Summarization of Historical Articles Using Temporal Event Clustering

James Gung
Department of Computer Science
University of Colorado at Colorado Springs
Colorado Springs, Colorado 80918

***Abstract*—In this paper, we investigate the use of temporal information for improving extractive summarization of historical articles. Our method timestamps every sentence in an article, then clusters the sentences based on their temporal similarity. Each resulting cluster is assigned an importance score which can then be used as a weight in traditional sentence ranking techniques. We cluster thirteen Wikipedia articles describing historical events. Twelve out of thirteen of the clusterings correctly identify the main events of the articles. Temporal importance weighting offers consistent improvements over baseline systems.**

I. Introduction

As the sheer quantity of available information grows, the ability to rapidly locate the salient points in documents becomes increasingly valuable. Manually filtering through large documents for relevant information is a difficult and time-consuming task. The automatizing of text summarization has therefore become an extremely important area of research.

Extractive summarization involves selecting the k sentences that best summarize a document. Extensive research has gone into determining which features of text documents are useful for calculating the importance of sentences, as well as how to use these features [1]. Little work, however, has considered the importance of temporal information towards single document summarization. This is likely because many text documents have very few explicit time features and do not necessarily describe topics in chronological order, thus making reliable time feature interpolation for each sentence an impractical expectation.

Historical articles, such as Wikipedia articles describing wars, battles, or other major events, possess characteristics that lend themselves towards time interpolation for each sentence. Historical articles tend to contain many explicit time features relative to other kinds of articles. Additionally, historical articles tend to describe events in chronological order. This further increases the reliability of time feature interpolation.

The motivation of our investigation is based on two basic assumptions pertaining to the structure of historical articles. First, historical articles tend to focus on a single central event, despite mentioning numerous other events of lesser importance. The importance of other events can then be judged by their temporal distance from this central event. Second, important events in an article will be described in greater detail, employing more sentences than less important events. We propose a method that exploits these assumptions.

Given an article where every sentence is assigned an explicit timestamp, we cluster the sentences based on their temporal similarity. That is, each cluster should contain sentences describing events that occurred around the same general timespan. We cluster the sentences by creating a hierarchical set of event boundaries using novelty scores as discussed in [2]. Based on the previous two assumptions, the largest cluster pertains to the central event of the article. Each cluster is assigned an importance score based on cluster size, spread, and distance from the cluster describing the central event of the article.

This paper investigates the value of this temporal-based score towards automatic summarization, specifically focusing on historical articles. We investigate whether or not the score can be used as a weight in traditional sentence ranking techniques to improve summarization quality. For testing purposes, we implement TextRank as a baseline system.

II. Related Work

Considerable work has gone into designing and improving extractive summarization techniques. These techniques look at various features of the text, such as word or phrase frequency, sentence position, or sentence to sentence cohesion.

Event-based summarization is a more recent approach to summary generation. Filatova et al. [3] introduced atomic events as a useful extractable feature for extractive summarization. Atomic events are defined as named entities connected by a relation such as a verb or action noun. Events are then selected for summary by applying a maximum coverage algorithm to minimize redundancy while maintaining coverage of the major concepts of the document. Vanderwende et al. [4] identity events as triples (consisting of two nodes and a relation) similarly to [3]. PageRank is then used to determine the relative importance of these triples represented in a graph. Sentence generation techniques are applied towards summarization, achieving results competitive with extractive summarization. We identify events in sentences for temporal extraction, but consider only one time interval per sentence.

Limited work has explored the use of temporal information for summarization. Lim et al. take advantage of the explicit time information given in multi-document summarization (MDS) for sentence extraction and detection of redundant sentence, ordering input documents by time [6]. They base their technique on the observation that important sentences

tend to occur in in time slots containing more documents and time slots occurring at the end and beginning of the documents set. Using traditional methods for extraction of important sentences, they select topic sentences for each time slot, giving higher weights based on the above observation.

Wu et al. use time features towards extractive summarization [7]. They extract events from the text that consist of event elements, the arguments in an event, and event terms, the actions. Each event is then placed on a timeline divided into intervals consistent with the timespan of the article. Each element and event term receives a weight corresponding to the total number of elements and event terms located in each time interval the event element or term occupies. Each sentence is then scored by the total weight of event elements and terms it contains. Encouraging results are reported.

Clustering of events based on time has also received little attention. Cooper et al. investigate clustering towards organizing timestamped digital photographs [5]. They present a method that first calculates the temporal similarity between all pairs of photographs at multiple time scales. These values are stored in a chronologically ordered matrix. Cluster boundaries are determined by calculating novelty scores for each set of similarity matrices. These are then used to form the final clusters. We adopt this clustering method for clustering our timestamped sentences.

III. Approach

A. Overview

The goal of our method is to give each sentence in an article a temporal importance score that can be used as a weight in traditional sentence ranking techniques. To do this, we need to gain an idea of the temporal structure of events in an article. In other words, we want to identify groups of sentences describing events that occurred in the same general timespan. A score must then be assigned to each group corresponding to the importance of the group's timespan to the article as a whole. Each sentence in a particular group will be assigned the same temporal importance score, necessitating the use of a sentence ranking technique to find a complete summary.

B. Temporal Information Extraction

Relatively accurate timestamps for events in an article are needed for this method to be applicable. Timestamp interpolation accuracy depends on the temporal linearity and number of explicit time features in a particular article. Thus, this method's usefulness is dependent on these factors.

For the purposes of this article, we use a temporal expression normalizer to extract the explicit time features. Heideltime is a rule-based system that uses sets of regular expressions to extract time features [8]. Events that occur between each Heideltime-extracted timestamp are naively assigned timestamps consisting of when the prior timestamp ends and the subsequent timestamp begins. This method of temporal extraction is not reliable, but serves the purposes of testing as a reasonable baseline for temporal extraction systems. As the precision increases, the performance of our system should also improve.

C. Temporal Clustering

Clustering is a method for discovering structure in unstructured datasets. To cluster our sentences into temporally-related groups, we adopt a clustering method proposed by Cooper et al. for grouping digital photograph collections based on time.

$$S_K(i,j) = exp\left(-\frac{|t_i - t_j|}{K}\right) \tag{1}$$

Inter-sentence similarity is calculated between every pair of sentences. The similarity measure is based inversely upon the distance between the central time of the sentences (shown in 1). The similarity scores are calculated at varying granularities of time. If the article focuses on a central event that occurs over only a few hours, such as the assassination of John F. Kennedy, the best clustering will generally be found from similarities calculated using a smaller time granularity. Conversely, articles with central events spanning several years, such as the American Civil War, will generally be clustered using similarities calculated at larger time granularities.

The similarities are placed in a matrix and organized chronologically in order of event occurrence time. The resulting matrix is structured such that entries close to the diagonal of the matrix are among the most similar and the actual diagonal entries are maximally similar (diagonal entries correspond to similarities between the same sentences).

To calculate temporal event boundaries, Cooper et al. describe a method for calculating novelty scores [5]. A checkerboard kernel in which diagonal regions contain all positive weights and off-diagonal regions contain all negative weights is correlated along the diagonal of the similarity matrix. The weights of each entry in the kernel are calculated from a Gaussian function such that the most central entries have the highest (or lowest in the off-diagonal regions) values. The result is maximized when the kernel is located on temporal event boundaries. In relatively uniform regions, the positive and negative weights will cancel each other out, resulting in small novelty scores. Where there is a gap in similarity, presumably at an event boundary, off diagonal squares will be dissimilar, thus increasing the novelty score [2]. In calculating novelty scores with each set of similarity scores, we obtain a hierarchical set of boundaries. With each time granularity, we have a potential clustering option.

In order to choose the best clustering, we calculate a confidence score for each boundary set, then choose the clustering with the highest score, as suggested in [5]. This score is the sum of intercluster similarities between adjacent clusters subtracted from the sum of intracluster similarities as seen in Equation 4. A high confidence score then suggests low intercluster similarity and high intracluster similarity.

$$IntraS(B_K)_S = \sum_{l=1}^{|B_k|-1} \sum_{i,j=b_l}^{b_l+1} \frac{S_K(i,j)}{(b_{l+1} - b_l)^2} \tag{2}$$

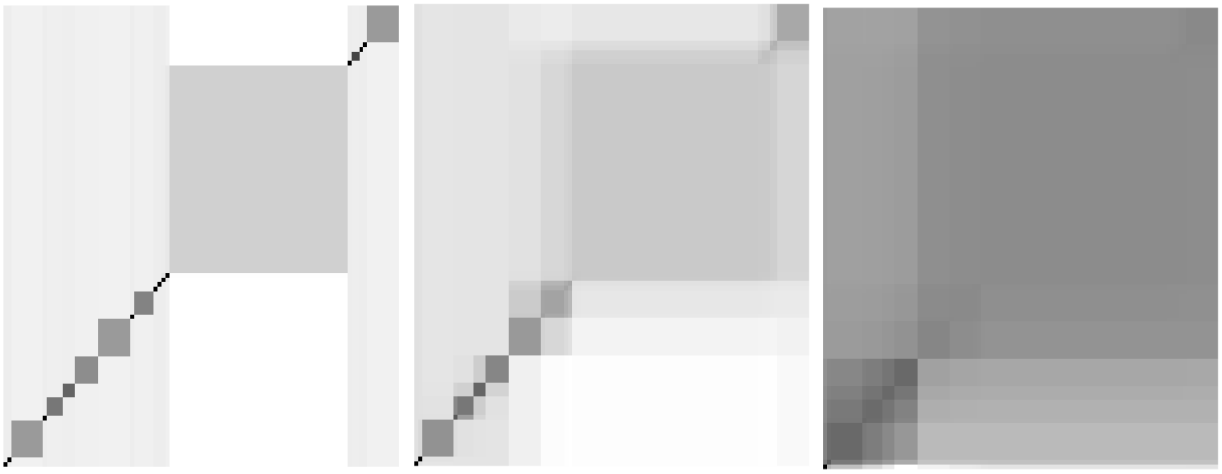

Fig. 1. Similarity matrices at varying *K* displayed as heat maps, darker representing more similar entries. Similarities scores calculated with higher values of *K* correspond to broader time scales (months vs. days). Left to right, *K* is increased by a factor of 10 at each iteration.

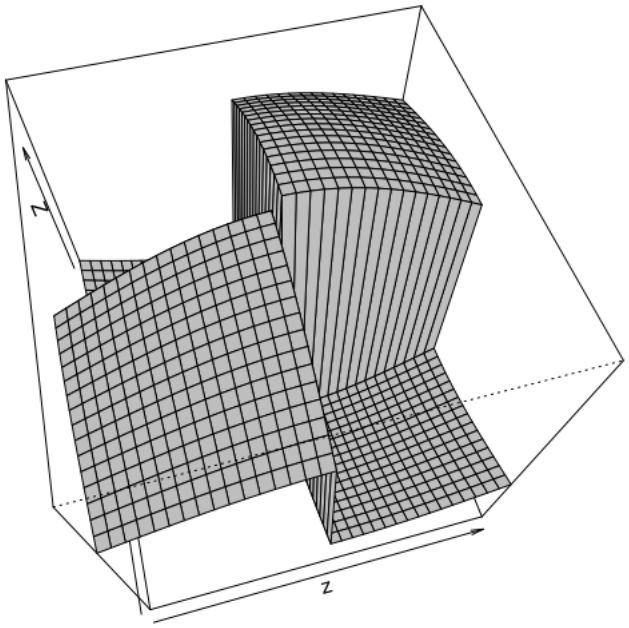

Fig. 2. A gaussian-tapered kernel used to calculate novelty scores. This is slid along the diagonal of each similarity matrix, calculating a novelty score for each sentence. Positive diagonal regions correlate with high intracluster similarity entries. Negative off-diagonal regions multiply by the low intercluster similarity entries, resulting in higher total novelty scores at temporal event boundaries.

$$InterS(B_K)_S = \sum_{l=1}^{|B_k|-2} \sum_{i=b_l}^{b_l+1} \sum_{j=b_l+1}^{b_l+2} \frac{S_K(i,j)}{(b_{l+1}-b_l)(b_{l+2}-b_{l+1})} \tag{3}$$

$$Confidence_S(B_K) = IntraS(B_K)_S - InterSB_K)_S \tag{4}$$

D. Estimating Clustering Paramaters

There are several parameters we must consider before clustering the sentences. Historical articles describing wars will generally have much larger timespans than articles describing battles. Looking at battles at a broad time granularity applicable to wars may not produce a meaningful clustering. Thus, it is worthwhile for us to estimate the temporal structure of each article before clustering. The time granularity for each clustering is controlled by the *K* parameter in the similarity function between sentences. To find multiple clusterings, we

start at a base *K*, then increment *K* by a multiplier for each new clustering. We calculate the base *K* using the standard deviation for event times in the article. Measuring the spread of events in the article gives us an estimate of what time scale we should use for measuring similarity.

E. Calculating Temporal Importance

We use three metrics to calculate the importance of a cluster towards a summary. The first metric is based on the size of the cluster (5). This is partially motivated by the assumption that more important events will be described in greater detail, thus producing larger clusters. The second metric (6) is based on the distance from the cluster's centroid to the centroid of the largest cluster, corresponding to the central event of the article. This metric is motivated by the assumption that historical articles have a central event which is described in the greatest detail. The third metric is based on the spread of the cluster (7). Clusters with large spreads are unlikely to pertain to the same event, and should therefore be penalized.

$$Size(C_i) = \frac{|C_i|}{|C_{max}|} \tag{5}$$

$$Sim(C_i) = exp\left(-\frac{|t_{C_iCentroid} - t_{MaxClusterCentroid}|}{m}\right) \tag{6}$$

$$Spread(C_i) = exp\left(-\frac{\sigma_{C_i}}{n * (t_{max} - t_{min})}\right) \tag{7}$$

The parameters *m* and *n* serve to weight the importance of these measures and are assigned based on the spread of events in an article.

The three measures are weighted and multiplied together to obtain a final importance score, working in tandem to ensure that the importance measure will still be valid even if the largest cluster does not correspond to the central event of the article. If the central event is broken up into multiple clusters, they will likely be located nearby each other. If the largest cluster does not correspond to the central event, following our assumption that the central event is described in the greatest detail, it will likely consist of many spread out smaller events, resulting in a greater spread and a decreased importance score. Conversely, clusters corresponding to the central event will be more concentrated relative to the length of the article, and be rated as more important.

F. Final Sentence Ranking

Each sentence is assigned a temporal importance score equal to the importance score of the cluster to which it belongs. To find a complete ranking of the sentences, we need to apply a traditional sentence ranking technique. Any automatic summarization technique that ranks its sentences with specific numerical scores can potentially be augmented with our temporal importance weight.

$$WS(V_i) = (1-d)+d*\sum_{V_j \in In(V_i)} \frac{w_{j,i}}{\sum_{v_k \in Out(V_j)} w_{j,k}} WS(V_j) \tag{8}$$

Existing graph-based methods for sentence ranking apply Google's PageRank algorithm to rank sentence importance [9]. Sentences are represented as nodes on a graph. Some measure of similarity between each of the sentences is calculated, such as cosine similarity or the number of shared open-class words between each pair of sentences [10]. Edges are placed between sentence pairs with similarities above a particular threshold. After giving each node an arbitrary score, the algorithm iteratively calculates the scores of each node based on the scores of neighboring nodes. This score is weighted by the ratio of the weight of the edge between a neighboring node and the current node divided by the total weight of all edges leaving the neighboring node. These weighted scores are then summed to determine the score of the current node (Equation 8). We set a damping factor *d* to 0.85, which in the context of PageRank is used to model the probability of randomly jumping to another page. The scores of each node after convergence indicate the relative importance of each sentence.

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)} \tag{9}$$

We choose to use TextRank in our experiments. Our similarity measure is calculated using the number of shared named entities and nouns between sentences as seen in Equation 9. For identification of named entities, we use Stanford NER [11]. It is straightforward to weight the resulting TextRank scores for each sentence using their cluster's temporal importance.

IV. Experimental Results

Evaluation of summaries is traditionally accomplished using ROUGE, Recall-Oriented Understudy for Gisting Evaluation [12]. ROUGE automatically determines the quality of summaries by comparing them to human-created ideal summaries based on measures such as number of overlapping n-grams, word sequences or word pairs. To apply ROUGE, we use human-annotated summaries of the articles we wish to evaluate. These were obtained by asking volunteers to choose what they consider to be the most important sentences from each article.

Using these human-annotated summaries as gold standards, we compare the performance of sentence ranking systems with and without temporal weighting. ROUGE-N computes the number of co-occurring N-grams in the system summary and set of gold standard reference summaries. This is divided by the total number of N-grams in the set of reference summaries. Thus this is a recall-oriented measure. We evaluate using ROUGE-2 bigram matching.

A. Clustering

The Wikipedia articles we test each contain a topic sentence stating the timespan of the main event described by the article. This provides an easy way to determine whether or not a clustering is successful. If the largest cluster contains the timespan of the main event described by the topic sentence, we consider the clustering to be successful (so long as the clustering isn't trivial). The articles vary greatly in length. Also, the ratio of sentences with time features to sentences without is considerably varied. We would expect the temporal-weighted summaries of articles with larger ratios to have more reliable clusterings than those of articles with smaller ratios as they contain more temporal information to interpolate from.

Out of thirteen articles, twelve were successful clusterings by the above criterion. Only Nickel Grass was clustered poorly, the clustering algorithm dividing the main event into two clusters. It is of interest to note that Nickel Grass had one of lowest ratios of sentences containing time features to sentences without, which possibly explains the poor clustering.

B. Temporal Importance Weighting

We test our TextRank implementation on thirteen Wikipedia articles, with and without temporal importance weighting. The articles vary widely in length and ratio of sentences containing time features to sentences without. Each article has at least two human-annotated gold standard summaries for use in ROUGE.

We observe consistent improvements for the articles using the TextRank system with temporal importance weighting over the bare TextRank implementation. In general, articles containing sentences that TextRank ranked highly, but that contain sentences occurring at significantly different times than the central events of the articles observe significant improvements. Although the content of these sentences is highly related to the rest of the article, they should not be included in the summary since the events they contain occur nowhere near the main event temporally.

In 1904, the United Spanish War Veterans was created from smaller groups of the veterans of the Spanish American War.

Fig. 3. An initially highly ranked sentence excluded from the final summary due to low temporal importance

TABLE I
RESULTS OF CLUSTERING ON *the Battle of Fredericksburg*, ONLY EXPLICIT TIME FEATURES.

Centroid: 09/01/1862	
09/01/1862	Although McClellan had stopped Lee at the Battle of Antietam in September, President Abraham Lincoln believed...
Centroid: 11/14/1862	
11/01/1862	Burnside was appointed commander of the Army of the Potomac in November, replacing Maj. Gen. George...
11/09/1862	Burnside, in response to prodding from Lincoln and General-in-Chief Maj. Gen. Henry W. Halleck, planned...
11/16/1862	The Union Army began marching on November 15, and the first elements arrived in Falmouth on November 17.
11/21/1862	By November 21, Lt. Gen. James Longstreet's Corps had arrived near Fredericksburg, and Jackson's was....
11/25/1862	The first pontoon bridges arrived at Falmouth on November 25, much too late to enable the Army of the Potomac...
Centroid: 12/13/1862	
12/13/1862	The Battle of Fredericksburg, fought in and around Fredericksburg, Virginia, from December 11 to December 15...
12/13/1862	The Union Army suffered terrible casualties in futile frontal assaults on December 13 against entrenched Confederate...
12/09/1862	On December 9, he wrote to Halleck, "I think now the enemy will be more surprised by a crossing immediately...
12/11/1862	Union engineers began to assemble six pontoon bridges on the morning of December 11, two just north of the town...
12/11/1862	Eventually his subordinates convinced Burnside to send landing parties over in the boats that evening to secure...
12/12/1862	Over the course of December 11 to December 12, Burnside's men deployed outside the city and prepared to attack...
12/13/1862	The battle opened south of the city at 8:30 a.m. on December 13, when Franklin sent two divisions from the Left...
12/13/1862	By 10 a.m., a thick fog began to lift, and the initially sluggish movements picked up speed.
12/13/1862	The initial assaults west of Fredericksburg began at 11 a.m. as French's division moved along the Plank Road...
12/13/1862	Griffin's division renewed the attack at 3:30 p.m., followed by Humphrey's division at 4 p.m.
12/13/1862	At dusk, Getty's division assaulted from the east and was also repulsed.
12/13/1862	Thousands of Union soldiers spent the cold December night on the fields leading to the Heights, unable to move...
12/14/1862	The armies remained in position throughout the day on December 14, when Burnside briefly considered leading...
12/14/1862	That afternoon, Burnside asked Lee for a truce to attend to his wounded, which Lee graciously granted.
12/15/1862	The next day the Federal forces retreated across the river, and the campaign came to an end.
12/13/1862	Stationed at the stone wall by the sunken road below Marye's Heights, Kirkland had a close up view to the suffering...
12/13/1862	The Cincinnati "Commercial" wrote, "It can hardly be in human nature for men to show more valor...

Similarly to the TextRank system, our random ranking system observes small improvements when augmented with temporal importance weighting. The results, however, are more mixed. It is likely that additional human-annotated summaries are necessary for conclusive results. The gold standard summaries widely vary in length and content, displaying the inherent subjectivity and difficulty involved in evaluation.

V. CONCLUSIONS AND FUTURE WORK

The novelty-based clustering method worked extremely well for our purposes. Out of thirteen articles, twelve were clustered such that the temporal bounds of the main events composed or belonged to the largest clusters. These results can likely be improved upon using more advanced temporal extraction and interpolation methods, as our baseline method used a very naive heuristic for interpolating between time features prone to error.

The temporal importance weighting had mixed results with both TextRank and random ranking. The average ROUGE score between all articles was modestly increased, but not significantly. Several single articles showed significant improvement when the ranked sentences were weighted with temporal importance measures. However, improvement was not uniform across all thirteen articles. We attribute decreases in ROUGE scores to poor clusterings. This demonstrates the importance to this method of finding good clusterings, and consequently correctly extracting and interpolating temporal

TABLE II
ROUGE SCORES FOR AN IMPLEMENTATION OF TEXTRANK WITH AND WITHOUT TEMPORAL WEIGHTING.

ROUGE-2		
System	TextRank Weighted	TextRank
Chancellorsville	**0.26495**	0.26305
Chickamauga	**0.23206**	0.22856
Coral Sea	**0.34436**	0.29591
First Barbary	**0.17499**	0.14087
Fredericksburg	**0.12713**	0.05555
Gulf War	**0.33408**	0.32225
Hampton Roads	0.21486	0.21486
Korean War	**0.26084**	0.23666
Nickel Grass	**0.38962**	0.33268
Spanish American	**0.32889**	0.32373
Vicksburg	**0.25000**	0.23118
War of 1812	**0.20970**	0.20960
Whiskey Rebellion	0.21573	0.21573

TABLE III
ROUGE SCORES FOR RANDOMLY SCORED SUMMARIES WITH AND WITHOUT TEMPORAL WEIGHTING.

ROUGE-2		
System	Random Weighted	Random
Chancellorsville	**0.25159**	0.23051
Chickamauga	0.13787	0.16213
Coral Sea	0.17349	0.25397
First Barbary	**0.16882**	0.12637
Fredericksburg	**0.10565**	0.09929
Gulf War	**0.20082**	0.16227
Hampton Roads	**0.29714**	0.19302
Korean War	**0.23441**	0.21803
Nickel Grass	0.14003	0.14003
Spanish American	**0.27194**	0.23872
Vicksburg	0.15585	0.19626
War of 1812	**0.28849**	0.27814
Whiskey Rebellion	0.10308	0.14556

information. Further testing and additional human-annotated summaries are necessary for conclusive results with regard to temporal importance weighting.

It may also be fairly easy to predict the success of using this temporal weight a priori to summarization of an article. A small ratio of explicit time features to sentences (less than 0.15) indicates that the temporal interpolation process may not be very accurate. Many other measures can be considered. The linearity of time features is also a good indication of the success of temporal extraction. More chronological event description will reduce the risk of errors in temporal interpolation. The spread of time features in an article is also a clue to the success of our weighting method. A greater spread indicates that more events will occur farther from the main event of the article necessitating the use of our weighting scheme to filter out unimportant sentences from the summary. Prediction of temporal weighting success would allow for the potential of improving summarization without a great risk of reducing the quality of the summaries by assigning incorrect importance weights.

We have naively assigned a time interval to each sentence. Individual events within sentences are not considered separately. Future work might individually extract events from each sentence, assigning time intervals to each event. For summarization purposes, the most representative event should be chosen for clustering.

Given the success of clustering major temporal events in historical articles, many directions in future work can be taken. It would be useful to augment timeline generation techniques using the hierarchical set of temporal event boundaries produced by the clustering algorithm. Timelines might be constructed at multiple scales, selecting important events representative of each cluster to display at each granularity, allowing the user to progressively zoom in on temporal regions and be provided with more detailed information representative of the region.

Additional sentence importance measures might be explored using the temporal clusterings. Summarization in the vein of maximum coverage, based upon maximally covering topics using a minimal number of sentences, might be explored using temporal boundaries to designate topics.

An alternative approach to clustering the sentences would be to incorporate a content-based similarity measure in the distance measure for the clustering algorithm. This additional dimension would allow for identification of major events that occurred simultaneously but in different clusters. Such an approach would be useful in articles describing events that occurred in parallel.

We presented a method for weighting sentences based on their temporal importance. Sentences are clustered based on the times of events occurring within them. The largest cluster is designated the major event of the article, and other clusters are scored based upon their distance from this cluster, their size, and their spread. Sentences are weighted based on the score of the cluster to which they belong. This weight is used to augment a traditional sentence ranking method, TextRank. We test summarization systems with and without this temporal importance weight, observing modest improvements.

ACKNOWLEDGEMENTS

The research reported in this document has been funded partially by NSF grants CNS-0958576 and CNS-0851783.

REFERENCES

[1] V. Gupta and G. S. Lehal. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3):258–266, August 2010.

[2] J. Foote and M. Cooper. Media segmentation using self-similarity decomposition. *Proceedings of SPIE*, 5021:167–175, 2003.

[3] E. Filatova and V. Hatzivassiloglou. Event-based extractive summarization. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 104–111, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[4] L. Vanderwende. Event-centric summary generation. In *DUC 2004*, 2004.

[5] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digitial photo collections. *ACM Transactions on Multimedia Computing, Communications and Applications*, 1(3):269–288, August 2005.

[6] J.-M. Lim, I.-S. Kang, J.-H. J. Bae, and J.-H. Lee. Sentence extraction using time features in multi-document summarization. *Information Retrievel Technology*, pages 82–93, 2005.

[7] M. Wu, W. Li, Q. Lu, and K.-F. Wong. Event-based summarization using time features. In *CICLing 2007*, pages 563–574, 2007.

[8] J. Strötgen and M. Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 321–324, Uppsala, Sweden, July 2010.

[9] D. R. Radev and G. Erkan. Lexrank: graph-based centrality as salience in text summarization. In *Journal of Artificial Intelligence Research*, volume 22, pages 457–479, 2004.

[10] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *Proceedings of EMNLP*, volume 4, pages 404–411. Barcelona: ACL, 2004.

[11] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pages 363–370, 2005.

[12] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Workshop On Text Summarization Branches Out*, 2004.