Analyses of Characters in Dramatic Works by Using Document Embeddings

Mehmet Can Yavuz[1][0000−0003−1677−9496]

Faculty of Engineering and Natural Science, Sabancı University, Tuzla
Physics Department, Boğaziçi University, Bebek
İstanbul, Türkiye
mehmetyavuz@sabanciuniv.edu

Abstract. Shakespearean tragedies show clear antagonisms and the resolutions are rational which means they obey the Aristotelian unity-of-action principle. Any tragic play must rely upon its own movement. Therefore, it is complete and there should not be extra characters written only for the sake of resolution, so called ”Deus-ex-Machina”. In this work, Deus-ex-Machina characters are automatically detected using machine learning methods.

We first train unsupervised Doc2Vec network by using all plays of Shakespeare. Then, we collected all the lines uttered by each character in a separate document and extracted the document vectors. Thus, each character is represented with a vector in the semantic space of Shakespeare. We measure the semantic similarity between characters using the cosine difference, the angle between normalized vectors of each character document and we observe characters form a cluster.

According to this work, it is possible to detect Deus-ex-Machina characters. Examples of strong unity-of-action principle plays could be demonstrated as well as distinct characters. Dis/similar characters between the plays could also be shown.

Keywords: Document Embedding · Dramatic Works · Character Similarity

1 Introduction

The Shakespearean dramas are the most outstanding examples of theatrical pieces. Most of the plays show clear antagonisms and the resolutions are very rational which means the inter-character relations are consistent within the plays. This is so called the ”unity-of-action” principle. Any tragic play must rely upon its own movement. There was no more need for a Deus-ex-Machina for the stage of Shakespeare, though there are exceptions. The term is an invention of Greeks, indicating a weakness according to Aristotle [2]. At the end of a play, a distinct character helps the resolution which violates within the play consistency, or unity-of-action. According to Aristotle, the solutions of plots should come about as a result of the plot. Most notable examples follow such an idea. Forming a rational play is essential to all modern playwriting, therefore it is important to measure within play consistency. A second important issue is to detect character similarities among the dramas, if we can match characters between the plays. In this study, we would like to answer such questions by using contemporary machine learning algorithms.

The above analyses is directly related to the recent state of the field. The literary criticism recently meets computerized analysis, [13]. The main inspiration of our analysis is based on the previous works by literary critics, [15]. The current and widely accepted machine learning algorithms frequently used in the purpose of verifying literary discussion. The other reason behind our interest is mainly due to the technical advancements. Advance chatbots, conversational AI helps to generate realistic speeches [8]. With the developing machine learning techniques, nowadays it seems very possible to generate realistic dialogues for drama, or in other words artificial literature starts to seem possible [11, 25]. There needs quality measures, or evaluation metrics for such texts either of hand-made or computer generated.

In this paper, we would firstly like to overview our mathematical approach to have a document vector and then show off the experiments we carried out on Shakespearean characters as representing each character in a n-dimensional space. The last section is left to discussions on the characters and the relationship between the plays.

1.1 Related Works

The field of digital humanities (DH) mostly with the quantitative analyses of literary and cultural studies [3, 4]. There is specifically a sub-fields of DH, the so called "Drametrics", that deals with the quantitative analysis of the literary genre of drama [16]. Digital Shakespeare projects have gotten attention since the 2000s [6, 14]. The dramatic structures in the form of antagonisms are revealed by topic modeling algorithms, [25]. Machine learning based text analyses are also carried out for genre classifications [25, 1, 7, 18, 22, 26]. In literature, structural elements such as *dramatis persone* are also analyzed and applications are developed for further analyses [5, 9, 19, 21, 23, 24].

2 Methodology

The method proposed is an unsupervised neural learning algorithm. By this way, each document can be represented by a document vector. The fixed length representation of each document helps to find semantic relations between documents. Similar documents are represented in a similar location in latent space. The cosine similarity between vectors is used as a measure of similarity.

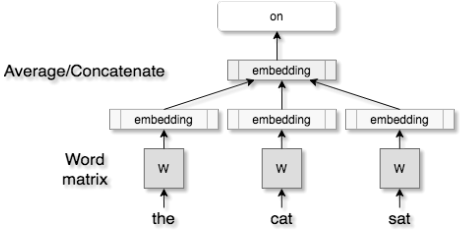

Fig. 1. The framework for learning paragraph vector. Context of the vector ”the”, ”cat” and ”sat” predicts the next word ”on”.

2.1 Document Vector

The document vector extraction is an unsupervised neural operation that recursively predicts the next word [10]. The idea is very similar to language modeling. By using the context of all the previous input tokens, the next word is predicted and errors are minimized by using back-propagation [17]. The framework is represented as Figure 1.

For example, the context of given three words, ”the”, ”cat” and ”sat”, the next word ”on” is predicted. In this framework, every document is mapped to a unique vector, represented by a column in matrix D and every word is also mapped to a unique vector, represented by a column in matrix W. In the experiments, we use concatenation as the method to combine the vectors.

As an unsupervised process, there needs texts to train such algorithms. After training, at inference stage, the input texts can be mapped to a N-dimensional latent space. Semantically similar documents would have similar features, such as orientation or location. Consistency is important between training and inference texts, when constructing such latent space.

2.2 Similarity Measure

The cosine similarity of any two N dimensional vector x and y computes L2-normalized dot product,

$$k(x, y) = \frac{x^T y}{\|x\| \, \|y\|} \tag{1}$$

The Euclidean (L2) normalization projects the vectors onto the unit sphere, the angle between normalized vectors is the similarity measure $k(x, y)$.

3 Experiments

Training dataset is the collection of Shakespeare dramas. Lines of each play treated as a document and trained for 40 epochs. A vocabulary is created out of all the plays. The number of documents is 37, it is low, the texts are long on average, around 20K. Vector length is 50. We use Gensim-Doc2Vec package.

At the inference stage, lines uttered by each character treated as a document and have a fixed length vector representation in pretrained semantic space of Shakespeare.

4 Discussion

In this section, the characters that have 40 lines or above are chosen for consistency analyses. The first subsection is on detection of Deus-ex-Machina characters in each play. The distinct characters are identified in order to analyze the strength of a play, the unity-of-action principle. The second subsection left to similar characters between the plays.

4.1 Unity-of-Action

Shakespearean comedies as well as tragedies are so strong in terms of dramatic structures. All can be thought of as Aristotelian. According to Aristotle, a play consists of complications, crisis and resolution. In the beginning of a play, the problems occur on protagonists. The crisis is the peak point which all problems need for a solution. According to Aristotle, the resolution should emerge from the plot itself. If the resolution comes as an outside force, for instance a Greek god appears at the end and kills the antagonists and saves the protagonists, this is so called Deus-ex-Machina. This is a weakness according to Aristotle. The play should be complete among the chain of events.

By thinking unity-of-action principle, we can assume that all characters in a play should be related, their speeches should be coherent. In the previous section, we construct a semantic space in an unsupervised manner by using Shakespeare's all plays. In this semantic space, similar characters have similar orientations. The position of the vector of a character has a semantic meaning. In this semantic space, all characters of a play lie inside a cluster of a play, while the Deus-ex-Machina characters would be distinct. Since Shakespeare's plays are in dialog form, as the characters talk to each other, the semantic cluster of a play forms. Following the conjuncture idea on literature by Moretti, [12], similarly, this reasoning leads to Deus-ex-Machina conjuncture to be tested,

Conjecture. *The semantics of lines uttered by characters are coherent within the play, except Deus-ex-Machina, if it obeys the unity-of-action principle.*

In order to verify above conjecture, the cosine difference between all characters is measured in pairs. By this way, similarities of all characters could be graphed in Figure 2 as a lower triangle. Each row and each column corresponds to a character, x and y in similarity function $k(x, y)$, respectively. The document vector projections would be maximized, if there are semantic similarities between characters. Then, strongly similar characters, the characters that have similarity close to 1, are represented with reddish colors, while dissimilar characters are blueish.

In Figure 2, there graphed six plays by Shakespeare. Bottom row is Macbeth, Othello and The Tempest, respectively. These are very good examples of unity-of-action principle. Three plays by Shakespeare demonstrate strong semantic similarity between characters. Each character's dialogues are at least 0.5 related to the others. Some of the characters are apparently more similar, for example, two main protagonists Macbeth and Lady Macbeth, in addition to Lady Macbeth and Banquo. Similarly, Othello and Iago are also intensely related, as well as Iago and Casio. These character pairs have importance for the play from a dramatic perspective. Strong semantic similarity is a good indicator.

The top row is Measure for Measure, Merchant of Venice and Richard III. These plays have a distinct character, who has nearly no semantic similarity with the rest of the characters. Mariana in Measure for Measure, Duke in Merchant of Venice and Richmond in Richard III demonstrates a difference from the rest. In Richard III, Richmond clearly presents himself as a Deus-ex-Machina, as drop to resolve the play. Mariana in Measure for Measure also has a similar function in the play. The Duke in Merchant of Venice is literally a Deus-ex-Machina [20]. Algorithms identify them successfully. However, the actual Deus-ex-Machina that turns tragedy to comedy is the Duke in Measure for Measure. Algorithms could not be identified. It is due to the length of the speeches by the Duke, 1/3th of the play is his lines. Changing the play from tragedy to comedy

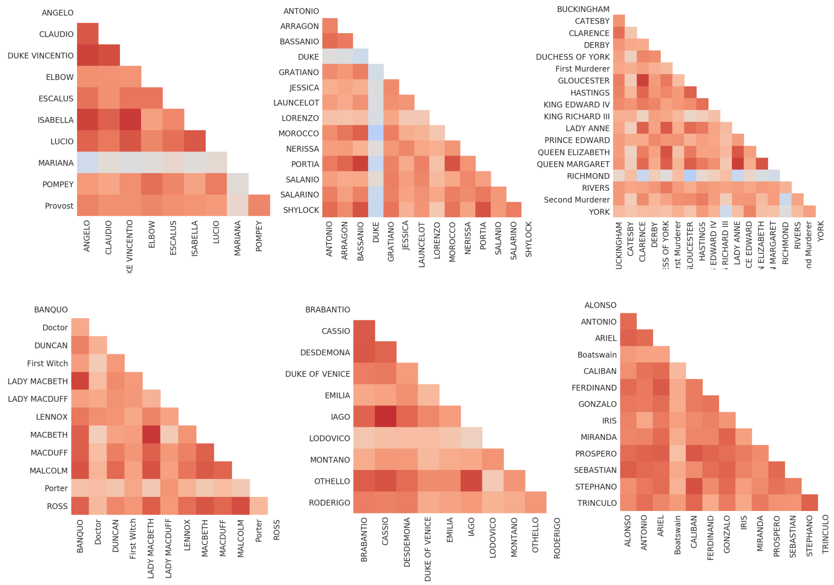

Fig. 2. Within play consistencies are graphed. The first row is the examples of Deus-ex-Machina characters. A single character is different than the rest of characters. The second row shows great plays by Shakespare. Within play consistency is very high, therefore the dramatic structures are more powerful.

is not his only function in the play. Thus far, we cover a successful identification of a character who breaks the unity-of-action principle and a failure case of the algorithm.

4.2 Character Similarity between Tragedies

In order to identify inter-play relations of characters, we measured and graphed semantic similarities among four plays, Macbeth, Richard III, Merchant of Venice, Othello, respectively. It is very apparent that within-play consistency is much denser than inter-play consistency. Each reddish triangle is a play, while the blue rectangles are inter-play relationships.

If we examine the dominant colors of rectangles that show inter-play relations, Macbeth and Merchant of Venice are least related plays, the dominant color is blue. On the other hand, Merchant of Venice and Othello, as having topics on revenge, have warmer colors. Besides, Macbeth and Richard III also have warmer colors, as their topics are mostly on taking the throne. The similarities between the plays can be observed from the dominant colors.

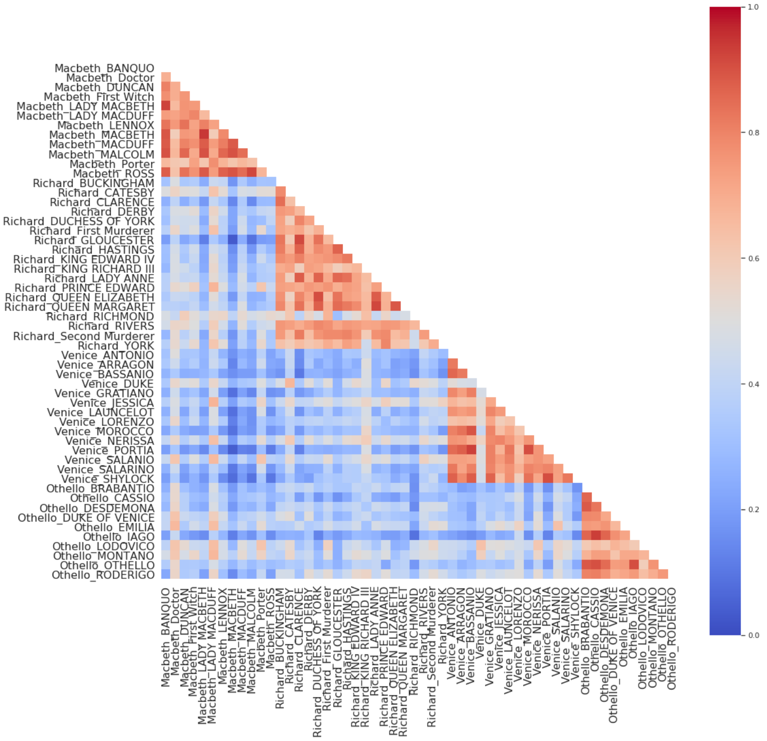

Fig. 3. Inter-play similarities are graphed. Each reddish triangle is a unique play, while blueish rectangles are inter-play similarities: Macbeth, Richard III, Merchant of Venice, Othello, respectively.

There are other observations on Figure 3. Some of the characters are almost similar to each character in different plays. Lady Macduff of Macbeth, Iago of Othello are examples of such characters, this is probably due to their compatible nature. In addition to that, Macbeth is dissimilar to all the characters in the rest of the other plays. Another interesting observation is the Duke in Merchant of Venice. Although the Duke is dissimilar to everyone within the play, it has similarities with other characters in other plays.

All in all, inter-play relations are demonstrated to a certain extent. It is very crucial to detect similar characters between dramas, for the purpose of play writing. Characters like Iago are compatible with everyone, while Duke-like characters can be found in other texts. Character similarities are observable by the analysis we proposed.

5 Conclusion

Thus far, we cover the analyses on the characters in each play and between the plays. These two analyses can be useful from the perspective of play writing as well as artificial literature. It is kind of hard to have an evaluation metric for any kind of generative model, however literary criticism helps to identify basic characteristics of a play. Drama has a well-defined form that shaped the beginning with Aristotle. In a previous work, we had identified a way of showing antagonisms, [25]. Our methodology in this work helps to find detecting the characters that break the unity-of-action principle. Inter-play similarities also gives much insight into the characters.

Richmond in Richard III and The Duke in Merchant of Venice are given as a successful identification of distinct characters. Algorithms almost always successfully detect these distinct characters, though exceptions. The weakness of such a method is based on the assumption of the characters. Instead of judging a character based on the lines, we treat the characters as a whole. All lines by a character are token and represented in a latent space. This holistic view of a character fails when dealing with multi-function characters like the Duke in Measure and Measure. In addition to these detection, the strongly similar characters are another observation. The most important characters of a play always show a strong similarity. We also observed that some of the characters are compatible with every other character. Lady Macduff and Iago are examples. Other types of characters are dissimilar to every other character in other plays, such as Macbeth. These observations are very important for the insight into the plays.

It is important to develop further evaluation metrics for drama. The generative models have a promising future in terms of dialogs, [8], and Shakespare wrote only dialog form plays. The writing of a play is possible with the knowledge of authorship. These metrics can be thought of as knowledge of the authorship of computers. Aristotle as the first critic, gives much insight into the authorship and states that the principle of unity-of-action is the most important feature of a play.

Acknowledgments This work was supported by grant 12B03P4 of Boğaziçi University. The author would like to thank Muhittin Mungan for suggesting this Master of Science topic as his advisor and Meltem Gürle Mungan for her kind opinion. The author would also like to thank Prof. Berrin Yanıkoğlu for reviewing article and actor Güneş Yakın for talks together on Shakespeare.

References

1. Ardanuy, M. C., & Sporleder, C. Structure-based clustering of novels. In: Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), pp. 31-39. Gothenburg, Sweden. (2014, April).
2. Aristotle. Aristotle's Poetics. Hill and Wang:New York (1961)
3. Clement, T., Steger, S., Unsworth, J. and Uszkalo, K. *How not to read a million books.* (2008). Available online at http://people.brandeis.edu/ unsworth/hownot2read.html
4. Crane, G. *What do you do with a million books?* D-Lib Magazine. (2006). Available online at http://www.dlib.org/dlib/march06/crane/03crane.html
5. Dennerlein, K. Measuring the average population densities of plays. A case study of Andreas Gryphius, Christian Weise and Gotthold Ephraim Lessing. Semicerchio. Rivista di poesia comparata LIII: 80–88. (2015).
6. Hirsch, B., & Craig, H. "Mingled Yarn": The State of Computing in Shakespeare 2.0. In T. Bishop, & A. Huang (Eds.), The Shakespearean International Yearbook (Vol. 14: Special Section, Digital Shakespeares, pp. 3-35). Ashgate Publishing Limited, United Kingdom. (2014).
7. Hope, J., & Witmore, M. The Hundredth Psalm to the Tune of "Green Sleeves": Digital Approaches to Shakespeare's Language of Genre. Shakespeare Quarterly, **61**(3), 357-390. (2010).
8. Jianfeng Gao, Michel Galley: "Neural Approaches to Conversational AI", 2018; arXiv:1809.08267.
9. Krautter, B. Quantitative microanalysis? Different methods of digital drama analysis in comparison. In: Book of Abstracts, DH 2018. Mexico-City, Mexico, pp. 225-228. (2018).
10. Le, Quoc; Mikolov, Tomas. Distributed representations of sentences and documents. In: Xing, Eric P. and Jebara, Tony (eds.) Proceedings of the 31st International Conference on International Conference on Machine Learning, Vol. 32, p. 1188-1196. (2014)
11. Lebrun T.: Who Is the Artificial Author?. In: Mouhoub M., Langlais P. (eds) Advances in Artificial Intelligence. Canadian AI 2017. Lecture Notes in Computer Science, vol 10233. Springer, Cham (2017)
12. Moretti, F. Conjectures on world literature. New left review, 54-68 (2000)
13. Moretti, F.: Distant reading. 1st edn. Verso Books, London (2013)
14. Mueller, M. Digital Shakespeare, or towards a literary informatics. Shakespeare **4**, 284-301 (2008)
15. Ramsay, S. Reading Machines: Toward an Algorithmic Criticism. 1st edn. University of Illinois Press, Champaign IL (2011)
16. Romanska, M. Drametrics: what dramaturgs should learn from mathematicians. In Romanska, M. (ed.), The Routledge Companion to Dramaturgy, pp. 472-481. Routledge, New York (2015).

17. Rumelhart, D., Hinton, G. Williams, R. Learning representations by back-propagating errors. Nature 323, 533–536 (1986) doi:10.1038/323533a0
18. Schöch, Christof. (2016). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. Digital Humanities Quarterly, (2016).
19. Schmidt, T., Burghardt, M., Dennerlein, K. & Wolff, C. Katharsis – A Tool for Computational Drametrics. In: Book of Abstracts, DH 2019. (2019).
20. Sennet, R. The Foreigner: Two Essays on Exile. Notting Hill Editions Ltd, London. (2011).
21. Trilcke, P., Fischer, F. and Kampkaspar, D. Digital Network Analysis of Dramatic Texts. In: Book of Abstracts, DH 2015. Sidney, Australia. (2015).
22. Underwood, T., Black, M.L., Auvil, L., & Capitanu, B. Mapping mutable genres in structurally complex volumes. IEEE International Conference on Big Data, 95-103. (2013)
23. Wilhelm, T., Burghardt, M., and Wolff, C. "To See or Not to See" - An Interactive Tool for the Visualization and Analysis of Shakespeare Plays. In R. Franken-Wendelstorf, E. Lindinger, and J. Sieck (Eds.), Kultur und Informatik: Visual Worlds & Interactive Spaces. Glückstadt: Verlag Werner Hülsbusch, pp. 175–185. (2013).
24. Xanthos, A., Pante, I., Rochat, Y and Grandjean, M. Visualising the dynamics of character networks. In: Book of Abstracts, DH 2016. Kraków, Poland, pp. 417-419. (2016).
25. Yavuz, M. C.: Analyses of Literary Texts by Using Statistical Inference Methods. In: Proceedings of the Sixth Italian Conference on Computational Linguistics, CLiC-it'19. Bari, Italy. (2019, November), CEUR-WS.org, online http://ceur-ws.org/Vol-2481/paper75.pdf.
26. Yu, B. An evaluation of text classification methods for literary study. Literary and Linguistic Computing **23**(3): 327-343. (2008).