HIGHRES: Highlight-based Reference-less Evaluation of Summarization

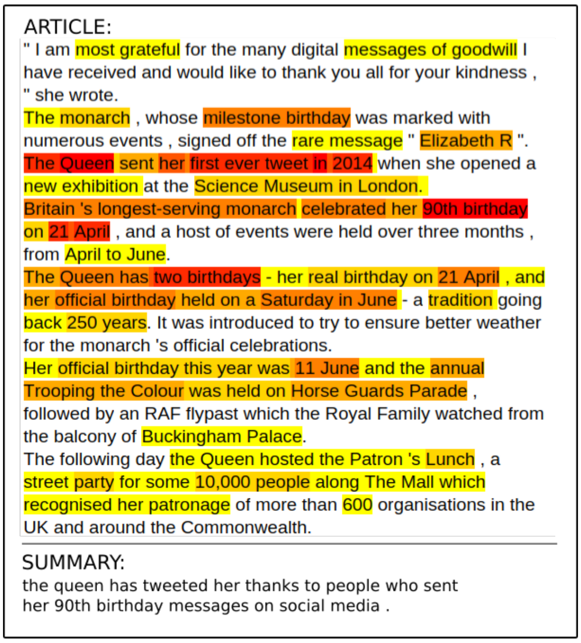
Hardy[1] **Shashi Narayan**[2*] **Andreas Vlachos**[1,3]
[1]Department of Computer Science, University of Sheffield [2]Google Research
[3]Department of Computer Science and Technology, University of Cambridge
hhardy2@sheffield.ac.uk, shashinarayan@google.com, av308@cam.ac.uk

Abstract

There has been substantial progress in summarization research enabled by the availability of novel, often large-scale, datasets and recent advances on neural network-based approaches. However, manual evaluation of the system generated summaries is inconsistent due to the difficulty the task poses to human non-expert readers. To address this issue, we propose a novel approach for manual evaluation, HIGHlight-based Reference-less Evaluation of Summarization (HIGHRES), in which summaries are assessed by multiple annotators against the source document via manually highlighted salient content in the latter. Thus summary assessment on the source document by human judges is facilitated, while the highlights can be used for evaluating multiple systems. To validate our approach we employ crowd-workers to augment with highlights a recently proposed dataset and compare two state-of-the-art systems. We demonstrate that HIGHRES improves inter-annotator agreement in comparison to using the source document directly, while they help emphasize differences among systems that would be ignored under other evaluation approaches.[1]

1 Introduction

Research in automatic summarization has made headway over the years with single document summarization as the front-runner due to the availability of large datasets (Sandhaus, 2008; Hermann et al., 2015; Narayan et al., 2018b) which has enabled the development of novel methods, many of them employing recent advances in neural networks (See et al., 2017; Narayan et al., 2018c; Pasunuru and Bansal, 2018, *inter alia*).



ARTICLE:
" I am most grateful for the many digital messages of goodwill I have received and would like to thank you all for your kindness , " she wrote.
The monarch , whose milestone birthday was marked with numerous events , signed off the rare message " Elizabeth R ".
The Queen sent her first ever tweet in 2014 when she opened a new exhibition at the Science Museum in London.
Britain 's longest-serving monarch celebrated her 90th birthday on 21 April , and a host of events were held over three months , from April to June.
The Queen has two birthdays - her real birthday on 21 April , and her official birthday held on a Saturday in June - a tradition going back 250 years. It was introduced to try to ensure better weather for the monarch 's official celebrations.
Her official birthday this year was 11 June and the annual Trooping the Colour was held on Horse Guards Parade , followed by an RAF flypast which the Royal Family watched from the balcony of Buckingham Palace.
The following day the Queen hosted the Patron 's Lunch , a street party for some 10,000 people along The Mall which recognised her patronage of more than 600 organisations in the UK and around the Commonwealth.

SUMMARY:
the queen has tweeted her thanks to people who sent her 90th birthday messages on social media .

Figure 1: Highlight-based evaluation of a summary. Annotators to evaluate a summary (bottom) against the highlighted source document (top) presented with a heat map marking the salient content in the document; the darker the colour, the more annotators deemed the highlighted text salient.

Measuring progress in summarization is difficult, as the task has as input a source document consisting of multiple sentences and methods need to generate a shorter text that expresses the salient information of the source fluently and succinctly. Thus there can be multiple equally good summaries for the same source document as not all salient information can fit in a given summary length, while even extractive methods that select complete sentences are not guaranteed to produce a coherent summary overall.

The most consistently used evaluation approach is comparison of the summaries produces against reference summaries via automatic measures such as ROUGE (Lin, 2004) and its variants. However,

[*]The work was primarily done while Shashi was still at School of Informatics, University of Edinburgh.

[1]Our dataset and code are available at https://github.com/sheffieldnlp/highres

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3381–3392
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

automatic measures are unlikely to be sufficient to measure performance in summarization (Schluter, 2017), also known for other tasks in which the goal is to generate natural language (Novikova et al., 2017). Furthermore, the datasets typically considered have a single reference summary, as obtaining multiple ones increases dataset creation cost, thus evaluation against them is likely to exhibit reference bias (Louis and Nenkova, 2013; Fomicheva and Specia, 2016), penalizing summaries containing salient content different from the reference.

For the above reasons manual evaluation is considered necessary for measuring progress in summarization. However, the intrinsic difficulty of the task has led to research without manual evaluation or only fluency being assessed manually. Those that conduct manual assessment of the content, typically use a single reference summary, either directly (Celikyilmaz et al., 2018; Tan et al., 2017) or through questions (Narayan et al., 2018b,c) and thus are also likely to exhibit reference bias.

In this paper we propose a novel approach for manual evaluation, HIGHlight-based Reference-less Evaluation of document Summarization (HIGHRES), in which a summary is assessed against the source document via manually highlighted salient content in the latter (see Figure 1 for an example). Our approach avoids reference bias, as the multiple highlights obtained help consider more content than what is contained in a single reference. The highlights are not dependent on the summaries being evaluated but only on the source documents, thus they are reusable across studies, and they can be crowd-sourced more effectively than actual summaries. Furthermore, we propose to evaluate the clarity of a summary separately from its fluency, as they are different dimensions. Finally, HIGHRES provides absolute instead of ranked evaluation, thus the assessment of a system can be conducted and interpreted without reference to other systems.

To validate our proposed approach we use the recently introduced eXtreme SUMmarization dataset (XSUM, Narayan et al., 2018b) to evaluate two state-of-the-art abstractive summarization methods, Pointer Generator Networks (See et al., 2017) and Topic-aware Convolutional Networks (Narayan et al., 2018b), using crowd-sourcing for both highlight annotation and quality judgments.

We demonstrate that HIGHRES improves inter-annotator agreement in comparison to using the source document directly, while they help emphasize differences among systems that would be ignored under other evaluation approaches, including reference-based evaluation. Furthermore, we show that the clarity metric from the DUC (Dang, 2005) must be measured separately from "fluency", as judgments for them had low correlation. Finally, we make the highlighted XSUM dataset, codebase to replicate the crowd-sourcing experiments and all other materials produced in our study publicly available.

2 Literature Review

In recent years, summarization literature has investigated different means of conducting manual evaluation. We study a sample of 26 recent papers from major ACL conferences and outline the trends of manual evaluation in summarization in Table 1. From 26 papers, 11 papers (e.g., See et al., 2017; Kedzie et al., 2018; Cao et al., 2018) did not conduct any manual evaluation. Following the Document Understanding Conference (DUC, Dang, 2005), a majority of work has focused on evaluating the content and the linguistic quality of summaries (Nenkova, 2005). However, there seems to be a lack of consensus on how a summary should be evaluated: (i) Should it be evaluated relative to other summaries or standalone in absolute terms? and (ii) What would be a good source of comparison: the input document or the reference summary? The disagreements on these issues result in authors evaluating their summaries often (11 out of 26 papers) using automatic measures such as ROUGE (Lin, 2004) despite of its limitations (Schluter, 2017). In what follows, we discuss previously proposed approaches along three axes: evaluation metrics, relative vs. absolute, and the choice of reference.

Evaluation Metrics Despite differences in the exact definitions, the majority (e.g., Hsu et al., 2018; Celikyilmaz et al., 2018; Narayan et al., 2018b; Chen and Bansal, 2018; Peyrard and Gurevych, 2018) agree on both or either one of two broad quality definitions: *coverage* determines how much of the salient content of the source document is captured in the summary, and *informativeness*, how much of the content captured in the summary is salient with regards to the original document. These measures correspond to "*recall*" and "*precision*" metrics respectively in Table 1, notions that are commonly used

Systems	No Manual Eval	Pyramid	QA	Correctness	Fluency	Clarity	Recall	Precision	Absolute	Relative	With Reference	With Document	With Ref. & Doc.
See et al. (2017)	✓												
Lin et al. (2018)	✓												
Cohan et al. (2018)	✓												
Liao et al. (2018)	✓												
Kedzie et al. (2018)	✓												
Amplayo et al. (2018)	✓												
Jadhav and Rajan (2018)	✓												
Li et al. (2018a)	✓												
Pasunuru and Bansal (2018)	✓												
Cao et al. (2018)	✓												
Sakaue et al. (2018)	✓												
Celikyilmaz et al. (2018)							✓	✓	✓	✓	✓	✓	
Chen and Bansal (2018)				✓			✓	✓		✓			✓
Guo et al. (2018)					✓			✓		✓			✓
Hardy and Vlachos (2018)					✓				✓				
Hsu et al. (2018)					✓		✓	✓	✓			✓	
Krishna and Srinivasan (2018)							✓			✓		✓	
Kryściński et al. (2018)					✓		✓		✓			✓	
Li et al. (2018b)				✓					✓				
Narayan et al. (2018a)					✓					✓		✓	
Narayan et al. (2018b)			✓		✓		✓		✓	✓	✓	✓	
Narayan et al. (2018c)			✓		✓		✓		✓	✓	✓	✓	
Peyrard and Gurevych (2018)							✓	✓	✓		✓		
ShafieiBavani et al. (2018)		✓											
Song et al. (2018)				✓	✓		✓		✓			✓	
Yang et al. (2017)							✓			✓	✓		
HIGHRES (ours)					✓	✓	✓	✓	✓			✓	

Table 1: Overview of manual evaluations conducted in recent summarization systems. We categorize them in four dimensions: the first columns presents papers that do not report on human evaluation; the second column identifies matrices used for evaluating content ("*Pyramid*", "*QA*", "*Correctness*", "*Recall*" and "*Precision*") and quality ("*Clarity*", "*Fluency*") of summaries; the third column focuses if the system ranking reported by humans on content evaluation were "*Absolute*" or "*Relative*"; and finally, the fourth column evaluates if summaries were evaluated against the input document ("*With Document*"), the reference summary ("*With Reference*") or both ("*With Ref. & Doc.*").

in information retrieval and information extraction literature. Clarke and Lapata (2010) proposed a question-answering based approach to improve the agreement among human evaluations for the quality of summary content, which was recently employed by Narayan et al. (2018b) and Narayan et al. (2018c) (QA in Table 1). In this approach, questions were created first from the reference summary and then the system summaries were judged with regards to whether they enabled humans to answer those questions correctly. ShafieiBavani et al. (2018), on the other hand, used the "Pyramid" method (Nenkova and Passonneau, 2004) which requires summaries to be annotated by experts for salient information. A similar evaluation approach is the factoids analysis by Teufel and Van Halteren (2004) which evaluates the system summary against factoids, a representation based on atomic units of information, that are extracted from multiple gold summaries. However, as in the case of the "Pyramid" method, extracting factoids requires experts annotators. Finally, a small number of work evaluates the "Correctness" (Chen and Bansal, 2018; Li et al., 2018b; Chen and Bansal, 2018) of the summary, similar to fact checking (Vlachos and Riedel, 2014), which can be a challenging task in its own right.

The linguistic quality of a summary encompasses many different qualities such as fluency, grammatically, readability, formatting, naturalness and coherence. Most recent work uses a single human judgment to capture all linguistic qualities of the summary (Hsu et al., 2018; Kryściński et al., 2018; Narayan et al., 2018b; Song et al., 2018; Guo et al., 2018); we group them under "Fluency" in Table 1 with an exception of "Clarity" which was evaluated in the DUC evaluation campaigns (Dang, 2005). The "Clarity" metric puts emphasis in easy identification of noun and pronoun phrases in the summary which is a different dimension than "Fluency", as a summary may be fluent but difficult to be understood due to poor clarity.

Absolute vs Relative Summary Ranking. In relative assessment of summarization, annotators are shown two or more summaries and are asked to rank them according to the dimension at question (Yang et al., 2017; Chen and Bansal, 2018; Narayan et al., 2018a; Guo et al., 2018; Krishna and Srinivasan, 2018). The relative assessment is often done using the paired comparison (Thurstone, 1994) or the best-worst scaling (Woodworth and G, 1991; Louviere et al., 2015), to improve inter-annotator agreement. On the other hand, absolute assessment of summarization (Li et al., 2018b; Song et al., 2018; Kryściński et al., 2018; Hsu et al., 2018; Hardy and Vlachos, 2018) is often done using the Likert rating scale (Likert, 1932) where a summary is assessed on a numerical scale. Absolute assessment was also employed in combination with the question answering approach for content evaluation (Narayan et al., 2018b; Mendes et al., 2019). Both approaches, relative ranking and absolute assessment, have been investigated extensively in Machine Translation (Bojar et al., 2016, 2017). Absolute assessment correlates highly with the relative assessment without the bias introduced by having a simultaneous assessment of several models (Bojar et al., 2011).

Choice of Reference. The most convenient way to evaluate a system summary is to assess it against the reference summary (Celikyilmaz et al., 2018; Yang et al., 2017; Peyrard and Gurevych, 2018), as this typically requires less effort than reading the source document. The question answering approach of Narayan et al. (2018b,c) also falls in this category, as the questions were written using the reference summary. However, summarization datasets are limited to a single reference summary per document (Sandhaus, 2008; Hermann et al., 2015; Grusky et al., 2018; Narayan et al., 2018b) thus evaluations using them is prone to reference bias (Louis and Nenkova, 2013), also a known issue in machine translation evaluation (Fomicheva and Specia, 2016). A circumvention for this issue is to evaluate it against the source document (Song et al., 2018; Narayan et al., 2018a; Hsu et al., 2018; Kryściński et al., 2018), asking judges to assess the summary after reading the source document. However this requires more effort and is known to lead to low inter-annotator agreement (Nenkova and Passonneau, 2004).

3 HIGHRES

Our novel highlight-based reference-less evaluation does not suffer from reference bias as a summary is assessed against the source document with manually highlighted salient content. These highlights are crowd-sourced effectively without the need of expert annotators as required by the Pyramid method (Nenkova and Passonneau, 2004) or to generate reference summaries. Our approach improves over the "Correctness" or "Fluency" only measure for summarization by taking salience into account. Finally, the assessment of summaries against the document with highlighted pertinent content facilitates an absolute evaluation of summaries with high inter-annotator agreement.

Our evaluation framework comprises three main components: document highlight annotation, highlight-based content evaluation, and clarity and fluency evaluation. The second component, which evaluates the notions of "Precision" and "Recall" requires the highlights from the first one to be conducted. However, the highlight annotation needs to happen only once per document, and it can be reused to evaluate many system summaries, unlike the Pyramid approach (Nenkova and Passonneau, 2004) that requires additional expert annotation for every system summary being evaluated. The third component is independent of the others and can be run in isolation. In all components we employ crowd-workers as human judges, and implement appropriate sanity checking mechanisms to ensure good quality judgements. Finally, we present an extended version of ROUGE (Lin, 2004) that utilizes the highlights to evaluate system summaries against the document; this demonstrates another use of the highlights for summarization evaluation.

3.1 Highlight Annotation

In this part, we ask human judges to read the source document and then highlight words or phrases that are considered salient. Each judge is allowed to highlight parts of the text at any granularity, from single words to complete sentences or even paragraphs. However we enforce a limit in the number of words to $\mathcal{K}$ that can be highlighted in total by a judge in a document, corresponding to the length of the summary expected. By employing multiple judges per document who are restricted in the amount of text that can be highlighted we expect to have a more diverse and focused highlight from multiple judges which cover different viewpoints of the article. To ensure that each highlight is reliable, we performed a sanity check at the end of the task where we ask the judges to answer a True/False question based on the article. We rejected all annotations that failed to correctly answer the sanity check question.

3.2 Highlight-based Content Evaluation

In this component, we present human judges a document that has been highlighted using heatmap coloring and a summary to assess. We ask our judges to assess the summary for (i) *'All important information is present in the summary'* and (ii) *'Only important information is in the summary.'* The first one is the recall (content coverage) measure and the second, the precision (informativeness) measure. All the ratings were collected on a 1-100 Likert scale (Likert, 1932). Figure 2 shows the content evaluation user interface where salient parts of the document are highlighted. As with the highlight annotation, we performed the same form of sanity check to the one in the highlight annotation task.

3.3 Clarity and Fluency Evaluation

In this part, we give the judges only the summary and ask them to rate it on clarity and fluency. For *clarity*, each judge is asked whether the summary is easy to be understood, i.e. there should be no difficulties in identifying the referents of the noun phrases (every noun/place/event should be well-specified) or understanding the meaning of the sentence. For *fluency*, each judge is asked whether the summary sounds natural and has no grammatical problems. While fluency is often evaluated in recent work, clarity, while first introduced in DUC evaluations, has recently been ignored in manual evaluation, despite that it captures a different dimension of summarization quality.

To ensure that the judgments for clarity and fluency are not affected by each other (poor fluency can affect clarity, but a summary can have perfect fluency but low clarity), we evaluate each metric separately. We ask the judges to evaluate multiple summaries per task with each dimension in its own screen. For sanity checking, we insert three artificial summaries of different quality (good, mediocre and bad summaries). The good summary is the unedited one, while the others are generated from sentences randomly sampled from the source document. For the mediocre summary, some words are edited to introduce some grammatical or syntactic errors while for the bad summary, the words are further scrambled. We reject judgements that failed to pass this criteria: bad $<$ mediocre $<$ good.

3.4 Highlight-based ROUGE Evaluation

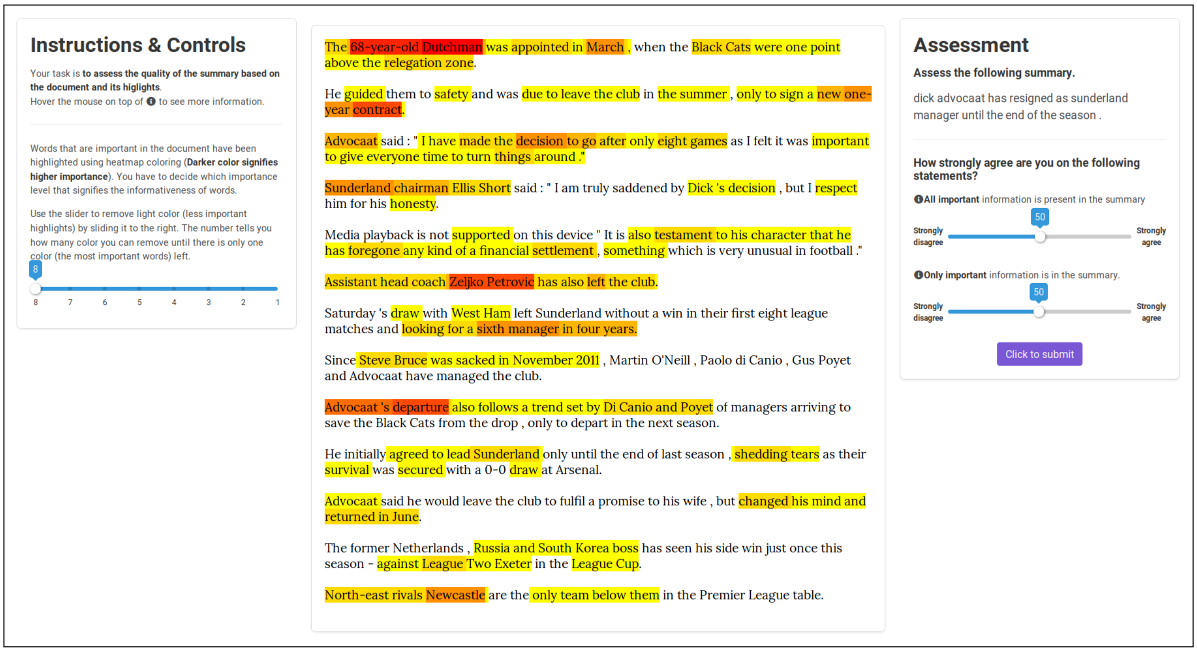
Our Highlight-based ROUGE (we refer to it as HROUGE) formulation is similar to the original ROUGE with the difference that the n-grams are weighted by the number of times they were highlighted. One benefit of HROUGE is that it introduces saliency into the calculation without being reference-based as in ROUGE. Implicitly HROUGE considers multiple summaries as the highlights are obtained from multiple workers.

Given a document $\mathcal{D}$ as a sequence of m tokens $\{w_1, \ldots, w_m\}$, annotated with $\mathcal{N}$ highlights, we define the weight $\beta_g^n \in [0,1]$ for an n-gram g as:

$$\beta_g^n = \frac{\displaystyle\sum_{i=1}^{m-(n-1)} \left[\frac{\sum_{j=i}^{i+n-1} \frac{\mathrm{NumH}(w_j)}{\mathcal{N}}}{n} \right]_{w_{i:i+n-1}==g}}{\displaystyle\sum_{i=1}^{m-(n-1)} [1]_{w_{i:i+n-1}==g}}$$

where, $[x]_y$ is an indicator function which returns x if y is true and 0, otherwise. $\mathrm{NumH}(w_j) = \sum_{k=1}^{\mathcal{N}} \frac{\mathrm{len}(H_k)}{\mathcal{K}} [1]_{w_j \in H_k}$ is a function which returns the number of times word w_j is highlighted by the annotators out of $\mathcal{N}$ times weighted by the lengths of their highlights; H_k is the highlighted text by the k-th annotator and $\mathcal{K}$ is the maximum allowed length of the highlighted text (see Section 3.1). $\mathrm{NumH}(w_j)$ gives less importance to annotators with highlights with few words. In principle, if an n-gram is highlighted by every crowd-worker and the length of the highlight of each crowd-worker is $\mathcal{K}$, the n-gram g will have a maximum weight of $\beta_g^n = 1$.

The HROUGE scores for a summary $\mathcal{S}$ can then



Instructions & Controls

Your task is **to assess the quality of the summary based on the document and its higlights**.
Hover the mouse on top of ⓘ to see more information.

Words that are important in the document have been highlighted using heatmap coloring (**Darker color signifies higher importance**). You have to decide which importance level that signifies the informativeness of words.

Use the slider to remove light color (less important highlights) by sliding it to the right. The number tells you how many color you can remove until there is only one color (the most important words) left.

The 68-year-old Dutchman was appointed in March , when the Black Cats were one point above the relegation zone.

He guided them to safety and was due to leave the club in the summer , only to sign a new one-year contract.

Advocaat said : " I have made the decision to go after only eight games as I felt it was important to give everyone time to turn things around ."

Sunderland chairman Ellis Short said : " I am truly saddened by Dick 's decision , but I respect him for his honesty.

Media playback is not supported on this device " It is also testament to his character that he has foregone any kind of a financial settlement , something which is very unusual in football ."

Assistant head coach Zeljko Petrovic has also left the club.

Saturday 's draw with West Ham left Sunderland without a win in their first eight league matches and looking for a sixth manager in four years.

Since Steve Bruce was sacked in November 2011 , Martin O'Neill , Paolo di Canio , Gus Poyet and Advocaat have managed the club.

Advocaat 's departure also follows a trend set by Di Canio and Poyet of managers arriving to save the Black Cats from the drop , only to depart in the next season.

He initially agreed to lead Sunderland only until the end of last season , shedding tears as their survival was secured with a 0-0 draw at Arsenal.

Advocaat said he would leave the club to fulfil a promise to his wife , but changed his mind and returned in June.

The former Netherlands , Russia and South Korea boss has seen his side win just once this season - against League Two Exeter in the League Cup.

North-east rivals Newcastle are the only team below them in the Premier League table.

Assessment

Assess the following summary.

dick advocaat has resigned as sunderland manager until the end of the season .

How strongly agree are you on the following statements?

ⓘ**All important** information is present in the summary

ⓘ**Only important** information is in the summary.

Figure 2: The UI for content evaluation with highlight. Judges are given an article with important words highlighted using heat map. Judges can also remove less important highlight color by sliding the scroller at the left of the page. At the right of the page judges give the recall and precision assessment by sliding the scroller from 1 to 100 based on the given summary quality.

be defined as:

$$\mathrm{HR}^n_{\mathrm{rec}} = \frac{\displaystyle\sum_{g \in n\text{-gram}(\mathcal{S})} \beta^n_g \operatorname{count}(g, \mathcal{D} \cap \mathcal{S})}{\displaystyle\sum_{g \in n\text{-gram}(\mathcal{D})} \beta^n_g \operatorname{count}(g, \mathcal{D})}$$

$$\mathrm{HR}^n_{\mathrm{pre}} = \frac{\displaystyle\sum_{g \in n\text{-gram}(\mathcal{S})} \beta^n_g \operatorname{count}(g, \mathcal{D} \cap \mathcal{S})}{\displaystyle\sum_{g \in n\text{-gram}(\mathcal{S})} \operatorname{count}(g, \mathcal{S})}$$

$\mathrm{HR}^n_{\mathrm{rec}}$ and $\mathrm{HR}^n_{\mathrm{pre}}$ are the HROUGE recall and precision scores; $\operatorname{count}(g, \mathcal{X})$ is the maximum number of n-gram g occurring in the text $\mathcal{X}$. The weight in the denominator of $\mathrm{HR}^n_{\mathrm{pre}}$ is uniform ($\beta^n_g = 1$) for all g because if we weighted according to the highlights, words in the summary that are not highlighted in the original document would be ignored. This would result in $\mathrm{HR}^n_{\mathrm{pre}}$ not penalizing summaries for containing words that are likely to be irrelevant as they do not appear in the highlights of the document. It is important to note HROUGE has an important limitation in that it penalizes abstractive summaries that do not reuse words from the original document. This is similar to ROUGE penalizing summaries for not reusing words from the reference summaries, however the highlights allow us to implicitly consider multiple references without having to actually obtain them.

4 Summarization Dataset and Models

We use the extreme summarization dataset (XSUM, Narayan et al., 2018b)[2] which comprises BBC articles paired with their single-sentence summaries, provided by the journalists writing the articles. The summary in the XSUM dataset demonstrates a larger number of novel n-grams compared to other popular datasets such as CNN/DailyMail (Hermann et al., 2015) or NY Times (Sandhaus, 2008) as such it is suitable to be used for our experiment since the more abstractive nature of the summary renders automatic methods such as ROUGE less accurate as they rely on string matching, and thus calls for human evaluation for more accurate system comparisons. Following Narayan et al. (2018b), we didn't use the whole test set portion, but sampled 50 articles from it for our highlight-based evaluation.

We assessed summaries from two state-of-the-art abstractive summarization systems using our highlight-based evaluation: (i) the Pointer-Generator model (PTGEN) introduced by See et al. (2017) is an RNN-based abstractive systems which allows to copy words from the source text,

[2] https://github.com/EdinburghNLP/XSum

and (ii) the Topic-aware Convolutional Sequence to Sequence model (TConvS2S) introduced by Narayan et al. (2018b) is an abstractive model which is conditioned on the article's topics and based entirely on Convolutional Neural Networks. We used the pre-trained models[3] provided by the authors to obtain summaries from both systems for the documents in our test set.

5 Experiments and Results

All of our experiments are done using the Amazon Mechanical Turk platform.We develop three types of Human Intelligence Tasks (HITs): highlight annotation, highlight-based content evaluation, and fluency and clarity evaluation. In addition, we elicited human judgments for content evaluation in two more ways: we assessed system summaries against the original document (without highlights) and against the reference summary. The latter two experiments are intended as the comparison for our proposed highlight-based content evaluation.

5.1 Highlight Annotation

We collected highlight annotations from 10 different participants for each of 50 articles. For each annotation, we set $\mathcal{K}$, the maximum number of words to highlight, to 30. Our choice reflects the average length (24 words) of reference summaries in the XSum dataset. To facilitate the annotation of BBC news articles with highlights, we asked our participants to adapt the 5W1H (Who, What, When, Where, Why and How) principle (Robertson, 1946) that is a common practice in journalism. The participants however were not obliged to follow this principle and were free to highlight content as they deem fit.

The resulting annotation exhibits a substantial amount of variance, confirming the intuition that different participants are not expected to agree entirely on what is salient in a document. On average, the union of the highlights from 10 annotators covered 38.21% per article and 33.77% of the highlights occurred at the second half of the article. This shows that the judges did not focus only on the beginning of the documents but annotated all across the document.

Using Fleiss Kappa (Fleiss, 1971) on the binary labels provided by each judge on each word (highlighted or not) we obtained an average agreement of 0.19 for the 50 articles considered. The low agreement score does not indicate a poor annotation process necessarily; we argue that this is primarily due to the annotators having different opinions on which parts of an article are salient. The article with the highest agreement (0.32) has more focused highlights, whereas the article with the lowest agreement (0.04) has highlights spread all over (both articles can be seen in the supplementary materials). Interestingly, the reference summary on the highest agreement article appears to be more informative of its content when the annotator agreement is high; the reference summary on the lowest agreement article is more indicative, i.e., it does not contain any informative content from the article but only to inform the reader about the article's topic and scope. These results confirm that the annotation behaviour originates from the nature of the document and the summary it requires, and validates our highlight annotation setup.

Model	Highlight -based		Non Highlight-based		Reference -based	
	Prec	Rec	Prec	Rec	Prec	Rec
TConvS2S	57.42	49.95	52.55	41.04	46.75	36.45
PtGen	50.94	44.41	48.57	39.21	44.24	38.24
Reference	67.90	56.83	66.01	52.45	—	—

Table 2: Results of content evaluation of summaries against documents with highlights, documents without highlights and reference summaries.

Model	Highlight-based		Non Highlight-based	
	Prec	Rec	Prec	Rec
TConvS2S	0.67	0.80	0.75	0.83
PtGen	0.73	0.86	0.73	0.90
Reference	0.49	0.63	0.48	0.67

Table 3: Coefficient of variation (lower is better) for evaluating summaries against documents with and without highlights.

5.2 Content Evaluation of Summaries

We assessed the summaries against (i) documents with highlights (Highlight-based), (ii) original documents without highlights (Non Highlight-based) and (iii) reference summaries (Reference-based). For each setup, we collected judgments from 3 different participants for each model summary. Table 2 and 3 presents our results.

Both the highlight-based and non-highlight based assessment of summaries agree on the ranking among TConvS2S, PtGen and Reference. Perhaps unsurprisingly human-authored

[3] Both models were trained using the standard cross-entropy loss to maximize the likelihood of the reference summary given the document.

summaries were considered best, whereas, TConvS2S was ranked 2nd, followed by PtGen. However, the performance difference in TConvS2S and PtGen is greatly amplified when they are evaluated against document with highlights (6.48 and 5.54 Precision and Recall points) compared to when evaluated against the original documents (3.98 and 1.83 Precision and Recall points). The performance difference is lowest when they are evaluated against the reference summary (2.51 and -1.79 Precision and Recall points). The superiority of TConvS2S is expected; TConvS2S is better than PtGen for recognizing pertinent content and generating informative summaries due to its ability to represent high-level document knowledge in terms of topics and long-range dependencies (Narayan et al., 2018b).

We further measured the agreement among the judges using the coefficient of variation (Everitt, 2006) from the aggregated results. It is defined as the ratio between the sample standard deviation and sample mean. It is a scale-free metric, i.e. its results are comparable across measurements of different magnitude. Since, our sample size is small (3 judgements per summary), we use the unbiased version (Sokal and Rohlf, 1995) as $cv = (1+\frac{1}{4n})\frac{\sigma}{\bar{x}}$, where σ is the standard deviation, n is the number of sample, and $\bar{x}$ is the mean.

We found that the highlight-based assessment in general has lower variation among judges than the non-highlight based or reference-based assessment. The assessment of TConvS2S summaries achieves 0.67 and 0.80 of Precision and Recall cv points which are 0.08 and 0.03 points below when they are assessed against documents with no highlights, respectively. We see a similar pattern in Recall on the assessment of PtGen summaries. Our results demonstrate that the highlight-based assessment of abstractive systems improve agreement among judges compared to when they are assessed against the documents without highlights or the reference summaries. The assessment of human-authored summaries does not seem to follow this trend, we report a mixed results (0.49 vs 0.48 for precision and 0.63 vs 0.67 for recall) when they are evaluated with and without the highlights.

Model	Fluency	Clarity
TConvS2S	69.51	67.19
PtGen	55.24	52.49
Reference	77.03	75.83

Table 4: Mean ”Fluency” and ”Clarity” scores for TConvS2S , PtGen and Reference summaries. All the ratings were collected on a 1-100 Likert scale.

Model	Unigram		Bigram	
	Prec	Rec	Prec	Rec
ROUGE (Original document)				
TConvS2S	**77.17**	4.20	26.12	1.21
PtGen	77.09	**4.99**	**28.75**	**1.64**
Reference	73.65	4.42	22.42	1.17
HROUGE (Highlights from the document)				
TConvS2S	**7.94**	5.42	3.30	2.11
PtGen	7.90	**6.46**	**3.37**	**2.64**
Reference	7.31	5.73	2.39	1.84

Table 5: HROUGE-1 (unigram) and HROUGE-2 (bigram) precision, and recall scores for TConvS2S , PtGen and Reference summaries.

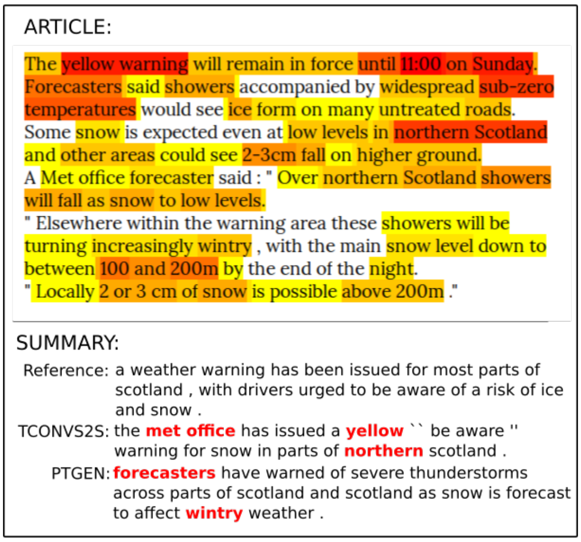
5.3 Clarity and Fluency Evaluation

Table 4 shows the results of our fluency and clarity evaluations. Similar to our highlight-based content evaluation, human-authored summaries were considered best, whereas TConvS2S was ranked 2nd followed by PtGen, on both measures. The Pearson correlation between fluency and clarity evaluation is 0.68 which shows a weak correlation; it confirms our hypothesis that the ”clarity” captures different aspects from ”fluency” and they should not be combined as it is commonly done.

5.4 Highlight-based ROUGE Evaluation

Table 5 presents our HROUGE results assessing TConvS2S , PtGen and Reference summaries with the highlights. To compare, we also report ROUGE results assessing these summaries against the original document without highlights. In the latter case, HROUGE becomes the standard ROUGE metric with $\beta_g^n = 1$ for all n-grams g.

Both ROUGE and HROUGE favour method of copying content from the original document and penalizes abstractive methods, thus it is not surprising that PtGen is superior to TConvS2S, as the former has an explicit copy mechanism. The fact that PtGen is better in terms of HROUGE is also an evidence that the copying done by PtGen selects salient content, thus confirming that the copying mechanism works as intended. When comparing the reference summaries against the original documents, both ROUGE and HROUGE confirm that the reference summaries are rather



ARTICLE:

The yellow warning will remain in force until 11:00 on Sunday. Forecasters said showers accompanied by widespread sub-zero temperatures would see ice form on many untreated roads. Some snow is expected even at low levels in northern Scotland and other areas could see 2-3cm fall on higher ground. A Met office forecaster said : " Over northern Scotland showers will fall as snow to low levels. " Elsewhere within the warning area these showers will be turning increasingly wintry , with the main snow level down to between 100 and 200m by the end of the night. " Locally 2 or 3 cm of snow is possible above 200m ."

SUMMARY:

Reference: a weather warning has been issued for most parts of scotland , with drivers urged to be aware of a risk of ice and snow .

TCONVS2S: the **met office** has issued a **yellow** `` be aware '' warning for snow in parts of **northern** scotland .

PTGEN: **forecasters** have warned of severe thunderstorms across parts of scotland and scotland as snow is forecast to affect **wintry** weather .

Figure 3: Highlighted article, reference summary, and summaries generated by TConvS2S and PtGen. Words in red in the system summaries are highlighted in the article but do not appear in the reference.

abstractive as reported by Narayan et al. (2018b), and they in fact score below the system summaries. Recall scores are very low in all cases which is expected, since the 10 highlights obtained per document or the documents themselves, taken together, are much longer than any of the summaries.

6 Qualitative Analysis

HighRES eliminates reference bias. The example presented in Figure 3 demonstrates how our highlight-based evaluation eliminates reference bias in summarization evaluation. The summaries generated by TConvS2S and PtGen are able to capture the essence of the document, however, there are phrases in these summaries that do not occur in the reference summary. A reference-based evaluation would fail to give a reasonable score to these system summaries. The HighRES however, would enable the judges to better evaluate the summaries without any reference bias.

Fluency vs Clarity. Example in Table 6 shows disagreements between fluency and clarity scores for different summaries of the same article. From the example, we can see that the TConvS2S summary is fluent but is not easily understood in the context of 'the duration of resignation', while the PtGen summary has word duplication which lower the fluency and also lacking clarity due to several unclear words.

Model	Summary Text	Fluency	Clarity
TConvS2S	dick advocaat has resigned as sunderland manager *until the end of the season* .	92.80	44.33
PtGen	sunderland have appointed *former sunderland boss* dick advocaat as manager *at the end of the season* to sign a *new deal* .	41.33	6.00

Table 6: TConvS2S and PtGen showing a disagreement between fluency and clarity scores. We italicized words that are not clear in the summaries.

7 Conclusion and Future Work

In this paper we introduced the HIGHlight-based Reference-less Evaluation Summarization (HighRES) framework for manual evaluation. The proposed framework avoids reference bias and provides absolute instead of ranked evaluation of the systems. Our experiments show that HighRES lowers the variability of the judges' content assessment, while helping expose the differences between systems. We also showed that by evaluating clarity we are able to capture a different dimension of summarization quality that is not captured by the commonly used fluency. We believe that our highlight-based evaluation is an ideal setup of abstractive summarization for three reasons: (i) highlights can be crowd sourced effectively without expert annotations, (ii) it avoids reference bias and (iii) it is not limited by n-gram overlap. In future work, we would like to extend our framework to other variants of summarization e.g. multi-document. Also, we will explore ways of automating parts of the process, e.g. the highlight annotation. Finally, the highlights could also be used as training signal, as it offers content saliency information at a finer level than the single reference typically used.

Acknowledgments

Hardy would like to thank the Indonesian government that has sponsored his studies through the Indonesia Endowment Fund for Education (LPDP). Shashi Narayan and Andreas Vlachos were supported by the EU H2020 SUMMA project (grant agreement number 688139). The latter is also supported by the EPSRC grant eNeMILP (EP/R021643/1).

References

Reinald Kim Amplayo, Seonjae Lim, and Seung-Won Hwang. 2018. Entity Commonsense Representation for Neural Abstractive Summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 697–707.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, volume 2, pages 131–198.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 152–161.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675.

Yen-Chun Chen and Mohit Bansal. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 675–686.

James Clarke and Mirella Lapata. 2010. Discourse constraints for document compression. *Computational Linguistics*, 36(3):411–441.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).*, pages 615–621.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference*, volume 2005, pages 1–12.

Brian S Everitt. 2006. *The Cambridge Dictionary of Statistics*. Cambridge University Press.

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological bulletin*, 76(5):378–382.

Marina Fomicheva and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 77–82.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft Layer-Specific Multi-Task Summarization with Entailment and Question Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697.

Hardy and Andreas Vlachos. 2018. Guided Neural Language Generation for Abstractive Summarization using Abstract Meaning Representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 768–773.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Neural Information Processing Systems*, pages 1–14.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1.

Aishwarya Jadhav and Vaibhav Rajan. 2018. Extractive Summarization with SWAP-NET: Sentences and Words from Alternating Pointer Networks. In

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 142–151.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content Selection in Deep Learning Models of Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Kundan Krishna and Balaji Vasan Srinivasan. 2018. Generating Topic-Oriented Summaries Using Neural Attention. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1697–1705.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving Abstraction in Text Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817.

Chenliang Li, Weiran Xu, Si Li, and Sheng Gao. 2018a. Guiding Generation for Abstractive Text Summarization Based on Key Information Guide Network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 55–60.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018b. Ensure the Correctness of the Summary: Incorporate Entailment Knowledge into Abstractive Sentence Summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for Multi-Document Summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, 1, pages 25–26.

Junyang Lin, Shuming Ma, and Qi Su. 2018. Global Encoding for Abstractive Summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 163–169.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39(2):267–300.

Jordan J Louviere, Terry N Flynn, Anthony Alfred Fred, and John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Afonso Mendes, Shashi Narayan, Sebastião Miranda, Zita Marinho, André F. T. Martins, and Shay B. Cohen. 2019. Jointly extracting and compressing documents with summary state representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, US.

Shashi Narayan, Ronald Cardenas, Nikos Papasarantopoulos, Shay B Cohen, Mirella Lapata, Jiangsheng Yu, and Yi Chang. 2018a. Document Modeling with External Attention for Sentence Extraction. In *Proceedings of the 56st Annual Meeting of the Association for Computational Linguistics*, pages 2020–2030, Melbourne, Australia.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Don't Give Me the Details, Just the Summary! Topic-aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018c. Ranking Sentences for Extractive Summarization with Reinforcement Learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1747–1759, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ani Nenkova. 2005. Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, pages 1436–1441.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why We Need New Evaluation Metrics for {NLG}. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-Reward Reinforced Summarization with Saliency and Entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 646–653.

Maxime Peyrard and Iryna Gurevych. 2018. Objective Function Learning to Match Human Judgements for Optimization-Based Summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 654–660.

D. W. Robertson. 1946. A Note on the Classical Origin of " Circumstances " in the Medieval Confessional. *Studies in Philology*, 43(1):6–14.

Shinsaku Sakaue, Tsutomu Hirao, Masaaki Nishino, and Masaaki Nagata. 2018. Provable Fast Greedy Compressive Summarization with Any Monotone Submodular Function. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1737–1746.

Evan Sandhaus. 2008. The New York Times Annotated Corpus.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 41–45.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the ACL*, pages 1073–1083.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018. Summarization Evaluation in the Absence of Human Model Summaries Using the Compositionality of Word Embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 905–914.

R R Sokal and F J Rohlf. 1995. *Biometry. 3rd ed.* WH Freeman and Company.

Kaiqiang Song, Lin Zhao, and Fei Liu. 2018. Structure-Infused Copy Mechanisms for Abstractive Summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1717–1729.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive Document Summarization with a Graph-Based Attentional Neural Model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1171–1181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simone Teufel and Hans Van Halteren. 2004. Evaluating information content by factoid analysis: Human annotation and stability. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 419–426.

L L Thurstone. 1994. A Law of Comparative Judgment. *Psychological review*, 101(2):255–270.

Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Jordan J Louviere Woodworth and George G. 1991. Best-worst scaling: A model for the largest difference judgments. *University of Alberta: Working Paper*.

Yinfei Yang, Forrest Sheng Bao, and Ani Nenkova. 2017. Detecting (Un)Important Content for Single-Document News Summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 707–712.