Evaluation of Automatic Caption Segmentation

James M. Waller
University of Chicago
5801 S Ellis Ave
Chicago, IL 60637
jmwaller@uchicago.edu

Raja S. Kushalnagar
Gallaudet University
800 Florida Ave NE
Washington, DC 20002
raja.kushalnagar@gallaudet.com

ABSTRACT

Captions are typically segmented in a way that respects grammatical boundaries and makes them more readable. However, the growth of online video content with captions generated from transcripts means that this segmentation process is often ignored. This study evaluates the effects of text segmentation on caption readability, and proposes a program to automatically segment captions using a parser. The parser-segmented captions readability is also evaluated and compared to human-segmented captions and arbitrarily-segmented captions. Results indicate segmentation influences sentence recall, though other wise little difference is found between the different kinds of captioning.

CCS Concepts

•Human-centered computing → Accessibility design and evaluation methods;

Keywords

Caption Segmentation

1. INTRODUCTION

Closed Captioning is used to represent spoken and audio information as written language in real-time. Primarily used by Deaf and hard-of-hearing people, closed captions are beneficial for a number of people in many situations. The current study focuses on how to automatically divide up lines on captions to maximize readability and comprehension. Current guidelines for captions specify a number of rules explaining where lines breaks should and should not occur âĂŞ at the end of a sentence, never in the middle of a name or compound sentence, etc. [1] The goal is to keep words in the same grammatical phrase together as much as possible. While this is useful, currently captionists must do this work manually, and on some websites these conventions are ignored entirely. This study develops a prototype based on text parsing information to segment captions at the optimal line breaks, and evaluate the usefulness of the program.

2. RELATED WORK

2.1 Impact of Segmentation on Readability

Research suggests caption segmentation can have a positive impact, but is not conclusive. Rajendran et al. [6] found participants were able to focus more on the video itself when captions appeared one phrase or sentence at a time, rather than by one word. On the other hand Perego et al. [5] found that segmenting subtitles in inappropriate places had no impact on sentence recall or eye movement. One explanation for he difference is that Rajendran et al. looked at breakpoints between sequential captions segments, when one appeared after the other, while Perego et al.'s "ill-formed" segmentations mostly occur between two simultaneous lines that appear together. This suggests that good breakpoints are more important when one segment appears after the other. The current study involves breakpoints at between both simultaneous and sequential lines.

2.2 Automatic Segmentation of Captions

Only two studies to date have tested programs that automatically segment captions. Murata et al. [4] developed a program for segmenting Japanese captions, while another study by Álvarez et al. [2] developed a general program, which they applied to Basque subtitles. Both used machine learning to teach their programs to identify correct segmentation points. Both programs had about 70-75% agreement with human captionists, and were judged as more readable than a baseline. Álvarez et al. did not involve grammatical analysis, and noted it was limited by the lack of parsers available for Basque. For English, many parsers are openly available, such as the Stanford parser used in this study. The Stanford parser can parse a transcript and produce a hierarchical phrase structure for each sentence [3], making it easy to identify optimal breakpoints.

3. PROGRAM AND EXPERIMENT

3.1 Program Development

The ACS (Automatic Caption Segmentation) program developed for this experiment first applies the Stanford Parser to produce syntax trees for the sentences in the transcript. The sentence is broken into phrases all the way down to the level of words. The program uses this information to identify optimal point breaks. All potential breakpoints between

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ASSETS '16 October 23-26, 2016, Reno, NV, USA

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4124-0/16/10. . . $15.00

DOI: http://dx.doi.org/10.1145/2982142.2982205

words are assigned indexes depending on how many parent nodes the words share. Since a separate syntax tree is created for each sentence, sentence boundaries have the lowest index and are the best candidates for breakpoints. If no sentence boundary is available, then the program searches within the program to the breakpoint with the next lowest index. If two adjacent words are deeply embedded together in a sentence, they will have a higher index, and will not be chosen, preventing phrases from being broken up.

The parser also puts additional emphasis on punctuation. Captioning guidelines also emphasize the importance of breakpoints at punctuation and “natural pauses" [1]. The program is modified to reflect this, by automatically reducing the index of punctuated breakpoints, thereby increasing the possibility they are selected as breakpoints.

3.2 Experimental Methodology

3.2.1 Participants 18 Deaf/hh participants were recruited for the first experiment, ranging from 20 to 28 years in age. All participants regularly use captions when watching online videos, TV, and other audio-video content. The entire experiment takes approximately 30-45 minutes and the participants are compensated for their time.

3.2.2 Procedure Each participant first watches two different 4-minute videos, each with a one of the following caption conditions:

- A: Segmented by a human captionist
- B: Segmented by the ACS program
- C: Segmented only by line-length (no parsing)

Thus each participant saw two of these three styles - condition A Human-generated captions follow standard conventions for breaking caption lines in suitable ways, while condition C captions simply obey a character-per-line limit and ignore phrase structure. There are eight comprehension questions given to the participant after the video, as well as two to three pre-test comprehension questions to check previous knowledge of the subject. The comprehension section consists of two kinds of multiple-choice questions: four general comprehension questions that asked about information presented in the text (i.e. What is the purpose of the video?) and four recall questions that asked gave the beginning of a sentence from the caption transcript and asked the participant to select the correct ending of the sentence from memory. Two subjective Likert-scale questions ask the participant to rate the readability of the captions and their satisfaction with the captions. There is also an open-ended question asking about general feedback on the questions, and if they noticed a difference between the two styles presented.

3.3 Results

No significant differences were found in subjective preference from the Likert question data, either in terms of satisfaction or readability across the three conditions, using Mann-Whitney tests. Parametric t-tests were used for data from the test questions. For comprehension question data (shown in Figure 1), there was no differences in general comprehension for the three conditions, but for sentence recall, participants performed significantly better for condition B ($M = 2.08$ out of 4 correct) than for C ($M = 1.00$, $t(22) = 2.17$, $p < 0.05$). There was no significant difference in recall between conditions A ($M = 1.58$) and either B or C.

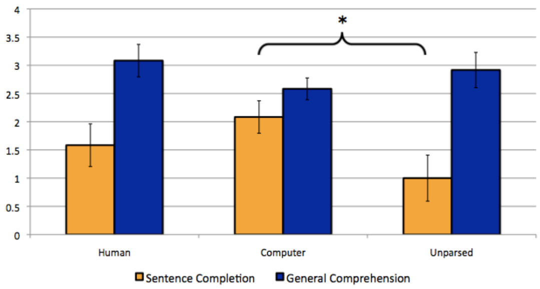

Figure 1: Average Number of Correct Questions

4. CONCLUSION

The general lack of differences between human-segmented (A) and arbitrarily segmented captions (C) may support Perego et al.'s conclusion [5] - that segmentation in captioning has little or no impact on readability. This study supports the notion that automatic segmentation could replace manual segmentation, since the two types performed very similarly.

Also, the difference between conditions B and C for sentence recall suggests that segmentation may indeed have an impact on our memory of the text. This result points to future ways to measure caption readability. Performance on content comprehension questions can involve many variables unrelated to the captions including background knowledge and inference from video graphics. Sentence recall questions more directly measure whether or not the participants recall the actual caption text itself. Future research could explore the viability of this kind of measurement, and improve it.

Any future work on caption segmentation must involve Deaf/hh people from a wider range of educational backgrounds, ages, and English reading skill to see if all benefit equally from caption segmentation.

5. REFERENCES

[1] Captioning key - text. https://www.dcmp.org/captioningkey/text.html. Accessed: 2015-06-08.

[2] Aitor Álvarez, Haritz Arzelus, and Thierry Etchegoyhen. Towards customized automatic segmentation of subtitles. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 229–238. Springer, 2014.

[3] Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750, 2014.

[4] Masaki Murata, Tomohiro Ohno, and Shigeki Matsubara. Automatic linefeed insertion for improving readability of lecture transcript. In *New Directions in Intelligent Interactive Multimedia Systems and Services-2*, pages 499–509. Springer, 2009.

[5] Elisa Perego, Fabio Del Missier, Marco Porta, and Mauro Mosconi. The cognitive effectiveness of subtitle processing. *Media Psychology*, 13(3):243–272, 2010.

[6] Dhevi J Rajendran, Andrew T Duchowski, Pilar Orero, Juan Martínez, and Pablo Romero-Fresco. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives*, 21(1):5–21, 2013.