Métodos Eficientes para la Generación de Lenguaje Natural: Aplicaciones a la simplificación lingüística y al pronóstico multilingüe

A. Vivel-Couso

Directores: J.M. Alonso-Moral, A. Bugarín-Diz
Fecha de inicio: 01/03/2024

Centro Singular de Investigacion en Tecnoloxías Intelixentes (CiTIUS)
Departamento de Electrónica y Computación
Universidade de Santiago de Compostela
Santiago de Compostela, Spain

{ainhoa.vivel.couso, josemaria.alonso.moral, alberto.bugarin.diz}@usc.es

Resumen Esta investigación aborda el desarrollo de sistemas de generación de lenguaje natural eficientes desde el punto de vista computacional y aplicables a lenguas de bajos recursos. Para ello, se estudian estrategias de orquestación de modelos orientadas a optimizar el compromiso entre calidad textual y consumo energético en tareas de generación textual. Paralelamente, se desarrolla MeteoGalEus, un recurso multilingüe alineado en gallego, euskera y español destinado a facilitar la investigación en generación automática de pronósticos meteorológicos en escenarios multilingües.

1. Resumen de la investigación

La presente tesis doctoral se enmarca en el ámbito del Procesamiento del Lenguaje Natural (PLN), con especial atención a dos líneas de investigación complementarias: (i) la optimización de sistemas de PLN desde una perspectiva de eficiencia computacional y sostenibilidad energética, y (ii) el desarrollo de recursos y modelos multilingües para lenguas de bajos recursos.

En este contexto, el objetivo principal es desarrollar métodos y recursos de PLN multilingüe y eficiente que permitan mejorar el rendimiento, la robustez y la sostenibilidad de los sistemas de PLN, con especial énfasis en lenguas de pocos recursos (en particular, el gallego). Para alcanzar este objetivo general, se plantean los siguientes objetivos específicos:

- Diseñar y desarrollar técnicas eficientes de simplificación textual mediante estrategias de orquestación de modelos basadas en LLM, incluyendo mecanismos de *routing* y *cascading*, con el objetivo de optimizar el equilibrio entre la calidad de los resultados y la eficiencia computacional. Se aprovecharán lenguas con una amplia disponibilidad de recursos lingüísticos, como el inglés y el español, para servir de base a la investigación.

- Analizar la evaluación de sistemas de simplificación textual mediante estudios de evaluación humana, estudiando su correlación con métricas automáticas y con la eficiencia energética de los modelos.
- Construir un conjunto de datos de simplificación textual orientado a lenguas de bajos recursos y en un contexto en el que, hasta la fecha, no existen recursos ni benchmarks disponibles para esta tarea en dicha lengua, y reproducir estudios previos de simplificación eficiente con el fin de comparar los resultados.
- Desarrollar un dataset multilingüe para tareas de generación de lenguaje en el dominio meteorológico multilingüe basado en la integración de fuentes meteorológicas regionales.
- Crear sistemas PLN para pronóstico meteorológico multilingüe, explorando estrategias de orquestación eficiente de LLMs y técnicas de transferencia de conocimiento para mejorar la eficiencia y la robustez del sistema en escenarios multilingües y de baja disponibilidad de recursos.

2. Metodología y plan de trabajo

A la fecha de redacción de este documento, la investigación doctoral se encuentra en una fase avanzada. Varios de los subobjetivos planteados ya han sido alcanzados, mientras que el resto continúa desarrolándose conforme a la planificación establecida. En este contexto, se han obtenido resultados relevantes en el diseño de estrategias de simplificación textual eficientes desde el punto de vista computacional y energético, así como en la construcción de recursos multilingües especializados para el dominio meteorológico.

En esta fase, la investigación se centra en varias líneas de trabajo complementarias que permitirán consolidar las principales aportaciones de la tesis. Por una parte, se está llevando a cabo una evaluación humana de sistemas de simplificación textual con el objetivo de analizar la relación entre las valoraciones de los usuarios, las métricas automáticas empleadas habitualmente y los indicadores de eficiencia energética. Este estudio permitirá comprender con mayor profundidad los compromisos existentes entre la calidad de la simplificación y el consumo energético.

Paralelamente, se está desarrollando el primer conjunto de datos para simplificación textual en gallego, un recurso destinado a impulsar la investigación en este ámbito y a facilitar la reproducibilidad y la comparación de futuras propuestas para lenguas con recursos limitados. Asimismo, se continúa avanzando en sistemas de procesamiento del lenguaje natural para la generación automática de predicciones meteorológicas multilingües, explorando técnicas de transferencia de conocimiento orientadas a mejorar tanto la eficiencia como la capacidad de generalización de estos sistemas.

Durante los próximos meses se completarán los experimentos pendientes, se realizará un análisis integral de los resultados obtenidos y se contrastarán las hipótesis de investigación planteadas. Asimismo, está prevista la publicación de las contribuciones más relevantes en revistas y congresos internacionales de

referencia, con el fin de favorecer su difusión y su impacto en la comunidad científica.

De acuerdo con la planificación prevista, se prevé completar todos los objetivos científicos de la tesis durante el primer trimestre de 2027. A partir de ese momento, la fase final del doctorado estará dedicada principalmente a la integración de los resultados obtenidos, la redacción de la memoria doctoral, la incorporación de las revisiones derivadas del proceso de supervisión y la preparación de la defensa. Esta se prevé para diciembre de 2027, siempre que el desarrollo de la investigación continúe conforme a lo previsto y se alcancen satisfactoriamente los objetivos científicos definidos. En cualquier caso, la fecha límite para su defensa es marzo de 2028.

A continuación, se presenta un resumen de los dos trabajos científicos derivados de esta investigación doctoral y finalizados durante 2026.

2.1. Framework de orquestación sensible a la complejidad para la simplificación textual eficiente en términos energéticos

En "A Complexity-Aware Orchestration Framework for Energy-Efficient Text Simplification" (actualmente en revisión), proponemos un marco de orquestación que combina diferentes tamaños de modelos de lenguaje de gran escala (LLM) para la simplificación textual adaptativa, incorporando estrategias de *routing* ($\vee$, Figura 1) y *cascading* ($\rightarrow$, Figura 2). Los sistemas de simplificación textual se han evaluado tradicionalmente principalmente en términos de la calidad de la salida, mientras que su coste computacional ha sido en gran medida desatendido. En este contexto, nuestro enfoque se basa en la idea de que las entradas más simples pueden ser procesadas por modelos más ligeros, mientras que las más complejas se benefician del uso de modelos de mayor capacidad.

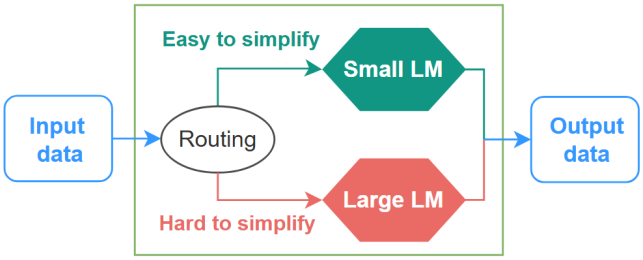


Figura 1: Descripción general del marco de orquestación adaptativa eficiente de LLM para la estrategia de *routing*. El sistema toma un texto de entrada en bruto y lo enruta en función de la dificultad estimada de su simplificación. Las instancias más sencillas son procesadas por un modelo de lenguaje pequeño (SLM), mientras que las más complejas son gestionadas por un modelo de lenguaje grande (LLM). Posteriormente, las salidas se combinan para producir el texto final simplificado.

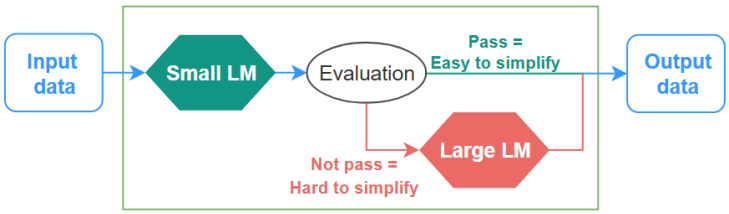


Figura 2: Descripción general del marco de orquestación adaptativa eficiente de LLM para la estrategia de *cascading*. El sistema toma un texto de entrada en bruto y, en primer lugar, genera una posible simplificación utilizando un modelo de lenguaje pequeño (SLM). A continuación, un módulo de evaluación analiza la calidad de la salida. Si se considera suficiente, se devuelve como resultado final. En caso contrario, la entrada se envía a un modelo de lenguaje grande (LLM) para su procesamiento adicional, y la salida del LLM se utiliza como texto final simplificado.

Decidimos utilizar conjuntos de datos basados en Wikipedia debido a que Wikipedia es el recurso predominante para la simplificación textual en inglés, gracias a la existencia de dos versiones en este idioma: la Wikipedia en inglés estándar y la Wikipedia en inglés simple. En concreto, utilizamos Wiki-Auto [2]. Este conjunto de datos contiene aproximadamente 488k pares de oraciones alineadas automáticamente. En comparación con otros recursos como WikiLarge [4], presenta una mayor calidad debido a una mejor alineación de los pares oración compleja–simple. Dividimos aleatoriamente el conjunto de datos (n=488,332) en subconjuntos de entrenamiento (n=341,832; 70 %), validación (n=73,250; 15 %) y prueba (n=73,250; 15 %). Debido al alto coste computacional de la evaluación, finalmente realizamos las pruebas utilizando los primeros 1,000 elementos del subconjunto de prueba, lo que preserva la distribución original de ejemplos fáciles y difíciles de simplificar.

En este trabajo evaluamos las distintas estrategias desde dos dimensiones complementarias: métricas automáticas de calidad y eficiencia energética, analizando el compromiso entre la calidad de la simplificación y el consumo energético con el objetivo de identificar configuraciones óptimas de Pareto (Figura 3). Los resultados muestran que la orquestación mejora el rendimiento en escenarios zero-shot (ZS), pero incrementa el consumo energético, mientras que el fine-tuning (FT) reduce tanto los costes de entrenamiento como de inferencia, obteniendo en general mejores resultados (Tabla 1).

Este estudio evidencia el compromiso entre calidad de simplificación y eficiencia computacional, proponiendo la orquestación de modelos como una alternativa eficiente frente al uso de un único modelo estático.

El trabajo introduce un marco de simplificación textual eficiente basado en señales de dificultad a nivel de instancia para guiar el *routing* y el *cascading* entre modelos de distinta escala. Los resultados muestran que, aunque el FT implica un coste adicional de entrenamiento frente a enfoques ZS, este se compensa

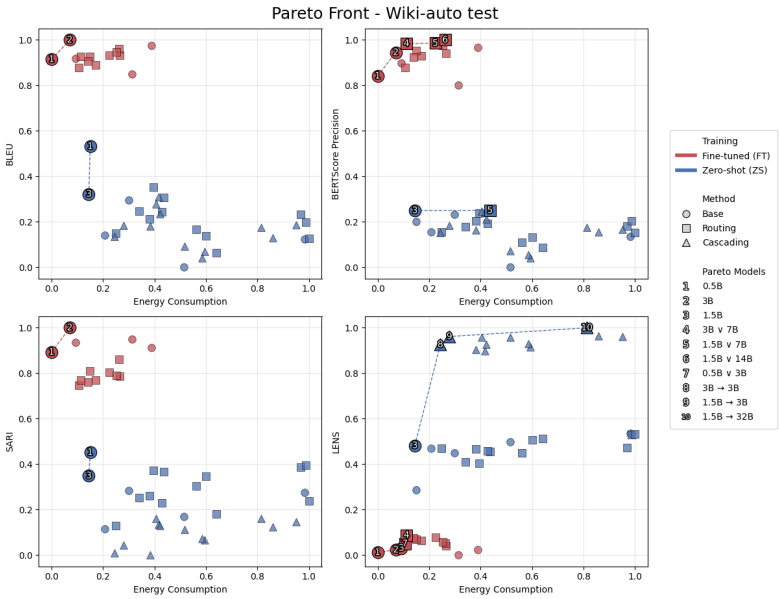


Figura 3: Comparación del frente de Pareto (FT vs ZS) a través de métricas de evaluación en el conjunto de prueba Wiki-Auto.

con mejoras consistentes en calidad (BLEU, BERTScore y SARI) y una mayor eficiencia en inferencia en Wiki-Auto.

En la práctica, los modelos FT ofrecen mejores compromisos energía-calidad, destacando variantes pequeñas como FT 0.5B, que igualan o superan a modelos mayores con menor coste computacional, lo que sugiere que la adaptación a la tarea puede sustituir parcialmente el escalado de modelos.

Además, las estrategias de orquestación, especialmente el *routing*, permiten reducir el coste energético manteniendo buena calidad y ocupando regiones competitivas del frente de Pareto. En contraste, el *cascading* mejora aspectos cualitativos como la simplicidad, pero con mayor coste en inferencia y beneficios menos consistentes en métricas de solapamiento.

Aunque la evaluación presentada en el estado actual de la investigación se basa principalmente en métricas automáticas, se reconoce que estas no capturan de forma completa todos los aspectos de la calidad de los modelos. Esto se puede observar examiando los textos generados por los modelos situados en el frente de Pareto (Tabla 2). Por ello, trabajos futuros incorporarán evaluación humana con el fin de proporcionar una valoración más integral de las salidas generadas, incluyendo dimensiones cualitativas como la fluidez, la utilidad y la robustez.

Modelo	Configuración	BLEU	BERTScore	SARI	LENS	EC_t	EC_i
0.5B	ZS	27.43	90.44	42.74	42.10	—	0.65
	FT	33.81	91.65	46.34	32.77	662.94	0.22
1.5B	ZS	23.94	90.53	41.90	48.69	—	0.63
	FT	33.87	91.76	46.69	33.41	657.76	0.49
3B	ZS	20.92	90.35	39.98	48.27	—	0.81
	0.5B ∨ 3B (ZS)	+8.37 %	+0.06 %	+2.83 %	-4.14 %	—	+48.15 %
	1.5B ∨ 3B (ZS)	+5.78 %	+0.10 %	+2.98 %	-0.15 %	—	+61.73 %
	3B ∨ 3B (ZS)	+0.62 %	0.00	+0.33 %	+0.02 %	—	+14.81 %
	0.5B → 3B (ZS)	+3.25 %	+0.02 %	-2.35 %	+30.56 %	—	+61.73 %
	1.5B → 3B (ZS)	+3.39 %	+0.07 %	-1.43 %	+34.49 %	—	+25.93 %
	3B → 3B (ZS)	-0.48 %	0.00	-2.15 %	+32.03 %	—	+13.58 %
	FT	35.23	91.85	47.22	33.14	403.26	0.42
	0.5B ∨ 3B (FT)	-5.71 %	-0.14 %	-4.36 %	+2.81 %	-27.12 %	+23.81 %
	1.5B ∨ 3B (FT)	-3.46 %	+0.01 %	-3.28 %	+4.89 %	-23.18 %	+54.76 %
	3B ∨ 3B (FT)	-4.34 %	-0.04 %	-4.13 %	+5.25 %	-4.80 %	+47.62 %
7B	ZS	23.48	90.50	41.37	47.57	—	1.08
	0.5B ∨ 7B (ZS)	+4.05 %	+0.01 %	+1.74 %	-3.24 %	—	+25.00 %
	1.5B ∨ 7B (ZS)	+0.94 %	+0.03 %	+1.62 %	+0.46 %	—	+36.11 %
	3B ∨ 7B (ZS)	-3.62 %	-0.09 %	-1.09 %	+0.63 %	—	+33.33 %
	0.5B → 7B (ZS)	+1.15 %	+0.02 %	-2.97 %	+32.04 %	—	+30.56 %
	1.5B → 7B (ZS)	-1.19 %	+0.03 %	-2.47 %	+36.24 %	—	+27.78 %
	3B → 7B (ZS)	-4.13 %	-0.04 %	-3.09 %	+34.01 %	—	+32.41 %
	FT	32.75	91.58	46.80	32.38	658.09	1.12
	0.5B ∨ 7B (FT)	+2.02 %	+0.26 %	-3.12 %	+6.55 %	-20.40 %	-36.61 %
	1.5B ∨ 7B (FT)	+4.18 %	+0.37 %	-2.50 %	+8.18 %	-17.99 %	-23.21 %
	3B ∨ 7B (FT)	+3.82 %	+0.37 %	-3.10 %	+9.17 %	-6.73 %	-51.79 %
14B	ZS	18.58	91.89	46.50	33.19	—	1.33
	0.5B ∨ 14B (ZS)	+14.91 %	+0.23 %	+2.75 %	-3.35 %	—	+8.28 %
	1.5B ∨ 14B (ZS)	+12.49 %	+0.28 %	+3.59 %	+0.61 %	—	+14.79 %
	3B ∨ 14B (ZS)	+5.60 %	+0.19 %	+0.27 %	+0.89 %	—	+21.89 %
	0.5B → 14B (ZS)	+6.14 %	+0.09 %	-2.08 %	+28.68 %	—	+13.61 %
	1.5B → 14B (ZS)	+8.18 %	+0.16 %	-1.16 %	+31.58 %	—	+0.59 %
	3B → 14B (ZS)	+3.66 %	+0.11 %	-1.95 %	+29.71 %	—	+12.43 %
	FT	34.82	91.94	46.16	33.41	1137.54	1.43
	0.5B ∨ 14B (FT)	-2.01 %	-0.05 %	-2.19 %	+1.75 %	-30.09 %	-26.32 %
	1.5B ∨ 14B (FT)	-0.66 %	+0.07 %	-0.90 %	+3.01 %	-28.69 %	-27.07 %
	3B ∨ 14B (FT)	-1.35 %	+0.02 %	-2.15 %	+3.59 %	-22.18 %	-29.32 %
32B	ZS	20.62	90.32	41.29	50.52	—	3.04
	0.5B ∨ 32B (ZS)	+8.87 %	+0.09 %	+2.23 %	-4.28 %	—	-1.32 %
	1.5B ∨ 32B (ZS)	+6.06 %	+0.13 %	+2.37 %	-0.49 %	—	+0.33 %
	3B ∨ 32B (ZS)	+0.39 %	+0.03 %	-0.75 %	-0.16 %	—	+1.64 %
	0.5B → 32B (ZS)	+5.19 %	+0.06 %	-2.52 %	+28.50 %	—	-3.29 %
	1.5B → 32B (ZS)	+4.17 %	+0.08 %	-2.23 %	+31.08 %	—	-16.12 %
	3B → 32B (ZS)	+0.53 %	+0.03 %	-2.98 %	+28.76 %	—	-11.84 %

Tabla 1: Evaluación en Wiki-Auto

2.2. MeteGalEus: Un conjunto de datos meteorológicos multilingüe ibérico en gallego, euskera y español

En "MeteoGalEus: An Iberian Multilingual Weather Dataset in Galician, Euskera, and Spanish" [3] se presenta un conjunto de datos multilingüe de predicción meteorológica que integra observaciones procedentes de dos agencias meteorológicas regionales españolas, Euskalmet y MeteoGalicia. El conjunto de da-

Configuración	Texto original en inglés	Traducción al español
Original	Although captive raccoons have been known to live over 20 years, their life expectancy in the wild is only 1.8 to 3.1 years.	Aunque los mapaches en cautiverio pueden vivir más de 20 años, su esperanza de vida en estado salvaje es de solo entre 1,8 y 3,1 años.
Referencia	In the wild, they usually live only 1-3 years	En estado salvaje, suelen vivir solo entre 1 y 3 años.
0.5B ZS	Although captive raccoons have been known to live over 20 years, their life expectancy in the wild is only 1.8 to 3.1 years.	Aunque los mapaches en cautiverio pueden vivir más de 20 años, su esperanza de vida en estado salvaje es de solo entre 1,8 y 3,1 años.
0.5B FT	Captive raccoons have been known to live over 20 years.	Los mapaches en cautiverio pueden vivir más de 20 años.
1.5B∨3B FT	Captive raccoons have been known to live over 20 years, but their life expectancy in the wild is only 1.8 to 3.1 years.	Los mapaches en cautiverio pueden vivir más de 20 años, pero su esperanza de vida en estado salvaje es de solo entre 1,8 y 3,1 años.

Tabla 2: Texto original vs. referencia humana vs. salidas de los modelos en el frente de Pareto. Ejemplo de la línea 730 del conjunto de prueba Wiki-Auto.

tos contiene registros diarios que abarcan un periodo de 4 años y 3 meses, con observaciones alineadas entre ambas fuentes. MeteoGalEus incluye variables meteorológicas clave como temperatura, viento y estado del cielo. El conjunto de datos está disponible públicamente en el repositorio oficial del proyecto[1] y en la plataforma Hugging Face[2].

El conjunto de datos se proporciona en un formato estructurado, lo que facilita su análisis e integración, e incluye predicciones textuales disponibles en las lenguas oficiales de cada región (gallego y español en el caso de MeteoGalicia; euskera y español en el caso de Euskalmet). Mediante la fusión y armonización de datos procedentes de ambas agencias, MeteoGalEus constituye un recurso único para el análisis meteorológico interregional y el estudio del clima en contextos multilingües.

MeteoGalEus abarca un periodo de cuatro años y tres meses, cubriendo un total de 1557 días entre el 25 de noviembre de 2020 y el 1 de marzo de 2025. Para cada fecha, el conjunto de datos incluye entradas procedentes de dos regiones, garantizando una representación equilibrada: la mitad de los registros proviene de Galicia y la otra mitad del País Vasco, lo que da lugar a un total de 3084 entradas meteorológicas.

El conjunto de datos incluye únicamente aquellas fechas presentes en ambas fuentes, lo que garantiza la alineación temporal de los registros. Cada entrada

[1] https://gitlab.citius.gal/ainhoa.vivel/meteogaleus/
[2] https://huggingface.co/datasets/ainhoa-vivel/MeteoGalEus

contiene variables meteorológicas clave como temperatura, velocidad y dirección del viento, y estado del cielo para entre 33 y 36 ciudades en cada una de las dos regiones. Las predicciones se proporcionan en español junto con una de las dos lenguas cooficiales (gallego o euskera), dependiendo de la región de origen. Mediante la armonización y estandarización de observaciones procedentes de dos agencias meteorológicas regionales, MeteoGalEus facilita la investigación meteorológica interregional y multilingüe, habilitando aplicaciones en análisis de tendencias climáticas, predicción meteorológica y procesamiento del lenguaje natural multilingüe.

El conjunto de datos MeteoGalEus abarca dos regiones de España con contextos lingüísticos y culturales diferenciados. Por parte de MeteoGalicia, incluye predicciones en español y gallego correspondientes a la comunidad autónoma de Galicia, mientras que desde Euskalmet incorpora información en español y euskera para el País Vasco. Al incluir la lengua oficial del Estado (español) junto con la lengua cooficial de cada región, el conjunto de datos refleja el carácter multilingüe natural de la comunicación meteorológica en estos territorios. Aunque las predicciones son bilingües, los datos meteorológicos se proporcionan exclusivamente en español, y todas las claves de los archivos JSON están estandarizadas en inglés para garantizar la consistencia y facilitar el procesamiento computacional.

El conjunto de datos MeteoGalEus se construyó mediante un proceso multietapa diseñado para garantizar la consistencia y la comparabilidad entre fuentes. En primer lugar, se recopilaron datos meteorológicos diarios procedentes de dos agencias oficiales regionales: MeteoGalicia (que cubre Galicia, con predicciones en español y gallego) y Euskalmet (que cubre el País Vasco, con predicciones en español y euskera). Se descartó aquella información que no estaba presente en ambos conjuntos de datos, limitando el corpus a los días para los cuales ambas agencias proporcionaban registros, lo que garantiza que el conjunto final contenga únicamente observaciones sincronizadas entre las dos regiones.

Tras la recopilación, se realizó un proceso de armonización de los valores procedentes de Euskalmet y MeteoGalicia hacia un formato común. Mientras que Euskalmet proporciona predicciones horarias para varias variables, MeteoGalicia ofrece datos organizados en tres franjas temporales diarias: mañana, tarde y noche. Por ello, se convirtieron los datos horarios de Euskalmet a estas mismas franjas temporales, teniendo en cuenta los cambios de horario estacional[3].

Además, MeteoGalicia proporciona observaciones para 33 localizaciones dentro de Galicia, mientras que Euskalmet incluye 319 localizaciones en el País Vasco. Fue necesario identificar un subconjunto comparable de ubicaciones en ambos conjuntos de datos. Para ello, se eliminaron aquellas localizaciones sin datos observacionales y se seleccionaron 36 localizaciones de Euskalmet que representan las tres principales zonas del País Vasco.

[3] El horario de invierno utiliza mañana (7:00–14:00), tarde (14:00–20:00) y noche (20:00–7:00), mientras que el horario de verano utiliza mañana (6:00–14:00), tarde (14:00–21:00) y noche (21:00–6:00).

Asimismo, se estandarizaron los nombres de las zonas en los datos de Euskalmet, pasando de denominaciones geográficas (por ejemplo, zona costera) a denominaciones provinciales (por ejemplo, Bizkaia o Gipuzkoa, según la localización). En el caso de la velocidad del viento, los valores fueron discretizados en cinco categorías[4]. En cuanto a la dirección del viento, Euskalmet utiliza un sistema de 16 puntos cardinales, mientras que MeteoGalicia emplea un sistema de 8 direcciones. Por ello, se unificó la representación utilizando el sistema de 8 puntos cardinales en ambos conjuntos de datos.

Finalmente, todos los datos se integraron en un único formato JSON estructurado, preservando tanto las variables meteorológicas comunes como las descripciones textuales multilingües generadas por meteorólogos humanos. Este proceso de alineación permite el análisis directo interregional e interlingüístico, manteniendo al mismo tiempo la riqueza y autenticidad de los registros originales.

Este conjunto de datos es adecuado para tareas que requieren datos meteorológicos alineados, de alta calidad y estandarizados en múltiples idiomas y regiones. Se han realizado experimentos base utilizando modelos basados en LLaMA en escenarios zero-shot y fine-tuning, con el objetivo de ilustrar el uso de MeteoGalEus en tareas de generación de lenguaje natural (NLG). El ajuste fino (fine-tuning) produce mejoras consistentes en todas las métricas evaluadas, con incrementos de BERTScore de 0.68 a 0.79, ROUGE de 0.20 a 0.35 y BLEU de 0.02 a 0.17 en el mejor modelo. Los resultados experimentales muestran que MeteoGalEus puede utilizarse como un benchmark para tareas de generación de lenguaje multilingüe y en contextos meteorológicos interregionales.

3. Relevancia

Esta investigación tiene una relevancia significativa tanto a nivel científico como ambiental. Desde una perspectiva científica, contribuye a la expansión del Procesamiento del Lenguaje Natural (PLN) hacia lenguas infrarepresentadas, así como a la comprensión del impacto del multilingüismo en tareas de generación y predicción de lenguaje. Asimismo, introduce un enfoque sistemático para el análisis de la eficiencia computacional y energética en sistemas de generación de lenguaje, con especial énfasis en tareas de simplificación textual, donde el equilibrio entre calidad lingüística, coste computacional y robustez del sistema resulta especialmente crítico.

Desde una perspectiva ambiental, este trabajo se enmarca en los principios de la *Green AI*, abordando el elevado coste energético asociado al entrenamiento e inferencia de modelos de lenguaje de gran escala. Este problema es particularmente relevante en el contexto actual de despliegue masivo de modelos basados en arquitecturas profundas, donde el consumo energético no se limita a la fase de entrenamiento, sino que se extiende de forma continua durante la inferencia

[4] Valores de velocidad del viento: *calma* ($\leq$ 5 km/h), *viento débil* ($\leq$ 20 km/h), *viento moderado* ($\leq$ 40 km/h), *viento fuerte* ($\leq$ 70 km/h) y *viento muy fuerte* ($>$ 70 km/h).

en producción, que constituye el principal componente del ciclo de vida computacional de estos sistemas.

En este marco, se estudian estrategias de orquestación de modelos y mecanismos de transferencia de conocimiento que permiten reducir la dependencia de modelos de alta complejidad, activándolos únicamente cuando la dificultad del *input* lo requiere. Este paradigma de selección dinámica (routing y cascading de modelos) permite reducir el consumo computacional agregado, optimizando el uso de recursos tanto en entrenamiento como, especialmente, en inferencia, donde se concentra la carga operativa real de los sistemas.

Este tipo de enfoques tiene implicaciones directas en la reducción de la huella energética asociada a sistemas de PLN, al disminuir el uso innecesario de cómputo intensivo y mejorar la eficiencia en la utilización de infraestructuras basadas en GPU y TPU. Además, contribuye a una utilización más sostenible del hardware especializado, prolongando su vida útil y reduciendo la necesidad de escalado continuo de infraestructura para soportar modelos de gran tamaño.

En paralelo, el entrenamiento de modelos de inteligencia artificial tiene implicaciones tanto económicas como medioambientales [1], contribuyendo a varios de los 17 Objetivos de Desarrollo Sostenible[5] de las Naciones Unidas, entre los que destacan la Energía asequible y no contaminante (Objetivo 7), la Producción y consumo responsables (Objetivo 12) y la Acción por el clima (Objetivo 13). Estas consideraciones reflejan la necesidad de desarrollar métodos más eficientes que reduzcan tanto el coste computacional como el impacto ambiental asociado.

El entrenamiento de modelos de IA requiere habitualmente una elevada capacidad computacional y un consumo energético significativo, especialmente en modelos basados en redes neuronales profundas. Este consumo se ve amplificado por la dependencia de fuentes de energía no renovables, lo que contribuye a la huella de carbono asociada al desarrollo de sistemas de inteligencia artificial. A ello se suma el coste económico del uso de hardware especializado (GPU y TPU) y recursos de computación en la nube, que condiciona la escalabilidad y accesibilidad de estos sistemas.

En este contexto, técnicas como el aprendizaje por transferencia han demostrado ser efectivas para reducir la necesidad de entrenamiento desde cero mediante el aprovechamiento de modelos preentrenados, disminuyendo así tanto el coste computacional como el energético. De forma complementaria, la investigación en inteligencia artificial eficiente se centra en el desarrollo de modelos que mantengan un rendimiento competitivo con menores requisitos de recursos, consolidando el paradigma de la *Green AI* como una línea clave de investigación.

Motivado por estas consideraciones, este trabajo se alinea con los principios de la *Green AI*, con el objetivo de estudiar métodos de transferencia de conocimiento entre modelos y estrategias de orquestación consciente de la complejidad. En particular, se investiga si un marco de selección dinámica de modelos puede alcanzar un rendimiento comparable o superior al de un modelo base completamente entrenado, utilizando simultáneamente menos recursos computacionales. El modelo base se emplea como referencia y se compara con alternativas más

[5] Disponible en: https://www.un.org/sustainabledevelopment/

eficientes mediante métricas automáticas estándar en tareas de modelado del lenguaje.

En conjunto, esta investigación contribuye al desarrollo de sistemas de PLN más eficientes, sostenibles y escalables, promoviendo un equilibrio entre rendimiento, coste computacional e impacto ambiental, con el objetivo de avanzar hacia sistemas de inteligencia artificial más responsables, accesibles y energéticamente eficientes.

Agradecimientos

Ainhoa Vivel Couso es una investigadora predoctoral financiada por la Consellería de Educación, Ciencia, Universidades e Formación Profesional de Galicia (ED481A-2024-059). Los autores reconocen el apoyo de la Consellería de Educación, Ciencia, Universidades e Formación Profesional de Galicia (proyecto CiTIUS 2024-2027 ED431G2023/04) y de la Unión Europea (Fondo Europeo de Desarrollo Regional - FEDER). Este trabajo también está financiado por el Ministerio de Ciencia, Innovación y Universidades de España (MCIN/AEI/10.13039/501100011033/) a través de los proyectos XAI4SOC PID2021-123152OB-C21, MAIXAI4Trust PID2024-157680NB-I00, y COLLAGE PID2023-149549NB-I00.

Referencias

1. Chen, Z., Wu, M., Chan, A., Li, X., Ong, Y.S.: Survey on AI sustainability: Emerging trends on learning algorithms and research challenges. IEEE Computational Intelligence Magazine **18**(2), 60–77 (2023), https://doi.org/10.1109/MCI.2023.3245733
2. Jiang, C., Maddela, M., Lan, W., Zhong, Y., Xu, W.: Neural CRF model for sentence alignment in text simplification. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7943–7960. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.709, https://aclanthology.org/2020.acl-main.709/
3. Vivel-Couso, A., Pramata, N.Z., Robredo, D., Soroa, A., Alonso-Moral, J.M.: MeteoGalEus: An iberian multilingual weather dataset in galician, euskera, and spanish. In: Piperidis, S., Bel, N., van den Heuvel, H., Ide, N., Krek, S., Toral, A. (eds.) Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026). pp. 9905–9919. European Language Resources Association (ELRA), Palma, Mallorca, Spain (May 2026). https://doi.org/10.63317/29v6u9pgo67x
4. Zhang, X., Lapata, M.: Sentence simplification with deep reinforcement learning. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 584–594. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1062, https://aclanthology.org/D17-1062/