A White-Box Model for Detecting Author Nationality by Linguistic Differences in Spanish Novels

Albin Zehe, Daniel Schlör, Ulrike Henny-Krahmer, Martin Becker, and Andreas Hotho

University of Würzburg, 97074 Würzburg, Germany
{zehe,schloer,becker,hotho}@informatik.uni-wuerzburg.de
ulrike.henny@uni-wuerzburg.de

1 Introduction

Automatic nationality detection of authors writing in the same language (such as Spanish) can be used for many tasks, like author attribution, building large corpora to analyse nationality specific writing styles, or detecting outliers like exiled or bilingual authors. While machine learning provides many methods in this area, the corresponding results are usually not directly interpretable. However, in the Digital Humanities, explainable models are of special interest, as the analysis of selected features can help to confirm assumptions about differing writing styles among countries, or reveal novel insights into country-specific formulations. In this work, we aim to bridge this gap: Our assumption is that nationality or country of origin of an author is strongly connected to their writing style. Thus, we first present a machine learning approach to automatically classifying literary texts regarding their author's nationality. We then provide an analysis of the most relevant features for this classification and show that they are well interpretable from a literary and linguistic standpoint.

2 Related Work

The problem of detecting regional linguistic differences is at the core of Digital Humanities, as it touches research questions in both traditional linguistics and modern computer science. In Spanish philology and linguistics, the analysis of different regional varieties has a long tradition (see for example Alvar 1969, Eberenz 1995, Noll 2001). There are well-known differences between the Spanish spoken and written in Spain itself and the variations used in the former colonies, for example in forms of address ("vosotros/ustedes" vs. just "ustedes", voseo) and articles (le/la vs. lo).[1] More recently, these differences have been investigated with quantitative methods, for example by applying Zeta to find distinctive words for novels from Spain and from Latin America, respectively (Schöch et al. 2018).

[1] http://lema.rae.es/dpd/?key=voseo, http://lema.rae.es/dpd/?key=loísmo&lema=loísmo

3 Model

3.1 Baseline SVM-Model for classifying author nationality

We assume that writers from different countries are distinguishable by a) their vocabulary and b) phrases that are more or less popular in different regions (cf. Section "Related Work"). Thus, we choose to use an n-gram model to represent our corpus in a computer readable way: First, we determine all word n-grams of length 1 to 4 in the corpus. Then, we select the 1000 most frequent n-grams of each length.[2] We represent a piece of text as tf*idf vectors of these n-grams (see Manning 2008). We then train a linear SVM (see Steinwart 2008) to predict the nationality of an author given a piece of text. The linear SVM is known for good results in text classification (Joachims 1998) and - essential for interpretability - allows to inspect the importance of specific features.

3.2 Enhancing Feature Interpretability

When examining our classification model, we observed an over-representation of geographical entities (e.g., frequent locations like Buenos Aires) as well as names. To instead enforce linguistic properties, we replaced all uppercase tokens by distinct UNKNOWN-tokens (except at the beginning of a sentence). For example "¡Oh, María, María! ¡Cómo deseaba triunfar, conquistar Buenos Aires [...]", becomes "¡Oh, UNK_1, UNK_2! ¡Cómo deseaba triunfar, conquistar UNK_3 UNK_4 [...]". This ensures that n-grams with proper nouns will never be frequent enough to be used as a feature in our classification task.

3.3 Augmenting Training Examples

The success of machine learning algorithms depends largely on the amount of training data. Thus, to increase the number of training samples, we split each novel into multiple segments of equal length,[3] assigning each segment the same label as the entire novel. The classifier is then trained and evaluated on individual segments, resulting in a set of "votes" for the nationality of each novel in the test set. The nationality is then established by majority vote.

4 Corpus

We use a corpus composed of 100 novels from four Spanish-speaking countries, specifically Spain, Argentina, Cuba and Mexico, written in the 19th and early 20th century (Calvo Tello 2017, Henny-Krahmer 2017). Fig. 1 shows the distribution over countries and the distributions over subgenres in the countries. All countries

[2] We also tried selecting the 100 or 10000 most frequent n-grams, which led to slightly worse results.

[3] The cross validation split was performed before segmentation, ensuring that no novel was present in both training and test set.

are represented by a roughly equal number of texts. We note that our corpus may have a bias towards a specific subgenre in some countries, which will later be addressed in the analysis of the features.

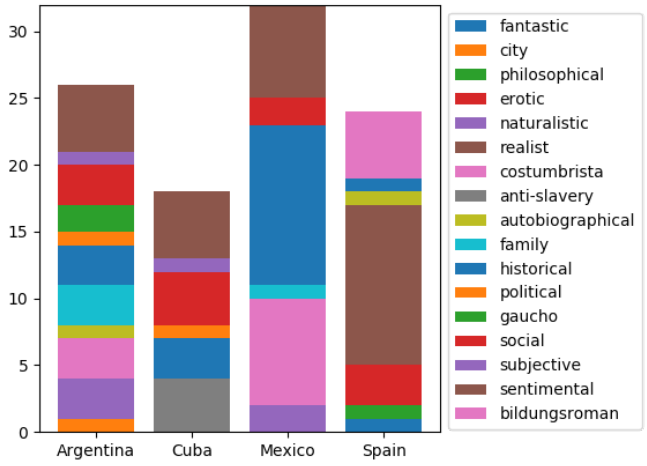

Figure 1: Distribution of countries and subgenres in our corpus

5 Experiments

We performed extensive experiments on the dataset to determine the accuracy of our approach. The main hyper-parameters of our model are the segment size s, determining how many words a segment contains, and the parameter C of the SVM. We performed parameter optimisation by grid search, choosing from $s \in \{100, 200, 500, 1000, 5000, 10000, 100000, \infty\}$ and $C \in \{10^{-5}, 10^{-4}, ...10^{5}\}$. $s = \infty$ does not perform segmentation. We also varied the maximum length of n-grams: unigrams ($n = 1$) vs. n-grams of length 1 to 4. All scores reported below are weighted average F1-scores[4] over 10-fold cross validation.

Generally, our model performed best when using only unigrams, removing uppercase tokens and splitting the novels into segments of length $s = 1000$ (see Table 1 for details). This can be explained by the small dataset: Unigrams are likely to occur in multiple samples even in a small corpus, while higher-order n-grams possibly only occur once and can therefore not be used for classification. Fig. 2 shows the results for varying s and C. Segments of a length around 1000

[4] see `http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html`

perform best, yielding F1-scores of up to 86.8%. Very small segments fail to deliver satisfying results, while larger segments still provide reasonable classification accuracy. The value for C must be set high enough, but the specific value does not matter for $C > 10$.

Using all n-grams of length 1 to 4 also delivered good accuracy (highest F1-score of 80.4% for $C = 10000$, $s = 1000$). Removing uppercase tokens had a positive effect when using unigrams, while it hardly influenced the accuracy using all n-grams.

A detailed view of all results can be found on GitHub.[5]

Table 1: Classification report for the best configuration, using only unigrams, segments of length $s = 1000$ and $C = 10000$

	precision	recall	f1-score	support
0	0.800	1.000	0.889	24
1	0.923	0.667	0.774	18
2	0.824	0.875	0.848	32
3	1.000	0.885	0.939	26
avg / total	0.882	0.870	0.868	100

6 Feature Analysis

Using a linear SVM enables us to analyse the 10 n-grams that provide the strongest evidence for and against a country (according to internal weights). In the following, we focus on features that are weighted strongly in all or at least multiple folds of the cross validation.

Generally, we identify three feature groups: topical features, features related to the geographical setting and linguistic features. The presence of topical features can be explained by the bias in subgenres that is present in our corpus and is not necessarily representative. The geographical features seem to point to a tendency of the authors to base their stories in their respective home countries rather than other countries.

With regard to the different model variants, the model based on *unigrams without removing uppercase tokens* tends to select names as its top-features such as the country itself or characteristic cities, for example "Madrid" for Spain. While these features are surely helpful for classification (yielding an F1-score of 81.7%), they are not particularly interesting for linguistic analysis. The features selected after *removing uppercase tokens*, on the other hand, seem more relevant from a linguistic viewpoint, while at the same time providing the best accuracy. Table 2 shows features that are among the highest weighted for more than 5 folds for each country in this setting.

[5] `https://github.com/cligs/projects2018/tree/master/country-dh`

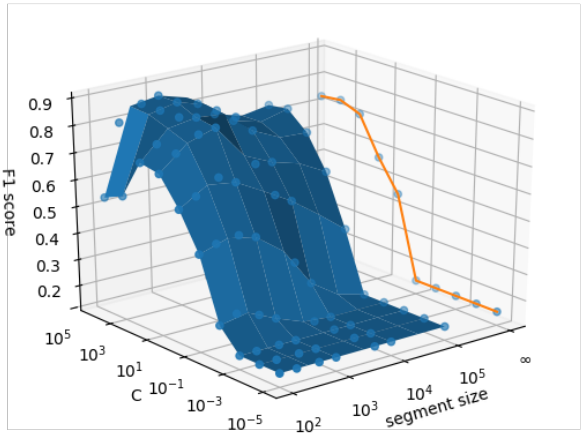

Figure 2: Weighted average F1-score depending on the segment size s and the cost parameter C of the SVM. The separated line denotes no segmentation ($s = \infty$). Only unigrams were used as features.

Table 2: Unigrams with large weights assigned by the SVM. Features marked with + and - are signals for and against a country, respectively.

Country	Unigrams	Comments
Spain	+ ello	linguistic (personal pronoun)
	+ señorito	linguistic (diminutive)
	+ duros	currency
	+ señores	linguistic/topical (noun)
	- pesos	currency
Cuba	+ esclavo/esclava	topical
	+ mulato	topical (ethnic group in Cuba)
	+ añadió	linguistic/narrative (verb, probably used to mark direct speech)
	- quizá	linguistic (adverb)
	- huerta	topical/linguistic (noun)
Mexico	+ hacienda	topical (haciendas are typical of Spanish colonies
	+ mexicano	
Argentina	+ entretanto	linguistic (temporal adverb)
	+ gaucho	topical
	+ misia	linguistic (form of address typical in South America)
	+ mate	topical (drink typical to Argentina)

Using all n-grams without removing uppercase tokens, we again find a preference for geographical phrases like "de la Habana". As with unigrams, linguistic features become more important than topical features when *uppercase tokens are removed*. Table 3 shows some particularly interesting n-grams with high weights.

Table 3: N-grams with large weights assigned by the SVM. Features marked with + and - are signals for and against a country, respectively.

Country	n-grams	Comments
Spain	+ se me figura que	linguistic (locution)
	+ de la huerta	topical ("huerta" is common in Spain)
Cuba	- de cuando en cuando	linguistic (temporal phrase)
Mexico	+ de/en la casa de	topical (probably due to a subgenre bias)
	+ al cabo de	linguistic (temporal phrase)
	+ de la hacienda	topical (typical of Spanish colonies)
	+ así es que	linguistic (locution)
	- al mismo tiempo	linguistic (temporal phrase)
	- la	linguistic (leísmo)
Argentina	+ en ese momento	linguistic (temporal phrase)
	+ se puso de pie	linguistic (verb)
	+ de vez en cuando	linguistic (temporal phrase)
	+ el hecho es que	linguistic (locution)
	+ al fin al cabo	linguistic (temporal phrase)
	- al cabo de un	linguistic (temporal phrase)

7 Discussion

7.1 Technical Aspects

We found that segmenting novels to augment the training data does improve results, but only if the segments are not too short and thus do not contain enough information to detect the author's nationality. Removing uppercase tokens improves the classification accuracy and makes the selected features more interesting from a linguistic standpoint. We assume that otherwise proper nouns are picked up by the classifier as important clues on the training set, which fail to generalise to the test set.

7.2 Feature Interpretation

The words and phrases that our algorithms selects for differentiating between nationalities strongly resemble features that humans would consider given the same task. These include well-known linguistic differences (leísmo) as well as country-specific words (hacienda/huerta). However, it also finds phrases, such as temporal expressions, that are not very well known to be specific for some

countries, but should be further investigated in future work. We also observe that authors in our corpus appear to have a strong tendency towards writing about their respective home countries, as evidenced by the selection of city or country names.

8 Conclusion and Future Work

We have presented a classifier that is able to distinguish between novels from different countries based on word n-grams. Our experiments show that this classifier is able to select features that are interpretable and reveal interesting insights into the language used in novels from different Spanish-speaking countries. We note that our findings are only based on a limited dataset. However, the tools we have built enable us to replicate the experiments and confirm our findings as soon as larger collections of text become available. Thus, our work is an important step towards combining machine learning with in-depth analysis and discovery of novel concepts in corpus-based linguistic studies through interpretable models. In future work, we believe that replacing the majority vote over segments by more sophisticated methods can further improve our results. We also believe that incorporating linguistic information like parse-trees into our features can help to reveal more interesting insights into subtle linguistic differences between countries.

9 Bibliography

Alvar, Manuel (1969). Variedad y unidad del español: estudios lingüísticos desde la historia. Editorial Prensa Española.

Calvo Tello, José (ed.) (2017). Corpus of Spanish Novel from 1880-1940. Würzburg: CLiGS. `https://github.com/cligs/textbox/blob/master/es/novela-espanola`.

Eberenz, Rolf (1995): "Norm und regionale Standards des Spanischen in Europa und Amerika". In: Oskar Müller, Dieter Nerius, Jürgen Schmidt-Radefeldt (eds.). Sprachnormen und Sprachnormenwandel in gegenwärtigen europäischen Sprachen. Rostock: Universität Rostock, 47-58.

Henny-Krahmer, Ulrike (ed.) (2017). Collection of 19th Century Spanish-American Novels (1880-1916). Würzburg: CLiGS, 2017. `https://github.com/cligs/textbox/master/spanish/novela-hispanoamericana/`.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98, 137–142.

Manning, C. D., Raghavan, P., Schütze, H. (2008). Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press. ISBN: 0521865719, 9780521865715

Noll, Volker (2001). Das amerikanische Spanisch: ein regionaler und historischer Überblick. Tübingen: Niemeyer.

Schöch, C., Calvo, J., Zehe, A., Hotho, A. (2018). Burrows Zeta: Varianten und Evaluation. DHd 2018

Siskind, Mariano (2010): "The Globalization of the Novel and the Novelization of the Global. A Critique of World Literature." In: Comparative Literature 62 (4), 336-360. https://doi.org/10.1215/00104124-2010-021

Steinwart, I., Christmann, A. (2008). Support Vector Machines. Springer Publishing Company, Incorporated. ISBN: 0387772413