Text Zoning of Theater Reviews: How Different are Journalistic from Blogger Reviews?

Mylène Maignant, Gaëtan Brison, Thierry Poibeau

Text Zoning of Theater Reviews: How Different are Journalistic from Blogger Reviews?

Mylène Maignant[1], Gaëtan Brison[2], Thierry Poibeau[1]

[1]Laboratoire LATTICE (CNRS & ENS-PSL & Université Sorbonne nouvelle)
1, rue Maurice Arnoux, 92120 Montrouge, France
firstname.lastname@ens.psl.eu

[2]Institut Polytechnique de Paris, 5 Av. Le Chatelier, 91764 Palaiseau
gaetan.brison@ip-paris.fr

Abstract

This paper aims at modeling the structure of theater reviews on contemporary London performances by using text zoning. Text zoning consists in tagging sentences so as to reveal text structure. More than 40 000 theater reviews going from 2010 to 2020 were collected to analyze two different types of reception (journalistic vs digital). We present our annotation scheme and the classifiers used to perform the text zoning task, aiming at tagging reviews at the sentence level. We obtain the best results using the random forest algorithm, and show that this approach makes it possible to give a first insight of the similarities and differences between our two subcorpora.

1 Introduction

Since 2010 in England, a wave of blogs written by authors coming from various horizons has arisen on the Internet. Students, theater professionals but also mere amateurs began publishing their own theater reviews. These new independent voices in the digital space progressively redefine the shape of classic journalistic criticism. Although discreet, they offer a new vision of the history of Londonian theaters. By doing so, it sets itself apart from the canon of mainstream journalism.

The emergence of this digital culture triggered a lot of controversies on the status of the review as a literary object. Michael Billington, reviewer for *The Guardian* since 1971, states that a blog 'is more like an informal letter: a review, if it's to have any impact, has to have a definable structure.' For Danielle Tarento, co-founder of the Menier Chocolate Factory, 'a lot of people [bloggers] are not 'proper writers'.' At the other hand of the spectrum, some of these bloggers claim the stylistic singularity of their publications. In the description of Exeuntmagazine.com for instance, the editors claim that: 'Exeunt believes in making beautifully written, *experimental*, fierce and *longform* writing about theatre available for free.'

A review is traditionally organized according to several sections: an introduction, a presentation of the plot, a few lines on the stage, etc. In order to compare the two subcorpora, it is first necessary to segment the reviews into textual zones corresponding to these thematic sections. We assume the two subcorpora will share the same zones, as they are all about theater, but the content of the zones may differ from one subcorpus to the other: e.g., the two communities may not focus on the same aspects of the plays. From a technical point of view, this experiment is also an opportunity to test the robustness and relevance of text zoning across different domains. Text zoning has been mainly used to segment scientific texts so far, but can this technique also be used in the humanities? Can it be used for performance reviews, where critics do not follow a fixed structure, contrary to scientific writing?

This short paper is structured as follows. We first give a brief overview of text zoning. We then present our corpus, the different features and machine learning techniques used for the task. We then comment our results and give some hints on the way these could be used to get a better understanding of the content of the corpus and the differences between the two communities at stake (official critics vs amateur bloggers).

2 Previous Work

The notion of text zoning was first introduced by Simone Teufel in her PhD (Teufel, 1999). Teufel was targeting the automatic analysis of scientific papers. In this context, argumentative zoning refers to the 'rhetorical status of a sentence with respect to the communicative function of the whole paper.' It is for example quite useful to distinguish 'back-

ground information' from 'statements of the particular aim of the current paper', to take an example from Teufel's work (Teufel and Moens, 2002). While text zoning has been mainly applied to scientific texts so far, one can also find this technique applied to other domains where it is relevant, for example email messages (Lampert et al., 2009), or job ads (Gnehm and Clematide, 2020).

The number of zones considered varies but is generally around ten or less (7 for example in (Teufel, 1999) and (Guo et al., 2011), or 8 in (Gnehm, 2018)). Zone annotation is generally performed by a group of experts at the sentence level (more rarely at the paragraph level). Inter-annotator agreement on the task is generally high: (Guo et al., 2011) for example reports a score of 0.85 for Cohen's κ (Cohen, 1960).

Once a representative corpus is available, it is possible to train a classifier for the task. Features considered are generally low level (unigrams, bigrams, sometimes specific terms also receive a specific weight) (Teufel and Moens, 2002) but higher level features are also sometimes considered (like syntactic relations in (Guo et al., 2011)). Contextual information (like the previous zone) is also often taken into consideration, since a specific zone tends to appear in typical positions in scientific abstracts. As for training, most recent ML techniques have been explored, from Naive Bayes (Teufel, 1999) to LSTM (Gnehm, 2018), through CRF and SVM (Guo et al., 2011). In this last paper, the authors also investigate semi-supervised learning and active learning, in order to reduce the amount of data needed for training, which often constitutes a bottleneck for the task. More recently, large language models like Bert have also been explored (Gnehm and Clematide, 2020), but they require large amount of data for training.

Here our goal is partly the same as the one in these previous studies. However, our corpus is very different since we analyze theater reviews, which may not be as regular as scientific papers. In our context, zones are important to determine whether the critic is addressing acting, staging or the general setting of the play (we use the rather neutral term 'text zoning' instead of 'argument zoning', since the zones we consider do not always correspond to arguments). Analyzing the overall organization of theater reviews will also make it possible to determine whether these have a rather fixed structure or not, if reviews in newspapers differ a lot from those directly written for blogs on the Web.

3 Corpus Creation

To answer these questions, the first step consisted in collecting the necessary data to create two subcorpora. The first subcorpus is made of journalistic reviews only, while the second one is based on digital theater reviews written by bloggers.

3.1 Subcorpus 1: Journalistic Theater Reviews

The first subcorpus was created thanks to the online database *Theatre Record*. *Theatre Record* is a biweekly paper magazine which reprints in full all the national drama reviews of the productions in London and its regions. Its archives were digitized in 2019 and each newspaper published since 1981 is now available online (in PDF format).

All the newspapers issues have the same characteristics. For each of the shows, a certain number of reviews is given as well as a series of details on the production, such as the cast, the credits and the photographs. The theater in which the play was performed as well as the opening and the closing dates of the show are also indicated. Most of the newspapers represented in this database are well-known among the general public: *The Times, The Guardian, The Independent*, etc. Out of the 84 newspapers available on *Theatre Record*, we have selected 32 of them in total. A number of sources had to be removed. Since this corpus focuses on printed newspapers, online news websites had to be excluded. We also removed newspapers whose reviews were not about London performances and all the newspapers which had a too limited number of reviews.

3.2 Subcorpus 2: Digital Theater Reviews

The second subcorpus is based on 18 English blog platforms whose authors' publications deal with London plays only. The content of these websites was extracted using webscraping techniques. These 18 blog platforms have the following characteristics: they have no printed equivalent, their content is entirely free and their authors are not paid for their activity. They are either run by one person, or by multiple authors.

The selection of these blogs was made according to the top 10 most popular British theater blogs established by Vuelio in 2020. A majority of them also came from the platform *MyTheatreMates*. All

the authors who have their reviews published on *MyTheatreMates* share the following characteristics: They have their own personal website, they post original theatre-related content on their personal website at least once a fortnight, they can provide three professional arts references (e.g. artists they have interviewed or, if they review, producers or publicists who already regularly provide them with complimentary press tickets to shows) and they are active on Twitter.

When this subcorpus was created (September 2020 – April 2021), 52 bloggers were members of *MyTheatreMates*. We selected the blogs which had the highest number of reviews (at least 200 reviews) as well as the ones which were mainly focusing on the Londonian stage.

3.3 Overview of the Corpus

	Newspapers	Blogs
Number of sources	33	18
Number of words	8,831,160	10,364,855
Number of reviews	22781	19045

Table 1: The Descriptive statistics of each corpus (source refers to newspapers vs blog platforms).

Table 1 gives an overview of the two datasets. The corpus is available in textual format (PDFs from *Theatre Record* have been converted and manually corrected) so that NLP tools coming from Stanford could be directly applied. It is to our knowledge the first corpus collecting so many reviews of theater performances. The corpus is freely available online, on the website dedicated to this project: `Dramacritiques.com`.

4 Experiment Description

4.1 Annotation Scheme and Data Labeling

Once the data were collected, the first step consisted in labeling a random sample of reviews that could be used for training. The annotation scheme corresponds to the 8 different possible sections of a review.

The definition of these sections is based on (Fisher, 2015). In his analysis, Fisher examines the various possibilities for one critic to structure his arguments, which leads to the following 8 different categories with 8 different colours:

For this first experiment, the data were labeled by an expert with a strong background in theatre studies. This expert spent more than 15 minutes per review, or 250 hours in total, annotating the sample. Each of the sentences was carefully analyzed to propose the best category it belonged to.

Zone category	Associated colour
Introduction	Purple
Reviewer analysis	Blue
Visual and audio details	Green
Conclusion	Yellow
Performance of actors	Orange
Plot	Red
Structure of the play	Brown
Related to the audience	Grey

Table 2: Delimitation of the different zones and their colours used in the model.

However, some of the sentences could have been classified in two different categories. These cases were recorded and resolved following explicit rules to ensure the consistency of the annotation. 1000 reviews were manually annotated, which was deemed enough for training.

4.2 Data Preparation

Several preprocessing steps were applied to the corpus, following previous experiments in text zoning. Texts were first segmented into sentences, tokenized and tagged (with POS and morphological features) and empty words were removed. Named Entity Recognition and Term Frequency-Inverse Document Frequency (TF-IDF) were also applied on the corpus. Annotations were performed using Stanford tools and were then used as features for training.

In the end, more than eighty variables were created, following previous work in the domain (among others (Teufel, 1999) and (Guo et al., 2011)):

- Statistical variables: average word length, average sentence length, frequency of personal pronouns, etc.
- Tense variables: proportion of verbs in future, present and past tenses
- Grammar variables: top verbs, adjectives, superlatives, nouns, etc.
- Parts of Speech variables: position of the words and their roles in the sentences

- Named Entity Recognition variables: organization, characters, etc.

If the creation of these different types of variables helped to improve the prediction of the algorithms, only a few of them were relevant for the models. We thus applied a feature selection process to reduce the number of variables used during training (we went from more than 100 different features to a little less than 20 main features). We used a correlation matrix, Principal Component Analysis (PCA) to reduce the number of variables we originally had and other tools in function of the models.

4.3 Models for Sentence Classification

We selected 4 traditional classification models that seemed relevant for the task (Naive Bayes, random forest, KNN and RNN). Most of them have already been used for text zoning (see the previous work section), but their relevance in the context of reviews remains to be assessed.

- Naive Bayes is a simple Bayesian model. It is known to perform well on small datasets and will thus constitute a baseline.

- Random Forest (Ho, 1995) is an ensemble learning method that builds a multitude of decision trees at training time. Random forest generally performs better than a single decision tree and can take into account the multiple parameters of our problem.

- K-Nearest Neighbors (KNNs) (Altman, 1992) assumes that all data points in close proximity is labeled with the same class. KNNs may thus not work so well on heterogeneous and diversified data. For this model we tested a wide range of k ranging from 1 to 15 to find the optimal one.

- Recurrent Neural Network (RNN) (Sperduti and Starita, 1997) is relevant to find hidden dependencies patterns in the data. This model is the most powerful one in theory but generally requires more data for training.

We chose to limit ourselves to these well-known classification techniques. More recent approaches exist, for example based on deep learning techniques and using large language models like Bert (Devlin et al., 2019). As we wanted the approach to be portable and easily reproducible by people working in humanities, we excluded these more resource intensive approaches but that is something we should try in the future (see (Gnehm and Clematide, 2020) for an experiment with biLSTM and BERT).

5 Results

We applied each model on the data and computed their accuracy (computed using 10-fold cross validation and averaging the results across folds). Our results are reported in Table 3:

Models	Accuracy
Naive Bayes	.69
Random Forest	.80
K-Nearest Neighbors	.72
Recurrent Neural Networks	.61

Table 3: Performance of the different models.

According to Table 3, the top performing model in terms of accuracy is random forest. The result is comparable to previous studies (for example, (Guo et al., 2011) report .81 overall accuracy).

Note however that we had to find the optimal parameter for the depth of the tree and the number of trees. By doing so the random forest model uses a system of threshold for each important feature (Breiman, 2001). To improve the model and avoid overfitting, we also used cross validation during the training and test steps.

As planned, Naive Bayes is not able to take into account the complexity of the task and performs poorly. KNN also fails at capturing the variations of the different zones, as the texts to classify are quite short. Lastly, RNN performs worse as there are not enough data to train this model properly. For this part in particular, we used Long Short Term Memory Neural Networks which work sequence by sequence. We wanted to use pre-trained models but none of them had already been trained on a similar dataset for this task. The closest we could get was on the IMDB dataset. However, if it were annotated for sentiment analysis, it were not for zoning.

Figure 1 represents a review with the different zones identified, each color corresponding to a specific zone. This figure illustrates how the algorithm works. The model looks for each sentence and calculates its probability of being part of one of the 8 predefined categories. Of course it is possible that one sentence may have different recognizable pat-

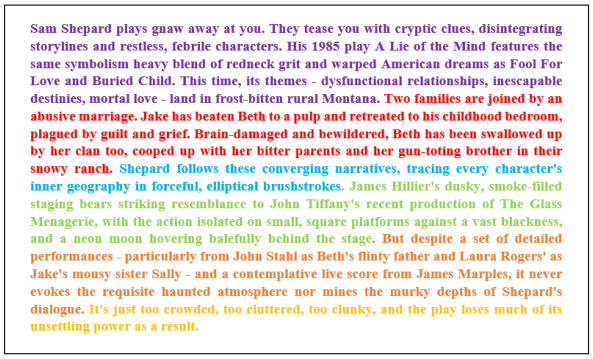

Figure 1: An Example of Annotated Text (each zone is annotated with a specific color). Color code: Purple: Introduction Red: Plot Blue: General Analysis of the Play Green: Visual, Auditory and Audible Details Orange: Actors' Performances Brown: Remarks on the Structure of the Play Yellow: Conclusion

terns that makes it belong to several classes. In this case, the model associates to the sentence the category which has the highest percentage. In function of the class assigned by the model, the sentence will then take the color related to its category.

6 Discussion

Although the accuracy of the algorithm could be improved, these first results are a reliable and relevant base to better understand the comparison between printed and digital theater criticism. If the debate in the artistic sphere highlights the differences between journalists and bloggers, the experiments actually prove that their reviews are more similar than what they claim. Each of the 8 categories we had defined are represented in the two datasets (see Table 4) which suggests that both of them employ similar lines of thought.

Zone category	Newspapers	Blogs
Introduction	15.9	17.0
Reviewer analysis	13.3	11.7
Visual and audio details	4.4	7.3
Conclusion	9.0	8.3
Performance of actors	15.2	18.4
Plot	32.5	28
Structure of the play	8.6	7.2
Related to the audience	0.9	2.1

Table 4: Relative coverage of each predicted zone.

The real differences are located at a subtler level. When we have a closer look at the percentages within each dataset and when we compare them, we can realize that bloggers tend to focus on categories related to affect. 'Visual and audio details', 'Performance of actors' and remarks 'Related to the audience' are all aspects of the review which put to the front the subjective perception of the critic. On the contrary, superior values in percentages for the subcorpus I are situated in categories linked to more factual arguments. 'Reviewer analysis', 'Plot' and 'Structure of the play' rather rely on descriptive and rational materials. This paves the way for further analysis, mixing text zoning and sentiment analysis for example, so as to get a better understanding of the content of the different zones and of the differences between the two corpora under study.

7 Conclusions and Perspectives

We have presented a study based on the automatic analysis of more than 40 000 theater reviews on the contemporary Londonian stage. We have shown that it is possible to segment these reviews into labeled text zones with a good accuracy. In the future, we want to investigate large language models and their potential benefit for the task.

Considering the classification obtained with text zoning, it seems that the two subcorpora considered in the study (journalists vs bloggers) are not as different as some actors of the domain may have claimed, at least from a distant reading perspective. However the content of some zones seems to be really different from one subcorpus to the other: the zoning experiment presented in this paper is thus a first necessary step in order to be able to perform a more precise analysis.

Acknowledgements

Mylène Maignant is partially supported by the EUR (École Universitaire de Recherche) Translitteræ (programme "Investissements d'avenir" ANR-10-IDEX-0001-02 PSL* and ANR-17-EURE-0025). Thierry Poibeau is supported in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute). Lastly, this research also benefited from the support of the CNRS International Research Network Cyclades (Corpora and Computational Linguistics for Digital Humanities).

References

Naomi Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46:175–185.

Leo Breiman. 2001. Random forests. *Statistics Department University of California Berkeley*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Fisher. 2015. *How to Write About Theatre*. Methuen Drama, London.

Ann-Sophie Gnehm. 2018. *Text Zoning for Job Advertisements with Bidirectional LSTMs*. University of Zurich.

Ann-Sophie Gnehm and Simon Clematide. 2020. Text zoning and classification for job advertisements in German, French and English. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 83–93, Online. Association for Computational Linguistics.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A Weakly-supervised Approach to Argumentative Zoning of Scientific Documents. In *Empirical Methods in Natural language Processing (EMNLP)*, Edinburgh, United Kingdom.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1.

Andrew Lampert, Robert Dale, and Cécile Paris. 2009. Segmenting email message text into zones. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 919–928, Singapore. Association for Computational Linguistics.

Alessandro Sperduti and Antonina Starita. 1997. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, 8/3:714—-735.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. University of Edinburgh.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computional Linguistics*, 28(4):409–445.