# How "Multi" is Multi-Document Summarization?

**Ruben Wolhandler*** **Arie Cattan*** **Ori Ernst** **Ido Dagan**
Computer Science Department, Bar Ilan University
{rwolhandler,arie.cattan,oriern}@gmail.com dagan@cs.biu.ac.il

## Abstract

The task of multi-document summarization (MDS) aims at models that, given multiple documents as input, are able to generate a summary that combines disperse information, originally spread *across* these documents. Accordingly, it is expected that both reference summaries in MDS datasets, as well as system summaries, would indeed be based on such dispersed information. In this paper, we argue for quantifying and assessing this expectation. To that end, we propose an automated measure for evaluating the degree to which a summary is "disperse", in the sense of the number of source documents needed to cover its content. We apply our measure to empirically analyze several popular MDS datasets, with respect to their reference summaries, as well as the output of state-of-the-art systems. Our results show that certain MDS datasets barely require combining information from multiple documents, where a single document often covers the full summary content. Overall, we advocate using our metric for assessing and improving the degree to which summarization datasets require combining multi-document information, and similarly how summarization models actually meet this challenge.[1]

## 1 Introduction

Multi-document Summarization (MDS) consists of creating a short and concise summary that includes the salient information in a set of related documents. Beyond the challenges in single-document summarization, a summary of multiple texts is expected to combine and assemble information spread *across* several input texts. Table 1 illustrates such an example where the summary combines multiple facts from the different documents about global warming. While the main fact ("melting ice") can be described in all source documents, secondary information such as "the rising water" often appear only in certain document(s).

| |
|---|
| **Doc 1**: Indigenous Arctic people urged European countries to step up the fight against global warming, saying it is threatening their societies. The Arctic Council said that the amount of sea ice around the North Pole has decreased about 8 percent in 30 years because of global warming |
| **Doc 2**: One of the topics discussed at the global warming conference is the decrease of the sea ice in the Arctic.. |
| **Doc 3**: Glaciologists worry most about the Arctic ice sheet: if gradually melted, to raise ocean levels worldwide by about five meters stems directly from global warming or from more localized conditions. |
| **Summary**: Global warming has caused the Arctic ice to melt considerably. These changes are threatening the indigenous Arctic population and could raise ocean levels worldwide. |

Table 1: An example of a summary of multiple documents. The proposition "melting ice" (in blue) appears in all source documents, while "the threat for the Arctic population" (in ochre) and "the rising water" (in red) are mentioned only in documents 1 and 3 respectively.

In order to develop MDS models that effectively merge information from various sources, it is necessary that reference summaries in MDS datasets should be based on such dispersed information across the source documents. However, to the best of our knowledge, while existing datasets assume that this property is realized, measuring (automatically) the degree of multi-text merging was not investigated in the literature.

In this work, we suggest quantifying the degree to which a summary is "disperse" in terms of the minimum number of documents needed to cover its content. Accordingly, we develop an automated method for measuring this aspect for any MDS summary. To that end, we first identify the potential provenance of the summary information in all source documents. Then, for each possible number of documents, we form the subset of documents that includes the largest amount aligned informa-

*Equal contribution.
[1]Our code is available in https://github.com/ariecattan/multi_mds.



*Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5761 - 5769
December 7-11, 2022 

tion with the summary. Finally, we define the degree of multi-text merging of an MDS summary as a function of the amount of summary information *not* covered by each subset of documents.

We apply our automated measure to evaluate the degree of multi-text merging in four prominent MDS datasets (DUC, TAC, MultiNews and WCEP) as well as the output of five recent systems. Our results show that some existing datasets barely involve multi-text merging because the reference summary information mostly appears in a single document. Unsurprisingly, the length of the summary has a substantial impact on the amount of multi-text merging since longer summaries cover more detailed information which tends to be spread across documents.

Taken together, our work is the first to measure and empirically analyze multi-text merging in MDS datasets and model summaries. We suggest that future work will use our methodology to develop better datasets and to improve the degree of multi-text merging in MDS models.

## 2 A Measure for Multi-text Merging

### 2.1 Motivating Analysis

The common dataset structure for an MDS instance is a topic that consists of a set of source documents $D = \{D_1, ..., D_n\}$ and a summary $S$. To motivate our measure, we first analyze the degree of multi-text merging on a sample of topics. To that end, we leverage the Summary-Source-Alignment dataset of Ernst et al. (2021), in which human annotators aligned all propositions in reference summaries with corresponding propositions in the source documents that cover the same information, as exemplified in Table 1. Given these alignments on 9 MDS topics from MultiNews (Fabbri et al., 2019), each composed of 4 source documents, we find that a single source document suffices to cover alone 70% of the summary propositions while 2 documents cover 95% of them. The remaining source documents thus hardly provide any substantial information to the summary.

Motivated by this analysis, we develop an automated measure that allows to evaluate the degree of multi-text merging in entire MDS datasets and in systems summaries. Our measure operates in the following steps. We first define the coverage score for a given subset of source documents (§2.2). Then, to approximate the minimum number of documents required to cover increasing portions of the summary information, we greedily construct, for each possible number of source documents, the subset of source documents with the highest coverage score (§2.3). Finally, we measure the total amount of summary information in all subset sizes, yielding a corresponding coverage curve (§2.4).

### 2.2 Relative Coverage Score

Let $D^*$ be a subset of source documents $D^* \subseteq D$. We define the *relative coverage* of $D^*$ as the proportion of information that is covered by $D^*$, normalized by the information covered by all source documents $D$:

$$cov(D^*, D, S) = \frac{s(D^*, S)}{s(D, S)} \quad (1)$$

For the absolute coverage score $s(D^*, S)$, we aim to approximate the human annotation of summary-source proposition alignment in (Ernst et al., 2021), which is based on the well established Pyramid scheme (Nenkova and Passonneau, 2004). Specifically, we follow their automated scheme: (1) we extract all propositions from the summary and all source documents using OpenIE (Banko et al., 2008),[2] (2) we compute the similarity score between the propositions in the summary and the source documents using SUPERPAL, an NLI model fine-tuned on proposition alignment (Ernst et al., 2021), (3) $s(D^*, S)$ is defined as the number of propositions in $S$ that are aligned with some proposition in $D^*$.

We consider the proportion $s(D^*, S)/s(D, S)$ and not the absolute coverage $s(D^*, S)$ for two main reasons. First, as both reference and system summaries are known to include hallucinated information (Maynez et al., 2020), we need to discard them in our measure in order to properly estimate the amount of information that each single source document actually provides to the summary. Second, normalizing the coverage score will mitigate the potential omissions of the alignment model.

### 2.3 Maximally-Covering Document Subsets

Given an MDS topic with $n$ source documents, we aim to measure the maximal content coverage of the summary content by a document subset of size $k \leqslant n$. To that end, we form $n$ subsets of source

[2] We use the AllenNLP implementation of (Stanovsky et al., 2018) to extract the OpenIE tuples. Following Ernst et al. (2021, 2022), we convert each OpenIE tuple into a proposition string by concatenating the predicate and its arguments by their original order.

documents $\{D_1^*, ..., D_n^*\}$, such that each subset $D_k^*$ includes an optimized set of $k$ source documents, covering the maximal amount of summary information. Specifically, we employ a greedy approach where we add one source document at a time to maximize the increase in the relative coverage score (Eq. 1), as follows:[3],

$$D_{k+1}^* = \underset{d \in D \setminus D_k^*}{\text{argmax}}\, cov(D_k^* \cup d, D, S) \quad (2)$$

That is, $D_1^*$ includes the single document $d_1$ with the maximal coverage score, $D_2^*$ adds the document $d$ that marginally contributes the most to the coverage, and so on.

### 2.4 Dispersion Score: Area Above the Curve

Given the $n$ optimal subsets (in the above greedy sense) of source documents for a given topic $t$, let $cov_k^t$ be the coverage of the subset $D_k^*$. We then define the overall coverage for an entire MDS dataset as the average coverage score for each topic: $cov_k = \frac{1}{T} \sum_{t \in T} cov_k^t$. Accordingly, $cov_k$ aims to measure the average amount of summary information that $k$ source documents cover.

To allow qualitative analysis of the degree of multi-text merging, we plot the curve of $cov_k$ for $k \in [1, n]$, as illustrated in Figure 1. Since the optimal subsets are formed incrementally, $cov_k$ is a monotonic non-decreasing function whose maximum is 1. As shown in Figure 1, the curve corresponding to WCEP-10 is close to 1 already when $k = 1$ (single source document), while the curve of DUC 2006-2007 gradually increases to 1, indicating that a larger number of source documents are required to cover the summary content.

While visual plots are insightful, we are also interested in *quantifying* how slowly $cov_k$ converges to 1. Analogously to $cov_k$, the difference $1 - cov_k$, expressed by the area above the $cov_k$ curve, aims to measure the average amount of summary information that $k$ documents do *not* cover. Therefore, we define our "dispersion" score as the Area Above the $cov_k$ Curve (AAC).[4] The higher the AAC, the more multi-text merging is required. To properly compare the degree of multi-text merging on various datasets with different numbers of source documents, we normalize the AAC by the maximum number of required source document, $n_{\text{MAX}}$.[5] The exact formula of the AAC becomes:[6]

$$AAC = \frac{1}{n_{\text{MAX}}} \sum_{k=1}^{n} 1 - cov_k \quad (3)$$

It is important to note that we avoid normalizing our dispersion metric by the number of source documents in a topic, aiming to better reflect the absolute degree of dispersity. To illustrate, consider one dataset with say 3 documents per cluster, where 2 are sufficient for covering the reference, vs. a dataset with 20 documents per cluster, where 5 are sufficient. Our (non-normalized) score would clearly favor the second dataset, fitting our main motivation in this paper, while a normalized score would counter-productively favor the first.

### 2.5 Related Evaluation Practices

When evaluating a new summarization model, beyond lexical similarity between the system and reference summary (ie. ROUGE), it is common to also report specific aspects such as relevance and informativeness (Jung et al., 2019; Peyrard, 2019), faithfulness (Kryscinski et al., 2020; Scialom et al., 2021), grammar, redundancy (Lloret and Palomar, 2013; Xiao and Carenini, 2020), or sentence fusion (Lebanoff et al., 2019). Other aspects such as extractiveness or compression were investigated for reference summaries (Grusky et al., 2018).

Our dispersion measure aims to measure, both for reference and system summaries, the degree of multi-text merging, which is an essential aspect in multi-document summarization.

## 3 Empirical Analysis

To assess our automated dispersion measure, we apply our method on the exact same MultiNews topics that we used for our motivating analysis (§2.1). As illustrated in Appendix B.1, we observe a similar behaviour in our automated measure, assessing its effectiveness. We cannot evaluate correlation due to a lack of statistical significance, as we have annotations only on 9 topics.

[3] We also tested the (exponential) optimal approach for finding $D_k^*$ on two datasets where this was feasible, namely Multi-news and DUC 2003-2004, and found no significant difference in AAC scores vs. the greedy approach (up to 0.1 difference), suggesting its sufficiency.

[4] It is easy to see that this definition of the AAC score is equivalent to the average of the individual per-topic AAC scores across the dataset.

[5] In particular, we set $n_{\text{MAX}} = 10$ based on Figure 1.

[6] Technically, our formula for calculating the AAC score considers the area above the *non-interpolated* curve. This is consistent with the common view by which the Average Precision score is often considered as the area under the non-interpolated recall-precision curve.

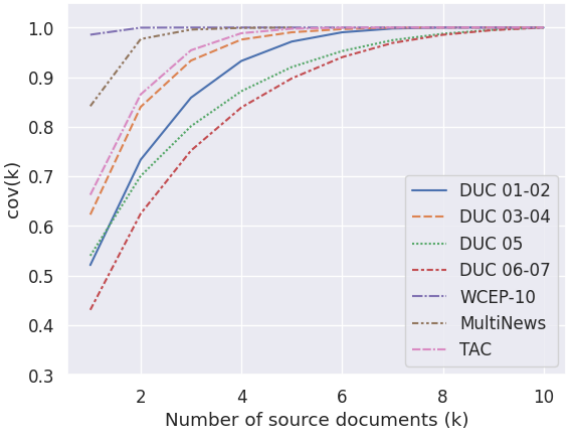


Figure 1: Curves of $cov_k$ on MDS datasets. Slow convergence to 1, as for DUC 06-07, reflects highe degree of multi-text merging.

| | #docs | length | AAC |
|---|---|---|---|
| WCEP-10 | 10 | 28 | 0.1 (±0.7) |
| MultiNews | 2.8 | 218 | 1.8 (±2.3) |
| TAC | 10 | 100 | 5.4 (±4.1) |
| DUC 2003-04 | 10 | 102 | 6.6 (±5.0) |
| DUC 2001-02 | 10 | 235 | 10.2 (±5.7) |
| DUC 2005 | 30 | 250 | 13.4 (±9.6) |
| DUC 2006-07 | 25 | 250 | 16.5 (±9.1) |

Table 2: AAC scores and standard deviation of the reference summary in several datasets. We group the DUCs with equal summary's length (number of tokens) and with equal number of source documents.

**Reference Summaries** We compute the AAC scores on the reference summaries of DUC 2001 to 2007 (NIST, 2014), TAC 2008 to 2011 (NIST), MultiNews (Fabbri et al., 2019) and WCEP-10 (Gholipour Ghalandari et al., 2020). We report the AAC score for each dataset, as well as the standard deviation of the individual AAC scores across the dataset topics, in Table 2 and plot the $cov_k$ curves in Figure 1.

The AAC score is 0.1 for WCEP-10 and 1.8 for MultiNews, showing a rather poor degree of multi-text merging, where the summary information mostly appears in a single document (see Figure 1). Early datasets such as DUC and TAC exhibit a higher degree of multi-text merging, with the highest 16.5 AAC score for DUC 06-07. We also observe, for all datasets, that the standard deviations of the per-topic AAC scores are of substantial magnitudes, proportional to the average AAC score across the dataset. This indicates that the various topics in each dataset present varying levels of dispersity, as can be expected in real-world scenarios.

| | DUC 2004 | TAC 2011 |
|---|---|---|
| Reference | **7.2** (±5.3) | **4.6** (±3.7) |
| RL-MMR | 2.6 (±2.4) | 1.9 (±2.9) |
| PG-MMR | 2.5 (±2.6) | 2.8 (±3.2) |
| LexRank | 3.7 (±3.3) | 3.2 (±2.9) |
| PG | 4.7 (±3.1) | 3.1 (±2.6) |
| ProCluster | 4.1 (±3.7) | 3.9 (±4.4) |

Table 3: AAC scores and standard deviation on different systems. For both DUC and TAC, reference summaries are more disperse than system summaries.

Further, we find that the length of the summary has an impact on the degree of multi-text merging. The longer the reference summary the more diverse it becomes. Indeed, DUC datasets with summary length of 100 tokens have a lower AAC score than those with length of more than 200 tokens. These results seem intuitive because the more concise the summary, the more salient the information in it is, and hence is more likely to appear in most source documents (e.g "melting ice" in Table 1). Nevertheless, the summary length is not the only factor, as on MultiNews we obtain a low AAC score, whereas the reference summary length is longer than 200 tokens on average.

**System Summaries** We also compute the AAC scores on the output of several system summaries: LexRank (Erkan and Radev, 2004), PG (See et al., 2017), PG-MMR (Lebanoff et al., 2018), RL-MMR (Mao et al., 2020) and ProCluster (Ernst et al., 2022), when tested on DUC 2004 and TAC 2011. The AAC scores are presented in Table 3 and the curves of $cov_k$ are shown in Appendix B.2. We notice that, for both DUC 2004 and TAC 2011, the AAC score of the reference summary is slightly higher than the AAC score of all systems summaries. The curves of systems summaries (Appendix B.2) have similar trends to the curves of the reference summaries, which leads us to believe that the more disperse the dataset will be, the more disperse the development of the systems will be.

**How MDS models benefit from training on multiple documents?** To answer this question, we fine-tune PRIMERA (Xiao et al., 2022), a state-of-the-art MDS model, to generate the summary given only a single document or two source documents as input, on multiple datasets. For this experiment,

| | $D_{\text{ALL}}$ | | | $D_1^*$ | | | $D_2^*$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | R-1 | R-2 | R-L | R-1 | R-2 | R-L | R-1 | R-2 | R-L |
| DUC | 36.4 | 8.7 | 19.7 | 31.3 | 7.0 | 17.6 | 34.3 | 8.2 | 18.8 |
| TAC | 36.6 | 10.0 | 20.7 | 39.3 | 14.4 | 23.9 | 37.7 | 11.0 | 21.2 |
| MultiNews (ours) | 47.9 | 19.4 | 25.0 | 48.0 | 19.5 | 25.2 | 48.3 | 19.6 | 25.1 |
| WCEP-10 | 45.8 | 25.0 | 37.6 | 46.4 | 24.8 | 37.3 | 46.7 | 25.2 | 37.8 |

Table 4: Results of PRIMERA when fine-tuning on all source documents ($D_{\text{ALL}}$), the source document with highest coverage ($D_1^*$) and the two source documents with highest coverage ($D_2^*$). Our results on MultiNews slightly differ from the original PRIMERA paper (Xiao et al., 2022) because we fine-tune with a smaller number of optimization steps (See App. C)

we select the source document(s) with the highest coverage score (§2.2), for both training and inference. More implementation details are presented in Appendix C. We report the results for each dataset in Table 4. For all datasets except for DUC, models trained on a subset of documents achieve comparable or higher ROUGE scores than models trained on the entire dataset. This finding hints that it may be worthwhile to explore MDS modeling architectures that first identify such a salience-covering subset, and then feed this reduced set to the summarization system, with the aim of easing the summarization step.

**Challenge: Applicability to non-news domains** As a preliminary attempt to examine the applicability of our dispersity measure implementation to non-news domains, we examined MSˆ2 (DeYoung et al., 2021), an MDS dataset from the biomedical domain. However, we observe that the results are not reliable for this type of data due to insufficient quality of the proposition alignment step, for two main reasons. First, abbreviations of technical terms are very common in scientific papers, which presents a challenge to the SuperPAL aligner that we use. For example, consider the document proposition *"The use of a Quadriceps tendon graft in primary ACL reconstruction leads to equal or better functional outcomes."*. In this case, SuperPAL failed to predict alignment with the summary proposition *"**QT** autograft represent a feasible option for primary ACL reconstruction."*, which uses the abbreviation *"QT"*. However, when we replaced the abbreviation with the full term *"Quadriceps Tendon"*, which appears in the document proposition, then SuperPAL successfully predicted alignment. The abundance of such cases indicates the need to adopt proposition alignment tools to their target domain, particularly with respect to technical terminology and abbreviations.

A second challenge stems from our use of a proposition alignment tool in order to track the evidence supporting a certain summary proposition. In fact, taking this approach assumes that the evidence in the source document should entail the summary proposition, which is mostly the case in the news domain. However, we observed that in the scientific domain of our data, a summary proposition often synthesizes information from multiple source evidences, yielding a novel consolidating proposition that is entailed only from the aggregation of multiple source evidences. For example, a typical such situation occurs when source documents include contradictory evidences or different perspectives, which is typical in scientific and other types of texts, while the summary synthesizes this evidences by pointing at the disagreements. To address this type of cases, more complex mechanisms of evidence tracking would be needed in order to compute dispersity, rather than just using 1:1 proposition alignment for this purpose.

## 4 Conclusion

We propose a measure to evaluate the degree of multi-text merging, an essential aspect in multi-document summarization. Using this measure, we found that some prominent MDS datasets contain summaries that hardly combine information from multiple input sources. Furthermore, we show that fine-tuning on a single effective document, achieves better summary performance than fine-tuning on the full set of documents.

## 5 Limitations

As described in Section 2.4, we use the SuperPAL model (Ernst et al., 2021) to predict summary-source alignment scores. Similarly to recent model-based evaluation metrics (e.g FactCC, BERTScore,

etc.), our measure can also include noise due to some inaccurate predictions or proposition extraction. In addition, the SuperPAL model is time- and computation hardware- consuming because it assigns a BERT-based score for every summary-source proposition pairs.

## Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work was supported in part by Intel Labs, the Israel Science Foundation grant 2827/21 and by a grant from the Israel Ministry of Science and Technology. Arie Cattan is partially supported by a fellowship for excellence in data science from the Bar-Ilan Data Science Institute (funded by the Israeli PBC).

## References

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2008. Open information extraction from the web. In *CACM*.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Wang. 2021. MSˆ2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan. 2022. A proposition-level clustering approach for multi-document summarization. In *NAACL*.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021. Summary-source proposition-level alignment: Task, datasets and supervised baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322, Online. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in 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 3324–3335, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Analyzing sentence fusion in abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Elena Lloret and Manuel Palomar. 2013. Tackling redundancy in text summarization through different levels of language analysis. *Comput. Stand. Interfaces*, 35:507–518.

Yuning Mao, Yanru Qu, Yiqing Xie, Xiang Ren, and Jiawei Han. 2020. Multi-document summarization with maximal marginal relevance-guided reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1737–1751, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings*

*of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

NIST. Document understanding conferences. https://tac.nist.gov. Accessed: 2019-12-02.

NIST. 2014. Document understanding conferences. https://duc.nist.gov. Accessed: 2019-12-02.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. 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, Vancouver, Canada. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. 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 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Wen Xiao and Giuseppe Carenini. 2020. Systematically exploring redundancy reduction in summarizing long documents. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 516–528, Suzhou, China. Association for Computational Linguistics.

## A Datasets

Here, we describe the datasets that we use in the paper.

**DUC** DUCs (2001 to 2007) (NIST, 2014) are modest MDS datasets in the news domain. Each year consists of 30-60 topics where each topic is composed of 10-30 source documents and 2-4 human-written (reference) summaries. Table 5 summarizes the dataset information for each separate year.

**TAC** TAC (2008, 2009, 2010, 2011) (NIST) are similar to DUC datasets. Each year consists of approximately 45 topics, where each topic includes 10 source documents and 4 human-written summaries.

**MultiNews (Fabbri et al., 2019)** This dataset includes news articles and human-written summaries taken from the site `newser.com`. MultiNews is very large and the training, test and validation are composed of 44972, 5622 and 5622 topics respectively. The average number of source documents is 2.8 documents (range from 2 to 11 documents).

**WCEP-10 (Gholipour Ghalandari et al., 2020)** This dataset includes news events article retrieved from the Wikipedia Current Events Portal. The reference summary is human-written. The summary must be short (30-40 words) and each summary is one complete sentence. WCEP-10 is a truncated version of WCEP with a maximum number of source documents fix to 10. WCEP-10 is composed of 8158, 1022, 1020 topics for the train, validation and test set respectively.

## B Additional Results

### B.1 Assessment of Our Measure

Figure 2 shows the graphs of $cov_k$ according to human annotation in SSA and our measure.

### B.2 Graphs of $cov_k$

Figures 3 and 4 shows the curves of $cov_k$ on DUC 2004 and TAC 2011 systems summaries respectively. We can notice differents between the different system summaries, which indicate us that some models have higher degree of multi-text merging than other.

| Year | #Topics | #Docs | #Sums by topic | #Sums |
|---|---|---|---|---|
| 2001 | 30 | 10 | 3 | 90 |
| 2002 | 60 | 10 | 2 | 120 |
| 2003 | 30 | 10 | 4 | 120 |
| 2004 | 50 | 10 | 4 | 200 |
| 2005 | 50 | 30 | 6 | 300 |
| 2006 | 50 | 25 | 4 | 200 |
| 2007 | 45 | 25 | 4 | 180 |

Table 5: DUC's characteristics for each year.

| Year | #Topics | #Docs | #Sums by topic | #Sums |
|---|---|---|---|---|
| 2008 | 48 | 10 | 4 | 192 |
| 2009 | 44 | 10 | 4 | 176 |
| 2010 | 46 | 10 | 4 | 184 |
| 2011 | 44 | 10 | 4 | 176 |

Table 6: TAC's characteristics for each year.

## C Training on a Subset of Source Documents

In order to properly assess the effect of fine-tuning an MDS model with only one or two source documents, we fine-tune also the MDS variant (given all source documents) with the same optimization steps. Table 7 presents the optimization and warm-up steps that were used for each dataset.

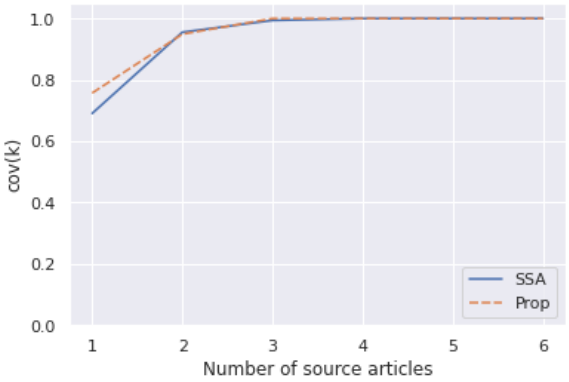


Figure 2: Curves of $cov_k$ using the SSA annotations (Ernst et al., 2021) vs. our measure, for 9 topics from MultiNews

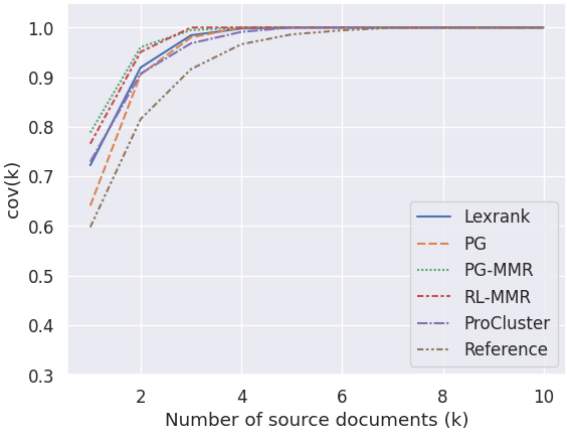


Figure 3: Curves of $cov_k$ on DUC 2004 systems summaries. The systems are trained on DUC 2003. The curves of ProCluster and PG are closer to the reference than Lexrank, PG-MMR and RL-MMR.

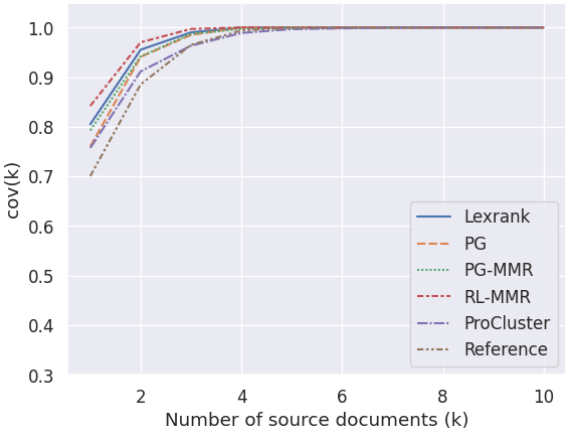


Figure 4: Curves of $cov_k$ on TAC 2011 systems summaries. The systems are trained on TAC 2008, 2009 and 2010. The curves of ProCluster and PG are closer to the reference than Lexrank, PG-MMR and RL-MMR.

| Dataset | Total Steps | Warmup Steps |
|---|---|---|
| DUC | 20 | 2 |
| TAC | 100 | 10 |
| MultiNews | 10k | 1k |
| WCEP-10 | 5k | 500 |

Table 7: Details of total and warm-up steps used for training the models with a single or all source documents, as described in Section 3. We use the same number of steps for both experiments.