Ajustando grandes modelos del lenguaje para el análisis de constituyentes mediante la traducción secuencia a secuencia

Fine-tuning of Large Language Models for Constituency Parsing Using a Sequence to Sequence Approach

Francisco José Cortés Delgado
Universidad de Murcia — MiSintaxis
fran.cortes@misintaxis.com

Resumen: Los avances recientes en el procesamiento del lenguaje natural mediante grandes modelos neuronales permiten explorar una nueva aproximación sintáctica del análisis de constituyentes basada en el aprendizaje automático. En este trabajo se propone el reentrenamiento de grandes modelos del lenguaje para el análisis de constituyentes empleando el enfoque de una traducción de una secuencia de entrada (la frase a analizar) a una secuencia de salida (su análisis de constituyentes). El objetivo final de esta técnica es ampliar las funcionalidades de la herramienta MiSintaxis (2023), desarrollada para la enseñanza de la sintaxis del español. Se ha realizado un ajuste de modelos disponibles en Hugging Face sobre los datos de entrenamiento generados a partir del corpus AnCora-ES y se han comparado los resultados mediante la métrica F1. Los resultados obtenidos indican una buena precisión en el análisis sintáctico de constituyentes, además de quedar patente el potencial de esta metodología.
Palabras clave: grandes modelos del lenguaje, reentrenamiento, traducción secuencia a secuencia, análisis sintáctico de constituyentes.

Abstract: Recent advances in natural language processing using large neural models enable investigating of a new syntactic approach to phrase-structure analysis based on machine learning. This work proposes fine tuning large language models for phrase-structure analysis by translating an input sequence (the sentence to be analyzed) into an output sequence (its phrase-structure analysis). The ultimate goal of this technique is to expand the functionalities of the MiSintaxis (2023) tool, designed for teaching Spanish syntax. Models available in Hugging Face have been fine-tuned on training data generated from the AnCora-ES corpus, and the results have been compared using the F_1 metric. The results indicate high precision in syntactic phrase-structure analysis while highlighting this methodology's potential.
Keywords: Large language models, Fine-tuning, Sequence to sequence, Constituency parsing.

1 Introducción

Tradicionalmente, el análisis de constituyentes ha empleado técnicas basadas en el algoritmo Cocke-Younger-Kasami (CYK). Sin embargo, la complejidad y ambigüedad inherentes a los lenguajes naturales han planteado desafíos significativos a esta aproximación. Este trabajo propone un nuevo enfoque basado en el análisis automático de la estructura gramatical de oraciones en español mediante grandes modelos del lenguaje, como Bloom o GPT-2, disponibles en Hugging Face. Estos modelos se pueden ajustar mediante un reentrenamiento para que realicen la tarea del análisis de constituyentes con un enfoque conocido como traducción *secuencia a secuencia*. Con los modelos ajustados es posible integrar un analizador sintáctico automático en las herramientas informáticas de enseñanza de la sintaxis del español, como MiSintaxis (2023), una aplicación dirigida a estudiantes de niveles preuniversitarios que cuenta actualmente con miles de usuarios en todo el mundo.

En la sección 2 se recogen las referencias más importantes sobre las que se ha desarrollado este trabajo. La sección 3 expone la preparación de los modelos ajustados y su evaluación. Por último, la sección 4 presenta las conclusiones y vías futuras de este trabajo.

2 Trabajos relacionados

Las herramientas tradicionales más importantes para el análisis de constituyentes han sido las gramáticas libres de contexto en forma

normal de Chomsky y los algoritmos basados en el método de programación dinámica Cocke–Younger–Kasami (CYK), que permiten establecer la estructura binaria del árbol sintáctico de una frase dada. La principal dificultad de este enfoque se encuentra en la elaboración de una gramática lo suficientemente expresiva como para describir todos los complejos fenómenos sintácticos de un lenguaje natural.

En la última década, el uso de redes neuronales profundas ha experimentado un crecimiento significativo en multitud de aplicaciones de procesado del lenguaje natural, especialmente a partir del desarrollo del mecanismo de atención (*self-attention*) que ha dado lugar a los grandes modelos del lenguaje actuales. Este mecanismo, que replica la atención cognitiva humana, fue introducido por Vaswani et al. (2017), superando en efectividad a técnicas similares previas que forman parte de las redes *Long Short-Term Model* (LSTM). Mediante la atención, presente en la arquitectura neuronal conocida como *transformer*, una parte de la red neuronal es capaz de determinar qué porciones del contexto previo son más relevantes para continuar generando la salida en el proceso de inferencia.

Las técnicas de atención no solo son útiles en la generación morfológica de las siguientes palabras, sino que también parecen aprender las categorías sintácticas del lenguaje, de acuerdo con Mrini et al. (2019). Por esta razón, los investigadores han comenzado a aplicarlas en el análisis de constituyentes. En Vinyals et al. (2014) se propone que la tarea del análisis de constituyentes se puede abordar de forma similar a la traducción entre lenguajes, con un enfoque conocido como *traducción secuencia a secuencia*. Dada una secuencia de entrada, la frase sin analizar, el modelo infiere una secuencia de salida que contiene el análisis de constituyentes de la entrada.

En concreto, como trabajo relacionado para el español destacamos la tesis doctoral de Chiruzzo (2020) y posterior trabajo de Chiruzzo y Wonsever (2020), en la que se comparan distintos métodos de análisis de constituyentes. Chiruzzo emplea una de las representaciones más ricas de los lenguajes naturales, denominada *Head-Driven Phrase Structure Grammar* (HPSG) Pollard y Sag (1994), con la que se anota no solo la estructura sintáctica sino también propiedades semánticas. Para el entrenamiento de sus modelos, Chiruzzo hace uso del corpus AnCora-ES, elaborado por Taulé, Peris, y Rodríguez (2016). Sus evaluaciones muestran que el enfoque basado en LSTM es el más eficiente.

Por otra parte, en el caso del inglés se ha observado una mejora notable cuando el codificador LSTM se reemplaza por una arquitectura basada en transformers con *self-attention*, permitiendo que el modelo capture el contexto global sin usar redes neuronales recurrentes, como se explica en el trabajo de Kitaev y Klein (2018). Por esta razón, en el presente trabajo se hace uso de modelos de transformers en lugar de redes LSTM para el análisis de constituyentes, siendo la primera contribución en esta línea para el caso particular del español.

Los grandes modelos del lenguaje disponibles en plataformas como Hugging Face, que han contribuido enormemente al progreso en el procesamiento del lenguaje natural, se desarrollan con un preentrenamiento autosupervisado a gran escala, como queda descrito en Devlin et al. (2018). Sin embargo, estos grandes modelos deben ser reentrenados para aumentar su efectividad en un ámbito de aplicación concreto, como se ha demostrado en numerosos trabajos. Baste citar Dai y Le (2015), Peters et al. (2018), Radford y Narasimhan (2018), Howard y Ruder (2018). Por esta razón, este trabajo comenzó con la selección y reentrenamiento de varios grandes modelos del lenguaje, como se explica en la siguiente sección.

3 Resolución del trabajo

Para reentrenar un modelo del lenguaje se requiere un corpus adecuado a la tarea en la que se quiere aplicar. En este trabajo se ha tomado como punto de partida el corpus AnCora-ES, con aproximadamente 500000 palabras y 17300 frases, en su mayoría compuesto por artículos periodísticos. Contiene niveles de anotaciones morfológicas, sintácticas y semánticas en formato XML, al igual que identificación de entidades y correferencias entre constituyentes. Para este trabajo se han empleado las etiquetas y atributos XML que identifican funciones sintácticas y, en algunos casos, morfológicas.

El corpus ha sido adaptado con el fin de realizar un análisis sintáctico similar al que tiene lugar en las aulas españolas, de acuerdo con la notación de la *Nueva gramática de la lengua española* (RAE, 2011). El corpus adaptado se ha representado con un formato semejante al del Penn Treebank (Marcus, Marcinkiewicz, y Santorini, 1993) en el que las estructuras sintácticas se delimitan con paréntesis que contienen un primer elemento que etiqueta la estructura y una lista de elementos separados por espacios simples que representan el contenido. Únicamente se ha tomado la precaución de cambiar los paréntesis por cor-

chetes porque en español los primeros pueden ser usados como signos de puntuación. Un ejemplo de la notación usada es la siguiente:

```
<s>La final de copa entre Inglaterra y
    Alemania ha tenido un efecto positivo,
    aunque haya pasado casi inadvertido.
</s>
<s>[O.Compuesta [GN/S [Det La] [N final]
    [GPrep/CN [E de] [GN/T [N copa]]]
    [GPrep/CN [E entre] [GN/T [N Inglaterra
    y Alemania]]]] [GV/PV [NP ha tenido]
    [GN/CD [Det un] [N efecto] [GAdj/CN
    [Adj positivo]]] [OS.Adverbial/AP [Punt
    ,] [nx aunque] [SO él] [GV/PV [NP haya
    pasado] [GAdj/PVO [GAdv [Adv casi]]
    [Adj inadvertido]]]]] [Punt .]]
</s>
```

Cuatro modelos de Hugging Face han sido reentrenados con este corpus: bigscience/bloom-560m, bigscience/bloom-1b1 (Scao et al., 2022), PlanTL-GOB-ES/gpt2-base-bne y PlanTL-GOB-ES/gpt2-large-bne (Gutiérrez-Fandiño et al., 2021). En la tabla 1 se pueden ver las características de los modelos. Es interesante comentar que un mayor número de parámetros no garantiza mejores resultados. Podemos ver que el número de tokens máximo de los modelos basados en GPT-2 es inferior al de los modelos basados en Bloom. Los primeros cuentan con la limitación de 512 tokens en la entrada, de modo que no es posible emplear el corpus en su totalidad, puesto que la mayor frase se compone de 1239 tokens. El conjunto de datos usado en los modelos GPT-2 consta de 15035 frases, mientras que en el caso de los modelos de Bloom se han usado 17300 frases. Esto ha hecho que realicemos el entrenamiento y evaluación de dos formas. Una con todas las frases del corpus y otra con las frases con la limitación de los 512 tokens. En ambos casos se han empleado el 80 % de las frases durante el reentrenamiento, dejando el 20 % restante para el test.

En la tabla 2 se pueden ver otros datos relativos al reentrenamiento: el tiempo en segundos que ha tardado, así como la memoria necesaria en el reentrenamiento y la perdida final en la última época del proceso. La figura 1 muestra la evolución de la medida de error en los cuatro modelos durante el reentrenamiento, que tuvo lugar en una máquina con una tarjeta NVIDIA A100 GPU con 40 GB de RAM. Se decidió emplear cinco épocas para evitar el *overfitting*. No obstante, queda como vía futura de este trabajo el realizar un estudio exhaustivo para comprobar si es posible mejorar el reentrenamiento evitando este problema. En la tabla 3 se puede ver el tiempo medio que tarda cada modelo en inferir una frase del conjunto de

Modelo	**Parámetros**	**Entrada máx.**
gpt2-base-bne	117 mills.	512 tokens
gpt2-large-bne	774 mills.	512 tokens
bloom-560m	559 mills.	2048 tokens
bloom-1b1	1065 mills.	2048 tokens

Tabla 1: Características de los modelos usados

Modelo	**Tiempo**	**Error**	**Memoria**
gpt2-base-bne	1997.65 s.	0.0472	4295 MB
gpt2-large-bne	11268.19 s.	0.0253	19883 MB
bloom-560m	22212.93 s.	0.0175	23307 MB
bloom-1b1	36971.62 s.	0.0177	32867 MB

Tabla 2: Reentrenamiento de los modelos

test, así como la cantidad de memoria requerida para la inferencia y su métrica F_1 para el corpus con todas las frases y para el corpus con la limitación de 512 tokens. Se puede observar que apenas hay cambio entre un corpus y otro. El mejor modelo es gpt2-large-bne en cuanto a F_1, si bien bloom-560m obtiene un resultado parecido con un tiempo medio de inferencia bastante inferior. La figura 2a muestra gráficamente el resultado correcto obtenido al analizar una frase compuesta, mientras la figura 2b ejemplifica la dificultad que puede encontrarse al analizar una frase ambigua (verbo en indicativo o imperativo).

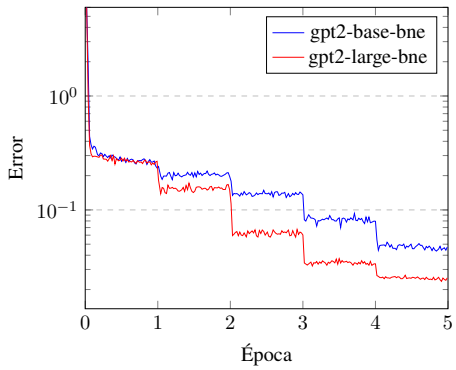
4 Conclusiones y vías futuras

Con este estudio se pretendía explorar la viabilidad del análisis de constituyentes mediante grandes modelos del lenguaje reentrenados y usados con el enfoque de traducción secuencia a secuencia. La conclusión principal es que es un método prometedor que apunta hacia el hecho de que los grandes modelos del lenguaje pueden representar los rasgos sintácticos del español. Las futuras investigaciones podrían explorar métodos alternativos, como la combinación de grandes modelos del lenguaje con el algoritmo de CYK.

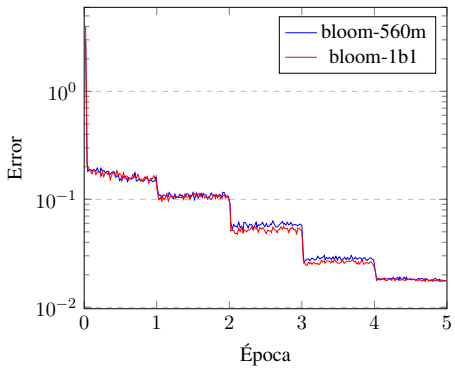
Con el fin de continuar mejorando los resultados obtenidos, se procurará enriquecer el corpus

Modelo	**Memoria**	**Tiempo**	F_1	F_1 **(512)**
gpt2-base-bne	1984 MB	1.9420 s.	0.7234	0.7222
gpt2-large-bne	4582 MB	5.2488 s.	0.8141	0.8183
bloom-560m	3606 MB	2.9910 s.	0.7963	0.7939
bloom-1b1	5584 MB	3.0467 s.	0.7792	0.7665

Tabla 3: Inferencia con el dataset de AnCora-ES sin limitación y con limitación de 512 tokens

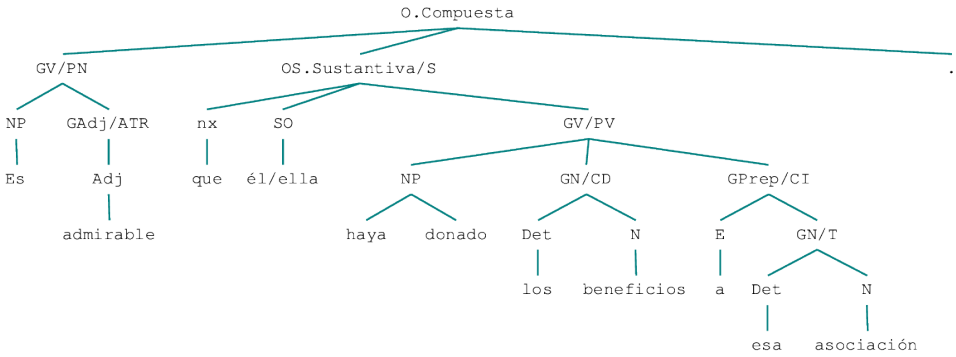


(a) Modelos gpt2

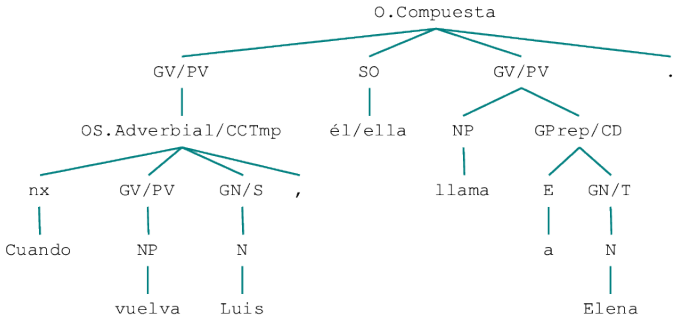


(b) Modelos bloom

Figura 1: Evolución del error en el entrenamiento



(a) Frase compuesta analizada con bloom-560m



(b) Frase ambigua analizada con gpt2-large

Figura 2: Ejemplos de análisis sintácticos realizados por los modelos

con más frases, especialmente las que pueden corresponder a un nivel próximo al estudio de la sintaxis a nivel preuniversitario. Además de lo mencionado anteriormente, también hay ciertos componentes lingüísticos de nueva aparición en la gramática del español, como el *Complemento Circunstancial de Compañía*, que no está etiquetado en AnCora-ES y que requiere, por tanto, de nuevos ejemplos en el corpus de reentrenamiento.

Agradecimientos

Quisiera agradecer a Eduardo Martínez Graciá y a Rafael Valencia García, por su ayuda como co-tutores del trabajo fin de grado que da lugar a este artículo. Agradezco igualmente a los lingüistas Pascual Cantos, Santiago Roca, Ana Bravo y Alejandra Valenciano, por haber compartido conmigo sus sugerencias. Y por último, quisiera agradecer el apoyo de los integrantes de MiSintaxis, Gonzalo Cánovas López de Molina, Laura Mateo Galindo, Tomás Bernal Beltrán y Mario

Rodríguez Béjar.

Bibliografía

[Chiruzzo2020] Chiruzzo, L. 2020. *Statistical Deep Parsing for Spanish*. Ph.D. tesis, Universidad de la República (Uruguay). Facultad de Ingeniería.

[Chiruzzo y Wonsever2020] Chiruzzo, L. y D. Wonsever. 2020. Statistical deep parsing for Spanish using neural networks. En *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, páginas 132–144, Online, Julio. Association for Computational Linguistics.

[Dai y Le2015] Dai, A. M. y Q. V. Le. 2015. Semi-supervised sequence learning. *CoRR*, abs/1511.01432.

[Devlin et al.2018] Devlin, J., M. Chang, K. Lee, y K. Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Gutiérrez-Fandiño et al.2021] Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, y M. Villegas. 2021. Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.

[Howard y Ruder2018] Howard, J. y S. Ruder. 2018. Fine-tuned language models for text classification. *CoRR*, abs/1801.06146.

[Kitaev y Klein2018] Kitaev, N. y D. Klein. 2018. Constituency parsing with a self-attentive encoder. *CoRR*, abs/1805.01052.

[Marcus, Marcinkiewicz, y Santorini1993] Marcus, M. P., M. A. Marcinkiewicz, y B. Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, jun.

[MiSintaxis2023] MiSintaxis. 2023. Misintaxis. `https://misintaxis.com/`. Accessed: 2023-05-17.

[Mrini et al.2019] Mrini, K., F. Dernoncourt, T. Bui, W. Chang, y N. Nakashole. 2019. Rethinking self-attention: An interpretable self-attentive encoder-decoder parser. *CoRR*, abs/1911.03875.

[Peters et al.2018] Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, y L. Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

[Pollard y Sag1994] Pollard, C. y I. A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

[Radford y Narasimhan2018] Radford, A. y K. Narasimhan. 2018. Improving language understanding by generative pre-training.

[RAE2011] RAE. 2011. *Nueva gramatica BASICA de la lengua española*. Espasa Libros.

[Scao et al.2022] Scao, T. L., A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, y others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

[Taulé, Peris, y Rodríguez2016] Taulé, M., A. Peris, y H. Rodríguez. 2016. Iarg-ancora: Spanish corpus annotated with implicit arguments. En *Language Resources and Evaluation, Vol. 50(3): 549-584, Springer-Verlag, Netherlands*.

[Vaswani et al.2017] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, y I. Polosukhin. 2017. Attention is all you need. En I. Guyon U. V. Luxburg S. Bengio H. Wallach R. Fergus S. Vishwanathan, y R. Garnett, editores, *Advances in Neural Information Processing Systems*, volumen 30. Curran Associates, Inc.

[Vinyals et al.2014] Vinyals, O., L. Kaiser, T. Koo, S. Petrov, I. Sutskever, y G. E. Hinton. 2014. Grammar as a foreign language. *CoRR*, abs/1412.7449.