Predicting metrical patterns in Spanish poetry with language models

LXAI @ NeurIPS 2020

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.

Syneresis

y al ver sonreír los astros, me prosterno

y al ver son-re-ír los as-tros, me pros-ter-no

− − − + − + − − − + − 11

(Manuel de Montoliu)

Dieresis

en cánticos y nácares süaves

en cán-ti-cos y ná-ca-res sü-a-ves

− + − − − + − − − + − 11

(Fray Jerónimo de San José)

Synalepha

Cuando el alba me despierta

*Cuan-do el-**al**-ba me des-**pier**-ta*

− − + − − − +− 8

(Miguel de Unamuno)

Rule-based approaches

1) PoS tagging
2) Syllabification
3) Stress and phonological groups
4) Metrical adjustment

Performance on "Golden Age Spanish Corpus"

Method	Accuracy (10 Epochs)	Accuracy (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
Gervas	70.88	
Navarro-Colorado	94.45	
Agirrezabal	90.84	
Rantanplan (SOTA)	**96.23**	

Javier de la Rosa (versae@linhd.uned.es) Elena González Blanco (egonzalezblanco@faculty.ie.edu) Salvador Ros (sros@scc.uned.es)