Complexity Measures and POS N-grams for Author Identification in Several Languages

SINAI at PAN@CLEF 2018

Rocío López-Anguita, Arturo Montejo-Ráez, and Manuel C. Díaz-Galiano

Centro de Estudios Avanzados en TIC
Universidad de Jaén
{rlanguit, amontejo, mcdiaz}@ujaen.es

Abstract This paper presents our approach and results for the 2018 PAN Author Identification Task. In this task, we are given a set of documents (known *fanfics*) by a small number (up to 20) of candidate authors. All documents are in the same language that may be English, French, Italian, Polish, or Spanish. The task consists in developing a system to identify the authors of another set of documents (unknown *fanfics*). We have used two strategies to solve this task. The first strategy has consisted in using several measures of the complexity of the fanfics texts for each candidate. In the second strategy, we analyzed the fanfics of each candidate by applying a Part-Of-Speech Tagger and a n-gram based vector space model.

1 Introduction

This year's Author Identification task [13] in PAN@CLEF [27] is divided into cross-domain authorship attribution and style change detection. In our work, we have focused on the first case. The goal of this task is to find which are the best approaches to model an author style of writing so new texts can be attributed. This year the task has moved to the classification of what is called *fanfics* which are narrative texts on topics and subjects originally created by other authors but written by fan readers of these referred authors. For example, we can have small stories on Harry Potter written by people different from J.K Rowling. This new challenges emphasizes the relevance of the writing style rather than the topic of the content. Nevertheless, content could be also important, as the richness of vocabulary or the use of certain words and expressions could be inherent characteristics of a certain author.

We wanted to study the effectiveness of well-known text complexity measures as features for this text classification task. Also, inspired by other features not derived from content meaning or related to content topics, we have applied a POS tagger to use n-grams of POS tag sequences as features in a vector space model. Our results show that complexity measures are not a good source of information for this task, whereas POS sequences behave reasonably well in certain languages.

The article is organized as follows: Section 2 introduces the complexity measures used for modeling texts; Section 3 describes in detail the approaches followed in this task; Section 4 describes the experiments carried out and shows the results obtained. Section 5 closes the article with some reflections.

2 Complexity measures

In this subsection, we will take a look at the different metrics of complexity that have been proposed by various authors. Some of these measures directly provide the recommended age for a reader, such as the *García López* [10] measure, others offer more difficult to interpret indexes, such as lexical complexity of *Anula* [1], the sentence complexity index or the depth dependency tree of *Saggion* [24], among others.

Punctuation Marks This measure was proposed by *Saggion* [24]. The average number of punctuation marks is used as one of the complexity indicators of the text.

Sentence Complexity The sentence complexity index was proposed by *Anula* [1]. It measures the number of words per sentence, thus obtaining an index on sentence length, and also measures the number of complex sentences per sentence, from an index of complex sentences. Among complex sentences are those with composed verbs, for instance.

Automated Readability Index *Senter and Smith* [25] proposed one of the most widely used indexes due to its ease of calculation. Measures the difficulty of a text from the average number of characters (letters and numbers) per word and the average number of words per sentence.

μ Readability The μ Readability is a formula for calculating the readability of a text. It provides an index between 0 and 100 and was developed by *Muñoz* [17]. This measure focuses on measuring the number of words, the average number of letters per word and their variance.

Dependency Tree Height This measure was also proposed by *Saggion* [24]. It is a very useful metric for capturing syntactic complexity: long sentences can be syntactically complex or contain a large number of modifiers (adjectives, adverbs or adverbial phrases). The latter do not increase syntactic complexity and do not lead to very deep trees, while the former have a strong tendency to produce deep trees.

Gunning Fog Score (FOG) This measure was developed in 1952 by *Robert Gunning* [4]. This index is a readability test for English and Polish writing. The index estimates the years of formal education a person needs to understand the text on the first reading.

Flesch The most popular English formula for calculating readability was proposed by *Rudolf Flesch* [8]. Measures the difficulty of a text from the average number of syllables per word and the average number of words per sentence.

Flesch-Kincaid *Rudolph Flesch*, is the co-author of this formula along with *John P. Kincaid* [4]. This Index improves upon the Flesch Reading Ease Readability formula and estimates the number of years of education in the American school system necessary for comprehending a given text.

SMOG *G. Harry McLaughlin* created the SMOG Readability Formula in 1969 in an article [15]. This formula estimates the years of education a person needs to understand a text through the number of sentences and words with three or more syllables.

Lexical Complexity This measure of complexity was proposed by *Anula* [1] to measure the lexical complexity of a text, determined in the basis of its frequency of use and lexical density. It is considered that the greater the lexical density (the greater the number of different words per text), the greater the difficulty of comprehension becomes.

Spaulding Readability The readability of Spaulding, commonly known as the SSR index, was proposed by *Spaulding* [26]. It focuses on measuring vocabulary and sentence structure to predict the relative readability of a text.

Fernández-Huerta Readability *Blanco Pérez* [3] and *Gutiérrez Couto* [22] propose this measure of complexity as an adaptation to Spanish of the Flesch readability test ([8]). It is based on the fact that in Spanish the words have more syllables on average and the sentences are also longer. Measures the average number of syllables per word and the average number of words per sentence in the text.

Flesch-Szigrist Readability (IFSZ) The works of *Granada Barrio-Cantalejo* [11] and *Ramírez-Puerta* [22] proposes the Flesch-Szigristzt readability index as a modification of the Flesch [8] formula adapted to Spanish. The IFSZ readability index is considered a reference for the Spanish language. Measures the number of syllables per word and the number of words per sentence in the text.

Gutierrez Readability It was created for Spanish by *Rodríguez* [23] and consists of a mathematical formula, generated by multiple regression methods, which includes certain linguistic characteristics of the material whose difficulty is to be evaluated. It focuses on measuring the average number of letters per word and the average number of words per sentence.

Minimum Age of Readability In the work of *García López* [10] we can find another formula to estimate the required age of a reader needed to understand a text. It is, again, an adaptation to Spanish of Flesch's original formula ([8]) for English. Measures the average number of syllables per word and the average number of words per sentence to obtain the minimum age necessary to understand a text.

SOL *Contreras* [5] proposes the SOL metric as an adaptation to Spanish and French of the SMOG formula proposed by *Mc Lauglin* [15]. It measures the readability of a text by the grade level, which is the number of years of school required to understand the text.

Crawford [6] This measure was proposed by *Alan N. Crawford* in 1989 [6]. It is used to calculate the years of school required to understand a text. Measures the number of sentences per hundred words and the number of syllables per hundred words.

Kandel-Models *Kandel and Models* [12] propose this measure of complexity as an adaptation to French of the Flesch readability test ([8]). It is based on the fact that in French the words have more syllables on average and the sentences are also longer. Measures the average number of syllables per word and the average number of words per sentence in the text.

Dale Chall This formula was inspired by Rudolf Flesch's Flesch-Kincaid readability and was created by *Edgar Dale and Jeanne Chall* [7]. Measures the complexity of a text, determined by the difficulty of the words and the average number of words per sentence.

Flesch-Vaca In 1972, *Roberto Vacca* and *Valerio Franchina* [9] propose this measure of complexity as an adaptation to Italian of the Flesch readability test ([8]). Measures the average number of syllables per word and the average number of words per sentence in the text.

Gulpease The Gulpease index is a readability index for Italian text. It was defined as part of the research of the GULP (Groupe Pedagogique Linguistique Universitaire) at the Seminar of Educational Sciences of the University of Rome "La Sapienza". [14] Compared to other indexes, it has the advantage of using the length of words in letters instead of in syllables, which simplifies the automatic calculation. It provides an index between 0 and 100, with "100" indicating the highest readability and "0" indicating the lowest readability.

Pisarek The most popular Polish formula for calculating readability was proposed in the 1960s by *Walery Pisarek* [4], based on research by *Rudolf Flesch* [8] and *Josef Mistrik* [16]. It takes into account only two characteristics of the text: the average length of a sentence and the percentage of "potentially difficult" words (longer than three syllables).

3 Method

In this section, we present the different approaches that we have applied in our participation in CLEF PAN 2018 Author Identification Task.

Our hypothesis is that, since complexity metrics capture different aspects of text complexity, they should be valid as characteristics in a model that represents each document in this author identification process. Additionally, we will compare these results with those obtained with other models of the text but not associated with complexity, such as the Part-Of-Speech Tagger vectors with TF and TF.IDF representations.

In short, we have tried two main approaches, both using supervised learning algorithms. The two approaches are as follows:

- Vectors of complexity measures of the fanfics.
- Vectors of n-grams of Part-Of-Speech tags of the fanfics.

For both cases, texts have been processed using the Freeling[1] toolkit [18] for tokenize words and punctuation, along with sentence splitting. Only in the case of Polish and Italian we have implemented our own tokenizer, using regular expressions and the NLTK[2] library [2]. All our scripts have been coded in Python. No special treatment on speech expressions have been performed.

The classification was performed by applying the Support Vector Classification provided by the SciKit-Learn[3] library [20]. The multiclass support is handled according to a one-vs-one scheme and the cost value was fixed to 1 to avoid discrimination among classes. The kernel function used was Radial Basis Function (RBF).

In subsections 2.1 and 2.2, the systems developed for the two approaches are described.

3.1 Complexity measures of the fanfics

In this approach, we obtain a vector of features for each text. This features vector is made up of the values obtained from the complexity measures. We have to distinguish between the different languages since complexity measures depend on them as can be seen in Table 1

Once we have the feature vector, an automatic classifier is trained with a SVM classifier, on known fanfics and applied to predict which are the closest authors for unknown fanfics.

3.2 Part-Of-Speech Tagger of the fanfics

In this approach, we obtain the POS tags for each fanfic text and apply the TF or TF.IDF to do the automatic classification the SVM algorithm. For this, we train with fanfics that have known candidates and predict the candidates of the unknown fanfics to obtain final measurements of the classifier's performance.

For English, Spanish and French we have used the *Freeling* tool to process texts and Python's *SciKit-Learn* libraries for automatic learning.

For Italian and Polish we used the *NLTK* Python library, and for Polish we used TreeTagger to get the POS tags for each text. We have also used Python libraries for automatic learning.

[1] http://nlp.lsi.upc.edu/freeling/

[2] http://www.nltk.org

[3] http://scikit-learn.org

Complexity measures	English	Spanish	French	Italian	Polish
Punctuation Marks	✓	✓	✓	✓	✓
Sentence Complexity Index	✓	✓	✓	✗	✗
Automated Readability Index	✓	✓	✓	✓	✓
μ Readability	✓	✓	✓	✗	✗
Dependency Tree Height	✓	✓	✗	✗	✗
FOG	✓	✗	✗	✗	✓
Flesch	✓	✗	✗	✗	✓
Flesch-Kincacid	✓	✗	✗	✗	✓
SMOG	✓	✗	✗	✗	✗
Lexical Complexity	✗	✓	✗	✗	✗
Spaulding Readability	✗	✓	✗	✗	✗
Fernández-Huerta Readability	✗	✓	✗	✗	✗
Flesch-Szigrist Readability	✗	✓	✗	✗	✗
Gutierrez Readability	✗	✓	✗	✗	✗
Minimum Age of Readability	✗	✓	✗	✗	✗
SOL	✗	✓	✓	✗	✗
Crawford	✗	✓	✗	✗	✗
Kandel-Models	✗	✗	✓	✗	✗
Dale Chall	✗	✗	✓	✗	✗
Flech-Vaca	✗	✗	✗	✓	✗
Gulpease	✗	✗	✗	✓	✗
Pisarek	✗	✗	✗	✗	✓

Table 1. Complexity measures computed by language

4 Experiments and Results

As indicated above, on the one hand we have evaluated the contribution of these complexity metrics to an author identification task through automatic learning. On the other hand, we have evaluated the contribution of Part-Of-Speech tags with TF and TF.IDF representation to this task.

The experiments we have conducted are parametrized on options like how the normalization of the vectors is computed (L2 over samples or over features), the weighting scheme used (TF or TF.IDF) or the maximum length of POS n-grams considered (from 2 up to 4). Thus, we have tried the runs, on the training set, shown in Table 2

Of all the experiments carried out on the training set (which results are shown here), we have selected those results that are more representative of our work. Our team submitted both systems, though only the script using complexity measures as features produced correct output with the test set.

The results obtained with the Complexity per sample L2 experiments can be seen in Table 3 and, as can be observed, they are quite low.

Features	L2-normalization	N-gram sizes	Weighting scheme
Complexity measures	per sample	-	-
Complexity measures	per feature	-	-
POS tags	per sample	1, 2	TF
POS tags	per feature	1, 2	TF
POS tags	per sample	1, 2	TF.IDF
POS tags	per feature	1, 2	TF.IDF
POS tags	per sample	1, 2, 3	TF
POS tags	per feature	1, 2, 3	TF
POS tags	per sample	1, 2, 3, 4	TF
POS tags	per feature	1, 2, 3, 4	TF

Table 2. Runs configurations

The highest results we on the training set are with the POS-TF (1,2,3,4-grams) with L2-normalization per sample, as shown in Table 4.

Problem	Language	Macro F1
Problem 00001	English	0.035
Problem 00002	English	0.143
Problem 00003	French	0.038
Problem 00004	French	0.27
Problem 00005	Italian	0.09
Problem 00006	Italian	0.299
Problem 00007	Polish	0.023
Problem 00008	Polish	0.358
Problem 00009	Spanish	0.021
Problem 00010	Spanish	0.287

Table 3. Complexity features with L2-normalization per sample

The tables above show the best results obtained on the two approaches studied in this work. We can see that there exists significant differences in the performance of these models depending on the problem and on the language. POS-tags based features behave better than complexity measures. Although F1 performance remains below 0.5 in most cases, there are some impressive results for Spanish in Problem00002, which reaches a 0.838 score.

4.1 Official results

Only the script where complexity measures where used as features for the classifier produced valid output on the test set in the TIRA [21] server (the platform used to run

Problem	Language	Macro F1
Problem 00001	English	0.435
Problem 00002	English	0.838
Problem 00003	French	0.475
Problem 00004	French	0.605
Problem 00005	Italian	0.365
Problem 00006	Italian	0.512
Problem 00007	Polish	0.123
Problem 00008	Polish	0.447
Problem 00009	Spanish	0.373
Problem 00010	Spanish	0.404

Table 4. POS-TF (1,2,3,4-grams) with L2-normalization per sample

experiments). As seen above for experiments with training data, this approach results in a poor performance. The results obtained reported an overall score of 0.149 in macro-F1. Table 5 shows official results obtained.

Problem	Macro F1
Problem 00001	0.110
Problem 00002	0.202
Problem 00003	0.078
Problem 00004	0.235
Problem 00005	0.102
Problem 00006	0.109
Problem 00007	0.052
Problem 00008	0.276
Problem 00009	0.032
Problem 00010	0.296

Table 5. Official results in PAN@CLEF 2018. Complexity features with L2-normalization per sample.

5 Conclusions

We have computed several complexity measures for different languages and tested their convenience as features for authorship identification. Also, POS-tag n-grams have been explored as features for this task. From our experiments and the results obtained we can conclude that the complexity metrics considered are not very helpful to identify the author of a text. This could be explained by the low number of aspects captured by these features, which basically rely on length sentences or the number of syllables in a word. Also, the merge of all this smaller characteristics (rare words, punctuation marks, sentence length...) into a final index of readability or complexity may have nothing to do with author style or characterization.

Our second approach, the use of POS-tags, seems a better approach to the problem, although results are from very bad (in the case of Polish, Problem00007) to very good (Spanish, Problem00002). These results need of further analysis.

As future work, we plan to combine both approaches and to use base metrics in complexity indexes rather than the final values proposed by complexity related formulas. Language modeling approaches appear as a natural way of author representation, but these models need of far more data to be trained than just 20 texts per class. In this case, a model based on bayesian networks [19] could be used, as these models do not need large training data sets.

6 Acknowledgments

This work has been partially funded by the Spanish Government under the REDES Project (TIN2015-65136-C2-1-R).

References

1. Anula, A.: Lecturas adaptadas a la enseñanza del español como l2: variables lingüísticas para la determinación del nivel de legibilidad. La evaluación en el aprendizaje y la enseñanza del español como LE L 2, 162–170 (2008)
2. Bird, S., Loper, E.: Nltk: the natural language toolkit. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. p. 31. Association for Computational Linguistics (2004)
3. Blanco Pérez, A., Gutiérrez Couto, U.: Legibilidad de las páginas web sobre salud dirigidas a pacientes y lectores de la población general. Revista española de salud pública 76(4), 321–331 (2002)
4. Broda, B., Niton, B., Gruszczynski, W., Ogrodniczuk, M.: Measuring readability of polish texts: Baseline experiments. In: LREC. pp. 573–580 (2014)
5. Contreras, A., Garcia-Alonso, R., Echenique, M., Daye-Contreras, F.: The sol formulas for converting smog readability scores between health education materials written in spanish, english, and french. Journal of health communication 4(1), 21–29 (1999)
6. Crawford, A.N.: A spanish language fry-type readability procedure: Elementary level. bilingual education paper series, vol. 7, no. 8. (1984)
7. Dale, E., Chall, J.S.: A formula for predicting readability: Instructions. Educational research bulletin pp. 37–54 (1948)
8. Flesch, R.: A new readability yardstick. Journal of applied psychology 32(3), 221 (1948)
9. Franchina, V., Vacca, R.: Adaptation of flesh readability index on a bilingual text written by the same author both in italian and english languages. Linguaggi 3, 47–49 (1986)
10. García López, J.: Legibilidad de los folletos informativos. Pharmaceutical Care España 3(1), 49–56 (2001)
11. de Granada Barrio-Cantalejo, D.S., Simón-Lorda, P., Melguizo, M., Escalona, I., Marijuán, M., Hernándo, P., et al.: Validación de la escala inflesz para evaluar la legibilidad de los textos dirigidos a pacientes (2008)

12. Kandel, L., Moles, A.: Application de lâĂŹindice de flesch à la langue française. Cahiers Etudes de Radio-Télévision 19, 253–274 (1958)
13. Kestemont, M., Tschugnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)
14. Lucisano, P., Piemontese, M.E.: Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana. Scuola e città 3(31), 110–124 (1988)
15. Mc Laughlin, G.H.: Smog grading-a new readability formula. Journal of reading 12(8), 639–646 (1969)
16. MISTRÍK, J.: Meranie zrozumitel'nosti prehovoru. Slovenská reč 33 (1968)
17. Muñoz, M.: Legibilidad y variabilidad de los textos. Boletín de Investigación Educacional, Pontificia Universidad Católica de Chile, 21 2, 13–26 (2006)
18. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: LREC2012 (2012)
19. Pearl, J.: From bayesian networks to causal networks. In: Mathematical models for handling partial knowledge in artificial intelligence, pp. 157–182. Springer (1995)
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research 12(Oct), 2825–2830 (2011)
21. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
22. Ramírez-Puerta, M., Fernández-Fernández, R., Frías-Pareja, J., Yuste-Ossorio, M., Narbona-Galdó, S., Peñas-Maldonado, L.: Análisis de legibilidad de consentimientos informados en cuidados intensivos. Medicina Intensiva 37(8), 503–509 (2013)
23. Rodríguez, T.: Determinación de la comprensibilidad de materiales de lectura por medio de variables lingüísticas. Lectura y vida 1(1), 29–32 (1980)
24. Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., Drndarevic, B.: Making it simplext: Implementation and evaluation of a text simplification system for spanish. ACM Transactions on Accessible Computing (TACCESS) 6(4), 14 (2015)
25. Senter, R., Smith, E.A.: Automated readability index. Tech. rep., CINCINNATI UNIV OH (1967)
26. Spaulding, S.: A spanish readability formula. The Modern Language Journal 40(8), 433–441 (1956)
27. Stamatatos, E., Rangel, F., Tschuggnall, M., Kestemont, M., Rosso, P., Stein, B., Potthast, M.: Overview of PAN-2018: Author Identification, Author Profiling, and Author Obfuscation. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Initiative (CLEF 18). Springer, Berlin Heidelberg New York (Sep 2018)