Exploiting Timelines to Enhance Multi-document Summarization

Jun-Ping Ng[1,2], Yan Chen[3], Min-Yen Kan[2,4], Zhoujun Li[3]

[1]DSO National Laboratories, Singapore
[2]School of Computing, National University of Singapore, Singapore
[3]State Key Laboratory of Software Development Environment, Beihang University, China
[4]Interactive and Digital Media Institute, National University of Singapore, Singapore
njunping@dso.org.sg

Abstract

We study the use of temporal information in the form of timelines to enhance multi-document summarization. We employ a fully automated temporal processing system to generate a timeline for each input document. We derive three features from these timelines, and show that their use in supervised summarization lead to a significant 4.1% improvement in ROUGE performance over a state-of-the-art baseline. In addition, we propose TIMEMMR, a modification to Maximal Marginal Relevance that promotes temporal diversity by way of computing time span similarity, and show its utility in summarizing certain document sets. We also propose a filtering metric to discard noisy timelines generated by our automatic processes, to purify the timeline input for summarization. By selectively using timelines guided by filtering, overall summarization performance is increased by a significant 5.9%.

1 Introduction

There has been a good amount of research invested into improving the temporal interpretation of text. Besides the increasing availability of annotation standards (e.g., TIMEML (Pustejovsky et al., 2003a)) and corpora (e.g., TIDES (Ferro et al., 2000), TimeBank (Pustejovsky et al., 2003b)), the community has also organized three successful evaluation workshops — TempEval-1 (Verhagen et al., 2009), -2 (Verhagen et al., 2010), and -3 (Uzzaman et al., 2013). As the state-of-the-art improves, these workshops have moved away from the piecemeal evaluation of individual temporal processing tasks and towards the evaluation of complete end-to-end systems in TempEval-3. We believe our understanding of the temporal information found in text is sufficiently robust, and that there is an opportunity to now leverage this information in downstream applications. In this paper, we present our work in incorporating the use of such temporal information into multi-document summarization.

The goal of multi-document summarization is to generate a summary which includes the main points from an input collection of documents with minimal repetition of similar points. We hope to improve the quality of the summaries that are generated by considering temporal information found in the input text. To motivate how temporal information can be useful in summarization, let us refer to Figure 1. The three sentences describe a recent cyclone and a previous one which happened in 1991. Recognizing that sentence (3) is about a storm that had happened in the past is important when writing a summary about the recent storm, as it is not relevant and can likely be excluded.

It is reasonable to expect that a collection of documents about the recent storm will contain more references to it, compared with the earlier one that happened in 1991. Visualized on a timeline, this will translate to more events (bolded in Figure 1) around the time when the recent storm occurred. There should be fewer events mentioned in the collection for the earlier 1991 time period. Figure 2 illustrates a possible timeline laid out with the events found in Figure 1. The events from the more recent storm are found together at the same time. There are fewer events which talk about the previous storm. Thus, temporal information does assist in identifying which sentences are more relevant to the final summary.

Our work is significant as it addresses an important gap in the exploitation of temporal information. While there has been prior work making use of temporal information for multi-document

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 923–933,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

(1) A fierce cyclone **packing** extreme winds and torrential rain **smashed** into Bangladesh's southwestern coast Thursday, **wiping** out homes and trees in what officials **described** as the worst storm in years.
(2) More than 100,000 coastal villagers have been **evacuated** before the cyclone made landfall.
(3) The storm matched one in 1991 that **sparked** a tidal wave that **killed** an estimated 138,000 people, Karmakar told AFP.

Figure 1: Modified extract from a news article which describes a cyclone landfall. Several events which appear in Figure 2 are bolded.

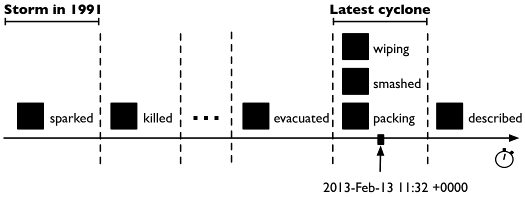


Figure 2: Possible timeline for events in Figure 1.

summarization, they 1) have been largely confined to helping to chronologically order content within summaries (Barzilay et al., 1999), or 2) focus only on the use of recency as an indicator of saliency (Goldstein et al., 2000; Wan, 2007). In this work we construct timelines (as a representation of temporal information) automatically and incorporate them into a state-of-the-art multi-document summarization system. This is achieved with 1) three novel features derived from timelines to help measure the saliency of sentences, as well as 2) TIMEMMR, a modification to the traditional Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998). TimeMMR promotes diversity by additionally considering temporal information instead of just lexical similarities. Through these, we demonstrate that temporal information is useful for multi-document summarization. Compared to a competitive baseline, significant improvements of up to 4.1% are obtained.

Automatic temporal processing systems are not perfect yet, and this may have an impact on their use for downstream applications. This work additionally proposes the use of the lengths of timelines as a metric to gauge the usefulness of timelines. Together with the earlier described contributions, this metric further improves summarization, yielding an overall 5.9% performance increase.

2 Related Work

Barzilay et al. (1999) were one of the first to use time for multi-document summarization. They recognized the importance of generating a summary which presents the time perspective of the summarized documents correctly. They estimated the chronological ordering of events with a small set of heuristics, and also made use of lexical patterns to perform basic time normalization on terms like "today" relative to the document creation time. The induced ordering is used to present the selected summary content, following the chronological order in the original documents.

In another line of work, Goldstein et al. (2000) made use of the temporal ordering of documents to be summarized. In computing the relevance of a passage for inclusion into the final summary, they considered the recency of the passage's source document. Passages from more recent documents are deemed to be more important. Wan (2007) and Demartini et al. (2010) made similar assumptions in their work on TIMEDTEXTRANK and entity summarization, respectively.

Instead of just considering the notion of recency, Liu et al. (2009) proposed an interesting approach using a temporal graph. Events within a document set correspond to vertices in their proposed graph, while edges are determined by the temporal ordering of events. From the resulting weakly-connected graph, the largest forests are assumed to contain key topics within the document set and used to influence a scoring mechanism which prefers sentences touching on these topics.

Wu (2008) also made use of the relative ordering of events. He assigned complete timestamps to events extracted from text. After laying out these events onto a timeline by making use of these timestamps, the number of events that happen within the same day is used to influence sentence scoring. The motivation behind this approach is that days which have a large number of events should be more important and more worthy of reporting than others.

These prior works target either 1) sentence reordering, or 2) the use of recency as an indicator of saliency. In sentence re-ordering, final summaries are re-arranged so that the extracted sentences that form the summary are in a chronological order. We argue that this may not be appropriate for all summaries. Depending on the style of writing or journalistic guidelines, a summary can arguably be written in a number of ways. The use of recency as an indicator of saliency is useful, yet disregards other accessible temporal information. If a summary of a whole sequence of events is desired, recency becomes less useful.

The work of Wu (2008) is closely related to one of the features proposed in this paper. He had also made use of temporal information to weight sentences to generate summaries. However his approach is guided by the number of events happening within the same time span, and relies on event co-referencing. In this work, we have simplified this idea by dropping the need for event co-referencing (removing a source of propagated error), and augmented it with two additional features derived from timelines. By doing so, we are able to make better use of the available temporal information, taking into account all known events and the time in which they occur.

A useful note here is that this work is arguably different from the Temporal Summarization (TmpSum) track at the Text Retrieval Conference (Aslam et al., 2013). Given a large stream of data in real-time, the purpose of the TmpSum track is to look out for a query event, and retrieve specific details about the event over a period of time. Systems are also expected to identify the source sentences from which these details are retrieved. This is not the same as our approach here, which makes use of temporal information encoded in timelines to generate prose summaries.

3 Methodology

To incorporate temporal information into multi-document summarization, we adopt the workflow in Figure 3, which has two key processes: 1) temporal processing, and 2) summarization.

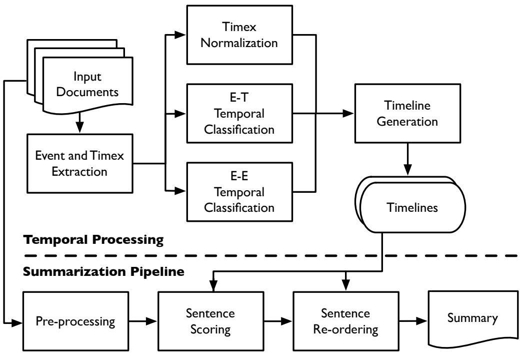


Figure 3: Incorporating temporal information into the SWING summarization pipeline.

Temporal Processing generates timelines from text, one for each input document. Timelines are well-understood constructs which have often been used to represent temporal information (Denis and Muller, 2011; Do et al., 2012). They indicate the temporal relationships between two basic temporal units: 1) events, and 2) time expressions (or timexes for short). In this work, we adopt the definitions proposed in the standardized TIMEML annotation (Pustejovsky et al., 2003a). An event refers to an eventuality, a situation that occurs or an action; while a timex is a reference to a particular date or time (e.g. "*2013 December 31*").

Following the "divide-and-conquer" approach described in Verhagen et al. (2010), results from the three temporal processing steps: 1) timex normalization, 2) event-timex temporal relationship classification, and 3) event-event temporal relationship classification, are merged to obtain timelines (top half of Figure 3). We tap on existing systems for each of these steps (Ng and Kan, 2012; Strötgen and Gertz, 2013; Ng et al., 2013).

Summarization. We make use of a state-of-the-art summarization system, SWING (Ng et al., 2012) (bottom half of Figure 3). SWING is a supervised, extractive summarization system which ranks sentences based on scores computed using a set of features in the *Sentence Scoring* phase. The Maximal Marginal Relevance (MMR) algorithm is then used in the *Sentence Re-ordering* phase to re-order and select sentences to form the final summary. The timelines built in the earlier temporal processing can be incorporated into this pipeline by deriving a set of features used to score sentences in *Sentence Scoring*, and as input to the MMR algorithm when computing similarity in *Sentence Re-ordering*.

3.1 Timelines from Temporal Processing

A typical timeline used in this work has been shown earlier in Figure 2. The arrowed, horizontal axis is the timeline itself. The timeline can be viewed as a continuum of time, with points on the timeline referring to specific moments of time. Small solid blocks on the timeline itself are references to absolute timestamps along the timeline (e.g., "*2013-Feb-13 11:32 +0000*" in the figure).

The black square boxes above the timeline denote events. Events can either occur at a specific instance of time (e.g., an explosion), or over a period of time (e.g. a football match). Generalizing, we refer to the time period an event takes place in as its *time span* (vertical dotted lines). As a simpli-

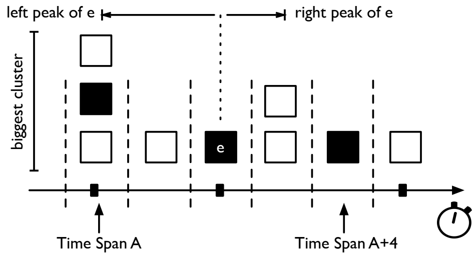


Figure 4: A simplified timeline illustrating how the various timeline features can be derived.

fying assumption, events are laid out on the timeline based on the starting time of their time span. Note that in our work, time spans may not correspond to specific instances of time, but instead help in inferring an ordering of events. Events which appear to the left of others take place earlier, while events within the same time span happen together over the same time period.

3.2 Sentence Scoring with Timelines

We derive three features from the constructed timelines, which are then used for downstream *Sentence Scoring*. Figure 4 shows a simplified timeline, along with annotations that are referenced in this section to help explain how these timeline features are derived.

1. Time Span Importance (TSI). We hypothesize that when more events happen within a particular time span, that time span is potentially more relevant for summarization. Sentences that mention events found in such a time span should be assigned higher scores. Referring to Figure 1, whose timeline is shown in Figure 2, we see that the time span with the most number of events is when the latest cyclone made landfall. Assigning higher scores for sentences which contain events in this time span will help us to select more relevant sentences if we want a summary about the cyclone.

Let TS_L be the time span with the largest number of events in a timeline. The importance of a time span TS_i is computed by normalizing the number of events in TS_i against the number of events in TS_L. The TSI of a sentence s is then the sum of the time span importance associated to all the words in s:

$$TSI(s) = \frac{\sum_{w \in s} \frac{|TS_w|}{|TS_L|}}{|s|} \qquad (1)$$

where TS_w denotes the time span which a word w is associated with, and $|TS_w|$ is the number of events within the time span.

2. Contextual Time Span Importance (CTSI). The importance of a time span may not depend solely on the number of events that happen within it. If it is near time spans which are "important" (i.e., one that has a large number of events), it should also be of relative importance. A more concrete illustration of this can also be seen in Figure 1. Sentence (2) explains that a lot of people have been evacuated prior to the cyclone making landfall. It is imaginable that this can be useful information to be included in a summary, even though from looking at the corresponding timeline in Figure 2, the "*evacuated*" event falls in a time span with a low importance score (i.e., the time span only has one event). CTSI seeks to promote sentences such as this.

We derive the CTSI of a sentence by first computing the contextual importance of words in the sentence. We define the contextual importance of a word found in time span TS_i as a weighted sum of the time span importance of the two nearest peaks TS_{lp} and TS_{rp} found to the left and right of TS_i, respectively. In Figure 4, taking reference from event e (shaded in black), the left peak to the time span which e is in happens to be time span A, while the right peak is time span $A + 4$. The contribution of each peak to the weighted sum is decayed by its distance from TS_i. Formally, the contextual time span importance of a word w can be expressed as:

$$\zeta(w) = \alpha \left(\frac{I_{lp}}{|TS_w - TS_{lp}|} \right) \times (1-\alpha) \left(\frac{I_{rp}}{|TS_{rp} - TS_w|} \right) \qquad (2)$$

where TS_w is the time span associated with w. I_{lp} and I_{rp} are the time span importance of the peaks to the left and right of TS_w respectively, while $|TS_w - TS_{lp}|$ and $|TS_{rp} - TS_w|$ are the number of time spans between the left and right peaks of TS_w respectively. α balances the importance of the left and right peaks, intuitively set to 0.5. The CTSI of a sentence is computed as:

$$CTSI(s) = \frac{\sum_{e \in \mathbb{E}_s} \zeta(e)}{|\mathbb{E}_s|} \qquad (3)$$

where $\mathbb{E}_s$ denotes the set of events words in s.

3. Sentence Temporal Coverage Density (TCD). We first define the *temporal coverage* of a sentence. This corresponds to the number of time spans that the events in a sentence talk about. Suppose a sentence contains events which are associated with time spans TS_a, TS_b, TS_c. The time spans are ordered in the sequence they appear on the timeline. Then the temporal coverage of a sentence is defined as the number of time spans between the earliest time span TS_a and the latest time span TS_c. Referring to Figure 4, suppose a sentence contains the three events which have been shaded black. The temporal coverage in this case includes all the time spans from time span A to time span $A + 4$, inclusive.

The constraint on the number of sentences that can be included in a summary requires us to select compact sentences which contain as many relevant facts as possible. Traditional lexical measures may attempt to achieve this by computing the ratio of keyphrases to the number of words in a sentence (Gong and Liu, 2001). Stated equivalently, when two sentences are of the same length, if one contains more keyphrases, it should contain more useful facts. TCD parallels this idea with the use of temporal information, i.e. if two sentences are of the same temporal coverage, then the one with more events should carry more useful facts.

Formally, if a sentence s contains events $\mathbb{E}_s = \{e_1, \ldots, e_n\}$, where each event is associated with a time span TS_i, then TCD is computed using:

$$TCD(s) = \frac{|\mathbb{E}_s|}{|TS_n - TS_1|} \quad (4)$$

where $|\mathbb{E}_s|$ is the number of events found in s, and $|TS_n - TS_1|$ is the temporal coverage of s.

3.3 Enhancing MMR with TimeMMR

In the sentence re-ordering stage of the `SWING` pipeline, the iterative MMR algorithm is used to adjust the score of a candidate sentence, s. In each iteration, s is penalized if it is lexically similar to other sentences that have already been selected to form the eventual summary $S = \{s_1, s_2, \ldots\}$. The motivating idea is to reduce repeated information by preferring sentences which bring in new facts.

Incorporating temporal information can potentially improve this. In Figure 5, the sentences describe many events which took place within the same time span. They describe the destruction caused by a hurricane with trees uprooted and buildings blown away. A summary about the hurricane need not contain all of these sentences as they are all describing the same thing. However it is not trivial for the lexically-motivated MMR algorithm to detect that events like "*passed*", "*uprooted*" or "*damaged*" are in fact repetitive.

Thus, we propose further penalizing the score of s if it contains events that happen in similar time spans as those contained in sentences within S. We refer to this as TIMEMMR. Modifying the MMR equation from Ng et al. (2012):

$$TimeMMR(s) = Score(s) - \gamma R2(s, S) - (1 - \gamma)\mathcal{T}(s, S) \quad (5)$$

where $Score(s)$ is the score of s, S is the set of sentences already selected to be in the summary from previous iterations, and $R2$ is the predicted ROUGE-2 score of s with respect to the already selected sentences (S). γ is a weighting parameter which is empirically set to 0.9 after tuning over a development dataset. $\mathcal{T}$ is the proportion of events in s which happen in the same time span as another event in any other sentence in S. Two events are said to be in the same time span if one happens within the time period the other happens in. For example, an event that takes place in "*2014 June*" is said to take place within the year "*2014*".

While TIMEMMR is proposed here as an improvement over MMR, the premise is that incorporating temporal information can be helpful to minimize redundancy in summaries. In future work, one could apply it to other state-of-the-art lexical-based approaches including that of Hendrickx et al. (2009) and Celikyilmaz and Hakkani-Tur (2010). We also believe the same idea can be transplanted even to non-lexical motivated techniques such as the corpus-based similarity measure proposed by Xie and Liu (2008). We chose to use MMR here as a proof-of-concept to demonstrate the viability of such a technique, and to easily integrate our work into `SWING`.

3.4 Gauging Usefulness of Timelines

Temporal processing is imperfect. Together with the simplifying assumptions that were made in timeline construction, our generated timelines have errors which propagate into the summarization process. With this in mind, we selectively employ timelines to generate summaries only when we are confident of their accuracy. This can be done by computing a metric which can be used to decide whether or not timelines should be used for a particular input document collection. We refer to this as *reliability filtering*.

We postulate that the length of a timeline can serve as a simple reliability filtering metric. The intuition for this is that for longer timelines (which contain more events), possible errors are spread over the entire timeline, and do not overpower any useful signal that can be obtained from the timeline features outlined earlier. Errors are however

(1) An official in Barisal, 120 kilometres south of Dhaka, spoke of severe **destruction** as the 500 kilometre-wide mass of cloud **passed** overhead.
(2) "Many trees have been **uprooted** and houses and schools **blown** away," Mostofa Kamal, a district relief and rehabilitation officer, told AFP by telephone.
(3) "Mud huts have been **damaged** and the roofs of several houses **blown** off," said the state's relief minister, Mortaza Hossain.

Figure 5: Extract from a news article which describes several events (bolded) happening at the same time.

very easily propagated into summary generation for shorter timelines, leading to less useful results.

We incorporate this into our process as follows: given an input document collection (which consists of 10 documents), the average size of all the timelines for each of these 10 documents is computed. Only when this value is larger than a threshold value are the timelines used.

4 Experiments and Results

The proposed timeline features and TimeMMR were implemented on top of SWING, and evaluated on the test documents from TAC-2011 (Owczarzak and Dang, 2011). SWING makes use of three generic features and two features targeted specifically at guided summarization. Since the focus of this paper is on multi-document summarization, we employ only the three generic features, i.e., 1) sentence position, 2) sentence length, and 3) interpolated n-gram document frequency in our experiments below. Summarization evaluation is done using ROUGE-2 (R-2) (Lin and Hovy, 2003), as it has previously been shown to correlate well with human assessment (Lin, 2004) and is often used to evaluate automatic text summarization.

The results obtained are shown in Table 1. In the table, each row refers to a specific summarization system configuration. We also show the results of two reference systems, CLASSY (Conroy et al., 2011) and POLYCOM (Zhang et al., 2011), as benchmarks. CLASSY and POLYCOM are top performing systems at TAC-2011 (ranked 2nd and 3rd by R-2 in TAC 2011, respectively; the full version of SWING was ranked 1st with a R-2 score of 0.1380). From these results, we can see that SWING is a very competitive baseline.

Rows 9 to 16 additionally incorporate our timeline reliability filtering. We assume that the various input document sets to be summarized are available at the time of processing. Hence in these experiments, the threshold for filtering is set to be the average of all the timeline sizes over the whole input dataset. In a production environment where this assumption may not hold, this threshold could be set by empirical tuning over a development set.

	Configuration	**R-2**	**Sig**
R	SWING	0.1339	NA
B1	CLASSY	0.1278	-
B2	POLYCOM	0.1227	**
Without Filtering			
1	SWING+TSI+CTSI+TCD	0.1394	*
2	SWING+TSI+CTSI	0.1372	-
3	SWING+TSI+TCD	0.1372	-
4	SWING+CTSI+TCD	0.1387	*
5	SWING+TSI+CTSI+TCD+TMMR	0.1389	-
6	SWING+TSI+CTSI+TMMR	0.1374	-
7	SWING+TSI+TCD+TMMR	0.1343	-
8	SWING+CTSI+TCD+TMMR	0.1363	-
With Filtering			
9	SWING+TSI+CTSI+TCD	0.1418	**
10	SWING+TSI+CTSI	0.1378	**
11	SWING+TSI+TCD	0.1389	**
12	SWING+CTSI+TCD	0.1401	**
13	SWING+TSI+CTSI+TCD+TMMR	0.1402	**
14	SWING+TSI+CTSI+TMMR	0.1397	**
15	SWING+TSI+TCD+TMMR	0.1376	*
16	SWING+CTSI+TCD+TMMR	0.1390	**

Table 1: R-2 scores after incorporating temporal information into SWING. '**' and '*' denotes significant differences with respect to Row R (paired one-tailed Student's t-test; $p < 0.05$ and $p < 0.1$ respectively), and TMMR denotes TimeMMR.

Row 1 shows the usefulness of the proposed timeline-based features. A statistically significant improvement of 4.1% is obtained with the use of all three features over SWING. When we use reliability filtering (Row 9), this improvement increases to 5.9%.

The ablation test results in Rows 2 to 4 show a drop in R-2 each time a feature is left out. With the exception of Row 4, removing a feature lessens the improvement in R-2 to be insignificant from SWING's. The same drop occurs even when reliability filtering is used (Rows 9 to 12). These indicate that all the proposed features are important and need to be used together to be effective.

Rows 5 to 8 and Rows 13 to 16 show the effect of TimeMMR. While the results do not uniformly show that TimeMMR is effective, it can be helpful, such as when comparing Rows 2 and 6, or Rows 10 and 14, where R-2 improves marginally.

Looking at Rows 1 to 8, and Rows 9 to 16, we see the importance of reliability filtering. It is able to guide the use of timelines such that significant improvements in R-2 over SWING are obtained.

To help visualize what the differences in these ROUGE scores mean, Figure 7 shows two summaries[1] generated for document set D1117C of the TAC-2011 dataset. The left one is produced by the configuration in Row 9, and the right one is produced by SWING without the use of any temporal information.

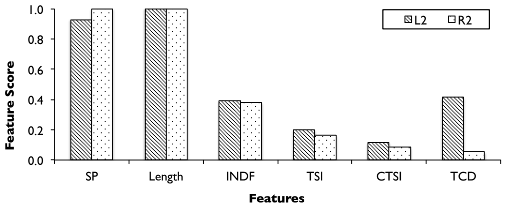


Figure 6: Breakdown of raw feature scores for sentences ($\mathbb{L}2$) and ($\mathbb{R}2$) from Figure 7.

The higher R-2 score obtained by the summary on the left (0.0873) compared to the one on the right (0.0723) suggests that temporal information can help to identify salient sentences more accurately. A closer look at sentences ($\mathbb{L}2$) and ($\mathbb{R}2$) and their R-2 scores (0.0424 and 0.0249, respectively) is instructive. Figure 6 shows the raw feature scores of both sentences. Both sentences score similarly for the SWING features of sentence position (SP), sentence length (Length), and interpolated n-gram document frequency (INDF); however, the scores for all three timeline features higher for ($\mathbb{L}2$) than ($\mathbb{R}2$). This helps our time sensitive system prefer ($\mathbb{L}2$).

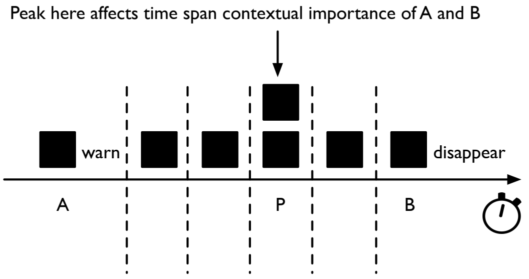
5 Discussion

We now examine the proposed 1) timeline features, 2) TIMEMMR algorithm, and 3) reliability filtering metric in greater detail to gain insight into their efficacy. For the analysis on timeline features, we only present an analysis for TSI and CTSI due to space constraints.

Time Span Importance. Figure 8 shows the last sentences from a pair of summaries generated with and without the use of TSI (all other sentences were the same). The original articles describe an accident where casualties were suffered when a crane toppled onto a building. It is easy to see why ($\mathbb{L}1$) scores higher for R-2 — it describes the cause of the accident just as it occurred. ($\mathbb{R}1$) however talks about events which happened before the accident itself (e.g., how much of the tower had already been erected). In this case time span importance is able to correctly guide summary generation by favoring time spans containing events related to the actual toppling.

Contextual Time Span Importance. CTSI recognizes that events which happen around the time of a big cluster of other events can be important too. The benefits of this feature can be clearly seen in Figure 9. The summary on the left achieved a R-2 score of 0.1215 while the one on the right achieved 0.0861. ($\mathbb{L}2$) and ($\mathbb{L}3$) were both boosted by the use of the contextual importance feature.

Figure 10 shows an extract of the timeline generated for the source document from which ($\mathbb{L}3$) is extracted. The two events inside ($\mathbb{L}3$) fall in time spans A and B marked in the figure. Their proximity to the peak P between them gives the sentence a higher score for CTSI. This boost results in the sentence being selected for inclusion in the final summary. It turns out that this sentence was lifted exactly in one of the model summaries for this document set, resulting in a very good R-2 score when contextual importance is used.



Peak here affects time span contextual importance of A and B
warn
disappear
A
P
B

Figure 10: Extract of timeline generated for document APW_ENG_20070615.0356 from the TAC-2011 dataset.

Is TIMEMMR Useful? The experimental results do not conclusively affirm the usefulness of TIMEMMR. However we believe it is because the ROUGE measures that are used for evaluation are not suited for this purpose. Recall that TIMEMMR seeks to eliminate redundancy based on time span similarities and not lexical likeness. ROUGE, however, measures the latter.

An interesting case in point is given in Figure 11. The summary on the left is generated using TIMEMMR and achieved a lower ROUGE score. The one on the right is generated without TIMEMMR and scores higher, suggesting that TIMEMMR is not helpful. The key difference in

[1] The produced summaries are truncated to fit within a 100-word limit imposed by the TAC-2011 guidelines.

R-2: **0.0873**		R-2: **0.0723**
(𝕃1,ℝ1) The Army's surgeon general criticized stories in The Washington Post disclosing problems at Walter Reed Army Medical Center, saying the series unfairly characterized the living conditions and care for soldiers recuperating from wounds at the hospital's facilities.		
(𝕃2) Defense Secretary Robert Gates says people found to have been responsible for allowing substandard living conditions for soldier outpatients at Walter Reed Army Medical Center in Washington will be "held accountable," although so far no one in the Army chain of command has offered to resign.	≠≠	(ℝ2) A top Army general vowed to personally oversee the upgrading of Walter Reed Army Medical Center's Building 18, a dilapidated former hotel that houses wounded soldiers as outpatients.
(𝕃3) Top Army officials visited Building 18, the decrepit former hotel housing more than 80 recovering soldiers, outside	≠≠	(ℝ3) "I'm not sure it was an accurate representation," Lt. Gen. Kevin Kiley, chief of the Army Medical Command which oversees Walter Reed and all Army health care, told reporters during a news conference.
	>>	(ℝ4) The Washington

Figure 7: Generated summaries for document set D1117C from the TAC-2011 dataset. Left summary is generated by SWING+TSI+CTSI+TCD with filtering; right summary is by SWING.

R-2: **0.1683**		R-2: **0.1533**
	....	
(𝕃1) A piece of steel fell and sheared off one of the ties holding it to the building, causing it to detach and topple, said Stephen Kaplan	≠≠	(ℝ1) About 19 of the 44 stories of the crane had been erected and it was to be extended when a piece of steel fell and sheared

Figure 8: Extract from summaries for document set D1137G from the TAC-2011 dataset. Left extract is generated by SWING+TSI+CTSI+TCD; right extract is by SWING+CTSI+TCD.

the two summaries is (ℝ3). (𝕃3) is the equivalent of (ℝ4), while (𝕃4) is the full version of the truncated (ℝ5). TIMEMMR penalizes (ℝ3). (ℝ3) reports that the shoe-throwing incident happened as the U.S. President Bush appeared together with the Iraqi Prime Minister Nouri al-Maliki. However their joint appearance is already reported in (ℝ1) (and similarly (𝕃1)). (ℝ3) repeats what had been presented earlier. Since (ℝ1) and (ℝ3) talk about the same time span, TIMEMMR down-weights (ℝ3). We argue that this is better even though the ROUGE scores indicate otherwise. In future work it will be worthwhile to consider the use of metrics like Pyramid (Passonneau et al., 2005) which are less bound to superficial lexicons.

Reliability Filtering. Table 2 shows the effect of varying the filtering threshold on R-2 for the best performing configuration from Table 1 (i.e., SWING+TSI+CTSI+TCD). The result obtained in Row 9 using a threshold of 42.68 is also re-produced for reference. **T**=0 means that timelines are used for all input document sets, whereas **T**=100 means that no timelines are used, as the length of the longest timeline is less than 100.

As the threshold increases from 0 to 40–50, summarization performance improves while the number of document sets where temporal information is used is reduced. This suggests that filtering is successful in identifying timelines that are not sufficiently accurate to be useful for summarization. R-2 performance peaks around a threshold of 40. This affirms our use of the average length of timelines as the threshold value in our earlier experiments. Beyond 60, the R-2 scores are still higher than that obtained by SWING, but no longer significantly different. At these higher thresholds, temporal information is still able to help get an improvement in R-2. However as this affects only very few out of the 44 document sets, statistical variances mean that these R-2 scores are no longer

T	R-2	Sig	#	T	R-2	Sig	#
0	0.1394	*	44	**50**	0.1386	**	13
10	0.1382	-	43	**60**	0.1361	*	7
20	0.1377	-	41	**70**	0.1351	-	3
30	0.1393	**	35	**80**	0.1351	-	2
40	0.1426	**	22	**90**	0.1353	-	1
42.68	*0.1418*	**	*21*	**100**	0.1339	-	0

Table 2: Effect of different reliability filtering thresholds for SWING+TSI+CTSI+TCD. '**T**' is the threshold used; '**#**' is the number of input collections (out of 44) where timelines are used; '**' and '*' is statistical significance over SWING of $p < 0.05$ and $p < 0.1$, respectively.

R-2: **0.1215**		R-2: **0.0861**
((𝕃1,ℝ1) Caribbean coral species essential to the region's reef ecosystems are at risk of extinction as a result of climate change.		
(𝕃2) But destructive fishing methods and over-harvesting have reduced worldwide catches by 90 percent in the past two decades.	≠≠	(ℝ2) The Coral Reef Task Force, created in the Clinton administration, regularly assesses coral health.
(𝕃3) Scientists warn that up to half of the world's coral reefs could disappear by 2045.	≠≠	(ℝ3) With a finished necklace retailing for up to 20,000 dollars (15,000 euros), red corals are among the world's most expensive wildlife commodities.
	....	

Figure 9: Extract from summaries for document set D1131F from the TAC-2011 dataset. Left extract is generated by SWING+TSI+CTSI+TCD; right extract is by SWING+TSI+TCD.

R-2: **0.2643**		R-2: **0.2772**
(𝕃1,ℝ1) – An Iraqi reporter threw his shoes at visiting U.S. President George W. Bush and called him a "dog" in Arabic during a news conference with Iraqi Prime Minister Nuri al-Maliki in Baghdad		
(𝕃2,ℝ2) "All I can report is it is a size 10,.		
(𝕃3) Muntadhar al-Zaidi, reporter of Baghdadiya television jumped and threw his two shoes one by one at the president, who ducked and thus narrowly missed being struck, raising chaos in the hall in Baghdad's heavily fortified green Zone.	≠≠	(ℝ3) The incident occurred as Bush was appearing with Iraqi Prime Minister Nouri al-Maliki.
(𝕃4) The president lowered his head and the first shoe hit the American and Iraqi flags behind the two leaders.	≠≠	(ℝ4) Muntadhar al-Zaidi, reporter of Baghdadiya television jumped and threw his two shoes one by one at the president, who ducked and thus narrowly missed being struck, raising chaos in the hall in Baghdad's heavily fortified green Zone.
(𝕃5) The	≠≠	(ℝ5) The president lowered his head and the

Figure 11: Summaries for document set D1126E from the TAC-2011 dataset. Left summary is generated by SWING+TSI+CTSI+TCD+TimeMMR; right summary is by SWING+TSI+CTSI+TCD.

significant from that produced by SWING.

6 Conclusion

We have shown in this work how temporal information in the form of timelines can be incorporated into multi-document summarization. We achieve this through two means, using: 1) three novel features derived from timelines to measure the saliency of sentences, and 2) TimeMMR which considers time span similarity to enhance the traditional MMR's lexical diversity measure.

To overcome errors propagated from the underlying temporal processing systems, we proposed a reliability filtering metric which can be used to help decide when temporal information should be used for summarization. The use of this metric leads to an overall 5.9% gain in R-2 over the competitive SWING baseline.

In future work, we are keen to study our proposed timeline-related features more intrinsically in the context of human-generated summaries. This can help us better understand their value in improving content selection. As noted earlier, it will be also be useful to repeat our experiments with less lexicon-influenced measures like the Pyramid method (Passonneau et al., 2005). Manual assessment of the generated summaries can also be done to give a better picture of the quality of the summaries generated with the use of timelines. Finally, given the importance of reliability filtering, a natural question is if there are other metrics that can be used to get better results.

Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

This work is also partially supported by the National Natural Science Foundation of China (Grant Nos. 61170189, 61370126, 61202239), the Fund of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2013ZX-19), and the Innovation Foundation of Beihang University for Ph.D. Graduates (YWF-13-T-YJSY-024).

References

Javed Aslam, Matthew Ekstrand-Abueg, Virgil Pavlu, Fernado Diaz, and Tetsuya Sakai. 2013. TREC 2013 Temporal Summarization. In *Proceedings of the 22nd Text Retrieval Conference (TREC)*, November.

Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1999. Information Fusion in the Context of Multi-document Summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL)*, pages 550–557, June.

Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 335–336, August.

Asli Celikyilmaz and Dilek Hakkani-Tur. 2010. A Hybrid Hierarchical Model for Multi-document Summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 815–824, July.

John M. Conroy, Judith D. Schlesinger, Jeff Kubina, Peter A. Rankel, and Dianne P. O'Leary. 2011. CLASSY 2011 at TAC: Guided and Multi-lingual Summaries and Evaluation Metrics. In *Proceedings of the Text Analysis Conference (TAC)*, November.

Gianluca Demartini, Malik Muhammad Saad Missen, Roi Blanco, and Hugo Zaragoza. 2010. Entity Summarization of News Articles. In *Proceedings of the 33rd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 798–796, July.

Pascal Denis and Philippe Muller. 2011. Predicting Globally-Coherent Temporal Structures from Texts via Endpoint Inference and Graph Decomposition. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, July.

Quang Xuan Do, Wei Lu, and Dan Roth. 2012. Joint Inference for Event Timeline Construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*, pages 677–689, July.

Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. 2000. Instruction Manual for the Annotation of Temporal Expressions. Technical report, The MITRE Corporation.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document Summarization by Sentence Extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*, volume 4, pages 40–48, April.

Yihong Gong and Xin Liu. 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *Proceedings of the 24th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 19–25, September.

Iris Hendrickx, Walter Daelemans, Erwin Marsi, and Emiel Krahmer. 2009. Reducing Redundancy in Multi-document Summarization using Lexical Semantic Similarity. In *Proceedings of the Workshop on Language Generation and Summarisation (UCNLG+Sum)*, pages 63–66, August.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL)*, volume 1, pages 71–78, May.

Chin-Yew Lin. 2004. Looking for a Few Good Metrics: ROUGE and its Evaluation. In *Working Notes of the 4th NTCIR Workshop Meeting*, June.

Maofu Liu, Wenjie Li, and Huijun Hu. 2009. Extractive Summarization Based on Event Term Temporal Relation Graph and Critical Chain. In *Information Retrieval Technology*, volume 5839 of *Lecture Notes in Computer Science*, pages 87–99. Springer Berlin Heidelberg.

Jun-Ping Ng and Min-Yen Kan. 2012. Improved Temporal Relation Classification using Dependency Parses and Selective Crowdsourced Annotations. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 2109–2124, December.

Jun-Ping Ng, Praveen Bysani, Ziheng Lin, Min-Yen Kan, and Chew-Lim Tan. 2012. Exploiting Category-Specific Information for Multi-Document Summarization. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 2093–2108, December.

Jun-Ping Ng, Min-Yen Kan, Ziheng Lin, Wei Feng, Bin Chen, Jian Su, and Chew-Lim Tan. 2013. Exploiting Discourse Analysis for Article-Wide Temporal Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12–23, October.

Karolina Owczarzak and Hoa Dang. 2011. Overview of the TAC 2011 Summarization Track: Guided Task and AESOP Task. In *Proceedings of the Text Analysis Conference (TAC)*, November.

Rebecca J. Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigelman. 2005. Applying the Pyramid Method in DUC 2005. In *Proceedings of the Document Understanding Conference Workshop on Text Summarization*, October.

James Pustejovsky, José Castano, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS)*, January.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics*, pages 647–656, March.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2):269–298.

Naushad Uzzaman, Hector Llorens, Leon Derczynski, Marc Verhagen, James F. Allen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TEMPEVAL-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval)*, June.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The TempEval Challenge: Identifying Temporal Relations in Text. *Language Resources and Evaluation*, 43(2):161–179.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 57–62, July.

Xiaojun Wan. 2007. TimedTextRank: Adding the Temporal Dimension to Multi-Document Summarization. In *Proceedings of the 30th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 867–868, July.

Mingli Wu. 2008. *Investigations on Temporal-Oriented Event-Based Extractive Summarization*. Ph.D. thesis, Hong Kong Polytechnic University.

Shasha Xie and Yang Liu. 2008. Using Corpus and Knowledge-based Similarity Measure in Maximum Marginal Relevance for Meeting Summarization. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4985–4988, March.

Renxian Zhang, You Ouyang, and Wenjie Li. 2011. Guided Summarization with Aspect Recognition. In *Proceedings of the Text Analysis Conference (TAC)*, November.