A Comparison Study of Human-Evaluated Automated Highlighting Systems

Sasha Spala[1], **Franck Dernoncourt**[2], **Walter Chang**[2], **Carl Dockhorn**[1]
[1]Adobe Systems, [2]Adobe Research
345 Park Ave
San Jose, CA 95110-2704
{sspala,dernonco,wachang,cdockhor}@adobe.com

Abstract

Automatic text highlighting aims to identify key portions that are most important to a reader. In this paper, we explore the use of existing extractive summarization models for automatically generating highlights; automatic highlight generation has not previously been addressed from this perspective. Evaluation studies typically rely on automated evaluation metrics as they are cheap to compute and scale well. However, these metrics are not designed to assess automated highlighting. We therefore focus on human evaluations in this work. Our comparison of multiple summarization models used for automated highlighting accompanied by human evaluation provides an approximate upper bound of the quality of future highlighting models.

1 Introduction

Automatic text highlighting aims to identify key portions that are most important to a first-time reader. Our motivation for pursuing highlighting analysis stems from an apparent lack of current computational research and machine learning applications dedicated to this type of reading aid. For clarity, we define document highlights as a brightly-colored overlay placed on top of a span of text in order to attract readers' attention to the content of that text. We do not define any one specific color those highlights must be, but our practice, as outlined in the methodology, makes use of yellow, green, and red highlights.

Highlighting, when used correctly, can be an important part of the reading process, and, in many cases, can aid a reader in information retention (Fowler and Barker, 1974). Because of its similarity to extractive summarization tasks, where highlights reflect the sentences retrieved for an extractive summary, it follows that this is a relevant repurpose of these summarization methods. We must recognize, however, that this is *not* a summarization task, as evaluation of a highlighting model must be cruicially different: Summarization tasks produce summaries that are evaluated as relevant, but separate from the document whereas the evaluation of highlight models must be done within the direct context of the document. Though the approach may be similar, the evaluation is fundamentally different and likely influences the structure of the model applied.

In order to confirm this is both useful and possible, however, we must first answer two questions: 1) Can humans agree on which sentences in a document should be highlighted? Knowing whether human readers distinguish a specific set of sentences from a document will inform to what extent this task differs from extractive summarization tasks, as it would identify how human highlights and baseline summarization model "highlights" differ, if at all. 2) How do current summarization models, trained for their original application, perform when used out-of-the-box as highlight generators?

To answer these two research questions, we created a novel method of collecting data from human annotators via Amazon Mechanical Turk or any other crowdsourcing platform. This paper presents our experimental methodology as well as the results we obtained from the annotation tasks. Our results reflect promising developments in the application of

Copyright 2018 by the authors

summarization models and the possibility for a gold dataset.

2 Related Work

Highlighting is one of the most common methods of annotation (Baron, 2009), making it a popular content annotation choice for many industry leaders as well. For our purposes, we are interested in the effect of passive highlights, or highlights that already appear in text. Passive highlighting has been shown in several studies (Fowler and Barker, 1974; Lorch Jr., 1989; Lorch Jr et al., 1995) to be a useful tool for information retention and comprehension.

Rath et al. (1961) asked human annotators to retrieve the "most representative" sentences in a document, but failed to find significant human agreement for both human-retrieved and machine-retrieved sentences. Daumé III and Marcu (2004) identified that when instructed to choose the "most important" sentences from a passage, humans still fell short of significant agreement. Though Daumé III and Marcu had low expectations for human agreement in the summarization domain, we believe that the effect of *inline* content, such as highlights, changes the parameters of this task enough to warrant further inspection. In an effort to find a more narrow task definition without influencing annotators' concept of "highlighting importance", we centered our task on the expected concrete and finite results of in-line highlights (e.g., increased reading speed, enhanced comprehension).

The connection between highlights and extractive summarization has yet to be established in the computational linguistics community, but we believe the two have a significant relationship. The two tasks, as mentioned above, require the same fundamental process: retrieving important sentences from a document. Certainly, there exists a large body of research on extractive summarization, utilizing many different approaches (see (Nenkova and MecKeown, 2012; Yogan et al., 2016) for details on extractive summarization techniques), but these works focus on summarization, and thus, evaluation is done out of the direct context of the document. That is, summaries are scored based on their completeness, coherence, and importance as a standalone paragraph rather than in line with the text. Because of this shortage of computational research, we must define a proper metric for evaluating machine-generated highlights. While ROUGE scores (Lin, 2004) continue to be the standard metric for evaluating the coherence and correctness of extractive and abstractive summaries, it is difficult to apply them to a highlighting task. Instead, we are interested in human reactions to and interest in these highlights.

3 Human Agreement Task

Before developing a capable highlight generation model, we must answer whether it is possible to find a ground truth to highlights in a given document. We designed an Amazon Mechanical Turk experiment to test whether, given an appropriate stimulus question, humans could agree on which sentences in a document should be highlighted.

3.1 Single-Document Interface Design

Our annotation method is based on a novel framework to gathering highlight data from human subjects. Participants, as discussed above, were asked to highlight a specific number of sentences that would make document comprehension easier and faster for another reader. A counter in the left column updated the number of highlights remaining as annotators worked through each document. Clicking anywhere within the boundaries of a sentence would highlight the entire sentence in yellow. Annotators were allowed to highlight and un-highlight as often as needed, but were not able to revisit the same document after moving on.

The agreement task was designed to elicit highlighting data without confounds. Highlighting is simple and only requires a single left click, and there are no color variations. Text size for both the left panel and the document title and content are consistent for the duration of the task.

3.2 Methodology

We asked 40 human annotators to highlight sentences that would make understanding the "main point" of the same document "easier and faster for a new reader". All annotators were required to have a US high school diploma or equivalent to ensure fluency in English. Annotators were instructed to assume their new reader would have no prior knowledge of the content of the background. Each anno-

Copyright 2018 by the authors

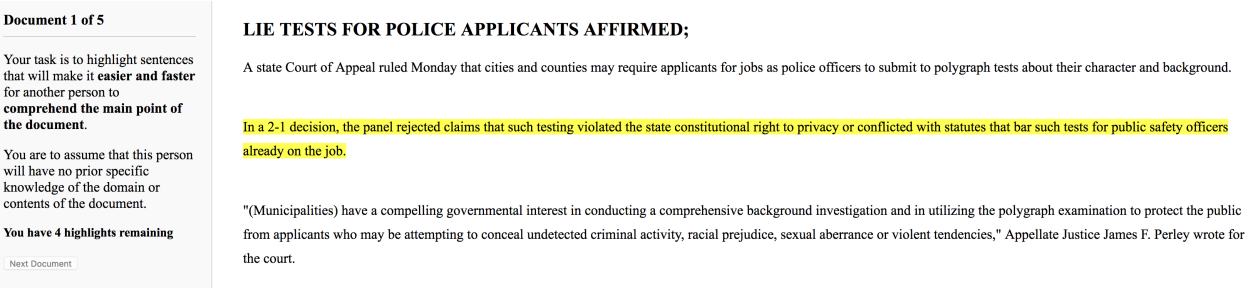

Figure 1: Interface for the human agreement task. Annotators may highlight or unhighlight any sentence by clicking on it. The interface verifies that the annotator has highlighted the proper amount of highlights.

tator was given 5 randomly ordered documents from a selection of 10 documents from the 2001 Document Understanding Conference (DUC) data (DUC, 2001) and asked to highlight a number of sentences approximating 20 percent of the total number of sentences in each document.

All annotators received a brief tutorial of the task controls before beginning the experiment in addition to a brief demographic survey before the experiment. Annotators were allowed as long as needed to complete the task but were instructed to remain engaged with the experiment until completed.

Though this experiment's task remained relatively general, identifying multiple different highlighting "tasks" is important here, as some human annotators may highlight different content based on their interpretation of what highlights may help comprehension. Previous psychology research (Lindner and others, 1996; Fowler and Barker, 1974; Lorch Jr., 1989) leads us to believe there are many highlighting tasks, including content organization and novel concept retention, both of which would require more narrow experiment parameters and annotator pools than our human agreement task.

3.3 Agreement Task Results

Using Krippendorff Alpha scores (Krippendorff, 2011), we measured the agreement across the 40 annotators for each document. For visualization and scoring purposes, each document is represented as a binary mapping of annotator highlights (see Fig. 2). Because of the skew of non-highlighted to highlighted sentences, and the binary annotation method in the task, Krippendorff values most closely represent actual human agreement. All alpha scores returned $\alpha > 0$, and two high-performing documents

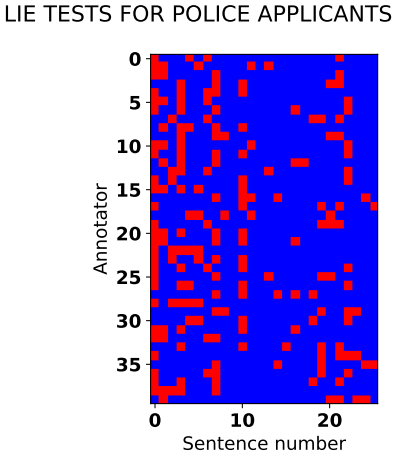

Figure 2: Binary mapping of annotator highlights in a single document where each row represents an annotator, and each column represents a sentence. Red cells are highlighted sentences, blue cells are unhighlighted sentences.

scored $\alpha > 0.2$, indicating "fair" agreement. The average score was $\alpha = 0.1596$. Though we used 40 annotators per document, Krippendorff values reached convergence after 20 annotators (see Fig. 3).

Though these values may seem relatively low, especially compared to those of a standard annotation task where trained annotators mark the corpus, these values indicate a trend towards possible agreement for highlighting tasks in part due to the multiple task dilemma. Because of the number of previously explored highlighting tasks and the relative generality of our own highlighting task, these agreement scores are likely to be diluted by multiple interpretations of the experiment definition. We believe that with more specific highlighting tasks (i.e., highlighting novel information, highlighting for information re-

Copyright 2018 by the authors

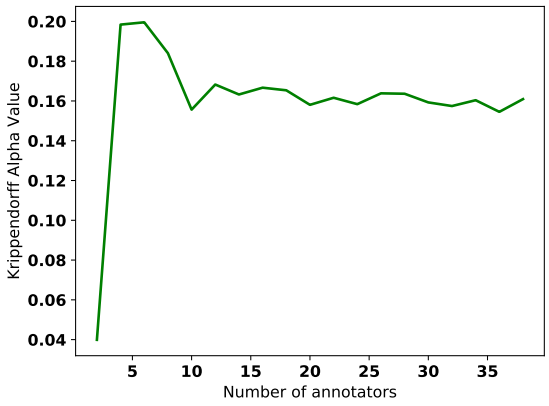

Figure 3: Convergence of Krippendorff Alpha scores reflecting the agreement between annotators as a function of the number of annotators.

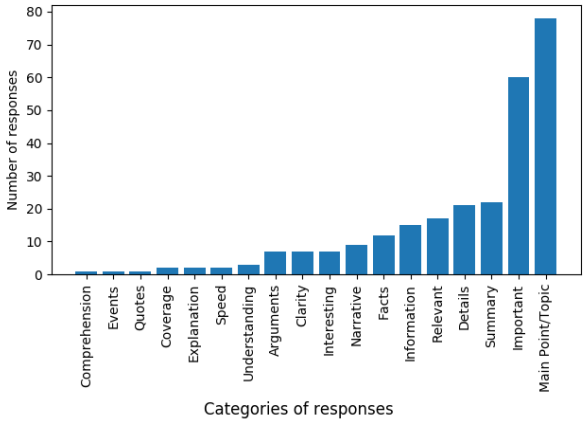

Figure 4: Initial paraphrased categories of user highlighting intentions for each document view.

trieval, etc.), these scores will represent higher levels of agreement.

The requirement that annotators highlight the same amount of sentences may also affect results. Lorch (1989) suggests that, in agreement with the von Restorff effect (von Restorff, 1933) which states that highlights stand out against a less crowded background, fewer highlights will more positively affect retention. It is possible that the exact number of highlights is an unnecessary constraint, however, it is important to note that this exercise examines annotators' sentence retrieval whereas this previous research evaluates the effect of *passive* highlights on readers.

3.4 Highlighting Intentions

We recognize that the low Krippendorff scores do not necessarily indicate strong highlighting agreement, but the idea that multiple highlighting intentions may dilute agreement metrics pushed us to explore methods of quantifying those potential intentions. In a separate but similar experiment, we asked 40 new Mechanical Turk annotators to complete the same agreement task with additional questions regarding highlighting intent. After each document, annotators were asked to consider the highlights they had just created and detail in a short response why they chose those sentences and what made them different from other sentences. Many annotators, in their own words, repeated the task instructions, though others provided in-depth details about individual sentences chosen.

In order to quantify these responses, we summarized each response by its main subject, represented in Fig. 4. While the vast majority of responses were characterized as "important" or pertaining to a "main topic" as expected from the experiment definition, many annotators also noted that their highlights summarized the content, contained details, or were especially relevant to the overall topic. Some annotators also argued that their highlights contained interesting facts or narratives that related to the main point of the article. This data supports our hypothesis that many annotators may highlight articles for many different reasons, which may impact agreement levels unless specific intentions are provided. And, even then, humans may not be very good at identifying the exact sentences that should be highlighted. However that doesn't necessarily mean humans will not agree on highlighted sentences when those highlights are already presented to them inline with the text.

4 Comparison Task

The original human annotation task showed that choosing highlights is subjective to some extent. In this section, we investigate whether annotators express clear preferences when presented documents with a few sentences highlighted using summarization models. To test this, we developed another interface that allows annotators to compare the output

Copyright 2018 by the authors

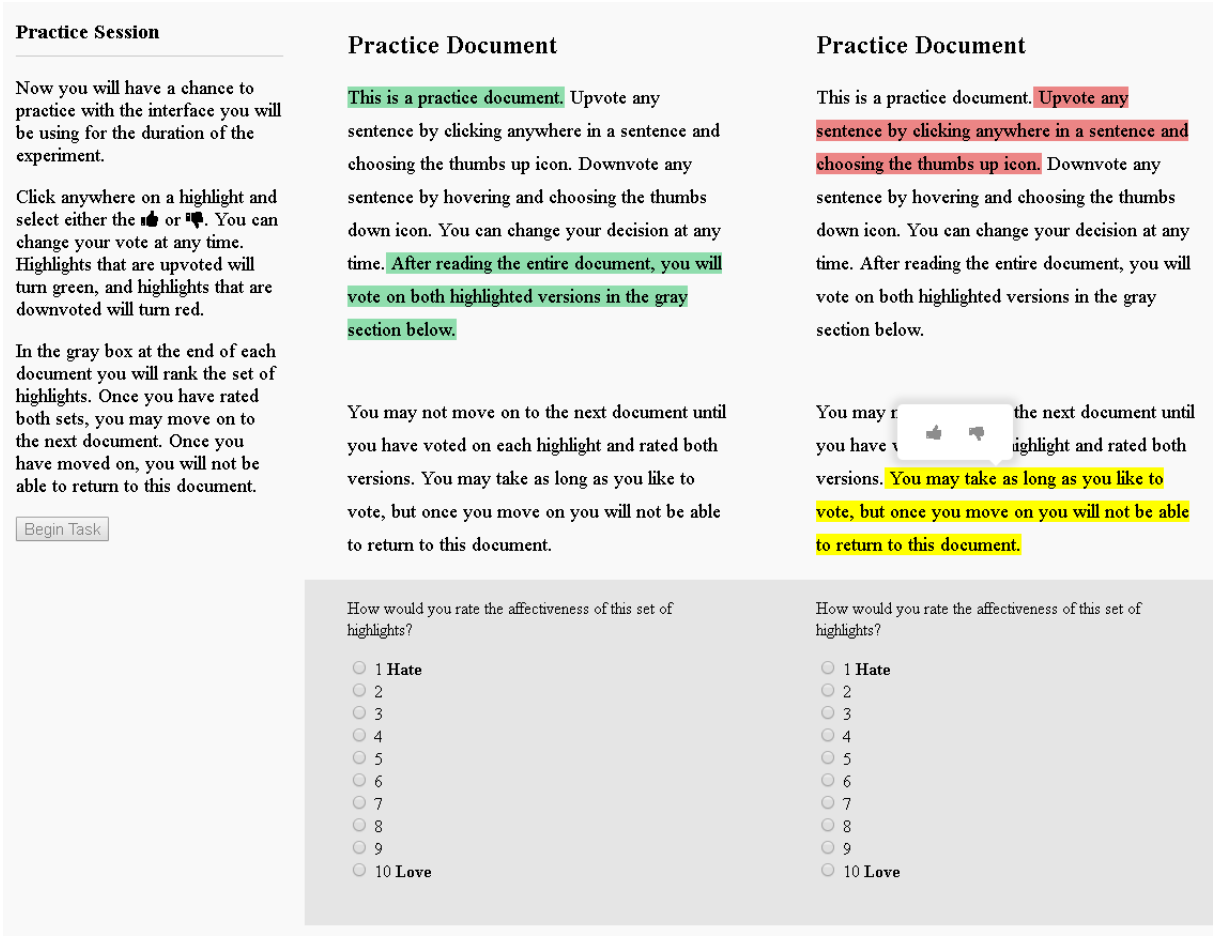

Figure 5: Comparison task tutorial with color-coded highlights and rating questionnaire at bottom.

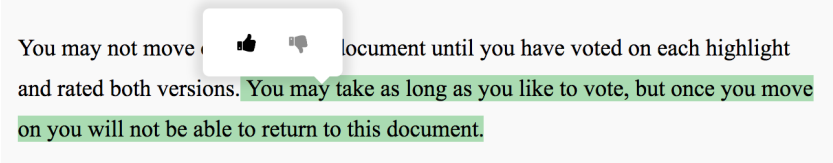

Figure 6: Voting options in comparison task. Participants are presented with two versions of the same text with different highlights. Participants must upvote (green) or downvote (red) each highlight, then give a global grade between 1 and 10 for each of the two versions.

of multiple highlighting models and provide feedback on the quality of each highlight.

4.1 Multi-Document Comparison Design

The comparison task is similar in style to the previous annotation task; though annotators were presented with two side-by-side versions of the same document for this task in order to be able to compare two highlighting systems, aesthetic details such as font, sizing, and distribution of the content and instruction panel remain the same (see Fig. 5).

To handle annotation of positive and negative votes on individual highlights, we introduced the "thumbs up" and "thumbs down" buttons, displayed after left clicking anywhere within the boundaries of a highlighted sentence (see Fig. 6). Annotators were asked to vote on every highlight displayed on the page, in addition to rating both versions of highlights on a one to ten scale before moving on. In

Copyright 2018 by the authors

an effort to avoid "lazy" annotation and elicit user preference, annotators were required to give the two versions different ratings, as raters tend to be more consistent in pairwise comparisons than when scoring directly (Agarwala, 2018).

4.2 Methodology

Using two batches, each consisting of 140 Mechanical Turk users with the same English fluency requirement as the human agreement task, annotators were instructed to "upvote" and "downvote" individual highlights that they believed help identify the main point(s) of the document. Annotators were shown two different highlighted versions, generated from a set of 5 models: the summarization models Recollect (Modani and others, 2015; Modani et al., 2016) and Sedona (Elhoseiny et al., 2016), the SMMRY summarizer (`smmry.com`), and two models derived from the data collected in the previous human agreement task representing the most common and least common human-selected highlights. Models were completely randomized and anonymized, both in location (e.g., left or right side of the content frame) and pairing.

Each batch of annotators worked on 5 documents of the same subset of 10 documents from the 2001 DUC data. Annotators had as long as needed to complete the task, as well as a brief demographics survey before the task. Similar to the human agreement task, a shortened version of the task definition remained in the left panel for the duration of the task. All annotators interacted with a controls tutorial before beginning the experiment. Annotators who worked on the same document sets from the human annotation task were ineligible for this task.

In addition to the demographic survey, annotators were asked to provide answers to a document type ranking question at the end of the task after considering the "best" versions of highlights they interacted with during the task. This data was used to gauge the models' impact on user preferences for highlighted documents.

4.3 Comparison Results

4.3.1 Human Evaluation Results

Results from this task reflect a clear preference towards human-generated highlights (see Fig. 7.).

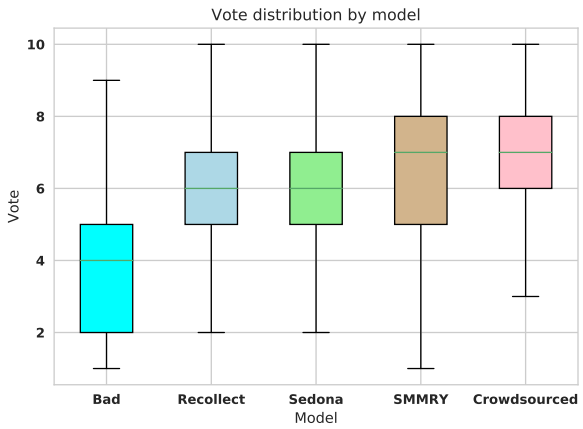

Figure 7: Boxplots representing the user vote distributions for each model.

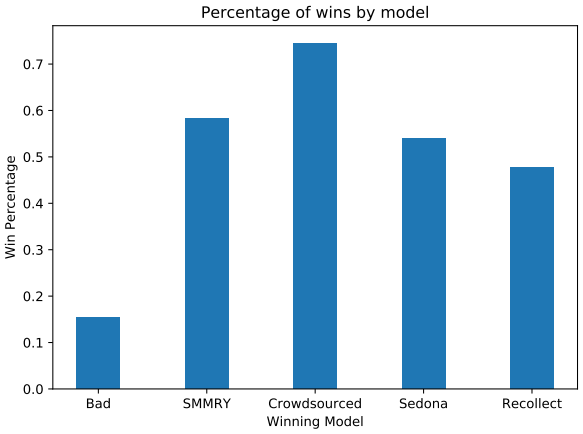

Figure 8: Percentage of "wins" by model, where a win is defined as the model with the higher vote score in each comparison.

As expected, votes for the least common highlights performed consistently lower than any of the other models ($\forall t, p, -25 < t < -14, p < 0.01$, Kruskal-Wallis h test $h = 515.6, p < 0.01$). Human-generated highlights performed significantly better ($\forall t, p, 6 < t < 25, p < 0.01$) than all other models, indicating that there must be information captured by humans that our standard summarization models do not identify. This is a clear signal that not only does there exist agreement among readers that "good" highlights exist, but also that highlighting may require more research as an independent area of study.

Results from the document type ranking question

Copyright 2018 by the authors

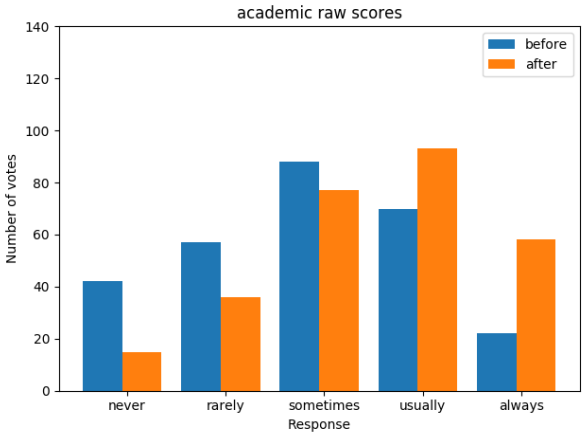

Figure 9: Before and after task document type ranking results for academic papers.

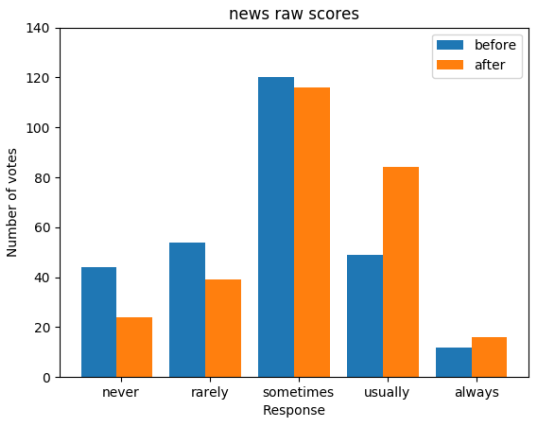

Figure 10: Before and after task document type ranking results for news articles.

reflected clear appreciation of highlights shown, despite annotators having potentially seen "bad" highlights. For academic papers and news articles especially, annotators seemed to show increased desire to see highlights after the task (see Fig. 9 and Fig. 10). We believe this is consistent with actual annotator belief rather than a confound of the task structure as results for fiction documents showed consistent disinterest across both before-task and after-task responses.

In Fig. 8, we present the percentage of times each model wins, normalized by the amount of views that particular model received. In Fig. 11, we expand the granularity of this data to show the distribution of differences between the winning and losing model

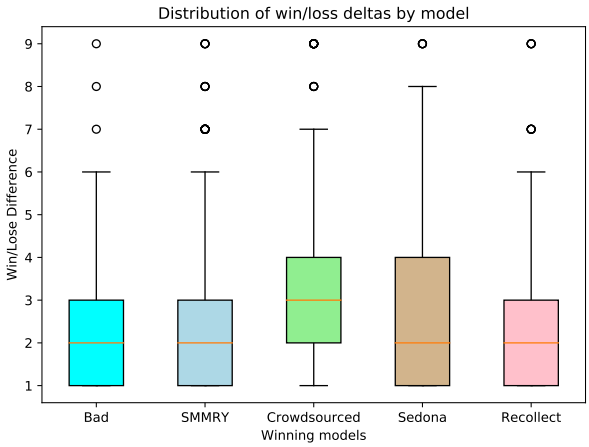

Figure 11: Distribution of win/loss deltas for each win by model.

for each comparison. Both visualizations further the same conclusions made above: the crowdsourced model consistently outperforms all competing models in every human evaluation-based metric we apply.

The relatively compact nature of the win/loss distributions in Fig. 11 gives an interesting hindsight on the annotator results. Ideally, we would see expansive differences between two models, where a user perceives a large difference between model A and model B. However, the figure indicates that in most cases, users only rank the two models within 1 and 3 ratings apart from each other.

Interestingly, the percentage of overlap with human highlights by each model does not reflect the model ordering seen in the above evaluations. When compared to the crowdsourced model, Recollect overlaps 27%, SMMRY 30%, and Sedona 35% of the time. This may be a reflection of small sample size, or it could reveal new insights of "good" highlights: perhaps some sentences are more salient to a human evaluator, such that some sentences should be weighted as more important to an evaluation metric than others. We would need a larger dataset and further exploration of this topic to confirm.

4.3.2 Automatic Evaluation Results

Human evaluations such as the one we have presented in the previous sections do not scale well. This section explores an automated metric for evaluating the quality of highlighted sentences. We con-

Copyright 2018 by the authors

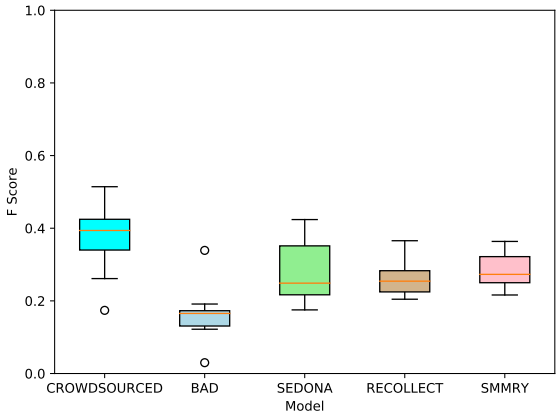

Figure 12: Distribution of ROUGE-1 F1-scores with stopword removal, using document respective DUC 50 word summaries.

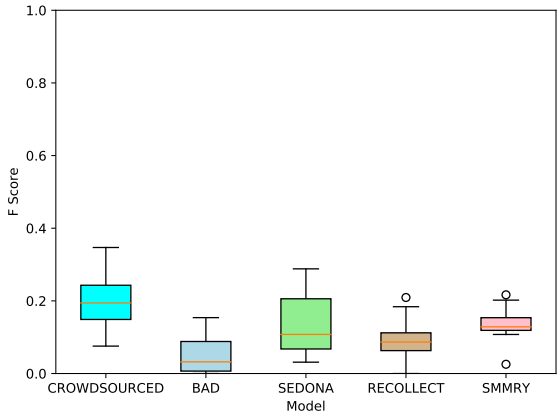

Figure 13: Distribution of ROUGE-2 F1-scores with stopword removal, using document respective DUC 50 word summaries.

sider the results of our human ratings and ROUGE metrics applied to this dataset. Since our documents come directly from the 2001 DUC dataset, each document has a matching 50-word human-generated and standardized summary, which we use as a gold version for ROUGE comparisons.

It is important to note that our system highlight sets may differ in number of sentences from the DUC gold files (DUC summaries are a standardized 50 words, whereas our models produce $n = .2L$ sentences, where L is the total number of sentences in the document. Because of this mismatch between summary lengths, recall would be a biased metric for our evaluation. Instead, we report here the distribution of precision scores according to the ROUGE metric.

ROUGE-1 results (see Fig. 12) show promising data, reflecting a similar model ranking pattern as our human voting evaluation. Again, the uncommon highlights, our "bad" model, show a significantly lower average ROUGE score than all other models, and the "good" crowdsourced model shows a significantly higher average ROUGE score in comparison to all other models ($\forall t, p, -7 < t < -5, p < 0.01$. See Fig. 12).

If higher ROUGE scores are indeed indicative of better user approval, then we should see this relative pattern occur across n-gram sizes. Unfortunately, this pattern deteriorates when using ROUGE-2, and, though our crowdsourced model still performs significantly better than all others with $p < 0.01$, the distance between the average F1-score of the uncommon highlights compresses, and no longer passes a paired t-test against both the Sedona and Recollect models at the same confidence interval ($t = 2.7, p = 0.02$ and $t = 2.3, p = 0.04$, respectively. See Fig. 13). This is a problem considering exactly how much different the "bad" model has been compared to other models in previous evaluations.

5 Conclusion

In this paper, we have explored the use of existing extractive summarization models for automatically generating highlights. The results of our experiments indicate that while the agreement on the choice of highlights to select is low, the positive results of the comparison task suggest that humans find machine generated highlights useful, which indicates that highlight generation is a valuable field of study.

Human evaluation is a necessary step in creating an automated metric to evaluate the quality of highlighting sets. While ROUGE scores show potential to be a valuable, inexpensive evaluation metric, it is clear that a reliable method requires more than n-gram overlap and will call for further investigation into what really differentiates human-produced highlights from our baseline models.

Copyright 2018 by the authors

References

Aseem Agarwala. 2018. Automatic photography with Google Clips. https://ai.googleblog.com/2018/05/automatic-photography-with-google-clips.html.

Dennis Baron. 2009. *A better pencil: Readers, writers, and the digital revolution*. Oxford University Press.

Hal Daumé III and Daniel Marcu. 2004. Generic sentence fusion is an ill-defined summarization task.

DUC. 2001. Document understanding conference.

Mohamed Elhoseiny, Scott Cohen, Walter Chang, Brian Price, and Ahmed Elgammal. 2016. Automatic annotation of structured facts in images. In *Proceedings of the 5th Workshop on Vision and Language*.

Robert L. Fowler and Anne S. Barker. 1974. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology*, 59(3):358–364.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, July.

Reinhard W. Lindner et al. 1996. Highlighting text as a study strategy: Beyond attentional focusing. In *Annual Meeting of the American Educational Research Association*.

R.F. Lorch Jr, E. Pugzles Lorch, and M.A. Klusewitz. 1995. Effects of typographical cues on reading and recall of text. *Contemporary Educational Psychology*, 20(1):51–64.

R. F. Lorch Jr. 1989. Text-signaling devices and their effects on reading and memory processes. *Educational Psychology Review*, 1(3):209–234.

Natwar Modani et al. 2015. Creating diverse product review summaries: a graph approach. *International Conference on Web Information Systems Engineering*.

Natwar Modani, Balaji Vasan Srinivasan, and Harsh Jhamtani. 2016. Generating multiple diverse summaries. *International Conference on Web Information Systems Engineering*.

A. Nenkova and K. MecKeown. 2012. A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer, Cham.

G.J. Rath, A. Resnick, and T.R. Savage. 1961. The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines. *Journal of the Association for Information Science and Technology*, 12(2):139–141.

H. von Restorff. 1933. Über die wirkung von bereichsbildungen im spurenfeld. *Psychologische Forschung*, 18(1):299–342.

J. K. Yogan, O. S. Goh, B. Halizah, H. C. Ngo, and C. Puspalata. 2016. A review on automatic text summarization approaches. *Journal of Computer Science*, 12(4):178–190.

Copyright 2018 by the authors