How to Compare Summarizers without Target Length? Pitfalls, Solutions and Re-Examination of the Neural Summarization Literature

Simeng Sun[1] **Ori Shapira**[2] **Ido Dagan**[2] **Ani Nenkova**[1]
[1]Department of Computer and Information Science, University of Pennsylvania
[2]Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel
{simsun, nenkova}@seas.upenn.edu
obspp18@gmail.com, dagan@cs.biu.ac.il

Abstract

Until recently, summarization evaluations compared systems that produce summaries of the same target length. Neural approaches to summarization however have done away with length requirements. Here we present detailed experiments demonstrating that summaries of different length produced by the same system have a clear non-linear pattern of quality as measured by ROUGE F1 scores: initially steeply improving with summary length, then starting to gradually decline. Neural models produce summaries of different length, possibly confounding improvements of summarization techniques with potentially spurious learning of optimal summary length. We propose a new evaluation method where ROUGE scores are normalized by those of a random system producing summaries of the same length. We reanalyze a number of recently reported results and show that some negative results are in fact reports of system improvement once differences in length are taken into account. Finally, we present a small-scale human evaluation showing a similar trend of perceived quality increase with summary length, calling for the need of similar normalization in reporting human scores.

1 Introduction

Algorithms for text summarization of news developed between 2000 and 2015, were evaluated with a requirement to produce a summary of a pre-specified length.[1] This practice likely followed the DUC shared task, which called for summaries of length fixed in words or bytes (Over et al., 2007) or influential work advocating for fixed summary length around 85-90 words (Goldstein et al., 1999).

With the advent of neural methods, however, the practice of fixing required summary length was summarily abandoned. There are some exceptions (Ma and Nakagawa, 2013; Kikuchi et al., 2016; Liu et al., 2018), but starting with (Rush et al., 2015), systems produce summaries of variable length. This trend is not necessarily bad. Prior work has shown that people prefer summaries of different length depending on the information they search for (Kaisser et al., 2008) and that variable length summaries were more effective in task-based evaluations (Mani et al., 1999).

There are, at the same time, reasons for concern. The confounding effect of output length has been widely acknowledged for example in earlier work on sentence compression (McDonald, 2006; Clarke and Lapata, 2007); for this task a meaningful evaluation should explicitly take output length into account (Napoles et al., 2011). For summarization in general, prior to 2015, researchers reported ROUGE *recall* as standard evaluation. Best practices for using ROUGE call for truncating the summaries to the desired length (Hong et al., 2014) [2]. (Nallapati et al., 2016) suggested using ROUGE F1 instead of recall, with the following justification "*full-length recall favors longer summaries, so it may not be fair to use this metric to compare two systems that differ in summary lengths. Full-length F1 solves this problem since it can penalize longer summaries.*". The rest of the neural summarization literature adopted F1 evaluation without further discussion.

In this paper we study how ROUGE F1 scores

[1]Here is a list of the most cited 'summarization' papers of that period according to Google Scholar (Erkan and Radev, 2004; Radev et al., 2004; Gong and Liu, 2001; Conroy and O'leary, 2001; Lin and Hovy, 2000; Mihalcea, 2004; Goldstein et al., 2000). All of them present evaluations in which alternative systems produce summaries of the same length, with two of the papers fixing the number of sentences rather than number of words.

[2]As a matter of fact, the established practice was to require human references of different lengths in order to evaluate system outputs of the respective length, a practice that has recently been shown unnecessary (Shapira et al., 2018).

Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen), pages 21–29
Minneapolis, Minnesota, USA, June 6, 2019.

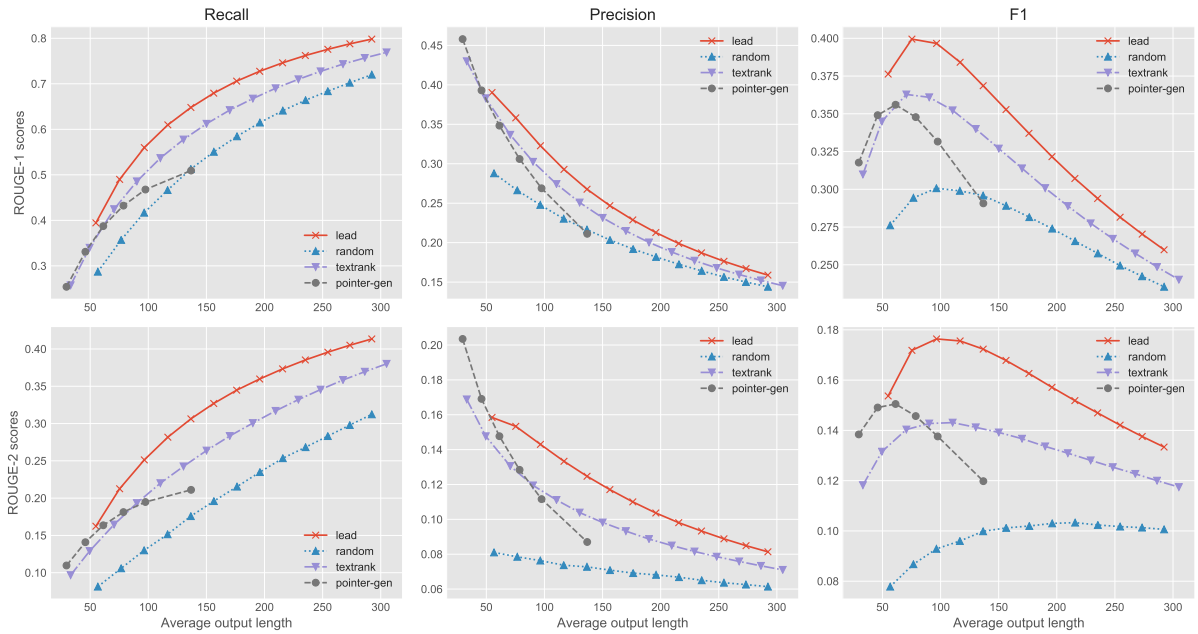

Figure 1: ROUGE recall, precision and F1 scores for lead, random, textrank and Pointer-Generator on the CNN/DailyMail test set.

change with summary length, finding that in the ranges of typical lengths for neural systems it in fact does not penalize longer summaries. We propose an alternative evaluation that appropriately normalizes ROUGE scores and reinterpret several recent results to show that not taking into account differences in length may have favored misleading conclusions. We also present a pilot analysis of summary length in human evaluation.

2 ROUGE and Summary Length

First we examine the behavior of four systems and their respective ROUGE-1 scores (overlap of unigrams between the summary and the reference), on the CNN/DailyMail test set (Nallapati et al., 2016). ROUGE F1 scores have a non-linear pattern with respect to summary length. The graphs for ROUGE-2 (bigram) have the same shape as can be seen from the second row of graphs. Of the four systems, three non-neural baselines are evaluated for lengths between 50 and 300, with a step of 20. Both sentence and word tokenization are performed using nltk (Bird et al., 2009) and words are lowercased. The four systems are as follows:

Lead Extracts full sentences from the beginning of the article with a total number of tokens no more than the desired length. Many papers on neural abstractive methods produce summaries with ROUGE scores worse than this baseline, usually comparing with the version of extracting the first three sentences of the article.

Random Randomly and non-repetitively selects full sentences with a total number of tokens that is no more than the desired length.

TextRank Sentences are scored by their centrality in the graph with sentences as the nodes (Erkan and Radev, 2004; Mihalcea, 2004). We use the Gensim.summarization package (Barrios et al., 2016) to produce these summaries.

Pointer-gen: We use the pre-trained Pointer-Generator model of (See et al., 2017) to get outputs with varying lengths by restricting both minimum and maximum decoding steps.[3] The largest values for min and max decoding step are set to 130 and 150 respectively due to limited computing resources.

Figure 1 shows that ROUGE recall keeps increasing as the summary becomes longer, while precision decreases. For recall, it is clear that even the random system produces better scoring summaries if it is allowed longer length. For all four systems, ROUGE F1 curves first rise steeply, then decline gradually. For summaries longer than 100 words, none of the systems produces a better score than corresponding system with shorter

[3] https://github.com/abisee/pointer-generator, we used the Tensorflow 1.0 version pre-trained pointer-generator model. The pre-trained model performs slightly worse than what was reported in their paper.

summaries. For the range of less than 100 words however, where most of the current systems fall as we will soon see, the trend is unclear since curves overlap and cross. In that range, differences in length may be responsible for differences in ROUGE scores.

It is possible that such behavior is related to the fact that ROUGE uses *word overlap* for comparison. Given the current trends of using text representations and similarity, we also check the shape of curves when representing the lead baseline and reference summary in *semantic space* using different methods. A higher cosine similarity between the two representations indicates a better baseline summary.

We represent summary and reference in embedding space using five methods: (1,2) two universal sentence encoders (Cer et al., 2018); (3) the Infersent (Conneau et al., 2017) model; (4) average and (5) max over each dimension of every word in the input with word2vec word embeddings (Mikolov et al., 2013).

Figure 2 shows the change in similarity between the lead baseline and the reference summary. For all representations, for summary lengths below 100 words, the similarity increases with length. After 100 words, the similarities plateau or slightly decrease for one representation. This indicates that when the number of words is not explicitly tracked, length is still a confounding factor and may affect the evaluations that are based on embedding similarities.

3 Normalizing ROUGE

In the data we saw so far, it is clear that difference in length may account for difference in system performance, while in some pairs of system, one is better than the other irrespective of the length of their summaries, as with the lead and random systems. Therefore, it is of interest to adopt a method that normalizes ROUGE scores for summary length and then re-examine prior literature to see if any of the conclusions change once summary length is taken into account. [4]

Simply dividing by summary length is unwarranted given the non-linear shape of the F1 curve. Instead, we choose to normalize the F1 score of a system by that of a random system which produces same average output length. The output length of a random baseline is easily controllable and any system is expected to be at least as good in content selection as the random baseline.

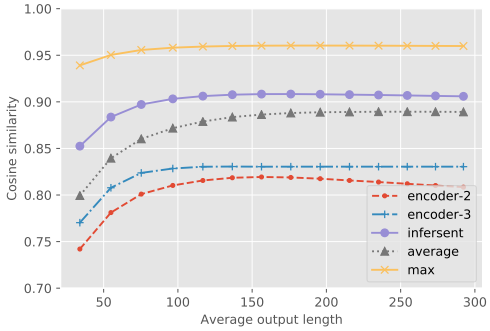

Figure 2: Cosine similarities between summaries generated by lead systems and reference in embedding space on the CNN/DailyMail test set.

The new score also has a useful intuitive interpretation. The score minus one is the percentage of a system improving upon a random system which has same average summary length. In general, it is easier for a system that produces shorter summaries to improve a lot upon a random baseline which has equally short summaries , and more difficult for systems that produce long summaries. The normalized ROUGE score can thus distinguish a poor system which achieves higher ROUGE scores because of generating longer texts from a system which has good summarization techniques but tends to generate shorter summaries. In addition, the random system is independent of the systems to be evaluated, thus the normalizing values can be computed beforehand.

4 Evaluation on CNN/DailyMail Test Set

We re-test 16 systems on the CNN/DailyMail test set:
(1) Pointer-Generator (See et al., 2017) and its variants: a baseline sequence-to-sequence attentional model (*baseline*), a Pointer-Generator model with soft switch between generating from vocabulary and copying from input (*pointer-gen*) and the same Pointer-Generator with coverage loss (*pointer-cov*) for preventing repetitive generation. There are three other content-selection variants proposed in (Gehrmann et al., 2018) which are also based on Pointer-Generator: (i) aligning ref-

[4] We could penalize summaries that are shorter or longer than the reference, similar to the brevity penalty in BLEU (Papineni et al., 2002). Such an approach however assumes that the reference summary length is ideal and deviations from that are clearly undesirable, a fairly strong assumption.

erence with source article (*mask-hi*, *mask-lo*) (ii) training tagger and summarizer at the same time (*multitask*), and (iii) a differentiable model with a soft mask predicted by selection probabilities (*DiffMask*).

(2) Abstractive system with bottom-up attention (*bottom-up*) (Gehrmann et al., 2018) and the same model using Transformer (*BU_trans*) (Vaswani et al., 2017).

(3) Neural latent extractive model (*latent_ext*) and the same model with compression over the extracted sentences (*latent_cmpr*) (Zhang et al., 2018). This setting is important to study, because compression naturally produces a shorter summary and a meaningful analysis of the effect is needed.

(4) TextRank system used in previous section, with maximum summary length set to 50 and 70.

(5) Lead-3 related systems: the first 3 sentences of each article (*lead3*); compressed first 3 sentences of each article which has length of corresponding *pointer-gen* (*lead-pointer*) and *pointer-cov* (*lead-cov*) output, similar to (3). The compression model we used is a Pointer-Generator trained on 1160401 aligned sentence/reference pairs extracted from CNN/DailyMail training data and Annotated Gigaword (AGIGA) (Napoles et al., 2012). We extract the pairs from CNN/DailyMail when every token from the summary sentence can be found in the article sentence. The pairs are extracted from AGIGA when over 70% tokens of a lead sentence are also in the headline. The minimum and maximum decoding step are set to be equal so that the output lengths are fixed. Specifically, let c_i be the length of a summary produced by *pointer-gen*, l_i be the length of lead 3 sentences for the same article and $l_i^{(j)}$ be the length of j^{th} sentence ($j \leq 3$). The j^{th} lead sentence is forced to have output length of $l_i^{(j)} c_i / l_i$ tokens. The average number of tokens are not exactly the same since the size after scaling may be off by at most 1 token.

The random scores are the average over n activations of random systems introduced in §2 ($n = 10$ in our setting). The instability of random systems can be mitigated by setting n to be large enough. Besides, the average over large amounts of test articles can also weaken this issue since we focus on system-level comparison instead of input-level. Given a system output length, we use linear interpolation of the two closest points to estimate the ROUGE score of a random system which has the same average output length.[5]

System	Len.	Sys. F1	Rand. F1	Norm.
latent_cmpr	43	$0.362_{[+2]}$	$0.245_{[+0]}$	$1.473_{[+13]}$
baseline	48	$0.311_{[-1]}$	$0.257_{[+0]}$	$1.209_{[-1]}$
textrank_50	50	$0.345_{[-1]}$	$0.259_{[+0]}$	$1.331_{[+1]}$
mask_lo	51	$0.371_{[+2]}$	$0.263_{[+0]}$	$1.410_{[+9]}$
BU_trans	53	$0.410_{[+10]}$	$0.266_{[+0]}$	$\mathbf{1.541}_{[+11]}$
bottom_up	55	$\mathbf{0.412}_{[+10]}$	$0.272_{[+0]}$	$1.517_{[+9]}$
pointer-gen	56	$0.362_{[-3]}$	$0.273_{[+0]}$	$1.327_{[-4]}$
lead-pointer	56	$0.377_{[+0]}$	$0.273_{[+0]}$	$1.381_{[+2]}$
mask_hi	58	$0.377_{[+0]}$	$0.276_{[+0]}$	$1.366_{[-2]}$
DiffMask	58	$0.380_{[+0]}$	$0.277_{[+0]}$	$1.373_{[-1]}$
lead-cov	61	$0.383_{[+0]}$	$0.279_{[+0]}$	$1.369_{[-3]}$
pointer-cov	62	$0.392_{[+0]}$	$0.280_{[+0]}$	$1.403_{[+0]}$
multitask	63	$0.376_{[-6]}$	$0.281_{[+0]}$	$1.341_{[-8]}$
textrank_70	71	$0.363_{[-9]}$	$0.288_{[+0]}$	$1.259_{[-12]}$
latent_ext	82	$0.409_{[-1]}$	$\mathbf{0.296}_{[+0]}$	$1.384_{[-4]}$
lead3	**85**	$0.401_{[-3]}$	$\mathbf{0.296}_{[+0]}$	$1.351_{[-10]}$
Rank change	-	48	0	90
Spearman	-	0.500	1.000	0.205
Pearson	-	0.491	0.949	0.194

Table 1: System performance on the CNN/DailyMail test set, including average summary length, system ROUGE-1 F1 score, ROUGE-1 F1 for the random system with same average length. Systems are ordered by length. Values in the last three columns are subscripted by the difference in rank when sorted by corresponding item as compared to when sorted by length. In the bottom of the table, we show the sum of absolute rank change, Spearman and Pearson correlation between corresponding values and length.

Table 1 shows the average length of summaries produced by each system, the system ROUGE-1 F1 score, the corresponding ROUGE-1 F1 score of a random system with the same average summary length, and the proposed normalized ROUGE-1 evaluation score. The bottom of the table gives the sum of absolute system rank change with respect to the ordering by summary length and correlations between corresponding values with summary length.

All systems produce summaries in the 43–85 word range, where we already established that ROUGE F1 increases steeply with summary length. Another important observation is that the scores of random systems follow exactly the ordering by length; here summary length alone is responsible for the over 5 ROUGE point improvement. Next to notice is that the normalization

[5] We also explored another kind of random baseline where the last sentence is truncated to get a summary of fixed length. The effect of that normalization is the same as to that presented here. Detailed results can be found in our supplementary material.

leads to about double the difference in rank change with respect to length than regular ROUGE F1. Hence, these scores give information about summary quality that is less related to summary length.

Now we get to revisit some of the conclusions drawn solely from ROUGE scores, without taking summary length into account. Many of the neural abstractive systems produce outputs with scores worse than the lead3 baseline. However this baseline results in the longest summaries. Moreover, after normalization, it becomes clear that *lead3* is in fact considerably worse than *pointer-cov*. As presented in Fig. 1, the TextRank system with summary length of 70 has better ROUGE scores than the same system with summary length of 50. Once these are normalized, however, the system with shorter summaries appears to be more effective (6 points better in normalized score). Finally, we compare the two pairs of extractive systems as well as their versions in which the extracted sentences are compressed. The compressed summaries are about 40 words shorter for the systems in **(3)** and 30 words shorter in **(5)**. Plain ROUGE scores decidedly indicate that compression worsens system performance. When normalized however, *latent_cmpr* emerges as the third most effective system, immediately follow the bottom-up systems (Gehrmann et al., 2018). This is not the case for the simplistic compression variant in *lead3*, which produces shorter summaries but barely changes its rank in the normalized score ranking.

Finally, we compare the systems that reported outperforming the *lead3* baseline. The *latent_ext* system results in summaries very similar in length to *lead3*. Given previous analysis, one might think the ROUGE improvement is due to summary length. However, the normalized score shows that this is not the case and that the *latent_ext* is indeed better than *lead3*. Even more impressive is the analysis of the bottom-up system, which has better ROUGE scores than lead even though it produces shorter summaries. It keeps its first place position even after normalization.

Overall, the analyses we present provide compelling evidence for the importance of summary length on system evaluation. Relying only on ROUGE would at times confound improvement in content selection with the learned ability to generate longer summaries.

Dim.	Question
IN	How well does the summary capture the key points of the article?
RL	Are the details provided by the summary consistent with details in the article?
VE	How efficient do you think the summary conveys the main point of the article?
UC	How much unnecessary content do you think the summary contains?
SR	To what degree do you think the summary is a perfect surrogate of the article?
CN	How much additional informative information can a reader find from the article after reading the summary?

Table 2: Prompts presented to Amazon Mechanical Turk workers

System	CN	IN	RL	SR	UC	VE	LE
frag	4.58	2.96	3.79	2.88	3.46	3.59	31.32
lead3	4.32	3.36	4.11	3.27	3.39	3.72	78.80
ptr_c	4.43	3.22	3.98	3.05	3.33	3.95	71.37
ptr_n	4.40	3.10	4.00	3.11	3.49	3.69	41.50
ptr_s	4.37	3.28	3.96	3.26	3.47	3.89	68.42
textrank	4.51	3.16	4.18	3.18	3.54	3.68	49.13

Table 3: Human ratings for each system. LE stands for summary length. The rest dimensions are described in table 2.

	CN	IN	RL	SR	UC	VE	LE
CN	1.00*	-	-	-	-	-	-
IN	-0.87*	1.00*	-	-	-	-	-
RL	-0.40	0.59	1.00*	-	-	-	-
SR	-0.81	0.88*	0.74	1.00*	-	-	-
UC	-0.36	0.42	-0.16	-0.06	1.00*	-	-
VE	-0.52	0.61	0.08	0.36	0.60	1.00*	-
LE	-0.79	0.96*	0.44	0.71	0.64	0.73	1.00*

Table 4: Correlation among the six human rating dimensions defined in Table 2 and summary length **LE**. Each dimension is the same as in Table 3. Entries with p-value smaller than 0.05 are marked with *.

5 Human Evaluation on Newsroom

We also conduct a pilot human evaluation experiment using the same data as in (Grusky et al., 2018). The human evaluation data are 60 articles from the Newsroom test set and summaries generated by seven systems. These are (1) extractive systems: first three sentences of the article (*lead3*), textrank with word limit of 50 (*textrank*) and the 'fragments' system (*frag*) representing the best performance an extractive system can achieve. (2) an abstractive system (Rush et al., 2015) (*abstractive*) trained on Newsroom data and (3) systems with mixed strategies: Pointer-Generator trained

Max Len.	Informativeness	Verbosity
50	4.13	4.58
70	4.55	4.35
90	4.94	4.42
110	5.22	4.32

Table 5: Average informativeness and verbosity rating for lead system with max length of 50, 70, 90 and 110.

System	CN	IN	RL	SR	UC	VE
frag	1.16	0.75	0.96	0.73	0.88	0.86
lede3	0.86	0.67	0.82	0.65	0.68	0.66
ptr_c	0.91	0.66	0.82	0.63	0.68	0.63
ptr_n	1.04	0.73	0.95	0.74	0.83	0.78
ptr_s	0.91	0.68	0.82	0.68	0.72	0.64
textrank	1.02	0.71	0.94	0.72	0.80	0.75

Table 6: Human ratings normalized by interpolated informativeness rating in table 5.

on CNN/DailyMail data set (*ptr_c*), on subset of Newsroom training set (*ptr_s*) and a subset of Newsroom training data (*ptr_n*). After examining the outputs of each system, the *abstractive* system was excluded because the model was not properly trained. Human evaluation results for each system are shown in Table 3.

We ask annotators to rate six aspects of summary content quality informativeness (*IN*), relevance (*RL*), verbosity (*VE*), unnecessary content (*UC*), making people want to continue reading the original article after reading the summary (*CN*) and being a sufficient substitute for the original article (*SR*) and compute the correlation among these dimensions as well as with summary length. Instead of rating in the range of 1 to 5 as in the original article, we ask the workers to rate in a range of 1 to 7, with higher value corresponds to summary is informative and relevant to the source article, not verbose, has no unnecessary content, much information to be attained after reading summary and can serve as a perfect surrogate to the article. The correlation among six aspects and with summary length are shown in table 4.

Some of the newly introduced questions, such as unnecessary content and verbosity, were intended to capture aspects of the summary which may favor shorter summaries. Relevance is the score introduced in the original (Grusky et al., 2018) study and measures to faithfulness of content, as neural systems tend to include summary content that is not supported by the original article being summarized.

We find that in general people favor systems that produce longer summaries. However, similar to our initial experiment with ROUGE, there is no way to know if the improvement is due simply to the longer length, in which more content can be presented, or in the content selection capabilities of the system. The highest correlation between summary length and a human rating is that for informativeness, which in hind sight is completely intuitive because the longer the summary, the more information it includes. The exact same informativeness definition is used for the Newsroom leaderboard (Grusky et al., 2018)[6]. Clearly, a meaningful interpretation of the human scores will require normalization similar to the one we presented for ROUGE, with human ratings for random or lead summaries of different length, so the overall effectiveness of the system over these is measured in evaluation.

To mirror the analysis of ROUGE scores, we conduct another experiment where we present the workers with lead system of max length 50, 70, 90 and 110 as well as the reference. Complete sentences are extracted so that readability is maintained. Each HIT is assigned to 3 workers and only contains one summary-reference pair. The average length of these four systems are 38.0, 53.4, 75.1, 92.5 respectively. Workers are told that they may assume the reference summary captures all key points of the article, then we ask them to rate the informativeness and verbosity question again. Average ratings for each length can be seen in Table 5. Much like ROUGE, human evaluation of informativeness is also confounded by summary length and requires normalization for meaningful evaluation. We normalize the original human ratings for each system with the interpolated (*IN*) rating in table 5 and present it in table 6.

We also evaluated how the verbosity score behaves when applied to summaries of that length. We chose that because it has the lowest overall correlation with the informativeness and relevance evaluations introduced in prior work. Its (and its related evaluation of unnecessary content) correlation with length is not significant but still appears high. Better sense of the relationship can be obtained in future work when a larger number of system can be evaluated.

Unlike informativeness, verbosity human scores fluctuate with length, increasing and

[6] https://summari.es

decreasing without clear pattern. This suggests future human evaluations should involve more similar judgments likely to capture precision in content selection, which are currently missing in the field.

6 Conclusion

We have shown that plain ROUGE F1 scores are not ideal for comparing current neural systems which on average produce different lengths. This is due to a non-linear pattern between ROUGE F1 and summary length. To alleviate the effect of length during evaluation, we have proposed a new method which normalizes the ROUGE F1 scores of a system by that of a random system with same average output length. A pilot human evaluation has shown that humans prefer short summaries in terms of the verbosity of a summary but overall consider longer summaries to be of higher quality. While human evaluations are more expensive in time and resources, it is clear that normalization, such as the one we proposed for automatic evaluation, will make human evaluations more meaningful. Finally, human evaluations related to content precision are needed for fully evaluating abstractive summarization systems.

References

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174. Association for Computational Linguistics.

James Clarke and Mirella Lapata. 2007. Modelling compression with discourse constraints. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

John M Conroy and Dianne P O'leary. 2001. Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 406–407. ACM.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109. Association for Computational Linguistics.

Jade Goldstein, Mark Kantrowitz, Vibhu O. Mittal, and Jaime G. Carbonell. 1999. Summarizing text documents: Sentence selection and evaluation metrics. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 121–128.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, pages 40–48. Association for Computational Linguistics.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25. ACM.

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.

Kai Hong, John M. Conroy, Benoît Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 1608–1616.

Michael Kaisser, Marti A. Hearst, and John B. Lowe. 2008. Improving search results quality by customizing summary lengths. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 701–709.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 495–501. Association for Computational Linguistics.

Yizhu Liu, Zhiyi Luo, and Kenny Zhu. 2018. Controlling length in abstractive summarization using a convolutional neural network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4110–4119.

Tengfei Ma and Hiroshi Nakagawa. 2013. Automatically determining a proper length for multi-document summarization: A bayesian nonparametric approach. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 736–746.

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 1999. The tipster summac text summarization evaluation. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 77–85. Association for Computational Linguistics.

Ryan McDonald. 2006. Discriminative sentence compression with soft syntactic evidence. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 20. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290. Association for Computational Linguistics.

Courtney Napoles, Matthew R. Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *AKBC-WEKEX@NAACL-HLT*.

Courtney Napoles, Benjamin Van Durme, and Chris Callison-Burch. 2011. Evaluating sentence compression: Pitfalls and suggested remedies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 91–97, Portland, Oregon. Association for Computational Linguistics.

Paul Over, Hoa Dang, and Donna Harman. 2007. DUC in context. *Inf. Process. Manage.*, 43(6):1506–1520.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics.

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 Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics.

Ori Shapira, David Gabay, Hadar Ronen, Judit Bar-Ilan, Yael Amsterdamer, Ani Nenkova, and Ido Dagan. 2018. Evaluating multiple system summary lengths: A case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 774–778. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784. Association for Computational Linguistics.

A Normalizing ROUGE with truncated random selection

System	Len.	Sys. F1	Rand. F1	Norm.
latent_cmpr	43	$0.362_{[+2]}$	$0.256_{[+0]}$	$1.413_{[+13]}$
baseline	48	$0.311_{[-1]}$	$0.267_{[+0]}$	$1.165_{[-1]}$
textrank_50	50	$0.345_{[-1]}$	$0.270_{[+0]}$	$1.280_{[+0]}$
mask_lo	51	$0.371_{[+2]}$	$0.273_{[+0]}$	$1.359_{[+7]}$
BU_trans	53	$0.410_{[+10]}$	$0.275_{[-1]}$	$1.491_{[+10]}$
bottom_up	55	$\mathbf{0.412}_{[+10]}$	$0.274_{[+1]}$	$\mathbf{1.505}_{[+10]}$
pointer-gen	56	$0.362_{[-3]}$	$0.279_{[+0]}$	$1.295_{[-3]}$
lead-pointer	56	$0.377_{[+0]}$	$0.280_{[+0]}$	$1.347_{[+2]}$
mask_hi	58	$0.377_{[+0]}$	$0.281_{[+0]}$	$1.344_{[-1]}$
DiffMask	58	$0.380_{[+0]}$	$0.283_{[+0]}$	$1.344_{[-1]}$
lead-cov	61	$0.383_{[+0]}$	$0.286_{[-1]}$	$1.340_{[-5]}$
pointer-cov	62	$0.392_{[+0]}$	$0.285_{[+1]}$	$1.378_{[+1]}$
multitask	63	$0.376_{[-6]}$	$0.286_{[+0]}$	$1.317_{[-8]}$
textrank_70	71	$0.363_{[-9]}$	$0.291_{[+0]}$	$1.245_{[-12]}$
latent_ext	82	$0.409_{[-1]}$	$0.298_{[+0]}$	$1.374_{[-3]}$
lead3	**85**	$0.401_{[-3]}$	$\mathbf{0.299}_{[+0]}$	$1.342_{[-9]}$
Rank change	-	48	4	86
Spearman	-	0.500	0.944	-0.115
Pearson	-	0.491	0.994	-0.047

Table 7: System performance on the CNN/DailyMail test set, including average summary length, system ROUGE-1 F1 score, ROUGE-1 F1 for the random system with same average length. Systems are ordered by length. Values in the last three columns are subscripted by the difference in rank when sorted by corresponding item as compared to when sorted by length. In the bottom of the table, we show the sum of absolute rank change, Spearman and Pearson correlation between corresponding values and length.