CREATIVESUMM: Shared Task on Automatic Summarization for Creative Writing

Divyansh Agarwal[1] **Alexander R. Fabbri**[1] **Simeng Han**[3] **Wojciech Kryściński**[1]
Faisal Ladhak[2] **Bryan Li**[5] **Kathleen McKeown**[2]
Dragomir Radev[3] **Tianyi Zhang**[4] **Sam Wiseman**[6]
[1]Salesforce Research [2]Columbia University [3]Yale University
[4]Stanford University [5]University of Pennsylvania [6]Duke University

Abstract

This paper introduces the shared task of summarizing documents in several creative domains, namely literary texts, movie scripts, and television scripts. Summarizing these creative documents requires making complex literary interpretations, as well as understanding non-trivial temporal dependencies in texts containing varied styles of plot development and narrative structure. This poses unique challenges and is yet underexplored for text summarization systems. In this shared task, we introduce four sub-tasks and their corresponding datasets, focusing on summarizing books, movie scripts, primetime television scripts, and daytime soap opera scripts. We detail the process of curating these datasets for the task, as well as the metrics used for the evaluation of the submissions. As part of the CREATIVESUMM workshop at COLING 2022, the shared task attracted 18 submissions in total. We discuss the submissions and the baselines for each sub-task in this paper, along with directions for facilitating future work in the field.

1 Introduction

Contemporary research in text summarization systems has focused mainly on a select few domains such as news, scientific and legal articles. While summarizing these domains is important, the corresponding documents are limited in length and the diversity of the discourse structure.

There is a large body of creative texts available on the web that pose greater challenges for text summarization in NLP. These creative documents like books, movie scripts, and television scripts are usually written by 'subject matter experts', are substantial in length and contain complex inter-dependencies between characters in the plotline (Kryscinski et al., 2021; Ladhak et al., 2020; Mihalcea and Ceylan, 2007). Summarization of such text requires understanding many different genres, and offers the possibility of improving creating writing platforms (Aparício et al., 2016; Toubia, 2021; Gorinski and Lapata, 2015).

Such creative documents are often accompanied by a synopses, short-description or a summary, which allows readers to gauge their interest in the corresponding artifact. Summaries of creative texts are widely used by students as study guides, and could be useful to experts that need to grade the quality of these documents for a particular audience. Furthermore, researchers may be interested in using the challenges posed by these creative texts, to identify the limitations of large language models at understanding complex discourse structure.

To further summarization research of creative writing, we identified four particular domains, each with their own set of challenges. Building off of datasets recently released in the community, we develop a shared task, composed of four sub-tasks, and encouraged participants to further research in this direction.

For sub-task 1, we focus on summarization of chapters from literary texts like books, novels and stories. We curate a combined version of Booksum (Kryscinski et al., 2021) and NovelChapters (Ladhak et al., 2020) for this purpose.

For sub-task 2, we use the Scriptbase dataset (Gorinski and Lapata, 2015), which pairs movie transcripts with their corresponding Wikipedia summaries.

Sub-tasks 3 and 4 both relate to summarization of transcripts for TV shows, both derived from SummScreen (Chen et al., 2021), which contains two non-overlapping set of TV shows, from different sources. The two sources provide transcripts for two different domains - prime time TV shows and daytime 'soap operas', respectively forming the sub-task 3 and 4.

We share the instructions to access the data for each of the four sub-tasks and associated scripts

on our github repo [1].

We detail the process of evaluation of these sub-tasks, presenting the metrics as well as the results from the submissions by the shared-task participants alongside several baseline, long-document summarization models. We discuss ways to expand the sub-tasks to include more genres of creative writing for automatic summarization systems in the future.

2 Datasets

In this section we describe the sources and pre-processing steps taken to curate the data from each sub-task of CREATIVESUMM. The data samples for each of the sub-tasks are available on the shared task github repo.

2.1 Sub-Task 1: Summarizing book chapters

The dataset for sub-task 1 (*BookSumm-chapters*) pairs chapters of novels released as part of Project Gutenberg with corresponding summaries from different online study guides. Source texts for these chapters have been made available in accordance with Project Gutenberg's guidelines. [2] For each book-chapter, we provide a link to the online study guide on Web Archive[3] where the corresponding ground truth summary can be found, for training and validation.

We combine the book titles and the study guide sources used by Kryscinski et al. (2021) and Ladhak et al. (2020), and ensure that we remove redundant titles and filter out unreliable study guides. We also filter out book-chapters that were identified as plays (such as those by Shakespeare), as they differ significantly from the other literary genres in the dataset. The book titles in each of our resulting data splits are non-overlapping.

2.2 Sub-Task 2: Summarizing movie transcripts

Scriptbase (Gorinski and Lapata, 2015) compiles a corpus of 1276 movie scripts/screenplays spanning 1909–2013, by crawling relevant web-sites. They pair the scripts with corresponding user-written summaries for the task of summarizing movie transcripts. The movie scripts comprise 23 different genres, and also include rich information about the multimodal aspects of the various scenes. The authors introduce the task of summarizing movie transcripts, as selecting a chain of scenes that accurately represents a film's story. In addition to the original Scriptbase data, we provide a *new* test set containing transcripts and summaries that did not appear in the original Scriptbase release.

2.3 Sub-Task 3: Summarizing transcripts from primetime tv-shows

We utilize the SummScreen dataset (Chen et al., 2021) for the task of summarizing transcripts of episodes from primetime tv-shows. The dataset (*SummScreenFD*) contains community-contributed transcripts for 4348 episodes from 88 shows, gathered from ForeverDreaming (FD).[4] The transcripts are paired with human-written summaries. These transcripts are characterized by long storylines, often involving parallel subplots and with great emphasis on character development. For this subtask, we use the data and preprocessing associated with the SCROLLS benchmark (Shaham et al., 2022), but we provide a *new* test set containing transcripts and summaries not released in either the original SummScreen or SCROLLS datasets.

2.4 Sub-Task 4: Summarizing transcripts from daytime 'soap-operas'

The dataset for sub-task 4 is made available in the exact same format as 2.3, except the transcripts are specific to daytime 'soap opera' type shows. Introduced by Chen et al. (2021), *SummScreenTMS* contains transcripts from 22.5k episodes, from TV MegaSite, Inc. (TMS).[5] Here again we provide a *new* test set containing transcripts and summaries not released in the original SummScreen dataset.

2.5 Data Splits

As noted above, while we used the original data splits for Sub-Task 1, we provided new unseen test inputs for Sub-Tasks 2, 3 & 4. For *SummScreenFD* & *SummScreenTMS* we suggested participants use the original test set for validation, and train on the combination of the original train and validation sets. We provide the data splits and some statistics for each sub-task in Table 1.

[1] github.com/fladhak/creative-summ-data
[2] https://www.gutenberg.org/policy/robot_access.html
[3] https://web.archive.org/
[4] transcripts.foreverdreaming.org
[5] http://tvmegasite.net/

Dataset	Train	Val	Test	Coverage	Density	Comp. Ratio	1-grams	2-grams
BookSumm-chapters	6754	983	851	0.7062	1.4078	16.4287	0.4456	0.8163
Scriptbase	1149	127	216	0.7637	1.2624	63.2474	0.3123	0.667
SummScreenFD	4011	337	459	0.7203	1.1138	87.5306	0.2640	0.7066
SummScreenTMS	20710	1793	679	0.7665	1.1954	21.0675	0.2499	0.6969

Table 1: Statistics of the documents in the dataset curated for each sub-task. We report the number of documents in each split, along with the coverage, density and compression ratio b/w the documents and the summaries in the full dataset. We also present the percentage of novel uni and bi-grams present in the reference summaries for documents in each sub-task.

3 Evaluation

We received 18 submissions in total for our shared task, with two, eleven, three and two submissions for shared tasks 1, 2, 3 and 4 respectively. We describe the metrics used for evaluating the submitted outputs. Participants had 12 weeks to sign up for the shared task and train their models, before we released the test set(s) for the different subtasks. We allowed another week after releasing the test set for the submission of system outputs. We encouraged multiple submissions for each subtask as long as they reflected different models or approaches.

3.1 Metrics

We used the same metrics for the evaluation of the four sub-tasks for easier comparison:

ROUGE (Lin, 2004): We apply the standard F1 variations of ROUGE-1, ROUGE-2, and ROUGE-L using the original PERL-based implementation.

BERTScore (Zhang et al., 2020): We compute the precision, recall, and F1 variations of reference-based BERTScore metric. We provide the hash of our BERTScore runs. [6]

LitePyramid (LP) (Zhang and Bansal, 2021): This metric has reported state-of-the-art correlations for relevance estimation on news summarization. This metric is fully automatic and first extracts semantic textual units from the reference summaries and then calculates the entailment score between the model summary and these units. We report the three-class entailment probability using an entailment model fine-tuned on TAC 2008 (TAC).

SummaC (Laban et al., 2022): We chose this metric as it relies on aggregating sentence-level computations, allowing us to score long input texts. We use zero-shot variation of the model, whose NLI component is trained on the VitaminC dataset (Schuster et al., 2021).

Summary Statistics: We report the average length and percentage of novel uni and bi-grams present in the model summaries. We also calculate the extractive coverage and density from Grusky et al. (2018), which measure the extent of the overlap between the summary and input texts.

3.2 Baselines

For each sub-task, we train three Longformer Encoder-Decoder (LED) (Beltagy et al., 2020) models that have a maximum input size of 1024, 4096, and 16384, respectively.

3.3 Results

The results for each of the sub-tasks are presented in Appendix A.

The submissions on all tasks surpass the performance of the baseline LED models by a large margin in terms of ROUGE and BERTScore. However, baselines and submissions score poorly in LP and SummaC, with several system scores near zero. These two metrics have been primarily studied within the news domain and with relatively short input, and additional analysis is required to validate these metrics within creative writing and longer input/output summarization.

Within the *BookSum-chapters* dataset, we find a sizable variation in the length and level of abstraction among systems. We also note that among the tasks, LP and SummaC give the highest scores to the systems here. Notably, we do not see a clear correspondence between the level of abstraction and the SummaC factual consistency score, which has often been observed in the news domain (Ladhak et al., 2022).

For the *Scriptbase* task, we received many variations of a base model from one team with different performance, although the gap between system statistics is not very large. The ROUGE perfor-

[6] microsoft/deberta-xlarge-mnli_L40_no-idf_version=0.3.9(hug_trans=4.20.1)

mance on this dataset is highest among all tasks, although we again note the low performance of all systems according to LP and SummaC.

For the *SummScreenFD* task, the system that performed best consisted of much shorter, fairly abstractive summaries. For the *SummScreenTMS* task, the LED baselines perform very poorly, and we found the output to be largely repetitive, likely due to optimization problems during training. The resulting baselines were much longer and more extractive than the submitted systems. The difference in system performance on the two SummScreen tasks (0.29 vs. 0.39 ROUGE-1 on SummScreen FD and TMS, respectively) demonstrates important data differences even within the same domain.

4 Related Work

Although summarization of newswire text has dominated research for over a decade, there have been key efforts at encouraging summarizing research in other domains. Nenkova et al. (2011) introduced a workshop on summarizing different genres, media and languages, with the aim of defining new tasks and corporas in these domains.

Toubia (2017) and Toubia (2021) highlight the need for summarization of creative documents, by arguing that summaries may serve as a "lubricant" in the market for creative content, making it easier for consumers to decide which information to consume.

TVRecap (Chen and Gimpel, 2021) introduced a story generation task that requires generating detailed recaps from a brief summary and a set of documents describing the characters involved in the episode. The dataset contains 26k episode recaps of TV shows gathered from fan-contributed websites.

Past work has demonstrated the importance of character analysis for understanding and summarizing the content of movie scripts (Sang and Xu, 2010; Tran et al., 2017).

With increased focus on characters in creative documents like movies and fictional stories, the task of generating character descriptions using automatic summarization been proposed recently (Zhang et al., 2019). The accompanying dataset contains 1,036,965 fictional stories and 942,218 summaries.

While there has been prior work on summarizing short stories (Kazantseva, 2006), more recent methods in summarizing books utilize the hierarchical structure of documents for understanding the complex inter-dependencies in the text. Wu et al. (2021) incorporate human feedback along with recursive task decomposition, using summaries of small sections of the book to produce an overall summary at inference time. Pang et al. (2022) propose a novel top-down and bottom-up inference framework, which is effective on a variety of long document summarization benchmarks, including books.

5 Discussion

This workshop proposes the task of summarizing documents containing creative textual content. We consider sub-tasks focusing on the summarization of book chapters, movie scripts and transcripts from TV shows. Each sub-task highlights key challenges in automatic summarization of creative texts in a different genre. We also discuss how other efforts in the literature have attempted to approach this area of research. In its first version as part of COLING 2022, our sub-tasks attracted 18 submissions in total. We present the results from these submissions, along with some baselines and compare the performance of the different systems.

Summarization of creative texts opens the door for the development of computer-based tools to aid authors and marketers in creative industries (Toubia, 2021). Automatic summaries can serve as an important component of screenwriting and book-writing tools, helping grade the quality and gauge interest amongst the consumers (Gorinski and Lapata, 2015).

Advances in Creative Summarization will also assist and augment research in other related tasks and areas. Cues from textual summaries and automated content analysis in movie scripts can be helpful in creating movie-image summaries (Tsoneva et al., 2007).

There are two main directions to improve efforts in this direction in the future - incorporating more datasets/sub-tasks that relate to creative summarization, and improving metrics/strategies to effectively evaluate these systems. Other sources on the web, such as the one used by Zhang et al. (2019) can be used for harnessing datasets for summarizing creative texts. Similarly, our evaluation scheme can be expanded to include more entity-centric metrics (Chen et al., 2021), which can be crucial for identifying the presence of key charac-

ters in summaries of creative texts.

References

Marta Aparício, Paulo Figueiredo, Francisco Raposo, David Martins de Matos, Ricardo Ribeiro, and Luís Marujo. 2016. Summarization of films and documentaries based on subtitles and scripts. *Pattern Recognition Letters*, 73:7–12.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*.

Mingda Chen and Kevin Gimpel. 2021. Tvrecap: A dataset for generating stories with character descriptions. *arXiv preprint arXiv:2109.08833*.

Philip Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076.

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.

Anna Kazantseva. 2006. An approach to summarizing short stories. In *Student Research Workshop*, pages 55–62.

Wojciech Kryscinski, Nazneen Fatema Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir R Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen R. McKeown. 2020. Exploring content selection in summarization of novel chapters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5043–5054. Association for Computational Linguistics.

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

Rada Mihalcea and Hakan Ceylan. 2007. Explorations in automatic book summarization. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 380–389. ACL.

Ani Nenkova, Julia Hirschberg, and Yang Liu, editors. 2011. *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*. Association for Computational Linguistics, Portland, Oregon.

Bo Pang, Erik Nijkamp, Wojciech Kryściński, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. 2022. Long document summarization with top-down and bottom-up inference. *arXiv preprint arXiv:2203.07586*.

Jitao Sang and Changsheng Xu. 2010. Character-based movie summarization. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 855–858.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. 2022. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*.

TAC. 2008. *Proceedings of the First Text Analysis Conference, TAC 2008, Gaithersburg, Maryland, USA, November 17-19, 2008*. NIST.

Olivier Toubia. 2017. The summarization of creative content. *Columbia Business School Research Paper*, (17-86).

Olivier Toubia. 2021. A poisson factorization topic model for the study of creative documents (and their summaries). *Journal of Marketing Research*, 58(6):1142–1158.

Quang Dieu Tran, Dosam Hwang, O Lee, Jai E Jung, et al. 2017. Exploiting character networks for movie summarization. *Multimedia Tools and Applications*, 76(8):10357–10369.

Tsvetomira Tsoneva, Mauro Barbieri, and Hans Weda. 2007. Automated summarization of narrative video on a semantic level. In *International conference on semantic computing (ICSC 2007)*, pages 169–176. IEEE.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Weiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019. Generating character descriptions for automatic summarization of fiction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7476–7483.

A Shared Task Submissions

The evaluation results for all the submissions to the sub-tasks are presented below in Tables 2-5. We use the original name of the system outputs as submitted by the participants to identify each submission.

Models	R_{1f}	R_{2f}	R_{Lf}	BS_P	BS_R	BS_{F1}	LP_{p3c}	$SummaC_{ZS}$	Length	Density	Coverage	1-grams	2-grams
Baselines													
LED_{1024}	0.1413	0.0214	0.1329	0.4318	0.4697	0.4446	0.2684	0.2599	937	1.5149	0.7230	0.2886	0.7094
LED_{4096}	0.1537	0.0221	0.1426	0.4232	0.4727	0.4420	0.2642	0.4718	866	2.9900	0.7617	0.2409	0.6096
LED_{16384}	0.1487	0.0218	0.1385	0.4282	0.4928	0.4544	0.2221	0.4132	755	2.2112	0.7372	0.2532	0.6541
Sub-Task 1 submissions													
MHPK	0.2975	0.0789	0.2833	0.5303	0.5553	0.5410	0.1071	0.1562	1465	2.1802	0.7735	0.3055	0.6775
LRL_NC	0.2643	0.0471	0.2436	0.4717	0.5264	0.4955	0.0669	0.3499	3345	12.6336	0.8336	0.1600	0.3457

Table 2: Sub-task 1 results for the baselines and the submissions on the *BookSumm-chapters* dataset.

Models	R_{1f}	R_{2f}	R_{Lf}	BS_P	BS_R	BS_{F1}	LP_{p3c}	$SummaC_{ZS}$	Length	Density	Coverage	1-grams	2-grams
Baselines													
LED_{1024}	0.1492	0.0146	0.1373	0.4298	0.4238	0.4258	0.2517	0.0000	903	1.1809	0.7021	0.3357	0.7485
LED_{4096}	0.1416	0.0130	0.1299	0.4245	0.4137	0.4179	0.2674	0.0000	877	1.3432	0.7311	0.3092	0.7273
LED_{16384}	0.1368	0.0125	0.1277	0.4322	0.3924	0.4099	0.2312	0.0000	551	1.4103	0.7266	0.3024	0.7211
Sub-Task 2 submissions													
MovING_scriptbase	0.4144	0.0823	0.3963	0.5163	0.5233	0.5194	0.0356	0.0476	729	3.4398	0.8759	0.1621	0.4827
UdS_scriptbase	0.4285	0.1088	0.4125	0.5543	0.5410	0.5474	0.0587	0.0323	883	1.3489	0.7778	0.2911	0.7346
UdS_base_bs4_0.2noisy	0.4634	0.1158	0.4405	0.5723	0.5680	0.5700	0.0372	0.0250	791	1.2722	0.7502	0.3520	0.7636
UdS_base_bs4_Bitfit	0.4344	0.1091	0.4178	0.5533	0.5402	0.5465	0.0566	0.0321	815	1.3475	0.7753	0.2879	0.7301
UdS_base_bs4_Bitfit_Mix01	0.4580	0.1162	0.4376	0.5664	0.5601	0.5631	0.0459	0.0316	770	1.3226	0.7590	0.3235	0.7457
UdS_base_bs4_Bitfit_Mix10	0.4576	0.1158	0.4380	0.5690	0.5620	0.5653	0.0418	0.0301	781	1.3194	0.7628	0.3167	0.7459
UdS_base_bs4	0.4639	0.1152	0.4408	0.5703	0.5672	0.5686	0.0408	0.0269	780	1.2733	0.7513	0.3456	0.7604
UdS_base_bs5_Bitfit	0.4316	0.1088	0.4151	0.5542	0.5414	0.5475	0.0567	0.0309	838	1.3451	0.7774	0.2890	0.7327
UdS_base_bs5_Bitfit_Mix01	0.4550	0.1149	0.4344	0.5664	0.5606	0.5634	0.0508	0.0322	776	1.3344	0.7621	0.3205	0.7438
UdS_base_bs5_Bitfit_Mix10	0.4553	0.1146	0.4353	0.5677	0.5619	0.5646	0.0435	0.0295	777	1.3235	0.7596	0.3218	0.7462
UdS_base_bs5	0.4604	0.1140	0.4369	0.5699	0.5668	0.5682	0.0410	0.0281	783	1.2709	0.7493	0.3485	0.7649

Table 3: Sub-task 2 results for the baselines and the submissions on the *Scriptbase* dataset.

Models	R_{1f}	R_{2f}	R_{Lf}	BS_P	BS_R	BS_{F1}	LP_{p3c}	$SummaC_{ZS}$	Length	Density	Coverage	1-grams	2-grams
Baselines													
LED_{1024}	0.1428	0.0154	0.1236	0.4100	0.4107	0.4052	0.0987	0.0559	330	1.1440	0.7148	0.3060	0.7801
LED_{4096}	0.1694	0.0209	0.1501	0.4591	0.4752	0.4600	0.0304	0.1052	188	1.4378	0.7343	0.2803	0.7314
LED_{16384}	0.1514	0.0170	0.1334	0.4485	0.4632	0.4489	0.0304	0.1644	192	1.5474	0.7108	0.2904	0.7285
Sub-Task 3 submissions													
inotum	0.2860	0.0624	0.2529	0.5934	0.5609	0.5750	0.0559	0.0272	86	1.0321	0.6664	0.3715	0.8251
team_ufal	0.2469	0.0408	0.2300	0.5038	0.5590	0.5285	0.0406	0.1282	289	2.0821	0.7127	0.2484	0.6498
AMRTVSumm	0.2307	0.0303	0.2106	0.4906	0.5344	0.5108	0.0138	0.024	256	0.8789	0.6137	0.4924	0.8569

Table 4: Sub-task 3 results for the baselines and the submissions on the *SummScreenFD* dataset.

Models	R_{1f}	R_{2f}	R_{Lf}	BS_P	BS_R	BS_{F1}	LP_{p3c}	$SummaC_{ZS}$	Length	Density	Coverage	1-grams	2-grams
Baselines													
LED_{1024}	0.0727	0.0063	0.0652	0.4100	0.4107	0.4052	0.0304	0.0905	984	0.9530	0.6302	0.3781	0.8379
LED_{4096}	0.0822	0.0062	0.0753	0.4122	0.4086	0.4057	0.0304	0.0837	879	0.8457	0.5822	0.4091	0.8674
LED_{16384}	0.0722	0.0047	0.0656	0.4096	0.3916	0.3960	0.0304	0.0800	885	0.7544	0.5430	0.4387	0.8931
Sub-Task 4 submissions													
AdityaUpadhyay	0.3921	0.0909	0.3794	0.5507	0.5550	0.5516	0.0625	0.1133	316	1.9436	0.7618	0.2026	0.6688
AMRTVSumm	0.3426	0.0717	0.328	0.5385	0.5318	0.5347	0.0152	0.0499	259	1.1024	0.7291	0.3514	0.7931

Table 5: Sub-task 4 results for the baselines and the submissions on the *SummScreenTMS* dataset.