An initial study of topical poetry segmentation

Chris Fournier
University of Ottawa
cfour037@eecs.uottawa.ca

Introduction

Objectives This study seeks to answer questions about topics in poetry, including:
1. Does a linear topic structure fit poetry?
2. What annotation instructions should be used?
3. What relationships exist between poem & topic features?

URL For data and related software, see http://github.com/cfournie/initial.poetry.segmentation

Annotation

Scope This study collected observations upon only one poem (Figure 1), so no generalizations can be made. This work seeks to inform future larger studies.

Subjects 9 subjects from Amazon’s Mechanical Turk.

Grainularity Poem line level.

Collection procedure Segmenters were asked to:
1. Read the poem;
2. 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; and
3. Describe, in one sentence, each segment.

Time 9 subjects took 35.15±6.87 minutes to read and segment the poem. Each was remunerated $8 USD or $18.91 ± 11.03 USD per hour.

Segmentations 9 coders placed 17.6667±6.2716 boundaries. The number of segmentations produced by each coder is shown in Figure 2.

Agreement Between coders, \( \pi_{ij} = 0.3789 \) low, but not unexpected given the different coding behaviours (e.g., boundary placement frequency) in Figure 2.

Similarity Despite the low agreement, subsets (i.e., clusters) of coders may agree more than others. Using \( 1 − B \) (Fournier, 2013) as a distance function, hierarchical agglomerative clustering produces Figure 3, with the agreement for reach cluster shown in Figure 4.

Line
1 In Xanadu did Kubla Khan
2 A stately pleasure-dome decree;
3 Where Alph, the sacred river, ran
4 Through caverns measured with man
5 Down to a sunless sea.
6 So twice five miles of fertile ground;
7 With walls and towers were girdled round;
8 And here were gardens bright with sinuous rills;
9 Where blossoms many an incense-bearing tree;
10 And here were forests ancient as the hills.
11 Enfolding sunny spots of greenery.
12 But oh! that deep romantic chasm which slanted
13 Down the green hill thriftless to a solemn cave!
14 A savage place! as holy and enchanted
15 As e’er beneath a morning moon was haunted
16 By woman wailing for her demon-lover!
17 And from this charm, with ceaseless turning going;
18 As if this earth in fast thick pants were breathing;
19 A mighty fountain momently was forced;
20 Its voice was heard as the thunder roaring;
21 Its fury unabated was kept up by the sea;
22 And lost in the greatabolic surge of all;
23 And on sink in tumult to a lifeless ocean.
24 And here were forests ancient as the hills.
25 Five miles meandering with a mazy motion
26 Through wood and dale the sacred river ran;
27 Then reached the cavern measureless to man,
28 And sank in tumult to a lifeless ocean.
29 And ‘mid this tumult Kubla heard from far
30 Ancestral voices prophesying war.
31 The shadow of the dome of pleasure
32 In Xanadu did Kubla Khan
33 A stately pleasure-dome decree;
34 From the fountain and the caves.
35 Ancestral voices prophesying war!
36 And here were forests ancient as the hills.
37 A damsel with a dulcimer
38 It was an Abyssinian maid;
39 She built her satisfaction;
40 She built her satisfaction.
41 She built her satisfaction.
42 It was an Abyssinian maid
33 She built her satisfaction;
34 She built her satisfaction.
43 She built her satisfaction.
44 She built her satisfaction.
45 She built her satisfaction.
46 She built her satisfaction.
47 She built her satisfaction.
48 She built her satisfaction.
49 She built her satisfaction.
50 She built her satisfaction.
51 She built her satisfaction.
52 She built her satisfaction.
53 For he on honey-dew hath fed,
54 And drunk the milk of Paradise.

Analysis & Discussion

Labels Segment summaries were annotated to label:
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).

Mean label Jaccard-similarity between all coders is 0.5330 ± 0.4567, but some clusters are higher (Figure 4).

Feature correlations The four stanzas at lines 11–12, 30–31, and 36–37 appear to correlate with topological shifts, but not the indentation of line 5 and lines 31–34 (see Figure 1).

Low agreement This study’s low agreement must be explained in future work, and may have been caused by:
1. Coders levels of education, proficiency, or motivation;
2. Instruction clarity;
3. Linear segmentation imposing artificial constraints; and
4. Inherent difficulty interpreting the chosen poem.

Clusters of agreement Pockets of agreement exist within clusters (Figure 4). If more data is collected, but inter-coder agreement stays steady; perhaps instead these clusters will remain and become more populated and may represent coder-types which could be individually explained.

Conclusion

This work serves an initial guiding study in topical segmentation of Kubla Khan, but more work is required to:
1. Obtain higher inter-coder agreement;
2. Explain clusters of agreement;
3. Find more correlations between poem-features & topics.

References
