Discovering Typical Histories of Entities by Multi-Timeline Summarization

Yijun Duan
Graduate School of Informatics,
Kyoto University, Japan
yijun@dl.kuis.kyoto-u.ac.jp

Adam Jatowt
Graduate School of Informatics,
Kyoto University, Japan
adam@dl.kuis.kyoto-u.ac.jp

Katsumi Tanaka
Graduate School of Informatics,
Kyoto University, Japan
tanaka@dl.kuis.kyoto-u.ac.jp

ABSTRACT

Categorization is a common solution used for organizing entities. For example, there are over 1.13 million categories in Wikipedia which group various types of entities such as persons, locations, etc. What is however often lacking when it comes to understanding categories is a clear information about the common aspects of the entities in a given category, for example, information on their shared histories. We propose in this paper a novel task of automatically creating summaries of typical histories of entities within their categories (e.g., a typical history of a Japanese city). The output summary is in the form of key representative events together with the information on their average dates. We introduce 4 methods for the aforementioned task and evaluate them on Wikipedia categories containing several types of cities and persons. The summaries we generate can provide information on the common evolution of entities falling into the same category as well as they can be compared with the summaries of related categories for providing contrastive type of knowledge.

CCS CONCEPTS

•**Information systems → Digital libraries and archives;**

KEYWORDS

entity summarization, digital history, typicality, Wikipedia

1 INTRODUCTION

Categorization is a common strategy applied for organizing and understanding entities. Wikipedia, which is considered these days to be the most comprehensive encyclopedia, contains over 1.13 million categories [3]. Each category typically consists of multiple related members that share some common traits (e.g., list of cities in Japan, list of American scientists active in the 19th century, etc.). To obtain a good understanding of a category, one needs to know well about its members, which is definitely a difficult task, especially, for larger categories. For example to fully understand the category of Japanese cities a user would need to read over 500 Wikipedia articles about different Japanese cities.

HT'17, July 4-7, 2017, Prague, Czech Republic.

DOI: http://dx.doi.org/10.1145/3078714.3078725

In particular our interest is on historical knowledge. Wikipedia abounds in knowledge about the histories of entities or concepts. Many articles contain dedicated history sections. For instance, an article about a person typically contains his/her biography, and an article about a city usually includes extended section covering its history. In fact, most entities cannot be properly understood without the knowledge of their histories. The same can be said about their categories.

What is the history of Japanese cities? How is it different from, e.g., the history of Chinese or UK cities? Which events frequently occurred during the life of French scientists? How different was the life of a French scientist in the 19th century from that of an American scientist at that time? Questions of this type are not easy to be answered as they usually require substantial knowledge of history, or at least necessitate much effort. For an average user, to answer them he/she would need to read histories of many individual instances.

Straightforward approaches to automatically creating such historical knowledge would be to formulate it as a standard multi-document summarization task. However, traditional multi-document summarization techniques are not suitable for our scenario. The first problem is that input documents in multi-document summarization are assumed to be similar to each other (e.g., news articles about the same event). This assumption is not guaranteed in the case of the category history summarization as entity histories can be quite different from each other. For example, while we expect to find some common events and tendencies within Japanese cities, each individual city has many specific events in its history, which have varying degree of resemblance to the events of other cities. The second problem stems from the strong temporal character of input documents in our task. Entity histories (e.g., biographies) typically have a sequential character and abound in multiple dates used to mark important events in time, delineate key periods, support explanation of causal-effect relationships and, in general, to provide logical progression and coherent account of entity's history.

A uniting feature of traditional multi-document summarization techniques is an implicit assumption that the importance of a sentence can be estimated based on its similarity to other sentences within the input document set. For instance, in MEAD system [25] a sentence is judged important if it is similar to the centroid sentence, or if it is similar to many important sentences as in LexRank method [12]. Considering the unique characteristics of our task, it is clear that the common approach of sentence selection used in multi-document summarization is not appropriate. To provide effective means for capturing shared traits in entity histories we make use of the following observations: (1) *Histories of many types of entities (e.g., countries, persons) can be often divided into particular eras.* For example, the history of Japan as well as the one of Japanese cities covers several dynasties, while a person's life can be divided into stages such as childhood, early education, early career, etc. (2) *Documents describing histories of entities often contain underlying themes.* These themes may be also correlated. (3) *Themes as well as eras are usually not equally important.* An event contained in an important era and being part of important topics can be regarded more salient than one in less important era or belonging to trivial topics. (4) *Finally, some entities are better representatives of a category than others.* This is known as the *graded structure* [26] of a category.

An event belonging to a typical entity is then deemed to be more salient than one of a trivial entity.

To reflect the above observations we rely on graph analysis. In particular, we adapt Markov Random Walk (MRW) [18] and Hyperlink-Induced Topic Search (HITS) [16] methods to our scenarios. To address the limitations of traditional techniques we propose 4 different models which are based on MRW and HITS and incorporate additional information about documents, eras, topics and topic correlation. The resulting summaries are in the form of timelines containing key events represented by set of words (Section 8.7 shows an example of output). Our experiments are performed on 7 Wikipedia category datasets (3 cities datasets and 4 persons datasets) with the results demonstrating higher effectiveness of our methods when compared to common multi-document summarization techniques.

To sum up, we make the following contributions in this paper:

(1) We introduce a new research problem of characterizing entity categories by generating typical histories of their entities.
(2) We propose 4 different models to discover typical histories of entities utilizing information about sentences, eras, topics and topic correlation. All our models work in an unsupervised way, which is important considering the lack of manually created summaries for most of the categories.
(3) The effectiveness of our methods is demonstrated in experiments on 7 Wikipedia category datasets.

The reminder of this paper is organized as follows. The related works are introduced in Section 2. We formulate our research problem and discuss different types of summaries in Section 3. Section 4 introduces the approach for event representation. Our summarization models are presented in Sections 5 and 6. Section 7 discusses the generalization of extracted events. In Section 8, we describe the experiments and evaluation results. Finally, we draw conclusions in Section 9.

2 RELATED WORK

Multi-Document Summarization. Multi-document summarization is the process of creating a summary that retains the most important information from multiple documents. Summarization methods can be coarsely divided into extractive summarization and abstractive summarization techniques. Extractive type of methods, to which our techniques belong, aims to select a subset of units (e.g. words, sentences) of original documents to form a summary. As an example of extractive methods, the centroid-based method MEAD [25] scores sentences based on sentence level and inter-sentence level features including cluster centroid, position, and TF-IDF, etc. Graph-based ranking methods, such as LexRank [12] and TextRank [22], have been developed to estimate sentence importance using random walks and eigenvector centrality. In order to remove redundancy in final summaries, Maximal Marginal Relevance (MMR) technique [10] has been commonly used. Wan *et al.* have improved the graph-ranking algorithm by utilizing sentence-to-sentence and sentence-to-topic relationships [29], and intra-document and inter-document links between sentences [30]. In contrast, abstractive methods create summary containing words not explicitly present in the original documents. In this process, information fusion [5], sentence compression [17] and reformulation [21] may be applied.

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 [31], Yan *et al.* proposed the evolutionary timeline summarization (ETS) to compute evolution timelines consisting of a series of time-stamped summaries. David *et al.* presented a method for discovering biographical structure based on a probabilistic latent variable model [4]. Abdalghani *et al.* [2] addressed the problem of identifying important events in the past, present, and future from semantically-annotated large-scale document collections. Tuan *et al.* [28] presented a novel approach for timeline summarization of high-impact events, which uses entities instead of sentences for summarizing the events.

The above-mentioned methods can not be directly applied to our task. While documents are timestamped in the timeline summarization task, in the task of category summarization, each document spans over a certain range of time. Due to this, existing timeline summarization techniques are unable to estimate well the representativeness of a document and the correlation between sentences, which are important factors considered for generating summaries.

3 PROBLEM STATEMENT

3.1 Input

The input are documents containing histories of entities belonging to the same category. Each history-related document spans over a certain range of time and each sentence is assumed to refer to a historical event. The dates of events can be either explicitly mentioned in the sentence or could be estimated based on nearby sentences.

We note that sometimes categories can consist of entities with very diverse histories. Naturally, the summarization task becomes then more difficult in those cases.

3.2 Research Problem

Given a set of history-related documents $[d_1, d_2, ..., d_n]$ each about a particular entity within the same category and a time window $[t_{begin}, t_{end}]$, the task is to select k most typical historical events $[e_1, e_2, ..., e_k]$ to form a summarized timeline reflecting typical history of the entities within $[t_{begin}, t_{end}]$. Each event in the summary is represented by words $[w_1, w_2, ..., w_i]$.

The events selected for inclusion into the summary should be:

(1) **typical**: we want to retain typical information of the history of category entities;
(2) **diverse**: events contained in the summary should be both diverse in their contents and in terms of their occurring time;

(3) **comprehensible**: events contained in the summary should be understandable to users.

3.3 Types of Output Summary

Cognitive science studies suggest two modes in which people understand categories: *prototype view* [27] and *exemplar view* [9]. The first one posits that a category be represented by a constructed prototype (sometimes called centroid), such that entities closer to the prototype are considered better examples of the category. The exemplar view is an alternative to the prototype view that proposes using real entities as exemplars instead of abstract prototypes that might actually not exist. Based on this division, we propose two types of summarization approaches:

Prototype-based summarization. In the prototype-based summarization, events may come from the history of arbitrary entity within the category. The prototype-based summary represents the category by constructing an imaginary prototype.

Exemplar-based summarization. In the exemplar-based summarization, events are drawn from a relatively small set of typical representatives among all entities. The size of the set depends on the size of summary. The exemplar-based summary thus uses a few typical representative instances to describe the whole category.

3.4 System Overview

Fig. 1 provides an overview of our approach. We first pre-process input documents and extract events. For the prototype-based summarization, we additionally detect eras and topics. Then we compute event importance by MRW and HITS for both the prototype-based summarization and exemplar-based summarization. Lastly, we remove redundancy and generalize the events for constructing final summary.

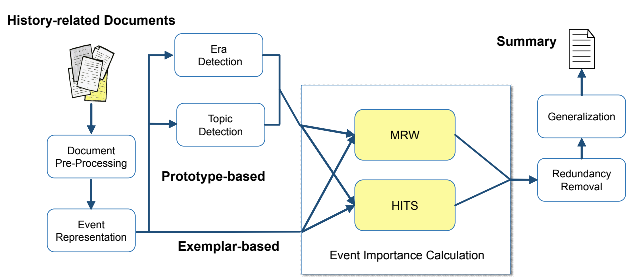

Figure 1: System Overview.

4 EVENT REPRESENTATION

A historical event is represented by a sentence and is associated with a date of its occurrence. As not all words of the original sentence are meaningful, each sentence is first normalized by pre-processing steps such as removing stopwords, stemming and retaining the most frequent 5,000 unigrams and bigrams. In the recent years, word2vec [23] was widely utilized for automatically learning the meaning behind words based on neural networks. We use word2vec to represent terms and events. The vector representation of an event is a weighted combination of the vectors of terms contained in the normalized sentence that represents the event. The weight of a term is its TF-IDF value calculated from the original corpus.

5 PROTOTYPE SUMMARY GENERATION

In this section we describe two methods which rely on prototype summarization strategy. Both use era and topic information. We start then with explaining the way to detect eras and themes underlying our datasets.

5.1 Eras Detection

Given a sequence of atomic time units $\xi = (t_1, t_2, ..., t_n)$, the task is to select a proper segmentation Θ containing m eras dividing 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 $\Theta = (T_1, T_2, ..., T_m)$, where $T_i = [\tau_b^i, \tau_e^i]$. In order to perform era detection we state two hypotheses:

Hypothesis 1 *A statistically significant increase or decrease in the number of events in two adjacent time units can be an indicator of the emergence of a new era.*

Hypothesis 2 *Events occurring in the same era tend to be more similar to each other than events occurring in other eras.*

Hypothesis 1 is similar to the one stated in [1] where authors utilized statistically significant changes in the frequencies of news articles in time segments in order to locate different stages of events.

The above hypotheses form the basis for the two stage process of era detection. We discuss both the steps below:

Chi-Square Test. The initial set of time units the category history is $\xi = (t_1, t_2, ..., t_n)$, where each time unit t_i represents a year. Based on Hypothesis 1, a chi-square test of independence is applied. We test for the independence of adjacent time units, and the lack of independence allows the adjacent time units to be combined. More concretely, the chi-square test is used to determine whether two neighboring time units exhibit a statistically significant association based on the number of contained events.

The default significance level is set to 0.05, thus a statistically significant change is defined where the χ^2 value exceeds the critical cut-off of 3.84. By this we obtain the intermediate set of segments $\varsigma = (\mu_1, \mu_2, ..., \mu_k)$, where $\mu_i = [\eta_b^i, \eta_e^i | \eta_b^i \in \xi, \eta_e^i \in \xi]$, is created after some of time units are combined.

Optimization. We next use an optimization formula to determine the final eras based on Hypothesis 2. Given the pre-set number of final periods, m, every possible combination Θ of segments from the intermediate set will be explored. Formally, let $\Theta = (T_1, T_2, ..., T_m)$ where $T_i = [\tau_b^i, \tau_e^i | \tau_b^i \in \varsigma, \tau_e^i \in \varsigma]$. In particular, we prefer the combination, in which the eras to be selected are characterized by high intra-similarity, low inter-similarity, and, in addition, they have high abundancy defined as the number of instances having their events in a given era. The era combination that has the highest score by applying Eq. (1) will be adopted.[1]

[1]We experimentally set weights for ω_1, ω_2 and ω_3 in Eq. (1) to be 0.4, 0.4 and 0.2, respectively.

$$
\begin{aligned}
\Theta \equiv \operatorname{argmax}[\omega_1 \sum_{i=1}^{m} IntraSimilarity(T_i) \\
- \omega_2 \sum_{i=1}^{m-1} InterSimilarity(T_i, T_{i+1}) \\
+ (1 - \omega_1 - \omega_2) \cdot \sum_{i=1}^{m} Abundancy(T_i)]
\end{aligned} \tag{1}
$$

Here, the intra-similarity measures the cosine similarity between events within a given era:

$$
IntraSimilarity(T_i) = \sum_{e_i \in T_i} \sum_{e_j \in T_i} \frac{Sim_{cosine}(e_i, e_j)}{|T_i|^2} \tag{2}
$$

The inter-similarity measures the cosine similarity between the events of the neighboring eras:

$$
InterSimilarity(T_i, T_{i+1}) = \sum_{e_i \in T_i} \sum_{e_j \in T_{i+1}} \frac{Sim_{cosine}(e_i, e_j)}{|T_i| \cdot |T_{i+1}|} \tag{3}
$$

Finally, the abundancy of an era indicates how many of the category instances have at least one event located in this era.

5.2 Topic Detection

Different entities may share similar historical events which however may not belong to the same eras. For instance, many Japanese cities were hit by earthquakes at different times throughout the last millennium. Thus, in addition to detecting eras we also conduct topic detection for better capturing event importance.

Clustering algorithms like K-means are popular techniques to detect topics. However, these are not appropriate as each event is expected to belong to only one topic. Latent Dirichlet Allocation (LDA) [7] allows for soft association of topics with events. However, LDA does not explicitly compute topic to topic association which could constitute another useful signal for estimating topic importance.

We then construct topics with Correlated Topic Model (CTM) [8] which captures both topic-event relations as well as topic-topic relations. Given a set of documents D=$[d_1, d_2, ..., d_N]$ and its vocabulary W=$[w_1, w_2, ..., w_M]$, CTM returns a set of latent topics Z=$[z_1, z_2, ..., z_K]$ (K is pre-specified). Each document d_j is considered as arising from the mixture of topics in Z, each of which is a distribution over the vocabulary W. In addition, the covariance structure among topics Z (which is a K-dimensional covariance matrix) is estimated via adopting the logistic normal distribution to model the latent topic proportions of a document.

Given CTM, we are able to obtain the per-document topic distributions $P(z_i|d_j)$, the per-topic word distributions $P(w_i|z_k)$ and the topic-topic correlations $Corr(z_i, z_j)$. We then incorporate this information to compute event importance as detailed next.

5.3 Prototype-based MRW

The Markov Random Walk Model (MRW) has been successfully exploited in multi-document summarization. MRW is a way of calculating the importance of a vertex within a graph based on global information recursively drawn from the entire graph. In MRW model, voting (or "recommendation") between two vertices is represented in the form of a link from one vertex to another, and the score associated with a vertex is determined by the votes received from adjacent vertices.

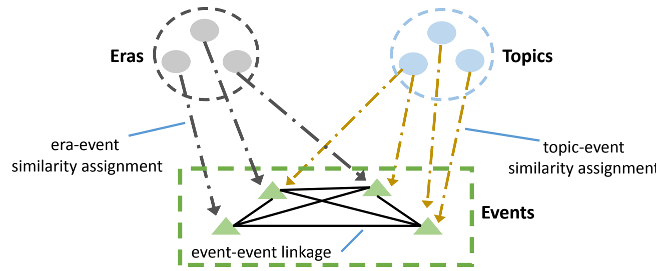

Figure 2: Illustration of Prototype-based MRW.

We use MRW for estimating the importance of events (see Fig. 2). In this process we make use of era and topic information. The underlying topics as well as eras are not equally important. An event contained in an important era and being part of important topics is deemed more salient than one in a less important era or belonging to trivial topics. Thus in order to calculate event importance with the prototype-based MRW, we state four hypotheses for determining important events as below:

Hypothesis 3 *An important event is contained in an important era and is similar to many events in this era.*

Hypothesis 4 *An important event is similar to important topics.*

Hypothesis 5 *An important event is similar to other important events.*

Hypothesis 6 *An important event is strongly correlated with other important events.*

Formally, let $G = (V_e, Q_{ee})$ be a graph (see Fig. 2) with the set of vertices V_e and the set of edges Q_{ee}. Let V_e=E=$\{e_i\}$, T=$\{T_j\}$, Z=$\{z_k\}$ denote the sets of events, detected eras and that of detected topics, respectively, and let $Q_{ee} = \{q_{ij}|e_i, e_j \in V_e\}$ represent the set of links between events. Below we are going to explain the way to assign initial scores to vertices in V_e and the way to compute edge weights.

First we compute the importance score of an era T_i denoted by $I(T_i)$ as follows:

$$
I(T_i) = Sim(T_i, D) = \sum_{e_i \in T_i} \sum_{e_j \in D} \frac{Sim_{cosine}(e_i, e_j)}{|T_i| \cdot |D|} \tag{4}
$$

where D is the document set. We also compute the importance score of a topic z_i denoted as $I(z_i)$:

$$
I(z_i) = \frac{\sum_{d \in D} P(z_i|d)}{|D|} \tag{5}
$$

$Sim(e_i, T_l)$ denotes the similarity between an event e_i and the era T_l that e_i is contained in. It is computed as follows:

$$
Sim(e_i, T_l) = \frac{1}{|T_l|} \cdot \sum_{e_j \in T_l} Sim_{cosine}(e_i, e_j) \tag{6}
$$

$Sim(e_i, z_k)$ denotes the similarity between an event e_i and a topic z_k:

$$Sim(e_i, z_k) = \frac{\sum_{w \in e_i} P(w|z_k)}{|e_i|} \quad (7)$$

Let $Score(e_i)$ be the score of a vertex e_i. We then compute the initial score $Score^0(e_i)$ of e_i used for MRW as follows:

$$Score^0(e_i) = I(T_l|e_i \in T_l) \cdot Sim(e_i, T_l) \cdot \sum_{z_k \in Z} I(z_k) \cdot Sim(e_i, z_k) \quad (8)$$

In other words, we assign high initial score to a given event if it belongs to an important era (Hypothesis 3), it is similar to this era (Hypothesis 3) and it is similar to important topics (Hypothesis 4).

We also associate each edge q_{ij} in Q_{ee} with an affinity weight w_{ij} between events e_i and e_j. Considering Hypothesis 5 and Hypothesis 6, this weight is computed using both the similarity $Sim(e_i, e_j)$ and the correlation $Corr(e_i, e_j)$ between the two events:[2]

$$w_{ij} = \alpha \cdot Sim_{cosine}(e_i, e_j) + (1 - \alpha) \cdot Corr(e_i, e_j) \quad (9)$$

$$Corr(e_i, e_j) = \sum_{z_i \in Z} \sum_{z_j \in Z} \frac{Sim(e_i, z_i) \cdot Sim(e_j, z_j) \cdot Corr(z_i, z_j)}{|Z|^2} \quad (10)$$

The transition probability p_{ij} from e_i to e_j is then computed by normalizing the corresponding affinity weight to ensure convergence:

$$p_{ij} = \frac{w_{ij}}{\sum_{e_k \in V_e} w_{ik}} \quad (11)$$

Based on the transition probability, the importance score $Score(e_i)$ for an event e_i can be deduced from all other events in a way similar to PageRank [24] algorithm by iteratively computing the following formula until convergence:

$$Score(e_i) = (1 - d) + d \cdot \sum_{e_j \in V_e, e_j \neq e_i} p_{ji} \cdot Score(e_j) \quad (12)$$

where d is a damping factor set by default to 0.85. The computation ends when the difference between the scores computed at the two successive iterations for the events is less than 0.0001.

5.4 Prototype-based HITS

Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that has been successfully used to rate web pages. HITS algorithm defines two types of nodes in a network, hubs and authorities, and computes their ranking scores in a mutually reinforcing way.

We are going to apply HITS for estimating the importance of events (see Fig. 3). We state the following hypotheses for measuring event importance with the prototype-based HITS method:

Hypothesis 7 *An important event is similar to many important eras.*

Hypothesis 8 *An important era is similar to many important events.*

[2]We empirically set the weight for α in Eq. (9) to be 0.6.

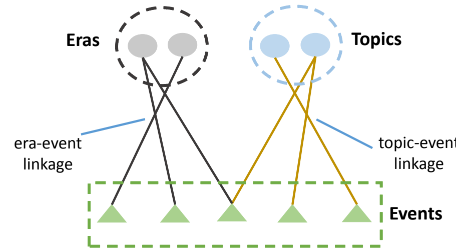

Figure 3: Illustration of the Prototype-based HITS.

Hypothesis 9 *An important event is similar to many important topics.*

Hypothesis 10 *An important topic is similar to many important events.*

Formally, we build a tripartite graph $G = (V_e, V_t, V_z, Q_{et}, Q_{ez})$ with three types of sets of vertices $\{V_e, V_t, V_z\}$ and two types of sets of edges $\{Q_{et}, Q_{ez}\}$. Let V_e=E=$\{e_i\}$, V_t=T=$\{T_j\}$, V_z=Z=$\{z_k\}$ denote the sets of events, detected eras and detected topics, respectively. Let $Q_{et} = \{q_{ij}|e_i \in V_e, T_j \in V_t\}$, $Q_{ez} = \{q_{ik}|e_i \in V_e, z_k \in V_z\}$ represent the set of links between an event and an era, and the set of the links between an event and a topic, respectively.

Era denoting vertices V_t and topic vertices V_z are regarded as hubs, while event vertices V_e are considered as authorities. Each edge q_{ij} in Q_{et} is associated with an affinity weight w_{ij} denoting the similarity between an event e_i and an era T_j, while each edge q_{ik} in Q_{ez} has an affinity weight w_{ik} representing the similarity between an event e_i and a topic z_k. Similarly, w_{ij} and w_{ik} are computed by Eqs. (6) and (7).

In the mutual reinforcement process, the authority scores of events (denoted as $Auth(e_i)$) and the hub scores of eras and topics (denoted as $Hub(T_j)$ and $Hub(z_k)$, respectively) are iteratively computed. The scores at the $(i+1)^{th}$ iteration are calculated based on the scores at the $(i)^{th}$ iteration as follows:

$$Auth^{(i+1)}(e_i) = \sum_{T_j \in T} w_{ij} \cdot Hub^{(i)}(T_j) + \sum_{z_k \in Z} w_{ik} \cdot Hub^{(i)}(z_k) \quad (13)$$

$$Hub^{(i+1)}(T_j) = \sum_{e_i \in E} w_{ij} \cdot Auth^{(i)}(e_i) \quad (14)$$

$$Hub^{(i+1)}(z_k) = \sum_{e_i \in E} w_{ik} \cdot Auth^{(i)}(e_i) \quad (15)$$

Both the authority scores and hub scores are normalized after each iteration in order to achieve convergence. The initial scores of all hubs and the ones of authorities are set to 1. The computation terminates when the difference between the scores computed at the two successive iterations for the hubs and the authorities is less than 0.0001.

6 EXEMPLAR SUMMARY GENERATION

The second type of summarization, the exemplar based summarization approach, assumes selecting a small number of representative entities and constructing the summary upon them. In this section we describe two methods that rely on the selection of the most representative entities.

6.1 Exemplar-based MRW

In the first method we decide the importance of entities using MRW with the following hypothesis:

Hypothesis 11 *An entity is important if it shares history similar to that of other important entities.*

To incorporate this hypothesis into event scoring we again use MRW model. Formally, let $G = (V, Q)$ be an undirected graph, where V=D=$\{d_i\}$ and $Q = \{q_{ij}|d_i, d_j \in D\}$ denote the set of entities (actually, documents representing their histories) and the set of links between entities, respectively. In view of Hypothesis 11, the affinity weight w_{ij} of edge q_{ij} between entities d_i and d_j is computed using the similarity $Sim(d_i, d_j)$.

Since cosine similarity is not a proper similarity measure for sequences such as sequences of events, we propose to use Dynamic Time Warping (DTW) for measuring distances between entities' histories (Eq. (16)). DTW calculates an optimal match between two sequences. Hence, entities' histories can be "warped" non-linearly in the time dimension so as their similar events become aligned. The advantage of DTW is that the order of events is considered when computing the similarity. Thus, histories containing identical events yet, positioned in different order will not be judged identical.

$$\begin{aligned} w_{ij} &= Sim_{DTW}(d_i, d_j) \\ &= \frac{1}{DTW(d_i, d_j) + 1} \end{aligned} \tag{16}$$

The transition probability p_{ij} from d_i to d_j is computed using Eq. (11), and the importance score $Score(d_i)$ for an event d_i is found by iteratively computing Eq. (12) until convergence.

After computing the entity importance scores we select the top m important entities. Let the expected summary size be k events and the number of events in the history of the *i-th* ranked entity d_i be $size(d_i)$. m is then decided as follows:

$$\sum_{i=1}^{m-1} size(d_i) < k, \sum_{i=1}^{m} size(d_i) \geqslant k \tag{17}$$

We next merge the histories of the selected m entities and pick up the top k important events from the merged history using MRW-based ranking method called LexRank [12].

6.2 Exemplar-based HITS

We now propose the last method that represents the exemplar based approach using a bipartite graph framework. In order to calculate document importance with the exemplar-based HITS, we state the following hypotheses:

Hypothesis 12 *An important document is similar to many important events.*

Hypothesis 13 *An important event is similar to many important documents.*

Formally, we build a bipartite graph $G = (V_e, V_d, Q_{ed})$ with two types of vertice sets $\{V_e, V_d\}$ and the set of edges $\{Q_{ed}\}$, where V_e=E=$\{e_i\}$, V_d=D=$\{d_j\}$ denote the set of events and the set of documents, respectively. $Q_{ed} = \{q_{ij}|e_i \in V_e, d_j \in V_d$ represents the set of links between events and documents representing entity histories.

Document vertices V_d are regarded as hubs, while event vertices V_e are regarded as authorities. Each edge q_{ij} in Q_{ed} is associated with an affinity weight w_{ij} denoting the similarity between an event e_i and a document d_j. It is is computed as follows.

$$w_{ij} = Sim(e_i, d_j) = maxSim_{cosine}(e_i, e_j|e_j \in d_j) \tag{18}$$

In the mutual reinforcement process, the authority scores of events and the hub scores of documents are iteratively calculated by Equations (19) and (20). All scores are normalized after each iteration. The initial score of all hubs and authorities are set to 1.

$$Auth^{(i+1)}(e_i) = \sum_{d_j \in D} w_{ij} \cdot Hub^{(i)}(d_j) \tag{19}$$

$$Hub^{(i+1)}(d_j) = \sum_{e_i \in E} w_{ij} \cdot Auth^{(i)}(e_i) \tag{20}$$

After document importance scores are calculated, events from the top m important documents are merged (where m is decided by Eq. (17)). Finally, the top k important events from the merged history are chosen according to ranking by their authority scores.

7 POST-PROCESSING

7.1 Redundancy Removal

After the historical events of a certain category are ranked by importance, we apply a modified version of MMR (Maximal Marginal Relevance) [10] denoted as TMMR (Temporally enhanced Maximal Marginal Relevance) to minimize redundancy. TMMR tries to avoid extracting similar (both semantically similar and temporally close) events in a summary by considering penalty based on the similarity between a newly extracted event and the already extracted events. TMMR allows extracting events which have high importance scores, yet, at the same time, are not semantically similar neither temporally close to the already extracted events.

$$\begin{aligned} \text{TMMR} \equiv \text{argmax}[&\alpha \cdot score(e_i) - \beta \cdot \text{max}Sim_{cosine}(e_i, e_j) \\ &- (1 - \alpha - \beta) \cdot min \frac{1}{|t_i - t_j| + 1}] \end{aligned} \tag{21}$$

Here, e_i denotes an event in the set of the candidate events which have not been selected, while e_j represents an event in the set of the already selected events. t_i and t_j denote the occurrence dates of e_i and e_j. The values of α and β are experimentally assigned to be 0.5 and 0.4, respectively.

7.2 Generalization

Each event in the final summary should be represented by a set of meaningful words. However, our models produce summaries in which each event is in the form of a sentence from the history of particular entity. The sentence representation may then contain too specific details which might be true only for the instance from which the given sentence has been extracted. For example, many cities in Japan have suffered from earthquakes, and, so, the sentence "earthquake hits city" would be a good general description of this type of event, instead of sentences giving detailed descriptions of specific circumstances or effects of earthquakes in particular cities. Thus, we choose to generalize the top-scored sentences to produce the set of descriptive words representing in a general way a given event type (see Tab. 2 for an example).

More concretely, for each sentence indicating an event to be included into the summary, we seek m most similar sentences in the corpus and construct a cluster of $m + 1$ sentences. Sentences within each cluster are semantically similar and each cluster represents the same event type. Then we compute Term Frequency-Inverse Cluster Frequency (TF-ICF) on the created clusters to extract the set of meaningful words describing each cluster (see Eq. (22)) . Those sets of words are used as the final representation of events to be included into the output summary. We set the number m of events used for building the event clusters to be 10.

$$tficf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot log \frac{|C|}{|c : c \ni t_i|} \tag{22}$$

8 EXPERIMENTS

In this section, we describe the experiments conducted to evaluate the effectiveness of our proposed methods.

8.1 Datasets

We test our methods on entities separated by both time and space dimensions. In particular, we perform experiments on 7 Wikipedia categories including 3 city categories and 4 person categories. The city categories are Japanese cities, Chinese cities and English cities (denoted by C_1, C_2, C_3 respectively), while for the person categories we have selected American scientists, French scientists, Japanese Prime Ministers till 1945 (i.e., the end of WW2) and Japanese Prime Ministers after WW2 (denoted by P_1, P_2, P_3, P_4, respectively). Note that our methods are not bound to Wikipedia categories as any listing of entities can form an input, provided the historical data about each instance is made available. In this work, we just use Wikipedia categories as a convenient data source.

For preparing the city categories, each city history is extracted from the "History" section in the corresponding Wikipedia article. To capture historical events, we collect all sentences with dates. As further preprocessing, we reduce inflected words to their word stems and retain only the terms that are among the most frequent 5,000 unigrams, excluding stopwords and numbers. Each historical event is then represented by the bag of unigrams extracted from its sentence along with the corresponding date.

For the person categories, we utilize a dataset of 242,970 biographies publicly released by Bamman *et al.* [4]. Every biography consists of several life events, each represented by bag of unigrams with a date. Unlike in the city datasets, the date here is measured as the difference between the observed date in the event and the date of birth of the entity (i.e., relative date for a person when counting from its birth date). In other words, the date of an event is a relative date here instead of an absolute one as in the city datasets. The basic statistics about our datasets are shown in Tab. 1.

Table 1: Summary of datasets (the time range of datasets C_1, C_2, C_3 are based on absolute time, while that of datasets P_1, P_2, P_3, P_4 are based on relative time.)

Dataset	Category	# Entity	Time Range
C1	Japanese Cities	532	40 - 2016
C2	Chinese Cities	357	12 - 2016
C3	UK Cities	68	1 - 2016
P1	American Scientists	141	0 - 103
P2	French Scientists	41	0 - 101
P3	Japanese PMs (pre WW2)	32	0 - 98
P4	Japanese PMs (post WW2)	30	0 - 93

8.2 Analyzed Methods

We test 4 proposed methods: ***Prototype-based MRW*** (***P-MRW***, see Sec. 5.3), ***Prototype-based HITS*** (***P-HITS***, see Sec. 5.4), ***Exemplar-based MRW*** (***E-MRW***, see Sec. 6.1) and ***Exemplar-based HITS*** (***E-HITS***, see Sec. 6.2). In addition, we set up 2 baselines as follows:

(1) ***LexRank + TMMR (LexRank)*** LexRank [12] method has been widely adopted in multi-document summarization tasks such as [15] and [11]. It constructs a sentence connectivity matrix and computes sentence importance based on the algorithm similar to PageRank. Same as in our methods we also use TMMR to remove redundancy. Finally, selected events are generalized from sentences to the sets of words following the generalization procedure described in Sec. 7.2.

(2) ***k-Means Clustering (K-Means)*** K-Means clustering is a popular method used for cluster analysis. It partitions all events into k clusters in which each event belongs to the cluster with the nearest mean (given k as the size of summary). Then, within each cluster, we pick up 10 sentences which are closest to the cluster centroid in order to build an event cluster. Finally, TF-ICF is applied to extract meaningful words for each event cluster.

8.3 Experiment Settings

We set the parameters as follows:
(1) **size of summary**: we experimentally set the summary size of the city datasets to be 20 events, and of the person datasets to be 10 events considering the sizes of the corresponding categories.
(2) **parameters in the prototype-based methods**: we empirically let the number of eras for the city datasets to be 10, and for the person datasets to be 5 considering the lengths of entity histories. In addition, the number of topics K is set to be identical to the size of summary (K = 10 for person datasets and K = 20 for city datasets).

8.4 Evaluation Criteria

Manually creating summaries of typical histories of categories is a difficult task. We then ask users to evaluate summary quality. To test our methods we conduct evaluation based on five criteria which we believe are crucial for a high quality summary. Each event in the summary is graded in terms of:

- **Saliency** which measures how sound and important the extracted events are.
- **Comprehensibility** which measures how easily the output words can be associated with real events.

Besides, each summary is graded in terms of:

- **Diversity** which measures how diverse the events in the summary are (both semantically and temporally).
- **Coverage** that quantifies the extent to which important events in a category history are included in summary.
- **Interestingness** which measures how interesting the results are. Intuitively, it represents the degree to which the extracted events are novel to annotators.

We have 6 methods to be tested (4 proposed methods and 2 baseline methods). 5 annotators (4 males, 1 female) who have significant interest in history were asked to evaluate 42 different summaries (6 methods, each on 7 datasets). Each summary was ensured to be evaluated by 3 annotators. Thus, 4 annotators were assigned to evaluate 25 summaries and 1 annotator received 26 summaries to evaluate. During the assessment, the annotators were allowed to utilize any external resources including the Wikipedia, Web search engines, books, etc. All of the scores were given in the range from 1 to 5 (1: not at all, 2: rather not, 3: so so, 4: rather yes, 5: definitely yes). After the annotation scores have been completed we averaged saliency and comprehensibility scores per each summary. Lastly, we averaged the individual scores given by the annotators to obtain the final scores per each summary.

8.5 Evaluation Results

Below we discuss the key experimental results.

Average results. Fig. 4 shows the average scores of summaries generated from all the datasets in 5 criteria by all the compared methods. We first note that our proposed methods outperform the baselines based on almost all the criteria (the only exception is that ***E-HITS*** achieves worse performance than ***LexRank*** in terms of coverage by 1.8%). On average, our proposed methods outperform ***LexRank*** by 10.5% and ***K-Means*** by 14.3% across all the metrics. In particular, ***P-MRW*** outperforms ***LexRank*** by 14.6% and ***K-Means*** by 18.5%. Especially, in terms of saliency, the proposed methods are better than ***LexRank*** by 22.4% and than ***K-Means*** by 21.7%. This proves that *incorporating the importance of eras, topics and entities helps to improve the saliency of events contained in summary.*

Furthermore, we note that the two prototype-based methods ***P-MRW*** and ***P-HITS*** are superior to the two exemplar-based methods ***E-MRW*** and ***E-HITS*** by on average 2.9% across all the metrics. On the other hand, MRW-based methods ***P-MRW*** and ***E-MRW*** generate better results than the HITS-based methods ***P-HITS*** and ***E-HITS*** on average by 2.8%. This suggests that *adding information about eras and topics may play more important role than selecting a few representative entities.* Moreover, these results support the conclusion that *event-to-event relationships (utilized in the MRW-based methods) could be more crucial than event-to-entity relationships (used in the HITS-based methods).*

Per dataset results. Fig. 5 shows the summation of evaluation scores in 5 criteria of each method on every dataset. The proposed methods ***P-MRW***, ***P-HITS***, ***E-MRW***, ***E-HITS*** and the two baselines ***LexRank***, ***K-Means*** are denoted by $M_1, M_2, M_3, M_4, M_5, M_6$ in Fig. 5, respectively. We note that all our proposed methods outperform the baselines across all the datasets. In particular, on the Japanese city dataset (C1), the two MRW-based methods ***P-MRW*** and ***E-MRW*** achieve the best performance outperforming the two baselines by 47.8%, while all the proposed methods have better results by on average 40.0% on this dataset. The worst performance is on the French scientists dataset (P2) on which our methods manage to output results better by, on average, only 7.9%.

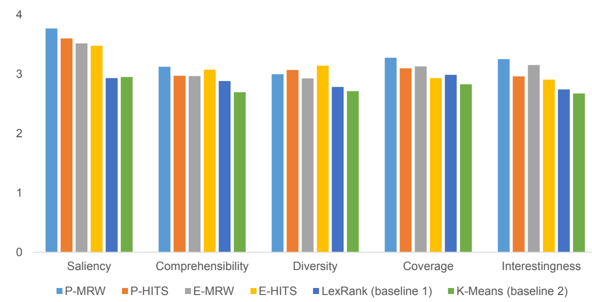

Figure 4: Average Results of All Datasets.

In addition, we note that the average summation scores of the three city datasets are higher than the ones for the four person datasets by 4.4%, which may support the observation that *the quality of summary could be influenced by the number of entities within the category.* Moreover, the average standard deviation of the summation scores by all methods of the city datasets is larger than the one for the person datasets by 16.9%. These both observations suggest that *person related datasets are more difficult and the performance is more uniform across all the methods including the baselines.*

8.6 Additional Observations

Diversity. The reason why the proposed methods work better on the city datasets than on the person datasets could be because city histories have longer time span, hence, their events may be characterized by higher diversity.

Coverage. The prototype-based methods achieve much better performance with regards to the coverage than the exemplar-based methods. It may be because events in exemplar-based summaries are extracted from a small set of typical representatives, which may miss some important information.

Interestingness. MRW-based methods in general outperform HITS-based methods in terms of interestingness. The reason can be due to MRW-based methods incorporating correlation between events and era information, which could make summary more coherent and consistent.

8.7 Example Summary

We present in Fig. 6 and Tab. 6 the summary of a typical history of Japanese cities generated by the prototype-based method **P-MRW**. The summary consists of a timeline containing 20 events ordered chronologically (see Fig. 6), followed by a table (see Tab. 2) which includes up to 15 top scored words representing each event. For every event, we manually create a label based on the words representing the event. In addition, each event is associated with two numbers indicating the median date and the standard deviation of occurrence dates of the corresponding events (these are computed from the event clusters that were discussed in Sec. 7.2).

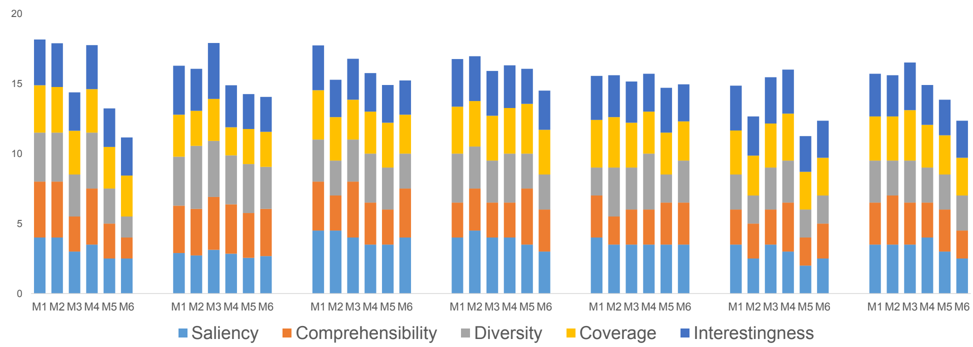

Figure 5: Results on Each Dataset (M1: P-MRW; M2: P-HITS; M3: E-MRW; M4: E-HITS; M5: LexRank; M6: K-Means). The datasets from left to right are C1 (Japanese Cities), C2 (Chinese Cities), C3 (UK Cities), P1 (American Scientists), P2 (French Scientists), P3 (Pre-WW2 Japanese PMs) and P4 (Post-WW2 Japanese PMs).

As we can notice, the ***Meiji Revolution*** is a key turning point in the history of Japanese cities, as most of the events occurred after it. Japanese cities were frequently at war, as reflected by the events: ***Battle*** with median date at around 1333, ***War*** at around 1876 and ***WW2*** at around 1939. The modern ***Education*** in Japan started from the early 20th century. After ***WW2***, Japanese cities enjoyed rapid economic and social development, embodied in the events of ***Population*** (which shows the rapid growth in population), ***Economics*** (which describes the economic boom of Japan in the late 20th century), ***Transportation*** (which reflects the advancement of transportation infrastructure) and ***Film*** (which shows the development of culture industry). Japan cities hosted many ***Sport*** events such as Summer Olympics in the 1960's and Winter Olympics in the 1970's and 1990's. In addition, it can be observed that Japanese cities frequently suffered from ***Natural Disasters*** such as earthquakes, tsunamis and typhoons (e.g. the *Hanshin Earthquake* in 1994), and Japan is paying particular attention to ***Nuclear*** issues (e.g. the *Fukushima Daichi Nuclear Disaster* in 2011). Many of these events can be found in books about Japan history [13] and [20].

9 CONCLUSIONS

It is natural for humans to categorize entities into meaningful groups based on their common traits. One way to better understand categories is by learning histories of their members. In this paper we have introduced a novel type of summarization task consisting in generating gists of histories of multiple entities. We then proposed 4 methods which utilize diverse kinds of signals such as information about documents, eras, topics and correlation between events, and incorporate them into graph-based ranking models. The output summary is in the form of key representative events represented by the sets of meaningful words and approximate event dates. The effectiveness of our models has been demonstrated by the experiments on 7 Wikipedia category datasets.

In the future, we plan to conduct more detailed evaluation on diverse types of entities as well as incorporate abstractive summarization strategies for increasing the readability of the generated summaries. The next step is to improve and extend methods for extracting and representing temporal information from input documents using techniques similar to the one presented in [14].

Acknowledgments This work was supported in part by MEXT Grant-in-Aid (#15H01718, #17H01828), and by the JST research promotion program Presto/Sakigake.

REFERENCES

[1] A. Abdulkareem, and H. Tchalian. Corpus Periodization Framework to Periodize a Temporally Ordered Text Corpus. (2016).
[2] A. Abujabal and K. Berberich. 2015. Important Events in the Past, Present, and Future. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 1315-1320.
[3] R. B. Bairi, M. Carman, and G. Ramakrishnan. On the Evolution of Wikipedia: Dynamics of Categories and Articles. In Ninth International AAAI Conference on Web and Social Media, 2015.
[4] D. Bamman, and N. A. Smith. Unsupervised discovery of biographical structure from text. Transactions of the Association for Computational Linguistics 2 (2014): 363-376.
[5] R. Barzilay, K. R. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, 1999.
[6] D. J. Berndt, and J. Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. KDD workshop. Vol. 10. No. 16. 1994.
[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of machine Learning research 3.Jan (2003): 993-1022.
[8] D. M. Blei, and J. Lafferty. Correlated topic models. Advances in neural information processing systems 18 (2006): 147.
[9] L. R. Brooks. Nonanalytic concept formation and memory for instances. In Cognition and categorization, 1973. Hillsdale.
[10] J. Carbonell, and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998.
[11] P. Flach, and E. T. Matsubara. A simple lexicographic ranker and probability estimator. European Conference on Machine Learning. Springer Berlin Heidelberg, 2007.
[12] E. Gnes, and D. R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research 22 (2004): 457-479.
[13] K. Henshall. A History of Japan: From Stone Age to Superpower. ISBN = 9780230346628, Palgrave Macmillan, 2012.
[14] A. Jatowt, C. M. Au Yeung, and K. Tanaka. Estimating document focus time. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (pp. 2273-2278). ACM, 2013.
[15] P. Kaustubh, and P. Brazdil. Text summarization: Using centrality in the pathfinder network. Int. J. Comput. Sci. Inform. Syst [online] 2 (2007): 18-32.
[16] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5), 604-632, 1999.
[17] K. Knight, and D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. Artificial Intelligence 139.1 (2002): 91-107.
[18] L. Lovsz. Random walks on graphs. Combinatorics, Paul erdos is eighty, 2, 1-46, 1993
[19] R. Lowry. Concepts and applications of inferential statistics. (2014).

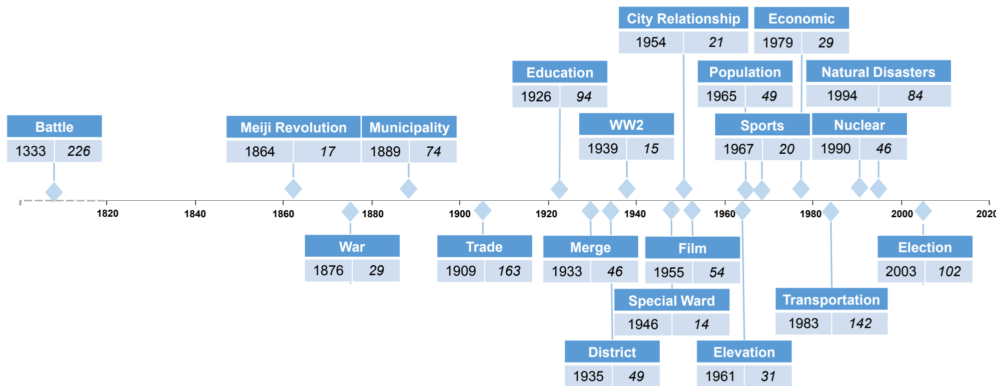

Figure 6: Typical history of a Japanese city learned from 532 instances presented in the form of a timeline. The timeline consists of 20 events ordered chronologically by their median occurrence time. Each event is illustrated by a manually created label along with its median (left value) and the standard deviation of occurrence time (right value).

Table 2: Events in the summary of Japanese cities (see Fig. 6). For each event we show up to top 15 words due to space limit.

Event	Terms
Battle	battle, kamakura, fought, took, kumegawa, area, komaki, period, site, place, war, zenkunen, yasutsune, ultimately, ujigawa
Meiji Revolution	people, peasant, escape, christianity, another, rebellion, damage, raid, air, war, yokkaichi, went, weakened, toyotomi, subsequent
War	war, naval, japan, school, russojapanese, kiyohara, fujiwara, japanese, rebellion, navy, english, end, meiji, major, period
Municipality	system, within, municipality, establishment, modern, created, district, saitama, prefecture, restoration, gunma, town, creation, meiji
Trade	first, tea, made, tsuen, shop, service, held, festival, completed, yoshimitsu, world, waraji, uji, telephone, still
Education	school, established, confucian, high, william, welfare, vories, university, ueshiba, tsujido, teacher, taught, taizen, studies, science
Merge	merged, district, form, create, village, town, tkamachi, numakuma, nakaminato, incorporated, both, neighboring, urasaki, toyosu
District	takikawa, ebeotsu, becomes, continued, tend, village, district, hekikai, town, domain, began, area, period, yamagata, utashinai
WW2	training, center, military, imperial, navy, naval, japanese, industry, facility, built, army, air, production, nagoya, development
Special Ward	ward, became, tokyo, special, founded, city, district, former, shinj, shinagawa, sanbu, sanbe, nine, minamiadachi, metropolis
City Relationship	founded, city, relationship, established, ueno, sister, yamatotakada, wales, tkai, takaishi, sistercity, raised, nanao, mitaka, lomita
Film	year, story, film, festival, shibuya, sakura, record, narita, master, mai, every, appear, place, name, one
Elevation	elevated, city, status, ska, seba, village, town, surrendered, sekigawa, sashima, neighoring, matsumoto, kunitachi, kitaadachi
Population	public, housing, population, real, estate, development, trading, revenue, rapidly, rapid, debt, bubble, large, koku, construction
Sports	olympics, summer, hosted, host, winter, walk, sport, played, park, marathon, events, event, athletics, part, national
Economic	toyota, line, city, opened, nagoya, largest, economic, detroit, aichi, expanded, local, aircraft, became, plant, new
Transportation	expressway, road, line, junction, connected, station, tokaihokoriku, thoku, kaid, highway, established, train, tokyo, opened
Nuclear	fukushima, school, nuclear, evacuee, accident, city, status, student, public, problem, high, caused, rapid, housing, population
Natural Disasters	damage, earthquake, tsunami, thoku, suffered, caused, due, typhoon, rain, flooding, city, isewan, fishing, extensive, although
Election	mayor, motomiya, former, elected, plan, hall, first, party, ochiai, mayoral, kitamura, harue, city, office, woman

[20] J. L. McClain. Japan: A Modern History. ISBN = 9780393041569, W. W. Norton & Company, 2012.
[21] K. McKeown, et al. Towards multidocument summarization by reformulation: Progress and prospects. AAAI/IAAI. 1999.
[22] R. Mihalcea, and P. Tarau. TextRank: Bringing order into texts. Association for Computational Linguistics, 2004.
[23] T. Mikolov, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).
[24] L. Page, et al. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.
[25] D. R. Radev, H. Jing, M. Sty, and D. Tam. Centroid-based summarization of multiple documents. Information Processing & Management, 40(6), 919-938, 2004.
[26] E. Rosch. On the internal structure of perceptual and semantic categories. Cognitive Development and Acquisition of Language, pages 111-114, 1973.
[27] E. Rosch. Cognitive representations of semantic categories. Journal of Experimental Psychology: General, 104:192-233, 1975.
[28] T. A. Tran, C. Niederee, N. Kanhabua, U. Gadiraju, and A. Anand. 2015. Balancing Novelty 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 (CIKM '15). ACM, New York, NY, USA, 1201-1210.
[29] X. Wan, and J. Yang. Multi-document summarization using cluster-based link analysis. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008.
[30] X. Wan, J. Yang, and J. Xiao. Manifold-Ranking Based Topic-Focused Multi-Document Summarization. IJCAI. Vol. 7. 2007.
[31] R. Yan, et al. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011.