Clustering voices in The Waste Land

Julian Brooke Dept of Computer Science University of Toronto jbrooke@cs.toronto.edu Graeme Hirst Dept of Computer Science University of Toronto gh@cs.toronto.edu

Adam Hammond

Dept of English University of Toronto adam.hammond@utoronto.ca

Abstract

T.S. Eliot's modernist poem *The Waste Land* is often interpreted as collection of voices which appear multiple times throughout the text. Here, we investigate whether we can automatically cluster existing segmentations of the text into coherent, expert-identified characters. We show that clustering *The Waste Land* is a fairly difficult task, though we can do much better than random baselines, particularly if we begin with a good initial segmentation.

1 Introduction

Although literary texts are typically written by a single author, the style of a work of literature is not necessarily uniform. When a certain character speaks, for instance, an author may shift styles to give the character a distinct voice. Typically, voice switches in literature are explicitly marked, either by the use of quotation marks with or without a *said* quotative, or, in cases of narrator switches, by a major textual boundary (e.g. the novel *Ulysses* by James Joyce). However, implicit marking is the norm in some modernist literature: a well-known example is the poem *The Waste Land* by T.S. Eliot, which is usually analyzed in terms of voices that each appear multiple times throughout the text. Our interest is distinguishing these voices automatically.

One of the poem's most distinctive voices is that of the woman who speaks at the end of its second section:

I can't help it, she said, pulling a long face, It's them pills I took, to bring it off, she said [158–159] Her chatty tone and colloquial grammar and lexis distinguish her voice from many others in the poem, such as the formal and traditionally poetic voice of a narrator that recurs many times in the poem:

Above the antique mantel was displayed As though a window gave upon the sylvan scene The change of Philomel [97–99]

Although the stylistic contrasts between these and other voices are clear to many readers, Eliot does not explicitly mark the transitions, nor is it obvious when a voice has reappeared. Our previous work focused on only the segmentation part of the voice identification task (Brooke et al., 2012). Here, we instead assume an initial segmentation and then try to create clusters corresponding to segments of the *The Waste Land* which are spoken by the same voice. Of particular interest is the influence of the initial segmentation on the success of this downstream task.

2 Related Work

There is a small body of work applying quantitative methods to poetry: Simonton (1990) looked at lexical and semantic diversity in Shakespearean sonnets and correlated this with aesthetic success, whereas Dugan (1973) developed statistics of formulaic style and applied them to the *Chanson de Roland* to determine whether it represents an oral or written style. Kao and Jurafsky (2012) quantify various aspects of poety, including style and sentiment, and use these features to distinguish professional and amateur writers of contemporary poetry. With respect to novels, the work of McKenna and Antonia (2001) is very relevant; they used principal components analysis of lexical frequency to discriminate different voices and narrative styles in sections of *Ulysses* by James Joyce.

Clustering techniques have been applied to literature in general; for instance, Luyckx (2006) clustered novels according to style, and recent work in distinguishing two authors of sections of the Bible (Koppel et al., 2011) relies crucially on an initial clustering which is bootstrapped into a supervised classifier which is applied to segments. Beyond literature, the tasks of stylistic inconsistency detection (Graham et al., 2005; Guthrie, 2008) and intrinsic (unsupervised) plagiarism detection (Stein et al., 2011) are very closely related to our interests here, though in such tasks usually only two authors are posited; more general kinds of authorship identification (Stamatatos, 2009) may include many more authors, though some form of supervision (i.e. training data) is usually assumed.

Our work here is built on our earlier work (Brooke et al., 2012). Our segmentation model for *The Waste Land* was based on a stylistic change curve whose values are the distance between stylistic feature vectors derived from 50 token spans on either side of each point (spaces between tokens) in the text; the local maxima of this curve represent likely voice switches. Performance on *The Waste Land* was far from perfect, but evaluation using standard text segmentation metrics (Pevzner and Hearst, 2002) indicated that it was well above various baselines.

3 Method

Our approach to voice identification in *The Waste Land* consists first of identifying the boundaries of voice spans (Brooke et al., 2012). Given a segmentation of the text, we consider each span as a data point in a clustering problem. The elements of the vector correspond to the best feature set from the segmentation task, with the rationale that features which were useful for detecting changes in style should also be useful for identifying stylistic similarities. Our features therefore include: a collection of readability metrics (including word length), frequency of punctuation, line breaks, and various parts-of-speech, lexical density, average frequency in a large

external corpus (Brants and Franz, 2006), lexiconbased sentiment metrics using SentiWordNet (Baccianella et al., 2010), formality score (Brooke et al., 2010), and, perhaps most notably, the centroid of 20dimensional distributional vectors built using latent semantic analysis (Landauer and Dumais, 1997), reflecting the use of words in a large web corpus (Burton et al., 2009); in previous work (Brooke et al., 2010), we established that such vectors contain useful stylistic information about the English lexicon (including rare words that appear only occasionally in such a corpus), and indeed LSA vectors were the single most promising feature type for segmentation. For a more detailed discussion of the feature set, see Brooke et al. (2012). All the features are normalized to a mean of zero and a standard deviation of 1.

For clustering, we use a slightly modified version of the popular k-means algorithm (MacQueen, 1967). Briefly, k-means assigns points to a cluster based on their proximity to the k cluster centroids, which are initialized to randomly chosen points from the data and then iteratively refined until convergence, which in our case was defined as a change of less than 0.0001 in the position of each centroid during one iteration.¹ Our version of k-means is distinct in two ways: first, it uses a weighted centroid where the influence of each point is based on the token length of the underlying span, i.e. short (unreliable) spans which fall into the range of some centroid will have less effect on the location of the centroid than larger spans. Second, we use a city-block (L_1) distance function rather than standard Euclidean (L_2) distance function; in the segmentation task, Brooke et al. found that city-block (L1) distance was preferred, a result which is in line with other work in stylistic inconsistency detection (Guthrie, 2008). Though it would be interesting to see if a good kcould be estimated independently, for our purposes here we set k to be the known number of speakers in our gold standard.

4 Evaluation

We evaluate our clusters by comparing them to a gold standard annotation. There are various metrics for extrinsic cluster evaluation; Amigó et al.

¹Occasionally, there was no convergence, at which point we halted the process arbitrarily after 100 iterations.

(2009) review various options and select the BCubed precision and recall metrics (Bagga and Baldwin, 1998) as having all of a set of key desirable properties. BCubed precision is a calculation of the fraction of item pairs in the same cluster which are also in the same category, whereas BCubed recall is the fraction of item pairs in the same category which are also in the same cluster. The harmonic mean of these two metrics is BCubed F-score. Typically, the 'items' are exactly what has been clustered, but this is problematic in our case, because we wish to compare methods which have different segmentations and thus the vectors that are being clustered are not directly comparable. Instead, we calculate the BCubed measures at the level of the token; that is, for the purposes of measuring performance we act as if we had clustered each token individually, instead of the spans of tokens actually used.

Our first evaluation is against a set of 20 artificially-generated 'poems' which are actually randomly generated combinations of parts of 12 poems which were chosen (by an English literature expert, one of the authors) to represent the time period and influences of The Waste Land. The longest of these poems is 1291 tokens and the shortest is just 90 tokens (though 10 of the 12 have at least 300 tokens); the average length is 501 tokens. Our method for creating these poems is similar to that of Koppel et al. (2011), though generalized for multiple authors. For each of the artificial poems, we randomly selected 6 poems from the 12 source poems, and then we concatenated 100-200 tokens (or all the remaining tokens, if less than the number selected) from each of these 6 poems to the new combined poem until all the poems were exhausted or below our minimum span length (20 tokens). This allows us to evaluate our method in ideal circumstances, i.e. when there are very distinct voices corresponding to different poets, and the voice spans tend to be fairly long.

Our gold standard annotation of *The Waste Land* speakers is far more tentative. It is based on a number of sources: our own English literature expert, relevant literary analysis (Cooper, 1987), and also *The Waste Land* app (Touch Press LLP, 2011), which includes readings of the poem by various experts, including T.S. Eliot himself. However, there is inherently a great deal of subjectivity involved in

literary annotation and, indeed, one of the potential benefits of our work is to find independent justification for a particular voice annotation. Our gold standard thus represents just one potential interpretation of the poem, rather than a true, unique gold standard. The average size of the 69 segments in the gold standard is 50 tokens; the range, however, is fairly wide: the longest is 373 tokens, while the shortest consists of a single token. Our annotation has 13 voices altogether.

We consider three segmentations: the segmentation of our gold standard (Gold), the segmentation predicted by our segmentation model (Automatic), and a segmentation which consists of equallength spans (Even), with the same number of spans as in the gold standard. The Even segmentation should be viewed as the baseline for segmentation, and the Gold segmentation an "oracle" representing an upper bound on segmentation performance. For the automatic segmentation model, we use the settings from Brooke et al. (2012). We also compare three possible clusterings for each segmentation: no clustering at all (Initial), that is, we assume that each segment is a new voice; k-means clustering (k-means), as outlined above; and random clustering (Random), in which we randomly assign each voice to a cluster. For these latter two methods, which both have a random component, we averaged our metrics over 50 runs. Random and Initial are here, of course, to provide baselines for judging the effectiveness of k-means clustering model. Finally, when using the gold standard segmentation and k-means clustering, we included another oracle option (Seeded): instead of the standard k-means method of randomly choosing them from the available datapoints, each centroid is initialized to the longest instance of a different voice, essentially seeding each cluster.

5 Results

Table 1 contains the results for our first evaluation of voice clustering, the automatically-generated poems. In all the conditions, using the gold segmentation far outstrips the other two options. The automatic segmentation is consistently better than the evenly-spaced baseline, but the performance is actually worse than expected; the segmentation metrics we used in our earlier work

Configuration	BCubed metrics		
	Prec.	Rec.	F-score
Initial Even	0.703	0.154	0.249
Initial Automatic	0.827	0.177	0.286
Initial Gold	1.000	0.319	0.465
Random Even	0.331	0.293	0.307
Random Automatic	0.352	0.311	0.327
Random Gold	0.436	0.430	0.436
k-means Even	0.462	0.409	0.430
k-means Automatic	0.532	0.479	0.499
k-means Gold	0.716	0.720	0.710
k-means Gold Seeded	0.869	0.848	0.855

Table 1: Clustering results for artificial poems

Table 2: Clustering results for The Waste Land

Configuration	BCubed metrics		
	Prec.	Rec.	F-score
Initial Even	0.792	0.069	0.128
Initial Automatic	0.798	0.084	0.152
Initial Gold	1.000	0.262	0.415
Random Even	0.243	0.146	0.183
Random Automatic	0.258	0.160	0.198
Random Gold	0.408	0.313	0.352
k-means Even	0.288	0.238	0.260
k-means Automatic	0.316	0.264	0.296
k-means Gold	0.430	0.502	0.461
k-means Gold Seeded	0.491	0.624	0.550

The results for *The Waste Land* are in Table 2. Many of the basic patterns are the same, including the consistent ranking of the methods; overall, however, the clustering is far less effective. This is particularly true for the gold-standard condition, which only increases modestly between the initial and clustered state; the marked increase in recall is balanced by a major loss of precision. In fact, unlike with the artificial text, the most promising aspect of the clustering seems to be the fairly sizable boost to the quality of clusters in automatic segmenting performance. The effect of seeding is also very consistent, nearly as effective as in the automatic case.

We also looked at the results for individual speakers in *The Waste Land*; many of the speakers (some of which appear only in a few lines) are very poorly distinguished, even with the gold-standard segmentation and seeding, but there are a few that cluster quite well; the best two are in fact our examples from Section 1,² that is, the narrator (F-score 0.869), and the chatty woman (F-score 0.605). The former result is particularly important, from the perspective of literary analysis, since there are several passages which seem to be the main narrator (and our expert annotated them as such) but which are definitely open to interpretation.

6 Conclusion

Literature, by its very nature, involves combining existing means of expression in surprising new ways, resisting supervised analysis methods that depend on assumptions of conformity. Our unsupervised approach to distinguishing voices in poetry offers this necessary flexibility, and indeed seems to work reasonably well in cases when the stylistic differences are clear. The Waste Land, however, is a very subtle text, and our results suggest that we are a long way from something that would be a considered a possible human interpretation. Nevertheless, applying quantitative methods to these kinds of texts can, for literary scholars, bridge the gab between abstract interpretations and the details of form and function (McKenna and Antonia, 2001). In our own case, this computational work is just one aspect of a larger project in literary analysis where the ultimate goal is not to mimic human behavior per se, but rather to better understand literary phenomena by annotation and modelling of these phenomena (Hammond, 2013; Hammond et al., 2013).

With respect to future enhancements, improving segmentation is obviously important; the best automated efforts so far provide only a small boost over a baseline approach to segmentation. However, independently of this, our experiments with goldstandard seeding suggest that refining our approach to clustering, e.g. a method that identifies good initial points for our centroids, may also pay dividends in the long run. A more radical idea for future work would be to remove the somewhat artificial delim-

²These passages are the original examples from our earlier work (Brooke et al., 2012), selected by our expert for their distinctness, so the fact that they turned out to be the most easily clustered is actually a result of sorts (albeit an anecdotal one), suggesting that our clustering behavior does correspond somewhat to a human judgment of distinctness.

itation of the task into segmentation and clustering phases, building a model which works iteratively to produce segments that are sensitive to points of stylistic change but that, at a higher level, also form good clusters (as measured by intrinsic measures of cluster quality).

Acknowledgements

This work was financially supported by the Natural Sciences and Engineering Research Council of Canada.

References

- Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12:461–486, August.
- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, May.
- Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING '98)*, pages 79–85, Montreal, Quebec, Canada.
- Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Corpus Version 1.1. Google Inc.
- Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, Beijing.
- Julian Brooke, Adam Hammond, and Graeme Hirst. 2012. Unsupervised stylistic segmentation of poetry with change curves and extrinsic features. In *Proceedings of the 1st Workshop on Computational Literature for Literature (CLFL '12)*, Montreal.
- Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), San Jose, CA.
- John Xiros Cooper. 1987. T.S. Eliot and the politics of voice: The argument of The Waste Land. UMI Research Press, Ann Arbor, Mich.
- Joseph J. Duggan. 1973. *The Song of Roland: Formulaic style and poetic craft*. University of California Press.

- Neil Graham, Graeme Hirst, and Bhaskara Marthi. 2005. Segmenting documents by stylistic character. *Natural Language Engineering*, 11(4):397–415.
- David Guthrie. 2008. Unsupervised Detection of Anomalous Text. Ph.D. thesis, University of Sheffield.
- Adam Hammond, Julian Brooke, and Graeme Hirst. 2013. A tale of two cultures: Bringing literary analysis and computational linguistics together. In *Proceedings of the 2nd Workshop on Computational Literature for Literature (CLFL '13)*, Atlanta.
- Adam Hammond. 2013. He do the police in different voices: Looking for voices in *The Waste Land*. Seminar: "Mapping the Fictional Voice" American Comparative Literature Association (ACLA).
- Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, sentiment, and imagery in contemporary poetry. In *Proceedings of the 1st Workshop on Computational Literature for Literature (CLFL '12)*, Montreal.
- Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon.
- Thomas K. Landauer and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.
- Kim Luyckx, Walter Daelemans, and Edward Vanhoutte. 2006. Stylogenetics: Clustering-based stylistic analysis of literary corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, Genoa, Italy.
- J. B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceed*ings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pages 281–297.
- C. W. F. McKenna and A. Antonia. 2001. The statistical analysis of style: Reflections on form, meaning, and ideology in the 'Nausicaa' episode of *Ulysses*. *Literary and Linguistic Computing*, 16(4):353–373.
- Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28:19–36, March.
- Dean Keith Simonton. 1990. Lexical choices and aesthetic success: A computer content analysis of 154 Shakespeare sonnets. *Computers and the Humanities*, 24(4):251–264.
- Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

- Benno Stein, Nedim Lipka, and Peter Prettenhofer. 2011. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, 45(1):63–82.
- Touch Press LLP. 2011. *The Waste Land* app. http://itunes.apple.com/ca/app/the-waste-land/ id427434046?mt=8.