An Anchor-Based Automatic Evaluation Metric for Document Summarization

Kexiang Wang[1*], **Tianyu Liu**[1*], **Baobao Chang**[1,2] and **Zhifang Sui**[1,2]
[1]Key Laboratory of Computational Linguistics, Ministry of Education,
School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[2]Peng Cheng Laboratory, Guangdong, China
{wkx,tianyu0421,chbb,szf}@pku.edu.cn

Abstract

The widespread adoption of reference-based automatic evaluation metrics such as ROUGE has promoted the development of document summarization. In this paper, we consider a new protocol for designing reference-based metrics that require the endorsement of source document(s). Following protocol, we propose an anchored ROUGE metric fixing each summary particle on source document, which bases the computation on more solid ground. Empirical results on benchmark datasets validate that source document helps to induce a higher correlation with human judgments for ROUGE metric. Being self-explanatory and easy-to-implement, the protocol can naturally foster various effective designs of reference-based metrics besides the anchored ROUGE introduced here.

1 Introduction

Automatic evaluation metric plays a vital role in evaluating system performance for the task of document summarization. Challenges remain in the design of an ideal evaluation metric and the off-the-shelf metrics have their own drawbacks (Schluter, 2017; Kryscinski et al., 2019). The widely adopted metrics, e.g. ROUGE (Lin, 2004), are reference-based in that they compare the output of some summarizer (namely peer summary) with one or multiple human-authored summaries (namely reference/model summary). The reference-free metrics are still not mature enough to be utilized for evaluation in a real-world setting since their correlations with human judgments have been reported to fall far behind reference-based metrics, especially for multi-document summarization (Peyrard et al., 2017; Gao et al., 2020)[1]. In this paper, we consider a new protocol of reference-based summarization metrics by rethinking the role of source document which is indeed a lost treasure neglected by most previous works. Furthermore, a specific implementation of the protocol (i.e. anchored version of ROUGE) will be discussed.

The reference-based metrics that already exist typically pursue a kinda computation of overlap between the peer and reference summary either at a lexical level (Lin, 2004) or at a semantic level (Ng and Abrecht, 2015; Sun and Nenkova, 2019; Zhang et al., 2019). However, to our knowledge, few of them consider the impact of source document (or documents in multi-document summarization) on the computation. This goes against common sense as source document is the true information source of both summaries and can be utilized to boost the discriminative power of metrics. Therefore, we advance a new protocol of reference-based metrics for the evaluation of document summarization. *More specifically, the direct participation of source document is a necessity to compute any reference-based metric for document summarization*. This makes source document endorse a certain metric and the advantage lies with the ability to fact-check the information of peer summary based on the information pool (i.e. source document). The protocol change is illustrated in Figure 1. Metrics designed under the new proto-

*: The authors contribute equally to this paper.

[1]Reference-free evaluation metrics inherently are more suitably used as the reward function in a reinforcement-learning-based summarizer (Böhm et al., 2019; Gao et al., 2020).

Proceedings of the 28th International Conference on Computational Linguistics, pages 5696–5701
Barcelona, Spain (Online), December 8-13, 2020

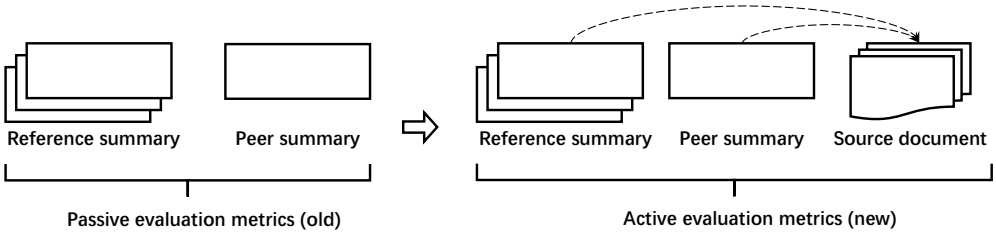

Figure 1: The transition from old-fashioned to newly-introduced protocol for designing reference-based automatic evaluation metrics in the document summarization task. The curved arrows on the right show that both summaries are derived from the source document.

col are called "active metrics" since they will be able to refer to the source. In a word, the new protocol has introduced a key dimension that can nurture reference-based summarization metrics.

For a verification purpose, we propose an anchored version of ROUGE metric under the new protocol. The anchors here mean a set of lexical items (called particles) in source document corresponding to a certain particle in the summary (peer or reference). Utilizing anchor set in the computation of ROUGE can introduce a weighted scheme that focuses more on the link to source document, as will be detailed in the next section.

2 A Specific Implementation: Anchored ROUGE

Following the new protocol, the ROUGE metric can be revised by introducing the anchor set for each particle (i.e. lexical item such as n-gram and skipping bigram) in both peer and reference summaries. The anchor set for a particle in the summary comprises k particles in source document, each of which is a good match for the summary particle. In other words, anchor sets serve as the ground of summary particles.

We build the anchor set $\mathcal{A}_s$ for summary particle s following the two steps: (1) Compute the cosine similarity of embedding vectors of s and d with d being any arbitrary document particle (s and d should be of the same lexical form such as bigram); (2) Extract top-k document particles based on similarity to form the anchor set, i.e. $\mathcal{A}_s = \{d_{s1}, d_{s2}, ..., d_{sk}\}$. Also, we record the similarity as the strength of anchor and denote the strength between s and d_{si} as q_{si} ($1 \leq i \leq k$). The embedding vector of the particle in this paper is obtained by averaging the contextualized embeddings of all tokens occurring in the particle. Specifically, in the following experiment, we will sum the last four hidden layers of the pre-trained uncased BERT Base model[2] (Devlin et al., 2019) to get the embedding for each token (dimension of embedding vector is 768). An example of anchor set can be found in Figure 2.

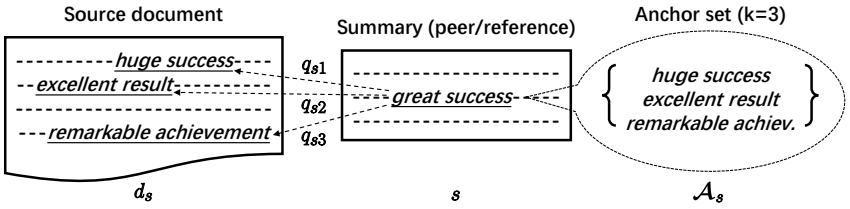

Figure 2: An example of anchor set for the bigram "great success" when top-3 results are extracted.

The anchored version of ROUGE can be defined as follows once all the anchor sets for summary particles (both in peer and reference) have been built. We calculate the union of anchor sets for all particles in a reference summary and denote it as $\mathcal{C}_{\text{ref}}$. Eqn. 1 gives the formula of anchored ROUGE and function T is defined by Eqn. 2. Notice that notation "RefSumm" is a collection of reference summaries, w_s is the count of particle s (with stemming) occurring in either peer or reference summary and δ is Kronecker delta function, i.e. it is 1 only when two relevant variables are equal and 0 otherwise.

[2] https://github.com/google-research/bert

		R-1	R-2	R-1-WE	R-2-WE	BERTScore	S^3_{full}	S^3_{best}	Mover	AncR-1	AncR-2
TAC 2008	r	.747	.718	.579	.556	.750	.753	.754	.760	**.772**	.756
	ρ	.632	.635	.458	.388	.649	.652	.652	.672	**.690**	.653
	τ	.501	.498	.329	.301	.492	.497	.495	.507	**.529**	.511
TAC 2009	r	.808	.803	.653	.671	.823	.838	**.842**	.831	.837	**.842**
	ρ	.692	.694	.516	.481	.703	.724	.731	.701	.730	**.738**
	τ	.533	.531	.384	.362	.545	.551	.557	.550	**.571**	.564

Table 1: Summary-level correlation results between reference-based automatic metrics and human judgments ($k=5$ and $n=4$). Best correlations are in bold and our proposed metrics are **AncR-1** and **AncR-2**.

$$\text{ROUGE-anchored} = \frac{\sum_{\text{ref}\in\text{RefSumm}} \sum_{d\in\mathcal{C}_{\text{ref}}} min(T(d,\text{peer}), T(d,\text{ref}))}{\sum_{\text{ref}\in\text{RefSumm}} \sum_{d\in\mathcal{C}_{\text{ref}}} T(d,\text{ref})}, \tag{1}$$

$$T(d,\text{summ}) = \sum_{s\in\text{summ}} \sum_{\substack{i=1 \\ d_{si}\in\mathcal{A}_s}}^{i=k} \delta_{d,d_{si}} \cdot w_s \cdot q_{si}, \quad \text{for} \quad \text{summ} \in \{\text{peer}, \text{ref}\}. \tag{2}$$

The anchored metric listed above has based the computation on anchor sets residing in the source document. For convenience, we will compare it with the original ROUGE metric, especially the equivalent definition of ROUGE-N given by Lin and Bilmes (2011) (Theorem 3 in the original paper). Function T replaces the count of summary particles, which is adopted in original ROUGE, and for a specific document particle sums the weighted contributions from different summary particles (the weight coefficient is the anchor strength q_{si} as shown in Eqn. 2). The factor w_s is used to assess the effect of multiple occurrences of the same summary particle s. In Eqn. 1, the min function is utilized to compute the weighted matching degree based on document particle d (thus the overall metric will be less than one), which revises the exact count of matching summary particles in original ROUGE. Based on these manipulations, anchored ROUGE is endorsed by source document and freed from the pattern of "hard matching" in original ROUGE, whose evaluation efficacy will be tested in the next section.

3 Evaluation Efficacy of Anchored ROUGE

Datasets. We select two datasets of topic-focused multi-document summarization (MDS), i.e. TAC 2008[3] and TAC 2009[4], for two main reasons: (1) MDS is more challenging than single document summarization and summarizers tend to behave more differently for evaluation, which fits the purpose to examine various metrics; (2) Multiple reference summaries are offered, which makes it possible to perform robustness test (see Table 2). The two datasets consist of 48 and 44 topics, respectively, each of which has 10 source documents and 4 reference summaries, i.e. n is 4. We only use document set A of official datasets in line with Louis and Nenkova (2013) and Gao et al. (2020). Additionally, TAC 2008 has 57 peer summaries for each topic while TAC 2009 has 55. All summaries are at most 100 words and each peer summary is associated with a Pyramid score (Nenkova and Passonneau, 2004), which serves as the human judgment. For tuning the anchor set size (i.e. k in Section 2), another dataset (DUC 2007[5]) will be used.

Comparing metrics. These reference-based metrics are involved in the experiment. (1) ROUGE (Lin, 2004): a traditional metric for counting lexical-level overlap. For comparison, two variants are considered based on either unigram (**R-1**) or bigram (**R-2**). (2) ROUGE-WE (Ng and Abrecht, 2015): a metric based on word2vec embeddings (Mikolov et al., 2013) to compute semantic similarity. ROUGE-WE with unigram (**R-1-WE**) and bigram (**R-2-WE**) are computed. (3) **BERTScore** (Zhang et al., 2019): a direct metric computing token similarity with BERT embeddings. (4) $\boldsymbol{S^3_{full}}$ and $\boldsymbol{S^3_{best}}$ (Peyrard et al.,

[3] https://tac.nist.gov/2008/summarization/update.summ.08.guidelines.html
[4] https://tac.nist.gov/2009/Summarization/update.summ.09.guidelines.html
[5] https://duc.nist.gov/duc2007/tasks.html#pilot

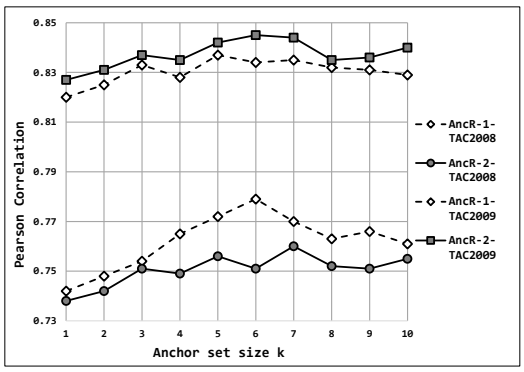

Figure 3: Exploring anchor set size k.

		TAC 2008		TAC 2009	
		r	ρ	r	ρ
AncR-1	n=4	.772	.690	.837	.730
	n=3	.770	.685	.836	.726
	n=2	.769	.686	.832	.724
	n=1	.764	.679	.831	.721
AncR-2	n=4	.756	.653	.842	.738
	n=3	.760	.658	.840	.736
	n=2	.754	.654	.835	.732
	n=1	.751	.652	.833	.729

Table 2: Correlations computed with n references.

2017): two learned metrics that combine different sets of existing metrics. (5) **Mover** (Zhao et al., 2019): a contextualized-embedding-based metric using Word Mover's Distance (Kusner et al., 2015). We report its best version with the BERT embeddings and the certain methods for fine-tuning and aggregation of embeddings according to the original paper. (6) ROUGE-anchored: our metric proposed under the new protocol as formulated in Section 2. Similar to ROUGE, we consider two variants with different particle granularities, i.e. unigram (**AncR-1**) and bigram (**AncR-2**). Tuning on DUC 2007 sets the anchor set size to 5.

Following the convention, we compute the average summary-level correlation with human judgments for each metric in terms of three correlation coefficients: Pearson r, Spearman ρ and Kendall τ.

Main results. As shown in Table 1, the overall correlation results prove the superiority of our anchored ROUGE metric. On both datasets, anchored ROUGE has achieved the highest correlations according to all three correlation coefficients. More specifically, both AncR-1 and AncR-2 have a correlation higher than their original counterparts (i.e. R-1 and R-2) and the gaps are over 2.5 and 1.3 percent, respectively. Even the most recent metric based on advanced contextualized embeddings, i.e. Mover, has fallen behind our metric (by over one percent as compared with AncR-1 on TAC 2008 and AncR-2 on TAC 2009). For a more convincing comparison, we have conducted the pairwise Williams significance test recommended by Graham (2015) between our metric (more precisely AncR-1 on TAC 2008 and AncR-2 on TAC 2009) and other competitors and the result shows that the increases of our metric over others except the supervised metric S^3_{best} are statistically significant (p-value < 0.05).

Hyperparameter effect & Robustness. Two extra tests have been performed to further analyze our metric. The effects of anchor set size k on Pearson correlations are illustrated in Figure 3, indicating that an anchor set with the proper size is needed to establish the efficacy of our metric. The correlations deteriorate when k is less than three and we see no substantial improvements with an extremely large k that causes more intensive computation. The effect of the number of reference summaries is shown in Table 2. We have used all available references to compute metrics when n is equal to four and used n randomly selected references with a smaller n (note that the average of $\binom{4}{n}$ results is reported). The observation is that our metric is relatively robust to n and it demonstrates that our metric is less prone to the reference noise observed in Kryscinski et al. (2019) or the reference bias introduced when very few reference summaries are available (Hermann et al., 2015; Grusky et al., 2018).

4 Related Work

There are various reference-based automatic evaluation metrics for the task of document summarization. The widely accepted metric is ROUGE (Lin, 2004) that focuses primarily on n-gram co-occurrence statistics. Some strategies are proposed to replace the "hard matching" of ROUGE, such as the adoption of WordNet (ShafieiBavani et al., 2018) and the fusion of ROUGE and word2vec (Ng and Abrecht, 2015). Another promising method of designing metrics is to directly compute the semantic similarity of peer and reference summary, including the metrics utilizing various word embeddings such as ELMo (Sun and Nenkova, 2019) and BERT (Zhang et al., 2019; Zhao et al., 2019). Furthermore, Zhang et

al. (2020) proposes a metric computing factual correctness based on information extraction. However, none of the above metrics fall into the newly-introduced protocol. The anchored ROUGE proposed by us is a refined metric that has followed the new protocol and enjoyed higher correlations with human judgments.

5 Conclusion

We propose a new protocol to foster the development of reference-based automatic metrics to evaluate document summarization. The protocol features the endorsement of source document and can be implemented as an anchored version of the ROUGE metric fixing each summary particle on the ground of source document. Experiments demonstrate that anchored ROUGE has a higher correlation with human judgments as compared to other metrics. Also, our metric is robust to the number of reference summaries, which can be applied to the challenging low-resource setting. Future works include extending the new protocol to get various workable evaluation metrics besides anchored ROUGE.

Acknowledgments

We would like to thank all the reviewers for their helpful advice on various aspects of this work. This work is supported by National Natural Science Foundation of China (No. 61936012 and No. 61772040) and Beijing Academy of Artificial Intelligence (BAAI).

References

Florian Böhm, Yang Gao, Christian M Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3101–3111.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online, July. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 128–137.

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.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

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

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.

Ani Nenkova and Rebecca J 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.

Jun Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930.

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

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*, pages 41–45.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018. A graph-theoretic summary evaluation for rouge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 762–767.

Simeng Sun and Ani Nenkova. 2019. The feasibility of embedding based automatic evaluation for single document summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1216–1221.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis P. Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Association for Computational Linguistics (ACL)*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.