Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 29-38 recibido 14-04-2023 revisado 08-05-2023 aceptado 10-05-2023

La perplejidad como herramienta para estimar la asignación de nivel de competencia en escritos de una lengua extranjera

Perplexity as a tool for the allocation of proficiency levels to utterances written by foreign language learners

M.P. Agustín-Llach,[1] J. Heras,[2] G. Mata,[2] J. Rubio[2]
[1]Departamento de Filologías Modernas. Universidad de La Rioja, España
[2]Departamento de Matemáticas y Computación. Universidad de La Rioja, España
{maria-del-pilar.agustin, jonathan.heras, gadea.mata, julio.rubio}@unirioja.es

Resumen: La asignación de niveles de competencia a escritos producidos por aprendices de una lengua es una tarea altamente subjetiva. Es por esto que el desarrollo de métodos que evalúen escritos de manera automática puede ayudar tanto al profesorado como al alumnado. En este trabajo, hemos explorado dos vías mediante el uso del corpus CAES. Dicho corpus está formado por escritos de aprendices de español y etiquetado con niveles CEFR (hasta el C1). La primera aproximación es un modelo de aprendizaje profundo llamado Deep-ELE que asigna niveles de competencia a las frases. La segunda aproximación llevada a cabo ha consistido en estudiar la perplejidad de las frases de los estudiantes de distintos niveles, para luego clasificarlos en niveles. Ambas aproximaciones han sido evaluadas, y se ha comprobado que pueden usarse de manera exitosa para clasificar frases por niveles. En concreto, el modelo Deep-ELE obtiene una *accuracy* de 81,3 % y un QWK de 0,83. Como conclusión, este trabajo es un paso para entender cómo las herramientas del procesado de lenguaje natural pueden ayudar a las personas que aprenden un segundo idioma.
Palabras clave: Perplejidad, Deep Learning, ELE.

Abstract: The allocation of proficiency levels to utterances written by foreign language learners is a subjective task. Therefore, the development of methods to automatically evaluate written sentences can help both students and teachers. In this work, we have explored two different approaches to tackle this task by using the corpus CAES, which contains written utterances of learners of Spanish labelled with CEFR levels (up to C1). The first approach is a deep learning model called Deep-ELE which assigns proficiency levels to sentences. The second approach consists in studying the perplexity of sentences written by students of different levels, to later allocate levels to those sentences based on such an analysis. Both approaches have been evaluated, and results confirm that they can be used to successfully classify written sentences into proficiency levels. In particular, the Deep-ELE model reaches an accuracy of 81.3 % and a weighted Cohen Kappa of 0.83. As a conclusion, this work is a step towards better understanding how natural language processing methods can help learners of a second language.
Keywords: Perplexity, Deep Learning, Spanish as a Foreign Language.

ISSN 1135-5948 DOI 10.26342/2023-71-2

1 Introducción

Las técnicas de *deep learning* han supuesto un cambio crucial en la mayoría de los campos de aplicación del procesamiento del lenguaje natural (por ejemplo, para la producción de textos (Narayan y Gardent, 2020), para el resumen de documentos (Kouris, Alexandridis, y Stafylopatis, 2021), para la traducción automática (Shao et al., 2021), para los sistemas de interrogación (Hao et al., 2022) o para la clasificación de textos (Minaee et al., 2021)). Siguiendo esta corriente, parece natural la creación de una herramienta automatizada que pueda determinar el nivel de competencia CEFR (*Common European Framework of Reference for Languages 2001*) de un fragmento de texto producido por aprendices de una lengua. Enunciado el problema de esta manera tan directa, se comprueba que se trata de un típico problema de clasificación en *machine learning*: dada una oración como punto de partida, se busca etiquetarla con uno de los niveles estándar CEFR (A1, A2, etc.). Sin embargo, esta versión simplista debe ser matizada, pues cuando quienes evalúan producciones de lengua asignan un nivel de conocimiento de dicha lengua se basan en algo más que en una oración despojada de cualquier contexto. Pese a eso, explorar las posibilidades del *deep learning* para abordar este marco simplificado tiene su propio interés, no tanto por el valor predictivo de las asignaciones realizadas por un programa de computador, como por analizar los límites de este tipo de acercamientos al problema, puesto que la evaluación humana no puede ser formalizada (en un sentido matemático) y, así, evaluar el funcionamiento de un programa para el mismo fin puede desvelar características relevantes que hayan podido ser pasadas por alto en aproximaciones previas.

Nuestro marco de trabajo se centra en ELE (Español como Lengua Extranjera) y se apoya en la existencia de un repositorio abierto de frases escritas por estudiantes de ELE llamado CAES (CAES, 2022). Tomando como base este *dataset*, describimos en este trabajo dos enfoques, ambos basados en *deep learning*. En primer lugar, hemos utilizado CAES para entrenar (y para validar) un modelo de *deep learning* (que hemos denominado Deep-ELE), basado en una arquitectura de *transformer*, y que permite, dado un texto en español, asignar un nivel de competencia al mismo.

Por otra parte, nos apoyamos, con el mismo propósito, en el concepto de *perplejidad* en un modelo de lenguaje. La noción de perplejidad (Jurafsky y Martin, 2021), que estima la distancia entre una frase introducida como argumento y las oraciones codificadas en un modelo de lenguaje dado, ha sido utilizada tradicionalmente para analizar la calidad de un modelo de lenguaje. Aquí proponemos, por primera vez, si no nos equivocamos, utilizar la perplejidad para determinar el nivel de competencia lingüística de producciones en ELE, con la hipótesis de que quienes tengan mayor conocimiento de una lengua, producirán oraciones con menor perplejidad. Aunque los resultados experimentales que hemos obtenido no permiten adoptar la perplejidad como criterio único para determinar el nivel de competencia lingüística, sí que obtenemos interesantes correlaciones que abren espacios para continuar la investigación.

2 Investigación relacionada

La evaluación de la producción de aprendices de una lengua ha sido abordada de dos modos: 1) por medio de evaluaciones holísticas que usan rúbricas estandarizadas (por ejemplo, *ESL Composition Profile* (Jacobs et al., 1981)) o 2) de forma analítica, supervisando la presencia o ausencia de aspectos específicos (en producciones orales o escritas); entre estos aspectos se puede citar la riqueza léxica, el número de errores, las estructuras sintácticas o las *T-units* (véanse, por ejemplo, (Wolfe-Quintero, Inagaki, y Kim, 1998; Jarvis y Paquot, 2015; Polio y Yoon, 2020)). Los estudios agrupados en la segunda clase hacen uso de diferentes *corpora* y de técnicas de *machine learning* (citemos, entre ellas, el análisis discriminante, *support vector machines* o regresión logística (Malmasi et al., 2017)) para crear sistemas de clasificación que predicen la asignación de niveles en muestras de texto escrito o fragmentos hablados, a través del cómputo de características, de frecuencias, de n-gramas, etc. Mientras que en la primera aproximación se tiende a confiar en el criterio de seres humanos que observan los textos como un todo. La personas que evalúan utilizan, en conjunción con su experiencia previa y su conocimiento especializado, rúbricas normalizadas que les ayudan a concentrarse en ciertos aspectos que la comunidad ha consensuado que son relevan-

tes para una buena escritura o habla, lo que les permite asignar un nivel a una producción lingüística dada (Jarvis, Alonso, y Crossley, 2019). Los aspectos más repetidos suelen ser: coherencia, cohesión y presencia de palabras u oraciones de enlace, la colocación de las palabras, la ausencia de errores, la corrección en la puntuación y en la ortografía, una gramática adecuada o ser capaces de comunicar eficazmente (veánse (Hamp-Lyons, 1991; COE, 2021; Weigle, 2002)). Además, las evaluaciones humanas pueden descubrir características importantes que han podido pasar desapercibidas o ser descartadas en las aproximaciones más algorítmicas. La puntuación holística, cuando involucra a personas evaluadoras, suele ser más rápida, más flexible y de miras más amplias, pero, por el contrario, suele ser menos fiable y objetiva, debida al cansancio generado por el mismo proceso de evaluación, y suele requerir más recursos, puesto que las evaluaciones deben ser revisadas por varios evaluadores para garantizar su validez y fiabilidad.

Hasta donde sabemos, los métodos de *deep learning* se han utilizado principalmente, en este contexto, para evaluar la pronunciación de estudiantes que aprenden inglés como segunda lengua (Fu, 2020; Metallinou y Cheng, 2014; Kobayashi y Wilson, 2020; Takai et al., 2020). También se han utilizado para analizar la calidad de documentos escritos por estudiantes de una segunda lengua, y algunas de estas investigaciones han dado lugar a aplicaciones comerciales como *e-rater* (Burstein, Tetreault, y Madnani, 2013), *Intelligent Essay Assessor* (Foltz et al., 2013) o *Research Writing Tutor* (Cotos, 2014) que pretende ayudar a las personas que quieren mejorar la calidad de sus escritos. Incluso han surgido competiciones para construir sistemas que midan el nivel de competencia de estudiantes de lengua extranjera (Lab, 2023).

En el ámbito más preciso que compete a nuestra investigación, se han utilizado técnicas de *machine learning* y de *deep learning* para clasificar muestras de lenguaje en niveles CEFR para el portugués (Santos et al., 2021), con una *accuracy* del 86.84 %; italiano (Santucci et al., 2020), con una *accuracy* del 71.88 %; inglés (Ding et al., 2021), con una *accuracy* del 90.11 %; chino (Sung et al., 2015), con una *accuracy* del 74.97 %; coreano (Lim, Song, y Park, 2022), con una *accuracy* del 96.85 %; o alemán (Hancke y Meurers, 2013), con una *accuracy* del 62.7 %; entre otros.

Sin embargo, de nuevo el foco de estas investigaciones esta puesto en identificar características textuales, ya sean cuantitativas o descriptivas, de cada nivel a partir de muestras ya clasificadas. Nuestro método para construir Deep-ELE es diferente, puesto que no está basado en ninguna característica previamente establecida, sino que se deja guiar por las evaluaciones holísticas almacenadas en CAES, con un enfoque probabilístico, como explicaremos en los siguientes apartados.

Por último, y como hemos adelantado en la introducción, consideramos que la perplejidad no ha sido utilizada en ningún trabajo previo como lo hacemos aquí.

3 Construcción y evaluación de Deep-ELE

Determinar el nivel de una oración puede ser interpretado como un problema de clasificación de textos (Jurafsky y Martin, 2021). Este tipo de tareas suele ser abordado construyendo un modelo (es decir: una función, en el sentido matemático) que, dado un texto como entrada, produce una etiqueta (en nuestro caso: el nivel del o la aprendiz) como salida. El proceso para construir tal modelo consta de los pasos que enumeramos a continuación (Tunstall, von Werra, y Wolf, 2022). El punto de partida es un repositorio (*dataset*) que está compuesto por pares: un texto junto a la etiqueta asociada a el. Estos *datasets* deben ser pre-procesados, para eliminar duplicados, evitar valores nulos o mal construidos y, en general, para limpiarlo. Una vez el *dataset* ha sido pre-procesado, es dividido en tres grupos disjuntos: un conjunto de entrenamiento, un conjunto de validación y un conjunto de test. Los conjuntos de entrenamiento y validación son empleados para ajustar los parámetros e hiper-parámetros del modelo, para intentar minimizar el número de errores al asignar una etiqueta a un texto; esta se conoce como *fase de entrenamiento*. Finalmente, el modelo es confrontado con el conjunto de test para determinar su precisión (*fase de testing*) y, más adelante, puede ser usado para predecir las características o el nivel de textos que no pertenezcan a ninguno de los tres conjuntos que acabamos de describir. En el resto de este apartado proporcionamos una descripción más detallada de cada uno de estos pasos para el problema concreto que

abordamos.

3.1 El corpus CAES

El corpus CAES (CAES, 2022) es una compilación de fragmentos de texto escritos por aprendices entre los niveles A1 y C1 de competencia en ELE, español como lengua extranjera. En CAES están representadas once lenguas maternas: alemán, árabe, chino mandarín, francés, griego, inglés, italiano, japonés, polaco, portugués y ruso. Se incluyen 6561 fragmentos de 2544 estudiantes que cursan estudios de ELE en el Instituto Cervantes y en otras universidades a lo largo del mundo. La asignación de nivel corresponde con el nivel de referencia que cada estudiante ya había superado, según el curso en que había formalizado matrícula en el momento de la recogida de datos; es decir, por ejemplo, a los textos de participantes que estaban en un curso de B1 se le asignaba el nivel A2, pues se supone que ese es el nivel en el que ya habían obtenido su certificado.

3.2 Adquisición de datos para Deep-ELE

La información extraída del corpus CAES fue almacenada como una serie de ficheros CSV (más concretamente, de ficheros TSV, datos separados por tabuladores, porque cualquier otro signo de puntuación, fuesen comas, puntos, puntos y comas, apóstrofes o signos de interrogación o exclamación, forman parte de algunos de los textos que CAES almacena). Cada fila está compuesta por un número de identificación (véase parte de una pantalla de CAES en la Figura 1, con mayor resolución en el material suplementario[1]), el nivel de competencia, la lengua materna de quien emitió la oración, la oración en sí y una palabra distinguida (o, más en general, un *token* distinguido; por ejemplo, en la Figura 1 aparece en negrita la palabra "para"). Ese token es utilizado por CAES para indexar las diferentes páginas web; es decir, las páginas web agrupan todas las oraciones que contienen ese ítem concreto (la palabra "para" en el ejemplo).

Globalmente, obtuvimos 649722 instancias del CAES; por el modo de organización del CAES, instancias se refieren a fragmentos de texto que pueden contener una o varias

[1]Material suplementario `https://bit.ly/45xZ7EP`

Figura 1: Captura de pantalla de una página web de CAES.

oraciones. Estas instancias estaban desequilibradas tanto respecto al nivel de competencia (véase la Tabla 1 y, de forma más descriptiva, en la Figura 2) como en lo relativo a las lenguas maternas de los aprendices que las produjeron (véase en la Tabla 1 del material suplementario en el enlace del pie de página).

Nivel	Número de instancias
A1	140974
A2	180699
B1	145400
B2	119737
C1	62912
Total	649722

Tabla 1: Número de instancias descargadas por nivel.

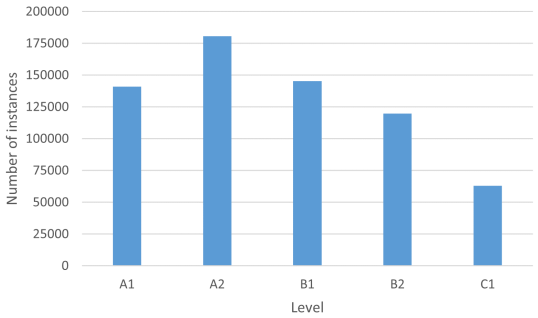

Figura 2: Número de instancias descargadas por nivel.

3.3 Preparación de datos para Deep-ELE

Los datos descargados no pudieron ser utilizados directamente para entrenar un modelo profundo por las razones explicadas a continuación. En primer lugar, se eliminaron ciertas instancias que presentaban problemas de

Lengua materna	Número de instancias
Árabe	143270
Chino mandarín	93189
Francés	79518
Inglés	123586
Portugués	154433
Ruso	55726
Total	649722

Tabla 2: Número de instancias descargadas por lengua materna.

codificación (debido a la presencia de comillas, apóstrofes y caracteres fuera del repertorio UTF-8). En segundo lugar, y esta fue la parte de pre-procesamiento más importante, fue necesario eliminar duplicados, dada la forma en que las oraciones son distribuidas en las páginas web de CAES. Como hemos mencionado anteriormente, las oraciones de cada página web son recuperadas desde un repositorio central a través de un *token* que actúa como índice. Como consecuencia, una misma oración puede aparecer en varias páginas web (indexadas con respecto a tokens o vocablos diferentes) y, por tanto, en varios ficheros TSV. Si organizamos el entrenamiento, la validación y el *testing* sin tener en cuenta esta característica, tomando los ficheros TSV iniciales, sin pre-procesamiento, las instancias duplicadas contaminarán los procesos, produciendo niveles de precisión que no corresponden a la realidad. Es obligatorio, así, eliminar duplicados de nuestro *dataset*.

Tras la eliminación de duplicados, el número total de instancias se redujo de 649722 a 46784. El *dataset* continuó estando, como era de esperar, desequilibrado (véanse las tablas 3 y 4, y la tabla proporcionada en el material suplementario).

Nivel	Número de instancias
A1	12294
A2	13930
B1	10839
B2	6593
C1	3128
Total	46784

Tabla 3: Número de instancias por nivel, tras la eliminación de duplicados.

Lengua materna	Número de instancias
Árabe	8002
Chino mandarín	7628
Francés	5780
Inglés	9837
Portugués	10694
Ruso	4843
Total	46784

Tabla 4: Número de instancias por lengua materna, tras la eliminación de duplicados.

3.4 BERT y RoBERTa

BERT (Devlin et al., 2019) es un modelo de lenguaje pre-entrenado orientado a tareas de procesamiento de lenguaje natural. Fue construido a partir de un enorme *dataset* y debe su gran éxito a haber sido utilizado en el contexto de las arquitecturas *transformer*. En este tipo de aproximaciones, un modelo pre-entrenado (como BERT) es *refinado* (o, en inglés, *fine-tuned*) por medio de un nuevo *dataset* de entrenamiento (habitualmente de tamaño mucho menor que el inicialmente utilizado para BERT) elegido para el problema específico de procesamiento de lenguaje natural que se quiera abordar. Puesto que BERT es una herramienta plurilingüe, pudimos utilizarlo para nuestro problema relacionado con la lengua española.

BERT fue entrenado con amplios *corpora* textuales utilizando dos tareas no-supervisadas. La primera tarea, MLM (*Masked Language Modeling*), hace que el modelo "adivine" qué palabra se esconde en una posición de un fragmento del texto, que es ocultada. La segunda tarea es NSP (*Next Sentence Prediction*), en la que el modelo tiene que predecir si es probable que dos oraciones puedan aparecer consecutivamente en el corpus o bien si han sido elegidas al azar en dicho corpus. Estas dos tareas permiten que el modelo cree representaciones internas sobre la organización del lenguaje, representaciones que pueden ser luego re-utilizadas en otros problemas relativos al procesamiento del lenguaje natural.

La arquitectura BERT fue revisada y mejorada en RoBERTa (Liu et al., 2019). RoBERTa vuelve a ser un modelo pre-entrenado, en el que se escudriñaron distintas decisiones de diseño adoptadas en BERT, con el objetivo de mejorar su rendimiento. En particular, se excluyó el paso NSP y, a cam-

bio, el paso MLM fue enriquecido alimentando el modelo con varias oraciones completas en cada entrada. En nuestro desarrollo hemos utilizado tanto BERT como RoBERTa.

3.5 Entrenamiento de Deep-ELE

Para obtener nuestro modelo de clasificación de texto, hemos aplicado un proceso de *fine-tuning* (Sharif Razavian et al., 2014) con las dos arquitecturas antes mencionadas: BERT y RoBERTa. Hemos llevado a cabo ese proceso de *refinado* reemplazando la "cabeza" de cada modelo del lenguaje (es decir, la última capa de la red neuronal correspondiente) por una nueva "cabeza" adaptada a nuestro problema específico. Entonces, hemos entrenado los modelos durante 10 *epochs*. Todas las redes utilizadas en nuestro desarrollo han sido implementadas con Pytorch (Paszke et al., 2019), y hemos entrenado apoyándonos en la funcionalidad de las librerías de Hugging Face (Wolf et al., 2020), FastAI (Howard y Gugger, 2020) y Blur (Gilliam, 2021), utilizando para ello una GPU Nvidia RTX 2080 Ti. Finalmente, ensamblamos los modelos así obtenidos para generar las predicciones sobre los niveles de competencia.

3.6 Evaluación de Deep-ELE

El *dataset* fue dividido, tras su preprocesamiento, en tres partes: entrenamiento, validación y test. El conjunto de entrenamiento constó de 33684 instancias (72 %) elegidas al azar, el conjunto de validación de 3743 instancias (8 %) y 9357 instancias (20 %) fueron usadas como conjunto de test; véase la Figura 3. Pese a que el *dataset* no estaba equilibrado ni respecto a nivel de competencia ni en lo relativo a las lenguas maternas de los aprendices, los tres conjuntos construidos fueron elegidos aleatoriamente, pero siguiendo una estrategia de estratificación por la que se respetan los porcentajes de cada nivel en cada uno de los tres subconjuntos de datos; véase la Figura 4.

La *accuracy* alcanzada en el conjunto de test es destacable: 81,3 %. Recordemos que mientras que el corpus CAES clasifica las producciones según el nivel del aprendiz que las produce, aquí, el modelo Deep-ELE, clasifica las mismas producciones directamente. En la Figura 5 puede verse la matriz de confusión obtenida por Deep-ELE.

Además, al poder ser ordenadas las etiquetas del corpus CAES como ordinales es conveniente usar el *weighted Kappa coefficient* (κ) con coeficientes cuadráticos para medir el acuerdo y desacuerdo. La interpretación de dicho coeficiente es 1 para una concordancia perfecta, mientras que el puro azar sería un 0. Los valores de este coeficiente se suelen interpretar del siguiente modo. Por debajo de $0,2$ se considera una concordancia pobre; entre $0,21$ y $0,4$ como justa; entre $0,41$ y $0,6$ como moderada; entre $0,61$ y $0,8$ como buena; y por encima de $0,8$ como muy buena. Es importante notar que el estadístico Kappa depende de la prevalencia de cada categoría y su número. En nuestro caso el valor de κ es de $0,83$ por lo que el nivel de concordancia se puede considerar como muy bueno.

Figura 3: Porcentaje de datos utilizados para cada uno de los tres subconjuntos.

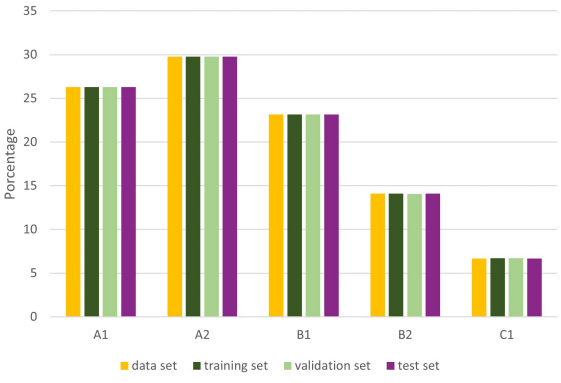

Figura 4: Porcentaje de datos de cada nivel para cada uno de los tres subconjuntos.

Finalmente, se ha desarrollado una aplicación de Gradio que se encuentra alojada en HuggingFace[2] y que permite usar Deep-ELE, ver Figura 6.

[2] `https://huggingface.co/spaces/joheras/DeepELE`

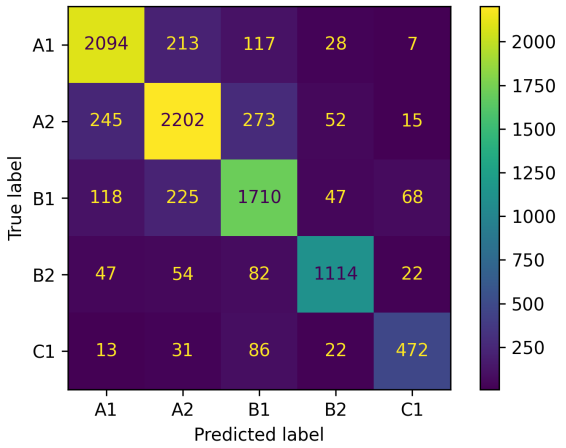

Figura 5: Matriz de confusión de Deep-ELE al ser aplicado al conjunto de test.

4 Perplejidad

En el contexto del procesamiento de lenguaje natural, la *perplejidad* es una medida de lo bien que el modelo de lenguaje predice una concatenación de palabras. Su uso más común ha sido para evaluar la calidad de un modelo de lenguaje (Jurafsky y Martin, 2021). Dado un modelo M y una oración separada en *tokens* $X = (x_0 x_1 \dots x_n)$, la perplejidad de X es:

$$PPL(X) = exp^{-\frac{1}{n}\sum_{i=0}^{t} \log M(x_i|x_{<i})}$$

donde $\log M(x_i|x_{<i})$ es la similitud logarítmica de la probabilidad asociada al *token* x_i por el modelo de lenguaje M condicionada a los *tokens* previos $x_{<i}$.

4.1 Pre-procesamiento de los datos de CAES para el cálculo de la perplejidad

Del *dataset* pre-pocesado para Deep-ELE se excluyeron aquellas oraciones que estuviesen escritas completamente con mayúsculas, debido a que comprobamos que la perplejidad era muy sensible a ese hecho (por cómo fue entrenado y evaluado el modelo para el cálculo de la perplejidad; véase el siguiente subapartado), y se daría la paradoja de que para una oración perfecta desde el punto de vista gramatical se calculase una perplejidad muy grande.

En este proceso retiramos 228 oraciones (161 del nivel A1, 34 del nivel A2, 13 del B1, 15 del nivel B2 y 5 del C1), quedándonos con 46556 instancias.

4.2 Modelo para el cálculo de la perplejidad

Para calcular la perplejidad, utilizamos un modelo en español para 5-gramas de Kneser-Ney (Ney, Essen, y Kneser, 1994) implementado en la librerías KenLM (Heafield, 2023). Los modelos KenLM (Heafield, 2023) pueden ser entrenados con diferentes *datasets* de distintas lenguas. En nuestro caso, el modelo en español fue entrenado con datos extraídos del corpus de la Wikipedia en español (y esto explica la extrema sensibilidad del modelo a las oraciones completamente escritas en mayúsculas, debido a los procesos de normalización a minúsculas de este corpus de referencia).

4.3 Resultados sobre la perplejidad

Nuestros resultados muestran que la perplejidad es una métrica que, efectivamente, está relacionada con los niveles de las distintas instancias (véase la Figura 7). Las instancias con menor nivel de competencia tienen un valor de perplejidad más elevado que las de nivel superior. Además, como se puede comprobar en la Figura 8, la perplejidad no solo discrimina en general, sino que si separamos por lengua materna, continúa ocurriendo que las oraciones tienen mayor perplejidad en los niveles inferiores. Solo en las oraciones cuya lengua materna es el árabe no se repite esta situación, ya que el nivel B2 tiene un valor de perplejidad mayor que los niveles A2 y B1.

5 Conclusiones

En este trabajo se ha estudiado el uso de la perplejidad y de distintos modelos de aprendizaje profundo para evaluar de manera automática el nivel de competencia de los estudiantes de ELE. Ambas aproximaciones han mostrado su viabilidad para ser empleadas con tal fin. Para ello se ha utilizado el corpus CAES, obteniendo con el modelo de aprendizaje profundo un nivel de acuerdo basado en el coeficiente de Kappa de 0,83 y una *accuracy* de 81,3 %, valores similares a los obtenidos en otras lenguas. Por su parte, hemos mostrado que la perplejidad es una métrica que está correlacionada con el nivel de las oraciones producidas por las personas aprendices.

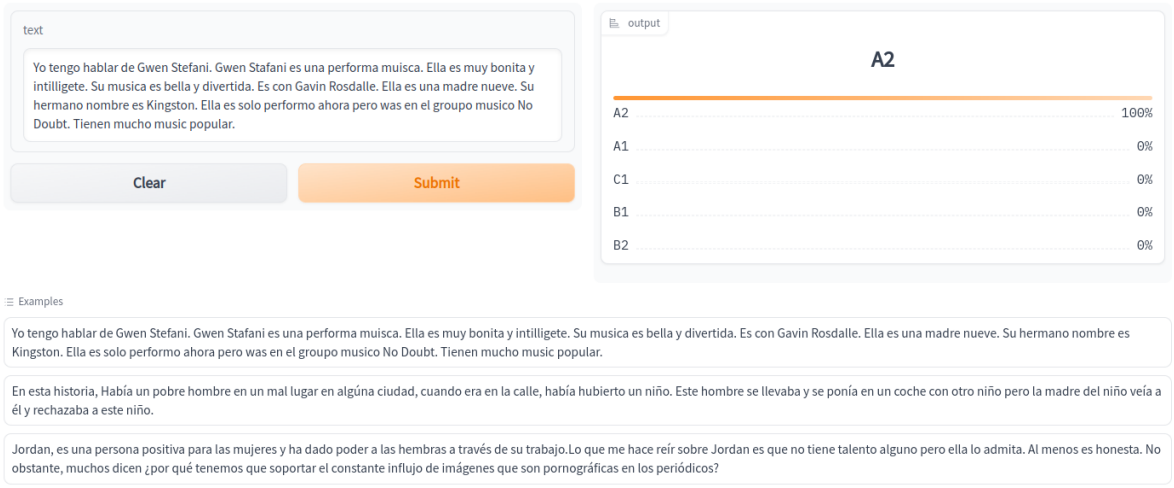

Figura 6: Aplicación alojada en HuggingFace para el uso del modelo Deep-ELE.

Nivel	Promedio Perplexity (KenLM)	Número de instancias
A1	19143.41	12133
A2	13674.07	13896
B1	13099.21	10826
B2	10843.43	6578
C1	9174.05	3123
Total		46556

Figura 7: Media de la perplejidad y el número de instancias para cada nivel.

Agradecimientos

Esta investigación ha sido parcialmente financiada por los proyectos AFIANZA 2022/02, PID2020-115225RB-I00 de MCIN/AEI/ 10.13039/501100011033 y PID2020-116641GB-I00 de MCIN/AEI/ 10.13039/501100011033.

Bibliografía

Burstein, J., J. Tetreault, y N. Madnani. 2013. The e-rater automated essay scoring system. En *Handbook of Automated Essay Evaluation*. Routledge, páginas 55—-67.

CAES. 2022. Corpus de aprendices de español (CAES). https://galvan.usc.es/caes/.

COE. 2021. CEFR: Common European Framework of Reference for Languages. Council of Europe. https://www.coe.int/en/web/common-european-framework-reference-languages.

Cotos, E. 2014. *Genre-based automated writing evaluation for L2 research writing: From design to evaluation and enhancement*. Macmillan.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, páginas 4171–4186. Association for Computational Linguistics.

Ding, H., Q. Zhong, S. Zhang, y L. Yang. 2021. Text difficulty classification by combining machine learning and language features. En *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, páginas 1055–1063. Springer.

Foltz, P. W., L. A. Streeter, K. E. Lochbaum, y T. K. Landauer. 2013. Implementation and applications of the Intelligent Essay Assessor. En *Handbook of Automated Essay Evaluation*. Routledge, páginas 68–88.

Fu, J. 2020. *Automatic Proficiency Evaluation of Spoken English by Japanese Learners for Dialogue-Based Language Learning System Based on Deep Learning*. Ph.D. tesis, Tohoku University.

Gilliam, W. 2021. Blur: A library that integrates huggingface transformers with version 2 of the fastai framework. https://github.com/ohmeow/blurr.

Nivel	Promedio Perplexity (KenLM)	Número de instancias
Árabe		
A1	17230.68	3070
A2	12677.18	1989
B1	11888.57	1664
B2	16110.95	771
C1	8247.62	408
Chino Mandarín		
A1	20597.07	1734
A2	16080.89	1815
B1	12171.41	3156
B2	9242.76	889
C1	3682.39	25
Francés		
A1	18008.29	1414
A2	14380.96	1282
B1	13930.31	1550
B2	9546.20	884
C1	7759.25	646
Inglés		
A1	19631.57	1270
A2	13187.71	4165
B1	13309.05	1852
B2	7801.63	1729
C1	8860.46	814
Portugués		
A1	19370.70	3936
A2	13403.85	2881
B1	14269.74	1527
B2	14005.27	1437
C1	11752.53	807
Ruso		
A1	23998.00	709
A2	13397.66	1764
B1	14471.98	1077
B2	9949.59	868
C1	8237.06	423

Figura 8: Media de la perplejidad y el número de instancias para cada lengua materna.

Hamp-Lyons, L., editor. 1991. *Assessing second language writing in academic contexts*. Ablex.

Hancke, J. y D. Meurers. 2013. Exploring CEFR classification for german based on rich linguistic modeling. *Learner Corpus Research*, páginas 54–56.

Hao, T., X. Li, Y. He, F. L. Wang, y Y. Qu. 2022. Recent progress in leveraging deep learning methods for question answering. *Neural Computing and Applications*, páginas 1–19.

Heafield, K. 2023. Kenlm language model toolkit. https://kheafield.com/code/kenlm/.

Howard, J. y S. Gugger. 2020. Fastai: A layered API for deep learning. *Information*, 11:108.

Jacobs, H. L., S. A. Zinkgraf, D. R. Wormuth, V. F. Hearfiel, y J. B. Hughey. 1981. *Testing ESL Composition: A Practical Approach. English Composition Program*. Newbury House Publishers, Inc.

Jarvis, S., R. Alonso, y S. Crossley. 2019. Native language identification by human judges. En *Cross-linguistic influence: From empirical evidence to classroom practice*. Springer, páginas 215–231.

Jarvis, S. y M. Paquot. 2015. *Native language identification*. Cambridge University Press.

Jurafsky, D. y J. H. Martin. 2021. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Kobayashi, A. y I. Wilson. 2020. Using deep learning to classify english native pronunciation level from acoustic information. En *SHS Web of Conferences*, volumen 77, página 02004. EDP Sciences.

Kouris, P., G. Alexandridis, y A. Stafylopatis. 2021. Abstractive text summarization: Enhancing sequence-to-sequence models using word sense disambiguation and semantic content generalization. *Computational Linguistics*, 47(4):813–859.

Lab, T. L. A. 2023. English language learning: Evaluating language knowledge of ell students from grades 8-12. https://www.kaggle.com/competitions/feedback-prize-english-language-learning.

Lim, K., J. Song, y J. Park. 2022. Neural automated writing evaluation for korean L2 writing. *Natural Language Engineering*, páginas 1–23.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, y V. Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Malmasi, S., K. Evanini, A. Cahill, J. Tetreault, R. Pugh, C. Hamill, D. Napolitano, y Y. Qian. 2017. A report on the 2017 native language identification shared task. En *12th Workshop on Innovative Use of NLP for Building Educational Applications*, páginas 62–75. Association for Computational Linguistics.

Metallinou, A. y J. Cheng. 2014. Using deep neural networks to improve proficiency assessment for children english language learners. En *Fifteenth Annual Conference of the International Speech Communication Association*.

Minaee, S., N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, y J. Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40.

Narayan, S. y C. Gardent. 2020. Deep learning approaches to text production. *Synthesis Lectures on Human Language Technologies*, 13(1):1–199.

Ney, H., U. Essen, y R. Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1–38.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, y S. Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. En *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., páginas 8024–8035.

Polio, C. y H. Yoon. 2020. Exploring multiword combinations as measures of linguistic accuracy in second language writing. En *Learner corpora and second language acquisition research*. Cambridge University Press, páginas 96–121.

Santos, R., J. Rodrigues, A. Branco, y R. Vaz. 2021. Neural text categorization with transformers for learning portuguese as a second language. En *EPIA Conference on Artificial Intelligence*, páginas 715–726. Springer.

Santucci, V., L. Forti, F. Santarelli, S. Spina, y A. Milani. 2020. Learning to classify text complexity for the italian language using support vector machines. En *International Conference on Computational Science and Its Applications*, páginas 367–376. Springer.

Shao, C., Y. Feng, J. Zhang, F. Meng, y J. Zhou. 2021. Sequence-level training for non-autoregressive neural machine translation. *Computational Linguistics*, 47(4):891–925.

Sharif Razavian, A., H. Azizpour, J. Sullivan, y S. Carlsson. 2014. CNN features off-the-shelf: An astounding baseline for recognition. En *CVPRW'14*, páginas 512–519.

Sung, Y.-T., W.-C. Lin, S. B. Dyson, K.-E. Chang, y Y.-C. Chen. 2015. Leveling l2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2):371–391.

Takai, K., P. Heracleous, K. Yasuda, y A. Yoneyama. 2020. Deep learning-based automatic pronunciation assessment for second language learners. En *International Conference on Human-Computer Interaction*, páginas 338–342. Springer.

Tunstall, L., L. von Werra, y T. Wolf. 2022. *Natural language processing with transformers*. O'Reilly Media, Inc.

Weigle, S. C. 2002. *Assessing writing*. Cambridge University Press.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, y A. Rush. 2020. Transformers: State-of-the-art natural language processing. En *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, páginas 38–45. Association for Computational Linguistics.

Wolfe-Quintero, K., S. Inagaki, y H.-Y. Kim. 1998. *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawai'i Press.