Rank Your Summaries Enhancing Bengali Text Summarization via Ranking based Approach

Paper ID: 312

Authors:

G. M. Shahariar *
Tonmoy Talukder *
Rafin Alam Khan Sotez
Md. Tanvir Rouf Shawon

Presented by —
Tonmoy Talukder
Ahsanullah University of Science and Technology

** denotes equal contribution*

Contents of Presentation

Introduction

Bengali Text Summarization

- **Languages with limited resources**, such as Bengali, face challenges in developing accurate text summarization systems.

- **Pre-trained transformer models** like BERT [1] and T5 [2] have improved Bengali text summarization by capturing contextual information.

Related Work

Sentence similarity measurement for bengali abstractive text summarization. [3]

- This paper applied sentence similarity measurement method using cosine similarity and word embeddings
- Creates and uses a Bengali news corpus
- **Limitations:** Small corpus, no semantic/syntactic analysis
- **Difference:**
  - This paper selects sentences, our paper generates summaries applying ranking-based approach
  - This paper: simple and fast, domain-independent, Our paper: complex and sophisticated, domain-specific

Automatic back transliteration of romanized bengali (banglish) to bengali. [4]

- This paper introduces automatic back transliteration of romanized Bengali (Banglish) to Bengali applying ranking-based approach
- Creates and used a Banglish-Bengali parallel corpus
- **Limitations:** limited corpus size, no handling of out-of-vocabulary words
- **Difference:**
  - This paper converts Banglish to Bangla, second paper generates summaries from Bengali texts
  - This paper: solves a transliteration problem, domain-independent, Our paper: solves a summarization problem, domain-specific

The evaluation of sentence similarity measures. [5]

- This paper evaluates various methods of measuring sentence similarity based on lexical, syntactic, semantic, and pragmatic features
- Uses two datasets of sentence pairs from different domains and languages
- Evaluates the methods based on correlation with human judgments and classification accuracy

- **Limitations:** no analysis of the impact of individual features, no comparison with state-of-the-art methods, no application to specific tasks

- **Difference:**
  - This paper measures sentence similarity, our paper generates summaries based on sentence ranking
  - This paper: surveys existing methods, domain- and language-independent, Our paper: proposes a new method, domain- and language-specific

Ranking paragraphs for improving answer recall in open-domain question answering. [6]

- This paper proposes a paragraph ranking model that uses query expansion and paragraph filtering techniques
- Uses a large-scale corpus of web documents and questions
- Evaluates the model based on answer recall and F1-score

- **Limitations:** no analysis of the impact of query expansion and paragraph filtering, no evaluation of answer quality or relevance

- **Difference:**
    - This paper ranks paragraphs for question answering, our paper ranks summaries for text summarization
    - This paper: solves a question answering problem, domain and language independent, Our paper: solves a text summarization problem, domain and language specific

Research Question

- How can we select the most suitable summary for a given document in Bengali, a language with limited resources?

Objectives

- Develop a novel ranking approach for summaries generated by pre-trained transformer models. This approach will select the most suitable summary based on its ranking score, which will allow for the identification of informative and coherent summaries.
- Evaluate the effectiveness of the ranking approach using multiple metrics. The metrics will be used to measure the accuracy, fluency, and informativeness of the summaries.

Outcomes and Impacts

A rank-based approach was proposed that uses multiple pre-trained models to generate summaries and ranks them based on quality.

The approach was evaluated using various metrics, and it was shown to outperform existing methods.

The implementation of the approach is available for further research.

Dataset

Data Statistics

- We have used two datasets mentioned below:

Table 01: Dataset Statistics

Dataset	Total Summaries
XL-Sum [7]	10126
Bangla Text Summarization [8]	5000

Methodology

Methodology Flowchart

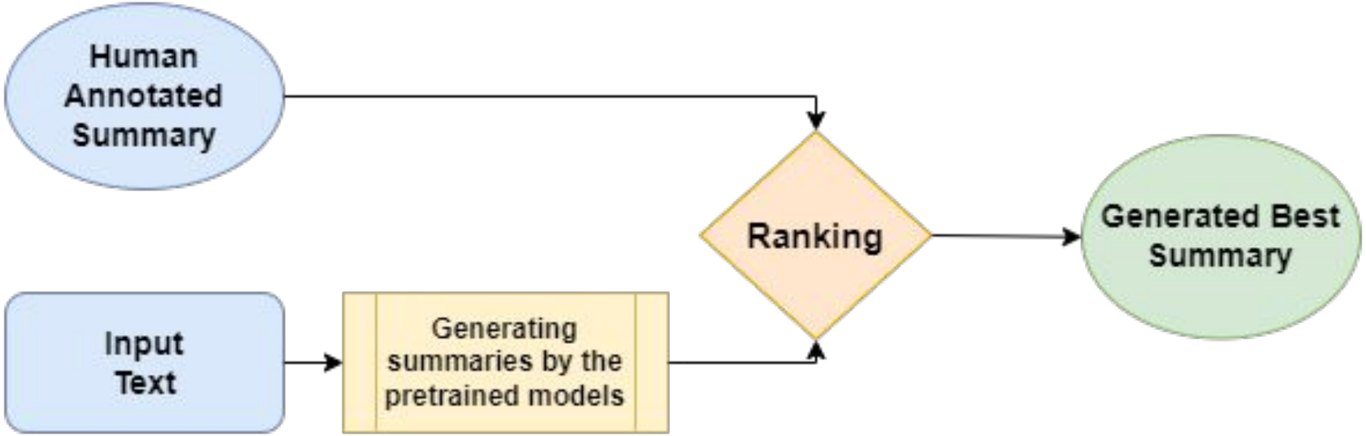

Figure 01: Flow chart of Summary Ranking.

Proposed Approach

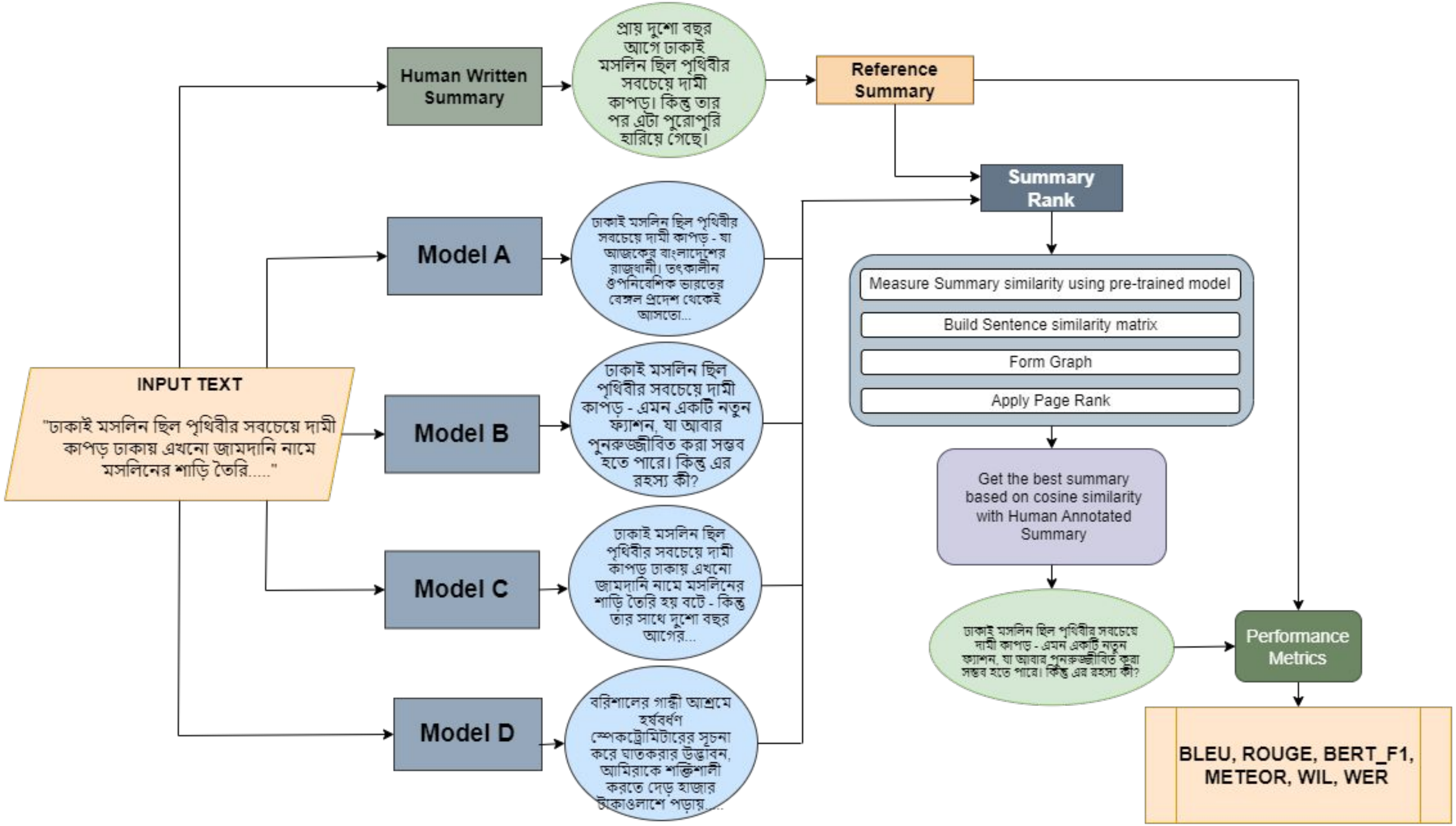

Figure 02: Proposed Methodology.

Output Examples

Human Annotated Summary	mT5 XLSum	mT5 CrossSum	scibert uncased	MT5 For Conditional Generation of Shahidul	Best Summary
বাংলাদেশ অনূর্ধ্ব-১৯ দলের এই ম্যাচে বাঁহাতি স্পিনার মাহফুজুর রহমান আফগানিস্তানের ব্যাটিং লাইন-আপ প্রায় একাই ধসিয়ে দিয়েছেন।	আবুধাবিতে আফগানিস্তানের বিপক্ষে ৬ উইকেটের জয়ে চ্যাম্পিয়ন হয়েছে বাংলাদেশ যুব দল। কিন্তু ব্যাটিং লাইন-আপে ধস না কাটিয়ে লক্ষ্য পৌঁছে দেয়নি বাংলাদেশের ব্যাটসম্যানরা।	আবুধাবির টোলেরেন্স স্টেডিয়ামে আজ আফগানিস্তানের বিপক্ষে ৬ উইকেটের জয়ে বাংলাদেশ যুব দলকে চ্যাম্পিয়ন করেছে। এই টি-টোয়েন্টি ক্রিকেট টুর্নামেন্টের ফাইনালে জয় পেয়েছে বাংলাদেশ।	রান তাড়ায় বাংলাদেশ অনূর্ধ্ব-১৯ দলের জয়ের লক্ষ্যটা থাকল দেড় শরও নিচে। এর পর টানা দুই জয় তুলে ফাইনালে জায়গা করে আহরার আমিনের দল। আরেক বাঁহাতি রাফিউজ্জামান নেন ৩৩ রানে ২ উইকেট। তার আগেই ডানহাতি এ উইকেটকিপার-ব্যাটসম্যান তিন চারে ১৩ বলে ১৭ রান করেন।	ক্ষতিকর কেমিকেলের চোরাকারবারি বন্ধে রামান স্পেকট্রোমিটারের সূচনা করে ঘাতকরার উদ্ভাবন, আমিরাকে শক্তিশালী করতে হবে : দীপু মনি-সুর আসিও : বদরুদ্দোজা জিয়া' র ্যালি, আলোচনা সভা ও সাঁতার প্রতিযোগিতার এমপি স্বপনওকে আত্মপ্রকাশও জিয়াও বয়ে বেড়াচ্ছেনকার ্যালিও' অনুষ্ঠিত অনুষ্ঠিতও অনুষ্ঠিত	আবুধাবির টোলেরেন্স স্টেডিয়ামে আজ আফগানিস্তানের বিপক্ষে ৬ উইকেটের জয়ে বাংলাদেশ যুব দলকে চ্যাম্পিয়ন করেছে। এই টি-টোয়েন্টি ক্রিকেট টুর্নামেন্টের ফাইনালে জয় পেয়েছে বাংলাদেশ।
বিয়ের পর কয়েকটা দিন পরস্পরের সঙ্গে একান্তে কাটানো খুব জরুরি। মধুচন্দ্রিমা সেই সুযোগই করে দেয়। এখানে একান্তে কাটানোর সময় বাকি জীবন দুজনের পথ চলাটাকে করে সহজ।	সারা জীবন যে মানুষটার সঙ্গে কাটাতে হবে, সবার আগেই তো তাকে জানা চাই। কিন্তু বর্ণিল বিয়ে ঠিক হওয়ার পর থেকেই এই সুযোগটা ছিল খুব জরুরি। বিয়ের আনুষ্ঠানিকতার শেষ দিন।	করোনাভাইরাস মহামারির ঠিক আগমুহূর্তে একজন দম্পতি বিয়ের পর কয়েকদিন পরস্পরের সঙ্গে একান্তে কাটাতে পারেন না বর-কনে। কিন্তু শেষ এক মাসের ধকল থেকেই তাকে জানাতে হয়।	সারা জীবন যে মানুষটার সঙ্গে কাটাতে হবে, সবার আগে তো তাকে জানা চাই। তবে শেষ এক মাসের ধকল যেন আর নেওয়া যায় না। কিন্তু টয়ার ভাষায়, 'ওটাকে ছোট্ট হানিমুন বলতে পারেন। কোনো কারণে একজন রেগে গেলে, আরেকজন চুপ থাকুন।	মুখ থুবড়ে পড়েছে 'গ্রিন ঠাকুরগাঁও, ক্লিন ঠাকুরগাঁও প্রকল্পও ইজিবাইক ফেলে দেওয়া হচ্ছে মা, পরেও পারের যাত্রামন্ত্রী এর লোকসানের গবেষণাটি হাইজ্যাকেরক আমিও সিসি আসো!, আমি সবারই সমস্যা... চলছে শেষ জিয়াও জিয়াতেও যান আসেও : দীপু মনিও	করোনাভাইরাস মহামারির ঠিক আগমুহূর্তে একজন দম্পতি বিয়ের পর কয়েকদিন পরস্পরের সঙ্গে একান্তে কাটাতে পারেন না বর-কনে। কিন্তু শেষ এক মাসের ধকল থেকেই তাকে জানাতে হয়।

Figure 03: Example of a few candidate summaries generated by all the models along with the reference and best-ranked summary on two randomly picked newspaper texts.

Experimental Results

Hyper Parameter Settings

Maximum output token length	400
Minimum output token length	64
Maximum input token length	512
no_repeat_ngram_size	2
Beam size	4

Evaluation Metrics

- BLEU Score (Bilingual Evaluation Understudy) [9]
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [10]
- BERTScore [11]
- METEOR (Metric for Evaluation of Translation with Explicit Ordering) [12]
- WER (Word Error Rate) [13]
- WIL (Word Information Lost) [13]

Performance Measurements

Summary	XLSum Dataset				Bangla Text Summarization Dataset			
	WIL	METEOR	WER	BERTScore (F1)	WIL	METEOR	WER	BERTScore (F1)
Given Summary	0.0099	0.196	0.0098	0.673	0.0098	0.278	0.0097	0.651
Best Summary	**0.0095**	**0.347**	**0.0094**	**0.723**	**0.0092**	**0.361**	**0.0090**	**0.725**
Model A	0.0098	0.320	0.0097	0.716	0.0095	0.332	0.0092	0.715
Model B	0.0098	0.296	0.0097	0.709	0.0095	0.326	0.0093	0.714
Model C	**0.0081**	**0.579**	**0.0081**	0.625	**0.0082**	**0.489**	**0.0079**	**0.765**
Model D	0.0100	0.025	0.0099	0.625	0.0099	0.032	0.0098	0.624

Table 1: Performance comparison between the input text and all the summaries on two different datasets

Table 2: Performance comparison between the reference and all other summaries (candidate and best ranked) on two different datasets

Summary	XLSum Dataset				Bangla Text Summarization Dataset			
	WIL	METEOR	WER	BERTScore (F1)	WIL	METEOR	WER	BERTScore (F1)
Best Summary	**0.0095**	**0.189**	**0.017**	**0.749**	**0.0094**	**0.192**	**0.040**	**0.708**
Model A	0.0095	0.182	0.012	**0.750**	0.0095	0.164	**0.031**	0.701
Model B	0.0097	0.143	0.012	0.735	0.0095	0.163	**0.031**	0.702
Model C	0.0099	0.108	0.051	0.679	0.0096	0.185	0.078	0.681
Model D	0.0100	0.007	0.019	0.619	0.0099	0.033	0.052	0.635

Table 3: BLEU and ROUGE scores comparison between the reference and all other summaries (candidate and best ranked) on two different datasets.

	XLSum Dataset					
Summary Model	BLEU 3	BLEU 4	ROUGE Version	Recall	Precision	F1 Score
Best Summary	0.783	0.0496	r-1	0.313	0.222	0.249
			r−2	0.132	0.096	0.107
			r−l	0.260	0.186	0.208
Best Summary	0.0300	0.0130	r−1	0.433	0.118	0.184
			r−2	0.176	0.044	0.069
			r−1	0.392	0.107	0.167
	Bangla Text Summarization					

	XLSum Dataset					
Summary Model	**BLEU 3**	**BLEU 4**	**ROUGE Version**	**Recall**	**Precision**	**F1 Score**
Model A	0.0765	0.0463	r-1	0.288	**0.227**	0.245
			r−2	0.125	**0.096**	0.105
			r−l	0.245	**0.191**	**0.208**
Model A	0.0253	0.0108	r−1	0.369	0.107	0.165
			r−2	0.144	0.038	0.060
			r−1	0.337	0.098	0.151
	Bangla Text Summarization					

	XLSum Dataset					
Summary Model	**BLEU 3**	**BLEU 4**	**ROUGE Version**	**Recall**	**Precision**	**F1 Score**
Model B	0.0502	0.029	r-1	0.235	0.187	0.201
			r−2	0.088	0.068	0.074
			r−l	0.197	0.155	0.168
Model B	0.0248	0.0102	r−1	0.367	0.108	0.166
			r−2	0.141	0.038	0.059
			r−1	0.332	0.098	0.151
	Bangla Text Summarization					

	XLSum Dataset					
Summary Model	**BLEU 3**	**BLEU 4**	**ROUGE Version**	**Recall**	**Precision**	**F1 Score**
Model C	0.0125	0.0064	r-1	0.277	0.075	0.112
			r−2	0.072	0.018	0.027
			r−l	0.202	0.055	0.082
Model C	0.0200	0.0089	r−1	**0.454**	0.080	0.132
			r−2	**0.187**	0.029	0.049
			r−1	**0.415**	0.073	0.121
	Bangla Text Summarization					

	XLSum Dataset					
Summary Model	**BLEU 3**	**BLEU 4**	**ROUGE Version**	**Recall**	**Precision**	**F1 Score**
Model D	1.13E-05	2.91E-82	r-1	0.017	0.010	0.012
			r−2	0.001	0.000	0.000
			r−l	0.016	0.010	0.012
Model D	0.0009	0.0001	r−1	0.099	0.020	0.034
			r−2	0.016	0.003	0.005
			r−1	0.093	0.019	0.032
	Bangla Text Summarization					

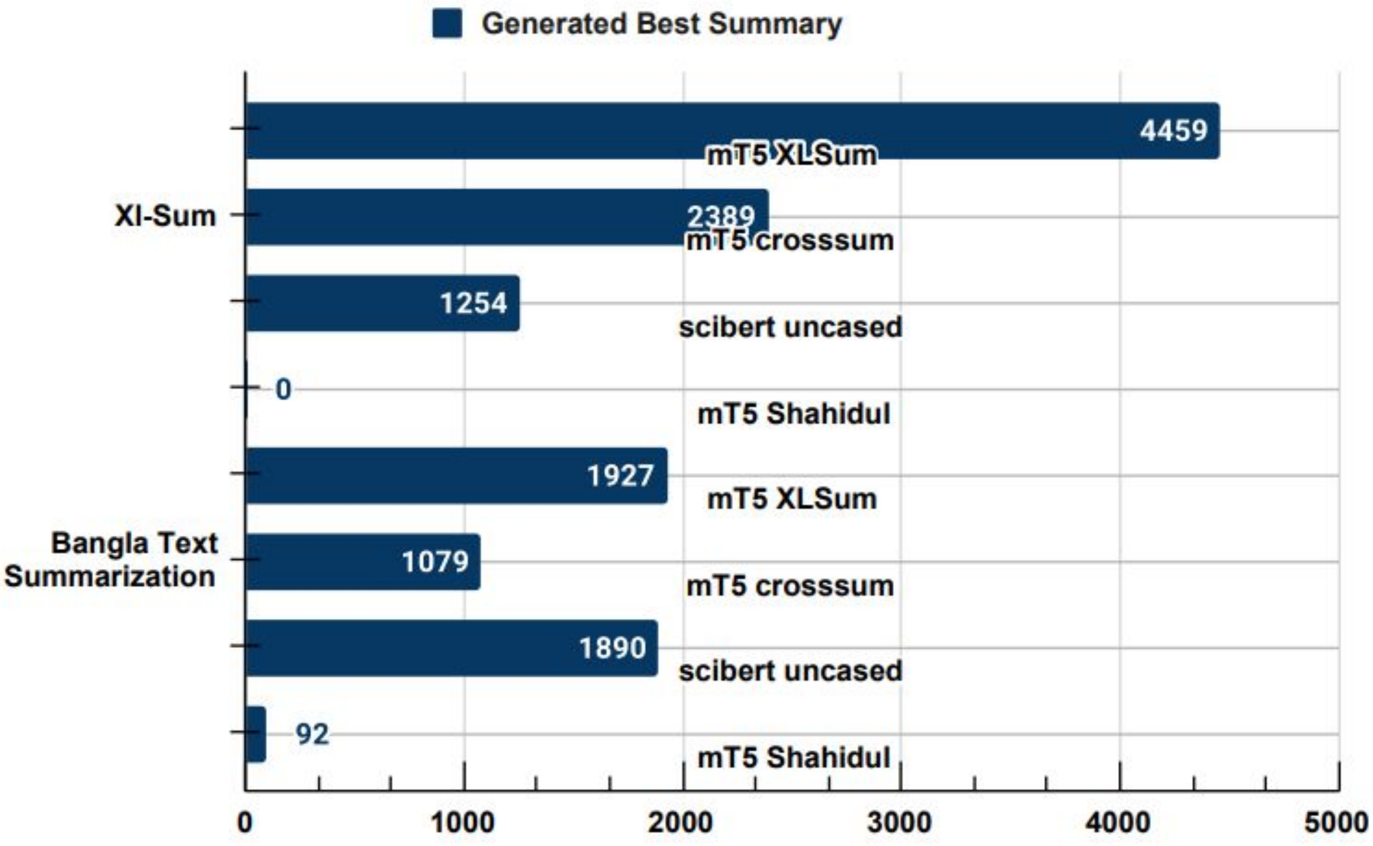

Figure 04: Statistics of the summaries per model that are selected by our approach on both datasets.

Best Summaries Statistics

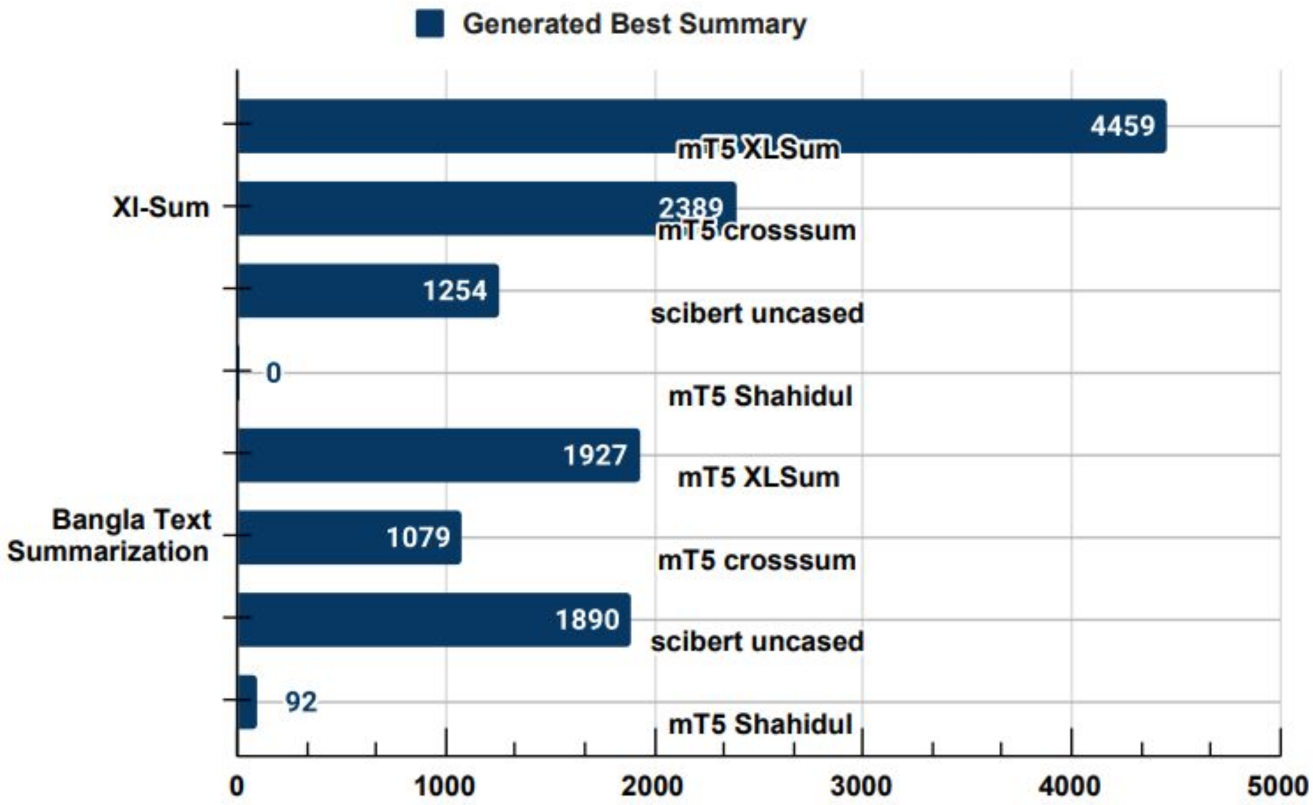

Figure 04: Statistics of the summaries per model that are selected by our approach on both datasets.

Conclusion

- Text summarization is a valuable tool for condensing large amounts of text and extracting key information.
- Low-resource languages like Bengali pose unique challenges for text summarization.
- A rank-based approach that leverages multiple models and selects the best summary can enhance the accuracy and quality of the generated summaries.

Future Research Direction

- Using different pre-trained transformer models to generate summaries.
- Developing more sophisticated ranking algorithms to select the best summary.
- Applying the rank-based approach to other low-resource languages.
- Investigating the impact of the rank-based approach on the accuracy and quality of the generated summaries.

References

[1] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). Association for Computational Linguistics (Jun 2019)

[2] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mT5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (Jun 2021)

[3] Masum, A.K.M., Abujar, S., Tusher, R.T.H., Faisal, F., Hossain, S.A.: Sentence similarity measurement for bengali abstractive text summarization. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). pp. 1–5. IEEE (2019)

[4] Shibli, G.S., Shawon, M.T.R., Nibir, A.H., Miandad, M.Z., Mandal, N.C.: Automatic back transliteration of romanized bengali (banglish) to bengali. Iran Journal of Computer Science pp. 1–12 (2022)

[5] Achananuparp, P., Hu, X., Shen, X.: The evaluation of sentence similarity measures. In: Data Warehousing and Knowledge Discovery: 10th International Conference, DaWaK 2008 Turin, Italy, September 2-5, 2008 Proceedings 10. pp. 305–316. Springer (2008)

[6] Lee, J., Yun, S., Kim, H., Ko, M., Kang, J.: Ranking paragraphs for improving answer recall in open-domain question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (Oct-Nov 2018)

References

[7] Hasan, T., Bhattacharjee, A., Islam, M.S., Mubasshir, K., Li, Y.F., Kang, Y.B., Rahman, M.S., Shahriyar, R.: XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics (Aug 2021)

[8] Bengali Text Summarization. (n.d.). Bengali Text Summarization | Kaggle. https:///datasets/hasanmoni/bengali-text-summarization

[9] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)

[10] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)

[11] Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SkeHuCVFDr

[12] Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)

[13] Morris, A.C., Maier, V., Green, P.: From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In: Eighth International Conference on Spoken Language Processing (2004)

THANKS!

Any questions?