Summarization of Historical Articles Using Temporal Event Clustering

James Gung
Department of Computer Science
Miami University
Oxford, Ohio 45056
gungjm@muohio.edu

Jugal Kalita
Department of Computer Science
University of Colorado
Colorado Springs CO 80920
jkalita@uccs.edu

Abstract

In this paper, we investigate the use of temporal information for improving extractive summarization of historical articles. Our method clusters sentences based on their timestamps and temporal similarity. Each resulting cluster is assigned an importance score which can then be used as a weight in traditional sentence ranking techniques. Temporal importance weighting offers consistent improvements over baseline systems.

1 Introduction

Extensive research has gone into determining which features of text documents are useful for calculating the importance of sentences for extractive summarization, as well as how to use these features (Gupta and Lehal, 2010). 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.

Historical articles, such as Wikipedia articles describing wars, battles, or other major events, tend to contain many explicit time features. Historical articles also tend to describe events in chronological order. In addition, historical articles tend to focus on a single central event. The importance of other events can then be judged by their temporal distance from this central event. Finally, important events in an article will be described in greater detail, employing more sentences than less important events.

This paper investigates the value of a temporal-based score towards automatic summarization, specifically focusing on historical articles. We investigate whether or not such a score can be used as a weight in traditional sentence ranking techniques to improve summarization quality.

2 Related Work

Event-based summarization is a recent approach to summary generation. (Filatova and Hatzivassiloglou, 2004) introduced atomic events, which are named entities connected by a relation such as a verb or action noun. Events are 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., 2004) identify events as triples consisting of two nodes and a relation. PageRank is then used to determine the relative importance of these triples represented in a graph. Sentence generation techniques are applied towards summarization.

Limited work has explored the use of temporal information for summarization. (Lim et al., 2005) use the explicit time information in the context of multi-document summarization for sentence extraction and detection of redundant sentences, ordering input documents by time. They observe 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. They select topic sentences for each time slot, giving higher weights based on the above observation.

(Wu et al., 2007) extract event elements, the arguments in an event, and event terms, the actions. Each event is 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 scored by the total weight of event elements and terms it contains.

Clustering of events based on time has also received little attention. (Foote and Cooper, 2003) investigate clustering towards organizing timestamped digital photographs. They present a method that first

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 631–635,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

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 used to form the final clusters. We adopt this clustering method for clustering timestamped sentences.

3 Approach

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. 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.

3.1 Temporal Information Extraction

We use Heideltime, a rule-based system that uses sets of regular expressions, to extract explicit time expressions in the article and normalize them (Strötgen and Gertz, 2010). Events that occur between each Heideltime-extracted timestamp are assigned timestamps consisting of when the prior timestamp ends and the subsequent timestamp begins. The approach is naive and is described in (Chasin et al., 2011). 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.

3.2 Temporal Clustering

To cluster sentences into temporally-related groups, we adopt a clustering method proposed by Foote et al. to group digital photograph collections.

Inter-sentence similarity is calculated between every pair of sentences using Equation (1).

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

The similarity measure is based inversely on the distance between the central time of the sentences. Similarity scores are calculated at varying granularities. 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 be clustered using similarities calculated at larger time granularities.

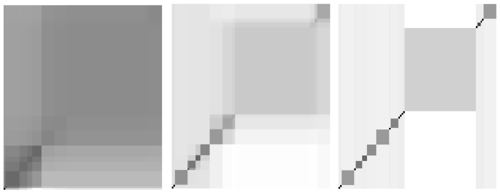

Figure 1: Similarity matrices at varying k displayed as heat maps, darker representing more similar entries

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

To identify temporal event boundaries, (Foote and Cooper, 2003) calculate novelty scores. 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 cancel each other out, resulting in small novelty scores. Where there is a gap in similarity, presumably at an event boundary, off diagonal squares are dissimilar, increasing the novelty score. 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 C for each boundary set, then choose the clustering with the highest score, as suggested in (Foote and Cooper, 2003). This score is the sum of intercluster similarities ($IntraS$) between adjacent clusters subtracted from the sum of intracluster ($InterS$) similarities as seen in Equation (4). A high confidence score suggests low intercluster similarity and high intracluster similarity.

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

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

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

3.3 Estimating Clustering Paramaters

Historical articles describing wars 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, we should 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.

3.4 Calculating Temporal Importance

We use three novel metrics to calculate the importance of a cluster towards a summary. The first metric is based on the size of the cluster (Eqn 5). This is motivated by the assumption that more important events will be described in greater detail, thus producing larger clusters. The second metric (Eqn 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 (Eqn 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. For *n*, we used the standard deviation of event times in the article. For *m*, we used the cluster similarity score from Equation (4). The three measures work in tandem to ensure that the importance measure will be valid even if the largest cluster does not correspond to the central event of the article.

3.5 Final Sentence Ranking

Each sentence is assigned a temporal importance weight equal to the importance score of the cluster to which it belongs. To find a complete ranking of the sentences, we apply a sentence ranking technique. Any automatic summarization technique that ranks its sentences with numerical scores can potentially be augmented with our temporal importance weight. We multiply the base scores from the ranking by the associated temporal importance weights for each sentence to find the final ranking.

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

Like several graph-based methods for sentence ranking for summarization (e.g., (Erkan and Radev, 2004)), we use Google's PageRank algorithm (Equation 8) with a damping factor d of 0.85.

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

We use TextRank (Mihalcea and Tarau, 2004) 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 (Finkel et al., 2005). It is straightforward to weight the resulting TextRank scores for each sentence using their cluster's temporal importance.

4 Experimental Results

We test on a set of 13 Wikipedia articles describing historical battles. The average article length is 189 sentences and 4,367 words. The longest article is 545 sentences and contains 11,563 words. The shortest article is 51 sentences and contains

1,476 words. Each article has at least two human-annotated gold standard summaries. Volunteers were asked to choose the most important sentences from each article. We evaluate using ROUGE-2 bigram matching (Lin, 2004).

4.1 Clustering

Each Wikipedia article contains a topic sentence stating the timespan of the main event in the article. This provides an easy way to determine whether 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. The articles vary greatly in length. Also, the ratio of sentences with time features to sentences without is considerably varied. In 92% of the articles, there were successful clusterings. An example of an article that didn't cluster is Nickel Grass, where the main event was divided into two clusters. It is of interest to note that this article had one of lowest time feature to sentence ratios, which possibly explains the poor clustering.

4.2 Temporal Importance Weighting

We test our TextRank implementation with and without temporal importance weighting.

We observe improvements in general using the TextRank system with temporal importance weighting. The ROUGE-2 score increased by 15.72% across all the articles. The lowest increase was 0% and the highest was 128.86%. The average ROUGE-2 scores were 0.2575 weighted and 0.2362 unweighted, a statistically significant increase with a 95% confidence interval of 0.0066 to 0.0360.

In particular, we see significant improvements in articles that contain sentences TextRank ranked highly but have events occurring at significantly different times than the central event of the article. Although the content of these sentences is highly related to the rest of the article, they should not be included in the summary since their events happen nowhere near the main event temporally.

Our random ranking system, which randomly assigns base importance scores to each sentence, observed only small improvements, of 4.27% on average, when augmented with temporal importance weighting. It is likely that additional human-annotated summaries are necessary for conclusive results.

5 Conclusions and Future Work

The novelty-based clustering method worked extremely well for our purposes. These results can likely be improved upon using more advanced temporal extraction and interpolation methods, since we used a naive method for interpolating between time features prone to error. The temporal importance weighting worked very well with TextRank and reasonably well with random ranking.

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. The linearity of time features is also a good indication of the success of temporal extraction. Finally, the spread of time features in an article is a clue to the success of our weighting method.

Acknowledgements

Research reported here has been funded partially by NSF grants CNS-0958576 and CNS-0851783.

References

R. Chasin, D. Woodward, and J. Kalita, 2011. *Machine Intelligence: Recent Advances*, chapter Extracting and Displaying Temporal Entities from Historical Articles. Narosa Publishing, Delhi.

G. Erkan and D.R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

E. Filatova and V. Hatzivassiloglou. 2004. Event-based extractive summarization. In *ACL Workshop on Summarization*.

J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, pages 363–370.

J. Foote and M. Cooper. 2003. Media segmentation using self-similarity decomposition. In *SPIE*, volume 5021, pages 167–175.

V. Gupta and G.S. Lehal. 2010. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3):258–268.

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

C.Y. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Workshop on text summarization*, pages 25–26.

R. Mihalcea and P. Tarau. 2004. Textrank: Bringing order into texts. In *EMNLP*, pages 404–411.

J. Strötgen and M. Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *5th International Workshop on Semantic Evaluation*, pages 321–324.

L. Vanderwende, M. Banko, and A. Menezes. 2004. Event-centric summary generation. *Working notes of DUC*.

M. Wu, W. Li, Q. Lu, and K.F. Wong. 2007. Event-based summarization using time features. *Computational Linguistics and Intelligent Text Processing*, pages 563–574.