiento del lenguaje que tengan los sistemas) diferente entre los dos idiomas, somos capaces de medirla y recalibrar la evaluación para que no contamine las diferencias observadas en el conjunto de test entre los dos idiomas. Sin embargo, conviene descartar los datasets en los que la dificultad intrínseca entre los dos idiomas es muy grande, porque esa es una señal de que la metodología con la que se han seleccionado y/o anotado en ambos idiomas seguramente no es equivalente.

- Los datasets y las tareas deben tener diversidad para ser representativos de las distintas aplicaciones del Procesamiento del Lenguaje Natural. Nos centraremos en tareas discriminativas que sean susceptibles de ser abordadas directamente por los modelos de lenguaje: en particular, de clasificación y de etiquetado. En el segundo año hemos empezado a trabajar también en tareas generativas por ser especialmente relevantes en el mercado actual de la IA, aunque su evaluación es mucho más compleja: o se invierte mucho esfuerzo en evaluaciones no automatizables, o se evalúan de forma poco precisa con medidas automatizadas de similitud textual con modelos realizados por humanos. También es deseable cierta diversidad en los dominios y en el tipo de textos.

- Accesibilidad. Los datos de entrenamiento deben ser fáciles de conseguir para los desarrolladores, idealmente mediante un sólo acuerdo centralizado con la UNED como proveedora.

Este conjunto de requisitos no se cumple en la mayoría de datasets disponibles, especialmente dada la necesidad que tenemos de expandirlos para disponer de un juego de pruebas privado que no haya sido publicado.

Los siguientes datasets se crearon para el leaderboard en el primer año de trabajo, y constituyen el leaderboard ODESIA CORE. En todos ellos se ha realizado al menos una anotación adicional para disponer de un juego de pruebas privado que no pueda ser leído por los modelos de lenguaje en su proceso de entrenamiento, garantizando así que la evaluación está libre de problemas de contaminación:

- DIPROMATS 2023. Este dataset fue creado desde cero para incorporarlo al Leaderboard ODESIA en la Versión 1. Se trata de un conjunto de tuits emitidos por diplomáticos de cuatro potencias mundiales (la Unión Europea, Rusia, China y Estados Unidos), anotados en función de las técnicas de propaganda que utilizan para transmitir una imagen determinada de sus países o de sus competidores a nivel global. Hay tres tareas asociadas con este dataset: identificación de propaganda, caracterización a grano grueso (cuatro técnicas) y caracterización a grano fino (15 técnicas subsumidas en las anteriores). Se trata de un problema de clasificación multiclase y multietiqueta. Se enmarca dentro de los problemas relacionados con la desinformación.

- EXIST 2022. Este dataset fue extendido para su incorporación al leaderboard en la Versión 1, creando un subconjunto adicional de datos anotados como conjunto de test privado. Se trata de un conjunto de tuits anotado en función de si contienen mensajes sexistas o no, y de qué tipo de sexismo se trata. Se enmarca dentro del problema de la toxicidad en redes sociales.

- DIANN 2023. Este dataset fue adaptado y extendido para su incorporación al Leaderboard. Se creó una partición de evaluación para la Versión 1 del Leaderboard. Los textos son resúmenes de artículos sobre biomedicina, y la tarea consiste en identificar menciones de discapacidades. Se trata por tanto de una tarea de etiquetación de secuencias.

En la Versión 2 del Leaderboard ODESIA CORE se han incorporado los siguientes datasets:

- EXIST 2023. Se trata de un dataset creado en su integridad para la Versión 2 del Leaderboard. Se compone de tuits etiquetados en función del tipo de sexismo expresado o descrito en ellos. Se trata, además, de un dataset desarrollado siguiendo el paradigma de "aprendizaje con desacuerdo" (Learning with Disagreement, LeWiDi) (Uma et al., 2021a), lo que lo convierte en el primer dataset para el entrenamiento y prueba de sistemas de detección de sexismo en textos construido conforme a este paradigma. Consta de tres particiones (entrenamiento, desarrollo, evaluación) y anotaciones para tres tareas: Detección de sexismo, categorización e identificación del emisor de sexismo. Se enmarca dentro del problema de la toxicidad en redes sociales.

- SQUAD/SQAC 2024. Este dataset contiene una partición de evaluación creada para la Versión 2 del Leaderboard ODESIA. Contiene artículos de divulgación científica del CSIC para el español y de Cambridge University para el inglés. La tarea que este dataset permite evaluar es la de comprensión de texto extractiva en sistemas de pregunta-respuesta. La tarea consiste en responder a preguntas sobre un texto, de tal manera que la respuesta sea un fragmento extraido directamente del texto. Se trata de una tarea de etiquetación de secuencias. El hecho de que los documentos anotados en los SQUAD/SQAC originales sean de fuentes distintas a los de nuestra anotación hace que, desde el punto de vista de los sistemas supervisados, este dataset sea particularmente complejo, ya que implícitamente se está midiendo la capacidad de transferencia del aprendizaje entre dominios. Además, es un dataset particularmente apropiado para evaluar modelos generativos en modo zero-shot (sin ejemplos) o few-shot (unos pocos ejemplos); de hecho, una de las motivaciones para incluirlo este año ha sido su proximidad a las aplicaciones más comunes a nivel empresarial de los modelos generativos: la capacidad de los modelos para extraer y sintetizar información de información corporativa en formato texto o semiestructurado, en una aproximación RAG (Retrieval-Augmented Generation) no supervisada.

La Tabla 6 muestra un resumen de los datasets del Leaderboard ODESIA CORE v2. Adicionalmente, se ha desarrollado otros datasets pensados exclusivamente para evaluar modelos generativos en modo zero-shot o few shot. De momento no se han añadido al leaderboard, ya que los modelos discriminativos no pueden abordarlos.

- UNED ACCESO 2024. El dataset contiene 1003 preguntas de opciones múltiples de once asignaturas del Curso de acceso para Mayores de 25 años de la UNED. Las preguntas y sus respuestas se han

Dataset	Tareas	Tarea abstracta	Dominio	Área de aplicación
DIANN 2023	detección de discapacidades	etiquetado	Biomedicina	Entidades nombradas
DIPROMATS 2023	Identificación de propaganda	Clasificación binaria	Geopolítica	Desinformación
	Caracterización de propaganda (gruesa)	Clasificación jerárquica multilabel	Geopolítica	Desinformación
	Caracterización de propaganda (fina)	Clasificación jerárquica multilabel	Geopolítica	Desinformación
EXIST 2022	Detección de sexismo	Clasificación binaria	Redes sociales	Toxicidad
	Categorización de sexismo	Clasificación jerárquica multiclase	Redes Sociales	Toxicidad
EXIST-2023	Identificación de sexismo	Clasificación binaria LeWiDi	Redes sociales	Toxicidad
	Intención de la fuente	Clasificación jerárquica LeWiDi	Redes Sociales	Toxicidad
	Categorización de sexismo	Clasificación jerárquica multilabel LeWiDi	Redes sociales	Toxicidad
SQUAD-SQAC 2024	Machine reading	Etiquetado de secuencias	Ciencia	Pregunta-respuesta

Tabla 6: Resumen de los datasets incorporados al Leaderboard ODESIA CORE v2.

traducido manualmente al inglés (sin intervención de ningún sistema de traducción automática, para evitar sesgos). Este dataset permite evaluar el conocimiento general de los modelos generativos, de forma similar a otros datasets como MMLU. Las diferencias con MMLU son: la disponibilidad de las preguntas en dos idiomas mediante traducciones manuales, y que el dataset no se hace público, de forma que se limitan los problemas de contaminación. Sobre este dataset se ha finalizado una experimentación exhaustiva sobre los mejores modelos generativos actuales (Claude 3 Opus y GPT-4) y sobre varios modelos abiertos (Llama-2, Mistral y Gemma).

- CURIA 2024. Este dataset contiene particiones de entrenamiento, desarrollo y evaluación. Los textos que lo conforman son sentencias de tribunales de justicia de la Unión Europea, que están acompañados de micro-resúmenes en lenguaje claro. En este caso, la tarea que permite evaluar el dataset CURIA-2024 es el resumen simplificado de textos legales, por tanto, se trata de una tarea de generación de texto. El dataset está finalizado y la experimentación se realizará en el tercer año de proyecto.

- Pron vs Prompt. El dataset consiste en 120 sinopsis para 60 títulos de películas imaginarias, 30 propuestos por GPT4 y 30 por un novelista de prestigio (Patricio Pron, premio Alfaguara de novela). Se solicitó a ambos, al escritor y GPT-4, que escribieran sinopsis de aproximadamente 600 palabras para cada título, incluyendo tanto los propuestos por ellos mismos como por su contraparte. En este caso, se está realizando una evaluación sistemática por parte de críticos y académicos, con la que se espera medir de forma precisa y fiable las capacidades de escritura creativa de los modelos de forma puntual.

En conjunto, los 5 datasets del Leaderboard ODESIA que tienen tests privados (ODESIA CORE) aportan 10 tareas distintas que abarcan problemas de clasificación (binaria, multiclase, multietiqueta, jerárquica, clasificación con disagreement) y etiquetado de secuencias. Hay varios tipos de textos: tuits en redes sociales, mensajes de autoridades y diplomáticos, resúmenes científicos y noticias de divulgación científica. Y en cuanto a dominios, se abarcan el biomédico, el de política y relaciones internacionales, las redes sociales, y dominios científicos variados.

Los detalles sobre la creación y contenido de cada dataset se especifican en los informes técnicos que también se entregan para cada dataset. A continuación se presenta un resumen de cada uno de ellos.

3.3.2. Datasets

Dataset 1: EXIST 2022

EXIST (sEXism Identification in Social neTworks) 2022 es un dataset desarrollado para facilitar la investigación en detección automática de sexismo en redes sociales. Se compone de textos cortos procedentes de redes sociales etiquetados en función del tipo de sexismo expresado o descrito en ellos. Contiene datos de dos redes sociales diferentes: Twitter [11] y Gab [12]. Se trata, por tanto, de mensajes cortos intercambiados en alguna de las dos redes anteriores y de dominio general (es decir, no versan, a priori, sobre ninguna temática en particular).

En total, el dataset se compone de 12,390 textos etiquetados: 6,226 en español y 6,164 en inglés. La distribución de los textos por partición (entrenamiento/test), fuente (Twitter/Gab) e idioma (inglés/español) se muestra en la Tabla 11. Sobre ellos se definen dos tareas:

- Tarea 1: detección de sexismo (clasificación binaria). Los sistemas deben decidir, para cada tweet, si contiene mensajes sexistas o no.
- Tarea 2: detección y caracterización de sexismo (clasificación multiclase). Los sistemas deben decidir, para cada tweet, si es o no sexista, y en caso afirmativo qué tipo de sexismo, entre las siguientes categorías: ideological and inequality, stereotyping and dominance, objectification, sexual violence, misogyny and non-sexual violence.

Entrenamiento				**Test**	
Twitter		**Gab**		**Twitter**	
Español	**Inglés**	**Español**	**Inglés**	**Español**	**Inglés**
5,211	5,152	490	492	522	513
10,363		982		1,035	

Tabla 7: Distribución de datos de EXIST 2022 por partición, fuente e idioma.

Cada texto del dataset tiene asignadas 1 o 2 etiquetas, dependiendo de si se trata de un texto sexista o no. La primera etiqueta responde a la pregunta: *¿Es el texto sexista, en cualquiera de sus formas, o describe conductas o situaciones en las que se produce discriminación sexista (es decir, es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista)?* En función de la respuesta a esta pregunta, la etiqueta puede tomar uno de los siguientes dos valores:

- `SEXIST`: el tuit es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista. Ejemplos de textos del dataset anotados con esta etiqueta son:
  (1) *Que materialistas se han vuelto las mujeres de hoy en día, aún recuerdo cuando las podíamos enamorar con puras mentiras.*
  (2) *Las mujeres no deberian ni maquillarse. Ya sabemos que son rompe bolas arregladas o desarregladas. Joda las amamos. JAJAJA.*
  (3) *Te tacharán de machista y misógino. Las mujeres pueden pegar los hombres no.*
- `NON-SEXIST`: el tuit no perjudica, discrimina o menosprecia a las mujeres ni se refiere a contenidos o situaciones en las que sí se haga. Ejemplos de textos del dataset anotados con esta etiqueta son:
  (4) *Mi mamá: Si te llegas a ir del país, vas a saber lo dura que es la vida, porque ni siquiera te dignas a fregar las cosas que usas.*
  (5) *Alguien por favor me regala un bate que yo entre los fachas los señoros de las nuevas masculinidades y los que mucho feminismo pero el porno no me lo toqueis quiero reventar una cabeza.*

[11] https://twitter.com/
[12] https://gab.com/

(6) *El impactante video de una mujer al volante que sale milagrosamente ilesa de un intento de asesinato...*

La segunda etiqueta responde a la pregunta: *De acuerdo con la faceta de la mujer que está siendo atacada, ¿qué tipo de sexismo encontramos en el tuit?*. El sexismo puede afectar a las mujeres en muchas facetas de sus vidas, incluidas las funciones domésticas y de crianza, las oportunidades profesionales, la imagen sexual y las expectativas de vida, por nombrar algunas. En función de la respuesta, la etiqueta puede tomar uno de los siguientes valores:

- `IDEOLOGICAL-INEQUALITY`: Descrédito ideológico, negación de la desigualdad y narrativa invertida. El tuit desacredita el movimiento feminista, rechaza la existencia de desigualdad entre hombres y mujeres, o presenta a los hombres como víctimas de la opresión de género. Ejemplos de tuits en esta categoría son:

  (7) *Mucho feminismo pero poco interés con casos de pedofilia, que hipócrita de mierda.*

  (8) *Mucho feminismo mucho feminismo pero andamos tirando historias burlándose de una piba solo porque el novio le metió los cuernos y volvió. Gente hipócrita en esta vida existen y muchas.*

  (9) *Mi hermana y mi madre se burlan de mí por defender todo el tiempo los derechos de todos y me acaban de decir feminazi, la completaron.*

- `STEREOTYPING-DOMINANCE`: Estereotipos y dominancia. El tuit expresa ideas falsas sobre las mujeres que sugieren que son más adecuadas o inapropiadas para ciertas tareas o afirma que los hombres son superiores a las mujeres. Ejemplos de textos del datasets con esta etiqueta son:

  (10) *Un taxista enojado con una mujer que pasaba en bicicleta le gritó "mujer tenias que ser" creo que la quiso insultar pero no se alto orgullo serlo bobo.*

  (11) *@RuloTanki @PrinncessRoyal Uy andas en tus días? Jajajja.*

  (12) *Paula2R faber_acuria A las mujeres hay que amarlas...solo eso... Nunca las entenderás.*

- `OBJECTIFICATION`: Objectificación. El tuit presenta a las mujeres como objetos al margen de su dignidad y personalidad, o asume o describe ciertas cualidades físicas que las mujeres deben tener para cumplir con los roles tradicionales de género. Ejemplos de textos anotados con esta etiqueta son:

  (13) *@EveryWhereTraps Tienes buen tamaño para ser chica, más tamaño ya no encuentras casi ropa y menos zapatos bonitos.*

  (14) *¿Cómo iba vestida?", "seguramente estaba provocando", "mmm ese short esta muy corto", "no te pongas eso que pareces una puta", "bueno iba sola por la calle a esa hora", "se lo busco", "bueno, para que toma entonces", "le gustó", "¿porque no gritó?*

  (15) *Pareces una puta con ese pantalón - Mi hermano de 13 cuando me vio con un pantalón de cuero.*

- `SEXUAL-VIOLENCE`: Violencia sexual. Se realizan sugerencias sexuales, solicitudes de favores sexuales o acoso de carácter sexual (violación o agresión sexual). Ejemplos de textos anotados con esta etiqueta son:

  (16) *@giulianarob @Gattarok @DemenArt ¿Misógino? Misoginia es odio a las mujeres y, en el porno los hombres las hacen disfrutar mucho.*

  (17) *#MeToo Estas 4 no han conseguido su objetivo. El juez estima que se abrieron de patas https://t.co/GSHiiwqY6Aánta lagartona hay en este https://t.co/8t5VmFIUFn.*

- `MISOGYNY-NON-SEXUAL-VIOLENCE`: Misoginia y violencia no sexual. El tuit expresa odio y violencia hacia la mujer. Ejemplos de textos anotados con esta etiqueta son:

(18) *Las mujeres de hoy en dia te enseñar a querer... estar soltero.*

(19) *Cualidades = odio a las mujeres, extranjeros, trabajadores, catalanes, vascos...*

(20) *Odio la misoginia más de lo que odio a las mujeres.*

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Informe del dataset EXIST 2022", que fue entregado en el Año 1.

Dataset 2: DIPROMATS 2023

DIPROMATS consiste en un conjunto de tweets en español y en inglés emitidos por diplomáticos y autoridades de cuatro potencias mundiales: EEUU, Rusia, China y la UE. Sobre ellos se ha realizado una anotación (de expertos) en la que se identifican los tweets que contienen algún tipo de propaganda, y una caracterización de las técnicas de propaganda usadas, tanto en grano grueso (cuatro categorías) como en grano fino (quince subcategorías). El dataset contiene 24,248 tweets anotados para las tres tareas, y se utiliza dentro del proyecto de dos formas: como parte del leaderboard bilingüe para evaluación comparada de modelos del lenguaje en español e inglés, y como parte de los datasets utilizados para la medición de la brecha de la IA entre ambos idiomas.

La Tabla 8 muestra la distribución de tuits y autoridades por áreas geopolíticas para el inglés, y la Tabla 9 para el español.

	China	Rusia	UE	EEUU	Total
Tuits	3,647	3,591	3,553	3,956	14,747
Autoridades	106	111	186	216	619

Tabla 8: Distribución del dataset DIPROMATS 2023 en inglés por áreas geopolíticas

	China	Rusia	UE	EEUU	Total
Tuits	2,997	1,391	2,465	2,738	9,501
Autoridades	25	22	48	40	135

Tabla 9: Distribución del dataset DIPROMATS 2023 en español por áreas geopolíticas

Utilizando las anotaciones manuales, se han definido tres tareas:

- **Tarea 1: Identificación de propaganda**, que se plantea como un problema de clasificación binaria que consiste en determinar si un tuit contiene o no técnicas de propaganda.
- **Tarea 2: Caracterización de la propaganda (grano grueso)**, que tiene como objetivo categorizar los mensajes propagandísticos según el tipo de propaganda que contienen. La categorización propuesta considera múltiples técnicas identificadas mediante revisión de la literatura existente, que se agrupan según sus características retóricas. Proponemos una tarea de clasificación multiclase y multietiqueta, en la que los sistemas tienen que asignar los tuits a una o más de las categorías definidas. Hay dos tipos de categorías:
  - Una categorización de **grano grueso** con cuatro clases de propaganda (más una clase negativa):
    - Grupo 0. `Not propaganda`
    - Group 1. `Appeal to Commonality`
    - Grupo 2: `Discrediting the opponent`
    - Group 3: `Loaded Language`
    - Grupo 4: `Appeal to authority`
- **Tarea 3: Caracterización de la propaganda (grano fino)**, idéntica a la anterior pero con un conjunto de subcategorías que refinan la clasificación anterior. En este caso se distinguen 15

subtipos de propaganda: `Flag Waving`, `Ad Populum / Ad antiquitatem`, `Name Calling`, `Undiplomatic Assertiveness / Whataboutism`, `Scapegoating`, `Propaganda Slinging`, `Appeal to Fear`, `Demonization`, `Personal Attacks`, `Doubt`, `Reductio Ad Hitlerum`, `Loaded Language`, `Appeal to False Authority` y `Bandwagoning`.

Como parte del leaderboard, el dataset tiene varias características interesantes:

- Se trata de un problema complejo (especialmente al nivel de caracterización de grano fino), difícil de resolver por cualquier sistema de PLN. Frente a otros datasets, éste tiene la ventaja de que no saturará fácilmente (momento en el cual no sirve para medir diferencias entre sistemas).
- Se trata de un problema de clasificación jerárquico, multiclase y multietiqueta. Aunque este tipo de problemas es muy habitual en aplicaciones prácticas de la IA, en entornos de laboratorio suelen simplificarse ignorando las características jerárquicas de las clases, o reduciendo el problema artificialmente a una sola etiqueta por ítem.
- Por tratarse de fuentes diplomáticas, abarca una gran variedad de registros dialectales, desde los propios del pais donde trabaja cada diplomático hasta su grado de bilingüismo (hay diplomáticos que son nativos en el idioma del pais de destino, otros lo han adquirido como segunda lengua).
- Se trata de una anotación de expertos, más costosa que el crowdsourcing pero que permite el uso de tipologías de anotación más sofisticadas, como es el caso de DIPROMATS.

Los detalles del dataset DIPROMATS 2023 están recogidos en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español: Informe del dataset DIPROMATS", entregado en el Año 1.

Dataset 3: DIANN 2023

El dataset DIANN-2023 se ha compilado y anotado en la UNED en el marco del proyecto. Consiste en un dataset bilingüe de resúmenes de artículos científicos relacionados con enfermedades raras anotados manualmente con discapacidades. El dataset ha sido concebido con el objetivo de entrenar sistemas de reconocimiento de entidades nombradas especializados en la detección de discapacidades. Una parte del corpus se creó en 2018 para la competición de IberLEF "Disability annotation on documents from the biomedical domain (DIANN)" [13](Fabregat et al., 2018), que proponía una tarea de reconocimiento de entidades centrada en la identificación de discapacidades. Otra parte se ha creado en 2023 con el fin de incorporarla como partición privada de test al leaderboard del proyecto ODESIA. La anotación de esta última parte es la que ha sido financiada por el proyecto.

El corpus se proporciona en dos particiones, una de entrenamiento y otra de evaluación. La partición de entrenamiento contiene 500 textos en cada lengua. Estos textos se corresponden con las particiones de entrenamiento y evaluación hechas públicas para la competición DIANN en Iberlef 2018, donde se proporcionaban 400 archivos de entrenamiento y 100 de evaluación por lengua. Además se dispone de una partición privada de test que contiene 100 textos para cada lengua. Puesto que esta es la partición que se usa para evaluar sistemas en el leaderboard, esta partición no se hará pública y no se proporciona información sobre sus contenidos más allá de la información referente al tamaño y la metodología de anotación.

En la Tabla 10 se muestran los detalles de tamaño para ambas particiones y el número anotaciones en la partición de entrenamiento. Los tokens se han calculado contando unidades separadas por espacios en blanco.

En el corpus se han anotado todas las discapacidades mencionadas en los textos. Por tanto, la única etiqueta usada es la de 'discapacidad'. La distribución por lengua en la partición de entrenamiento es la siguiente:

[13]Página web de la competición DIANN 2018: http://nlp.uned.es/diann/

Partición	Entrenamiento				Evaluación		
Idioma	**# docs**	**# tokens**	**# líneas**	**# discapacidades**	**# docs**	**# tokens**	**# líneas**
Español	500	98948	5923	1555	100	21103	1203
Inglés	500	89325	6901	1656	100	19087	1234

Tabla 10: Información sobre el tamaño del corpus DIANN-2023.

- **Español**: 1555 menciones de discapacidades anotadas.
- **Inglés**: 1656 menciones de discapacidades anotadas.

Dentro del leaderboard, este dataset es representativo de uno de los dominios de aplicación más relevantes del PLN, el dominio biomédico; y de las tareas de etiquetado, que son las más prototípicas del PLN discriminativo junto con las de clasificación.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español, Informe del dataset DIANN-2023", entregado en el Año 1.

Dataset 4: EXIST-2023

En la línea de EXIST 2022, EXIST (sEXism Identification in Social neTworks) 2023 es un dataset desarrollado para facilitar la investigación en detección automática de sexismo en redes sociales. Se compone de textos cortos, tuits, procedentes de redes sociales etiquetados en función del tipo de sexismo expresado o descrito en ellos. A diferencia de EXIST 2022, se trata de un dataset desarrollado siguiendo el paradigma de "aprendizaje con desacuerdo" (Learning with Disagreements, LeWiDi) (Uma et al., 2021a), lo que lo convierte en el primer dataset para el entrenamiento y prueba de sistemas de detección de sexismo en textos construido conforme a este paradigma, y el primer dataset de este tipo que se incorpora al Leaderboard ODESIA.

En este paradigma, en lugar de depender de una única etiqueta "correcta" para cada ejemplo o instancia del dataset, el modelo se entrena para aprender de anotaciones conflictivas o diversas. De esta manera, las perspectivas, sesgos o interpretaciones diferentes de los anotadores pueden ser tenidas en cuenta por los sistemas, permitiendo un aprendizaje más justo y equitativo. Esto se deriva del hecho de que el dataset ha sido anotado por diferentes anotadores, con distintas características sociodemográficas, de manera que todas las notaciones (aun siendo en algunos casos contradictorias) forman parte del dataset final.

El dataset se compone de 10,034 textos etiquetados, 5,307 en español y 7,727 en inglés. Los textos proceden de conversaciones de Twitter. Cada texto está anotado con diferentes tipos de etiquetas. Como en EXIST 2022, en EXIST 2023 la primera etiqueta responde a la pregunta: *¿Es el texto sexista, en cualquiera de sus formas, o describe conductas o situaciones en las que se produce discriminación sexista (es decir, es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista)?* y la clasificación es binaria. A diferencia de EXIST 2022, en EXIST 2023 la segunda etiqueta responde a la pregunta: *¿Cuál crees que es la intención de la persona que escribió el tweet?* Las categorías anotadas son: `DIRECT`, `REPORTED`, `JUDGEMENTAL`. La tercera etiqueta responde a la pregunta: *De acuerdo con la faceta de la mujer que está siendo atacada, ¿qué tipo de sexismo encontramos en el tweet?* y las categoríaas son las mismas que en EXIST 2022.

El dataset presenta tres particiones para cada lengua: entrenamiento, desarrollo y test. La distribución de los textos por partición e idioma se muestra en la Tabla 11.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Año 2 - Informe del dataset EXIST 2023", entregado con este informe.

	Entrenamiento	Desarrollo	Test	Total
Español	3,660	549	1,098	5,307
Inglés	3,260	489	978	7,727
Total	6,920	1,038	2,076	10,034

Tabla 11: Distribución de EXIST 2023 por partición e idioma.

Dataset 5: SQUAD/SQAC-2024

En este caso, la tarea que este dataset permite evaluar es la de comprensión de texto en sistemas de pregunta-respuesta con respuestas extractivas. La tarea consiste en responder a preguntas sobre un texto, de tal manera que la respuesta sea un fragmento extraído directamente del texto. Los textos son noticias del CSIC para español y de Cambridge University para inglés. En todos los casos, las respuestas son fragmentos del texto y no se incluyen preguntas que no se puedan contestar a partir del texto. Esta tarea es interesante por su dificultad, ya que requiere tanto entender el lenguaje como tener una representación del conocimiento del mundo en general y del mundo representado en cada texto.

SQUAD/SQAC-2024 es una extensión de los datasets SQUAD/SQAC. SQAC (Spanish Question Answering Corpus) (Gutiérrez-Fandiño et al., 2021) es un dataset de pregunta-respuesta, con respuestas extractivas en español. En la tarea de PLN asociada, dada una pregunta y un párrafo, el sistema debe localizar el span (fragmento) más pequeño que contiene la respuesta. La metodología para crearlo está basada en la de SQuAD (Stanford Question Answering Dataset) v1.1 (Rajpurkar et al., 2016), un dataset de pregunta-respuesta extractivo en inglés. Si bien el dataset SQAC se creó por la necesidad de tener un corpus de pregunta-respuesta en español que no fuera una traducción del inglés, el dataset SQUAD/SQAC-2024 se ha creado para disponer de una partición de evaluación privada que permita evaluar Large Language Models (LLMs) sin riesgo de contaminación de datos, tanto en inglés como en español.

El dataset contiene noticias académicas del CSIC (Centro Superior de Invvestigaciones Científicas) para el español[14] y de Cambridge University para el inglés.[15] Las noticias son de dominios científicos variados y suelen ser cortas, entre 712 y 2,760 palabras en inglés, y entre 514 y 2,818 palabras en Español. Además, están dirigidas al público general, por lo que no se usa lenguaje especializado.

En la Tabla 12 se proporciona información sobre el tamaño del dataset por idioma.

	# textos	# tokens	μ tokens/texto	# pares pregunta-respuesta	μ preguntas/texto
Español	110	962,502	840	1,144	10,4
Inglés	110	1,235,638	1,045	1,182	10,7
Total	220	2,198,140	–	2,379	–

Tabla 12: Información sobre el tamaño del dataset por idioma.

Las unidades que conforman el dataset son pares de pregunta-respuesta. Hay 1,144 pares en español y 1,182 pares en inglés, realizados sobre 110 textos en cada lengua. Los textos en español son un poco más cortos, con una media de 840 palabras, en comparación con los textos en inglés, con una media de 1,045 palabras. Se ha realizado una media de aproximadamente 10 preguntas por texto tanto en inglés como en español.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Año 2 - Informe del dataset SQUAD/SQAC-2024", entregado con este informe.

Datasets públicos adicionales para la medición de la brecha

Para afinar la brecha de efectividad inglés/español hemos utilizado todos los datasets bilingües disponibles en el Leaderboard ODESIA EXTENDED, que incorpora cuatro datasets bilingües adicionales

[14] https://www.csic.es/es/actualidad-del-csic/noticias
[15] https://www.cam.ac.uk/news

construidos en inglés y español de dominio público. Hemos aplicado con ellos la misma metodología que con los de ODESIA CORE. Los datasets adicionales son:

- **DIANN Task 2** La segunda tarea de DIANN consiste en localizar y determinar el alcance de las negaciones en resúmenes de artículos biomédicos.
- **MLDoc** El Multilingual Document Classification Corpus (MLDoc) (Schwenk and Li, 2018) es un dataset de clasificación de noticias en varios idiomas, del que usamos el subconjunto de inglés y el subconjunto de español. La tarea tiene cuatro categorías: corporate/industrial, economics, government/social y markets.
- **MultiCONER 2022** (Malmasi et al., 2022) es un dataset multilingüe para reconocimiento de entidades nombradas complejas de seis categorías diferentes. Como en los demás casos, utilizamos los subconjuntos de español e inglés para nuestra experimentación. En este dataset, el conjunto de entrenamiento es un orden de magnitud más pequeño que el conjunto de evaluación (del orden de 5,000 vs 50,000).
- **STS-2017** (Cer et al., 2017) es un dataset multilingüe de similitud textual, en la que los sistemas deben predecir el grado de similitud entre un par de oraciones. Esta es la única tarea de regresión en nuestro diseño experimental, y también la única en la que se usa como métrica de evaluación la correlación Pearson entre las predicciones del sistema y las anotaciones manuales. De nuevo, utilizamos los subconjuntos de inglés y español.
- **SQUAD/SQAC**: en este caso se trata de dos datasets construidos de forma independiente pero con la misma metodología. SQAC (Spanish Question Answering Corpus) (Gutiérrez-Fandiño et al., 2021) es un dataset de Question Answering extractivo, en el que, dada una pregunta y un párrafo asociado, el sistema debe localizar el span más pequeño que contiene la respuesta. La metodología para crearlo está basada en la de SQuAD v1.1 (Rajpurkar et al., 2016), por lo que se pueden considerar datasets equivalentes. De hecho, nuestro algoritmo baseline obtiene un rendimiento equivalente en ambos idiomas (0,53 en ambos casos), lo que hace el par SQAC/SQUAD un conjunto ideal para comparar modelos de lenguaje en ambos idiomas. Nótese que este dataset también se utiliza como conjunto de entrenamiento en el SQUAD/SQAC 2024, pero en este último el conjunto de test se ha desarrollado en ODESIA sobre un tipo de documentos ligeramente diferente.

Para todos estos datasets utilizamos los datos públicos de entrenamiento y test, a diferencia de los datasets del Leaderboard ODESIA, en los que los conjuntos de evaluación no están disponibles públicamente. El proceso de entrenamiento para las tareas extendidas es idéntico al de las tareas de ODESIA CORE. En conjunto, la medición del gap se realiza sobre quince tareas.

3.3.3. Datasets adicionales no incorporados en el Leaderboard

Adicionalmente, se ha desarrollado los siguientes datasets que no se incorporan al Leaderboard público.

Dataset 6: CURIA-2024

CURIA-2024 es un dataset compuesto por sentencias de tribunales en inglés y en español, con sus correspondientes resúmenes simplificados. Se enmarca, por tanto, en el dominio jurídico. El dataset se ha elaborado con textos descargados de la página web del Tribunal de Justicia de la Unión Europea. Este organismo está formado por el Tribunal General (de primera instancia) y el Tribunal de Justicia. Todas las sentencias de ambos tribunales son publicadas en sus páginas web.[16] Aparte de las sentencias, CURIA-2024 contiene los micro-résumenes de las sentencias, que han sido creados por profesionales expertos en lenguaje legal.

Puesto que el dataset se compone de sentencias con sus resúmenes, la unidad de referencia es el par texto/resumen. Como se muestra en la Tabla 13, el dataset en inglés contiene 2,230 pares texto/resumen

[16] https://curia.europa.eu/jcms/jcms/j_6/es/

sumando un total de 18,153,858 tokens, mientras que el dataset en español contiene 1,961 pares sumando un total de 17,842,388 tokens. El dataset está dividido en tres particiones por lengua: entrenamiento, desarrollo y evaluación. Las particiones se han creado siguiendo el mismo procedimiento para las dos lenguas: se ha tomado un 80 % para entrenamiento, un 10 % para desarrollo y un 10 % para evaluación.

Partición	**Entrenamiento**		**Desarrollo**		**Evaluación**		**Total**	
Idioma	**# docs**	**# tokens**	**# docs**	**# tokens**	**# docs**	**# tokens**	**# docs**	**# tokens**
Español	1,568	14,173,589	196	1,820,298	197	1,848,501	1,961	17,842,388
Inglés	1,784	14,461,442	223	1,724,975	223	1,967,441	2,230	18,153,858

Tabla 13: Número de pares sentencia–resumen en el dataset CURIA-2024.

Mediane este dataset se pueden evaluar sistemas de resumen simplificado de textos legales, por lo que se trata de una tarea de generación de texto.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Informe del dataset CURIA-2024", entregado con este informe.

Dataset 7: UNED ACCESO 2024

Muchos benchmarks se han propuesto como evaluaciones de una sola tarea, pero con el surgimiento de modelos de lenguaje generales como BERT (Devlin et al., 2018), se ha popularizado el desarrollo de benchmarks más completos para poder medir las capacidades generales de estos modelos. GLUE (Wang et al., 2018) y SuperGLUE (Wang et al., 2019) son benchmarks populares que evalúan el rendimiento de los modelos de lenguaje con distintas tareas de PLN. Más recientemente se han introducido benchmarks que contienen una amplia gama de tareas de PLN para la evaluación, como Big-Bench (Srivastava, 2022), Big-Bench Hard (Suzgun et al., 2022), MMLU (Hendrycks et al., 2021) y HELM (et al, 2023). La mayoría de los datasets presentes en estos benchmarks proponen evaluaciones que se realizan con conjuntos de datos creados artificialmente para tareas concretas de PLN, en vez de proponer escenarios reales de evaluación como los exámenes con los que se evalúa a los humanos. El reconocimiento de esta limitación ha llevado a que en los últimos tiempos se haya puesto énfasis en la importancia de las evaluaciones centradas en evaluar capacidades humanas. Se han introducido así benchmarks como AGIEval (Zhong et al., 2023), que se centran en un tipo de evaluación que pone el énfasis en tareas cognitivas a nivel humano, en escenarios reales. Este benchmark incluye exámenes de acceso a la universidad, pruebas de admisión a la facultad de derecho, concursos de matemáticas y pruebas de cualificación de abogados. Por otro lado, se han desarrollado diversos datasets a partir de exámenes, que abarcan distintas tareas (no sólo de respuesta múltiple), como en RACE (Lai et al., 2017). Las preguntas de respuesta múltiple se han erigido como uno de los métodos preferidos para evaluar los nuevos modelos generativos, debido a la dificultad que presenta su evaluación y la ausencia de una métrica estándar.

Así, en la intersección entre las evaluaciones con preguntas de múltiple respuesta, y las evaluaciones centradas en capacidades humanas y basadas en exámenes, se enmarca el dataset EXÁMENES UNED 2024. El dataset se compone de preguntas tipo test con 3 o 4 respuestas extraídas de exámenes de los Cursos de Acceso para Mayores de 25 años de la UNED de los siguientes grados: Administración y Dirección de Empresas, Biología, Bioquímica, Economía, Fundamentos de Informática, Lengua Castellana, Literatura, Matemáticas, Matemáticas Aplicadas a las Ciencias Sociales, Matemáticas Avanzadas y Psicología. El dataset se presenta en español y en inglés. La versión en español se ha obtenido directamente de la transcripción de los exámenes de Acceso para mayores de 25 años, mientras que la parte en inglés se ha construído mediante la traducción (manual) de estos exámenes.

En la Tabla 14 se incluye el número de preguntas, el número de exámenes y el número de palabras por asignatura, así como el número de opciones en la respuesta que tienen los exámenes de cada asignatura.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Informe del dataset UNED ACCESO 2024", entregado con este informe.

Asignatura	# Preguntas	# Exámenes	# Respuestas por pregunta	# Palabras
Administración y Dirección de Empresas	87	6	3	3936
Biología	119	6	3	2872
Bioquímica	59	4	3	1466
Economía	51	3	4	1726
Fundamentos de Informática	63	6	4	1987
Lengua Castellana	94	4	4	2816
Literatura	91	6	4	5130
Matemáticas	73	11	3	1538
Matemáticas Aplicadas a las Ciencias Sociales	94	10	3	2941
Matemáticas Avanzadas	24	5	3	470
Psicología	248	14	4	5669
Total	1003	75	–	30551

Tabla 14: Distribución del número de preguntas y exámenes por asignatura, y número de opciones de respuesta por pregunta que tienen los exámenes de cada asignatura.

Dataset 8: PRON VS PROMPT

Motivación Históricamente, hitos en el desarrollo de la Inteligencia Artificial como las victorias de Deep Blue y AlphaGo en ajedrez y Go, respectivamente, han marcado el avance tecnológico en diversas área de investigación. En el caso del Go, el sistema de IA desarrolló estrategias de juego novedosas que han sido imitadas, desde entonces, por todos los maestros humanos: se demostró que la IA podía ser creativa. Pero los juegos de mesa tiene características muy particulares que los hacen muy adecuados para los sistemas de IA. Desde la aparición de ChatGPT, la atención se ha desplazado hacia tareas como la escritura creativa, mucho más complejas. Estos modelos de lenguaje desafían la concepción de la inteligencia humana y las fronteras de la creación artística tradicionalmente considerada exclusiva de los humanos. Tanto es así que la colaboración humano-máquina en la industrias creativas es cada vez mayor (Adelani et al., 2023). De hecho, tal es su relevancia que los propios desarrolladores de OpenIA, en la interfaz de ChatGPT, ofrecen como primera sugerencia de uso de su modelo de lenguaje la opción çrea un historia".

Objetivo A pesar de todo, no existen aproximaciones de evaluación objetiva y rigurosa en esta tarea. Este dataset tiene el propósito general medir comparativamente la calidad en la creación de textos creativos de GPT4, el modelo con mejor rendimiento hasta la fecha, y un escritor profesional. Así, el objetivo de este dataset es doble: por un lado, se quiere realizar un estudio comparativo entre las capacidades creativas de generación de texto de una Inteligencia Artificial avanzada (GPT-4) y un novelista consagrado, Patricio Pron (Premio Alfaguara de novela). Y, por otro lado, se diseñará una metodología rigurosa de evaluación de textos creativos que servirá para puntuar la salida de cualquier modelo. A medio plazo, esperamos que esta experimentación sea el punto de partida para el desarrollo de métricas de evaluación automática que permitan conocer de la manera más objetiva posible la calidad de las producciones literarias de los modelos de lenguaje.

Descripción El dataset consiste en 120 sinopsis para 60 títulos de películas imaginarias, 30 propuestos por GPT4 y 30 por Patricio Pron. Se solicitó a ambos, al escritor y a GPT-4, que escribieran sinopsis de aproximadamente 600 palabras para cada título, incluyendo tanto los propuestos por ellos mismos como por su contrincante.

Evaluación Los textos generados están (a fecha de finalización del año 2 del proyecto) siendo sometidos a una evaluación a ciegas por un panel de expertos, compuesto por críticos y académicos, para garantizar una valoración objetiva de la calidad, creatividad y coherencia narrativa de las sinopsis. Los resultados estarán disponibles a lo largo del tercer año del proyecto.

3.3.4. Modelos de lenguaje utilizados en la experimentación

Para establecer una comparación inicial entre modelos del lenguaje en inglés y español, hemos hecho una selección de modelos preentrenados en español, en inglés, y multilingües, escogiendo entre los más utilizados en la literatura y en Hugging Face. En conjunto se trata de cinco modelos en español, cuatro en inglés, y cinco modelos multilingües, lo que nos permite considerar diez modelos en el leaderboard español y nueve en inglés.

El leaderboard incluye tareas tanto discriminativas (EXIST, DIPROMATS, etc.) como generativas (CURIA, UNED Acceso). De manera natural, las tareas discriminativas suelen resolverse con modelos encoders. Y, las tareas generativas, con modelos decoder, también llamados modelos generativos; el objetivo de entrenamiento de los modelos decoder es generar texto fluido. Sin embargo, para resolver esta tarea, han demostrado adquirir una gran capacidad de comprensión de lenguaje tal que parecen ser capaces de resolver tareas discriminativas, aunque no estén originalmente diseñados para eso.

De momento en el leaderboard sólo se incluyen los modelos discriminativos. Respecto a modelos generativos, en este año se ha realizado una primera experimentación con el dataset UNED-ACCESO, utilizando los modelos decoder o generativos que ahora mismo son el estado del arte: GPT4 (Achiam et al., 2023), GPT3.5[17], Claude 3[18], Mistral (Jiang et al., 2023), Llama-2 (Touvron et al., 2023) y Gemma[19]. Además, está en proceso la evaluación de estos modelos con el dataset CURIA de resúmenes en texto claro. También se ha evaluado GPT-4 en modo zero-shot (proporcionando al sistema sólo la guía de anotación) para las tres tareas de DIPROMATS 2023.

Los modelos discriminativos incluidos en el leaderboard son los siguientes:

Español	Inglés	Multilingües
roberta-large-bne	roberta-large	xlm-roberta-large
roberta-base-bne	roberta-base	xlm-roberta-base
bert-base-spanish-wwm-cased	bert-base-cased	bert-base-multilingual-cased
distillbert-base-spanish-uncased	distilbert-base-uncased	distilbert-base-multilingual-cased
bertin-roberta-base-spanish		ixambert-base-cased

Tabla 15: Modelos de lenguaje evaluados.

Modelos en inglés

- **Bert-base-cased** BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) fue el primer modelo que utilizó la arquitectura transformer, que después ha sido la predominante en toda la investigación posterior. El modelo base tiene 12 capas transformer blocks, 12 attention heads, y 110 millones de parámetros. El preentrenamiento se realizó de forma autosupervisada en un gran corpus en inglés optimizando dos objectivos simultáneamente:
  - Masked Language Model (MLM). En cada frase el modelo oculta de forma aleatoria el 15 % de las palabras, e intenta predecir las palabras ocultas.
  - Next Sentence Prediction (NSP). En el preentrenamiento, dado un par de oraciones, el modelo debe predecir si las dos frases aparecian consecutivamente en el corpus de entrenamiento o no.
- **RoBERTa-base** RoBERTa (Liu et al., 2019) un modelo que introduce algunas variaciones respecto a BERT. La principal diferencia es que RoBERTa usa un masking dinámico, en el que en cada epoch (iteración de aprendizaje sobre el corpus de entrenamiento) se enmascaran distintas partes de la oración. RoBERTa-base tiene 123M de parámetros.

[17] https://chat.openai.com
[18] https://www.anthropic.com/news/claude-3-family
[19] https://ai.google.dev/gemma?hl=es-419

- **Roberta-large** La única diferencia entre ROBERTA-BASE y ROBERTA-LARGE es el número de parametros utilizado para su entrenamiento. Si bien ROBERTA-BASE fue entrenado con 123 millones de parámetros, ROBERTA-LARGE lo hizo con 354 millones de parámetros.
- **Distilbert-base-uncased** es un modelo preentrenado con el mismo corpus que BERT, pero se trata de un modelo más pequeño y más rápido (Sanh et al., 2019). El modelo fue preentrenado con tres objectivos:
  - Que tuviera la capacidad de devolver las misma probabilidades que el modelo BERT
  - Que fuera capaz de aprender como su predecesor al ocultar el 15 % de las palabras en la entrada.
  - Que fuera capaz de generar estados ocultos como su predecesor.

  Con esto se pretendía conseguir un modelo igual de potente que BERT, pero a su vez más rápido y pequeño.

Modelos en español

- **Roberta-base-bne** Este es un modelo basado en RoBERTa base que ha sido preentrenado con el mayor conjunto de datos disponible en español hasta la fecha de su entrenamiento. Forma parte del conjunto de modelos MarIA (Gutiérrez-Fandiño et al., 2021) desarrollados por el Barcelona Supercomputing Center y la Biblioteca Nacional de España dentro del Plan Nacional de Tecnologías del Lenguaje español. El corpus de entrenamiento está compuesto por un total de 570 GB de texto limpio, realizado por la Biblioteca Nacional de España a partir del rastreo de páginas web entre 2009 y 2019. Sin embargo, la cantidad de datos para entrenar roberta-base-bne fue inferior, para obtener un modelo más ligero.
- **Roberta-large-bne** Modelo RoBERTa large entrenado de forma similar al anterior.
- **Bertin-roberta-base-spanish** Este es un modelo entrenado desde cero sobre la porción en español de mC4 (Xue et al., 2020), que contiene unos 416M de documentos. Se trata de un modelo RoBERTa-base con 12 capas, 12 attention heads cada una, y 125M de parámetros.
- **Bert-base-spanish-wwm-cased**, conocido como *BETO* (Cañete et al., 2020), es un modelo entrenado en un gran corpus español y que tiene un tamaño similar a BERT-Base. Al igual que en las versiones anglosajonas, esta versión distingue entre mayúsculas y minúsculas.
- **Distillbert-base-spanish-uncased** Se trata de la versión "destilada"de BETO. Fue entrenada por sus mismos autores y tiene el propósito de ofrecer un modelo de un tamaño más reducido que BETO, pero con un rendimiento similar.

Modelos bilingües

- **xlm-roberta-base** (Conneau et al., 2019) Se trata de la versión multilingüe de RoBERTa, abarca más de 100 idiomas y, a pesar de que en la página de Huggingface asegura que está preentrenada con 2,5TB de datos CommonCrawl, está entrenado con menos cantidad de datos que la versión *large*.
- **xlm-roberta-large** Esta es la versión que sí utiliza toda la cantidad de datos de los 2,5TB. Al igual que el modelo inferior, su entrenamiento está realizado para más de 100 idiomas diferentes.
- **bert-base-multilingual-cased** Modelo BERT preentrenado con los 104 idiomas mejor representados en la Wikipedia.
- **distilbert-base-multilingual-cased** Modelo DistilBERT preentrenado con un corpus similar al modelo anterior.
- **ixa-ehu/ixambert-base-cased** (**?**) Modelo preentrenado multilingüe para inglés, español y euskera. El corpus de entrenamiento está compuesto por las Wikipedias en inglés, español y euskera, junto con artículos de noticias en euskera extraídos de periódicos online.

Entrenamiento de modelos e hiperparámetros

El entrenamiento eficaz de modelos de lenguaje avanzados es fundamental para que den su mejor rendimiento. Un aspecto fundamental de este proceso es la selección óptima de hiperparámetros, los cuales pueden influir significativamente en la capacidad del modelo para aprender de los datos. En este contexto, hemos implementado una estrategia de búsqueda exhaustiva, conocida como Grid Search, para optimizar los hiperparámetros de nuestros modelos de lenguaje.

Grid Search es una técnica de optimización de hiperparámetros que sistematiza el proceso de experimentación al construir y evaluar un modelo para cada combinación de parámetros especificada en una cuadrícula predefinida. Esto permite identificar la configuración de hiperparámetros que resulta en el mejor rendimiento del modelo.

La aplicación de Grid Search asegura una exploración exhaustiva del espacio de hiperparámetros, proporcionando una base sólida para la toma de decisiones informadas sobre la configuración óptima para el entrenamiento de modelos.

Para la optimización de los modelos de lenguaje, hemos definido la siguiente cuadrícula (grid) de hiperparámetros para explorar:

- Tamaño de *batch*: Especifica el tamaño del lote (batch size) durante el entrenamiento. Se consideraron tamaños de 32 y 16, ajustando el equilibrio entre la memoria consumida y la precisión del gradiente.
- Tasa de aprendizaje: Se utiliza para el ajuste de los pesos del modelo en cada iteración. Se exploraron valores de 0.00001, 0.00003 y 0.00005, buscando optimizar la velocidad y estabilidad del aprendizaje.
- Weight decay: Controla la regularización sobre los pesos del modelo. Se usaron los valores de 0.1 y 0.01. La regularización ayuda a prevenir el sobreajuste al penalizar pesos grandes.

Para cada combinación de hiperparámetros especificada en la grid, se entrenó un modelo de lenguaje desde cero, utilizando un conjunto de datos estandarizado. La evaluación del rendimiento de cada modelo se realizó a través de las métricas específicas que se detallan en la siguiente Sección, tratando de identificar la configuración de hiperparámetros que maximiza el rendimiento del modelo sobre el conjunto de desarrollo. Cada modelo se entrenó un total de 12 veces en diferentes configuraciones sobre cada tarea e idioma.

La aplicación del Grid Search ha permitido una exploración sistemática y completa del espacio de hiperparámetros, identificando la configuración que conduce al mejor rendimiento de los modelos de lenguaje en el contexto de nuestro proyecto. La metodología adoptada asegura que la selección de hiperparámetros se base en evidencia empírica, mejorando la confiabilidad y eficacia de los modelos entrenados, así como la fiabilidad del gap reportado.

Hemos realizado una configuración base para todos los modelos, con alguna alteración para los problemas de clasificación binaria y multiclase, ya que los modelos DistillBert no soportan el *'gradient_checkpoint'*. Sin embargo, algunos modelos requieren tener activada está opción debido a la carga de memoria que reciben a la hora de procesar la información en la capa de *Attention*, sobre todo en modelos grandes con muchos parámetros y en modelos que no convergían por el número de pasos o epochs. Según el conjunto de datos y sus tareas, hemos ajustado diferentes configuraciones para comprobar la consistencia de los modelos. En cualquier caso, la configuración final para cada tarea se estableció como única para todos los modelos.

Evaluación: Métricas y baselines

En esta sección resumimos las métricas y los algoritmos baseline, sin información lingüística, utilizados como referencia para calibrar la efectividad de los modelos del lenguaje entre inglés y español.

Para poder comparar resultados en datasets de dos idiomas es necesario estimar primero la dificultad intrínseca de cada dataset, de forma que puedan calibrarse los resultados en un idioma y otro para compararlos directamente. Para ello, en cada dataset hemos estimado la dificultad intrínseca mediante algoritmos de Machine Learning que no usan ningún tipo de información lingüística.

Respecto a las métricas, salvo que se indique lo contrario se ha utilizado la implementación de las métricas en la librería PyEvall desarrollada dentro del proyecto.

EXIST-2022

Métrica Para evaluar las dos tareas de EXIST-2022, se ha utilizado F1 Macro (ver informe correspondiente).

Baseline Para generar los baselines de las tareas de EXIST, vectorizamos el conjunto de datos y test sin ningún tipo de preprocesamiento, para evitar utilizar información lingüística. A partir de los conjuntos resultantes, entrenamos modelos de regresión logística, xgboots y Support Vector Machines (SVM). Tomamos como baseline la media de los resultados obtenidos. El proceso es el mismo para las dos tareas.

DIANN 2023

Métrica La métrica que hemos usado para evaluar la tarea de DIANN 2023 es F1, para ello hemos usado Seqeval,[20] un framework en python que evalúa el etiquetado de secuencias.

Baseline Para realizar la tarea de NER de DIANN, utilizamos Conditional Random Fields (CRF), una clase de métodos de modelado estadístico que se aplican a menudo en el reconocimiento de patrones. Lo hemos usado como fórmula de predicción estructurada. No se ha usado ningún tipo de información lingüística.

DIPROMATS 2023

Métrica La evaluación se ha llevado a cabo considerando las tareas como de clasificación, abarcando tanto clasificación binaria, para la Tarea 1, como multietiqueta, para las Tareas 2 y 3. Es importante destacar que la tarea de clasificación multietiqueta de estas dos últimas presenta una complejidad no trivial desde el punto de vista de las métricas de evaluación. Esto se debe a que las clases involucradas mantienen una relación jerárquica entre sí. Por ejemplo, en la Tarea 2, un error de clasificación entre el grupo 2 y el grupo 3 se considera menos grave que un error entre el grupo 2 y el grupo 0.

Por ello, además de las métricas estándar, se reporta la métrica ICM (Amigó and Delgado, 2022), la cual se adapta de manera óptima a las particularidades de nuestra tarea de clasificación. La métrica ICM está diseñada para considerar con severidad variable los errores de clasificación entre diferentes grupos, basándose en su relación jerárquica. Esto la hace especialmente adecuada para tareas en las que los errores entre ciertas clases son inherentemente menos críticos que entre otras, como en DIPROMATS.

Aparte, ya que son métricas estándar, se reportan la Precisión (P), Recall (R) y la puntuación F1 (F1) de cada clase, para proporcionar una evaluación básica del rendimiento de los modelos en las tareas de clasificación mencionadas.

Baseline Para la primera tarea de DIPROMATS, usamos los mismos algoritmos que hemos usado en la clasificación binaria de EXIST 2022. Para las tareas 2 y 3 de DIPROMATS, y continuando con la idea de evitar usar información semántica en los modelos, usamos un clasificador multietiqueta, basado en el algoritmo de KNN (K-Nearest Neighbors).

EXIST-2023

Métrica La métrica empleada fue la oficial de la competición en el modo de evaluación soft-soft (Plaza et al., 2023), ICM-soft. ICM-soft se ha desarrollado dentro del proyecto ODESIA, y es una extensión de la métrica ICM (Amigó and Delgado, 2022) que permite evaluar un sistema bajo el paradigma "learning-with-disagreements"(LeWiDi) (Uma et al., 2021b) comparando sus salidas (dadas como probabilidades de pertenencia a una o varias clases) con un “soft gold standard” especificado de esta misma forma. Además, ICM-soft permite evaluar distintos tipos de problemas de clasificación: binaria (Tarea 1), jerárquica multiclase (Tarea 2) y jerárquica multiclase multietiqueta (Tarea 3). Es la única métrica existente que permite evaluar tareas complejas de clasificación (multilabel, jerárquica o ambas) en modo learning with

[20] https://huggingface.co/spaces/evaluate-metric/seqeval

disagreement.

Baseline para cada tarea Para establecer un baseline para cada una de las tareas de EXIST-2023 se entrenó una red neuronal simple con una única capa oculta para la clasificación de los textos. La red se entrenó a 20 epochs con un learning rate de 0.001. El algoritmo de optimización escogido fue Adam. La función de pérdida usada por la red neuronal fue la entropía binaria para las tres tareas. La arquitectura de la red neuronal constó de tres componentes principales: dos capas totalmente conectadas y una función de activación unitaria lineal rectificada (ReLu). Los textos de entrada suministrados a la red se convirtieron primeramente a vectores de 10.000 dimensiones mediante el método TF-IDF La última capa de la red neuronal produce salidas que corresponden al número de capas objetivo de cada tarea específica: 2 para la tarea 1, 4 para la tarea 2 y 6 para la tarea 3. Para obtener un baseline robusto en cada tarea, se promediaron los resultados de 10 ejecuciones distintas para cada tarea.

SQUAD/SQAC-2024

Métrica Para evaluar la tarea de SQAC-SQUAD-2024 hemos usado la misma métrica que se utilizó para evaluar tanto SQAC como SQUAD v1.1. Para calcular el F1 score, primero se hace un preprocesamiento de las predicciones y gold standard, luego cada par de respuestas (predicción-gold standard) se tokenizan y se cuenta cuantas palabras coinciden. Con este dato, se calcula el F1.

Baseline El algoritmo baseline consiste en cotejar, mediante distancia coseno cada frase del texto con la pregunta, tomando la más semejante como respuesta candidata.

MLDoc

Métrica Para la tarea de MLDoc, al ser una tarea de clasificación multi-categoría, hemos usado la métrica estándar F1 macro.

Baseline Para este conjunto de datos usamos los algoritmos ya comentados de Regresión logística, Xgboots y SVM. Al igual que en los anteriores baseline, vectorizamos evitando utilizar información lingüística.

MultiCONER 2022

Métrica Al igual que DIANN, MultiCONER es una tarea NER, por lo que hemos usado la misma métrica que para DIANN (F1 en la implementación del script Seqeval).

Baseline Para MultiCONER 2022, usamos el mismo algoritmo baseline que para DIANN, CRF.

STS 2017

Métrica STS es una tarea de similitud, y para medir esa similitud entre dos frase, usamos la métrica de Pearson Correlation.

La correlación de Pearson se calcula con la siguiente fórmula:

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

donde:
- r es el coeficiente de correlación de Pearson,
- x_i e y_i son los valores individuales de las variables X e Y,
- $\overline{x}$ y $\overline{y}$ son los promedios de las variables X e Y,
- $\sum$ denota la suma sobre todos los datos de muestra.

Baseline En el caso de STS 2017, un problema de regresión, simplemente vectorizamos las dos frases de cada caso mediante la función `TfidfVectorizer` de sklearn (Pedregosa et al., 2011), y calculamos el coseno entre ambas representaciones como aproximación de su similitud.

SQAC/SQUAD

Métrica Para evaluar la tarea de SQAC-SQUAD hemos usado una versión flexible de F1 (lenient frente a strict). Para calcular este F1 score, primero se hace un preprocesamiento de las predicciones y gold standard, luego, con cada par de predicción-gold standard, se tokenizan y se cuenta cuantas palabras coinciden. Con este dato de todos los pares, se calcula el F1.

Baseline El algoritmo baseline consiste en cotejar, mediante distancia coseno como en el caso anterior, cada frase del contexto con la pregunta, quedándose con toda la frase como respuesta candidata. Nótese que al tomar toda la frase como respuesta (para evitar cualquier tipo de proceso lingüístico), la puntuación del baseline es muy baja.

DIANN

Métrica La métrica que hemos usado para evaluar la tarea es F1, para ello hemos usado Seqeval, un framework en python que evalúa el etiquetado de secuencias

Baseline Para obtener el baseline de DIANN Detección de negación utilizamos Conditional Random Fields (CRF), una clase de métodos de modelado estadístico que se aplican a menudo en el reconocimiento de patrones. Lo hemos usado como fórmula de predicción estructurada. La idea, al igual que en las demás experimentaciones, es que no haya información semánticasen los baselines.

3.3.5. Resultados experimentales: gap en efectividad

En la Tabla 16 pueden verse los resultados de nuestra experimentación sobre las tareas del Leaderboard ODESIA EXTENDED. Disponemos de 15 mediciones del gap sobre cada una de las 15 tareas contempladas. En cada una de ellas, los baseline se han calculado como el promedio de varios algoritmos de aprendizaje supervisado sin conocimiento lingüístico, como se ha explicado anteriormente. Las columnas *mejor ES* y *mejor EN* recogen la efectividad del modelo de lenguaje con mejores resultados en cada tarea en español e inglés, respectivamente.

Es interesante observar que, a pesar de la diversidad de datasets y tareas, en todos los casos menos uno (EXIST 2023 tarea 2, donde el rendimiento en ambos idiomas es similar) se ha medido una brecha favorable al inglés, lo que proporciona una fuerte evidencia estadística de la existencia de esa brecha. Puede apreciarse, también, que hay un valor muy diferente al resto en el caso de la tarea DIPROMATS Task 3. Esa tarea es la más difícil de las contempladas en nuestra experimentación, ya que es un problema de clasificación multiclase y multilabel con 13 clases distribuidas de forma muy desigual. Al ser la más difícil, también es la tarea en la que los modelos de lenguaje aportan una mejora más sustancial respecto a las aproximaciones baseline sin conocimiento lingüístico. Aunque es un resultado en el que merece la pena profundizar, a efectos del cálculo del gap debemos descartarlo por razones estadísticas, ya que se trata de un outlier ($p < 0,01$ según el test de Grubbs[21]).

Una vez eliminado el outlier, el indicador de brecha de efectividad EF,1 se ha calculado según se describe en la sección 2.1.6 sobre las otras catorce tareas. **El resultado final (con su error estándar) es una estimación de la brecha porcentual del 20 $\pm$ 06 a favor del inglés.** Aunque hay una variación apreciable entre tareas, el error estándar nos indica que, en cualquier caso, la brecha real promedio estará en una horquilla entre el 14 % y el 26 %. Estos datos son compatibles con los obtenidos el primer año.

Esta brecha está medida exclusivamente sobre modelos discriminativos. Nuestra experimentación con modelos generativos a dado lugar a primeras estimaciones de brecha que no hemos incluido en el indicador global, por las razones que se comentan en la sección 5.4.

3.4. Evolución de la brecha de efectividad

En el primer año de proyecto se midió una brecha de efectividad de modelos discriminativos del 18 ± 04 %, y en este segundo año la medición es del 20 ± 06 %. Las dos mediciones son compatibles dentro de su error estándar y, por tanto, **no se puede hablar de variación** en este aspecto. De hecho, no han surgido nuevos modelos discriminativos destacables durante este año, de modo que la pequeña diferencia observada proviene de la mejora del diseño experimental, al que hemos añadido cuatro tareas nuevas y una búsqueda exhaustiva de hiperparámetros (con más de 3600 experimentos).

[21] Hemos utilizado https://www.graphpad.com/quickcalcs/grubbs2/

Tarea	baseline ES	Baseline EN	mejor ES	mejor EN	brecha %
Core Tasks					
EXIST-2022 Task 1	0,693	0,674	0,770	0,810	16,54
EXIST-2022 Task 2	0,463	0,437	0,570	0,580	9,803
DIPROMATS-2023 Task 1	0,750	0,707	0,818	0,819	11,60
DIPROMATS-2023 Task 2	0,220	0,210	0,471	0,550	47,80
DIPROMATS-2023 Task 3	0,090	0,080	0,267	0,472	293,33
DIANN Task 1	0,747	0,665	0,840	0,790	0,38
EXIST-2023 Task 1	0,469	0,437	0,645	0,643	9,52
EXIST-2023 Task 2	0,251	0,220	0,421	0,363	-2,70
EXIST-2023 Task 3	0,218	0,206	0,400	0,403	12,10
SQUAD/SQAC-2024	0.132	0.125	0,464	0,463	18,60
Extended Tasks					
DIANN Task 2	0,878	0,575	0,960	0,917	71,80
MLDoc	0,930	0,883	0,970	0,980	39,86
MultiCoNER 2022	0,523	0,553	0,710	0,750	4,86
STS-2017	0,680	0,707	0,800	0,860	14,76
SQUAD/SQAC	0,533	0,528	0,770	0,883	24,57
Brecha efectividad					**20 $\pm$ 06**

Tabla 16: Cálculo de la brecha de efectividad. Los baseline se han calculado como el promedio de varias algoritmos de aprendizaje supervisado sin conocimiento lingüístico. Los LLM son la efectividad del modelo de lenguaje con mejores resultados en cada tarea. El indicador de brecha de efectividad E,1.a se ha calculado según se describe en la sección 2.1.6, eliminando del promedio el outlier *DIPROMATS Task 3* y reportando el error cuadrático medio sobre el resto de mediciones. Un número positivo indica una brecha porcentual a favor del inglés, y negativo a favor del español.

La razón de que no hayan aparecido modelos discriminativos destacables es que el esfuerzo investigador (tanto en la academia como en la industria) se ha volcado en los modelos generativos. Estos modelos no son apropiados para tareas discriminativas para las que se tienen datos de entrenamiento (que son muchas de las que tienen aplicación industrial), pero han abierto un gran abanico de aplicaciones en modo no supervisado que eran impensables hace pocos años. Nuestra experimentación inicial con modelos generativos no tiene suficiente representatividad como para incluirla en el cálculo de la brecha, pero indica, provisionalmente, que **en los modelos generativos no hay una brecha mayor que la que hemos medido con los discriminativos**:

- GPT-4 (el modelo más potente junto con Claude 3 en el momento de finalizar nuestra experimentación) en modo zero-shot, aplicado a tres tareas discriminativas de nuestro leaderboard, arroja una brecha promedio del 18 % entre el español y el inglés, muy similar a nuestro resultado en modelos discriminativos.

- En nuestro dataset UNED ACCESO de preguntas de exámenes de acceso a la universidad, los modelos generativos abiertos LLama-2, Gemma y Mistral tienen una brecha promedio del 12 % entre el español y el inglés. Y los modelos más potentes (Claude-3 y GPT-4) dan una brecha ligeramente negativa (-2 % en ambos casos), es decir, se comportan un poco mejor en español que en inglés. Teniendo en cuenta que las preguntas del dataset están originalmente en español (y son traducidas manualmente al inglés), es muy posible que los modelos hayan visto parte de las respuestas en su fase de entrenamiento (contaminación), y tampoco es descartable que haya algún artefacto de traducción (aunque son traducciones manuales realizadas por profesionales). Descartando estos efectos, no

sería raro que la medición del gap estuviera en niveles parecidos a los del experimento anterior con GPT-4. En cualquier caso, necesitamos profundizar en nuestro diseño experimental para obtener cifras fiables.

4. Agregación de resultados

La Figura 3 resume los valores obtenidos para todos los indicadores obtenidos durante el primer año de proyecto, así como los valores agregados para cada uno de los ámbitos del proyecto. Los valores entre paréntesis representan los indicadores obtenidos en el año 1. Los resultados aparecen subrayados y en negrita en los casos en los que ha habido un aumento de la brecha lingüística. En las siguientes secciones describimos el proceso llevado a cabo para la agregación de indicadores en los distintos ámbitos.

Aunque, como veremos en las siguiente secciones, se dan variaciones a nivel de indicadores específicos, a nivel de ámbitos los resultados son bastante consistentes respecto del año 1. En el Ámbito 1 hay una diferencia de 2 puntos, y en los tres ámbitos restantes hay una diferencia de un punto.

4.1. Ámbito 1: Brecha en diseminación y recursos

En cuanto a diseminación y recursos, se mantiene la tendencia observada en la primera iteración del proyecto, con algunas diferencias. Según los resultados de esta iteración el factor más desfavorable es la diseminación, con una brecha en publicaciones (D.1) y proyectos subvencionados (D.2) del 98 % y 96 % respectivamente. En concreto, la brecha en proyectos subvencionados ha ascendido del 88 % al 96 %. Sin embargo, esta volatilidad puede deberse a las pocas muestras de proyectos en español respecto al inglés.

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 muy similar (R.1). Se observa un aumento importante de la 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 en campañas de evaluación permanece bastante constante, a pesar de esa diferencia en 4 puntos respecto al año anterior, si tenemos en cuenta el reducido número de muestras y el consiguiente efecto en la volatilidad del indicador R.2.b.

Al igual que en la iteración anterior, la agregación de indicadores se ha planteado como una estimación del esfuerzo necesario para construir una aplicación de Procesamiento del Lenguaje Natural (el campo de la Inteligencia Artificial que nos ocupa) en español, respecto del inglés. Para ello hemos realizado un promedio de los siguientes indicadores de brecha: D.2 (proyectos subvencionados), R.0 (corpora disponible en Internet para desarrollar modelos de lenguaje pre-entrenados), 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.

4.2. Ámbito 1: Brecha en efectividad

Para la estimación del indicador de efectividad se han empleado 15 datasets, de los cuales 10 pertenecen al leader board desarrollado en el proyecto con datos de test no públicos, y 5 de ellos suponen una extesión del estudio de la brecha con datos públicos. La Tabla 17 muestra las brechas obtenidas en cada uno de los datasets sobre los que se ha realizado el estudio.

En cuanto a tareas abstractas se refiere, para la estimación de la brecha, se ha cubierto la clasificación binaria (EXIST 2022 tarea 1, EXIST 2023 tarea 1, DIPROMATS 2023 tarea 1), clasificación multiclase, jerárquica y/o multilabel (EXIST 2022 tarea 2, EXIST 2023 tareas 2 y 3, DIPROMATS 2023 tareas 2 y 3), *learning with disagreement* (EXIST 2023, tareas 1,2 y 3), regresión (STS 2017), etiquetado de secuencias (DIANN). 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).

Los dominios cubiertos en el segundo año en el cálculo de la brecha en efectividad incluyen: geopolítica (DIPROMATS), biomedicina (DIANN), publicaciones académicas (SQUAD/SQAC), y análisis de

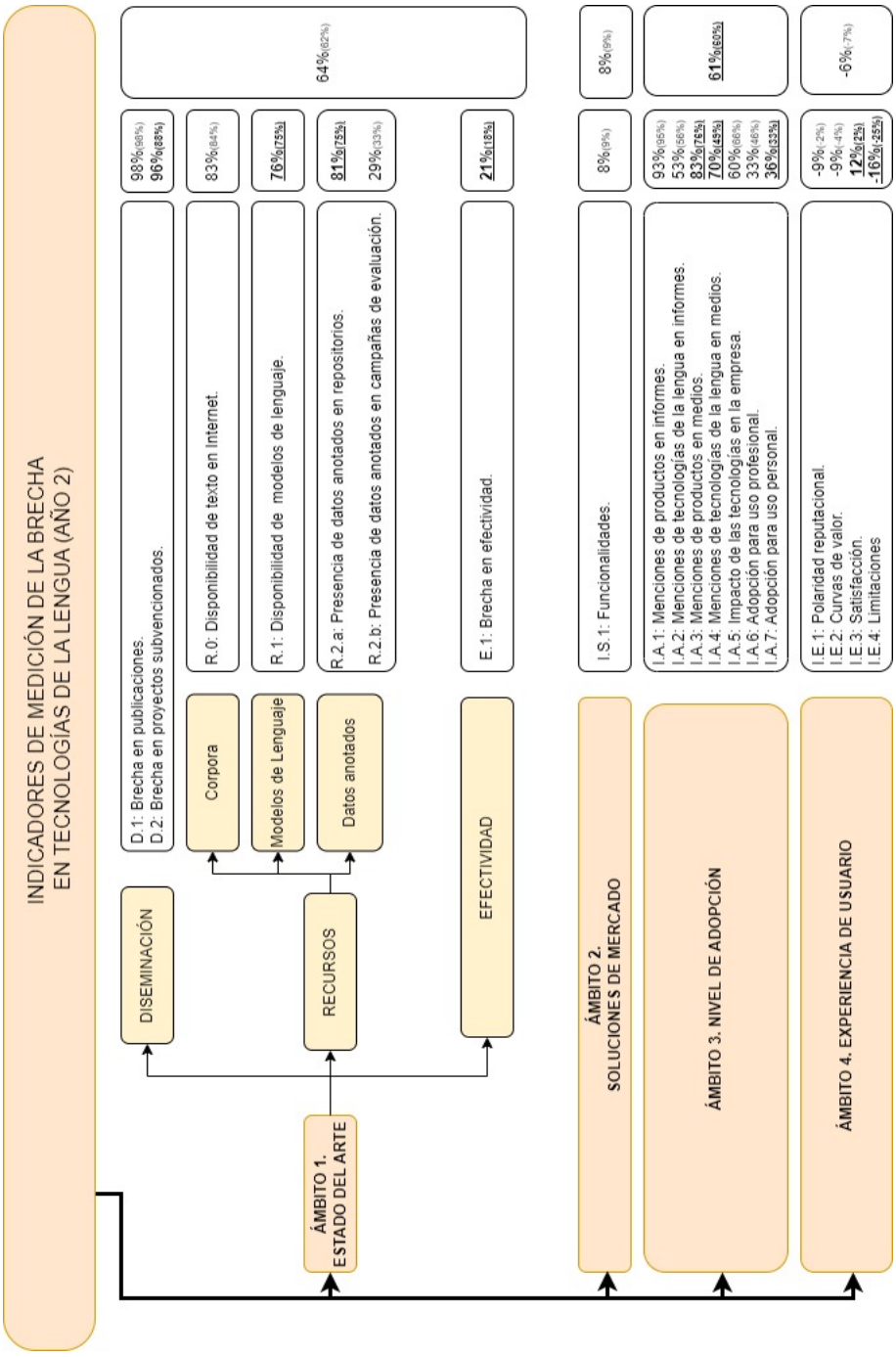

Figura 3: Estimación ODESIA de la brecha entre español e inglés, año 2

redes sociales, en concreto, identificación de sexismo (EXISTS). Además, para la experimentación se han incluido tres datasets adicionales de dominio público que incluyen textos de dominio periodístico (MLDOC), conocimiento enciclopédico y consultas en buscadores (MULTICONER) y resolución de similitud textual (STS).

Tabla 17: Descripción de tareas y modelos.

TAREA	Mejor LLM ES	Tipo de Tarea	Brecha
CORE TASKS			
EXIST 2022 Task 1	dccuchile/bert-base-spanish-wwm-cased	clasificación binaria	17 %
EXIST 2022 Task 2	PlanTL-GOB-ES/roberta-large-bne	clasificación multiclase	10 %
DIPROMATS 2023 Task 1	dccuchile/bert-base-spanish-wwm-cased	Clasificación binaria	11 %
DIPROMATS 2023 Task 2	xlm-roberta-large	clasificación multiclase multilabel	48 %
DIPROMATS 2023 Task 3	xlm-roberta-large	Clasificación multiclase multilabel	293 %
DIANN 2023 Task 1	PlanTL-GOB-ES/roberta-base-bne	Etiquetado de secuencias	0.4 %
SQUAD/SQAC 2024	PlanTL-GOB-ES/roberta-large-bne	Etiquetado	19 %
EXIST 2023 Task 1	xlm-roberta-large	clasificación binaria learning with disagreement	10 %
EXIST 2023 Task 2	xlm-roberta-large	clasificación jerárquica multiclase learning with disagreement	-3 %
EXIST 2023 Task 3	xlm-roberta-large	clasificación multiclase multilabel jerarquica learning with disagreement	12 %
EXTENDED TASKS			
DIANN Task 2	dccuchile/bert-base-spanish-wwm-cased	etiquetado de secuencias	72 %
MLDoc	PlanTL-GOB-ES/roberta-base-bne	clasificación multiclase	40 %
MultiCoNER CoNLL 2022	xlm-roberta-large	etiquetado de secuencias	5 %
STS-2017	dccuchile/bert-base-spanish-wwm-cased	regresión	15 %
SQUAD/SQAC 2016	xlm-roberta-large	etiquetado de secuencias	25 %

Es interesante observar que, a pesar de la diversidad de datasets y tareas, en todos los casos menos uno (EXIST 2023 tarea 2, donde el rendimiento en ambos idiomas es similar) se ha medido una brecha favorable al inglés, lo que proporciona una fuerte evidencia estadística de la existencia de esa brecha. Puede apreciarse, también, que hay un valor muy diferente al resto en el caso de la tarea DIPROMATS Task 3. Esa tarea es la más difícil de las contempladas en nuestra experimentación, ya que es un problema de clasificación multiclase y multilabel con 13 clases distribuidas de forma muy desigual. Al ser la más difícil, también es la tarea en la que los modelos de lenguaje aportan una mejora más sustancial respecto a las aproximaciones baseline sin conocimiento lingüístico. Aunque es un resultado en el que merece la pena profundizar, a efectos del cálculo del gap lo hemos descartado por razones estadísticas, ya que se trata de un outlier ($p < 0,01$ según el test de Grubbs).

Una vez eliminado el outlier, el indicador de brecha de efectividad EF,1 se ha calculado según se describe en la sección 2.1.6 sobre las otras catorce tareas. **El resultado final (con su error estándar) es una estimación de la brecha porcentual del 20 $\pm$ 06 a favor del inglés.** Aunque hay una variación apreciable entre tareas, el error estándar nos indica que, en cualquier caso, la brecha real promedio estará en una horquilla entre el 14 % y el 26 %. Estos datos son compatibles con los obtenidos el primer año.

Esta brecha está medida exclusivamente sobre modelos discriminativos. Nuestra experimentación con modelos generativos ha dado lugar a primeras estimaciones de brecha que no hemos incluido en el indicador global. Nuestra experimentación inicial con modelos generativos no tiene suficiente representatividad como para incluirla en el cálculo de la brecha, pero indica, provisionalmente, que **en los modelos generativos no hay una brecha mayor que la que hemos medido con los discriminativos**:

- GPT-4 (el modelo más potente junto con Claude 3 en el momento de finalizar nuestra experimentación) en modo zero-shot, aplicado a tres tareas discriminativas de nuestro leaderboard, arroja una brecha promedio del 18 % entre el español y el inglés, muy similar a nuestro resultado en modelos discriminativos.

- En nuestro dataset UNED ACCESO de preguntas de exámenes de acceso a la universidad, los modelos generativos abiertos LLama-2, Gemma y Mistral tienen una brecha promedio del 12 % entre

el español y el inglés. Y los modelos más potentes (Claude-3 y GPT-4) dan una brecha ligeramente negativa (-2 % en ambos casos), es decir, se comportan un poco mejor en español que en inglés. Teniendo en cuenta que las preguntas del dataset están originalmente en español (y son traducidas manualmente al inglés), es muy posible que los modelos hayan visto parte de las respuestas en su fase de entrenamiento (contaminación), y tampoco es descartable que haya algún artefacto de traducción (aunque son traducciones manuales realizadas por profesionales). Descartando estos efectos, no sería raro que la medición del gap estuviera en niveles parecidos a los del experimento anterior con GPT-4. En cualquier caso, necesitamos profundizar en nuestro diseño experimental para obtener cifras fiables.

4.3. Conclusiones

En este segundo año de proyecto 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.

Entre el primer año y el segundo se ha producido la irrupción definitiva de la IA generativa, y esperábamos ver cambios en los indicadores analizados. Sin embargo, aunque la investigación, el desarrollo y la implantación se han acelerado enormemente desde la aparición de ChatGPT el 30 de noviembre de 2022, parecen haberlo hecho de forma paralela en el mundo angloparlante y en el hispanoparlante, de tal manera que la brecha se ha mantenido, promediando sobre todos los ámbitos de análisis, similar a la que había antes de ChatGPT.

Es destacable el esfuerzo realizado para la medición de la brecha en efectividad de los modelos de lenguaje: se han desarrollado en el ámbito del proyecto nueve datasets, se ha experimentando con 14 modelos de lenguaje discriminativos abiertos, tres modelos generativos propietarios (Claude 3, GPT-4 y GPT-3.5) y tres modelos generativos abiertos (Mistral, Llama-2 y Gemma). En la experimentación se han realizado más de 3,500 procesos de fine-tuning y evaluación sobre cada una de las tareas discriminativas en cada uno de los idiomas. Este esfuerzo sienta las bases para un seguimiento continuado de los modelos de lenguaje aplicables al español, además de servir para estimar la brecha inglés-español.

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.

Bibliografía

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. ArXiv:2309.07445 [cs].

Enrique Amigó and Agustín Delgado. 2022. Evaluating extreme hierarchical multi-label classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume Volume 1: Long Papers, page 5809–5819, Dublin, Ireland. ACL.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Percy Liang et al. 2023. Holistic evaluation of language models.

Hermenegildo Fabregat, Juan Martínez-Romo, and Lourdes Araujo. 2018. Overview of the DIANN task: Disability annotation task. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 1–14. CEUR-WS.org.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. ArXiv:2009.03300 [cs].

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). In *Proceedings of the 16th international workshop on semantic evaluation (SemEval-2022)*, pages 1412–1437.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Laura Plaza, Jorge Carrillo de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization. Experimental IR Meets Multilinguality, Multimodality, and Interaction. In *Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Aarohi et al Srivastava. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. ArXiv:2206.04615 [cs, stat].

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. ArXiv:2210.09261 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021b. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. ArXiv:2304.06364 [cs].