Classifying Written Texts Through Rhythmic Features

Classifying Written Texts Through Rhythmic Features

Mihaela Balint[1], Mihai Dascalu[1(✉)], and Stefan Trausan-Matu[1,2]

[1] Computer Science Department, University Politehnica of Bucharest, Bucharest, Romania
{mihaela.balint, mihai.dascalu, stefan.trausan}@cs.pub.ro
[2] Research Institute for Artificial Intelligence of the Romanian Academy, Bucharest, Romania

Abstract. Rhythm analysis of written texts focuses on literary analysis and it mainly considers poetry. In this paper we investigate the relevance of rhythmic features for categorizing texts in prosaic form pertaining to different genres. Our contribution is threefold. First, we define a set of rhythmic features for written texts. Second, we extract these features from three corpora, of speeches, essays, and newspaper articles. Third, we perform feature selection by means of statistical analyses, and determine a subset of features which efficiently discriminates between the three genres. We find that using as little as eight rhythmic features, documents can be adequately assigned to a given genre with an accuracy of around 80 %, significantly higher than the 33 % baseline which results from random assignment.

Keywords: Rhythm · Text classification · Natural language processing · Discourse analysis

1 Introduction

Rhythm refers to the quest for harmonious proportions in all creative acts, which is essential for both human emotion and cognition. Rhythm brings thoughts and feelings to resonance, and facilitates understanding, remembering, and learning [1]. A creative piece is built as an ensemble of identical and different units, and rhythm emerges as a particular succession of these units. Examples of units are musical beats, linguistic phonemes, or colors and shapes used in paintings.

Text classification or categorization is the task of assigning a written document to a class from a set of predefined classes. The increasing importance of this task follows the increasing amount of textual information available online, and the need to efficiently index and retrieve such information. Researchers have approached the problem using statistical methods and machine learning, with the latter attaining accuracies comparable to the human expert standard. In machine learning, the distinctive features of individual classes are learned from a set of pre-classified documents. Preferred features include single words, syntactic phrases (two or more words plus the syntactic relationship between them), or n-grams [2]. A high number of words in the vocabulary leads to a high

C. Dichev and G. Agre (Eds.): AIMSA 2016, LNAI 9883, pp. 121–129, 2016.
DOI: 10.1007/978-3-319-44748-3_12

number of features, difficult or impossible to handle by classifiers. Even in the simplest case of single words, the resulting high dimensional feature space requires efficient algorithms of feature selection prior to entering inductive learning algorithms [3].

Our hypothesis is that the communicative purpose of a text influences significantly the rhythm of that text; thus, rhythmic features would become predictors for text categorization. The purpose of this work is to test this hypothesis, by evaluating how well rhythmic features extracted from already categorized text function as predictors. Section 2 presents relevant studies in rhythm analysis. Section 3 describes the first two steps of our method, namely the proposed set of features, and the feature extractor (together with the three corpora selected to demonstrate its use). The third step, namely feature selection, is discussed in Sect. 4, followed by the results of the classification using the selected features. Section 5 is dedicated to conclusions and future work.

2 Related Work

There are multiple perspectives on what constitutes linguistic rhythm analysis, and, most of the time, metrical phonology is implied. Phonology is the branch of linguistics that investigates the systematic organization of sounds in languages. Metrical phonology uses syllabification (at word level) and constituency parsing (at sentence level) to create a hierarchy of stresses inside clauses. The stress phenomenon refers to the relative emphasis placed on a syllable (word level) or syntactic category (sentence level). Rules for stress assignment in English are presented in seminal works written by Chomsky and Halle [4], and Liberman and Prince [5], while an analysis on French Literary text is performed by Boychuk et al. [6].

Several works compare rhythmic behavior in language and music, from which the concept of rhythm is derived. Jackendoff and Lerdahl [7] carry out a complete grammatical parallel between the tree structures used to represent rhythm in language and music. This profound similarity could be explained by Barbosa and Bailly's [8] theory that humans have an internal clock which needs to synchronize with the external clock of the stimulus (the meter in language, or the beat in music). The internal clock hypothesis is in accordance with Beeferman's [9] study, which demonstrates that sentences with a higher probability of occurrence, i.e. sentences that are actually preferred by writers, are more rhythmical. For this result, he uses a corpus of over 60 million syllables of Wall Street Journal text, in both its original form and in a second form, altered to randomize word order inside sentences. He finds that the stress entropy rate is higher in the second case. The rhythm of language appears to be culturally regulated. Galves et al. [10] extract streams of stresses from corpora of newspaper articles written in both European and Brazilian Portuguese, and use Variable Length Markov Chains [11] to model rhythmic realization in the two corpora, arriving at different final models. Where cultural background influences linguistic rhythm, it similarly influences musical rhythm, as shown by Patel and Daniele [12]. As a tool of comparison, they use the normalized Pairwise Variability Index (nPVI), introduced by Grabe and Low [13] to capture the difference in duration between successive vocalic intervals. Patel and Daniele contrast the nPVI's of spoken English and French with the nPVI's computed from English and French instrumental music scores, and obtain

statistically significant differences (in the same direction for both language and music, albeit smaller in music). Their conclusion is strengthened by London and Jones's [14] refined method to compute the nPVI of music.

However, linguistic rhythm does not have to be restricted to metrics. According to Boychuk et al. [6], a high degree of rhythmization is achieved whenever there are elements with a high frequency of occurrence and the occurrences are close to each other. They build a tool for the French language, with the option of highlighting repetitions of specific words, vowels, consonants, or phonemic groups. Other features include detection of coordinated units, same-length units, or affirmative, interrogative, exclamatory, and elliptical sentences. In our research, we adopt this more general view of rhythm as repetition and alternation of linguistic elements.

3 Method Description

This section describes our method for the rhythmic evaluation of texts. We model a text as a sequence of elementary units. To separate units, we use the loci where readers naturally insert pauses, and we obtain two kinds of units: sentences (separated by sentence boundaries), and punctuation units (separated by punctuation markers in general). For example, there are four punctuation units in the sentence "Shall we expand(1), be inclusive(2), find unity and power(3); or suffer division and impotence(4)". Rhythmic features will characterize individual units (e.g. the length of a unit in syllables) or interactions between neighboring units (e.g. the anaphora phenomenon – two or more units which start with the same sequence of words). Subsection 3.1 presents the pre-classified data chosen as ground truth for our model, while Subsect. 3.2 describes the full set of rhythmic features, prior to the step of feature selection.

3.1 Data Collection

In our method, any text corpus can be a data source. We opted for the comparison of three corpora, chosen to exhibit various degrees of rhetoric: a corpus of famous speeches (extracted from http://www.famous-speeches-and-speech-topics.info/famous-speeches/), student essays from the Uppsala Student English (USE) corpus (http://ota.ox.ac.uk/desc/2457), and the raw texts from the RST-DT corpus of Wall Street Journal articles [15]. Table 1 presents the relevant properties of the three datasets. In order to obtain accurate corpus statistics, the full datasets underwent feature extraction. Subsequently, when we evaluated the relevance of rhythmic features in text classification, we balanced the data, by keeping the longest (in number of sentences) 110 documents from each category. This does not eliminate imbalance pertaining to age, gender, or nationality, but this is a pilot study created to demonstrate the strength and scalability of our model. The model as it is now can be further used to look for significant differences in rhythmicity according to age, gender, or nationality.

Our feature extractor was implemented in the Python programming language, using the NLTK package for natural language processing (http://www.nltk.org/) and the SQLite3 package for interfacing with SQL databases (https://www.sqlite.org/). We

Table 1. Statistics of the three datasets.

Dataset	# of documents	# of sentences
Speeches	110	14,111
RST-DT	380	8,281
USE	1,266	49,851

loaded the raw content of documents into three distinct databases (one for each corpus), that were subsequently filled with the extracted document features.

3.2 Rhythmic Features

The analysis presented in this paper relies on five main categories of features: organizational, lexical, grammatical, phonetical, and metrical. They are refined versions of the features we introduced in previous work [16].

Organizational features include the average word length, the length of units in either words or syllables, and patterns of length variation along sequences of units. The number of syllables is computed using the CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). Rhythm can occur from a particular alternation of long and short units. We count rising (successive units keep getting longer), falling (they keep getting shorter), alternating (shorter and longer units alternate), or repetitive (same-length) patterns, and the maximum length of such patterns. For frequent words in a document, we determine how often they occur in the beginning (first third), at the middle (second third), or at the end (last third) of units.

Lexical features refer to types of lexical repetition. Words or n-grams in a document are considered frequent if their number of occurrences exceeds the value *(text_length * threshold/n_gram_length)*, where the threshold can be varied. Stop words are eliminated in the detection of frequent words, but accepted inside n-grams which contain at least two non stop words.

In the case of frequent words or n-grams, there is no restriction on the maximum distance between successive occurrences. We use a variable parameter *delta* to impose this kind of restriction when counting *duplicated units* (several identical units), *anaphora* (several units starting with one or more identical words), *epistrophes* (several units ending in the same word(s)), *symploces* (several units presenting a combination of the anaphora and epistrophe phenomena), *anadiploses* (a second unit starting the way a first unit ends), *epanalepses* (single units starting and ending with the same word(s)). We consider only the maximal and non-redundant occurrences of these phenomena. Therefore, if n neighboring units have the same start, that is considered to be a single anaphora. If they share w words, the anaphora is counted only once, not once for every initial substring of the maximal one.

Grammatical features consider the frequencies of parts-of-speech, commas, and types of sentence boundaries (full-stops, question marks, exclamation marks) in each document. Each sentence is parsed using the Stanford Parser (http://nlp.stanford.edu) and the resulting trees of constituents are used to detect syntactic parallelism between neighboring sentences (located within a given distance of each other). Parallelism can

be checked either for the entire or only up to a given depth of the tree. With the obvious exception of terminal nodes (corresponding to actual words), nodes in equivalent positions should be labelled with the same main part-of-speech category. Another kind of noun, verb, adjective, etc. is allowed in place of a kind of noun, verb, adjective, but a noun cannot be in place of a verb, for example. Figure 1 illustrates this point using an excerpt from Jesse Jackson's speech "Common ground and common sense". Non-identical nodes which still fulfill the standard for syntactic parallelism are shown in boldface.

```
(ROOT
   (S
     (NP (PRP We))
     (VP (VBP have) (NP (JJ public) (NNS accommodations)))
     (. .)))
  (ROOT
    (S (NP (PRP We)) (VP (VBP have) (NP (JJ open) (NN housing))) (. .)))
```

Fig. 1. Two syntax trees marked for syntactic parallelism.

Phonetical features refer to phonetical repetition, in much the same way that lexical features refer to lexical repetition The representative phenomena are the ones of *assonance* (the repetition of a vocalic phoneme over a small amount of text), *alliteration* (the same for consonants), and *rhyme* (defined here as the repetition of the same phonemic sequence, not necessarily at the end of words).

To compute *metrical* features, for each syllabified document the complete stream of stresses (primary, secondary, or no-stress) is extracted. We record the frequencies of units built with an odd number of syllables, and of units ending in a stressed syllable. For the latter feature, the stress from monosyllabic stop words is removed, because, in practice, monosyllabic words are sometimes stressed, sometimes not, and the CMU dictionary does not handle this problem satisfactorily.

4 Feature Selection, Results and Discussions

This section describes our approach to feature selection and testing for feature relevance, together with corpus statistics and classification results for the refined set of features.

Discriminant Function Analysis (DFA) is a statistical analysis which predicts a categorical dependent variable (a class from a set of predefined classes) from the behavior of several independent variables (called predictors). Performing a DFA over a given dataset requires that independent variables respect a normal distribution, and that no discriminating variables be a linear combination of other variables. These requirements guide the reduction of the feature space described in the previous section. First, we remove all features which demonstrate non-normality. Second, we assess multicollinearity based on pair-wise correlations with a correlation coefficient $r > .70$, and filter multicollinear features to keep only the feature with the strongest effect in the model (see Table 2 for the final list of rhythmic features and their descriptive statistics).

The results indicate rhetorical preferences. Conciseness and fluency are achieved through the usage of short words and the alternation of long and short units. The main themes of a document, captured in frequent words, tend to occupy the middle of units, with the beginning and end of units functioning as background and elaboration. Essays contain more frequently used words and fewer commas, which might be explained by the lower English proficiency of their authors. Speeches do not repeat many words, but they make the most use of figures of speech based on repetition, especially anaphora. Anaphora in reference to punctuation units, not sentences, are particularly indicative of a document's genre.

Table 2. General statistics of rhythmic features - *M*(*SD*).

Rhythmic feature	Article	Essay	Speech
# of syllables per word	1.576 (0.081)	1.444 (0.088)	1.475 (0.094)
% of rising word-length patterns	0.171 (0.052)	0.176 (0.051)	0.190 (0.057)
% of falling word-length patterns	0.182 (0.052)	0.173 (0.045)	0.170 (0.042)
% of repetitive word-length patterns	0.035 (0.031)	0.042 (0.029)	0.032 (0.027)
longest rising word-length sequence	2.140 (0.807)	2.340 (0.805)	2.640 (1.002)
longest falling word-length sequence	2.150 (0.826)	2.330 (0.692)	2.410 (0.881)
longest repetitive word-length sequence	0.840 (0.614)	1.060 (0.529)	1.030 (0.642)
% of falling syllable-length patterns	0.187 (0.052)	0.179 (0.044)	0.172 (0.045)
longest rising syllable-length sequence	2.050 (0.806)	2.350 (0.840)	2.620 (0.967)
longest repetitive syllable-length sequence	0.570 (0.582)	0.860 (0.438)	0.850 (0.618)
% of frequent words at the beginning of sentences	0.203 (0.096)	0.172 (0.061)	0.143 (0.092)
% of frequent words at the end of sentences	0.199 (0.086)	0.220 (0.068)	0.174 (0.103)
% of frequent words at the beginning of punctuation units	0.210 (0.080)	0.167 (0.056)	0.149 (0.078)
% of frequent words at the end of punctuation units	0.285 (0.089)	0.297 (0.065)	0.308 (0.088)
# of words deemed frequent	41.33 (20.26)	53.47 (19.70)	39.34 (22.44)
normalized # of sentence anaphora	0.005 (0.003)	0.006 (0.003)	0.007 (0.004)
normalized # of punctuation unit anaphora	0.005 (0.003)	0.008 (0.004)	0.013 (0.006)
normalized # of commas	0.060 (0.015)	0.041 (0.015)	0.060 (0.015)
% of sentences with an odd # of syllables	0.507 (0.068)	0.509 (0.062)	0.497 (0.068)

Table 3 denotes the features that vary significantly between the three datasets, in descending order of effect size, determined through a multivariate analysis of variance (MANOVA) [17, 18]. There is a significant difference among the three datasets in terms of rhythmic features, Wilks' $\lambda = 0.259$, $F(28, 628) = 21.635$, $p < .001$ and partial $\eta^2 = .491$.

We predict the genre of a given text using a stepwise Discriminant Function Analysis (DFA) [19]. Only eight variables from Table 2 (marked with italics) are deemed significant predictors, denoting complementary features of rhythmicity: the number of syllables per word, the normalized number of falling syllable-length patterns, the percentage of frequent words located at the end of sentences, the percentage of frequent words located at the beginning of punctuation units, the number of words deemed frequent, the normalized number of sentence anaphora, the normalized number of punctuation unit anaphora, and the normalized number of commas. Figure 2 depicts the two retained canonical discriminant functions ($\chi^2(df = 7) = 171.773$, $p < .001$).

Table 3. Tests of between-genre effects for significantly different rhythmic features.

Rhythmic feature	df	*F*	*p*	η^2 partial
normalized # of punctuation unit anaphora	2	96.433	<.001	.371
# of syllables per word	2	68.483	<.001	.295
normalized # of commas	2	55.41	<.001	.253
% of frequent words at the beginning of punctuation units	2	20.335	<.001	.111
# of words deemed frequent	2	14.84	<.001	.083
% of frequent words at the beginning of sentences	2	13.968	<.001	.079
longest rising syllable-length sequence	2	11.826	<.001	.067
longest repetitive syllable-length sequence	2	9.885	<.001	.057
longest rising word-length sequence	2	9.077	<.001	.053
normalized # of sentence anaphora	2	8.441	<.001	.049
% of frequent words at the end of sentences	2	7.646	.001	.045
longest repetitive word-length sequence	2	4.6	.011	.027
% of rising word-length patterns	2	3.909	.021	.023
% of repetitive word-length patterns	2	3.255	.040	.020

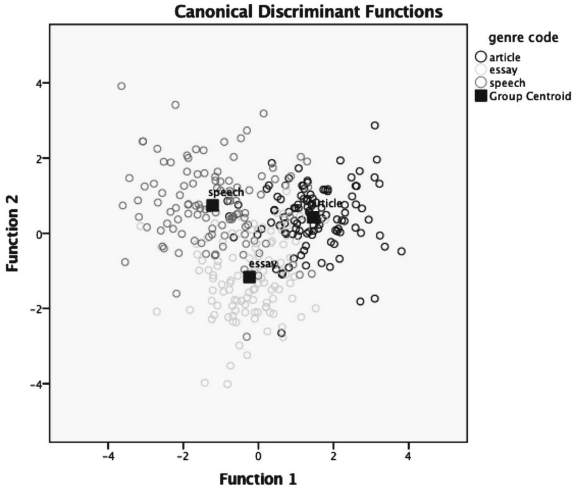

Fig. 2. Separation of genres based on canonical discriminant functions derived from rhythmic features.

The results presented in Table 4 show that the DFA based on these eight features correctly allocated 269 out of 330 texts, for an accuracy of 81.51 % (the chance level for this analysis being 33.33 %). Using leave-one-out cross-validation (LOOCV), the DFA achieved an accuracy of 79.69 % (see the confusion matrix in Table 4 for detailed results). The resulting weighted Cohen's Kappa of 0.723 demonstrates substantial agreement between the actual genre and the genre assigned by the model.

Table 4. Confusion matrix for DFA classifying texts pertaining to different genres.

	Genre	Predicted Group Membership		
		Article	Essay	Speech
Original	Article	99	7	4
	Essay	14	84	12
	Speech	8	16	86
Cross-validated	Article	97	8	5
	Essay	16	81	13
	Speech	8	17	85

5 Conclusions and Future Work

The main purpose of this paper was to test the ability to predict the genre of a given document based on rhythmic features. We used a dataset of 330 documents, equally distributed between three genres: famous speeches, student essays, and newspaper articles. A Discriminant Function Analysis based on the most predictive eight features of our model performed classification with an accuracy of around 80 %, significantly higher than the 33.33 % baseline represented by a trivial classifier which randomly assigns a document to one of the three genres. Our work is of interest to both linguists and computer scientists, as we provide both an automated method to study the rhythmic properties of English text, and a feature extractor that can be used in text categorization. Moreover, our method is highly extensible and can be used to study the rhythmic properties of other corpora. For example, it is possible to test the intuition that words are longer, on average, in a corpus of scientific articles.

We consider two directions for the development of this model. First, in terms of refining our rhythmic features, we intend to find a reliable solution to characterize words absent from the CMU dictionary. The number of anaphora was shown to greatly vary when calculated on punctuation units instead of sentences. Similar results may occur for syntactic parallelism or other stylistic devices, when we experiment with other types of units, such as elementary discourse units (EDUs). EDUs are units separated on rhetorical grounds, which leads us to our second intended development. Using the RST-DT corpus of newspaper articles, already annotated with rhetorical relations, we can study the correlation between the rhetorical role of an EDU and its rhythmic properties, with viable applications in rhetorical relation labelling.

Acknowledgements. The work presented in this paper was partially funded by the EC H2020 project RAGE (Realising and Applied Gaming Eco-System) http://www.rageproject.eu/ Grant agreement No 644187.

References

1. Lefebvre, H.: Rhythmanalysis: Space. Time and Everyday Life. Continuum, London (2004)
2. Fürnkranz, J.: A study using n-gram features for text categorization. Austrian Research Institute for Artificial Intelligence, Wien (1998)
3. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: 14th International Conference on Machine Learning (ICML 1997), pp. 412–420. Morgan Kaufmann Publishers Inc., San Francisco (1997)
4. Chomsky, N., Halle, M.: The Sound Pattern of English. Harper & Row, New York (1968)
5. Liberman, M., Prince, A.: On stress and linguistic rhythm. Linguist. Inq. **8**(2), 249–336 (1977)
6. Boychuk, E., Paramonov, I., Kozhemyakin, N., Kasatkina, N.: Automated approach for rhythm analysis of french literary texts. In: 15th Conference of Open Innovations Association FRUCT, pp. 15–23. IEEE, St. Petersburg (2014)
7. Jackendoff, R., Lerdahl, F.: A grammatical parallel between music and language. In: Clynes, M. (ed.) Music, Mind, and Brain, pp. 83–117. Springer, Heidelberg (1982)
8. Barbosa, P., Bailly, G.: Characterisation of rhythmic patterns for text-to-speech synthesis. Speech Commun. **15**(1–2), 127–137 (1994)
9. Beeferman, D.: The rhythm of lexical stress in prose. In: 34th Annual Meeting of the Association for Computational Linguistics (ACL). ACL, Santa Cruz (1996)
10. Galves, A., Galves, C., Garcia, J., Garcia, N., Leonardi, F.: Context tree selection and linguistic rhythm retrieval from written texts. Ann. Appl. Stat. **6**(1), 186–209 (2012)
11. Buhlmann, P., Wyner, A.J.: Variable length Markov chains. Ann. Stat. **27**(2), 480–513 (1999)
12. Patel, A.D., Daniele, J.R.: An empirical comparison of rhythm in language and music. Cognition **87**(1), B35–B45 (2003)
13. Grabe, E., Low, E.L.: Durational variability in speech and the rhythm class hypothesis. In: Gussenhoven, C., Warner, N. (eds.) Papers in Laboratory Phonology, pp. 515–546. Mouton de Gruyter, Berlin (2002)
14. London, J., Jones, K.: Rhythmic refinements to the nPVI measure: a reanalysis of Patel & Daniele (2003a). Music Percept. Interdisc. J. **29**(1), 115–120 (2011)
15. Carlson, L., Marcu, D., Okurowski, M.E.: Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In: 2nd SIGdial Workshop on Discourse and Dialogue (SIGDIAL 2001), vol. 16, pp. 1–10. Association for Computational Linguistics, Stroudsburg (2001)
16. Balint, M., Trausan-Matu, S.: A critical comparison of rhythm In music and natural language. Ann. Acad. Rom. Scientists Ser. Sci. Technol. Inf. **9**(1), 43–60 (2016)
17. Stevens, J.P.: Applied Multivariate Statistics for the Social Sciences. Lawrence Erblaum, Mahwah (2002)
18. Garson, G.D.: Multivariate GLM, MANOVA, and MANCOVA. Statistical Associates Publishing, Asheboro (2015)
19. Klecka, W.R.: Discriminant Analysis. Quantitative Applications in the Social Sciences Series, vol. 19. Sage Publications, Thousand Oaks (1980)