A Hybrid Approach to Stanza Classification in Spanish Poetry

EADH2021
"Interdisciplinary Perspectives on Data"
2nd International Conference of the European Association for Digital Humanities Krasnoyarsk, Russia | 21 - 25 September 2021

Javier de la Rosa (versae@linhd.uned.es)
Álvaro Pérez (alvaro.perez@linhd.uned.es)
Laura Hernández (laura.hernandez@scc.uned.es)
Mirella de Sisto (mdesisto@scc.uned.es)
Salvador Ros (sros@scc.uned.es)
Elena González Blanco (egonzalezblanco@faculty.ie.edu)

Outline

POSTDATA
Poetry Standardization
and Linked Open Data

- Introduction
- Classification of stanzas
- Corpus
- Methods
- Discussion
- Conclusions

Introduction

POSTDATA
Poetry Standardization and Linked Open Data

- Analysis of poetry relies on the extraction of information from the different structures found in a poem
- It is possible to identify these structures automatically with the help of a computer (Gervás, 2000; Araújo & Mamede, 2002; McAleese, 2007; Heuser & Antiila, 2010; Ibrahim & Plecháč, 2011; Agirrezabal, 2016; De la Rosa et al, 2020; De Sisto, 2020)
- Most approaches use rule-based systems
- These tools focus on automatic scansion: verse length and rhyme (except for Araújo and Mamede 2002)

Introduction

POSTDATA
Poetry Standardization and Linked Open Data

- A **stanza** is the minimal structural unit of a poem that usually encapsulates themes or ideas (Kirszner, 2013)
- Stanzas complement the metrical information of a poem
- Automatic identification of stanza types remains understudied

Introduction

POSTDATA
Poetry Standardization and Linked Open Data

- We framed this problem as a classification task
- Approaches from traditional computational methods to artificial intelligence-based solutions (NLP)
- On a corpus of Spanish poems

Classification of Stanzas

POSTDATA
Poetry Standardization and Linked Open Data

- Stanzas are structural units formed by verses
- Affected by author style and historical preferences
- Stanzas as expressive elements of a poem (Jauralde, 2020)

Classification of Stanzas

POSTDATA
Poetry Standardization and Linked Open Data

- Three aspects determine how a stanza is identified in the Spanish tradition (Domínguez Caparrós, 2014; Jauralde, 2020; Quilis, 2000; Torre, 2000)
  - verse length
  - rhyme type
  - rhyme pattern

Classification of Stanzas

POSTDATA
Poetry Standardization and Linked Open Data

- Stanza classification can be formulated in three stages (Domínguez Caparrós, 2014):

1. Calculation of per verse metrical length
2. Determining the rhyme type
3. Extraction of the rhyme pattern

Classification of Stanzas

POSTDATA
Poetry Standardization and Linked Open Data

1. Calculation of per verse metrical length

Original verse

Pongo estos versos en mi botella al mar **(I put these verses in my bottle to the sea)**

Length according to orthographic separation

1	2	3	4	5	6	7	8	9	10	11	12	13
Pon	go	es	tos	ver	sos	en	mi	bo	te	lla	al	mar

Metrical lengths for $n = 2$

Pon	goes	tos	ver	sos	en	mi	bo	te	llaal	mar	
Pon	goes	tos	ver	sos	en	mi	bo	te	lla	al	mar
Pon	go	es	tos	ver	sos	en	mi	bo	te	llaal	mar
Pon	go	es	tos	ver	sos	en	mi	bo	te	lla	al	mar

Introduction

POSTDATA
Poetry Standardization and Linked Open Data

- Synalepha

Cuando el alba me despierta

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

– – + – – – +– 8

(Miguel de Unamuno)

Introduction

POSTDATA
Poetry Standardization and Linked Open Data

- 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)

Introduction

POSTDATA
Poetry Standardization and Linked Open Data

- Dieresis

  en cánticos y nácares süaves

  en cán-ti-cos y ná-ca-res sü-a-ves

  – + – – – + – – – + – 11

  (Fray Jerónimo de San José)

Classification of Stanzas

POSTDATA
Poetry Standardization and Linked Open Data

2. Determining the rhyme type

Stanza	Consonant rhyme
Bravo león, mi coraz**ón**	-ón
Tiene apetitos, no raz**ón**	-ón
	Author: Alfonsina Storni

	Assonant rhyme
Ante una vidriera r**ota**	-ó-a
Coso mi lírica r**opa**	-ó-a
	Author: Federico García Lorca

Classification of Stanzas

POSTDATA
Poetry Standardization and Linked Open Data

3. Extraction of the rhyme pattern

Stanza	Rhyme pattern
Escribí en el aren**a**l	(a)
los tres nombres de la vida:	(–)
vida, muerte, am**o**r.	(b)
Una ráfaga de m**a**r,	(a)
tantas claras veces d**a**,	(a)
vino y nos borr**ó**.	(b)
	Author: Miguel Hernández

Corpus

POSTDATA
Poetry Standardization and Linked Open Data

- 1600 poems
- Early 15th century to contemporary poems
- 5005 stanzas extracted
- 45 stanza types (+ 1 misc.)

Corpus

POSTDATA
Poetry Standardization and Linked Open Data

- At least 10 stanzas per type, max. 30
- Manually reviewed by three experts
- Texts in modern Spanish
- No spelling or orthotypographic errors (Pérez Pozo et al., 2021)

Corpus

POSTDATA
Poetry Standardization and Linked Open Data

- 4,004 (80%) stanzas for training and evaluation
  - 3,204 training set
  - 800 evaluation set
- 1,001 (20%) test set

Methods: Baseline

POSTDATA
Poetry Standardization and Linked Open Data

- Expert system on top of Rantanplan (De la Rosa et al., 2020)

- Knowledge base based on the three stages

Sexta Rima stanza
Type of rhyme: Consonant rhyme
Rhyme pattern: “ababcc” or “aacbbc,”
Verse length pattern [11, 11, 11, 11, 11, 11]

Methods: Baseline

POSTDATA
Poetry Standardization and Linked Open Data

- Expert system on top of Rantanplan (De la Rosa et al., 2020)
- Knowledge base based on the three stages

```
(
    CONSONANT_RHYME,
    "sexta_rima",
    r"ababcc|aabccb|aabcbc",
    lambda ranges_list: has_fixed_length_verses("sexta_rima", ranges_list)
),
```

Methods: Baseline

POSTDATA
Poetry Standardization and Linked Open Data

```
STRUCTURES = (
    (
        CONSONANT_RHYME,
        "seguidilla",
        r"(-a-a)|(abab)",
        lambda ranges_list: has_fixed_length_verses("seguidilla", ranges_list,
                                                    fluctuation_size=1)
    ), (
        ASSONANT_RHYME,
        "seguidilla",
        r"(-a-a)|(abab)",
        lambda ranges_list: has_fixed_length_verses("seguidilla", ranges_list,
                                                    fluctuation_size=1)
    ), (
        CONSONANT_RHYME,
        "seguidilla_compuesta",
        r"((a-a-)|(-a-a)|(abab))((a-a)|(b-b)|(c-c))",
        lambda ranges_list: has_fixed_length_verses("seguidilla_compuesta",
                                                    ranges_list)
    ), (
        ASSONANT_RHYME,
        "seguidilla_compuesta",
        r"((a-a-)|(-a-a)|(abab))((a-a)|(b-b)|(c-c))",
        lambda ranges_list: has_fixed_length_verses("seguidilla_compuesta",
                                                    ranges_list)
    ), (
        ASSONANT_RHYME,
        "chamberga",
        r"(([^a]a[^a]a)|(abab)|[^a]a[^a]a)([^-]{2}){3}",
```

Methods: Baseline

POSTDATA
Poetry Standardization and Linked Open Data

- Expert system on top of Rantanplan (De la Rosa et al., 2020)
- Knowledge base based on the three stages
- Accuracy: **78.63%**

Methods: Tree-based

POSTDATA
Poetry Standardization and Linked Open Data

- Encode rules and their priorities
- Explainable and interpretable models
- Each independent rule in the knowledge base is added to a feature vector

Methods: Tree-based

POSTDATA
Poetry Standardization and Linked Open Data

La primavera ha venido.
Nadie sabe como ha sido.

Rantaplan

Check which structures trigger

{consonant, (8, 8), aa}

Knowledge base (structures)

Create feature vector

is_consonant	is_couplet	is_sexteto	...	lengths_are_8
1	1	0	...	1

Methods: Tree-based

POSTDATA
Poetry Standardization
and Linked Open Data

- We expected the method to infer the inner structure of rule firing

Methods: Tree-based

POSTDATA
Poetry Standardization and Linked Open Data

- We expected the method to infer the inner structure of rule firing

```
|--- is_cuaderna_vía <= 0.50
|   |--- is_soleá <= 0.50
|   |   |--- is_cuarteto_lira <= 0.50
|   |   |   |--- is_quinteto <= 0.50
|   |   |   |   |--- is_couplet <= 0.50
|   |   |   |   |   |--- class: unknown
|   |   |   |   |--- is_couplet >  0.50
|   |   |   |   |   |--- class: couplet
|   |   |   |--- is_quinteto >  0.50
|   |   |   |   |--- class: quinteto
|   |   |--- is_cuarteto_lira >  0.50
|   |   |   |--- is_estrofa_sáfica <= 0.50
|   |   |   |   |--- is_seguidilla <= 0.50
|   |   |   |   |   |--- class: cuarteto
|   |   |   |   |--- is_seguidilla >  0.50
|   |   |   |   |   |--- class: cantar
|   |   |   |--- is_estrofa_sáfica >  0.50
|   |   |   |   |--- class: cuarteta
|   |--- is_soleá >  0.50
|   |   |--- rhyme_type <= 0.50
|   |   |   |--- class: soleá
|   |   |--- rhyme_type >  0.50
|   |   |   |--- class: terceto
```

Methods: Tree-based

POSTDATA
Poetry Standardization and Linked Open Data

- We expected the method to infer the inner structure of rule firing

Method	Accuracy
Decision Tree	88.21%
Random Forest	88.51%

Methods: Tree-based

POSTDATA
Poetry Standardization and Linked Open Data

- We expected the method to infer the inner structure of rule firing

Method	Accuracy
Decision Tree	88.21%
Random Forest	88.51%

- An improvement of ~13% over baseline!

Methods: Neural Networks

POSTDATA
Poetry Standardization and Linked Open Data

- Neural networks capture patterns in datasets without having to specify the rules the govern them
- We expected neural networks to work without the knowledge base crafted by the experts
- Classifying stanzas is a multiclass single-label classification: stanzas as inputs, and one of the 46 possible stanza types as the output
- We used word embeddings and language models to extract feature vectors directly form plain text

Methods: Neural Networks

POSTDATA
Poetry Standardization and Linked Open Data

- Stacked BiLSTM layers, dropout, and classification head
- Grid search for dropout, number of BiLSTM layers, and epochs
- Non-contextual (GloVe) and contextual embeddings (BERT) (Pennington et al., 2014; Devlin et al., 2019)

Methods: Neural Networks

POSTDATA
Poetry Standardization
and Linked Open Data

- We expected the method to infer most of the rules crafted by experts

Methods: Neural Networks

POSTDATA
Poetry Standardization and Linked Open Data

- We expected the method to infer most of the rules crafted by experts

Method	Accuracy
GloVe	66.72%
BERT	42.12%

Methods: Neural Networks

POSTDATA
Poetry Standardization and Linked Open Data

- We expected the method to infer most of the rules crafted by experts

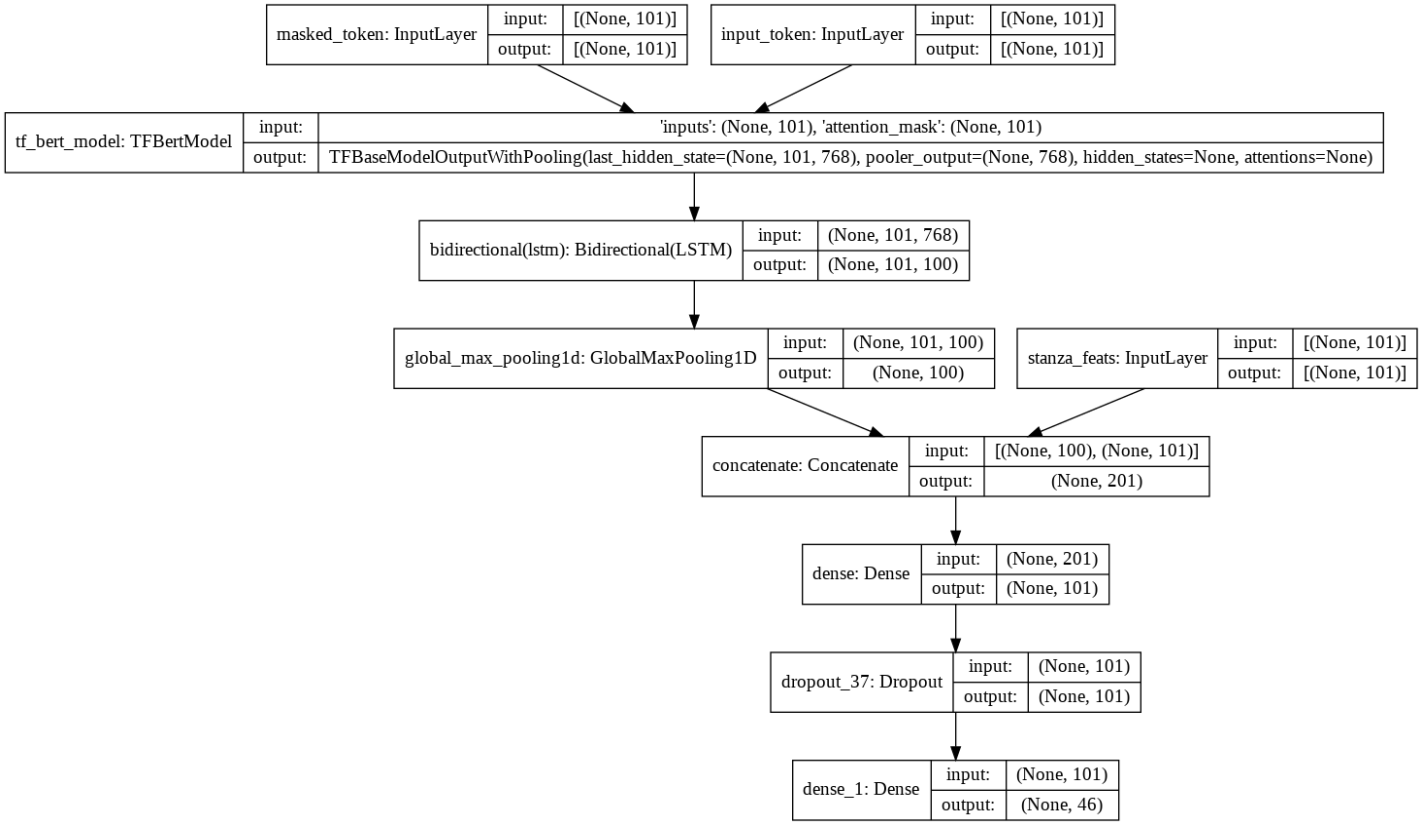
Method	Accuracy
GloVe	66.72%
BERT	42.12%

- A decrease from our baseline!

Methods: Hybrid

POSTDATA
Poetry Standardization and Linked Open Data

- Combining BiLSTMs + BERT + tree-based feature vectors (Rantanplan)



Methods: Hybrid

POSTDATA
Poetry Standardization and Linked Open Data

- Combining BiLSTMs + BERT + tree-based feature vectors (Rantanplan)

Method	Accuracy
BERT + features	91.91%

Methods: Hybrid

POSTDATA
Poetry Standardization and Linked Open Data

- Combining BiLSTMs + BERT + tree-based feature vectors (Rantanplan)

Method	Accuracy
BERT + features	91.91%

- An improvement of ~15% over baseline!

Methods: Summary

POSTDATA
Poetry Standardization and Linked Open Data

Method	Accuracy
Decision Tree	88.21%
Random Forest	88.51%
GloVe	66.72%
BERT	42.12%
BERT + features	**91.91%**

Discussion

POSTDATA
Poetry Standardization and Linked Open Data

- Limitations of baseline
  - Use of Old Spanish confused the scansion tool
  - The relaxation of some rules related to verse length allowing a small fluctuation in the fixed length of verses (Domínguez Caparrós, 2014; Jauralde, 2020; Quilis, 2000; Torre, 2000)
  - The presence of hemistiches, verses split in two halves with independent metrical lengths that affect that of the verse as a whole.

Discussion

POSTDATA
Poetry Standardization and Linked Open Data

- Tree-based solutions learned a better order of the knowledge base rules crafted by experts
- Neither contextual nor contextually-aware embeddings produced better results than baseline
- The combination of contextual embeddings with prior domain-specific knowledge performed remarkably well
- Embedding layers seem to carry insufficient structural information for this task, but it complements very well the feature set obtained from the 3-stage rules of each stanza type.

Conclusions

POSTDATA
Poetry Standardization and Linked Open Data

- We have contributed with a novel corpus, a knowledge base, and a baseline classifier
- Language models alone underperform simple methods such as decision trees (not enough structural information is encoded)
- Combining expert knowledge with contextual embeddings performs best (91%)
- Framing stanza identification as a classification task is challening. Could it be done in a multilingual setting?

Thanks

POSTDATA
Poetry Standardization
and Linked Open Data

Javier de la Rosa
versae@linhd.uned.es
@versae

POSTDATA

Poetry Standardization
and Linked Open Data