Evaluating *Multiple* System Summary Lengths: A Case Study

Ori Shapira[1], David Gabay[1], Hadar Ronen[2], Judit Bar-Ilan[2], Yael Amsterdamer[1], Ani Nenkova[3], and Ido Dagan[1]

[1]Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel
[2]Information Science Department, Bar-Ilan University, Ramat-Gan, Israel
[3]University of Pennsylvania, Philadelphia, PA

{obspp18, dawid.gabay, hadarg, juditb}@gmail.com
amstery@cs.biu.ac.il, nenkova@seas.upenn.edu, dagan@cs.biu.ac.il

Abstract

Practical summarization systems are expected to produce summaries of varying lengths, per user needs. While a couple of early summarization benchmarks tested systems across multiple summary lengths, this practice was mostly abandoned due to the assumed cost of producing reference summaries of multiple lengths. In this paper, we raise the research question of whether *reference summaries of a single length* can be used to reliably evaluate *system summaries of multiple lengths*. For that, we have analyzed a couple of datasets as a case study, using several variants of the ROUGE metric that are standard in summarization evaluation. Our findings indicate that the evaluation protocol in question is *indeed competitive*. This result paves the way to practically evaluating varying-length summaries with simple, possibly existing, summarization benchmarks.

1 Introduction

Automated summarization systems typically produce a text that mimics a manual summary. In these systems, an important aspect is the output summary length, which may vary according to user needs. Consequently, output length has been a common tunable parameter in pre-neural summarization systems and has been incorporated recently in few neural models as well (Kikuchi et al., 2016; Fan et al., 2017; Ficler and Goldberg, 2017).

It was originally assumed that summarization systems should be assessed across multiple summary lengths. For that, the earliest Document Understand Conference (DUC) (NIST, 2011) benchmarks, in 2001 and 2002, defined several target summary lengths and evaluated each summary against (manually written) reference summaries of the same length.

However, due to the high cost incurred, subsequent DUC and TAC (NIST, 2018) benchmarks (2003-2014), as well as the more recently popular datasets CNN/Daily Mail (Nallapati et al., 2016) and Gigaword (Graff et al., 2003), included references and evaluation for just one summary length per input text. Accordingly, systems were asked to produce a *single* summary, of corresponding length. This decision was partly supported by an observation that system rankings tended to correlate across different summary lengths (Over et al., 2007), even though, as we show in Section 2, this correlation is limited.

In this paper, we propose that the summarization community should consider resuming evaluating summarization systems over *multiple* length outputs, as it would allow better assessment of length-related performance within and across systems (illustrated in Section 3). To avoid the need in multiple-length reference summaries we raise the following research question: *can reference summaries of a single length be used to evaluate system summaries of multiple lengths, as reliably as when using references of multiple lengths, with respect to different standard evaluation metrics?* Recently, Kikuchi et al. (2016) evaluated system summaries of three different lengths against reference summaries of a single length. Yet, their evaluation methodology was not assessed through correlation to human judgment, as has been commonly done for other automatic evaluation protocols. Here, we provide a closer look into this methodology, given its potential value.

As a first accessible case study, we test our research question over the DUC 2001 and 2002 data (Section 2). To the best of our knowledge, these are the only two datasets that include multiple length reference and submitted system summaries, as well as manual assessment of the latter. Our analysis reveals that, for this data and with respect to various highly utilized automatic ROUGE metrics, the answer to our question is affirmative, in

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 774–778
Brussels, Belgium, October 31 - November 4, 2018.

	# refs	ref lengths (# words)	# clusters	# systems
2001	3	50, 100, 200, 400	30	14
2002	2	10, 50, 100, 200	59	10

Table 1: DUC 2001 and 2002. Number of reference summaries per length for each text cluster, reference lengths, number of clusters and number of evaluated systems.

terms of correlation with human judgment.

Our promising results suggest repeating the assessment methodology presented here in future work, to test our question over more recent and broader summarization datasets and human evaluation schemes. This, in turn, would allow the community to feasibly resume proper evaluation and deliberate development of systems that target effective summaries across a range of lengths.

2 Case Study Analysis

Here, we first examine the relevance of our proposal to reinstitute summarization evaluation over multiple summary lengths. Then, we investigate our research question of whether using reference summaries of a single length suffices for evaluating system summaries of multiple lengths. We turn to the DUC 2001 and 2002 multi-document summarization datasets, which, to the best of our knowledge, are the only available datasets that provide the necessary requirements for this analysis (see Table 1).

The importance of evaluating and comparing systems at several lengths is demonstrated with the observation that system rankings can change quite significantly at different summary lengths. In 2001, the Spearman correlation between the available human rankings of systems at the 50-word and 400-word lengths is 0.61. For example, the system ranked first at length 50 ranks sixth at lengths 200 and 400. Even for the human system ranking at the 100-word length, which deviates the least from human rankings at the other lengths, the correlation with system ranking at the 400 length is only 0.73. Generally, the larger the difference between a pair of summary lengths, the greater the fluctuation in system rankings. Similar trends were observed for DUC 2002, and when comparing system rankings by automatic ROUGE scoring (both rankings are elaborated below). Obviously, such performance differences are overlooked when evaluating systems over summaries of a single length.

Next, we turn to investigate our research question. In this paper, we examine it with respect to automatic summary evaluation, which has become most common for system development and evaluation, thanks to its speed and low cost. Specifically, we use several variants of the ROUGE metric (Lin, 2004), which is almost exclusively utilized as an automatic evaluation metric class for summarization. ROUGE variants are based on word sequence overlap between a system summary and a reference summary, where each variant measures a different aspect of text comparison. Despite its pitfalls, ROUGE has shown reasonable correlation of its system scores to those obtained by manual evaluation methods (Lin, 2004; Over and James, 2004; Over et al., 2007; Nenkova et al., 2007; Louis and Nenkova, 2013; Peyrard et al., 2017), such as SEE (Lin, 2001), responsiveness (NIST, 2006) and Pyramid (Nenkova et al., 2007).

We follow the same methodology of assessing the reliability of automatic evaluation scores by measuring their correlation to human evaluation scores. In our case, DUC 2001 and 2002 applied the SEE manual evaluation method. NIST assessors compared systems' summaries to reference summaries, which were all decomposed into a list of elementary discourse units (EDUs). Each reference EDU was marked throughout the system EDUs and was scored for how well it was expressed. The final manually evaluated scores, called the *human mean content coverage scores*, are provided in the DUC datasets. We can then correlate the human-based system ranking, attained from these provided scores, to the system ranking attained from the automatic scores that we calculate using our proposed methodology.

As a baseline, we consider the ROUGE Recall scores obtained by the standard reference summary configuration (Standard, first row in Table 2), that is, when system summaries of each length (table columns) are evaluated against reference summaries of the same length. This is the same configuration used by Lin (2004) when introducing and assessing ROUGE. Then, looking into our research question, we consider reference summary configurations in which system summaries of all lengths are evaluated against reference summaries of a single chosen length (OnlyNNN, subsequent rows of Table 2). In each configuration (each row), we repeat the evaluation twice: once using the complete set of available reference sum-

		System Summary Length									
		50		100		200		400		Avg. across lengths	
		3refs	1ref	3refs	1ref	3refs	1ref	3refs	1ref	3refs	1ref
Reference Set	Standard	0.72	0.65	0.88	0.85	0.9	0.86	0.95	0.94	0.86	0.83
	Only50	0	0	+0.02	0	+0.01	+0.04	+0.01	+0.02	**+0.010**	**+0.015**
	Only100	-0.01	+0.04	0	0	+0.01	-0.01	+0.02	0	+0.005	+0.008
	Only200	-0.09	-0.09	-0.06	-0.08	0	0	+0.01	-0.01	-0.035	-0.0045
	Only400	-0.06	+0.02	-0.09	-0.09	-0.01	+0.03	0	0	-0.040	-0.010

Table 2: Pearson correlations between ROUGE-1 and human scores over DUC 2001 for different system summary lengths (column pairs) and different reference summary configurations (rows), when using one reference or three. The first baseline row presents absolute correlations while successive rows show relative differences to the baseline.

		2001						2002					
		R-1		R-2		R-L		R-1		R-2		R-L	
		3refs	1ref	3refs	1ref	3refs	1ref	2refs	1ref	2refs	1ref	2refs	1ref
Reference Set	Standard	0.86	0.83	0.79	0.77	0.87	0.83	0.78	0.75	0.86	0.82	0.82	0.77
	Only10	-	-	-	-	-	-	0	-0.015	-0.100	-0.178	-0.003	-0.045
	Only50	**+0.010**	**+0.015**	-0.013	-0.038	**+0.008**	+0.010	**+0.035**	**+0.053**	-0.050	-0.038	**+0.020**	**+0.080**
	Only100	+0.005	+0.008	**-0.010**	**-0.003**	+0.005	**+0.013**	+0.023	+0.048	**-0.035**	**0**	-0.008	+0.040
	Only200	-0.035	-0.045	-0.055	-0.053	-0.033	-0.04	+0.013	+0.023	-0.068	-0.025	-0.028	+0.005
	Only400	-0.040	-0.010	-0.075	-0.075	-0.038	0	-	-	-	-	-	-

Table 3: Averaged correlations (across system summary lengths, equivalent to the rightmost columns in Table 2) for different ROUGE variants (column pairs) and reference summary configurations (rows), when using 1 reference or multiple. The first row presents absolute correlations, with relative differences in subsequent rows.

maries of the utilized reference length, and once with just one randomly chosen reference summary from that set (the 3refs and 1ref sub-columns).

For each reference summary configuration, we compute ROUGE Recall system scores[1] for the three common ROUGE variants R-1, R-2 and R-L, which compare unigrams, bigrams and the longest common subsequence, respectively. System scores, per summary length, are obtained by averaging across all summarized texts. We then calculate their Pearson correlation[2] with the available human mean content coverage scores for the systems. The first row of Table 2 shows these correlations, considering the R-1 scores for the DUC 2001 systems, per summary length. The subsequent rows show the corresponding figures for the single-reference-length configurations. For readability, we present in these rows the relative *differences* to the Standard baseline row. Hence, positive values indicate a configuration that is at least as good as the standard configuration.

Table 3 presents correlations averaged over all summary lengths, for the three ROUGE variants over both datasets. We see in the tables that evaluating system summaries of all lengths against references of a single length often performs on par with the standard configuration. In particular, the single fixed set of 50-word reference summaries performs overall as well as the standard approach, and, although not substantially, is the most effective configuration within the data analyzed. In other words, in this dataset, the 50-word reference summaries provide a "test sample" for evaluating the longer system summaries, which is as effective as the same length references used by the standard method.

We note that even when a single reference summary is available, reasonable correlations with human scores are obtained for the 50 word reference. This suggests that it may be possible to compare system summaries of multiple lengths even against a single reference summary, of a relatively short length. This observation seems to deserve further assessment over recent large scale datasets, such as CNN/DailyMail, which provide a single relatively short reference for each summarized text.

In addition to correlation to human assessment, we computed the correlations between system rankings calculated by Standard and those calcu-

[1] Omitting stop words.

[2] Following Lin (2004). Spearman ranking correlations provide similar results.

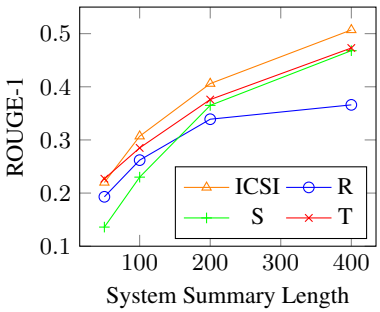

Figure 1: R-1 scores of a few systems, evaluated against the 50-word reference set of DUC 01. Systems R, S and T are from DUC 01; ICSISumm is a later competitive system (Gillick et al., 2008).

lated by Only50, at each system summary length. We find very high correlations (above 0.95 for all system summary lengths, in both datasets) when using multiple references and slightly lower (0.85 to 0.9) with one reference summary. These figures show that the Only50 configuration ranks systems very similarly to Standard.

To further verify our results, we computed correlations in two additional settings. First, we conducted the same analysis, excluding 2-3 of the worst systems, which might artificially boost the correlation (Rankel et al., 2013). Second, we computed score differences between all pairs of systems, for both human and ROUGE scores, and computed the correlation between these two sets of differences (Rankel et al., 2011). In both cases we observed rather consistent results, assessing that a single set of short reference summaries evaluates system summaries of different lengths just as well as the standard configuration.

3 Cross-length Summary Evaluation

This section illustrates how system performances can be measured and compared when evaluating them on outputs of varying lengths against a single reference point. Figure 1 presents the ROUGE scores of the Only50 configuration for three DUC-01 submitted systems, and for ICSISumm (Gillick et al., 2008), a later competitive system.

As expected when measuring ROUGE Recall against a fixed reference length, longer system summaries typically cover more of the reference summaries content than shorter ones, yielding higher scores. Yet, it can be noted, for example, that the value of the 400-word summary of system R in the figure is lower than that of the 200-word summaries of the other systems. Such a comparison is impossible in the standard setup, as each system length is evaluated against different reference summaries. We note that similar comparisons are embedded in the evaluations of Steinberger and Jezek (2004) and Kikuchi et al. (2016), who also evaluated multiple summary lengths.

Further, one can define the marginal value of longer summaries of a given system as the ROUGE score increase per number of additional words, namely the graph slope. This denotation allows measuring the effectiveness of producing longer summaries. For example, deploying system R, we might decide to output only summaries no longer than 200 words, since the marginal value of longer summaries becomes too small. The other systems, on the other hand, seem marginally effective also in 400 word summaries.

4 Discussion

We proposed the potential value of evaluating summarization systems at different summary lengths. Such evaluations would allow proper evaluation of systems' "length knob", tracking how their ranking changes across summary lengths as well as tracking the cross-length behavior of individual systems. Given that reference summaries of a single length are usually available in practice, we analyzed the potential use of *reference summaries of a single length* for evaluating *system summaries of multiple lengths*. We found, on the only two datasets readily available for such analysis, that this configuration is as reliable as the standard configuration, which evaluates each system summary against a reference of a matching length.

To broadly substantiate our findings, we propose future work that would follow our assessment methodology over test samples from current datasets (e.g. CNN/DailyMail), judging performance of current systems and utilizing current manual evaluation protocols. This would require preparing, for limited samples, additional manually crafted summaries of several lengths and manually evaluating system summaries of corresponding lengths. Using such data, it will be possible to repeat our analysis and test the broader validity of the single-reference-length configuration. If broadly assessed, it will be possible to start evaluating system summaries of multiple lengths over most currently available datasets, leveraging the available single-length reference summaries. Fu-

ture benchmarks could require systems to produce different length outputs, while feasibly evaluating them using the existing, single length, reference summaries. This, in turn, is likely to drive research to better address the need for producing high quality summaries flexibly across a range of summary lengths, a dimension that has been disregarded for long.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. We thank Yinfei Yang for his assistance in producing the ICSISumm summaries that we utilized in our analysis. This work was supported in part by grants from the MAGNET program of the Israel Innovation Authority; by the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1); by the BIU Center for Research in Applied Cryptography and Cyber Security in conjunction with the Israel National Cyber Bureau in the Prime Ministers Office; and by the Israel Science Foundation (grants 1157/16 and 1951/17).

References

Angela Fan, David Grangier, and Michael Auli. 2017. Controllable abstractive summarization. *CoRR*, abs/1711.05217.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104. Association for Computational Linguistics.

Daniel Gillick, Benoit Favre, and Dilek Hakkani-Tür. 2008. The ICSI Summarization System at TAC 2008. In *TAC*.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4:1.

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. 2001. Summary evaluation environment user guide. http://www1.cs.columbia.edu/nlp/tides/SEEManual.pdf.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, 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, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4.

NIST. 2006. Responsiveness assessment instructions. www-nlpir.nist.gov/projects/duc/duc2006/responsiveness.assessment.instructions.

NIST. 2011. Document Understanding Conferences. https://duc.nist.gov/.

NIST. 2018. Text Analysis Conferences. https://tac.nist.gov/.

Paul Over, Hoa Dang, and Donna Harman. 2007. DUC in context. *Information Processing & Management*, 43(6):1506–1520.

Paul Over and Yen James. 2004. An Introduction to DUC 2004 Intrisic Evaluation of Generic News Text Summarization Systems. duc.nist.gov/pubs/2004slides/duc2004.intro.pdf.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the EMNLP 2017 Workshop on New Frontiers in Summarization*, pages 74–84.

Peter Rankel, John M Conroy, Eric V Slud, and Dianne P O'Leary. 2011. Ranking human and machine summarization systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 467–473. Association for Computational Linguistics.

Peter A Rankel, John M Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 131–136.

Josef Steinberger and Karel Jezek. 2004. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4:93–100.