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

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

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 Introduction

Traditionally, constituent analysis has employed techniques based on the Cocke-Younger-Kasami (CYK) algorithm. However, the complexity and inherent ambiguity of natural languages have posed significant challenges to this approach. This paper proposes a new approach based on the automatic analysis of the grammatical structure of Spanish sentences using large language models, such as Bloom or GPT-2, available on Hugging Face. These models can be fine-tuned through retraining to perform the task of constituent analysis with an approach known as sequence-to-sequence translation. With the adjusted models, it is possible to integrate an automatic syntactic parser into the computer tools for teaching Spanish syntax, such as MiSintaxis (2023), an application aimed at pre-university level students that currently has thousands of users worldwide.

In section 2, the most important references on which this work has been developed are collected. Section 3 explains the preparation of the adjusted models and their evaluation. Finally, section 4 presents the conclusions and future paths of this work.

2 Related work

The most important traditional tools for constituent analysis have been context-free grammars in Chomsky's normal form and algorithms based on the dynamic programming method Cocke–Younger–Kasami (CYK), which establish the binary structure of the syntactic tree of a given

sentence. The main difficulty of this approach lies in developing a grammar expressive enough to describe all the complex syntactic phenomena of a natural language.

In the last decade, the use of deep neural networks has experienced significant growth in numerous applications of natural language processing, especially since the development of the self-attention mechanism that has led to the current large language models. This mechanism, which replicates human cognitive attention, was introduced by Vaswani et al. (2017), surpassing in effectiveness previous similar techniques that are part of the Long Short-Term Model (LSTM) networks. Through attention, present in the neural architecture known as transformer, a part of the neural network can determine which portions of the previous context are most relevant for continuing to generate the output in the inference process.

Attention techniques are not only useful in the morphological generation of the following words, but they also seem to learn the syntactic categories of the language, according to Mrini et al. (2019). For this reason, researchers have begun to apply them in constituent analysis. In Vinyals et al. (2014), it is proposed that the task of constituent analysis can be approached similarly to translation between languages, with an approach known as sequence-to-sequence translation. Given an input sequence, the unanalyzed sentence, the model infers an output sequence that contains the constituent analysis of the input.

Specifically, as related work for Spanish, we highlight the doctoral thesis of Chiruzzo (2020) and subsequent work of Chiruzzo y Wonsever (2020), in which different methods of constituent analysis are compared. Chiruzzo uses one of the richest representations of natural languages, called Head-Driven Phrase Structure Grammar (HPSG) Pollard y Sag (1994), with which not only the syntactic structure but also semantic properties are annotated. For the training of his models, Chiruzzo makes use of the AnCora-ES corpus, developed by Taulé, Peris, y Rodríguez (2016). His evaluations show that the LSTM-based approach is the most efficient.

On the other hand, in the case of English, a noticeable improvement has been observed when the LSTM encoder is replaced by an architecture based on transformers with self-attention, allowing the model to capture the global context without using recurrent neural networks, as explained in the work of Kitaev y Klein (2018). For this reason, in this work, we use transformer models instead of LSTM networks for constituent analysis, being the first contribution in this line for the particular case of Spanish.

The large language models available on platforms like Hugging Face, which have contributed enormously to progress in natural language processing, are developed with large-scale self-supervised pre-training, as described in Devlin et al. (2018). However, these large models need to be retrained to increase their effectiveness in a specific application field, as demonstrated in numerous works. Suffice to mention Dai y Le (2015), Peters et al. (2018), Radford y Narasimhan (2018), Howard y Ruder (2018). For this reason, this work began with the selection and retraining of several large language models, as explained in the next section.

3 Resolution of the work

To retrain a language model, a suitable corpus for the task at hand is required. In this work, we have taken as a starting point the AnCora-ES corpus, with approximately 500,000 words and 17,300 sentences, mostly composed of journalistic articles. It contains morphological, syntactic, and semantic annotation levels in XML format, as well as entity identification and coreference between constituents. For this work, we have used XML tags and attributes that identify syntactic functions and, in some cases, morphological ones.

The corpus has been adapted to carry out a syntactic analysis similar to that which takes place in Spanish classrooms, in accordance with the notation of the *New Spanish language grammar* (RAE, 2011). The adapted corpus has been represented in a format similar to that of the Penn Treebank (Marcus, Marcinkiewicz, y Santorini, 1993) in which syntactic structures are delimited by parentheses containing a first element that labels the structure and a list of elements separated by simple spaces that represent the content. The only precaution taken was to change the parentheses for brackets because in Spanish the former can be used as punctuation marks. An example of the notation used is the following:

```
<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>
```

Four Hugging Face models have been retrained with this corpus: bigscience/bloom-560m, bigscience/bloom-1b1 (Scao et al., 2022), PlanTL-GOB-ES/gpt2-base-bne, and PlanTL-GOB-ES/gpt2-large-bne (Gutiérrez-Fandiño et al., 2021). The characteristics of the models can be seen in Table 1. It is interesting to note that a larger number of parameters does not guarantee better results. We can see that the maximum token number of the GPT-2-based models is lower than that of the Bloom-based models. The former have a limitation of 512 tokens in the input, so it is not possible to use the entire corpus, as the longest sentence consists of 1239 tokens. The dataset used in the GPT-2 models consists of 15035 sentences, while in the case of the Bloom models, 17300 sentences have been used. This has led us to carry out the training and evaluation in two ways. One with all the sentences from the corpus and another with the sentences with the limitation of 512 tokens. In both cases, 80 % of the sentences were used during retraining, leaving the remaining 20 % for testing.

Table 2 shows other data related to retraining: the time in seconds it took, as well as the memory required in retraining and the final loss in the last epoch of the process. Figure 1 shows the evolution of the error measure in the four models during retraining, which took place on a machine with an NVIDIA A100 GPU with 40 GB of RAM. It was decided to use five epochs to avoid *overfitting*. However, it remains as a future direction of this work to conduct an exhaustive study to check if it is possible to improve retraining while avoiding this problem. Table 3 shows the average time it takes each model to infer a sentence from the test set, as well as the amount of memory required for inference and its F_1 metric for the corpus with all sentences and for the corpus with the limitation of 512 tokens. It can be seen that there is barely any change between one corpus and another. The best model is gpt2-large-bne in terms of F_1, although bloom-560m achieves a similar result with a much lower average inference time. Figure 2a graphically shows the correct result obtained when analyzing a compound sentence, while figure 2b exemplifies the difficulty that can be encountered when analyzing an ambiguous sentence (verb in indicative or imperative).

Model	**Parameters**	**Max. input**
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: Characteristics of the models used

Model	**Time**	**Error**	**Memory**
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: Model fine tuning

4 Conclusions and Future Directions

The goal of this study was to explore the feasibility of constituent analysis through large language models retrained and used with the sequence-to-sequence translation approach. The main conclusion is that it is a promising method that points towards the fact that large language models can represent the syntactic features of Spanish. Future research could explore alternative methods, such as combining large language models with the CYK algorithm.

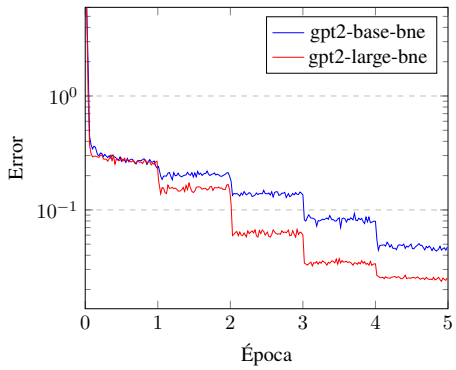
In order to continue improving the obtained results, efforts will be made to enrich the corpus with more sentences, especially those that could correspond to a level close to the study of syntax at a pre-university level. In addition to what has been mentioned above, there are also certain newly appearing linguistic components in the Spanish grammar, such as the *Circumstantial Complement of Company*, which is not labeled in AnCora-ES and therefore requires new examples in the retraining corpus.

Acknowledgments

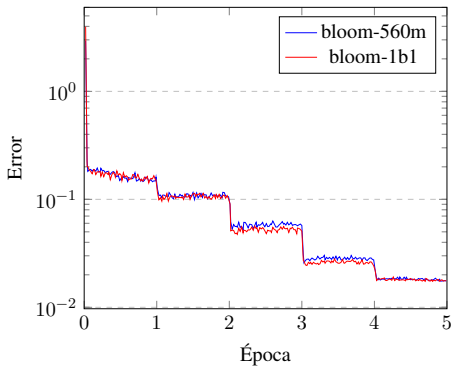
I would like to thank Professor Eduardo Martínez Graciá and Dr. Rafael Valencia García for their assistance as co-advisors of the final degree project that gives rise to this article. I also appreciate lin-

Model	**Memory**	**Time**	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: Inference with the AnCora-ES dataset without limitation and with 512 token limitation

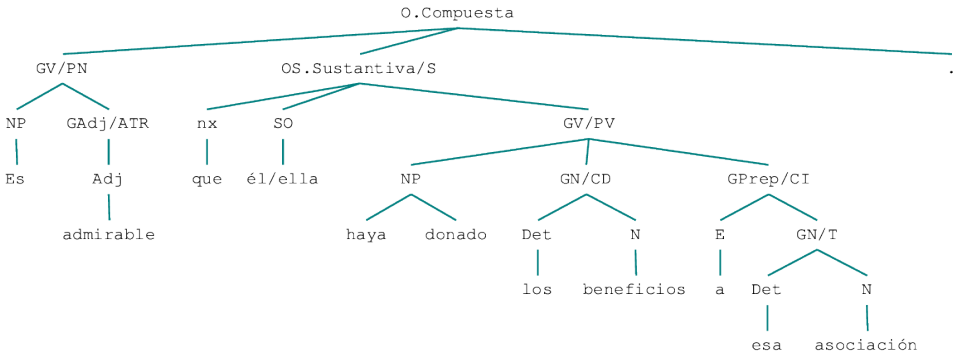


(a) GPT2 models

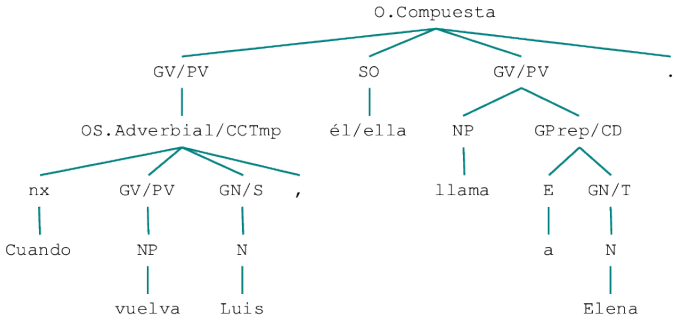


(b) Bloom models

Figura 1: Evolution of training error



(a) Composite phrase analyzed with bloom-560m



(b) Ambiguous sentence analyzed with gpt2-large

Figura 2: Examples of parsing performed by the models

guists Dr. Pascual Cantos, Dr. Santiago Roca, Dr. Ana Bravo, and Alejandra Valenciano, for sharing their suggestions with me. Finally, I would like to acknowledge the support of the members of MiSintaxis, Gonzalo Cánovas López de Molina, Laura Mateo Galindo, Tomás Bernal Beltrán, and 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.