BERTSIFICATION
Language modeling fine-tuning for Spanish scansion

Authors: Javier de la Rosa[1]*, Salvador Ros[1], Elena González-Blanco[2]

Affiliations:

[1] Universidad Nacional de Educación a Distancia, Spain.
[2] CoverWallet Inc., Spain.

*Corresponding author. E-mail: versae@linhd.uned.es

Abstract.

Since ancient times, cultures have maintained oral traditions by sharing stories in poetic form (Francis 2017). Although different traditions treat poetry according to their own thinking frameworks, most of them perform some sort of analysis that requires the extraction of stress patterns of lines or verses. Scansion, as this procedure is broadly referred to, is usually language dependent. On the assumption that modern context-dependent language models are able to retain the structural properties of text, this paper explores whether metrical patterns are part of the information encoded in the embedding spaces of text. We used three of the best performing pre-trained language models –namely: BERT (Devlin et al 2019), DistilBERT (Sanh et al 2019), and RoBERTa (Liu et al 2019)–, since they are proven to work very well in a variety of tasks, from question answering to named entity recognition. Specifically, we evaluated accuracy after fine-tuning the models on the task of predicting syllabic stress on a corpus of hendecasyllable Spanish verses (Navarro-Colorado 2016), designed as a multilabel classification tasks where each syllable has a label indicating whether it is stressed or not. We then compared these results to previous rule-based systems for Spanish scansion. Table 1 shows the per line accuracy, measured as whole stress patterns matches, of fine-tuning the different pre-trained models.

Model	**Score**
Gervás 2000	88.73
Navarro-Colorado 2017	94.44
Agirrezabal 2017	90.84

distilbert-base-multilingual-cased	91.79
bert-base-multilingual-cased	93.71
roberta-base	94.64
roberta-large	**96.35**

Table 1. Scores on Navarro-Colorado’s fixed-metre poetry corpus. Best score in bold (ours).

Our results achieve a new state of the art on automatic scansion of fixed-metre for Spanish poetry and open a new exciting path for multilingual scansion via fine-tuning of language models.

Bibliography.

Agirrezabal, M., Alegria, I., and Hulden, M. (2017). A comparison of feature-based and neural scansion of po-etry.arXiv preprint arXiv:1711.00938

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Francis, Norbert. *Bilingual and Multicultural Perspectives on Poetry, Music, and Narrative: The Science of Art*. Lexington Books, 2017.

Gervas, P. (2000). A logic programming application for the analysis of spanish verse. InInternational Conference on Computational Logic, pages 1330–1344. Springer.

Navarro-Colorado, Borja, María Ribes Lafoz and Noelia Sánchez (2016) “Metrical annotation of a large corpus of Spanish sonnets: representation, scansion and evaluation”, Proceedings of the 10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016, Portorož (Slovenia).

Navarro-Colorado, B. (2017). A metrical scansion system for fixed-metre spanish poetry. Digital Scholarship in the Humanities, 33(1):112–127.

Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).

Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108* (2019).

Bios

Elena González-Blanco is the Director of the Digital Humanities Lab at UNED: LINHD. Her main research and teaching areas are Comparative Medieval Literature, Metrics and Poetry, and Digital Humanities.

Salvador Ros is Vice-Dean of Technologies at the School of Computer Science. His research and professional activity in general is focused on enhanced learning technologies for distance learning scenarios,learning analytics and big data in education and text analysis.

Javier de la Rosa holds a Ph.D. in Hispanic Studies by the University of Western Ontario and a M.Sc. in Artificial Intelligence and Logic by the University of Seville. He worked as a Research Engineer at Stanford University, and as Technical Lead at the CulturePlex in Canada. At UNED LINHD, he is a member of the POSTDATA Project where he works as a Natural Language Processing Post-doctoral Fellow. His interests span from Corpus Linguistics or Authorship Attribution to Artificial Intelligence and Machine Learning applied to the Humanities.