M2DS: Multilingual Dataset for Multi-document Summarization

Presented by:
Kushan Hewapathirana[1,2]

Co-authors:
- Nisansa de Silva[1]
- C.D. Athuraliya[2]

1. Department of Computer Science and Engineering, University of Moratwua, Sri Lanka.
2. ConscientAI, Sri Lanka.

Contents

INTRODUCTION

Introduction to Multi-document Summarization (MDS)

MDS is an automatic process that aims to extract relevant information from multiple texts written about the same topic and represent it in a short piece of text [1]

Complex relationships between different documents, making it more intricate than single-document summarization (SDS) [1,2]

Current MDS datasets are largely English-centric, limiting their global applicability [2]

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022

RESEARCH PROBLEM

Multilingual Requirement in MDS

- The world has over 7,000 languages, but most MDS research focuses only on English [3]
- With only 380 million native English speakers, there's a critical need for multilingual datasets to serve a broader global audience [3,4]
- M2DS addresses this gap by introducing a multilingual dataset that includes document-summary pairs in five languages
- Facilitate the development of robust MDS models across diverse languages, including low-resource languages

[3] Eberhard, D.M., G.F.S., Fennig, C.D.: Ethnologue: languages of the Americas and the pacific (2023)
[4] Giannakopoulos, G.: Multi-document multilingual summarization and evaluation tracks in ACL 2013 multiling workshop. In: Proceedings of the Multiling 2013 Workshop on Multilingual Multi-document Summarization, pp. 20–28 (2013)

Research Objectives

- To Introduce the first comprehensive multilingual MDS dataset, M2DS
- Evaluate state-of-the-art MDS models on this multilingual dataset

RELATED WORK

Existing MDS Datasets

- DUC and TAC: Early benchmarks primarily focused on the news domain [2].
- Multi-News: Provides substantial size and traceability in the news domain [4].
- WikiSum: Leverages Wikipedia and search engine results for abstractive summarization [5].
- BigSurvey: Contribute to scientific writing, focusing on comprehensive summaries [6].

[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[4] Fabbri, A.R., Li, I., She, T., Li, S., Radev, D.: Multi-news: a large-scale multi- document summarization dataset and abstractive hierarchical model. In: ACL, pp. 1074–1084 (2019)
[5] Liu, P.J., Saleh, M., et al.: Generating wikipedia by summarizing long sequences. arXiv preprint arXiv:1801.10198 (2018)
[6] Liu, S., Cao, J., Yang, R., Wen, Z.: Generating a structured summary of numerous academic papers: dataset and method. arXiv preprint arXiv:2302.04580 (2023)

Challenges in MDS Research

- MDS datasets are relatively scarce compared to single-document summarization (SDS) datasets [1,2].

- Most datasets and research efforts are limited to English, creating a gap in multilingual summarization [4]

- Progress in multilingual MDS is limited, with most research focusing on single-document summarisation (SDS) and cross-lingual summarisation (CLS) [7].

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[4] Giannakopoulos, G.: Multi-document multilingual summarization and evaluation tracks in ACL 2013 multiling workshop. In: Proceedings of the Multiling 2013 Workshop on Multilingual Multi-document Summarization, pp. 20–28 (2013)
[7] Elhadad, M., Miranda-Jim´enez, S., Steinberger, J., Giannakopoulos, G.: Multi- document multilingual summarization corpus preparation, Part 2: Czech, Hebrew. In: Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pp. 13–19 (2013)

Existing MDS Models

Transformer-Based Models [1,8]:

- BERTSUM: Uses a hierarchical encoder for summarization tasks
- BART: Designed as a denoising auto-encoder
- PEGASUS: Leverages self-supervised learning
- T5: A text-to-text transformer model
- PRIMERA: Based on LongFormer Encoder-Decoder (LED) architecture, known for synthetic summary generation during pre-training

Challenges [1]:

- Difficulty in reflecting conflicting information
- Limited research in multilingual MDS, with most models focusing on English

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[8] Xiao, W., Beltagy, I., Carenini, G., Cohan, A.: Primera: pyramid-based masked sentence pre-training for multi-document summarization. In: ACL, pp. 5245–5263 (2022)

Existing MDS Models on different domain datasets

Dataset		PRIMERA	PEGASUS	LED
Multi-News	R-1	**42.0***	32.0*	17.3*
	R-2	**13.6***	10.1*	3.7*
	R-L	**20.8***	16.7*	10.4*
Multi-Xscience	R-1	**29.1***	27.6*	14.6*
	R-2	**4.6***	4.6*	1.9*
	R-L	**15.7***	15.3*	9.9*
WikiSum	R-1	**28.0***	24.6*	10.5*
	R-2	**8.0***	5.5*	2.4*
	R-L	**18.0***	15.0*	8.6*
Rotten Tomatoes	R-1	25.4$^\bullet$	**27.4**$^\bullet$	25.6$^\bullet$
	R-2	8.4$^\bullet$	**9.5**$^\bullet$	8.0$^\bullet$
	R-L	19.8$^\bullet$	**21.1**$^\bullet$	19.6$^\bullet$

METHODOLOGY

Experimental Setup

Data:

Dataset used:

M2DS

Baselines:

Fine-tuned PRIMERA, PEGASUS, LD models.

Zero-shot – Llama 2

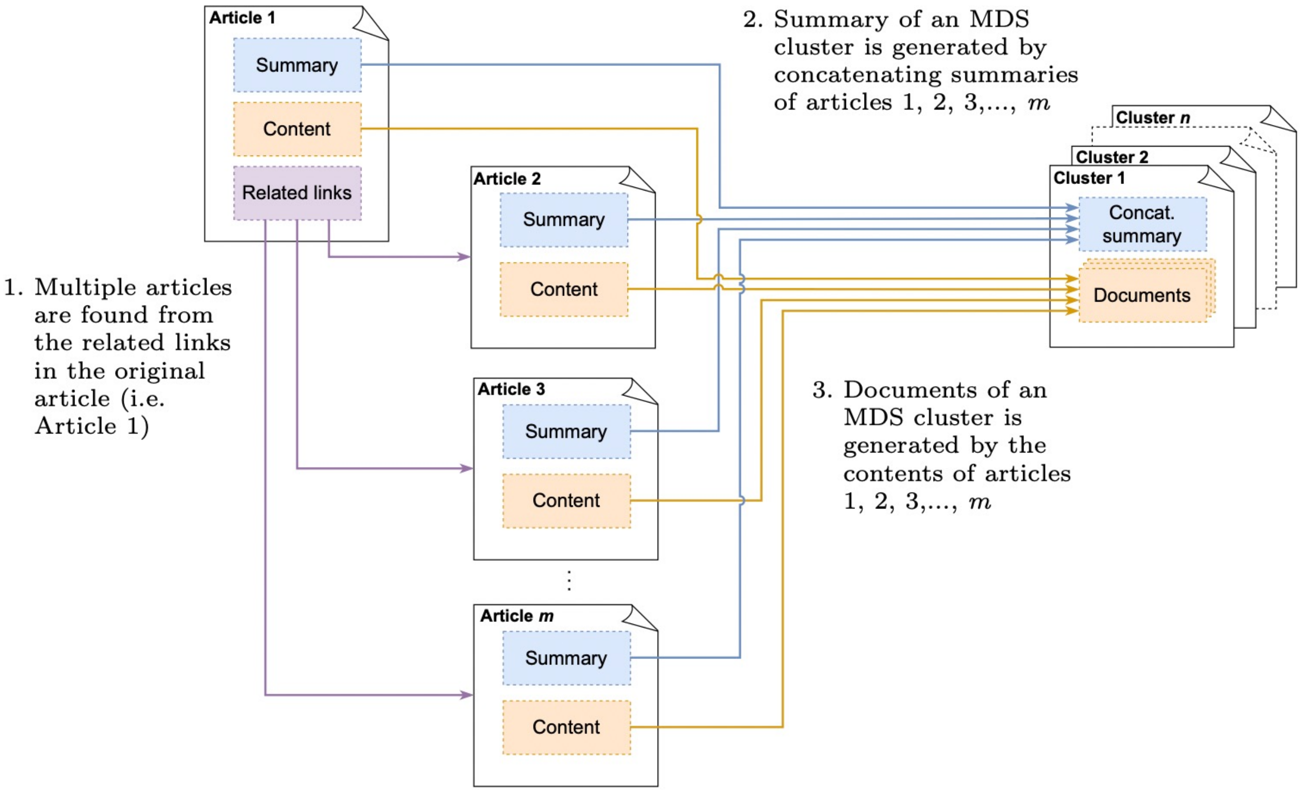
Experimental Process:

Fine-tuned all evaluated models with cross-entropy loss on all datasets. Used Adam optimizer with a learning rate of 5e 5, and without any warm-up or weight decay.

Experimental Environment:

on single A100 GPU.

Dataset Construction



M2DS Dataset

- Data Collection:
  - Sources: M2DS dataset is derived from BBC news articles in five languages: English, Japanese, Korean, Tamil, Sinhala.
  - Timeframe: Articles span from 2010 to 2023, providing a comprehensive dataset for multilingual MDS research.
- Dataset Structure:
  - Document-Summary Pairs: Each language includes document-summary pairs where summaries are generated by concatenating the summaries of related articles
  - Cluster Formation: Articles are grouped into clusters based on related content, and summaries are generated by combining the related summaries.

M2DS Dataset...

- Preprocessing and Data Standardization:
  - Consistency Across Languages: Ensured consistent formatting and preprocessing across all languages to facilitate effective model training.
  - Focus on Clean Data: Structured the dataset to provide clean, organized data suitable for training robust MDS models.
- Baseline Model Evaluation:
  - Models Evaluated: PRIMERA, PEGASUS, and LED.
  - Languages Evaluated: Performance was measured across all five languages in the M2DS dataset.

M2DS Dataset...

Dataset	No. of documents	No. of clusters	Avg. no. of documents per cluster	Domain
Multi-News•	56.0k*	16.0k	3.5*	News articles•
Multi-Xscience◦	40.0k*	14.0k	2.8*	Related work section in scientific articles◦
Wikisum¦	1.5M*	37.5k	40.0*	Wikipedia articles¦
BigSurvey-MDS¢	430.0k*	7.0k	61.4*	Human-written survey papers on various domains¢
PEERSUM‖	11.9k‖	1.5k	7.8‖	Peer reviews of scientific publications
MS^2†	470.0k†	20.0k	23.5†	Reviews of scientific publications in medical domain†
Rotten Tomato Dataset↑	244.0k‡	9.0k	26.8‡	Movie reviews‡
M2DS	180.0k	51.5k	3.5	News articles
- English	67.0k	17.0k	3.9	
- Tamil	32.0k	10.0k	3.2	
- Japanese	29.0k	11.0k	2.6	
- Korean	27.0k	8.0k	3.4	
- Sinhala	23.5k	5.5k	4.2	

The sources are as follows: ∗Xiao et al.(2022) [41], †DeYoung et al.(2021) [6], ‡DeYoung et al.(2023) [7], •Fabbri et al.(2019) [10], ◦Lu et al.(2020) [27], ‖Li et al.(2022) [22], ¦Liu et al.(2018) [24], ¢Liu et al.(2023) [25], ↑Leong et al.(2020) [19].

EVALUATION
&
RESULTS

Performance Comparison (English): Zero-shot vs Fine-tuned

Language		Models					
		PRIMERA	PRIMERA (fine-tuned)	PEGASUS	PEGASUS (fine-tuned)	LED	LED (fine-tuned)
English	R-1	23.6	**28.7**	18.6	22.5	17.1	20.5
	R-2	8.8	**12.3**	9.1	9.9	7.1	10.1
	R-L	13.6	**17.1**	12.4	14.7	13.2	15.2

Performance Comparison (All Languages): Fine-tuned

Language		Models						
		LEAD-3	RANDOM	CENTROID	PRIMERA	PEGASUS	LED	Llama 2
Sinhala	R-1	0.06	5.7	4.5	5.7	4.1	3.6	**20.2**
	R-2	0.0	0.05	0.1	2.2	2.1	1.9	**6.5**
	R-L	0.06	5.1	3.9	3.2	2.8	2.9	**17.3**
Japanese	R-1	3.5	2.3	1.9	6.3	5.7	5.9	**7.7**
	R-2	0.0	0.01	0.05	**3.2**	1.3	1.4	0.8
	R-L	3.5	1.9	1.7	4.1	3.3	2.7	**6.8**
Korean	R-1	2.4	1.4	1.3	5.4	5.5	4.6	**8.5**
	R-2	0.4	0.02	0.03	1.1	**1.4**	0.8	1.0
	R-L	2.3	1.3	1.3	2.3	2.9	1.9	**8.1**
Tamil	R-1	6.8	1.6	2.2	4.4	3.8	3.7	**10.2**
	R-2	0.9	0.0	0.06	1.1	0.7	0.4	**3.1**
	R-L	6.2	1.6	1.9	2.2	1.7	1.3	**9.8**
English	R-1	1.2	6.4	7.6	**28.7**	22.5	20.5	20.8
	R-2	0.0	0.05	3.8	12.3	9.9	10.1	**13.5**
	R-L	1.1	5.7	7.6	17.1	14.7	15.2	**19.2**

Observations

- Llama 2 7B Performance: Outperformed all models, showcasing its robustness across the dataset.
- PRIMERA's Strength in English: Slightly better performance in English, highlighting its ability to capture language-specific nuances.
- Performance Drop in Multilingual Dataset: Models fine-tuned on our dataset showed a noticeable decline in performance compared to English-centric datasets
- LEAD-3 Lower Scores: Our dataset's higher quality, compared to TAC/DUC datasets, is evident from the lower LEAD-3 scores, indicating less bias toward the first three sentences

Model Insights

- Task-Specific Models vs. LLMs: PRIMERA's superior English performance suggests that simpler, task-specific models can outperform large language models (LLMs) like Llama 2 without extensive fine-tuning

- Fine-Tuning Benefits: Fine-tuning improved model performance across the board, with PRIMERA showing the largest increase from 23.6 to 28.7

- Scalability and Transfer Learning: Llama 2's scalability suggests potential for handling larger datasets. Future research should explore Transfer Learning to minimize performance drops and enhance model adaptability across languages

CONCLUSIONS

Conclusions and Future Research Directions

- M2DS Dataset: Introduced as a pioneering multilingual MDS dataset, filling the gap in multilingual representation.
- Five-Language Coverage: M2DS stands out with document-summary pairs across five languages, contributing uniquely to MDS research
- Model Performance: Llama 2 7B demonstrated robust performance, while PRIMERA excelled slightly in English, capturing language-specific nuances effectively

REFERENCES

[1] C. Ma, W. E. Zhang et al., “Multi-document summarization via deep learning techniques: A survey,” ACM Computing Surveys (CSUR), 2020
[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., “A survey on multi-document summarization and domain- oriented approaches,” Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[3] Eberhard, D.M., G.F.S., Fennig, C.D.: Ethnologue: languages of the Americas and the pacific (2023)
[4] Giannakopoulos, G.: Multi-document multilingual summarization and evaluation tracks in ACL 2013 multiling workshop. In: Proceedings of the Multiling 2013 Workshop on Multilingual Multi-document Summarization, pp. 20–28 (2013)
[5] Liu, P.J., Saleh, M., et al.: Generating wikipedia by summarizing long sequences. arXiv preprint arXiv:1801.10198 (2018)
[6] Liu, S., Cao, J., Yang, R., Wen, Z.: Generating a structured summary of numerous academic papers: dataset and method. arXiv preprint arXiv:2302.04580 (2023)
[7] Elhadad, M., Miranda-Jim´enez, S., Steinberger, J., Giannakopoulos, G.: Multi- document multilingual summarization corpus preparation, Part 2: Czech, Hebrew. In: Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pp. 13–19 (2013)
[8] Xiao, W., Beltagy, I., Carenini, G., Cohan, A.: Primera: pyramid-based masked sentence pre-training for multi-document summarization. In: ACL, pp. 5245–5263 (2022)
[9] . Dos Santos and M. Gatti, “Deep convolutional neural networks for sentiment analysis of short texts,” in Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, 2014, pp. 69–78
[10] Z. Cao, F. Wei, S. Li, W. Li, M. Zhou, and H. Wang, “Learning summary prior representation for extractive summarization,” in ACL, 2015, pp. 829–833
[11] H. Jin, T. Wang, and X. Wan, “Multi-granularity interaction network for extractive and abstractive multi-document summarization,” in Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 6244–6254
[12] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text summarization branches out, 2004, pp. 74–81
[13]P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, “Generating wikipedia by summarizing long sequences,” arXiv preprint arXiv:1801.10198, 2018
[14] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, “Multi-news: A large- scale multi-document summarization dataset and abstractive hierarchical model,”
[15] M. Koupaee and W. Y. Wang, “Wikihow: A large scale text summarization dataset,” arXiv preprint arXiv:1810.09305, 2018
[16] S. Leon, “Rotten tomatoes movies and critic reviews dataset,” https://bit.ly/RTdataset, 2020, (Accessed on 06/24/2023)
[17]C. Raffel, N. Shazeer, A. Roberts, K. Lee, and others, “Exploring the limits of transfer learning with a unified text-to-text transformer,” The Journal of Machine Learning Research, vol. 21, no. 1, pp. 5485–5551, 2020
[18] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” in International Conference on Machine Learning. PMLR, 2020, pp. 11 328–11 339
17] W. Xiao, I. Beltagy et al., “Primera: Pyramid-based masked sentence pre-training for multi-document summarization,” in ACL, 2022, pp. 5245–5263
[18] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” in International Conference on Machine Learning. PMLR, 2020, pp. 11 328–11 339
[19] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” arXiv preprint arXiv:2004.05150, 2020.
[20] . Lu, Y. Dong, and L. Charlin, “Multi-xscience: A large- scale dataset for extreme multi-document summarization of scientific articles,” in EMNLP, 2020, pp. 8068–8074
[21] J. DeYoung, S. C. Martinez, I. J. Marshall, and B. C. Wallace, “Do multi- document summarization models synthesize?” arXiv preprint arXiv:2301.13844, 2023
[22] S. Liu, J. Cao, R. Yang, and Z. Wen, “Generating a structured summary of numerous academic papers: Dataset and method,” arXiv preprint arXiv:2302.04580, 2023.
[23] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, “MSˆ2: Multi-Document Summarization of Medical Studies,” in EMNLP, 2021, pp. 7494–7513.

Q & A

THANK YOU…

Evaluation Metrics

- ROUGE [1,2,24]
  - Measures overlap between generated and reference summaries
  - Calculates score based on recall and precision of the generated summary
  - ROUGE-N measures n-gram recall while ROUGE-L uses longest common subsequence algorithm which are variants of ROUGE
  - ROUGE-W, ROUGE-S, and ROUGE-SU are extensions of ROUGE-N that incorporate weighting and skip-bigram statistics
- BLUE [25]
  - Measures n-gram overlap between generated and reference summaries
  - Gives higher score for more matching n-grams and penalize for incorrect word order
- Other metrics [1,2]
  - Precision: measures the proportion of generated summary that is relevant to the reference summary
  - Recall: measures the proportion of the reference summary that is covered by the generated summary
  - Pyramid: evaluates summaries based on how many content units they cover, where content units are defined based on their position in the source document.

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[2]M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[24] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74–81
[25] Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318

Evaluation Metrics

- ROUGE [1,2,12]
  - ROUGE-N measures n-gram recall while ROUGE-L uses longest common subsequence algorithm which are variants of ROUGE
  - ROUGE-W, ROUGE-S, and ROUGE-SU are extensions of ROUGE-N that incorporate weighting and skip-bigram statistics

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[2]M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[12] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74–81
[