Studying Summarization Evaluation Metrics in the Appropriate Scoring Range

Maxime Peyrard*
EPFL
maxime.peyrard@epfl.ch

Abstract

In summarization, automatic evaluation metrics are usually compared based on their ability to correlate with human judgments. Unfortunately, the few existing human judgment datasets have been created as by-products of the manual evaluations performed during the DUC/TAC shared tasks. However, modern systems are typically better than the best systems submitted at the time of these shared tasks. We show that, surprisingly, evaluation metrics which behave similarly on these datasets (average-scoring range) strongly disagree in the higher-scoring range in which current systems now operate. It is problematic because metrics disagree yet we can't decide which one to trust. This is a call for collecting human judgments for high-scoring summaries as this would resolve the debate over which metrics to trust. This would also be greatly beneficial to further improve summarization systems and metrics alike.

1 Introduction

The progress in summarization is tightly intertwined with the capability to quickly measure improvements. Thus, a significant body of research was dedicated to the development of automatic metrics (Lloret et al., 2018). Yet, this remains an open problem (Rankel et al., 2013).

Typically, evaluation metrics are compared based on their ability to correlate with humans (Lin and Hovy, 2003). Then, the selected metrics heavily influence summarization research by guiding progress (Lloret et al., 2018) and by providing supervision for training summarization systems (Yogan et al., 2016).

Despite their central role, few human judgment datasets have been created. The existing ones are the result of the manual evaluations performed during shared tasks (Dang and Owczarzak, 2008, 2009).

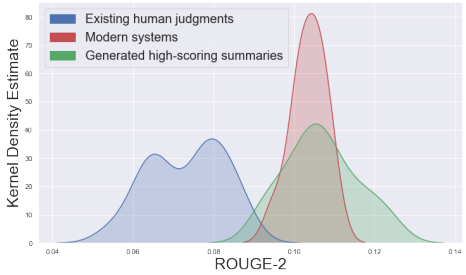


Figure 1: The blue distribution represents the score distribution of summaries available in the human judgment datasets of TAC-2008 and TAC-2009. The red distribution is the score distribution of summaries generated by mordern systems. The green distribution corresponds to the score distribution of summaries we generated in this work as described in section 3.

Thus, the annotated summaries are mostly *average* compared to nowadays standards. Indeed, the best systems submitted at the time of these shared-tasks have typically served as baselines for subsequent works. This is illustrated by figure 1, which compares the score distribution of summaries in human judgment datasets with the score distribution of modern summarization systems.[1] The score distribution on which evaluation metrics are tested (blue zone) differs from the one in which they now operate (red zone). Thus, there is no guarantee that evaluation metrics behave according to human judgments in the high-scoring range. Yet, summarization systems explicitly target high-scoring summaries (Radev et al., 2003).

In this work, we study several evaluation metrics in this high-scoring range based on an automatically generated dataset. We show that, even though current evaluation metrics correlate well

*Research partly done at UKP Lab from TU Darmstadt

[1]for modern systems, we used the scores of summaries from Hong et al. (2014) and other recent approaches (Cao et al., 2015; Nallapati et al., 2017).

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5093–5100
Florence, Italy, July 28 - August 2, 2019.

with each other in the average range, they strongly disagree for high-scoring summaries. This is related to the *Simpson paradox*, where different conclusions are drawn depending on which slice of the population is considered (Wagner, 1982). This is problematic because current metrics cannot be distinguished based solely on an analysis of available human judgments. Nevertheless, they will promote very different summaries and systems.

These results call for the gathering of human judgments in the high-scoring range. We provide data and code to reproduce our experiments.[2]

Contributions:

(i) We present a simple methodology to study the behavior of metrics in the high-scoring range.

(ii) We observe low and even some negative correlations in this range.

(iii) This work serves as a motivation to gather human annotations in the relevant scoring range.

2 Background

Usually, evaluation metrics are compared based on their ability to correlate with human judgments (Lin and Hovy, 2003). Several works followed this principle and provided different recommendations about which metric to use. For instance, Owczarzak et al. (2012) used a *signed Wilcoxon test* to find significant differences between metrics and recommended to use ROUGE-2 recall with stemming and stopwords not removed. In a wider study, Graham (2015) found ROUGE-2 precision with stemming and stopwords removed to be the best. Rankel et al. (2013) used accuracy and found ROUGE-4 to perform well. They also observe that the correlation between ROUGE and human judgments decreases when looking at the best systems only. This is in agreement with our work, except that we look at summaries better than the current state-of-the-art. Radev et al. (2003) also observed that the high-scoring range is the most relevant for comparing evaluation metrics because summarizers aim to extract high-scoring summaries. However, they performed analysis on the best scoring summaries from 6 systems which remain average compared to nowadays standard.

Our analysis differs from such meta-evaluation (evaluation of evaluation metrics) because we do not provide another metric recommendation. Instead, we start from the observation that human judgments are limited in their coverage and analyze the behavior of existing candidate metrics in the high-scoring range not available in these datasets.

These previous works computed correlations between metrics and humans, we compute correlations between pairs of metrics in scoring ranges for which there are no human judgments available.

3 Data Generation

In this work, we study the following metrics:

ROUGE-2 (**R-2**): measures the bigram overlap between the candidate summary and the pool of reference summaries (Lin, 2004).

ROUGE-L (**R-L**): a variant of ROUGE which measures the size of the longest common subsequence between candidate and reference summaries.

ROUGE-WE (**R-WE**): instead of hard lexical matching of bigrams, **R-WE** uses soft matching based on the cosine similarity of word embeddings (Ng and Abrecht, 2015).

JS divergence (**JS-2**): uses Jensen-Shannon divergence between bigram distributions of references and candidate summaries (Lin et al., 2006).

S3: a metric trained explicitly to maximize its correlation with manual Pyramid annotations (Peyrard et al., 2017).

We chose these metrics because they correlate well with available human judgments (about .4 Kendall's τ; the exact numbers are provided in appendix A) and are easily available. For a recent overview of evaluation metrics, we recommend Lloret et al. (2018).

Once an evaluation metric becomes standard, it is optimized, either directly by supervised methods or indirectly via repeated comparisons of unsupervised systems. To mimic this procedure, we optimized each metric using a recently introduced genetic algorithm for summarization (Peyrard and Eckle-Kohler, 2016).[3] The metric m is used as the fitness function. The resulting population is a set of summaries ranging from random to upper-bound according to m. For both TAC-2008 and TAC-2009, we used a population of 400 summaries per topic (per metric). The final dataset contains $160,523$ summaries for an average of

[2] https://github.com/PeyrardM/acl-2019-Compare_Evaluation_Metrics

[3] https://github.com/UKPLab/coling2016-genetic-swarm-MDS

		R-WE	R-L	JS-2	S3
R-2	(W)	.774	.708	.871	.799
	(A)	.644	.532	.887	.744
	(T)	.016	-.187	.284	.096
R-WE	(W)		.692	.703	.824
	(A)	-	.462	.530	.752
	(T)		-.254	-.145	.131
R-L	(W)			.647	.709
	(A)	-	-	.492	.571
	(T)			-.274	-.200
JS-2	(W)				.738
	(A)	-	-	-	.659
	(T)				-.046

Table 1: Pairwise correlation (Kendall's τ) between evaluation metrics on various scoring range. (T) is the high-scoring range, (A) is the average-scoring range (human judgment datasets) and (W) is the whole scoring range

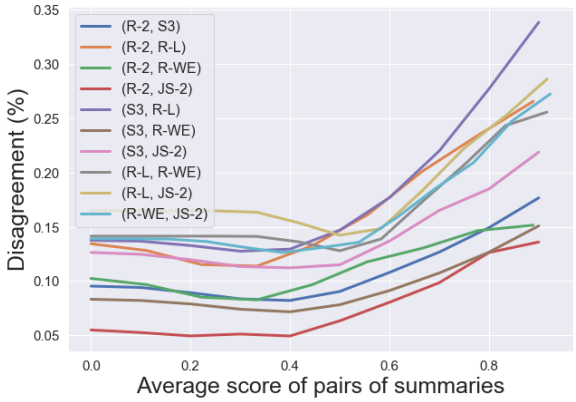


Figure 2: Percentage of disagreement between metrics for increasing scores of summary pairs (Scores have been normalized).

$1,763$ summaries per topic (less than $5 * 400$ due to removed duplicates). We refer to this dataset as (W) as it covers the whole scoring range.

In order to focus on the top-scoring summaries, we preserve the summaries scoring higher than the *LexRank* baseline (Erkan and Radev, 2004) for at least one metric. LexRank is a graph-based extractive summarizer often used as a baseline. Thus, most current and future summarization systems should perform better and should be covered by the selected scoring range. Besides, LexRank is strong enough to discard a large number of average scoring summaries. The resulting dataset contains an average of 102 summaries kept per topic. This dataset of top-scoring summaries is noted (T). The ROUGE-2 score distribution of (T) is depicted by the green area in figure 1.

We provide the pseudo-code and other details concerning the data generation procedure in appendix B.

Additionally, we refer to the summaries available as part of the human judgment datasets as (A) because they cover the average-scoring range.

4 Correlation Analysis

We compute the pairwise correlations between evaluation metrics averaged over all topics for different scoring ranges and report the results in table 1. For (A) and (W), we observe high correlations between pairs of metrics ($>$ .6 Kendall's τ). **JS-2** and **R-2** have the strongest correlation, while **R-L** is less correlated with the others. It is worth remembering that **JS-2** and **R-2** both operate on bigrams which also explain their stronger connection.

However, in the high-scoring range (T), correlations are low and often negative. Even, **R-2** and **JS-2** only retain little correlation (< 0.3 τ).

For most pairs, the correlations are close to what would be expected from random behavior. Additionally, **R-L** has negative correlations with other metrics. It indicates that there is no global agreement on what constitutes improvements when the summaries are already better than the baseline.

This is akin to the *Simpson paradox* because considering different sub-populations yields different conclusions (Wagner, 1982). In fact, it is simple to distinguish obviously bad from obviously good summaries, which results in superficially high correlations when the whole scoring range is considered (Radev et al., 2003). However, summarization systems target the high-scoring range and evaluation metrics should accurately distinguish between high-scoring summaries. Unfortunately, existing metrics disagree wildly in this range.

Disagreement increases with higher-scoring summaries:
We also visualize the gradual change in metrics agreement when moving from the average to the high-scoring range in figure 2. For each pair of metrics, the disagreement clearly increases for higher scoring summary pairs. This confirms that metrics disagree for high-scoring summaries. It is more pronounced for some pairs like the ones involving **R-L** as already observed in table 1.

The problem with reporting several disagreeing metrics:
It is a common good practice to report the results of several evaluation metrics. In the average scoring range, where metrics generally agree, this creates a robust measure of progress. The specificities of each metric are averaged-out putting the focus on the general trend.

However, when the metrics do not correlate, improvements according to one metric are rarely improvements for the other ones.

Let $M = \{m_1, \ldots, m_n\}$ be the set of evaluation metrics. Then, for a topic $\mathcal{T}$ from the dataset (W), we select a summary s and ask: among the summaries which are better than s for one metric (N), how many are better for all metrics (F)? This is given by:

$$\frac{F}{N} = \frac{|\{x \in \mathcal{T} |\ \forall m \in M,\ m(x) > m(s)\}|}{|\{x \in \mathcal{T} |\ \exists m \in M,\ m(x) > m(s)\}|} \quad (1)$$

Here, $m(x)$ is the score of the summary x according to the metric m. Thus, $\frac{F}{N}$ measures the difficulty of finding consistent improvements across metrics.

The process is repeated for $5,000$ randomly sampled summaries in the sources. In figure 3, the resulting $\frac{F}{N}$ ratios are reported against the normalized average score of the selected summaries s.

We observe a quick decrease in the ratio $\frac{F}{N}$. The proportion of consistent improvements (agreed by all metrics) is dropping when the average score of summaries increases. When the baseline scores go up, the disagreement between metrics is strong enough that we cannot identify summaries which are considered better than the baseline for each metric. Thus, there is no common trend between metrics that can be exploited by reporting them together.

Discussion:
Intuitively, smaller populations and narrow scoring ranges can also lead to lower correlations. However, (T) displays low correlations with 102 summaries per topic whereas (A) has strong correlations with 50 summaries per topic. Also, the high-scoring range covers 38% of the full scoring range (from LexRank to upper-bound), while human judgments cover 35% of the full scoring range. Thus, the width of the scoring range and the population size do not explain the observed differences.

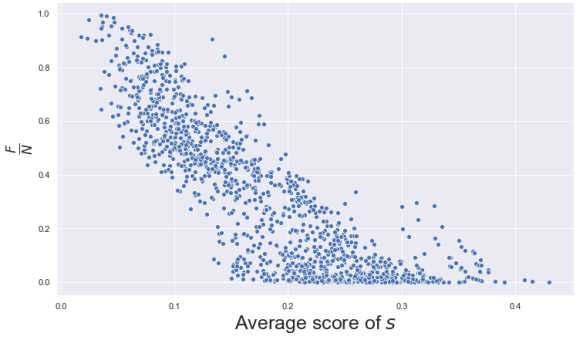


Figure 3: The x-axis is the score of the normalized average score of s given by $\frac{1}{n}\sum_i m_i(s)$ after the metrics have been normalized between 0 and 1. On the y-axis: $\frac{F}{N}$ associated to the sampled summary s. We also report the average performance of current systems.

As a limitation of this study, we can note that the data generation procedure simulates further progress in summarization by stochastically optimizing each evaluation metric. While this constitutes a good approximation, there is no guarantee that high-scoring summaries are sampled with the same distribution as future summarization systems. However, the sampling still covers a large diversity of high-scoring summary and reveal general properties of evaluation metrics.

Other tasks:
Our analysis is performed on TAC-2008 and TAC-2009 because they are benchmark datasets typically used for comparing evaluation metrics. However, our approach can be applied to any dataset. In particular, for future work, this study could be replicated for related fields like *Machine Translation* or *Natural Language Generation*.

5 Conclusion

Evaluation metrics behave similarly on the average scoring range covered by existing human judgment datasets. Thus, we cannot clearly decide which one is the best. Yet, we showed that they will promote very different summaries in the high-scoring range. This disagreement is strong enough that there is no common trend which could be captured by reporting improvements across several metrics. This casts some doubts on the evaluation methodologies in summarization and calls for the collection of human annotations for high-scoring

summaries.

Indeed, since metrics strongly disagree in the high-scoring regime, at least some of them are deviating largely from humans. By collecting human judgments in this specific range, we could identify the best ones using standard meta-evaluation techniques. Such annotations would also be greatly beneficial to improve summarization systems and evaluation metrics alike.

Acknowledgements

This work was partly supported by the German Research Foundation (DFG) as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1, and via the German-Israeli Project Cooperation (DIP, grant No. GU 798/17-1). We also thank the anonymous reviewers for their comments.

References

Ziqiang Cao, Furu Wei, Sujian Li, Wenjie Li, Ming Zhou, and WANG Houfeng. 2015. Learning Summary Prior Representation for Extractive Summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 2, pages 829–833.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 Update Summarization Task. In *Proceedings of the First Text Analysis Conference (TAC 2008)*, pages 1–16.

Hoa Trang Dang and Karolina Owczarzak. 2009. Overview of the TAC 2009 Summarization Track. In *Proceedings of the First Text Analysis Conference (TAC 2009)*, pages 1–12.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *Journal of Artificial Intelligence Research*, pages 457–479.

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. Association for Computational Linguistics.

Kai Hong, John M. Conroy, benoit 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'14)*, pages 1608–1616.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An Information-Theoretic Approach to Automatic Evaluation of Summaries. In *Proceedings of the Human Language Technology Conference at NAACL*, pages 463–470, New York City, USA.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 71–78.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The Challenging Task of Summary Evaluation: An Overview. *Language Resources and Evaluation*, 52(1):101–148.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In *AAAI*, pages 3075–3081.

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, Lisbon, Portugal. Association for Computational Linguistics.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montreal, Canada. Association for Computational Linguistics.

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

Maxime Peyrard and Judith Eckle-Kohler. 2016. A General Optimization Framework for Multi-Document Summarization Using Genetic Algorithms and Swarm Intelligence. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 247 – 257.

Maxime Peyrard and Judith Eckle-Kohler. 2017. A principled framework for evaluating summarizers: Comparing models of summary quality against human judgments. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, volume Volume 2: Short Papers, pages 26–31. Association for Computational Linguistics.

Maxime Peyrard and Iryna Gurevych. 2018. Objective function learning to match human judgements for optimization-based summarization. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 654–660. Association for Computational Linguistics.

Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Çelebi, Danyu Liu, and Elliott Drabek. 2003. Evaluation Challenges in Large-scale Document Summarization. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 375–382.

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*, pages 131–136, Sofia, Bulgaria. Association for Computational Linguistics.

Clifford H Wagner. 1982. Simpson's Paradox in Real Life. *The American Statistician*, 36(1):46–48.

Jaya Kumar Yogan, Ong Sing Goh, Basiron Halizah, Hea Choon Ngo, and C Puspalata. 2016. A Review on Automatic Text Summarization Approaches. *Journal of Computer Science*, 12(4):178–190.

A Correlation with Human Judgments

For each metric that we consider in the paper, we computed its correlation with human judgments in both TAC-2008 and TAC-2009 datasets (Peyrard and Eckle-Kohler, 2017). We used two kinds of human annotations available in these datasets: Responsiveness which is score given by human on 5-point LIKERT scale, and Pyramid where annotators follow the Pyramid annotation guideline to annotate content selection. The correlations are computed with Kendall's τ for each topic and averaged over all topics in both datasets. The results are reported in table 2.

	responsiveness	Pyramid
R-2	.391	.451
R-WE	.378	.431
R-L	.353	.392
JS-2	.379	.444
S3	.403	.477

Table 2: Correlation of automatic metrics with human judgments for TAC-2008 and TAC-2009. The correlation is measured with Kendall's τ.

B Data Generation Algorithm

The general data generation procedure is described by algorithm 1. The function $Score(S, M)$ takes a list S of summaries and a list M of evaluation metrics and outputs a list where each summary has been scored by each evaluation metric in M. The $SampleSummaries$ function is the genetic algorithm introduced genetic algorithm for summarization (Peyrard and Eckle-Kohler, 2016; Peyrard and Gurevych, 2018). The evaluation metric is optimized by the genetic algorithm and the resulting population is a set of summaries ranging from random to upper-bound.

We used a population of $k = 400$. Then, the final dataset contains $160,523$ summaries for an average of $1,763$ summaries per topic (less than $5 * 400$ due to removed duplicates).

This algorithm results in a dataset covering the whole scoring range. In order to filter out low and average scoring summaries, we employ the procedure described by algorithm 2. In this algorithm, the function $Score(\mathcal{T}, m)$ returns a list of all the summaries in the topic $\mathcal{T}$ scored by the metric m. The baseline B is an existing algorithm used as a threshold: for each metric, we keep every summary scoring higher than B. The final set of top-scoring summaries is the union of the top-scoring summaries of each metric.

Algorithm 1: Generate a Dataset of Scored Summaries

Input : $D = \{s_1, \ldots, s_n\}$: document as a set of sentences
L: length constraint
k: number of summaries to generate
$M = \{m_1, \ldots, m_e\}$: evaluation metrics considered

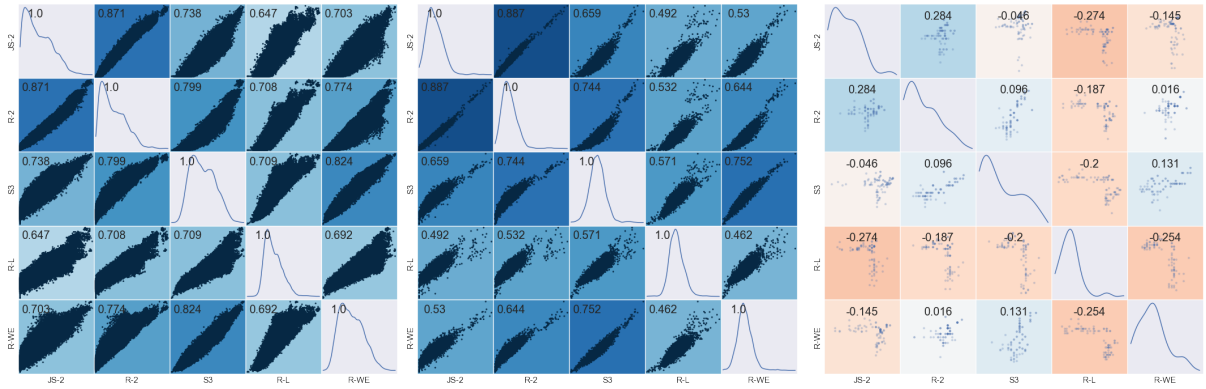
Output: $C = [S_1, \ldots, S_k]$: a set of scored summaries

Function *GenerateData*(D, L, k, M)*:*
  $C := []$
  for $m \in M$ **do**
    $S := SampleSummaries(D, L, k, m)$
    $S := RemoveDuplicate(S)$
    $C \leftarrow Score(S, M)$
  end

For the thresholding, we chose LexRank (Erkan and Radev, 2004), because it is a heavily used baseline. Therefore, most current and future summarization systems should perform better and should be covered by the selected scoring range. Besides, LexRank is strong enough to discard a large number of average scoring summaries. After the selection, we ended up with an average of 102 summaries kept per topic.

C Scatter Matrix Plots: TAC-2008 and TAC-2009

We compute the pairwise correlation between metrics using the existing human judgments (TAC-2008 and TAC-2009). Figure 4 is the scatter matrix plot describing the correlations between pairs of candidate metrics. The number and the cell background color indicate the Kendall's τ between the two metrics. This measures the proportion of pairs of summaries ranked in the same order by both metrics. Thus, the kendall's τ are the ones depicted in the paper in table 1. Diagonal cells represent the score distribution of summaries for the given metric.



(a) Whole scoring range (W). (b) Average-scoring range (A). (c) High-scoring range (T).

Figure 4: Pairwise correlation between evaluation metrics on various scoring range. The generated dataset uses the topics from TAC-2008 and TAC-2009. The human judgments are the ones available as part of TAC-2008 and TAC-2009.

Algorithm 2: Select Top-Scoring Summaries

Input : $D = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$: dataset as a list of topics (each topic contains a list of summaries)
B: baseline algorithm used to decide the high-scoring summaries
$M = \{m_1, \ldots, m_e\}$: evaluation metrics considered

Output: $D^{(top)}$: dataset which contain only top-scoring summaries

Function *SelectTopSummaries*(D, B, M)*:*
  $D^{(top)} := []$
  for $\mathcal{T} \in D$ **do**
    $T^{(top)} := []$
    for $m \in M$ **do**
      $S := []$
      for $s \in Score(\mathcal{T}, m)$ **do**
        if $m(s) > m(B(\mathcal{T}.source))$ **then**
          $S \leftarrow s$
        end
      end
      $T^{(top)} := T^{(top)} \cup S$
    end
    $D^{(top)} \leftarrow T^{(top)}$
  end