Propuesta recibida en febrero 2024 y aceptada para publicación en mayo 2024.

Explorando las capacidades de los modelos de lenguaje neuronales en la identificación y clasificación de colocaciones léxicas

Exploring the capabilities of neural language models for the identification and classification of lexical collocations

Radovan Milović
Universidad de Santiago de Compostela

Resumen

La mayoría de las investigaciones sobre el procesamiento automatizado de colocaciones se ha centrado en el uso de medidas de asociación. Sin embargo, el enfoque se ha ido cambiando lentamente hacia la exploración de la efectividad de los modelos de lenguaje neuronales o *neural language models* (NLMs). En este artículo, investigamos el último método mediante el ajuste fino de modelos de la familia BERT en inglés, español y portugués utilizando recursos léxicos anotados con Funciones Léxicas (FL). Examinamos así las capacidades de los modelos de lenguaje para la identificación y clasificación de colocaciones léxicas tanto en escenarios monolingües como multilingües. Los resultados de los desempeños generales variaron, con valores F que oscilan entre 0.30 y 0.51. Concluimos que el modelo multilingüe sobresale en el aprendizaje cruzado al emplear un conjunto de entrenamiento combinado de los tres idiomas. Además, a pesar de la posible variabilidad, los resultados demuestran una mejor identificación de las Funciones Léxicas con un mayor número de instancias en el conjunto de entrenamiento. Por último, realizamos un análisis cualitativo para investigar posibles patrones de identificación errónea exhibidos por el modelo.

Palabras clave

colocaciónes léxicas, funciones léxicas, modelos de lenguaje neuronales, ajuste fino

Abstract

The majority of research on automated collocation processing has focused on using association measures. However, the focus has been slowly shifting to exploring the effectiveness of neural language models (NLMs). In this paper, we investigate the latter by fine-tuning BERT family models in English, Spanish, and Portuguese using annotated lexical resources with Lexical Functions (LFs). We examine the capabilities of language models for the identification and classification of lexical collocation in both monolingual and multilingual scenarios. The results of the overall performances varied, with f1 scores ranging from 0.30 to 0.51. We conclude that the multilingual model excels in cross-lingual learning by employing a combined training set of all three languages. Moreover, despite possible variability, the results demonstrate improved identification of Lexical Functions with a larger number of instances in the training set. Lastly, we conduct a qualitative analysis to investigate possible patterns of misidentification exhibited by the model.

Keywords

lexical collocations, lexical functions, neural language models, fine-tuning

1. Introducción

Según Pawley (1985, p. 102), el lenguaje debe considerarse "una colección de formas de hablar sobre las cosas [...], expresando ideas de una manera convencional (gramatical, idiomática, etc.)"[1], destacando la noción de que la estructura del lenguaje se construye a través de patrones recurrentes que surgen de su uso. Estos patrones, influenciados por las convenciones de un idioma específico, exhiben una arbitrariedad inherente. A nivel léxico, esto es particularmente evidente en combinaciones de palabras como *lluvia torrencial*, *dormir profundamente*, *prestar atención* y otras similares. Estas expresiones son conocidas por su complejidad lingüística, ya que el significado de los constituyentes de las frases "torrencial", "profundamente" y "prestar" se desvía de sus definiciones literales cuando se encuentran en contacto con una palabra específica. Además, la variación en la expresión de conceptos similares a través de diferentes idiomas, como *prestar atención* en español, *Aufmerksamkeit schenken* en alemán o *pay attention* en inglés,

[1] Cita original: "[...] a collection of ways of talking about things [...], expressing ideas in a manner that is conventional (grammatical, idiomatic, etc.)" (Pawley, 1985)

DOI: 10.21814/lm.16.1.428

los hace impredecibles. Tales frases idiosincráticas son comúnmente conocidas como "colocaciones."

La adquisición de colocaciones representa un desafío significativo específicamente para los estudiantes de segundo idioma, ya que requiere un adecuado dominio de los matices lingüísticos específicos de cada idioma. Para superar este desafío, los aprendices suelen recurrir a diccionarios. Sin embargo, antes de incorporar información sobre colocaciones en los diccionarios, es necesario elaborar listas de colocaciones extrayéndolas de corpus. Por lo tanto, los desafíos planteados por las colocaciones también se extienden al ámbito del procesamiento del lenguaje natural (PLN). Además del procesamiento computacional de colocaciones con fines lexicográficos, estas son un factor crucial en una variedad de tareas de PLN como el aprendizaje de idiomas asistido por ordenador, la traducción automática y la desambiguación del sentido de las palabras.

La investigación en PLN sobre colocaciones gira predominantemente en torno al empleo de diversas medidas de asociación para extraer listas de candidatos a colocaciones (Church & Hanks, 1990; Smadja, 1993; Rychlý, 2008), que posteriormente son evaluadas. Este enfoque se ha convertido en el método principal para el procesamiento automático de colocaciones. Sin embargo, descuida las características semánticas fundamentales de las colocaciones. Las medidas de asociación no logran discriminar las colocaciones de otras expresiones multipalabra, lo que lleva a una lista de candidatos que abarca una variedad de coocurrencias.

Investigaciones más recientes han comenzado a explorar estrategias para el descubrimiento de colocaciones que se centran en el uso de *word embeddings* (Rodríguez-Fernández et al., 2016; Enikeeva & Mitrofanova, 2017) y los modelos de lenguaje neuronales (Espinosa-Anke et al., 2021, 2022; Nisho, 2022). Los *word embbedings* y los modelos de lenguaje neuronales mapean la similitud semántica en un espacio vectorial multidimensional según sus patrones de distribución en los corpus, permitiendo así considerar las propiedades semánticas. Además, estos estudios hacen uso de recursos léxicos que contienen colocaciones anotadas con FL. Como resultado, la tarea de detectar colocaciones en corpus con respecto a sus propiedades semánticas se ha entrelazado con su categorización. El objetivo de este enfoque, que también es el punto focal de nuestra investigación, es identificar instancias de colocaciones en corpus y clasificarlas simultáneamente según el modelo teórico de las Funciones Léxicas.

En este artículo, llevamos a cabo un experimento exploratorio que implica el ajuste fino de modelos de lenguaje neuronales con recursos léxicos anotados con FL en tres idiomas: inglés, español y portugués. Estos corpus fueron creados por García et al. (2019), especialmente para abordar el área poco explorada del procesamiento multilingüe de colocaciones. Por consiguiente, queremos tener en cuenta tanto configuraciones monolingües como multilingües, lo que nos permite obtener percepciones sobre el rendimiento de estos modelos en diferentes contextos lingüísticos. Observamos que los datos multilingües generalmente mejoran la identificación de colocaciones en la lengua meta. Sin embargo, algunas FL parecen más difíciles de aprender que otras, debido a la cantidad de instancias de esa FL a las que el modelo está expuesto durante el entrenamiento. A través del análisis cualitativo, también intentamos descubrir conocimientos lingüísticos mediante la identificación de patrones de error presentes en las predicciones del modelo.

2. Marco teórico

2.1. La noción de "colocación"

La definición de la noción de "colocación" no tiene consenso, principalmente debido a diferencias en las perspectivas y métodos de investigación. Sin embargo, el discurso académico actual reconoce dos interpretaciones principales: la interpretación estadística y la interpretación semántica.

Desde el punto de vista estadístico, las colocaciones se perciben como un fenómeno empírico que nos permite aprender sobre los patrones de comportamiento de una palabra (el nodo) en relación con las palabras adyacentes (los colocativos) (Evert, 2004). Estas colocaciones estadísticas se definen como coocurrencias que aparecen en textos más de lo esperado por azar (Sinclair, 1991). La noción de "colocación estadística" sirve como principio fundamental para su extracción automatizada mediante la aplicación de medidas de asociación.

Nuestro trabajo, sin embargo, gira en torno a la interpretación semántica de las colocaciones, las cuales se definen en función de la composición estructural de sus componentes. Comúnmente se les denomina "colocaciones léxicas" (Krenn, 2000) para distinguirlas de la definición estadística. Desde este punto de vista, las colocaciones son construcciones binarias léxicamente vinculadas, que consisten en una **base** y un **colocativo**, cuya relación muestra disparidad en términos

de aspectos sintácticos y semánticos (Mel'čuk, 1996; Hausmann, 1998). La base se elige libremente según la intención del hablante de transmitir un significado particular, mientras que la elección del colocado está restringida por las convenciones del lenguaje.

El principio fundamental de este enfoque es definir las colocaciones diferenciándolas de las combinaciones libres y las locuciones, principalmente en función de los niveles de composicionalidad y sustituibilidad (Nesselhauf, 2005).

Mientras que una combinación libre no tiene restricciones y opacidad semántica, una locucion, por ejemplo *dar en el clavo*, carece de transparencia semántica, ya que su significado "hacer o decir algo correctamente"no puede determinarse solo analizando sus componentes, lo que resulta en una completa falta de composicionalidad. Además, sus componentes no pueden ser sustituidos por elementos sinónimos para transmitir el mismo significado. Por otro lado, una colocación como *dar un paseo* muestra una composicionalidad parcial[2], ya que uno de sus constituyentes ("paseo") lleva un significado semántico transparente, mientras que la interpretación del otro constituyente ("dar") sigue siendo más ambigua. La elección del colocativo también está influenciada por la convención de un idioma particular y no puede ser sustituida sin alterar el significado de toda la frase. En consecuencia, las colocaciones están limitadas a una combinación interdependiente específica y se caracterizan por una transparencia semántica parcial y una falta de libertad combinatoria.

2.2. Clasificación de colocaciones

La clasificación de colocaciones suele basarse en sus patrones sintácticos, según la categoría gramatical de sus constituyentes: *lluvia torrencial* (adjetivo + sustantivo), *prestar atención* (verbo + sustantivo), *dormir profundamente* (verbo + adverbio), etc. (Hausmann, 1998, p. 1010; Benson et al., 2010). Sin embargo, existe otro sistema de clasificación más exhaustivo y detallado desarrollado por Igor Mel'čuk (1996), llamado Funciones Léxicas (FL).

[2]Las perspectivas sobre la noción de composicionalidad también pueden variar de un autor a otro. Por ejemplo, Mel'čuk (2012, p. 39) considera las colocaciones como frases enteramente composicionales, ya que su significado puede dividirse en dos partes de modo que correspondan a los dos constituyentes. Sin embargo, la comprensión general de este enfoque es que las colocaciones representan el área intermedia de un continuo de transparencia, con combinaciones libres y locuciones ubicadas en los dos extremos (Dražić, 2014, p. 17).

Las Funciones Léxicas proporcionan una clasificación concisa de las genuinas relaciones léxicas, que pueden ser tanto sintagmáticas (relaciones de colación) como paradigmáticas (relaciones semánticas). En el contexto de este trabajo, nuestro enfoque se centra en las FL sintagmáticas, es decir, las colocaciones léxicas.

Para su representación sistemática, se ilustran diferentes tipos de relaciones como una notación matemática $\mathbf{f(L) = Li}$. En términos de FL sintagmáticos, el símbolo **f** representa una función léxica particular, **L** denota el constituyente elegido libremente o *keyword* (base), y **Li** denota los elementos restringidos o *values* (colocativos).

El objetivo de esta clasificación es capturar el significado subyacente inferido por los colocativos y su relación de dependencia de la base. Por ejemplo, la FL **Magn** representa un grupo de colocativos que comparten el mismo significado cuando se asocian a una base. Así, en inglés, *deeply, heartily* y *terribly* son colocativos de *sorry* que comparten el mismo significado subyacente de "intensidad" representado como **Magn**(sorry) = {deeply, heartily, terribly}, capturando simultáneamente la semántica de la colocación y la interdependencia entre los constituyentes. De la misma manera, FL **Bon** (p. ej., *diálogo fructífero*) expresa "bueno" o FL **Oper1** (p. ej., *prestar atención*) "hacer". Además, aunque la elección de los colocativos puede diferir entre idiomas, puesto que expresan el mismo significado central, las Funciones Léxicas pueden aplicarse universalmente a todos los idiomas:

$$\mathbf{Oper1}(\text{atención}) = \{\textit{prestar}/\textit{pay}/\textit{schenken}\}\,.$$

Este modelo teórico ha encontrado relevancia en el procesamiento del lenguaje natural debido a sus diversas propiedades (Kolesnikova, 2011, p. 68–71): su encapsulación de las propiedades sintácticas y semánticas de las colocaciones las hace valiosas para resolver ambigüedades sintácticas y léxicas, mientras que su aplicabilidad universal las convierte en una herramienta ideal para la traducción automática. Además, la clasificación detallada de construcciones lingüísticas impredecibles las hace útiles para crear programas de aprendizaje de idiomas que podrían ayudar a los estudiantes a adquirir tales estructuras. Por último, se vuelven centrales para la identificación y clasificación automáticas en corpus, lo cual exploramos más a fondo en este trabajo.

3. Trabajos relacionados

La identificación de colocaciones en relación con las medidas de asociación ha ganado la mayor popularidad hasta el momento. Esta línea de investigación progresó desde la información mutua (Church & Hanks, 1990), hasta diversas medidas de asociación utilizadas hoy en día, como *log-Dice*, *log-likelihood*, *t-score* y otras (Evert, 2004; Rychlý, 2008). Además, estas medidas se han complementado con etiquetador morfosintáctico (Evert & Kermes, 2003), analizadores de dependencias (Lin, 1999; Seretan & Wehrli, 2006), y finalmente con medidas direccionales (Gries, 2013; Carlini et al., 2014). A través de esta metodología, los candidatos a colocaciones pasan por un riguroso proceso de filtrado, abordando también los problemas de discontinuidad y asimetría de los constituyentes colocacionales.

En cuanto a la tarea autónoma de clasificar automáticamente las colocaciones basadas en el modelo de Funciones Léxicas, los métodos tempranos utilizaron representaciones semánticas basadas en hiperónimos, como *WordNet*, en conjunto con técnicas de aprendizaje automático (Wanner, 2004; Wanner et al., 2006; Gelbukh & Kolesnikova, 2012). Sin embargo, con los avances en la representación semántica y la introducción de *word embeddings* (modelo *Word2vec* de Mikolov et al. (2013)), la tarea de identificar y clasificar colocaciones se unificó.

Uno de los primeros estudios en clasificar e identificar colocaciones simultáneamente fue realizado por Rodríguez-Fernández et al. (2016). Combinan *Word2vec* y el modelo teórico de Funciones Léxicas para identificar y clasificar colocaciones. El método del estudio consiste en utilizar el algoritmo *Word2vec* para generar *word embbedings* en las que las relaciones entre las bases (por ejemplo, "thought" en *deep thought* o "wind" en *strong wind*) y las glosas ("intensity") de sus colocativos correspondientes se mapean en consecuencia. El objetivo es utilizar esta información semántica capturada por los *word embbedings* para recuperar los colocativos potenciales dado una nueva base y glosa.

Los avances en PLN alcanzaron su punto máximo con el desarrollo de modelos de lenguaje neural utilizando la arquitectura de transformador (Vaswani et al., 2017). Lo que hace que estos modelos sean de última generación es su capacidad para enfocarse atentamente en diferentes segmentos de la secuencia de entrada y, por lo tanto, tener en cuenta todo el contexto. Espinosa-Anke et al. (2021) llevaron a cabo dos experimentos para evaluar la eficacia de los transformadores en el manejo de colocaciones. El primer experimento, realizado en una configuración no supervisada, implicó enmascarar el colocativo dentro de una colocación dada y evaluar la precisión predictiva del modelo de lenguaje. El experimento subsiguiente, realizado en una configuración supervisada, se centró en el ajuste fino de los modelos para predecir la función léxica asociada con las colocaciones.

A continuación, Espinosa-Anke et al. (2022) mejoraron el rendimiento de los tranformadores al tener en cuenta las relaciones de dependencia entre la base y el colocativo, integrando un *Graph-aware Transformer* (transformador sensible al grafo) en la estructura del modelo, diseñado específicamente para el análisis de dependencias. Además, incorporan un clasificador de oraciones para determinar la presencia de colocaciones en las oraciones, ofreciendo un contexto adicional para su identificación. Finalmente, utilizan grandes corpus anotados con funciones léxicas en inglés, español y francés para entrenar el modelo y evaluar la efectividad de la arquitectura modificada en la identificación y clasificación de colocaciones.

Junto a los estudios mencionados, es importante destacar la investigación que explora la utilización de *word embeddings* para la tarea respectiva en el idioma ruso realizada por Enikeeva & Mitrofanova (2017), así como la aplicación del *Graph-aware Transformer* para el procesamiento de colocaciones japonesas por Nisho (2022).

Los estudios descritos han mostrado niveles variables de éxito. Además, la falta de transparencia del funcionamiento interno de estos modelos presenta una barrera significativa para los investigadores que buscan comprender precisamente cómo operan estos modelos y por qué producen ciertas salidas. Dicho esto, la identificación y clasificación de colocaciones léxicas utilizando NLMs sigue siendo un territorio abierto que requiere una mayor exploración.

4. Experimento

El ajuste fino es un método popular en el aprendizaje automático, mediante el cual un modelo se entrena en conjuntos de datos más pequeños para adaptarlo específicamente a una tarea objetivo. En esta sección, proporcionamos una descripción detallada de los pasos tomados para el ajuste fino de modelos de lenguaje neural para aprender patrones de colocaciones.

	EN			ES			PT		
	train	**dev**	**test**	**train**	**dev**	**test**	**train**	**dev**	**test**
#tokens	41483	4993	5121	30316	3794	3796	46082	5711	5676
#orac. sin FL	2276	458	255	956	112	114	1710	248	288
#orac. con FL	342	29	51	232	29	25	391	49	45

Cuadro 1: Estadísticas del conjunto de datos para cada idioma.

4.1. Conjunto de datos

Para los experimentos, utilizamos corpus anotados con funciones léxicas en inglés, español y portugués (García et al., 2019)[3], que constan de más de 155.000 tokens y 1.526 colocaciones clasificadas en 60 funciones léxicas. Con el fin de ajuste fino para la clasificación de tokens, convertimos los corpus en un sistema de etiquetado inspirado en el formato BIO. En este formato, el token que denota el elemento base de la colocación se anotó como "B-{FL}", mientras que el token que representa el colocativo se marcó como "C-{FL}". Todos los demás tokens dentro de una oración se etiquetaron como "O" ("outside") para indicar que no pertenecen a la colocación.

Token	**Etiqueta**
Harry	O
felt	O
a	O
hot	C-Magn
surge	B-Magn
of	O
anger	O
.	O

Cuadro 2: Oración de ejemplo con nuevas etiquetas.

Para evitar posibles confusiones, se eliminaron las colocaciones anidadas (p. ej., una expresión en inglés "take a deep breath", donde tanto "take (a) breath" como "deep breath" son colocaciones) y múltiples instancias de las mismas FL dentro de la misma oración (un total de 11 oraciones). Dejamos para futuros trabajos una exploración de estos casos, utilizando, por ejemplo, el análisis sintáctico de dependencia para identificar las relaciones base-colocativo. El Cuadro 1 muestra las estadísticas del conjunto de datos final para cada idioma, mientras que la Figura 1 presenta la frecuencia relativa de las funciones léxicas más comunes.

[3]Para obtener la clasificación completa de FL utilizada en los corpus, así como la descripción del enfoque de anotación y los enlaces a los corpus, consulte a García et al. (2019).

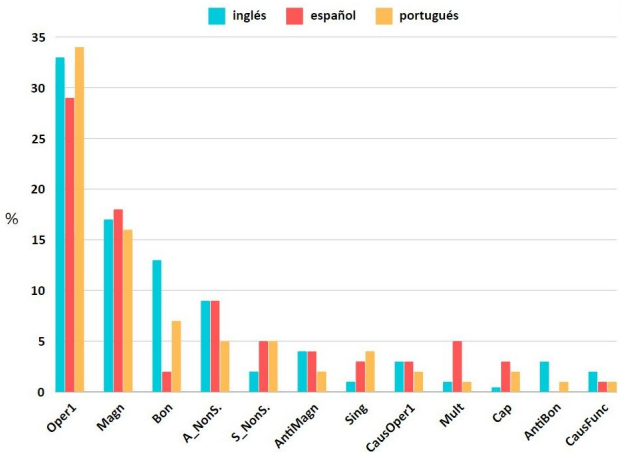

Figura 1: Distribución de las funciones léxicas más comunes para cada idioma.

Dividimos el conjunto de datos en subconjuntos de entrenamiento, validación y prueba según el número total de tokens para cada idioma. El propósito del conjunto de entrenamiento es permitir que el modelo aprenda y capture patrones y características que ayudarán en la identificación de colocaciones. Durante la fase de entrenamiento, el modelo utiliza el conjunto de validación para evaluación interna, refinando continuamente sus parámetros a lo largo de múltiples *epochs* (ciclos de aprendizaje). Una vez que el modelo ha seleccionado la mejor configuración de parámetros, se evalúa el rendimiento en datos de prueba desconocidos para evaluar sus capacidades de generalización.

El 80 % de los tokens se utilizaron para el conjunto de entrenamiento, seguido por el conjunto de validación con el 10 % subsiguiente y el 10 % restante para el conjunto de prueba. Además, nos aseguramos de que el conjunto de entrenamiento abarcara todas las funciones léxicas que aparecieran en los restantes subconjuntos. Los conjuntos de entrenamiento consistían en instancias de colocaciones únicas que no aparecían en los conjuntos de validación y prueba, asegurando así que los modelos no memorizaran simplemente colocaciones específicas, sino que aprendieran los patrones lingüísticos subyacentes.

4.2. Modelos

Para nuestra tarea, utilizamos una familia de modelos BERT y sus *base*[4] variantes adaptadas para diferentes idiomas:

- **BERT** (Devlin et al., 2018) para inglés;
- **RoBERTa** (Gutiérrez-Fandiño et al., 2022) para español;
- **BERTimbau** (Souza et al., 2020) para portugués;
- **mBERT** (Devlin et al., 2018) para la configuracion multilingüe.

Además, para comparar el rendimiento de los modelos de transformadores evaluados, entrenamos el modelo **BiLSTM-CNN-CRF** (Chernodub et al., 2019)[5] con los mismos datos. Este modelo se entrena durante un mínimo de 50 *epochs*.

4.3. Ajuste fino

Para el proceso de ajuste fino empleamos el script *run_ner*[6] (Wolf et al., 2020), diseñado específicamente para el ajuste fino de tareas de clasificación de tokens. En nuestro experimento, utilizamos los parámetros predeterminados, que incluyeron ejecutar el proceso de ajuste fino durante 3 *epochs*.

4.4. Evaluación

Para la evaluación utilizamos el script *conlleval*[7]. El código adopta como entrada las etiquetas verdaderas y predichas y calcula métricas de evaluación, la precisión, la exhaustividad y el valor F. Genera un resumen general del rendimiento y el rendimiento para cada tipo de entidad, que en nuestro caso se corresponde con las funciones léxicas.

4.5. Configuraciones experimentales

Con el propósito de explorar las capacidades de generalización de los modelos en reconocimiento de colocación, llevamos a cabo diferentes configuraciones experimentales. En primer lugar, realizamos el ajuste fino de modelos monolingües utilizando datos etiquetados para cada idioma. Después, reproducimos este proceso con un modelo multilingüe. Finalmente, diseñamos un enfoque de aprendizaje cruzado para entrenar un modelo multilingüe con datos combinados de diferentes idiomas. Por lo tanto, el experimento involucró las siguientes configuraciones:

- Ajuste fino de modelos monolingües **BERT**, **RoBERTa** y **BERTimbau** para cada idioma.
- Ajuste fino del modelo multilingüe mBERT con datos monolingües: **mBERT-mono**.
- Ajuste fino de mBERT utilizando conjuntos de entrenamiento combinados de inglés, español y portugués: **mBERT-multi**.

5. Resultados y análisis

El Cuadro 3 ofrece una visión general de los resultados de rendimiento, mostrando valores F en un rango de 0.30 a 0.51.

5.1. Análisis del rendimiento general

5.1.1. Modelos monolingües

El modelo de referencia, que utiliza la arquitectura BiLSTM-CNN-CRF, tuvo en rendimiento deficiente (promedio $F = 0,11$). Por el contrario, los modelos de transformadores ajustados superaron significativamente el modelo de referencia, lo que indica que la arquitectura de transformadores del modelo entrenado proporcionó beneficios en términos de reconocimiento de colocación. Teniendo en cuenta los parámetros de entrenamiento, la familia de modelos BERT se entrenó durante solo tres *epochs*, mientras que el modelo de referencia requirió alrededor de 30 *epochs* para lograr su máximo rendimiento en los tres lenguajes. Esta notable superioridad de los modelos BERT sobre el modelo de referencia enfatiza las capacidades significativamente mayores de los modelos transformadores para comprender señales contextuales y su adaptabilidad para capturar diversas estructuras lingüísticas.

También es importante destacar que el modelo RoBERTa mostró un rendimiento inferior ($F = 0,32$) en comparación con los modelos BERT inglés ($F = 0,45$) y portugués ($F = 0,47$). La diferencia principal entre los modelos BERT y RoBERTa es el tamaño de los datos de preentrenamiento. RoBERTa se entrena en un corpus de preentrenamiento más grande, lo que potencialmente puede proporcionar una representación lingüística más rica. Sin embargo, Pérez-Mayos et al. (2021) exploraron recientemente la

[4] La variante *base* se refiere a la configuración estándar de un modelo.

[5] https://github.com/achernodub/targer

[6] https://github.com/huggingface/transformers/tree/main/examples/pytorch/token-classification

[7] https://github.com/sighsmile/conlleval

	EN			ES			PT		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
(Ro)BERT(a/imbau)	0.69	0.34	0.45	0.44	0.25	0.32	0.56	0.40	**0.47**
mBERT-mono	0.51	0.27	0.35	0.46	0.23	0.30	0.41	0.37	0.39
mBERT-multi	0.69	0.41	**0.51**	0.48	0.44	**0.46**	0.48	0.42	0.45
BiLSTM-CNN-CRF	0.13	0.13	0.13	0.23	0.07	0.11	0.14	0.07	0.09

Cuadro 3: Rendimiento general de las configuraciones para cada idioma.

correlación entre el tamaño de los datos de preentrenamiento y el rendimiento de los modelos de lenguaje neuronales en la adquisición de patrones sintácticos. Concluyeron que "si bien los modelos preentrenados con más datos codifican más conocimientos sintácticos y tienen un mejor rendimiento en aplicaciones posteriores, no siempre ofrecen un mejor rendimiento en diferentes fenómenos sintácticos". Por lo tanto, el impacto del tamaño del corpus de preentrenamiento en el rendimiento del modelo puede variar según la tarea objetivo. Las puntuaciones más bajas de los modelos RoBERTa en comparación con los modelos BERT pueden sugerir que más datos de preentrenamiento no siempre conducen a mejores resultados para la identificación de colocaciones.

5.1.2. Modelos multilingües

La comparación entre los modelos monolingües y mBERT-mono revela que los modelos monolingües muestran un rendimiento superior en comparación con mBERT cuando se entrenan con datos monolingües. Esto concuerda con hallazgos anteriores (Singh & Lefever, 2022; Conneau et al., 2022) que indican que la abrumadora cantidad de idiomas en los datos de preentrenamiento de modelos multilingües puede conducir a la dilución de información en tareas monolingües. Sin embargo, dada su capacidad para codificar información multilingüe, parecen intrigantes para evaluar enfoques de aprendizaje cruzado, como se demostró en el experimento final (mBERT-multi).

La configuración mBERT-multi, en donde realizamos el ajuste fino de mBERT con conjuntos de entrenamiento combinados, muestra un aumento significativo de rendimiento en la mayoría de los aspectos (excepto la precisión de BERTimbau). Esta mejora podría atribuirse a dos factores clave. En primer lugar, la inclusión de conjuntos de entrenamiento adicionales lleva a un aumento sustancial en el tamaño de los datos. El conjunto de datos ampliado proporciona a mBERT una representación más amplia de patrones lingüísticos, mejorando su capacidad para generalizar en varios idiomas. En segundo lugar, a diferencia de la configuración mBERT-mono, donde llevamos a cabo el ajuste fino del modelo con datos monolingües, la configuración mBERT-multi expone al modelo a una variedad de idiomas, lo que le permite participar en el aprendizaje entre idiomas, capturando así mejor las propiedades sintáctico-semánticas de las colocaciones presentadas por Funciones Léxicas universalmente aplicables.

5.2. Análisis a nivel de FL

Para realizar un análisis a nivel de FL, nos centraremos en una configuración específica, mBERT-multi, que logró predominantemente los mejores resultados generales.

El Cuadro 4 muestra el valor F para cada FL. En algunos casos (Oper1 y Magn), los resultados son altos (0.60 o más), mientras que para varias FL, los modelos no pudieron reconocer correctamente ni una sola colocación. Para explorar más a fondo esta discrepancia, realizamos un análisis de correlación entre la frecuencia de las FL en los datos de entrenamiento y su rendimiento.

5.2.1. Análisis de correlación

Utilizando la correlación de *Spearman*, investigamos la correlación entre el número de tipos de colocaciones en los datos de entrenamiento y el rendimiento de los modelos en el reconocimiento de las respectivas funciones léxicas. Con este análisis, pretendemos determinar si el número de ejemplos de FL se correlaciona con el rendimiento del modelo en el reconocimiento de FL, o si el rendimiento se atribuye a las capacidades del modelo para generalizar las características lingüísticas intrínsecas de FL.

El coeficiente de correlación de 0.70 y 0.72 para inglés y español (Cuadro 4) indica una correlación positiva fuerte entre las FL. Esto sugiere que una mayor frecuencia de FL en los datos de entrenamiento corresponden con un mejor rendimiento en el reconocimiento de las FL respectivas para ambos idiomas. Por lo tanto, aumentar el número de ejemplos de FL en los datos de entrena-

		EN	ES	PT
	cuenta	**valor F**		
Oper1	352	0.60	**0.68**	0.60
Magn	189	**0.72**	0.60	**0.69**
Bon	88	0.56	-	0.44
A_NonS.	87	0.38	0.00	0.00
S_NonS.	47	0.00	0.25	0.00
AntiMagn	40	0.00	0.66	-
CausOper1	33	0.00	0.00	0.00
Sing	31	0.00	0.00	0.00
Cap	25	0.00	0.00	0.00
Mult	22	0.00	0.00	0.00
AntiBon	19	0.28	-	-
CausFunc	19	-	0.00	0.66
correlación		0.70	0.72	0.37
valor p		0.02	0.02	0.29

Cuadro 4: Cuentas y rendimientos de FL más frecuentes en la configuración mBERT-multi, resultados de la correlación de *Spearman* y valor p.

miento probablemente mejorará la competencia del modelo en el reconocimiento de FL. A pesar de que las correlaciones no son estadísticamente significativas ($p = 0{,}2$, probablemente debido al pequeño número de ejemplos), los valores de *Spearman* son notablemente altos. Sin embargo, el coeficiente de correlación de 0.37 en portugués revela una correlación positiva más débil entre la frecuencia de colocación de FL y el rendimiento del modelo. Como podemos ver, a pesar de que el conjunto de entrenamiento contenía un bajo número de ejemplos de CausFunc, el modelo pudo identificar sus instancias, dando como resultado un valor F elevado de 0.66. Aunque aún hay una asociación positiva, no es tan fuerte como se observa en inglés y español. Esto sugiere que, aunque un aumento en la frecuencia de colocación de FL en los datos de entrenamiento puede tener un efecto positivo en el rendimiento del modelo, otros factores como la capacidad de los modelos para aprender las propiedades sintáctico-semánticas de FL pueden contribuir a su identificabilidad.

5.2.2. Análisis de errores

El modelo mostró un rendimiento más alto al identificar la función léxica Magn, logrando un valor F promedio de 0.67 en los tres idiomas. FL Oper1 fue la segunda función léxica mejor identificada, con un promedio de valor F de 0.63. En los párrafos siguientes, nuestra atención se centrará en estas FL. Dado que ocurren con mayor frecuencia, nos proporcionarán información suficiente para el análisis de errores. Buscamos identificar posibles inclinaciones de aprendizaje erróneas del modelo. Las oraciones a continuación ejemplifican los principales desafíos encontrados por los modelos:[8]

(1) China has been a [**great** $_{C-Bon;\ C-Magn}$] [**help** $_{B-Bon;\ B-Magn}$] at getting North Korea to the table...[9]

Problemas para distinguir entre Magn y otras funciones léxicas semánticamente similares, como Bon. A medida que el modelo aprendía más sobre otras funciones léxicas, más dificultades tenía para distinguirlas. Las funciones léxicas Magn y Bon están intrincadamente conectadas y describen típicamente colocaciones que involucran patrones de "adjetivo + sustantivo". Estas funciones buscan capturar el significado fundamental transmitido por colocativos o modificadores que expresan diferentes modificaciones. Magn se refiere principalmente a modificaciones cuantitativas, mientras que Bon está asociada con calificaciones subjetivas. Debido a su similitud, los modelos a veces tienen dificultades para distinguir entre estas dos funciones léxicas. Por ejemplo, en casos como *great help*, los modelos reconocen erróneamente esta colocación como Magn. Dado que "great" funciona como un adjetivo que denota algo de gran magnitud, los modelos no logran discriminarlo de Bon, donde "great" se usa como un modificador positivo cualitativo. Esta confusión puede surgir de las diferencias poco claras en los significados semánticos centrales de los colocativos y la similitud en el patrón sintáctico entre estas dos funciones léxicas.

(2) ...mandam os casos mais graves para um [**grande** $_{O;\ C-Magn}$] [**depósito** $_{O;\ B-Magn}$], que conhecemos como presídio. [10]

Problemas para identificar colocativos con cambio semántico vago. El modelo tuvo éxito en general al reconocer las propiedades semánticas centrales de esta clase de colocaciones. Sin embargo, hubo algunas instancias de falsos positivos, como *grande depósito*. Podemos ver que el modelo tiene dificultades para distinguir estos adjetivos cuando se usan como parte de combinaciones libres versus como

[8] La primera etiqueta es la etiqueta original y la segunda es la etiqueta predicha por el modelo.

[9] Traducción: "China ha sido de gran ayuda para llevar a Corea del Norte a la mesa de negociaciones..."

[10] Traducción: "...mandamos los casos más graves a un gran depósito, que conocemos como presidio."

constituyentes en la colocación Magn. Una posible razón para esta confusión es que el cambio de significado entre su uso independiente y su papel como constituyentes de la colocación no es lo suficientemente pronunciado. Por ejemplo, tanto "grande" en una combinación libre como "grande" como parte de la colocación Magn transmiten la noción de algo de gran magnitud, aunque con matices sutiles. Otro posible factor es la frecuencia de ocurrencia de estos colocativos. Dado el uso frecuente de estos tipos de adjetivos como intensificadores en la colocación, el modelo puede aprender a asociarlos más fuertemente con esta función léxica, lo que podría llevar a falsos positivos.

(3) . . . que están realmente [**enfrentando** $_{C-Oper1;\ O}$] [**problemas** $_{B-Oper1;B-Oper1}$] por la pobreza", indicó.

Problemas para identificar colocativos de Oper1. Los colocativos de Oper1, comúnmente conocidas como verbos de apoyo, según Mel'čuk (1996, p. 53), se consideran "semánticamente vacíos". Aunque el cambio semántico en este caso es muy evidente, podemos ver la incertidumbre del modelo al identificar específicamente los colocativos.

(4) Assim como pode impulsionar as vendas de uma empresa, uma celebridade pode [**causar** $_{C-CausFunc1;\ C-CausOper1}$] o [**efeito** $_{B-CausFunc1;\ B-Oper1}$] contrário.[11]

Problemas para distinguir entre Oper1 y otras FL que representan verbos de apoyo, como Func. A diferencia de Magn y Bon, que se distinguen por criterios semánticos, la diferenciación entre Oper1 y otras funciones léxicas de verbos de apoyo se basa en patrones sintácticos. La función léxica Caus ("iniciar una situación") se combina a menudo con otras funciones léxicas para formar expresiones complejas de FL. Aunque el modelo demostró un mejor aprendizaje de las implicaciones semánticas de la función léxica Caus, este éxito llevó a una confusión adicional al diferenciar entre Oper1, CausFunc y CausOper1.

(5) . . . General Chen Zaido del EPL (militar chino) decidió [**dar** $_{O;\ C-Oper1}$] [**espalda** $_{O;\ B-Oper1}$] en la facción de Guardas Rojas moderado Millón Heroes[. . .]

Problemas para diferenciar entre colocaciones y locuciones. Algunas instancias presentan desafíos para determinar si una frase debe categorizarse como una colocación o una locucion. Para el modelo, un ejemplo de dicho caso es la expresión en español *dar (la) espalda*. Esta frase se caracteriza por una total idiomaticidad, ya que su significado de "rechazar" no se puede inferir a partir de sus partes constituyentes, siendo por lo tanto considerada una locucion. Sin embargo, el patrón estructural de la frase se asemeja al patrón sintáctico de una función léxica Oper1, lo que causa confusión en el modelo. Como resultado, el modelo la identifica como Oper1, teniendo en cuenta el patrón sintáctico de la colocación e incluso reconociendo su naturaleza idiomática, pero sin identificarla correctamente.

Una de las características destacadas de los modelos de transformadores es su capacidad para considerar todo el contexto oracional en lugar de observar únicamente palabras vecinas. Teniendo eso en cuenta, también es necesario señalar la posibilidad de que los patrones de aprendizaje del modelo estén influenciados por las señales contextuales que rodean a las colocaciones. Ya sea Oper1 o Magn, o cualquier otra función léxica, estas señales contextuales desempeñan un papel significativo en la formación de la comprensión e interpretación del modelo sobre las colocaciones. Sin embargo, debido a la diversidad de oraciones en los corpus, no está claro cómo afecta exactamente el contexto a la identificación precisa de funciones léxicas.

6. Conclusiones

El reconocimiento automático de colocaciones léxicas, particularmente con una pequeña cantidad de datos de entrenamiento, es una tarea difícil. Las colocaciones son expresiones lingüísticas complejas y existen innumerables combinaciones únicas a considerar. Sin embargo, podemos ver que los modelos con arquitectura transformadora son capaces de aprender mejor que la generacion de modelos anterior.

Los enfoques de aprendizaje cruzado generalmente mejoran el rendimiento de los modelos multilingües, lo que indica su potencial para capturar propiedades semántico-sintácticas de colocaciones entre idiomas. Se observa también un mayor éxito en el reconocimiento de diferentes tipos de colocaciones con un mayor número de instancias de una función léxica particular en los datos de entrenamiento, aunque con excepciones.

[11] Traducción: "Así como puede impulsar las ventas de una empresa, una celebridad puede causar el efecto contrario."

Con los rápidos avances en el aprendizaje automático, la investigación futura aún espera explorar nuevas arquitecturas y combinaciones de datos de entrenamiento, teniendo en cuenta todos los desafíos potenciales que los modelos puedan enfrentar. Además, investigar los mecanismos internos de los NLMs nos acercará a mejorar y comprender las capacidades de los modelos de lenguaje neuronal en la identificación y clasificación de colocaciones.

Agradecimientos

Este artículo surgió a partir de TFM del Máster Europeo en Lexicografía. Quiero expresar mi agradecimiento a todos los afiliados del programa por su orientación y asistencia a lo largo de todo el proceso.

Referencias

Benson, Morton, Evelyn Benson & Robert Ilson. 2010. *The BBI combinatory dictionary of English.* John Benjamins Publishing Company. 10.1075/z.bbi

Carlini, Roberto, Joan Codina-Filba & Leo Wanner. 2014. Improving collocation correction by ranking suggestions using linguistic knowledge. En *3rd Workshop on NLP for Computer-Assisted Language Learning*, 1–12.

Chernodub, Artem, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann & Alexander Panchenko. 2019. TARGER: Neural argument mining at your fingertips. En *57th Annual Meeting of the Association for Computational Linguistics*, 195–200. 10.18653/v1/P19-3031

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2022. Unsupervised cross-lingual representation learning at scale. En *58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. 10.18653/v1/2020.acl-main.747

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. En *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. 10.18653/v1/N19-1423

Dražić, Jasmina. 2014. *Leksičke i gramatičke kolokacije u srpskom jeziku.* Filozofski fakultet Novi Sad.

Enikeeva, Ekaterina & Olga Mitrofanova. 2017. Russian collocation extraction based on word embeddings. En *International Conference Dialogue 2017: Computational Linguistics and Intellectual Technologies*, 52–64.

Espinosa-Anke, Luis, Joan Codina-Filba & Leo Wanner. 2021. Evaluating language models for the retrieval and categorization of lexical collocations. En *16th Conference of the European Chapter of the Association for Computational Linguistics*, 140–104. 10.18653/v1/2021.eacl-main.120

Espinosa-Anke, Luis, Alexander Shvets, Alireza Mohammadshahi & Leo Wanner. 2022. Multilingual extraction and categorization of lexical collocations with graph-aware transformers. En *11th Joint Conference on Lexical and Computational Semantics*, 89–100. 10.18653/v1/2022.starsem-1.8

Evert, Stefan. 2004. *The statistics of word cooccurrences: Word pairs and collocations*: University of Stuttgart. Tesis Doctoral

Evert, Stefan & Hannah Kermes. 2003. Experiments on candidate data for collocation extraction. En *10th Conference of The European Chapter of the Association for Computational Linguistics*, 83–86

García, Marcos, Marcos García-Salido, Susana Sotelo, Estela Mosqueira & Margarita Alonso-Ramos. 2019. Pay attention when you pay the bills. a multilingual corpus with dependency-based and semantic annotation of collocations. En *57th Annual Meeting of the Association for Computational Linguistics*, 4012–4019. 10.18653/v1/P19-1392

Gelbukh, Alexander & Olga Kolesnikova. 2012. *Semantic analysis of verbal collocations with lexical functions.* Springer. 10.1007/978-3-642-28771-8

Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next.... *International Journal of Corpus Linguistics* 18(1). 137–166. 10.1075/ijcl.18.1.09gri

Gutiérrez-Fandiño, Asier, Jordi Armengol-Estape, Marc Pamies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre & Marta Villegas. 2022. MarIA: Spanish language models. *Procesamiento Del Lenguaje Natural* 68. 39–60. 10.26342/2022-68-3

Hausmann, Franz Josef. 1998. Le dictionnaire de collocations. En *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*, 1010–1019. Walter de Gruyter

Kolesnikova, Olga. 2011. *Automatic extraction of lexical functions*: Instituto Politecnico Nacional — Centro de Investigacion en Computacion. Tesis Doctoral.

Krenn, Brigitte. 2000. CDB: A database of lexical collocations. En *2nd International Conference on Language Resources and Evaluation*,

Lin, Dekang. 1999. Automatic identification of noncompositional phrases. En *37th Annual of the Association for Computational Linguistics (ACL)*, 317–324. 10.3115/1034678.1034730

Mel'čuk, Igor. 1996. Lexical functions: A tool for the description of lexical relations in a lexicon. En *Lexical Functions in Lexicography and Natural Language Processing*, 37–102. John Benjamins

Mel'čuk, Igor. 2012. Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology* 3(1). 31–56. 10.1515/phras-2012-0003

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. En *Advances in Neural Information Processing Systems*, 3111–3119

Nesselhauf, Nadja. 2005. *Collocations in a learner corpus*. John Benjamins. 10.1075/scl.14

Nisho, Kosuke James. 2022. *Extraction and categorization of Japanese lexical collocations with graph-aware transformers*: Universidad Pompeu Fabra. Trabajo de Fin de Máster

Pawley, Andrew. 1985. On speech formulas and linguistic competence. *Lenguas Modernas* 12. 84–104

Pérez-Mayos, Laura, Miguel Ballesteros & Leo Wanner. 2021. How much pretraining data do language models need to learn syntax? En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7927–7934. 10.18653/v1/2021.emnlp-main.118

Rodríguez-Fernández, Sara, Luis Espinosa-Anke, Roberto Carlini & Leo Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. En *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 499–505. 10.18653/v1/P16-2081

Rychlý, Pavel. 2008. A lexicographer-friendly association score. En *Recent Advances in Slavonic Natural Language Processing*, 6–9.

Seretan, Violeta & Eric Wehrli. 2006. Accurate collocation extraction using a multilingual parser. En *21st International Conference on Computational Linguistics and 44th Annual of the Association for Computational Linguistics*, 317–324. 10.3115/1220175.1220295

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford University Press

Singh, Pranaydeep & Els Lefever. 2022. When the student becomes the master: Learning better and smaller monolingual models from mBERT. En *29th International Conference on Computational Linguistics*, 4434–4441.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1). 143–178.

Souza, Fabio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. En *Brazilian Conference on Intelligent Systems*, 403–417. 10.1007/978-3-030-61377-8_28

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aaidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. En *31st Conference on Neural Information Processing Systems*, 6000–6010. 10.5555/3295222.3295349

Wanner, Leo. 2004. Towards automatic fine-grained semantic classification of verb-noun collocations. *Natural Language Engineering* 10(2). 95–143. 10.1017/S1351324904003328

Wanner, Leo, Bernd Bohnet & Mark Giereth. 2006. Making sense of collocations. *Computer Speech & Language* 20(4). 609–624. 10.1016/j.csl.2005.10.002

Wolf, Tomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf,

Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. En *Conference on Empirical Methods in Natural Language Processing*, 38–45. 10.18653/v1/2020.emnlp-demos.6