Are Fictional Voices Distinguishable? Classifying Character Voices in Modern Drama

Krishnapriya Vishnubhotla
Department of Computer Science
University of Toronto
Toronto, Canada
vkpriya@cs.toronto.edu

Adam Hammond
Department of English
University of Toronto
Toronto, Canada
adam.hammond@utoronto.ca

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, Canada
gh@cs.toronto.edu

Abstract

According to the literary theory of Mikhail Bakhtin, a *dialogic* novel is one in which characters speak in their own distinct voices, rather than serving as mouthpieces for their authors. We use text classification to determine which authors best achieve dialogism, looking at a corpus of plays from the late nineteenth and early twentieth centuries. We find that the SAGE model of text generation, which highlights deviations from a background lexical distribution, is an effective method of weighting the words of characters' utterances. Our results show that it is indeed possible to distinguish characters by their speech in the plays of canonical writers such as George Bernard Shaw, whereas characters are clustered more closely in the works of lesser-known playwrights.

1 Introduction

The concept of *dialogism* has been a notable focus in recent computational literary scholarship (Brooke et al., 2017; Hammond and Brooke, 2016; Muzny et al., 2017). As theorized by Russian literary critic Mikhail Bakhtin (2013), a dialogic novel is one in which characters present "a plurality of independent and unmerged voices and consciousnesses, a genuine polyphony of fully valid voices". Bakhtin presents Dostoevsky as the preeminent dialogic author, arguing that his novels are "multi-accented and contradictory in [their] values", whereas the works of other novelists like Tolstoy are *monologic* or homogeneous in their style, with characters reflecting the prejudices as well as the distinctive mannerisms of their authors.

While previous computational studies of dialogism take this definition of dialogism for granted and seek to model it, here we take a step back to pose a series of fundamental questions: Can the voices of characters be distinguished in fictional texts? Which computational techniques are most effective in making these distinctions? Are certain authors better than others at creating characters with distinctive voices and do these authors tend to be more canonical? Focusing, for pragmatic purposes, on plays rather than novels, we argue here that character voices can, in the work of certain authors, be readily distinguished; that SAGE (Sparse Additive Generative) models (Eisenstein et al., 2011) are especially powerful in making these distinctions; and that canonical authors are, in our small sample, more successful in creating distinctive character voices than are less canonical authors.

2 Related Work

Computational approaches to the authorship attribution problem involve using certain textual features, called *style markers*, to build a representation of an author's texts, which is then passed to a classification algorithm. Stop-word frequencies, part-of-speech trigrams, and structural features such as sentence lengths have been shown to be good indicators of author identity (Stamatatos, 2009). The earliest work in authorship attribution focused on discovering the stylistic markers that would reveal the identity of the author or authors of disputed works (Mosteller and Wallace, 1963), and the bulk of contemporary work in authorship attribution continues in this vein (Rybicki, 2018). Our work draws on an alternative tradition that uses the techniques of authorship attribution to investigate what J. F. Burrows, in a study of the novels of Jane Austen, calls *idiolects*, the distinctive stylistic patterns of individual speakers within texts (Burrows, 1987). Whereas Burrows's approach focuses on very common words and relies on statistical methods whose results are not easily interpretable, our particular application requires us

Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 29–34
Minneapolis, MN, USA, June 7, 2019.

to employ methods that are sensitive to rare and infrequent words, and whose results allow us to distinguish between stylistic and topical phenomena.

Recently, machine learning methods have been applied in computational stylometry for authorship attribution tasks, and also in the context of style transfer for texts. Bagnall (2015) uses a recurrent neural network (RNN) based model for the author identification task. Since neural architectures massively overfit the training set unless used with large datasets, the authors propose a shared recurrrent layer, with only the final softmax layer being author-specific. Shrestha et al. (2017) use convolutional neural networks (CNNs) over character n-grams for authorship attribution, which proves to be more interpretable than the former in identifying important features.

3 Corpus

Our corpus consists of plays published in the late 19th and early 20th centuries by George Bernard Shaw, Oscar Wilde, Cale Young Rice, Sydney Grundy, Somerset Maugham, Arthur Wing Pinero, and Hermann Sudermann (whose plays are translated from German) — giving a total of 63 plays. We would ideally have examined character dialogue in novels, Bakhtin's preferred genre, but the problem of sufficiently reliable quote attribution for novels remains unsolved. However, in plays, each utterance is explicitly labeled with the name of the character who speaks it. We use GutenTag (Brooke et al., 2015) to extract all plays from the specified authors, restricting the year of publication to 1880–1920 to roughly capture the literary period from which Bakhtin developed his theory of dialogism.

4 Methodology

Our primary method of measuring the distinguishability of character voices is classification. Our task is to build a classifier able to correctly discriminate between the speech of different characters. We perform experiments using several feature sets, in order to capture stylistic aspects that are syntactic as well as lexical. These include surface, syntactic, and generative topic-modeling induced features. Generative models that we used include latent Dirichlet allocation (Blei et al., 2003), naive Bayes, and SAGE models (Eisenstein et al., 2011). Accuracy of classification is measured using the F_1 score, which strikes a balance between precision and recall. We experiment with both support vector machine (SVM) and logistic regression classifiers.

In addition, we experiment with vector representations of words as features. We use distributed word vectors trained on the Wikipedia corpus using the word2vec algorithm (Mikolov et al., 2013). Each dialogue is represented as a weighted average of the individual word vectors, where the weights are TF-IDF weights, or obtained from the SAGE algorithm.

We also look at representations obtained from lexicons that score words across a discrete set of stylistic dimensions. Brooke and Hirst (2013) pick three dimensions to rate words along, the opposing polarities of which give us six styles: colloquial vs. literary, concrete vs. abstract, and subjective vs. objective. We also use the NRC Emotion Intensity Lexicon (EmoLex) (Mohammad, 2018b) and the NRC Valence, Arousal, and Dominance Lexicon (VAD Lexicon) (Mohammad, 2018a). The former provides real-valued intensity scores for four basic emotions — anger, fear, sadness, and joy, and the latter for the three primary dimensions of word meaning — valence, arousal, and dominance. The scores along each dimension are normalized to give us a set of values ranging from 0 to 1. Principal component analysis (PCA) of these vectors gives us an insight into which authors are the most successful at creating characters whose style is highly mutually distinguishable.

We repeat these experiments for "artificial plays" constructed by sampling a random subset of characters either across plays (strategy 1) or across authors (strategy 2). Intuitively, we expect the character speech in these artificial plays to be more readily distinguishable than in actual plays, because the characters are likely to discuss a wider variety of topics and to come from a wider variety of classes, professional milieus, and dialect communities than a group of characters in any actual play (strategies 1 and 2), and because the characters are the creations of different authors, each with their own distinct stylistic fingerprints (strategy 2).

5 Classification Models

In this section, we describe the two main models of classification that we employed. All hyperparameters in both models are tuned using grid-search, along with 5-fold cross validation.

5.1 Lexical and Syntactic features

Our first feature set consists of lexical, syntactic and structural features. These include average sentence and word lengths, type-token ratio, and proportion of function words in each sentence. We also use n-gram frequencies of word and part-of-speech tags, where $n \in \{1,2,3\}$, and dependency triples of the form *(head-PoS, child-PoS, DepRel)* from the dependency parse of each sentence, where *child-PoS* and *head-PoS* are the parts-of-speech of the current word and its parent node, and *DepRel* is the dependency relation between them. All proper nouns in our sentences are masked, as they often serve as indicative clues as to who the speaker is or is not.

Because word and PoS n-grams are very sparse features, the resulting feature vector has a relatively high dimensionality. We therefore pass it through a feature selection pipeline before classification. Two main feature selection algorithms are used: variance threshold and k-best selection. The former removes all features with a zero variance across samples — i.e, features that have the same value at each datapoint. The k-best selection algorithm then picks the top-k features according to some correlation measure. Here, we use the chi-squared statistic, which gets rid of the features that are the most likely to be independent of class and therefore irrelevant for classification. We pass this feature vector through a support vector machine (SVM) classifier.

5.2 Sentence Vectors with SAGE

Since we are dealing with a dataset that can contain very few samples per class, we need a model that is sensitive to low-frequency word features. We use the Sparse Additive Generative (SAGE) model of text, proposed by Eisenstein et al. (2011), which models the word distribution of each class as a vector of log-frequency deviations from a background distribution. We take the background distribution to be the average of the word frequencies across all classes. An alternative to the naive Bayes and LDA-like models of text generation, the SAGE model enforces a sparse prior on its parameters, which biases it towards rare and infrequent terms in the text.

We use the SAGE model to derive weights for each sentence (i.e, each quote) in our dataset. Sentence vectors are obtained by averaging the vector representation of each word in the sentence with its corresponding SAGE weight. Classification is performed by passing these sentence vectors to a logistic regression classifier.

Author	#Plays	Baseline	Avg F_1
Shaw	29	.153	.400
Wilde	6	.116	.376
Maugham	8	.137	.318
Grundy	4	.107	.283
Pinero	5	.090	.272
Sudermann	5	.084	.253
Rice	6	.151	.234
Weighted Avg.		.133	.342

Table 1: F_1 scores for classification of individual characters, by author, using lexical and syntactic features. Baseline is random classification with the class distribution of the training data. The final row reports the weighted average of the scores for each author, where the weights are proportional to the number of their plays in our dataset.

Author	Baseline	Avg F_1
Shaw	.148	.573
Wilde	.194	.376
Maugham	.182	.318
Grundy	.184	.283
Pinero	.140	.272
Sudermann	.119	.253
Rice	.186	.234
Average	.165	.329

Table 2: F_1 scores for classification, using lexical and syntactic features, of characters by each author in artificial plays generated by sampling characters from the all plays of that author. Baseline is computed in the same manner as in Table 1.

6 Results

We first present results for classification of individual characters with our lexical and syntactic features in Tables 1 and 2. We compare scores with a baseline that randomly generates predictions that respect the class distributions of the training data.

The classification scores are above the baseline for almost all the plays, though the absolute numbers themselves are not very high. Table 1 shows the average scores across all plays for each author, while Table 2 contains the average scores for the artificial plays. Shaw achieves the highest average score.

As expected, the scores for artificial plays are,

Author	Average F_1 Original plays	Artificial plays
Wilde	.641	.669
Shaw	.635	.630
Maugham	.662	.645
Sudermann	.538	.574
Grundy	.517	.517
Pinero	.458	.543
Rice	.181	.208
Weighted Avg.	.561	.540

Table 3: F_1 scores for classification of characters in original artificial plays using out SAGE classification model.

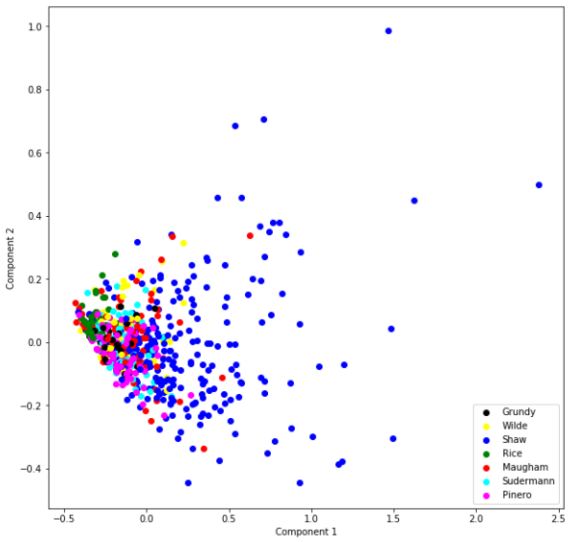

Figure 1: Plot of first two components of PCA on the lexical style vectors of each character of each author.

on average, higher than those of actual plays. We generate a maximum of 50 artificial plays for each author by sampling 7 characters from the complete set of characters, without repetition.

We achieve the best classification results, however, using the SAGE+word2vec classification algorithm described in Section 5.2. Table 3 shows the author-wise average F_1 scores for both original and artificial (strategy 1) plays. The average F_1 is higher still, at .605, for strategy 2 artificial plays (not presented in the table).

As an additional test, we performed PCA on vectors constructed using the style lexicons from Section 4. To construct our vectors, we replace our word2vec embeddings with a concatenated vector of the scores for each word along each of the 14 dimensions. Missing dimensions for words are assigned a score of zero. All the vectors are normalized along each dimension to account for variations in scale.

The results are shown in Figure 1, which plots the first two principal components. The two components combined account for 74.7% of the variance of the data. Each dot corresponds to a character in an actual play, and wider spacing between them indicates a wider range of styles and emotions. Even taking into account the fact that Shaw has significantly more plays, and thus more characters, than the other playwrights, he is nonetheless evidently the most successful, followed by Maugham, at creating characters with a wide range across all of the dimensions.

7 Discussion

Our work presents insights into a series of fundamental questions related to the phenomenon of literary dialogism and its tractability for computational analysis. The most fundamental is whether the voices of individual characters can be distinguished at all in literary texts. In a provocative argument in *Enumerations*, Andrew Piper uses computational methods to argue that "character-text" (the words used to describe characters) is — contrary to the intuitions of many literary scholars — relatively uniform within and across novels (Piper, 2018). Our work suggests that the same cannot be said of "dialogue-text" (the words that characters say). In a finding more in line with the intuitions of critics and the theories of Bakhtin, our experiment shows that the voices of characters can indeed be distinguished from one another, sometimes with quite high precision.

As to the question of whether certain authors are better able to distinguish their characters' voices than others, our results suggest that this is clearly the case. Although we approach the classification task from a variety of methodological perspectives, each of these reveals a continuum along which some playwrights are able to create distinctive character voices (e.g., Shaw) and some are not (e.g, Rice). That this continuum separates well-known playwrights like Shaw and Wilde from mostly forgotten playwrights like Pinero and Rice suggests that the ability to distinguish voices may be a property of more canonical — and, per-

haps, more talented — writers.[1] A larger sample size would be necessary to draw such conclusions definitively, however, as would an investigation of the effect of genre on the distinctiveness of character speech — for instance, whether comedy, which tends to put characters of different classes (and class dialects) in conversation, produces higher distinctiveness scores.

Our experiments with different feature sets also provide insights into how these characters are distinguishable from one another. SAGE, as an alternative to TF-IDF and naive Bayes measures of vocabulary usage, proves to be a very good indicator of which words are most distinctive for a particular character. At the character level, looking at the top features from the SAGE algorithm provides insights into the easiest types of stylistic distinction one can make while creating characters. Servants and butlers are easily recognizable by their use of words such as *'sir'*, *'yes'*, and *'please'*, and achieve a high classification score despite having relatively fewer quotes. In Shaw's *Pygmalion*, the character of The Flower Girl is distinguished by her unique vocabulary of words like *'ow'*, *'ai'*, *'–'*, *' 'm'*, *'ah'*, *'oo'*, etc. These kinds of lexical, dialectal features seem to be the most popular way of creating unique character voices.

The semantic and syntactic information captured by word2vec vectors forms the other key component of our analysis. While these dense vectors are not directly interpretable, we did attempt an initial clustering experiment with the word embeddings, which resulted in some insightful clusters. Proper nouns were grouped into one, another had words associated with tragedy (*sad, dreadful, miserable, awful, horrible, terrible, unfortunate*), and yet another cluster had *duty, servants, rank, ideals*. These are indicative of some stylistic aspect of words being captured by the embeddings which, when combined with the SAGE weights, boosts our classification performance. However, we reiterate that quantifying this is a hard-to-solve problem. Our analysis with lexicon-based vectors more concretely illustrates some of the stylistic dimensions along which characters and authors differ.

[1]Nonetheless, we acknowledge the alternative viewpoint expressed by one of the reviewers: "It could be that the characters from Rice are so rich and diverse that they cannot be classified and that Shaw's or Wilde's are so exaggerated or archetypal that even simple classification mechanisms can recognize them."

An interesting observation we make is that the artificial plays do not achieve a significantly higher score when compared to the original ones, despite the intuition that they must deal with more disparate topics. The number of sources of variance in creating these plays makes it hard to interpret this; performing more controlled experiments in the future might provide a better explanation.

8 Conclusion

We propose new techniques for classifying character speech in the works of seven modern dramatists. We show that SAGE models achieve the highest classification scores. Our results suggest that, in many dramatic works, characters are distinguishable with relatively high precision; that certain playwrights are better able to create distinctive character voices; and that these playwrights tend to be more canonical. Given the small size and restricted domain of our dataset, we treat these results are preliminary. Further investigation with a wider range of authors and genres, including novels, would aid us in drawing more decisive conclusions.

Acknowledgements

This work was supported financially by the Natural Sciences and Engineering Research Council of Canada. We are grateful to the anonymous reviewers for their helpful comments.

References

Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks. *arXiv preprint arXiv:1506.04891*.

Mikhail Bakhtin. 2013. *Problems of Dostoevsky's Poetics*. University of Minnesota Press.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2017. Using models of lexical style to quantify free indirect discourse in modernist fiction. *Digital Scholarship in the Humanities*, 32(2):234–250.

Julian Brooke and Graeme Hirst. 2013. A multi-dimensional Bayesian approach to lexical style. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 673–679.

J. F. Burrows. 1987. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford University Press.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1041–1048. Omnipress.

Adam Hammond and Julian Brooke. 2016. Project Dialogism: Toward a computational history of vocal diversity in English-language literature. In *Digital Humanities*, pages 543–544, Kraków.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Saif Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 174–184.

Saif Mohammad. 2018b. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Frederick Mosteller and David L. Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309.

Grace Muzny, Mark Algee-Hewitt, and Dan Jurafsky. 2017. Dialogism in the novel: A computational model of the dialogic nature of narration and quotations. *Digital Scholarship in the Humanities*, 32(supplement 2):ii31–ii52.

Andrew Piper. 2018. *Enumerations: Data and Literary Study*. University of Chicago Press.

Jan Rybicki. 2018. Partners in life, partners in crime? In Arjuna Tuzzi and Michele A. Cortelazzo, editors, *Drawing Elena Ferrante's Profile*, pages 109–119. Padova University Press.

Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 669–674.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.