DBSJ Journal
Vol. 18, Article No. 2, March 2020

Comparative Summarization of Temporal Document Collections

Yijun Duan[1] Adam Jatowt[2]
Masatoshi Yoshikawa[3]

A significant fraction of documents these days are accounts of historical events. Biographies or descriptions of entity histories such as histories of geographical places or organizations are examples of such timeline documents. Their content is explicitly or implicitly associated with timestamps indicating occurrence time of described past events. However, the collections of such timeline documents can be quite large and can pose challenge for readers trying to make sense of them. We introduce in this paper a novel research task, *Comparative Timeline Summarization* (CTS), as an effective strategy to discover important similarities and differences in collections of timeline documents for providing contrastive type of knowledge. We then propose a three-step approach which relies on a dynamic affinity-preserving mutually reinforced random walk for the CTS task and evaluate it on Wikipedia categories. The summaries we generate are in the form of timelines consisting of salient and discriminative sentences ordered chronologically.

1 Introduction

Multi-Document Summarization (MDS) plays an important role in combating the problem of information overload caused by the exponential growth of documents these days, especially, ones posted on the Web. Quite many documents however contain chronologically ordered content describing histories of entities or detailed accounts of past events. Biographies and history sections of Wikipedia articles are two examples of such documents in which paragraphs or sentences are associated with time indicators[*1] denoting the chronological order of discussed events. We call such documents, timeline documents, or, in short, timelines. Traditional MDS techniques are not suitable for summarizing timeline documents, due to their too general assumptions on sentence importance, as shown in [2].

In the recent years, Multi-Timeline-Summarization (MTS) [2, 6] has been introduced to generate the common traits and to capture the common temporal patterns within a set of timeline documents. Unlike MDS, MTS approaches make explicit use of the temporal character of input timelines.

However, sometimes, users would like to compare two collections of timeline documents in order to discover commonalities and differences between them. For example, they may be interested in the comparison of histories of Chinese cities with the histories of Japanese cities, or they would like to know how different were the lives of French scientists in the 19^{th} century from those of American scientists at the same century. Note that such contrasting knowledge is difficult to be manually obtained by users due to relatively large number of documents that would have to be analyzed. Another reason is the difficulty in judging the significance of encountered information by average users.

Comparative timeline summarization (CTS) is then a solution that we propose to provide a condensed and informative timeline summary consisting of major contrasting events chronologically ordered for better understanding of the compared sets of timeline documents. Users could benefit from this kind of information for a myraid of needs such as gaining insights from history, analyzing trends, making better decisions, as well as for educational or entertaining purposes. For example, for a student who is interested in the comparison between histories of Chinese and Japanese cities, our study could greatly facilitate her understanding of these two sets.

Straightforward approach to this problem would be to formulate it as a comparative summarization task [7–11]. However, existing comparative summarization models are not suitable for our scenario, due to the strong temporal character of the input documents in CTS task. Considering the unique characteristic of our task, we have to deal with two problems: (1) First, how to rank important information along the timeline?; (2) Second, how to generate a proper summary at various time scopes, within the time range spanned by the input documents?

To provide effective summarization method we use the following hypotheses: (1) Timeline documents in the same set can be divided into latent eras, where sentences within the same era tend to be more similar to each other than sentences contained in other eras; (2) A sentence is locally important if it is similar to many other sentences within the same era, and is globally important if it frequently appears in different eras of the same document set; (3)An important sentence is semantically dissimilar to the content of the contrasting document set.

To reflect the above hypotheses we rely on a three-step approach. We first discover latent eras of each compared document set by adopting the widely-used linear text segmentation algorithm [16]. We then propose an affinity-preserving mutually reinforced random walk (APMRRW) method to appropriately model the relationship among sentences belonging to different document sets for selecting locally important sentences. Finally, we develop a dynamic ranking model to extract globally important sentences. The dynamic ranking model is also able to flexibly generate a proper summary at various time scopes along the timeline.

To sum up, we make the following contributions in this paper: (1) We introduce a new research task of *comparative timeline summarization.* (2) We propose a three-stage approach including a novel dynamic affinity-preserving mutually reinforced random walk model (D-APMRRW) for solving the above task. The proposed model is flexible over different granularities of time units. (3) The effectiveness of our approach is demonstrated by experiments on diverse Wikipedia categories.

2 Related Work

2.1 Comparative Summarization

Comparative summarization requires providing short summaries from multiple comparative aspects. Wang *et al.* [7] propose a discriminative sentence selection method based on a multivariate normal generative model aiming to extract sentences that are best describing the unique characteristics of each document group. Huang *et al.* [11] formulate the task of comparative news summarization as an optimization problem of selecting sentences to maximize the score of comparative and representative evidences based on an integer linear programming (ILP) model. Ren *et al.* [8] explicitly consider contrast, relevance and diversity for summarizing contrastive themes by adopting a hierarchical nonparametric Bayesian model to

[1] Non member Graduate School of Informatics, Kyoto University
yijun@db.soc.i.kyoto-u.ac.jp
[2] Regular member Graduate School of Informatics, Kyoto University
adam@dl.kuis.kyoto-u.ac.jp
[3] Regular member Graduate School of Informatics, Kyoto University
yoshikawa@i.kyoto-u.ac.jp

[*1] Typically, temporal expressions serve as such indicators. They can be however implicit when time is obvious from context.

infer hierarchical relations among topics for enhancing the diversity of themes. Gillenwater *et al.* [9] develop a probabilistic approach to address the problem of finding diverse and salient threads in graphs of large document collections. Recently, differential topic models have also been explored to measure sentence discriminative capability for comparative summarization [10].

2.2 Timeline Summarization

Timeline Summarization defined as the summarization of sequences of documents (typically, news articles about the same event) has been actively studied in the recent years. In [4], Yan *et al.* propose the evolutionary timeline summarization (ETS) to compute evolution timelines consisting of a series of time-stamped summaries. David *et al.* [6] present a method for discovering biographical structure based on a probabilistic latent variable model. His approach summarizes timestamped biographies to a set of event classes along with the typical times when those events occur. Satoko *et al.* [3] present a graph-based algorithm for online summarization of time-series documents. In that work, the authors iteratively calculate sentence importance using random walks and pass important sentences to the next calculation. Abdalghani *et al.* [5] address the problem of identifying important events in the past, present, and future from semantically-annotated large-scale document collections. Tuan *et al.* [1] present a novel approach for timeline summarization of high-impact events, which uses entities instead of sentences for summarizing the events. Recently, Duan *et al.* [2] have introduced a novel type of summarization task consisting in generating gists of histories of multiple entities.

To the best of our knowledge, we are the first to work on comparative summarization of timeline documents. Unlike general comparative summarization tasks, we focus on the temporal characteristics of input documents. Finally, different from timeline summarization tasks, we aim to discover discriminative and contrasting information for comparing sets of timeline documents.

3 Problem Definition

We begin by defining the notion of *timeline document*. A *timeline document* spans over a certain range of time and its sentences are assumed to refer to an event within the time range. Each event is associated with a date, which can be either explicitly mentioned in the sentence or could be estimated based on context (e.g., nearby sentences).

Formally, let D_A and D_B denote two sets of timeline documents to be compared. The purpose of comparative timeline summarization task is to form two timeline documents $S_A = \{s_1^A, s_2^A, ..., s_m^A\}$ and $S_B = \{s_1^B, s_2^B, ..., s_m^B\}$ consisting of meaningful sentences summarizing the representative and contrasting information in input document sets, where s_i^A and s_i^B are sentences extracted from D_A and D_B respectively, and m denotes the summary size.

4 Method

In order to compute sentence importance we state the following assumptions:

Assumption 1 *Most timeline documents can be divided into eras. An important sentence is then semantically similar to the era it belongs to.*

Assumption 2 *An important sentence is similar to sentences in the same set while dissimilar to sentences in the contrasting set. Especially, it is more important if it is similar to important events in the same set while being dissimilar to important events in the contrast set.*

Assumption 3 *An important sentence can appear in many eras rather than appear in one era.*

Based on the above hypotheses, we first focus in Sec. 4.1 on detecting latent eras of input documents embodying the idea of **Assumption 1**. We then propose in Sec. 4.2 an affinity-preserving random walk model to locally score sentences following **Assumption 2**. Finally, motivated by **Assumption 3**, we present in Sec. 4.3 the idea of a dynamic ranking algorithm which flexibly allows for generating globally important summary.

4.1 Eras Detection

We first formulate the problem of eras detection as follows. Given a document set D and a sequence of atomic time units $\xi = (t_1, t_2, ..., t_n)$, the task is to select a proper segmentation Θ containing k eras that divide the entire time span $[t_1, t_n]$, where each era T_i is expressed by two time points representing its beginning date τ_b^i and the ending date τ_e^i. Formally, let $\Theta = (T_1, T_2, ..., T_k)$, where $T_i = [\tau_b^i, \tau_e^i]$. To detect eras, we refer to the C99 algorithm [16], which is a linear text segmentation algorithm achieving promising results. The algorithm assumes that sentences within the same era tend to be more similar to each other compared to sentences contained in other eras. It takes as an input the pairwise cosine similarities between all atomic time unit pairs in order to generate a similarity matrix M, with entries $M_{ij} = Sim_{cosine}(t_i, t_j)$ (see Fig. 1 for an example). Since only the relative values are meaningful, C99 transfers each value of M into the fraction of its neighbors with smaller value, where the neighborhood is an $r \times r$ block:

$$M_{ij} = \sum_{l \in [i-r/2, i+r/2]} \sum_{k \in [j-r/2, j+r/2]} [M_{ij} > M_{lk}] \tag{1}$$

where the value of the expression in square brackets equals to 1 if the inequality holds, otherwise 0. Each time unit is now represented by a transformed vector of its cosine similarity with each other unit. The goal of the above matrix transformation is to enhance the contrast between different eras.

Given $\Theta = (T_1, T_2, ..., T_k)$ as a list of k coherent eras, let D_Θ denote the inside density of Θ as follows:

$$D_\Theta = \frac{\sum_{l=1}^{k} s_l}{\sum_{l=1}^{k} a_l} \tag{2}$$

where $s_l = \sum_{i \in [\tau_b^l, \tau_e^l]} \sum_{j \in [\tau_b^l, \tau_e^l]} M_{ij}$ and $a_l = (\tau_e^l - \tau_b^l)^2$ refer to the sum of all transformed cosine similarities in a era, and the squared length of the era, respectively. Then the final process is to determine the location of the era boundaries. To initialize the segmentation process, the entire time span $[t_1, t_n]$ is placed as one coherent era. Then the C99 algorithm builds up a segmentation into k eras by greedily inserting a new boundary at each step which maximizes D_Θ. Given the discovered latent eras Θ, we compute the preliminary importance score of a sentence v as its average similarity with events in the same era $T(v)$ which v belongs to as follows:

$$Imp(v, T_v) = \frac{1}{|T_v|} \cdot \sum_{e \in T_v} Sim_{cosine}(v, e) \tag{3}$$

The scores of sentences will be used as their initial scores in the APMRRW model introduced in the next step for selecting locally important sentences.

4.2 Affinity-Preserving Mutually Reinforced Random Walk

In this section we introduce our proposed APMRRW model for locally scoring sentences, which are assumed to be similar to sentences in the same set while dissimilar to sentences in the contrasting set, within the same time unit. We first briefly explain the Affinity-Preserving Random Walk model, which recently has been reported to achieve the state-of-the-art performance amont the graph-based

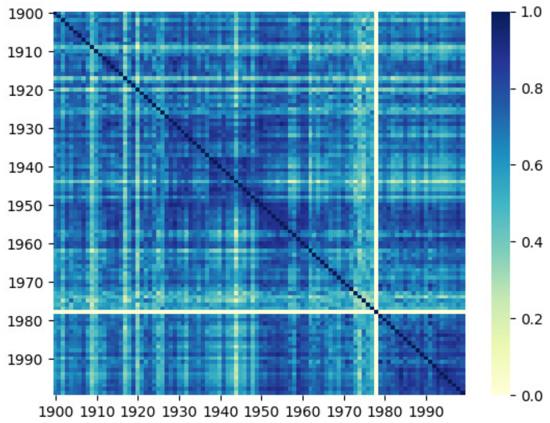

Fig. 1 The pairwise cosine similarities between the histories of Japanese cities (see Sec. 5) in any two years in the 20^{th} century. High similarity values are represented by dark pixels. Notice the matrix is symmetric and contains dark square regions along the diagonal. These regions represent cohesive time units.

summarization models [12], and then extend it to the our proposed APMRRW model for with the comparative summarization scenario.

4.2.1 Affinity-Preserving Random Walk

Given a graph $G = (V, E)$ with vertex set V and edge set $E = V^2$, where each edge $e = (v_i, v_j) \in E$ is associated with a weight w_{ij}, a random walk on graph G is defined as the discrete-time Markov chain with transition probability matrix P. P is a row-stochastic matrix computed by $P = D^{-1}W$, where D is a diagonal matrix with entries $D_{ii} = \sum_{v_j \in V} w_{ij}$ and W is the adjacency matrix with entries $W_{ij} = w_{ij}$.

However the traditional random walk suffers from democratic normalization mechanism [12] from W to P. The problem here is that the number of salient sentences (which are good candidates to be included in the summary) is usually far less than the number of bad candidate sentences given the summary length limit. Under the democratic normalization of $P = D^{-1}W$, a random surfer is highly likely to visit another neighboring bad candidate sentence when it is currently at a bad sentence given their high similarity. Such process can intrinsically suppress the effect of good candidates as well as amplify the adverse effect of bad sentences.

To make good and bad vertices more distinguishable, affinity-preserving random walk has been proposed [14]. With a new normalization technique, it is able to preserve the original affinity relations between vertices. Given a graph $G = (V, E)$, an augmented graph $G^{'}$ is constructed by adding an absorbing vertex v_0 to G. An affinity-preserving random walk on graph $G^{'}$ is defined as the discrete-time Markov chain with transition probability matrix P, where P is formulated as [14]:

$$P = \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{e} - \mathbf{v}/||W||_1 & W/||W||_1 \end{pmatrix} \tag{4}$$

Here, $\mathbf{0}$ and $\mathbf{e}$ are $1 \times |V|$ vectors with all elements 0 and 1, respectively. $\mathbf{v} = [D_{11}, D_{22}, ..., D_{|V||V|}]^T$ contains the total weights coming from each vertex. The affinity matrix W is normalized by its first norm (equal in value to $max_{i=1,2,...|V|}\mathbf{v}[i]$, or $max_{i=1,2,...|V|}D_{ii}$). During the affinity-preserving random walk, the surfer can not walk out of the absorbing vertex once it reaches there. On the other hand, it is less likely for the surfer at good sentences than bad sentences to walk into the absorbing vertex. Note that to ensure a good characterization of the sentence ranking distribution on graph $G^{'}$, we turn to the quasi-stationary distribution $\mathbf{x}$ [13, 14] as the distribution of unabsorbed random walkers, which is defined as $\mathbf{x}_i = Prob(X = v_i | X \neq v_0)$, and X represents the position of random walker.

4.2.2 Affinity-Preserving Mutually Reinforced Random Walk (APMRRW)

Given two input sets of timeline documents D_A and D_B, we construct a two-layer graph $\tilde{G} = (V_A, V_B, v_0^A, v_0^B, E_{AA}, E_{BB}, E_{AB})$, where V_A and V_B denote the sets of sentences contained in D_A and D_B respectively, while v_0^A and v_0^B denote the absorbing vertex in each layer. E_{AA}, E_{BB} and E_{AB} correspond to the edge sets reflecting the relation between sentences in V_A, the relation between sentences in V_B, and the relation between sentences in V_A and V_B, respectively.

We then compute affinity metrics W_{AA}, W_{BB}, W_{AB} and W_{BA}. Let $W_{AA} = [w_{v_i^A, v_j^A}]_{|V_A| \times |V_A|}$, where affinity $w_{v_i^A, v_j^A}$ is from $Sim_{cosine}(v_i^A, v_j^A)$. We define W_{BB}, W_{AB} and W_{BA} in a similar way, where $w_{v_i^A, v_j^B} = 1 - Sim_{cosine}(v_i^A, v_j^B)$, reflecting the *dissimilarity* between vertices v_i^A and v_j^B.

Next, affinity metrics W_{AA}, W_{BB}, W_{AB} and W_{BA} are normalized to transition metrics P_{AA}, P_{BB}, P_{AB} and P_{BA},, respectively without hurting the affinity information. Let $P_{AA} = W_{AA}/||W_{AA}||_1$, and W_{BB}, W_{AB}, W_{BA} are computed similarly. We then perform a two-layer affinity-preserving mutually reinforced random walk to propagate the node scores based on internal importance propagation within the same layer and external mutual reinforcement between different layers as follows:

$$\mathbf{x}_A^{t+1} = (1 - \alpha) \cdot \mathbf{x}_A^0 + \alpha \cdot P_{AA}^T P_{AB} \mathbf{x}_B^t \tag{5}$$

$$\mathbf{x}_B^{t+1} = (1 - \alpha) \cdot \mathbf{x}_B^0 + \alpha \cdot P_{BB}^T P_{BA} \mathbf{x}_A^t \tag{6}$$

Here, $\mathbf{x}_A^t$ and $\mathbf{x}_B^t$ denote respectively the scores of the sentence set V_A and of the sentence set V_B at the t-th iteration, which integrates the initial score and the score including within- and between-layer propagation. The initial importance scores $\mathbf{x}_A^0$ and $\mathbf{x}_B^0$ of the two sets are ones computed by Eq. (3) in Sec. 4.1 after normalization such that the scores sum to 1. In this paper, we empirically set $\alpha = 0.85$ as 0.85 is a commonly used value of the damping factor [15].

It can be proved that the closed-form solution $\mathbf{x}_A^*$ of Eq. (5) is the dominant eigenvector of $\mathbf{M}$ (see Appendix), or the eigenvector corresponding to the largest absolute eigenvalue of $\mathbf{M}$, where

$$\begin{aligned} \mathbf{M} = (1 - \alpha) \cdot \mathbf{x}_A^0 \mathbf{e}^T + \alpha \cdot (1 - \alpha) \cdot P_{AA}^T P_{AB} \mathbf{x}_B^0 \mathbf{e}^T \\ + \alpha^2 \cdot P_{AA}^T P_{AB} P_{BB}^T P_{BA} \end{aligned} \tag{7}$$

For numerical computation of the saliency scores, we can also iteratively run Eq. (5) and Eq. (6) until convergence. In order to guarantee the convergence of the iterative form, $\mathbf{x}_A$ and $\mathbf{x}_B$ are normalized after each iteration. However, we point out that anything that successfully solves Eq. (5) and Eq. (6) should lead to an equal value of $\mathbf{x}_A$ and $\mathbf{x}_B$.

4.3 Dynamic Ranking Algorithm

We now present the idea of a dynamic ranking framework for generating globally important sentences, which are assumed to be locally important in many time units. A comparative summary $\tilde{S}$ can be generated for any period when it is required. In each atomic time unit t, a two-layer graph whose vertices correspond to the sentences of the two compared document sets within t is constructed, and the aforementioned APMRRW algorithm is applied to the graph to locally rank the sentences. Then the sentence ranked by APMRRW is reranked by the commonly used MMR algorithm [18] to avoid redundancy in a summary, by providing penalty corresponding to the similarity between a newly extracted sentences and the already extracted sentences. Based on the ranking score, top m sentences (m is the pre-defined summary length introduced in Sec. 3) are selected as the summary at the time unit t, and will be passed over to the next time unit as past information. Then a new graph consisting of both the previous summary and sentences in a new unit is constructed and the same procedure is applied. In this way, the summarization process works dynamically and a comparative summary of input article sets can be generated for any given period within the entire time span if necessary.

The process is shown in Alg. 1.

Algorithm 1 Dynamic Ranking Algorithm

```
Input: $D$, $m$, $[t_{begin}, t_{end}]$
  $S \leftarrow \{\}$
  for $t \leftarrow t_{start}$ to $t_{end}$ do
    $\tilde{S} \leftarrow S + D_t$
    ranking $\tilde{S}$ with APMRRW
    ranking $\tilde{S}$ with MMR
    if $|\tilde{S}| > m$ then
      $S \leftarrow$ top $m$ sentences of $\tilde{S}$
    else
      $S \leftarrow \tilde{S}$
    end if
  end for
Output: $S$
```

5 Experiments

5.1 Datasets

We test our methods on diverse Wikipedia categories. Since our research problem is non-trivial and unexplored, hence to make the evaluation more feasible, we selected from existing Wikipedia categories and lists of moderate size, with which all the annotators were quite familiar. In particular, we perform experiments on generating the comparative summary of histories of 3 pairs of Wikipedia categories including location categories (Japanese cities vs. Chinese cities, denoted as L_1 and L_2), organization categories (western teams of The National Basketball Association (NBA) league in North America vs. eastern teams of NBA, denoted as O_1 and O_2) and person categories (Japanese Prime Ministers till the end of WW2 vs. Japanese Prime Ministers after WW2, denoted as P_1 and P_2), respectively. For preparing the documents, the history of each entity is extracted from the "History" section in the corresponding Wikipedia article. To capture historical events, we collect all sentences along with a single date following related works [2, 6, 25] using SUTime [22]. The basic statistics about our datasets are shown in Tab. 1.

Table 1 Summary of datasets

Dataset	Category	# Docs	# Sentences
L_1	Japanese Cities	532	22,045
L_2	Chinese Cities	357	6,444
O_1	Western NBA Teams	15	3,755
O_2	Eastern NBA Teams	15	3,701
P_1	Japanese PMs (pre WW2)	32	2,338
P_2	Japanese PMs (post WW2)	30	1,715

5.2 Experimental Settings

In this study, we experimentally set the summary size to be 20 sentences. For era detection (Sec. 4.1), we let the number of eras for the location and organization datasets to be 10, and for the person datasets to be 5 following [2]. The time unit is empirically set to 1 year (see Sec. 5.7). In this study we adopt the widely-used Skip-gram model [24] to represent terms and sentences. We obtain the distributed vector representations of each word by training the Skip-gram model on the entire English Wikipedia from 2016 using the gensim Python library [23]. The vector representation of a sentence is a TF-IDF weighted combination of the vectors of terms. The number of dimensions of word vectors is experimentally set to 200.

5.3 Reference Summary

To assess the quality of generated summaries by different test methods, three human judges manually annotated the experimental datasets. After reading the content of two compared document sets, the annotators were asked to write up to 300-words long[*2] *reference summary* for each document set that will help in grasping the contrastive content of input document collections. In particular, we pool the summaries created by all the analyzed methods, then the following instructions were given to the annotators: (1) select all the representative and discriminative sentences and (2) write a 300-word summary of the selected text.

5.4 Analyzed Methods

We list below all the analyzed methods as follows. We prepare D-APMRRW for the overall process as described in Sec. 4. We test D-APMRRW* for the model that only considers the steps in Sec. 4.2 and in Sec. 4.3, thus skipping the era detection in Sec. 4.1. Similarly, we test D-MRRW and APMRRW for the models that skip the affinity-preserving in Sec. 4.2, and the dynamic ranking in Sec. 4.3, respectively.

To assess the effectiveness of D-APMRRW, our baselines also include recent related work. For comparative summarization models (denoted as CS models), we use the discriminative sentence selection model (DSS, [7]) and the integer linear programming model (ILP, [11]) as baselines. For timeline summarization methods (denoted as TS models), we test the exemplar-based Markov random walk model (E-MRW, [2]), exemplar-based HITS model (E-HITS, [2]), and the online graph-based model (OGM, [3]). We also consider two popular multi-document summarization methods (denoted as MDS models): (1) LexRank [20] that ranks sentences via a Markov random walk strategy and (2) ClusterCMRW [19] which scores sentences by a clustering-based approach.

5.5 Evaluation Metrics

We evaluate all the models using *ROUGE-1.5.5 toolkit [17]*. The ROUGE is a widely used metric which has been officially adopted by DUC for automatic summarization evaluation. It measures summary quality by counting overlapping units between the candidate summary and the reference summary [17]. In the experiment, we report the f-measure values of ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W and ROUGE-SU, which are based on overlapping unigrams, bigrams, longest common subsequence (LCS), weighted LCS and skip-bigrams plus unigrams, respectively. The higher the ROUGE scores are, the more similar the machine-generated summary and the reference summary are, and thus the more effective the summarization approach is.

5.6 Results and Discussion

Tab. 2 shows the performance of summaries generated by all the methods in terms of ROUGE scores. From the results, we make the following observations. (1) D-APMRRW model exhibits the best performance in terms of all metrics. In addition, our proposed D-APMRRW model, D-APMRRW* model and D-MRRW model all achieve higher ROUGE scores than compared recent works, which proves the summarization effectiveness of the proposed models. (2)

[*2] The average length of an English sentence is around 15 words [7], so we choose 300 as the number of words for the summary size of 20 sentences.

D-APMRRW model, which considers the overall process as described in Sec. 4, outperform all its variants. More concretely, D-APMRRW shows a 27.2%, 27.4% and 326.7% increase in terms of average ROUGE score compared to D-APMRRW*, D-MRRW and APMRRW, respectively. Hence, we conclude that the proposed assumptions all help to enhance the performance of salient sentence extraction, and that dynamic ranking (Sec. 4.3) offers a more significant performance increase than era detection (Sec. 4.1) and affinity preservation (Sec. 4.2). This observation implies that globally important sentences are better candidates than locally imporant sentences to be included in the summary.

From Tab. 2, it can also be observed that the existing methods perform relatively poorly. For multi-document summarization methods, the plausible reason can be that they tend to select very general sentences similar to many other sentences in the same document set, while such sentences may not have enough discrimination nor may interest annotators. Contrastingly, the comparative summarization methods focus more on discovering sentences delivering set-specific information, which prevents them from embodying historic significance. Finally, regarding the timeline summarization methods, although they aim to discovering salient temporal information, they suffer from similar drawbacks of multi-document summarization methods.

5.7 Additional Analysis

We further investigate the influence of the value of α in Eq. (5) and Eq. (6) as well as the value of time unit in dynamic ranking, as shown in Fig. 2. We first test the α within the range [0, 1] and with a step of 0.1. We can see from the figure that the value of α has an effect on the performance of summarization. In this paper, we empirically set $\alpha = 0.85$ as 0.85 is a commonly used value of the damping factor [15]. In additon, the length of time unit is set in the range [1,10] with a step of 1 year. When time unit is larger than 1, the system achieves worse performance due to the plausible reason that many good candidate sentences are discarded during the dynamic ranking, as we enlarge the candidate sentences set while keeping the summary size fixed per local scoring. In general, we can see that both α and time unit need to be fine-tuned to achieve an optimal performance.

5.8 Limitations

This work has several limitations that we need to acknowledge. First, while we assume that many timeline documents can be divided into latent eras (e.g., dynasties in histories of cities), there may exist timelines which are not following such hypothesis. In such case our approach will fail in segmenting the whole time span of input documents, as well as initializing the importance of each candidate text unit. Moreover, most existing works involving linear text segmentation [2, 16, 26] rely on a pre-defined number of eras, while the number of latent eras should be automatically decided based on the input documents. Second, although we make use of sentences whose dates are explicitly mentioned, we discard sentences whose dates are implicitly associated. Such temporal information should also be estimated and utilized, for example, by using approaches similar to [21]. Third, our models produce summaries in which each event is in the form of a sentence from a particular timeline document. The sentence representation may however contain too specific details which might be true only for the instance from which the given sentence has been extracted. To improve the readability of generated summary, one would need to incorporate abstractive summarization strategies or generalization procedures.

6 Conclusions and Future Work

This work approaches the problem of a special kind of summarization task - comparative timeline summarization (CTS). The special character of our proposed summarization allows capturing important comparative aspects of evolutionary trajectories hidden in two sets of timeline documents. Users can benefit from such novel task for needs including gaining insights from history, analyzing trends, as well as for educational or entertaining purposes. To address the introduced problem we develop a three-step approach which consists of era detection, an affinity-preservining mutually reinforced random walk model and a dynamic ranking process. The effectiveness of our models has been demonstrated by the experiments on 3 pairs of Wikipedia category datasets using ROUGE toolkit.

In future, we plan to conduct evaluation on diverse types of timeline documents as well as incorporate abstractive summarization strategies for increasing the readability of the generated summaries. We will also extract and utilize temporal information in input text more thoroughly using an approach of [21].

7 Acknowledgements

This research has been partially supported by MEXT Kakenhi grants (#17H01828, #18K19841 and #19H04215).

References

[1] Tran, T.A., Nieder ´ee, C., Kanhabua, N., Gadiraju, U., Anand, A.: Balancing nov- elty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 1201–1210. ACM (2015)

[2] Duan, Y., Jatowt, A., Tanaka, K.: Discovering typical histories of entities by multi- timeline summarization. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media. pp. 105–114. ACM (2017)

[3] Suzuki, S., Kobayashi, I.: On-line summarization of time-series documents using a graph-based algorithm. In: Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (2014)

[4] Yan, R., Wan, X., Otterbacher, J., Kong, L., Li, X., Zhang, Y.: Evolutionary time- line summarization: a balanced optimization framework via iterative substitution. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 745–754. ACM (2011)

[5] Abujabal, A., Berberich, K.: Important events in the past, present, and future. In: Proceedings of the 24th International Conference on World Wide Web. pp. 1315–1320. WWW ’ 15 Companion, ACM, New York, NY, USA (2015).

[6] Bamman, D., Smith, N.A.: Unsupervised discovery of biographical structure from text. Transactions of the Association for Computational Linguistics 2, 363–376 (2014)

[7] Wang, D., Zhu, S., Li, T., Gong, Y.: Comparative document summarization via discriminative sentence selection. TKDD 6(3), 12 (2012)

[8] Ren, Z., de Rijke, M.: Summarizing contrastive themes via hierarchical non- parametric processes. In: SIGIR. pp. 93–102. ACM (2015)

[9] Gillenwater, J., Kulesza, A., Taskar, B.: Discovering diverse and salient threads in document collections. In: EMNLP. pp. 710–720. Association for Computational Linguistics (2012)

[10] He, L., Li, W., Zhuge, H.: Exploring differential topic models for comparative summarization of scientific papers. In: COLING. pp. 1028–1038 (2016)

[11] Huang, X., Wan, X., Xiao, J.: Comparative news summarization using linear pro- gramming. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2. pp. 648–653. Association for Computational Linguistics (2011)

[12] Wang, K., Liu, T., Sui, Z., Chang, B.: Affinity-preserving random walk for multi- document summarization. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 210–220 (2017)

Table 2 Overall performance comparison of all methods using ROUGE scores.

Type	Acronym	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU
CS models	DSS [7]	0.050	0.007	0.050	0.012	0.002
	ILP [11]	0.090	0.023	0.089	0.020	0.008
MDS models	LexRank [20]	0.159	0.045	0.156	0.032	0.020
	ClusterCRMW [19]	0.178	0.053	0.175	0.037	0.024
TS models	E-MRW [2]	0.184	0.104	0.181	0.036	0.023
	E-HITS [2]	0.105	0.031	0.104	0.022	0.006
	OGM [3]	0.182	0.103	0.180	0.033	0.022
Proposed models	APMRRW	0.080	0.022	0.079	0.017	0.003
	D-MRRW	0.218	0.163	0.217	0.045	0.029
	D-APMRRW*	0.209	0.188	0.208	0.041	0.027
	D-APMRRW	**0.258**	**0.251**	**0.258**	**0.049**	**0.040**

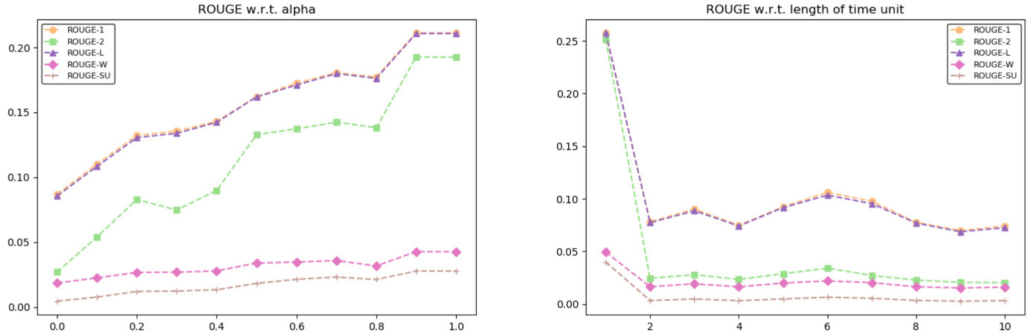

Fig. 2 Performance of D-APMRRW w.r.t. α (left) and time unit (right).

[13] Darroch, J.N., Seneta, E.: On quasi-stationary distributions in absorbing discrete- time finite markov chains. Journal of Applied Probability 2(1), 88–100 (1965)

[14] Cho, M., Lee, J., Lee, K.M.: Reweighted random walks for graph matching. In: European conference on Computer vision. pp. 492–505. Springer (2010)

[15] Brin,S.,Page,L.:Reprint of The anatomy of a large-scale hypertextual web search engine. Computer networks 56(18), 3825–3833 (2012)

[16] Choi, F.Y.: Advances in domain independent linear text segmentation. In: Pro- ceedings of the 1st North American chapter of the Association for Computational Linguistics conference. pp. 26–33. Association for Computational Linguistics (2000)

[17] Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co- occurrence statistics. In: Proceedings of the 2003 Conference of the North Ameri- can Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. pp. 71–78. Association for Computational Linguistics (2003)

[18] Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reorder- ing documents and producing summaries. In: SIGIR. pp. 335–336. ACM (1998)

[19] Wan, X., Yang, J.: Multi-document summarization using cluster-based link anal- ysis. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 299–306. ACM (2008)

[20] Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research 22, 457–479 (2004)

[21] Jatowt, A., Au Yeung, C.M., Tanaka, K.: Estimating document fo- cus time. In: Proceedings of the 22Nd ACM International Conference on Information Knowledge Management. pp. 2273–2278. CIKM ’ 13, ACM, New York, NY, USA (2013).

[22] Chang, A.X., Manning, C.D.: Sutime: A library for recognizing and normalizing time expressions. In: Lrec. vol. 2012, pp. 3735–3740 (2012)

[23] ˇeh uˇrek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Chal- lenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en

[24] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word repre- sentations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[25] Bamman, D., Smith, N.A.: Unsupervised discovery of biographical structure from text. Transactions of the Association for Computational Linguistics 2, 363–376 (2014)

[26] Alsudais, A., Tchalian, H.: Corpus periodization framework to periodize a tempo- rally ordered text corpus (2016

A Appendix

We now prove Eq. (11) in Sec. 4.2. Similar to the PageRank algorithm [15], Eq. (5) and Eq. (6) will converge satisfying Eq. (8) and Eq. (9) as follows.

$$\mathbf{x}_A^* = (1-\alpha)\cdot\mathbf{x}_A^0 + \alpha\cdot P_{AA}^T P_{AB}\mathbf{x}_B^* \tag{8}$$

$$\mathbf{x}_B^* = (1-\alpha)\cdot\mathbf{x}_B^0 + \alpha\cdot P_{BB}^T P_{BA}\mathbf{x}_A^* \tag{9}$$

Since $\mathbf{e}^T\cdot\mathbf{x}_A^* = \mathbf{e}^T\cdot\mathbf{x}_B^* = 1$, $\mathbf{x}_A^*$ can be then solved as below.

$$\begin{aligned}\mathbf{x}_A^* &= (1-\alpha)\cdot\mathbf{x}_A^0 \\ &\quad+\alpha\cdot P_{AA}^T P_{AB}\left\{(1-\alpha)\cdot\mathbf{x}_B^0 + \alpha\cdot P_{BB}^T P_{BA}\mathbf{x}_A^*\right\} \\ &= (1-\alpha)\cdot\mathbf{x}_A^0 + \alpha\cdot(1-\alpha)\cdot P_{AA}^T P_{AB}\mathbf{x}_B^0 \\ &\quad+\alpha^2\cdot P_{AA}^T P_{AB}P_{BB}^T P_{BA}\mathbf{x}_A^* \\ &= \mathbf{M}\cdot\mathbf{x}_A^*\end{aligned} \tag{10}$$

where

$$\begin{aligned}\mathbf{M} = (1-\alpha)\cdot\mathbf{x}_A^0\mathbf{e}^T + \alpha\cdot(1-\alpha)\cdot P_{AA}^T P_{AB}\mathbf{x}_B^0\mathbf{e}^T \\ +\alpha^2\cdot P_{AA}^T P_{AB}P_{BB}^T P_{BA}\end{aligned} \tag{11}$$

Similarly, the closed-form solution $\mathbf{x}_A^*$ of (6) is the dominant eigenvector of $\mathbf{N}$, where

$$\begin{aligned}\mathbf{N} = (1-\alpha)\cdot\mathbf{x}_B^0\mathbf{e}^T + \alpha\cdot(1-\alpha)\cdot P_{BB}^T P_{BA}\mathbf{x}_A^0\mathbf{e}^T \\ +\alpha^2\cdot P_{BB}^T P_{BA}P_{AA}^T P_{AB}\end{aligned} \tag{12}$$