Segmenting Oral History Transcripts

Ryan Shaw

School of Information and Library Science
University of North Carolina at Chapel Hill
ryanshaw@unc.edu
https://aeshin.org

Abstract. Dividing oral histories into topically coherent segments can make them more accessible online. People regularly make judgments about where coherent segments can be extracted from oral histories. But when different people are asked to extract coherent segments from the same oral histories, they often do not agree about where such segments begin and end.

Keywords: oral history, discourse segmentation, natural language processing, digital libraries

1 Introduction

Oral histories are rich and unique documents of the past and our memory of it. Putting oral histories on the web makes them more accessible, but they remain daunting to consume [4]. It requires a significant time commitment to listen to a one or two hour interview. This is why curators of oral history collections, when creating "exhibits" of their materials for the public, usually select short extracts from longer interviews. Scholars also select extracts from their recordings when presenting their work to a live audience [7, 265]. But are extracts merely subjective judgments, or do they reflect a topical structure about which consensus might be reached? An analysis of 829 judgments about oral history extract boundaries suggests that while these judgments are not purely subjective, consensus about topical structure is weak.

2 Data and Methods

Our corpus consisted of 19 transcripts of oral history interviews conducted by the Southern Oral History Program (SOHP) at the University of North Carolina.[1] In an earlier project, SOHP staff transcribed and selected salient extracts from each of the interviews. Each interview was divided into segments, of which some subset (the selected extracts) were judged to be topically coherent. On average, half of the segments were selected as salient extracts. The unselected segments

[1] All of the data and code discussed in this paper are available at `https://github.com/contours`.

tended to be longer, with a mean length of 70 sentences compared to 44 sentences for the selected extracts. This suggests that the extraction process may not identify some potential topic boundaries (those that appear within the segments of interviews judged to be less salient).

Two non-expert annotators were asked to imagine that they had been tasked with curating an online collection of oral histories and to select "the most important parts" from each transcript. Each transcript was presented as an HTML page showing only the names of the speakers and the transcribed text of their speech. The annotators could click on the text to split it into segments and then indicate which segments were to be selected as extracts. They were instructed that each extract "should cover a single topic or anecdote and should be understandable on its own." To give them a sense of the expected granularity of the extracts, the annotators were told that the length of extracts could vary considerably but would average around 30–50 sentences (the average extract length in the original project).[2] Extracts could not overlap, and not all of the text had to be extracted (i.e. it was permissible to "leave out" parts of the transcript between extracts). Extract boundaries were not limited to speaker changes or paragraph breaks and could be placed between any two adjacent sentences.

3 Comparing Segmentations

The three annotators placed a total of 829 boundaries between sentences of the 19 interview transcripts, creating a total of 886 segments. The distribution of segment lengths appears to be exponential (see figure 1). The mean segment length was 50 ± 52.7 sentences ($n = 886$). The shortest and longest segments created were one sentence and 621 sentences long, respectively. The longest segment may be an error, since it is far longer than any other segment in the dataset and far longer than the other segments created by that annotator for that interview. Some of the very short segments seem to be cases of "trimming" behavior, where a short segment is created specifically to be excluded (e.g. an interviewer unsuccessfully trying to interrupt an interviewee with a new question) or highlighted (e.g. a particularly vivid single quote that makes sense in isolation). Different tendencies to "trim" like this may explain some of the variations observed across the annotators' segmentations.

If the annotators had consistently segmented at the same level of granularity, one would expect to see little variance in the segment lengths and a positive correlation between each transcript's length and the number of segments into which it was divided. But segment lengths varied significantly. While there was a positive correlation between transcript length and the number of segments, there was also a positive correlation between transcript length and the median segment length. Thus while the annotators divided longer transcripts into more segments,

[2] The complete text of the instructions provided to the annotators is available at `https://github.com/contours/segment/blob/5404fce/public/instructions.html`.

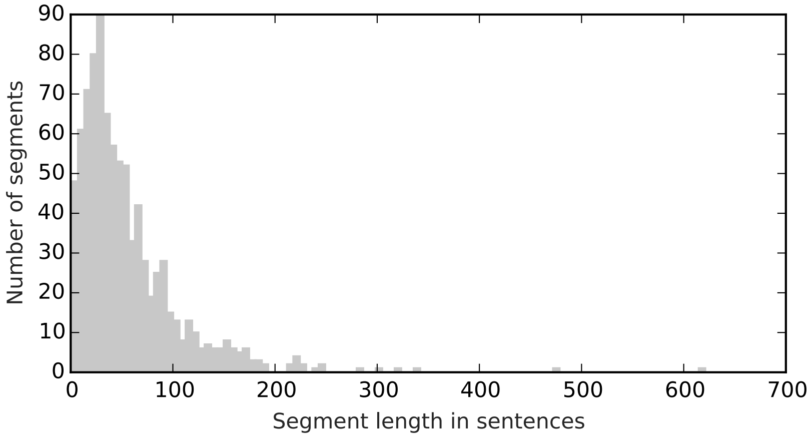

Fig. 1: Lengths in sentences of the manually-created segments.

they also created longer segments for longer transcripts. Longer interview transcripts might reflect the fact that some interviewees are more long-winded than others. If that were the case then one might expect longer segments for longer interviews: each topic takes longer to cover. However it is also possible that longer interview transcripts simply cover more topics than others. If so, then the division of longer interviews into longer segments may have been due to annotator exhaustion, resulting in interviews segmented at different levels of granularity.

Segmentation granularity also varied across annotators. The original extractor placed boundaries slightly more frequently than the overall rate, which was about one boundary per 59 potential boundaries. Annotator A placed fewer boundaries (creating longer segments) on average than either the original extractor or annotator B. Annotator B placed more boundaries (creating shorter segments) on average than the other two. This could indicate that annotator B engaged in more "trimming" than the other annotators. These differences in boundary placement rates across annotators might lead one to expect low inter-annotator agreement, and indeed that is the case.

To measure inter-annotator agreement the *boundary edit distance* metric proposed by Fournier [3] was used. Boundaries differing by more than eight sentences were treated as misses, while boundaries differing by eight or less sentences were treated as "near misses", scaled the distance between the boundaries. A boundary edit distance of one indicates perfect agreement. The micro-averaged edit distance between pairs of boundaries placed by two annotators was 0.27 ± 0.0232 (95% CI, $n = 1270$). This pairwise mean boundary edit distance measures actual agreement; to correct for chance agreement one can calculate Fleiss' π^* coefficient [2], which was also 0.27.

4 Discussion and Future Work

This study examined flat, non-overlapping segmentations of oral histories, but topical structure is generally believed to be hierarchical [6]. Ashplant [1, 107] suggests that this is true of oral histories in his analysis of part of *The Dillen* [5]. He discerns a three-level topical hierarchy, with *anecdotes* grouped into *narrative elements*, which are in turn grouped into broad topical *clusters*. Differing judgments about which of these levels is the appropriate one for selecting extracts from may have been a factor contributing to the varying segmentation granularities found in this study.

Segmentation and segment-level description could make oral histories more accessible. But even though identifying topically coherent segments within interviews is an accepted part of working with oral histories, people often do not agree on the boundaries of those segments. In this study, annotators agreed (exact match or "near miss") on less than half of the boundaries they placed. Further progress will depend on clearer definition of the segmentation task.

5 Acknowledgments

This work was funded by the Institute of Museum and Library Services. Thanks to the Southern Oral History Program for making the interview transcripts available and Kathy Brennan and Sara Mannheimer for their annotation work.

References

1. Ashplant, T.G.: Anecdote as Narrative Resource in Working-Class Life Stories: Parody, Dramatization and Sequence. In: Chamberlain, M., Thompson, P. (eds.) Narrative and Genre, pp. 99–113. Routledge, London (1998)
2. Fleiss, J.L., Nee, J.C., Landis, J.R.: Large sample variance of kappa in the case of different sets of raters. Psychological Bulletin 86(5), 974–977 (1979), `http://dx.doi.org/10.1037/0033-2909.86.5.974`
3. Fournier, C.: Evaluating Text Segmentation using Boundary Edit Distance. In: Proceedings of 51st Annual Meeting of the Association for Computational Linguistics. p. to appear. Association for Computational Linguistics, Stroudsburg, PA, USA (2013), `http://anthology.aclweb.org/P/P13/P13-1167.pdf`
4. Frisch, M.: Oral History and the Digital Revolution: Toward a Post-Documentary Sensibility. In: Perks, R., Thomson, A. (eds.) The Oral History Reader. Routledge, London, 2nd edn. (2006)
5. Hewins, G.H., Hewins, A.: The Dillen: Memories of a man of Stratford-Upon-Avon. Elm Tree Books, London (1981)
6. Manning, C.D.: Rethinking Text Segmentation Models: An Information Extraction Case Study. Tech. rep., University of Sydney (1998), `http://nlp.stanford.edu/cmanning/papers/SULTRY-980701.ps`
7. Thompson, P.: The Voice of the Past: Oral History. Oxford University Press, Oxford, 3rd edn. (2000)