An initial study of topical poetry segmentation

Chris Fournier
University of Ottawa
Ottawa, ON, Canada
cfour037@eecs.uottawa.ca

Abstract

This work performs some basic research upon topical poetry segmentation in a pilot study designed to test some initial assumptions and methodologies. Nine segmentations of the poem titled Kubla Khan (Coleridge, 1816, pp. 55-58) are collected and analysed, producing low but comparable inter-coder agreement. Analyses and discussions of these codings focus upon how to improve agreement and outline some initial results on the nature of topics in this poem.

1 Introduction

Topical segmentation is the division of a text by placing boundaries between segments. Within a segmentation, each segment should represent a coherent and cohesive topic. The decision to place a boundary between two segments of text is subjective and must often be determined manually. The factors involved in performing this subjective task are poorly understood, which motivates this work to begin the basic research required to understand this phenomenon.

For literature, topical segmentations have been produced for a short story (Kozima, 1993) and a novel (Kazantseva and Szpakowicz, 2012). Poetry, however, has had little attention in terms of topical segmentation. Brooke et al. (2012) collected segmentations of poetry that sought to delineate which voices communicate various segments of *The Wasteland* by T.S. Elliot (1888-1965), but a voice segment does not necessarily correlate with a topical segment. Because *The Wasteland*'s defining feature is its voice-shifts, more data is required to understand the variety of topical segments that could exist within poetry besides those delineated by changing voice — which this work aims to provide.[1]

[1]Available at http://nlp.chrisfournier.ca/

This work's goal is to begin to provide some initial information about what constitutes a topic in poetry by analysing the Romantic-era poem titled *Kubla Khan* (Coleridge, 1816, pp. 55-58) by Samuel Taylor Coleridge (1772–1834). Chosen for its beauty, variety, short length (54 lines), and lack of strict adherence to a prescribed structure (e.g., sonnets, odes, etc.), it is assumed that this purported fragment of a dream will contain a wide variety of different topics (as judged by manual coders).

This work aims to discover from reader's interpretations of topical segmentation in poetry the:

- Structure of these topics (e.g., are they linear, hierarchical, or something else?);
- Types and variety of topics (e.g., do topics shift when there are changes in time, place, description, exposition, etc.); and
- Relationship between poetic features and topical boundaries (e.g., do stanzas correlate with topical boundaries?).

Unfortunately, this work is simply a pilot study and it cannot make any generalizations about poetry overall, but inferences can be made about this single poem and its topical structure.

2 Related Work

Topical Segmentation Topical segmentation of expository texts such as popular science magazine articles have been well studied by Hearst (1993, 1994, 1997) while developing the automatic topical segmenter named *TextTiling*. On a parallel track, Kozima (1993) segmented a simplified version of O. Henry's (William Sydney Porter; 1862–1910) short story titled *Springtime à la Carte* (Thornley, 1816). Both bodies of work focused upon using lexical cohesion to model where topic boundaries occur and collected manual segmentations to study. This data,

Proceedings of the Second Workshop on Computational Linguistics for Literature, pages 47–51,
Atlanta, Georgia, June 14, 2013.

however, was never analysed for the types of segments contained, but only for the presence or absence of topic boundaries at specific positions.

Kazantseva and Szpakowicz (2012) delved deeper into topical segmentation of literature by collecting segmentations of Wilkie Collins' (1824–1883) romantic novel *The Moonstone* (Collins, 1868). In the novel, 20 of its chapters were segmented individually by 27 annotators (in groups of 4–6) into episodes. Episodes were defined as "topically continuous spans of text demarcated by the most perceptible shifts of topic in the chapter" (Kazantseva and Szpakowicz, 2012, p. 213). This work also analysed the boundaries placed by the coders themselves, but not the types of segments that they produced.

Brooke et al. (2012) collected voice-switch segmentations of *The Wasteland* by T.S. Elliot (1888-1965). Although voices are not topics, voice switching could constitute topical boundaries. Segmentations from 140 English literature undergraduate students and 6 expert readings were collected and used to compose one authoritative reference segmentation to test a large number automatic segmenters upon.

Agreement and Comparison Inter-coder agreement coefficients measure the agreement between a group of human judges (i.e. coders) and whether their agreement is greater than chance. Low coefficient values indicate that a task may have restricted coders such that their responses do not represent an empirical model of the task, or the task instructions did not sufficiently define the task. High coefficient values indicate the degree of reliability and replicability of a coding scheme and the coding collection methodology (Carletta, 1996). Although there is much debate about what coefficient value represents adequate agreement, any coefficient value can be used to compare studies of the same task that use different coding schemes or methodologies.

Many inter-coder agreement coefficients exist, but this work uses Fleiss' multi-π (π^*, Fleiss 1971; occasionally referred to as K by Siegel and Castellan 1988) to measure agreement because it generalizes individual coder performance to give a better picture of the replicability of a study. Specifically, an adaptation of the proposal by Fournier and Inkpen (2012, pp. 154–156) for computing π^* is used that is detailed by Fournier (2013).

Fournier (2013) modifies the work of Fournier and Inkpen (2012) to provide a more discriminative measure of similarity between segmentations called *boundary similarity* (B) — an edit distance based measure which is unbiased, more consistent, and more intuitive than traditional segmentation comparison methods such as P_k (Beeferman and Berger, 1999, pp. 198–200) and WindowDiff (Pevzner and Hearst, 2002, p. 10). Using the inter-coder agreement formulations provided in Fournier and Inkpen (2012), Fournier (2013) provides B-based inter-coder agreement coefficients including Fleiss' multi-π (referred to as π^*_B) which can discern between low/high agreement while still awarding partial credit for near misses.

3 Study Design

This work is a small study meant to inform future larger studies on topical poetry segmentation. To that end, a single 54 line poem, *Kubla Khan* (Coleridge, 1816, pp. 55-58), is segmented. Written in four stanzas (originally published in two) composed of tetra and penta-meter iambs, this well studied work appears to show a large variety of topical segment breaks, including time, place, scenery, narration, exposition, etc. Stripped of its indentation and with its stanzas compressed into one long sequence of numbered lines, this poem was presented to segmenters to divide into topics.

Objectives The objective of this study is to identify whether topics in poems fit well into a linear topic structure (i.e., boundaries cannot overlap) and to test the annotation instructions used. Additionally, a survey of the types and variety of topics is desirable to inform whether more than one boundary type might be needed to model segment boundaries (and to inspire statistical features for training an automatic topical poetry segmenter). Finally, the relationship between poem features and topic boundaries is of interest; specifically, for this initial work, do stanzas correlate with topical boundaries?

Subjects Nine subjects were recruited using Amazon's Mechanical Turk from the United States who had an exemplary work record (i.e., were "Master Tukers"). Segment text summaries were analysed for correct language use to ensure that coders

demonstrated English language proficiency.

Granularity Segmentations were solicited at the line level (arbitrarily assuming that a topic will not change within a line, but may between lines). This level is assumed to be fine enough to partition segments accurately while still being coarse enough to make the task short (only 54 lines can be divided into segments). Because there may be a great number of topics found in the poem by readers, it is assumed that a nearly missed boundary would only be those that are adjacent to another (i.e., n_t for B is set to 2).

Collection procedure Segmenters were asked to read the poem and to divide it into topical segments where a topic boundary could represent a change in time, scenery, or any other detail that the reader deems important. A short example coding was also provided to augment the instructions. Along with line number spans, a single sentence description of the segment was requested (for segment type analysis and to verify coder diligence and thoughtfulness) and overall comments on the task were solicited.

4 Study Results and Analysis

Time The 9 subjects took 35.1556 ± 18.6796 minutes to read and segment the poem.[2] Each was remunerated \$8 USD, or $\$18.91 \pm 11.03$ USD per hour.

Segmentations The 9 coders placed 17.6667 ± 6.2716 boundaries within the 54 lines of the poem. The number of segmentations produced by each coder is shown in Figure 1a, along with the mean and standard deviation (SD).

Agreement The segmentations provided by the 9 coders in this study have an inter-coder agreement coefficient value of $\pi^*_{\mathrm{B}} = 0.3789$. This value is low, but it is only slightly below that of Hearst (1997) (0.4405) and Kazantseva and Szpakowicz (2012) (0.20, 0.18, 0.40, 0.38, 0.23 for each of the 5 groups) as reported in Fournier (2013). This value is also not unexpected given the different coding behaviours (e.g., boundary placement frequency) in Figure 1a.

Similarity Using Boundary Similarity (B), taking $1 - \mathrm{B}$ can yield a simple distance function between segmentations. Because of the low agreement of this study, it is assumed that there must be subsets of coders who agree more with each other than with others (i.e., clusters). Using $1-\mathrm{B}$ as a distance function between segmentations, hierarchical agglomerative clustering was used to obtain the clusters shown in Figure 1b. Computing inter-coder agreement for these clusters produces subsets with significantly higher than overall agreement (Table 1).

Labels Taking the single-sentence descriptions of each topic, an attempt was made to label them as belonging to one or more of these categories:

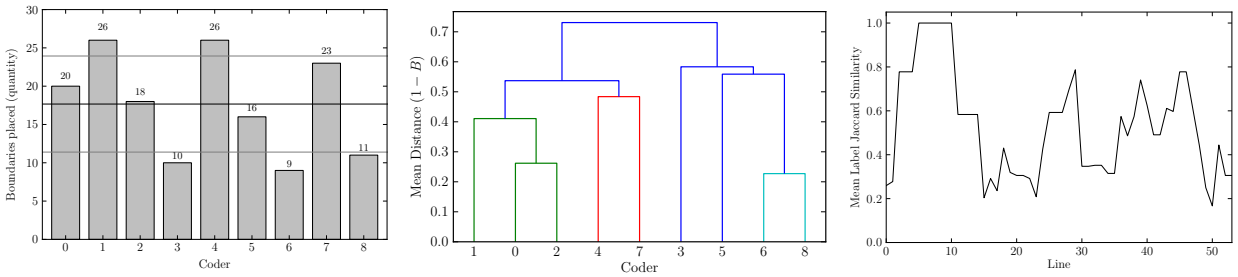
1. Exposition (e.g, story/plot development);
2. Event (e.g., an action or event occurred);
3. Place (Location is stated or changed);
4. Description (of an entity; can be specific): *a*) Scenery *b*) Person *c*) Sound *d*) Comparison (simile or metaphor)
5. Statement (to the reader).

These labels were decided by the author while reading the segmentations and were iteratively constructed until they suitably described the one-line segment topic summaries. Using Jaccard similarity, the labels placed on each position were compared to those of each other coder to obtain mean similarity of each line, as plotted in Figure 1c. This shows that in terms of topic types, actual agreement varies by position. The portions with the highest agreement are at the beginning of the poem and contain scenery description which appear to have been easy to agree upon (type-wise). Overall, mean label similarity between all coders was 0.5330 ± 0.4567, but some of the identified clusters exhibited even higher similarity (Table 1).

Feature correlations There is some evidence to suggest that boundaries between the four stanzas at lines 11–12, 30–31, and 36–37 correlate with topical shifts because $^6/_9$, $^9/_9$, and $^9/_9$ (respectively) coders placed boundaries at these locations. There is little evidence to suggest that the indentation of line 5 and lines 31–34 (not shown) correlate with topical shifts because only $^1/_9$ and $^5/_9$ (respectively) coders placed boundaries between these segments.

Topical structure One of the coders commented that they felt that the segments should overlap and

[2] One coder took far less time because they submitted part of their answers via email and time was not accurately recorded.



(a) Coder boundaries placed with mean and SD

(b) Hierarchical agglomerative clusters using 1 − B as a distance metric

(c) Mean Jaccard similarity of topic label types per line

Figure 1: Various analyses of the 9 manual segmentations of Kubla Khan

Coders	$\{4,7\}$	$\{0,2\}$	$\{6,8\}$	$\{1,0,2\}$	$\{1,0,2,4,7\}$	$\{3,5,6,8\}$	$\{5,6,8\}$
π^*_B	0.3704	0.6946	0.7625	0.5520	0.4474	0.4764	0.5389
E(J)	0.491 ± 0.495	0.460 ± 0.439	0.685 ± 0.464	0.508 ± 0.452	0.512 ± 0.467	0.593 ± 0.425	0.580 ± 0.432

Table 1: Inter-coder agreement (π^*_B) and mean Jaccard topic label similarity (with WD) for coder clusters

coded so. These codings were adjusted by the author to not overlap for analysis, but the coder's comment highlights that perhaps these segments should be able to overlap, or that linear segmentation may not be an adequate model for topics in poetry.

5 Discussion

Given the low (but comparable) inter-coder agreement values of this study, it is evident that some variables are not properly being controlled by the procedure used herein. Before a larger study is performed, the issue of low agreement must be explained; some hypotheses for this are that:

1. Coders may have been of varying levels of education, English proficiency, or motivation;
2. Instructions may have not been clear or exhaustive in terms of the potential topics types;
3. A linear segmentation not allowing for overlap may artificially constrain coders; and
4. The poem selected may simply be inherently difficult to interpret and thus segment.

This study has, however, catalogued a number of topic labels which can be used to better educate coders about the types of topical segments that exist, which could lead to obtaining higher inter-coder agreement. Pockets of agreement do exist, as shown in the clusters and their agreement and topic label similarity values (Table 1). If more data is collected, but inter-coder agreement stays steady, perhaps instead these clusters will remain and become more populated. Maybe these clusters will reveal that the problem was modelled correctly, but that there is simply a difference between the coders that was not previously known. Such a difference could be spotted using clustering, but what the actual difference is may remain a mystery unless more biographical details are available (e.g., sex, age, education, English proficiency, reading preferences, etc.).

6 Conclusions and Future Work

Although Kubla Khan is a beautiful poem, its topical segmentation is vexing. Low inter-coder agreement exemplified by this study indicates that the methodology used to investigate topical poetry segmentation may require some modifications, or more biographical details must be sought to identify the cause of the low agreement. Clustering was able to identify pockets of high agreement and similarity, but the nature of these clusters is largely unknown — what biographical details or subjective opinions of the task separate these groups?

Future work will continue with subsequent pilot studies to attempt to raise the level of inter-coder agreement or to explain the low agreement by looking for clusters of coders who agree (and attempting to explain the relationships between coders in these clusters). Also, more poems need to be analysed to make generalisations about poetry overall. The relationships between topical segments in poetry and other poetic features such as rhyme, meter, and expert opinions are also worth investigation.

References

Beeferman, Doug and Adam Berger. 1999. Statistical models for text segmentation. *Machine Learning* 34:177–210.

Brooke, Julian, Adam Hammond, and Graeme Hirst. 2012. Unsupervised Stylistic Segmentation of Poetry with Change Curves and Extrinsic Features. In *Proceedings of the 1st NAACL-HLT Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 26–35.

Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22(2):249–254.

Coleridge, Samuel Taylor. 1816. *Christabel, Kubla Khan, and the Pains of Sleep*. John Murray.

Collins, Wilkie. 1868. *The Moonstone*. Tinsley Brothers.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76:378–382.

Fournier, Chris. 2013. Evaluating Text Segmentation using Boundary Edit Distance. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA.

Fournier, Chris and Diana Inkpen. 2012. Segmentation Similarity and Agreement. In *Proceedings of Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 152–161.

Hearst, Marti A. 1993. TextTiling: A Quantitative Approach to Discourse. Technical report, University of California at Berkeley, Berkeley, CA, USA.

Hearst, Marti A. 1994. *Context and Structured in Automated Full-Text Information Access Context and Structure in Automated Full-Text Information Access*. Ph.D. thesis, University of California Berkeley.

Hearst, Marti A. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics* 23:33–64.

Kazantseva, Anna and Stan Szpakowicz. 2012. Topical Segmentation: a Study of Human Performance. In *Proceedings of Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 211–220.

Kozima, Hideki. 1993. Text segmentation based on similarity between words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '93, pages 286–288.

Pevzner, Lev and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics* 28:19–36.

Siegel, Sidney and N. J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York, USA, chapter 9.8. 2 edition.

Thornley, G. C., editor. 1816. *British and American Short Stories*. Longman Simplified English Series. Longman.