Comparison of Character n-grams and Lexical Features on Author, Gender, and Language Variety Identification on the Same Spanish News Corpus (preprint version)

Conference Paper · September 2017

DOI: 10.1007/978-3-319-65813-1_15

CITATIONS
8

READS
812

4 authors:

Miguel A. Sanchez-Perez
Instituto Politécnico Nacional
10 PUBLICATIONS **161** CITATIONS
SEE PROFILE

Ilia Markov
University of Antwerp
42 PUBLICATIONS **297** CITATIONS
SEE PROFILE

Helena Gomez Adorno
Universidad Nacional Autónoma de México
47 PUBLICATIONS **531** CITATIONS
SEE PROFILE

Grigori Sidorov
Instituto Politécnico Nacional
226 PUBLICATIONS **1,745** CITATIONS
SEE PROFILE

Comparison of Character N-grams and Lexical Features on Author, Gender, and Language Variety Identification on the Same Spanish News Corpus

Miguel A. Sanchez-Perez, Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov

Instituto Politécnico Nacional, Center for Computing Research, Mexico City, Mexico
miguel.sanchez.nan@gmail.com, imarkov@nlp.cic.ipn.mx, helena.adorno@gmail.com, sidorov@cic.ipn.mx

Abstract. We compare the performance of character n-gram features (n = 3–8) and lexical features (unigrams and bigrams of words), as well as their combinations, on the tasks of authorship attribution, author profiling, and discriminating between similar languages. We developed a single multi-labeled corpus for the three aforementioned tasks, composed of news articles in different varieties of Spanish. We used the same machine-learning algorithm, Liblinear SVM, in order to find out which features are more predictive and for which task. Our experiments show that higher-order character n-grams (n = 5–8) outperform lower-order character n-grams, and the combination of all word and character n-grams of different orders (n = 1–2 for words and n = 3–8 for characters) usually outperforms smaller subsets of such features. We also evaluate the performance of character n-grams, lexical features, and their combinations when reducing all named entities to a single symbol "NE" to avoid topic-dependent features.

Keywords: feature selection; authorship attribution; author profiling; discriminating between similar languages; lexical features; character n-grams

1 Introduction

This paper focuses on three natural language processing (NLP) tasks that have experienced an increase in interest in recent years: authorship attribution (AA), author profiling (AP), and discriminating between similar languages (DSL). Authorship attribution (AA) is the task that aims at automatically identifying the author of a text [1], when author profiling (AP) aims at identifying profiling aspects of an author, such as age, gender, or native language based solely on a sample of his or her writing.[1] Discriminating between similar languages (DSL) is the task of predicting the language variety in which a given text was written.

From the machine-learning perspective, all the three tasks can be viewed as a multi-class, single-label classification problem, where automatic methods have to assign class labels (e.g., author's name (AA); author's gender (AP); language variety (DSL)) to objects (text samples). Practical applications of these tasks vary from electronic commerce and forensics to machine translation and information retrieval systems.

[1] In this paper, we only address gender identification.

Character n-grams and lexical features (unigrams and bigrams of words), as well as their combinations, have proved to be predictive for these tasks, including when the Spanish language or its varieties are concerned [2, 3]. Thus the research question addressed in this work is to examine which features and feature combinations are the best predictive for author, gender, and language variety identification when evaluated on the same corpus in Spanish. Moreover, we evaluate the impact of NEs on these tasks.

2 Related Work

In this section, we will focus on the best approaches for the Spanish language published in the most recent editions of two widely known workshops: PAN[2] and VarDial[3]. These workshops provide a common platform for researchers interested in evaluating and comparing their systems' performance on the authorship identification-related and discriminating between similar languages tasks, respectively.

In the 2014 edition of the PAN Authorship Attribution (AA) competition [4], the task consisted in identifying the author of a text on a corpus composed of newspaper opinion articles. The winner approach for Spanish [5] used a modification of the *Impostors* method [6]. The author identification (author verification) task in PAN 2015 [7] focused on a cross-genre scenario, that is, when training and test sets are on different genre (e.g., tweets vs. news articles). The best approach for Spanish [8] relied on a variety of features, including character n-grams, words, POS tags, and sentence length.

In the 2015 edition of the PAN Author Profiling (AP) task [9], the winning approach [10] for gender identification on the Spanish tweets corpus was based on second order attributes technique. In 2016 [2], the shared task focused on cross-gender AP conditions. The best approach [11] in identifying the gender on the Spanish dataset relied on words, sentiment and topic derivation, and stylistic features.

The 2016 edition of the VarDial workshop for discriminating between similar languages (DSL) [12] used a corpus of short excerpts of news texts, covering Argentine, Castilian, and Mexican Spanish. The overall winner [13] employed character n-gram features (n = 1–7). This year edition [3] included Argentinian, Peruvian, and Peninsular Spanish. The winner [14] used character n-grams (n = 1–4) for predicting the language group and character n-grams of different order, POS n-grams, and proportions of capitalized letters, punctuation marks, and spaces for identifying the language varieties within the group.

The results for the DSL task are usually higher than those for AA or AP. For instance, the best performing system [14] in the VarDial 2017 workshop [3] achieved 92.74% of accuracy, while the results for AA and AP under single-genre conditions are usually around 80% [2,7]. As can be seen, the state-of-the-art approaches in both shared tasks employed character n-gram and lexical features. Therefore, it makes it important to evaluate the performance of these features on the same corpus for the three tasks.

[2] http://pan.webis.de

[3] http://ttg.uni-saarland.de/vardial2017/sharedtask2017.html

3 Corpus

There are numerous works that tackle the evaluation of character n-gram and lexical features' performance for the English language, since for English there is a large number of corpora and lexical resources. However, for Spanish, the availability of corpora is scarce, which limits the amount of research done for this language. For the evaluation of character n-gram and lexical features, we built a corpus composed of news articles in eight varieties of the Spanish language: Argentinian, Mexican, Colombian, Chilean, Venezuelan, Panamanian, Guatemalan, and Peninsular Spanish.

The corpus includes only the news with a minimum size of 750 characters. We removed all the news with distributed authorship, e.g., *AP*, *La prensa*, *Editorial*, etc. Overall, between 10 and 40 texts (news articles) were selected for each author in the corpus; these ranges were set so that the corpus is not highly unbalanced with respect to the number of documents per author. Additionally, we manually checked each news content and deleted names of authors, places, emails, and any other information that may help to reveal the authorship of a text. Finally, during the manual inspection of the corpus, we labeled each text with author's gender (male or female).

The corpus is composed of 5,187 news articles written by 232 different authors (2,968 articles written by male authors and 2,219 by female authors) and includes eight varieties of Spanish distributed as follows: Argentina: 449, Venezuela: 828, Colombia: 929, Guatemala: 598, Spain: 908, Mexico: 682, Panama: 418 and Chile: 375. The Spanish News Corpus is freely available on our website[4], where you will find more information about the corpus statistics.

4 Experimental Settings and Results

We evaluated the performance of character n-grams and lexical features, as well as some of their combinations. Character n-grams vary in order from 3 to 8, while lexical features include unigrams and bigrams of words. Each model was evaluated by measuring classification accuracy on the entire corpus under stratified 10-fold cross-validation. Following previous research [15], we removed features with a frequency less than 5 in the entire corpus, which significantly reduces the size of the feature set (on average by approximately 80%). As machine-learning algorithm, we selected Support Vector Machines (SVM); it was the classifier of choice of the majority of the teams in the previous editions of the PAN and VarDial competitions [2, 12]. Given that the number of features is much larger than the number of instances, we used Crammer and Singer's linear kernel algorithm with default parameters implemented in the WEKA's [16] Liblinear [17] package. Following the practice of the VarDial workshop [18], we conducted additional experiments reducing all named entities (NEs) to a single symbol (*#NE#*) in order to evaluate their impact on these tasks.

Table 1 shows the obtained results for the AA, AP (gender identification), and DSL tasks in terms of accuracy (%) before and after replacing the NEs with a symbol. For each experiment, the number of features (N) is provided. The top accuracy values for

[4] http://www.cic.ipn.mx/~sidorov/SpanishNewsCorpus.zip

Table 1. Accuracy results (%) for lexical features, character (char.) n-grams, and their combinations in the AA, AP, and DSL tasks, before and after reducing all NEs to a single symbol.

Word unigrams	Word bigrams	Char. 3-grams	Char. 4-grams	Char. 5-grams	Char. 6-grams	Char. 7-grams	Char. 8-grams	Accuracy with NEs (%)				Accuracy without NEs (%)			
Features								**AA**	**AP**	**DSL**	**N**	**AA**	**AP**	**DSL**	**N**
✓								73.74	73.99	92.92	38,360	68.77	70.79	86.70	31,884
	✓							70.04	73.13	91.05	94,501	63.97	70.43	84.36	89,448
		✓						74.01	69.87	91.50	25,631	70.43	66.96	86.97	19,209
			✓					76.60	72.80	93.75	83,917	74.15	70.10	90.57	62,514
				✓				76.71	73.92	93.75	189,240	74.55	71.43	91.34	148,026
					✓			76.13	74.94	94.04	336,422	73.85	72.45	91.65	285,365
						✓		75.30	75.11	94.04	498,014	73.16	73.55	91.77	445,546
							✓	74.17	**75.61**	93.64	628,180	72.12	**73.90**	91.61	579,349
Combinations								**AA**	**AP**	**DSL**	**N**	**AA**	**AP**	**DSL**	**N**
✓	✓							74.82	74.78	92.94	132,861	70.08	72.43	87.97	121,332
✓	✓	✓						75.32	71.68	92.60	158,492	72.53	70.50	89.47	140,541
✓	✓	✓	✓					76.46	72.97	93.14	242,409	73.97	71.16	90.17	203,055
✓	✓	✓	✓	✓				77.31	73.51	93.45	431,649	74.63	71.74	90.73	351,081
✓	✓	✓	✓	✓	✓			77.91	74.15	93.70	768,071	**75.13**	72.14	91.19	636,446
✓	✓	✓	✓	✓	✓	✓		**77.96**	74.57	93.99	1,266,085	*	72.86	92.30	1,081,992
✓	✓	✓	✓	✓	✓	✓	✓	77.93	75.28	**94.16**	1,894,265	*	73.05	**92.52**	1,661,341

each task are shown in bold typeface. The asterisks correspond to experiments that did not finish on time. We believe that the number of classes (238) for AA leads to a high computational cost for this SVM kernel. It is worth mentioning that when reducing the NEs this algorithm takes much more time to converge (about 6 times more).

As one can see from Table 1, higher-order character n-grams (n = 5–8) outperform both lower-order character n-grams and n-grams of words for all the three tasks when evaluated in isolation. The combination of all word and character n-grams provides the best results for two out of three considered tasks, AA and DSL, which is in line with the previous research [19]. These results are consistent with and without replacing NEs.

Moreover, it can be seen that the results continue to improve when adding higher-order character n-grams to the combination of features. However, higher-order character n-gram features significantly increase the size of the feature set, especially when used in combinations with each other, and consequently, the computational cost of the training process, while the accuracy improvement is only marginal. Therefore, we limited our experiments with the maximum order of 8 for character n-grams.

The best model for the AP and DSL tasks slightly outperforms the BOW approach when NE's are present (1.62% and 1.24%, respectively). However, when NEs are reduced the difference becomes higher (3.11% and 5.82%, respectively). The average drop in accuracy after reducing NEs is 2.52% for character n-grams and approximately

5% for lexical features. This confirms that lexical features are more topic-specific, which sometimes leads to unintended extraction of topic or domain information [20].

The average accuracy drop after reducing NEs is 3.52% for AA, 2.29% for AP, and 3.47% for DSL. For the AA task, the accuracy drop of 3.52% on our corpus is lower than the one of 5%–20% reported in [21], when for DSL the drop of 3.47% is higher than the one of around 2% reported in the VarDial workshop proceedings [18]. One of the possible explanations is the nature of our corpus, which contains shorter texts than the fiction novels corpus used for AA in [21], but much longer texts than the VarDail corpus of excerpts of journalistic texts.

5 Conclusions

In this paper, we examined the performance of character n-grams (n = 3–8), lexical features (unigrams and bigrams of words), and their combinations on the tasks of authorship attribution (AA), author profiling (AP) (only gender identification), and discriminating between similar languages (DSL) on a developed multi-labeled corpus of news articles in different varieties of Spanish.

The obtained results indicate that higher-order character n-grams outperform lower-order character n-grams for all the three tasks and provide the best results for gender identification when used in isolation (75.61% of accuracy). The combination of all word and character n-grams of different orders (n = 1–2 for words and n = 3–8 for characters) outperforms other combinations of such features and provides the best results for author and language variety identification (77.96% and 94.16%, respectively). We also evaluated the impact of named entities on these tasks. Our experiments showed that reducing them all to a single symbol "NE" to avoid topic-dependent features decreases accuracy by around 2.5%–3.4%, depending on the task. This work serves as a baseline for more complex methods based on dimensionality reduction or deep learning.

Acknowledgments

This work was partially supported by the Mexican Government (CONACYT projects 240844, SNI, COFAA-IPN, SIP-IPN 20171813, 20171344, 20172008 and CONACYT under the Thematic Networks program (Language Technologies Thematic Network Projects 260178 and 271622).

References

1. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society For Information Science and Technology **60** (2009) 538–556
2. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4^{th} author profiling task at PAN 2016: Cross-genre evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs, CLEF and CEUR-WS.org (2016)
3. Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., Aepli, N.: Findings of the vardial evaluation campaign 2017. In: Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017). (2017)

4. Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., Sánchez-Pérez, M.A., Barrón-Cedeño, A.: Overview of the author identification task at PAN 2014. In: Working Notes of CLEF 2014. (2014) 877–897
5. Khonji, M., Iraqi, Y.: A slightly-modified GI-based author-verifier with lots of features. In: Working Notes of CLEF 2014. (2014)
6. Koppel, M., Winter, Y.: Determining if two documents are by the same author. Journal of the American Society for Information Science and Technology **65** (2014) 178–187
7. Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., Stein, B.: Overview of the author identification task at PAN 2015. In: Working Notes of CLEF 2015. (2015)
8. Bartoli, A., Dagri, A., Lorenzo, A.D., Medvet, E., Tarlao, F.: An author verification approach based on differential features. In: Working Notes of CLEF 2015. (2015)
9. Rangel, F., Celli, F., Rosso, P., Pottast, M., Stein, B., Daelemans, W.: Overview of the 3^{rd} author profiling task at PAN 2015. In: CLEF 2015 Labs and Workshops, Notebook Papers. Volume 1391., CEUR (2015)
10. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y-Gómez, M., Villaseor-Pineda, L., Jair-Escalante, H.: INAOE's participation at PAN'15: Author profiling task. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. Volume 1391., CEUR (2015)
11. Gencheva, P., Boyanov, M., Deneva, E., Nakov, P., Georgiev, G., Kiprov, Y., Koychev, I.: PANcakes team: A composite system of genre-agnostic features for author profiling. In: Working Notes Papers of the CLEF 2016 Evaluation Labs, CLEF and CEUR-WS.org (2016)
12. Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J.: Discriminating between similar languages and arabic dialect identification: A report on the third DSL shared task. In: Proceedings of the 3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (VarDial 2016). (2016) 1–14
13. Çöltekin, C., Rama, T.: Discriminating similar languages: experiments with linear SVMs and neural networks. In: Proceedings of the 3rd Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2016). (2016) 15–24
14. Bestgen, Y.: Improving the character n-gram model for the DSL task with BM25 weighting and less frequently used feature sets. In: Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017). (2017)
15. Markov, I., Stamatatos, E., Sidorov, G.: Improving cross-topic authorship attribution: The role of pre-processing. In: Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017). (2017)
16. Witten, I., Frank, E., Hall, M., Pal, C.: Data Mining: Practical machine learning tools and techniques. 4th edn. Morgan Kaufmann (2016)
17. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. **9** (2008) 1871–1874
18. Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., Nakov, P.: Overview of the DSL shared task 2015. In: Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial 2015). (2015) 1–9
19. Gómez-Adorno, H., Markov, I., Baptista, J., Sidorov, G., Pinto, D.: Discriminating between similar languages using a combination of typed and untyped character n-grams and words. In: Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017). (2017)
20. Tsur, O., Rappoport, A.: Using classifier features for studying the effect of native language on the choice of written second language words. In: Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition (CACLA 2007), ACL (2007) 9–16
21. Ríos, G., Sidorov, G., Castro, N., Nava, A., Chanona-Hernández, L.: Relevance of named entities in authorship attribution. In: Proceedings of the 15th Mexican International Conference on Artificial Intelligence (MICAI 2016), LNAI, Springer (2017)