Julian Brooke* Graeme Hirst* Adam Hammond†

jbrooke@cs.toronto.edu, gh@cs.toronto.edu, adam.hammond@utoronto.ca *Department of Computer Science †Department of English University of Toronto

Clustering Voices in *The Waste Land*

1. Introduction

- *The Waste Land*, by T.S. Eliot (1922)
  - Long-form modernist poetry
  - Voices of differing styles throughout test, not explicitly marked
- Examples
  - *Chatty woman*

    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]
  - *Narrator*

    Above the antique mantel was displayed
    As though a window gave upon the sylvan scene
    The change of Philomel
    [97–99]
- Project goals
  - To segment according to changes in voice (Brooke et al. 2012) …
  - …and then cluster voice segments together (the present work)
- Related work
  - Quantitative poetry analysis (Dugan 1973; Simonton 1990)
  - Clustering in literature (Luyckx, 2006; Koppel et al., 2011)
  - Stylistic inconsistency detection (Graham et al., 2005)

2. Automatic Segmentation

- From our earlier work (Brooke et al. 2012)
- Unsupervised model
  - Consider each point in text
  - Stylistic change curve based on 50-token spans on either side
  - Select local maxima of curve as breakpoints
- Features
  - Readability metrics (e.g. word length, lexical density)
  - Frequency of punctuation
  - Frequency of part-of-speech
  - Frequency of line breaks
  - Sentiment metrics (Bacianella et al. 2010)
  - Formality score (Brooke et al, 2010)
  - Lexical LSA vectors from large web corpus, 20 dimensions
- Features normalized (mean = 0, standard deviation = 1)

3. Clustering Method

- Same feature vector as segmentation
- Clustering with *k*-means
  - Randomly choose *k* cluster centroids
  - Assign points to cluster
  - Iterate until convergence (less than 0.0001 change)
- Differences from standard *k*-means
  - Centroid is weighted by span length
  - Use city-block (L_1) distance instead of Euclidean
    - Based on our segmentation work
- *k* = 13, chosen based on expert annotation
  - Non-parametric model would be preferred

4. Evaluation

- BCubed metrics (Bagga and Baldwin, 1999)
  - Precision: fraction of same cluster pairs also in same category
  - Recall: fraction of same category pairs also in same cluster
  - F-score: harmonic mean of precision and recall
- Evaluation 1: 20 artificial mixed-style poems
  - Made from 12 poems representing *Waste Land* influences by taking 100-200 length spans from 6 of these poems
- Evaluation 2: *The Waste Land*
  - Expert annotation (not definitive)
- Segmentation baselines
  - Even spacing
  - Gold
- Clustering baselines
  - Initial (no clustering)
  - Random
- Seeded *k*-means
  - Use longest instance of each voice as initial centroid

5. Results

Table 1: Clustering results for artificial poems

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 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

- Similar results across both evaluations
  - Though *The Waste Land* is more difficult than artificial poems
- Automatic unsupervised better than even-spacing baseline
  - But not as good as suggested by segmentation metrics
- For most conditions, *k*-means is clearly better than baselines
  - Though marginal for gold condition in *The Waste Land*
- Starting with voice seeds is very helpful
- Voices most easily distinguished:
  - *Narrator* (F-score 0.869)
  - *Chatty woman* (F-score 0.605)

6. Conclusion

- Still a long way from a potential human interpretation
  - Though some correspondence between human and computer judgments of stylistic distinctiveness
- Improving segmentation seems key to future clustering gains
- Or is it possible to eliminate our separation of segmentation and clustering steps?

References

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.*

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.*

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics.*

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.*

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.

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.*

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.*

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.

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