Predicting metrical patterns in Spanish poetry with language models

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we compare automated metrical pattern identification systems available for Spanish against extensive experiments done by fine-tuning language models trained on the same task. Despite being initially conceived as a model suitable for semantic tasks, our results suggest that BERT-based models retain enough structural information to perform reasonably well for Spanish scansion.

1 Introduction

We can consider the metre of a verse as a sequence of stressed (strong) and unstressed (weak) syllables, which are sometimes denoted with the plus symbol '+' for stressed syllables and the minus '−' for the unstressed ones. Example 1 shows an hendecasyllabic Spanish verse and the resulting metrical pattern after applying rhetorical figures that might shrink (synalepha) or expand (syneresis) its length. The stress of the last word also affects the metrical length in Spanish poetry. The identification of metrical patterns is part of a larger procedure for the scansion of a poem.

(1) *cubra de nieve la hermosa cumbre*[1]
     ***cu**-bra-de-**nie**-ve-la-her-**mo**-sa-**cum**-bre*
     $+ - - + - - - + - + -$ 11
     (Garcilaso de la Vega)

Research has shown that neural models implicitly encode linguistic features ranging from token labeling to different kinds of segmentation [9]. There is also evidence that language models and embeddings are able to capture not only semantic and syntactic properties but structural information, as shown by Hewitt and Manning [8] in their work with structural probes for extracting syntax trees, and Conneau et al. [4] approximating the length in words of a sentence by its vector.

On the other hand, the earliest computational approach to Spanish scansion was introduced by Gervás [7], whose tool uses DCG (Definite Clause Grammars) to model word syllabification as well as additional predicates to define synalepha, syllable count and rhyme. More recently, the ADSO Scansion system introduced by Navarro-Colorado [10] first applies part of speech (PoS) tags to the words of every line in a poem to analyze hendecasyllables. Rantanplan [5] employs a similar system but focused on accuracy and speed and achieves state of the art results. The only neural approach was explored by Agirrezabal et al. [1, 2], who used bi-LSTM neural networks and CRF's to automatically scan poetry in three languages (i.e. English, Spanish and Basque).

---

[1]"[It] covers with snow the beautiful summit."

2 Experimental Design and Results

As a corpus for Spanish, we decided to use the *Corpus de Sonetos de Siglo de Oro* "Golden Age Spanish corpus" [11]. Although most of the poems included were annotated automatically, it includes 730 poems with manually annotated metrical information, consisting of over 71,000 lines. From this corpus, a subset of 100 poems was used to evaluate Navarro-Colorado [10]'s automated scansion system. We used only this subset, which we consider a gold standard for Spanish metrical annotation applied to hendecasyllables. This corpus, annotated in TEI-XML, contains sonnets from canonical Golden Age Spanish authors (16th and 17th centuries), featuring only hendecasyllabic verses.

Our downstream task is defined as metrical pattern prediction, that is, given a raw string of text representing a line of verse of a poem, a model is expected to predict a string of $+$ and $-$ symbols representing the stress of each syllable after any rhetorical device has been applied. We split the corpus into training (6,558 lines), evaluation (2,187 lines), and test (1,401 lines) sets. Using the training and evaluation sets, we fine-tuned a series of BERT-based models and a baseline classifier using fasttext [3] with multi-class binary classification. Specifically, we added a fully connected layer to predict the presence or absence of stress in each of the 11 positions of the hendecasyllabic verses in the corpus. We used the language Python 3, the library PyTorch, and the framework Transformers [12]. We pre-processed all texts removing duplicate verses and punctuation marks since they are irrelevant for metrical purposes. We trained the models using an AdamW optimiser with a learning rate set to 3e-5 [6], gradient clipping of 1.0, early stopping set at 5, and for 100 epochs. Training was run on a 30GB of RAM memory with one NVIDIA Tesla V100 GPU with 16GB of memory running on Debian 10. The maximum sequence length was set at 32 tokens and the batch size at 16.

Instead of calculating standard measures for multi-class classification, which in our case would produce per-syllable metrics, we decided to consider as a correct prediction only when all the individual syllable predictors were correct per-line. Therefore, we are reporting accuracy expressed as a percentage of the testing set.

| | Accuracy | |
Method	10 Epochs	100 Epochs
Baseline (fasttext)	10.89	11.20
Multilingual BERT	73.45	85.15
RoBERTa (base)	76.52	87.37
RoBERTa (large)	**87.58**	**93.08**
XLM RoBERTa (base)	61.53	82.16
XLM RoBERTa (large)	85.15	91.51
Spanish BERT	47.47	72.73
Rantanplan [5] (SOTA)	**96.23**	

Table 1: Accuracy of the different methods. Best scores in bold

As illustrated in Table 1, results show that BERT-based models are capable of predicting metrical patterns 9 times better tha baseline and reasonably well soon after training starts, performing close to state of the art rule-based systems for longer training times.

3 Conclusions and Further Work

In this paper we have evaluated the capabilities of BERT-based models when trained on the task of predicting the metrical pattern of a verse. Under the assumption that transformed-based models were capable of performing tasks of structural nature beyond those of the semantic kind, we show that BERT models perform reasonably well for Spanish. These results suggest that extended pre-training on poetic corpora and hyperparameters search could further improve on the task of metrical pattern prediction. Moreover, there seems to be advantages on the use of multilingual models over the monolingual Spanish BERT, which suggests that important information on the prosody and syllabification of languages is somehow shared for the task of stress assignment.

References

[1] Manex Agirrezabal, Aitzol Astigarraga, Bertol Arrieta, and Mans Hulden. Zeuscansion: a tool for scansion of english poetry. *Journal of Language Modelling*, 4, 2016.

[2] Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. A comparison of feature-based and neural scansion of poetry. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, Ranlp 2017*, pages 18–23, 2017.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017.

[4] Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.

[5] Javier de la Rosa, Álvaro Pérez, Laura Hernández, Salvador Ros, and Elena González-Blanco. Rantanplan, fast and accurate syllabification and scansion of spanish poetry. *Procesamiento del Lenguaje Natural*, 65:83–90, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Pablo Gervás. A logic programming application for the analysis of spanish verse. In *International Conference on Computational Logic*, pages 1330–1344. Springer, 2000.

[8] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *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)*, pages 4129–4138, 2019.

[9] Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*, 2019.

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

[11] Borja Navarro-Colorado, María Ribes Lafoz, and Noelia Sánchez. Metrical annotation of a large corpus of spanish sonnets: representation, scansion and evaluation. In *International Conference on Language Resources and Evaluation*, pages 4360–4364, 2016.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910, 2019.