Proceedings Paper:
Mao, Q, Li, J, Wang, J et al. (4 more authors) (2022) Explicitly Modeling Importance and Coherence for Timeline Summarization. In: Proceedings of ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 23-27 May 2022, Singapore. IEEE , pp. 8062-8066. ISBN 978-1-6654-0540-9

https://doi.org/10.1109/ICASSP43922.2022.9746383

EXPLICITLY MODELING IMPORTANCE AND COHERENCE FOR TIMELINE SUMMARIZATION

Qianren Mao[1], *Jianxin Li*[1*], *JiaZheng Wang*[1], *Peng Hao*[1], *Lihong Wang*[2], *Zheng Wang*[3]

[1]Beihang University, [2]CNCERT/CC, [3]University of Leeds
maoqr, lijx@act.buaa.edu.cn

ABSTRACT

Timeline summarization (TLS) identifies major events and generates short summaries on how the event evolves in a period of time. Existing timeline summarization methods generate summaries by considering the *coverage* and *diversity* of the content and temporized information but ignore the *importance* and *coherence* of sentences used in summary. However, ignoring such information often causes missing important facts in the generated TLS and confuses users. We propose a better approach for TLS by explicitly optimizing importance and coherence on top of coverage and diversity. We apply our approach to both direct and pipeline TLS frameworks. Experimental results show that our approach achieves better performance when compared with two state-of-the-art TLS methods.

Index Terms— Timeline, text summarization, text mining

1. INTRODUCTION

Timeline summarization (TLS) generates an overview of long-running events through dated daily summaries for the most important dates. It is a key enabling technology for NLP tasks like event tracking and information retrieval [1, 2, 3, 4, 5, 6, 7, 8]. TLS often requires understanding the evolving text content across multiple documents over months. Doing is challenging because TLS often requires processing a large number of various documents that are hard to comprehend.

Existing TLS methods are primarily concerned with the *coverage* [1, 7, 9] and *diversity* [1, 7] of content understanding. The former measures the 'fidelity' of a summary to the document by evaluating sentences chosen from the document, while the latter measures the newness of the sentences selected for a daily summary. Both properties are essential for generating informative and faithful daily summaries. However, existing approaches overlook (have not explicitly model) two other properties that are equally important for generating high-quality summaries: the *importance* and *coherence* of sentences. In the context of TLS, *importance* measures how much relevant information is presented within a daily summary, and *coherence* measures the consistency of sentences with the daily summaries over the dates. Both properties are crucial for generating informative and well-covered summaries and should be explicitly optimized in the generated summary.

To illustrate the usefulness of importance and coherence of summarized sentences, consider the representative example given in Table 1. This example is an excerpt of a timeline from the TL17 dataset [3, 4], covering the *BP oil spill* topic published by the Washington Post newspaper. For this example, the human-crafted summaries in Table 1 include important information that captures critical aspects of the major event and its evolution. Examples of such key information include '*explosion*', '*BP*', '*oil*' and '*injure*' on '04-20-2010', '*search*' and '*rescue*' on '04-22-2010', and '*risk*' and '*economic impacts*' on '04-23-2010'. Similarly, we want the machine-generated summary to include such information because it provides primary information of the major event and its development. Likewise, sentences can express critical information coherently if sentences are ordered, such as two adjacent sentences discuss similar content or topic. For examples, in the expert-generated daily summary of '04-20-2010', `S1` describes an explosion, and `S2` describes the casualties of this explosion (i.e., the consequence). Those coherent expressions can simplify the context for the major event 'BP oil spill'. In short, by ignoring these two properties, a TLS system could miss critical facts or hinder document understanding. Our work is designed to avoid such pitfalls.

Daily Summary (04-20-2010)
`S1`: Explosion and fire on the BP-licensed Transocean drilling rig Deepwater Horizon in the Gulf of Mexico.
`S2`: 11 people are reported missing and approximately 17 injured.
`S3`: The force of the sinking breaks off the rig 's drillpipe, allowing oil to spew out into the gulf.
Daily Summary (04-22-2010)
`S1`: Search-and-rescue operations by the US National Response Team begin.
`S2`:Multiple Coast Guard helicopters , planes and cutters responded to rescue the Deepwater Horizon 's 126 person crew.
`S3`:The US coast guard suspends the search for missing workers, who are all presumed dead.
Daily Summary (04-23-2010)
`S1`: A homeland security department risk analysis says the incident "poses a negligible risk to regional oil supply markets and will not cause significant national economic impacts".
Other daily summaries...

Table 1. The hand-crafted timeline on *BP oil spill* created by journalists from the Washington Post.

This paper presents a better approach for TLS by explicitly considering the content importance and coherence. We achieve this by introducing two new objective functions to reward the importance of selecting sentences of high word frequencies and coherence between two adjacent sentences in a daily summary. We show that our objective functions satisfy two crucial properties: monotone and submodular, which enables constructing an approximation algorithm to explore the vast optimization space of TLS effectively.

We evaluate our approach[1] using both automatic metric and hu-

* Jianxin Li is the corresponding author.

[1]Code and data available at: https://github.com/OpenSUM/TimeSum.

man evaluation. Experimental results show that by simultaneously considering content coverage, diversity, importance, and coherence, our approach delivers higher-quality daily summaries over the state-of-the-art method [10]. Our work is the first attempt in considering all four optimization properties for TLS simultaneously. It is also the first approach to consider the importance and coherence of content for TLS, leading to higher-quality TLS.

2. BACKGROUND AND PRELIMINARIES

Problem Scope and Notation. Like most TLS tasks, we start from a query phrase, $\mathcal{Q}$, (e.g., the major event of *BP oil spill* in Table 1) for a collection of documents, $\mathcal{D}$, (e.g., news articles). The $\mathcal{D}$ contains a set of dated sentences, $\mathcal{U}$, where the date can be either explicitly expressed in the sentence or derived from the article's publication date. From the dated sentences, our task is to generate a timeline $S = \{(d_1, s_1)...(d_i, s_i)...(d_l, s_l)\}$, where d_i is i-th date and s_i is a daily summary for date d_i. The number of dates (l) to be covered and the length (k) of a daily summary are configurable parameters given to the TLS system, and $m = k * l$ is the number of total sentences.

Formulation and Constraints. Assume we are given a function $\mathcal{F}$ to measure the quality of a summary S. Let $\mathcal{I}$ be a set of constraints such as the maximum number of sentences (cardinality constraint $\mathcal{I}_c : |S| \leq m$), or words of a summary (knapsack constraints $\mathcal{I}_k : \sum_{s \in S} |words(s)| \leq n$), or temporal constraint $\mathcal{I}_{tl}$[2]. The task of TLS can be formalized as a combinatorial optimization problem:

$$S^* \in \text{argmax}_{S \subseteq \mathcal{U}} \mathcal{F}(S), \textit{subject to} : \mathcal{I}. \tag{1}$$

A common approach for solving the NP-hard problem in Equation 1 is to adopt a greedy-based algorithm to obtain a near-optimal solution [11, 12, 1]. To do so, the objective function $\mathcal{F}$ for TLS needs to be monotone and submodular. In layman's terms, a monotone-submodular function has the property that adding more elements to a set cannot decrease the value of the set.

Objectives. Prior work for TLS has considered two monotone and submodular content objective functions for the optimization [1]. The first aims to quantify the content coverage of the summary set S to the document collection D: $\mathcal{F}_{cov}(S) = \sum_{s \in S} \sum_{v \in \mathcal{U}} \textit{Sim}(s, v)$, where *Sim* is a cosine similarity function. s is a candidate summary and $s \in S$. v is a dated sentence and $v \in \mathcal{U}$, in which $\mathcal{U}$ is the set of all dated sentences. The second aims to measure content diversity by selecting dated sentences from diverse clusters:$\mathcal{F}_{div}(S) = \sum_{i=1}^{k} \sqrt{\sum_{s \in P_i \cap S} r(s)}$, where P_i is a partition (i.e., obtained by semantic clustering). $r(\cdot)$ is a *singleton reward* function (i.e., the reward of adding i into the empty set).

3. OUR OPTIMIZATIONS

Our work aims to simultaneously consider multiple properties – coverage, diversity, importance, and coherence – for TLS content optimizations with $\mathcal{F}$ [3] $\equiv \sum_{i=1}^{m} \lambda_i \mathcal{F}_i$. This is achieved by employing a greedy-based approach to generate daily summaries. To meet the monotone and submodular constraints of a greedy-based solution, we carefully design our new objective functions for modeling the importance and coherence.

Importance Optimization Function. One of the basic requirements of a good summary is that it should contain the most crucial information across sentences. To this end, we introduce a monotone and submodular objective function to model this property of importance. This objective function is defined as:

$$\mathcal{F}_{imp}(S) = \sum\nolimits_i \textit{Frq}_i b_i, \tag{2}$$

where i is the i-th keyword in a summary sentence when stopwords are deleted. We assign an importance weight *Frq* to i-th keyword using its document frequency computed by the TF-IDF metric. b_i is a binary variable that indicates the presence of keywords i in the extracted sentences. Intuitively, sentences containing the most relevant keywords are essential for a summary. This objective is to maximize the weight of the keywords for selecting salient sentences.

Coherence Optimization Function. Ordering extracted sentences into a coherent summary is usually achieved by computing the lexical similarity (smoother connectivity) from one sentence to the next between two consecutive sentences [14, 15]. The greater similarity between adjacent sentences reflects that they have similar topics. Since there is no function satisfying the submodular property for coherence optimization, we revise the lexical similarity function into a submodular function $\mathcal{F}_{coh}(S)$, to optimize the closeness between two adjacent sentences:

$$\mathcal{F}_{coh}(S) = \sum\nolimits_{i=1}^{m-1} (\textit{Sim}(s_i, s_{i+1}) + 1)/l(i), \tag{3}$$

where the coherence is mainly formulated as the cosine similarity *Sim* between two adjacent sentences s_i and s_{i+1} of the ordered form of summary S. We set $l(i) = 2^i$ as a location function for the i-th summary in a timeline. Here, $l(i)$ is used to adjust a global coherence, where the front sentences in a sequence have a greater impact on the $\mathcal{F}$ than the latter ones since the prior sub-events are decisive to the event development. To plus the similarity score with 1 is to make $\mathcal{F}$ satisfy the monotonicity and submodularity.

Proofs of Monotonicity and Submodularity. Monotonicity and submodularity of $\mathcal{F}_{cov}(S)$ and $\mathcal{F}_{div}(S)$ are shown by Lin et al., [1]. We give the proof of $\mathcal{F}_{imp}(S)$ and $\mathcal{F}_{coh}(S)$.

Lemma 1. $\mathcal{F}_{imp}(S)$ is monotone and submodular.

Proof of Monotonicity. Let $A \subseteq B \subset \mathcal{U}$. It holds that,

$$\mathcal{F}(A) = \sum\nolimits_{i \in A} \textit{Frq}_i b_i. \tag{4}$$

$$\begin{aligned} \mathcal{F}(B) &= \sum\nolimits_{i \in A} \textit{Frq}_i b_i + \sum\nolimits_{i \in (B-A)} \textit{Frq}_i b_i \\ &= \mathcal{F}(A) + \sum\nolimits_{i \in (B-A)} \textit{Frq}_i b_i, \end{aligned} \tag{5}$$

Since all elements are non-negative, it follows that $\mathcal{F}(A) \leq \mathcal{F}(B)$. Therefore $\mathcal{F}_{imp}$ is monotone.

Proof of Submodularity. Let $A \subseteq B \subset \mathcal{U}$ and $v \in \mathcal{U} \backslash B$. Then,

$$\mathcal{F}(A \cup v) - \mathcal{F}(A) = \mathcal{F}(v). \tag{6}$$

$$\mathcal{F}(B \cup v) - \mathcal{F}(B) = \mathcal{F}(v). \tag{7}$$

Then, $\mathcal{F}(A \cup v) - \mathcal{F}(A) \geq \mathcal{F}(B \cup v) - \mathcal{F}(B)$. Therefore, $\mathcal{F}_{imp}$ is submodular.

Lemma 2. $\mathcal{F}_{coh}(S)$ is monotone and submodular. For brevity, we write f_i instead of $(\textit{Sim}(s_i, s_{i+1}) + 1)/2^i$.

Proof of Monotonicity. Let $A \subseteq B \subset \mathcal{U}$. It holds that,

$$\mathcal{F}(B) = \mathcal{F}(A) + \sum\nolimits_{i=m_A}^{m_B - 1} f_i. \tag{8}$$

[2]It contains $|\left\{s' | s' \in S, d(s') = d(s)\right\}| \leq k$ and $|\{d(s) | s \in S\}| \leq l$, where $d(\cdot)$ is the function that assigns each sentence to its date [7].

[3]If a collection of functions $\{\mathcal{F}_i\}_i$ is submodular, so is their weighted sum $\mathcal{F} \equiv \sum_{i=1}^{m} \lambda_i \mathcal{F}_i$ [1], where $\{\lambda_i\}_i$ is a nonnegative weight.

TL17		Concat F1		Agree F1		Align F1		Date-F1
		R-1	R-2	R-1	R-2	R-1	R-2	
Oracle(c)		*.513*	*.181*	*.320*	*.124*	*.320*	*.129*	*.926*
Oracle(k)		*.514*	*.180*	*.317*	*.124*	*.320*	*.129*	*.926*
Oracle(tl)		*.511*	*.176*	*.312*	*.122*	*.313*	*.124*	*.926*
Non-SUB	CLUST [13]	.352	.074	.059	.015	.082	.020	.407
	PUBCOUNT [13]	.377*	.093*	.102*	.030*	.105	.027	.481
	DATEWISE [13]	.378*	.093*	.103*	.029*	.120*	.035*	.544*
	WILSON [10]	.408*	.101*	.107*	.032*	.121*	.035*	**.563***
SUB	TILSE [7]	.369	.092*	.091*	.024	.105	.027	.513*
	DATEWISESUB	.377*	.092*	.102*	.030*	.117*	.033*	.544*
	TLSUBX†	.379*	.095*	.105*	.033*	.123*	.036*	.544*
	TLSUBX‡	**.411***	**.103***	**.109***	**.035***	**.123***	**.037***	**.563***

CRISIS		Concat F1		Agree F1		Align F1		Date-F1
		R-1	R-2	R-1	R-2	R-1	R-2	
Oracle(c)		*.513*	*.175*	*.367*	*.147*	*.360*	*.143*	*.974*
Oracle(k)		*.507*	*.172*	*.362*	*.142*	*.366*	*.142*	*.974*
Oracle(tl)		*.500*	*.171*	*.351*	*.146*	*.366*	*.147*	*.974*
Non-SUB	CLUST [13]	.340	.069	.044	.009	.061	.013	.226
	PUBCOUNT [13]	.340	.073	.069	.023	.067	.012	.233
	DATEWISE [13]	.347	.075	.072*	.023	.089*	.026*	.295*
	WILSON [10]	.361	.076	.068*	.020	.085*	.023*	**.302***
SUB	TILSE	.331	.069	.056	.013	.076*	.017*	.274
	DATEWISESUB	.344	.073	.072*	.024*	.089*	.026*	.295*
	TLSUBX†	.347	.077	.075*	.024*	.091*	.027*	.295*
	TLSUBX‡	**.362**	**.078**	**.077***	**.026***	**.092***	**.028***	**.302***

Table 2. Results of the Oracles, comparison pipeline systems and our frameworks. Italics represent Oracle results. Highest values per column/dataset are boldfaced. $*$ denotes sign.difference to CLUST.

Since $f_i \geq 0$ and all elements are non-negative, it follows that $\mathcal{F}(A) \leq \mathcal{F}(B)$. Therefore $\mathcal{F}_{coh}$ is monotone.

Proof of Submodularity. Let $A \subseteq B \subset \mathcal{U}$ and $v \in \mathcal{U} \backslash B$. Then,

$$\mathcal{F}(A \cup v) - \mathcal{F}(A) = \frac{Sim(s_{m_A}, v)+1}{2^{m_A}} \geq \frac{1}{2^{m_A}}. \quad (9)$$

$$\mathcal{F}(B \cup v) - \mathcal{F}(B) = \frac{Sim(s_{m_B}, v)+1}{2^{m_B}} \leq \frac{2}{2^{m_B}}. \quad (10)$$

Since $A \subseteq B$, for $A = B$, apparently, $\mathcal{F}(A \cup v) - \mathcal{F}(A) = \mathcal{F}(B \cup v) - \mathcal{F}(B)$. For $A \neq B$, $m_A \leq m_B - 1$, and we obtain $\frac{1}{2^{m_A}} \geq \frac{2}{2^{m_B}}$. We arrive at:, $\mathcal{F}(A \cup v) - \mathcal{F}(A) \geq \mathcal{F}(B \cup v) - \mathcal{F}(B)$. Therefore, $\mathcal{F}_{coh}$ is submodular.

4. OUR TIMELINE FRAMEWORKS

We integrate the aforementioned multi-submodular optimization in a pipeline framework. This approach decouples summarization in two stages by first selecting the dates to cover using salient date selection algorithm [4] and then generating a summary for each data cluster. We follow the best performing WILSON [10] to select the l most important dates given by the PageRank algorithm [17] on their constructed date graph [5]. Based on this date selection algorithm, our derivational method is denoted as **TLSUBX**‡. We also follow the works of DATEWISE [13] to select the important dates by SUPERVISED [3]. **TLSUBX**† uses this date selection algorithm. Since the salient date selection algorithm firstly selects dates, we can explore whether our content optimizations can bring performance improvement or not, compared to other content optimizations. Two kinds of TLSUBX are adopted with our multiple submodularity optimization with the temporal constraint $\mathcal{I}_{tl}$.

5. EXPERIMENTS

5.1. Experimental Setup

Datasets. We evaluate our approach on two datasets preprocessed by Martschat et al.,[7]: TL17 [4] and CRISIS [6]. The datasets are divided into multiple topics (e.g., *BP oil spill*, *Syrian civil war*) where each topic consists of a set of news articles spanning over 6 months. Each topic has at least one ground-truth timeline, which is manually created by several professional journalists.

Evaluation Metrics. For TLS, it is essential to evaluate the temporal aspect of the task. To this end, we report four variants of Rouge F1, following the methodology in [7], to allow alignments of dates. Specifically, we consider three metrics for summaries: concatenation-based Rouge F1 (denoted Concat R1 or R2), date-agreement Rouge F1 (denoted Agree R1 or R2), and alignment-based Rouge F1 (denoted align R1 or R2). We evaluate data selection on the Date F1 score.

Implementation Details. We follow the experimental settings [7, 13, 8, 10] by asking the TLS framework to summarize a single topic across multiple timelines (where the ground-truth timeline varies). The timelines are produced with l dates and k sentences per date, in which l and k are same as the ground-truth. The weight λ of importance and coherence are tuned by greedy research in $\{0.01, 0.1\}$.

Oracle Summaries. Previous studies have approximated the upbound performance (Oracle) of TLS [7, 18, 13]. Oracle results are obtained by greedily selecting all input documents' sentences. The work presented in Steen et al., [18] provides two Oracle results for direct TLS systems: Oracle (c) and Oracle (k), by applying the cardinality constraint and the knapsack constraint to the data. In this work, we also apply the temporal constraint to compute the Oracle, denoted as Oracle(tl). We first select the dates of ground truth and use our submodular approach for daily summarization.

5.2. Baseline Frameworks

Non-submodularity approaches. CLUST [13] is a textual cluster-based method which count date mentions to rank event clusters, and then generate daily summaries in each cluster. To cluster articles, they use Markov Clustering (MCL). PUBCOUNT [13] uses the publication count to rank dates, and use CENTROID-OPT [19] for summarization. DATEWISE [13] uses supervised date selection and CENTROID-OPT for summarization. WILSON [10] is the state-of-the-art method for timeline summarization. It is a three-stage pipeline framework: graph-based date selection, summarization with Textrank [20], and postprocessing to remove redundancy across the dates.

Submodularity approaches. TILSE [7] is submodularity-based multi-document summarization framework with temporal constraints. DATEWISESUB uses supervised date selection and uses ASMDS for summarization with temporal constraint $\mathcal{I}_{tl}$. DATEWISESUB is re-implemented by released code from Ghalandari et al., [13].

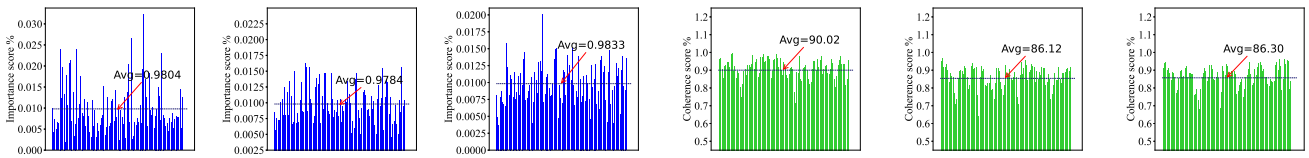
DATEWISE, DATEWISESUB and our TLSUBX† use the same data selection method. WILSON and our TLSUBX‡ use another same data selection method.

6. ANALYSIS AND DISCUSSION

Main Results. Tables 2 shows the results of TLS frameworks. All results are tested for significant differences using an approximate randomization test [21, 22] with a p-value smaller than 0.05 under the paired t-test. Compared with DATEWISESUB [13] obtained the

[4] In this work, the date selection is not the focus of the research. We verify whether our content-oriented optimization can improve the performance under the same date selection algorithm for the TLS.

[5] https://github.com/wilson-nts/WILSON.



(a) Ground-truth (b) WILSON Timeline (c) Our Model (d) Ground-truth (e) WILSON Timeline (f) Our Model

Fig. 1. Distribution of Importance and Coherence scores of all daily summaries come from the Golden timeline and the generation of WILSON and our TLSUBX for *BP oil spill* in TL17 dataset. The ticks of x-axis represent each date of a timeline. For an obvious comparison, we mark the average results.

Method	Automatic Evaluation		Human Evaluation							
	Imp.	Coh.	Imp.	Kappa(Imp.)	Coh.	Kappa(Coh.)	Cov.	Kappa(Cov.)	Div.	Kappa(Div.)
Oracle(tl)	*.9804*	*86.33*	*4.66*	*0.73*	*4.41*	*0.68*	*3.52*	*0.68*	*4.27*	*0.72*
DATEWISESUB	.9632	84.59	3.71	0.71	3.55	0.66	3.21	0.70	4.19	0.67
WILSON	.9784	86.12	3.88	0.70	4.01	0.70	3.34	0.66	4.20	0.68
TLSUBX‡	**.9833**	**86.30**	**4.73**	0.72	**4.11**	0.67	**3.41**	0.69	**4.21**	0.71

Table 3. Automatic and Human evaluation results of the average Coherence score (Coh.Avg) and Importance score (Imp.Avg) on the topic of *BP oil spill*. The properties of coverage and derversity are also evaluated by human annotators. We evaluate the agreement among human annotators by Fleiss' kappa-ratio [16].

same Date-F1 score, our TLSUBX† achieves performances under the same date selection algorithm. It shows that optimizing our four combinatorial objectives can produce better summaries than the TLS method that considers coverage and diversity only. Equipped with the best performing data selection algorithm, our content-optimation method TLSUBX‡ is superior among all baselines.

Automatic evaluation for importance. To further analyze the linguistic quality, we automatically test the *importance* property of a daily summary by calculating the *Imp* score. We give the results are shown in Table 3, and we also illustrate the distribution of the score on all dates about the topic *BP oil spill* as shown in Figure 1. Most of the daily summaries of TLSUBX‡ obtain general high importance scores and are even better than the ground truth. The results indicate our method can find some crucial sentences that are not selected by the handcrafted timelines. Besides, We find the importance score of all timelines (Ground-truth, WILSON and our generation) have periodic peaks over all dates, which indicates that a guarantee of importance for a daily summary is essential for some expected dates.

Automatic evaluation for coherence. We now analyze the linguistic *coherence* property by calculating an average of all adjacent two daily summaries' *Coh* score. As shown in Table 3, our method obtains 86.30 coherence distribution, which is close to that of the ground truth and the Oracle. As shown in Fig. 1, we find that all adjacent daily summaries are coherent with a high similarity score.

Human evaluation. We conduct a human evaluation [6] to analyze the properties of the generated summaries. We see that our method achieves the highest importance score among the baselines and the Oracle. The evaluation results of the other three properties are almost close to the ideal Oracle results and superior to the baselines. Cohen's kappa coefficients are average above 0.65, indicating a high correlation and agreement among the three human annotators.

Case study. We illustrate the first 6 generated daily summaries about events of *BP oil spill*, shown as Table 4. Our model produces summaries in which each sentence describes sequential sub-events. For an example of the first daily summary on 04-20-2010, `S1` describes an explosion, and `S2` describes the casualties of this explosion. The following summary describes events about the investigation ('04-23-2010'), measures ('04-25-2010', '04-26-2010', '04-28-2010'), congressional hearing ('11-05-2010'), and so on. Moreover, it shows that our PTLSUBX method, which arranges sentences through selecting sentences having the most high-frequency words, can produce better coherence or smoother connectivity among each daily summary.

· **04-20-2010**	
`S1`: Explosion and fire on the BP-licensed Transocean drilling rig Deepwater Horizon in the Gulf of Mexico. `S2`: Eleven people are reported missing and approximately 17 injured.	explosion☑ bp☑, oil mexico☑, injured☑
· **04-23-2010**	
`S1`: The Coast Guard says it had no indication that oil was leaking from the well 5,000 ft below the surface of the Gulf.	coast☑, leaking☑ oil☑, mexico, gulf☑
· **04-25-2010**	
`S1`: US coast guard remote underwater cameras report the well is leaking 1,000 barrels of crude oil per day. `S2`: It approves a plan for remote underwater vehicles to activate a blowout preventer and stop the leak.	oil☑, coast☑ stop☑, remote☑ leak☑
· **04-26-2010**	
`S1`: In a reverse, officials reveal the well is leaking an estimated 1,000 barrels of oil per day and warn of environmental disaster. `S2`: Meanwhile, BP sends undersea robots to the wellhead in an unsuccessful effort to activate the blowout preventer, a piece of heavy kit mounted on top of the well to stem the flow of oil.	oil☑, officials☑ disaster☑ preventer☑ flow☑
· **04-28-2010**	
`S1`: The US Coast Guard sets fire to patches of spilled oil in an effort to prevent the slick from reaching the vulnerable Louisiana coastal wetlands.	coast☑, louisiana☑ wetlands☑ obama, cleanup
· **05-11-2010**	
`S1`: At a series of congressional hearings, BP, Transocean and Halliburton, the three companies involved in the Deepwater Horizon drilling operations, all blame each other for the disaster.	congressional☑ companies☑ blame☑, oil, u.s.
Other daily summaries...	

Table 4. The timeline on *BP oil spill* generated by our model. Due to space limit, we show up to top 5 descriptive words for each cluster and mark ☑ for those words existed in our summaries (which are extracted from the cluster).

7. CONCLUSIONS

We have presented a new approach for timeline summarization (TLS) by explicitly optimizing the importance and coherence on top of coverage and diversity, which are flexible to optimize contents information. Our approach achieves absolute improvements and benefits from readability overhead because the submodular property offered by our method permits generating informative and coherent summaries.

[6] We choose the generated 10 successive daily summaries on *BP oil spill* from TL17. Here, we invite three human annotators (excluding the authors of this paper) who have good knowledge of natural language generation to assign scores to the samples. In the 5-point Likert scale, 5-point means 'very satisfying', while 1-point means' very terrible'. We further average the three annotated scores over the 10 daily summaries.

8. REFERENCES

[1] Hui Lin and Jeff A. Bilmes, "A class of submodular functions for document summarization," in *ACL, Volume 1 (Long Papers)*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, Eds. 2011, pp. 510–520, Association for Computational Linguistics.

[2] Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang, "Evolutionary timeline summarization: a balanced optimization framework via iterative substitution," in *SIGIR*, 2011, pp. 745–754.

[3] Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen, "Predicting relevant news events for timeline summaries," in *WWW*, 2013, pp. 91–92.

[4] Giang Binh Tran, Tuan A Tran, Nam-Khanh Tran, Mohammad Alrifai, and Nattiya Kanhabua, "Leveraging learning to rank in an optimization framework for timeline summarization," in *SIGIR 2013 Workshop on Time-aware Information Access (TAIA)*, 2013.

[5] Kiem-Hieu Nguyen, Xavier Tannier, and Véronique Moriceau, "Ranking multidocument event descriptions for building thematic timelines," in *Technical Papers, COLING*, Jan Hajic and Junichi Tsujii, Eds., 2014, pp. 1208–1217.

[6] Giang Tran, Mohammad Alrifai, and Eelco Herder, "Timeline summarization from relevant headlines," in *European Conference on Information Retrieval*. Springer, 2015, pp. 245–256.

[7] Sebastian Martschat, Katja Markert, and Sebastian Martschat, "A temporally sensitive submodularity framework for timeline summarization," in *CoNLL*, Anna Korhonen and Ivan Titov, Eds., 2018, pp. 230–240.

[8] Moreno La Quatra, Luca Cagliero, Elena Baralis, Alberto Messina, and Maurizio Montagnuolo, "Summarize dates first: A paradigm shift in timeline summarization," in *SIGIR*. 2021, pp. 418–427, ACM.

[9] Shinsaku Sakaue, Tsutomu Hirao, Masaaki Nishino, and Masaaki Nagata, "Provable fast greedy compressive summarization with any monotone submodular function," in *NAACL-HLT, Volume 1 (Long Papers)*. 2018, pp. 1737–1746, Association for Computational Linguistics.

[10] Yiming Liao, Shuguang Wang, and Dongwon Lee, "WILSON: A divide and conquer approach for fast and effective news timeline summarization," in *EDBT*. 2021, pp. 635–645, OpenProceedings.org.

[11] Maxim Sviridenko, "A note on maximizing a submodular set function subject to a knapsack constraint," *Oper. Res. Lett.*, vol. 32, no. 1, pp. 41–43, 2004.

[12] Ryan T. McDonald, "A study of global inference algorithms in multi-document summarization," in *ECIR*, Giambattista Amati, Claudio Carpineto, and Giovanni Romano, Eds., 2007, vol. 4425 of *Lecture Notes in Computer Science*, pp. 557–564.

[13] Demian Gholipour Ghalandari, Georgiana Ifrim, and Georgiana Ifrim, "Examining the state-of-the-art in news timeline summarization," in *ACL, Volume 1 (Long Papers)*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, Eds. 2020, pp. 1322–1334, Association for Computational Linguistics.

[14] Aqil M. Azmi and Suha Al-Thanyyan, "A text summarizer for arabic," *Comput. Speech Lang.*, vol. 26, no. 4, pp. 260–273, 2012.

[15] Yi Yu, Adam Jatowt, Antoine Doucet, Kazunari Sugiyama, and Masatoshi Yoshikawa, "Multi-timeline summarization (MTLS): improving timeline summarization by generating multiple summaries," in *ACL/IJCNLP, (Volume 1: Long Papers)*. 2021, pp. 377–387, Association for Computational Linguistics.

[16] Joseph L Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, pp. 378, 1971.

[17] L. Page, "The pagerank citation ranking : Bringing order to the web," 1998.

[18] Julius Steen and Katja Markert, "Abstractive timeline summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 2019, pp. 21–31.

[19] Demian Gholipour Ghalandari, "Revisiting the centroid-based method: A strong baseline for multi-document summarization," in *NFiS@EMNLP*, Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, Eds. 2017, pp. 85–90, Association for Computational Linguistics.

[20] Rada Mihalcea and Paul Tarau, "Textrank: Bringing order into text," in *EMNLP*. 2004, pp. 404–411, Association for Computational Linguistics.

[21] Philipp Koehn, "Statistical significance tests for machine translation evaluation," in *EMNLP*. 2004, pp. 388–395, Association for Computational Linguistics.

[22] Stefan Riezler and John T. Maxwell III, "On some pitfalls in automatic evaluation and significance testing for MT," in *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, Eds. 2005, pp. 57–64, Association for Computational Linguistics.