DEIM Forum 2017 F3-4

Discovering Typical Histories of Entities

Yijun DUAN†, Adam JATOWT†, and Katsumi TANAKA††

† Graduate School of Informatics, Kyoto University
36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
††
E-mail: †{yijun,adam,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract Categorization is a common solution used for organizing entities like persons or places. For example, there are over 1.13 million categories in Wikipedia which group various types of entities. What is however often lacking is a general, shared information about the entities with a category, for example, information on typical histories of the category entities. We propose in this paper a novel task of automatically creating summaries of shared 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 of entities together with the information on their typical 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 generating contrastive type of knowledge.

Key words entity summarization, digital history, typicality, Wikipedia

1. Introduction

Categorization is a common strategy applied for organizing and understanding entities. Wikipedia, which is considered the most comprehensive encyclopedia these days, contains over 1.13 million categories [1]. Each category typically consists of multiple related members that share some common traits (e.g., list of cities in Japan, list of American scientists born in 19th century, etc.). To obtain a good understanding of a given category, one needs to know much about its members, which is definitely a difficult task especially for larger categories. For example to understand about the category of Japanese cities a user would need to at least read about 500 of Japanese city instances.

In this paper we focus in particular on historical knowledge. Wikipedia abounds in knowledge about entity or concept histories. Nearly every article such as one about a person or place contains history section. In fact, many entities cannot be properly understood without the knowledge of their histories. Same observation applies to categories of these entities.

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 19th century from that of an American scientist? Questions of this type are not easy to be answered as they usually require substantial knowledge of history of much effort.

Straightforward approach to automatically extracting such historical knowledge would be to formulate it as a standard multi-document summarization task. However, traditional multi-document summarization techniques are not suited well for our scenario. The first problem is that the 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 entity 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 a group of Japanese cities, each individual city may have many unique events in its history. Another problem stems from the strong temporal character of 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 [2] 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 [4].

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 common 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. For example, a biochemist's biography is more likely to be about genetics than about x-ray astronomy, and genetics' theme may correlate strongly with genetics within the dataset.

（ 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） It is usual to consider some entities as better representatives of a category than others. This is known as the *graded structure* [5] of a category. An event belonging to a typical entity is then deemed to be more salient than ones of a trivial entity.

To reflect the above observations we rely on graph analysis. Graph based summarization approaches have been successfully used for multi-document summarization. In this work, we adapt Markov Random Walk (MRW) [3] model to our scenarios. To address the limitations of traditional techniques we propose one model for exemplar-based summary generation and another model for prototype-based summary generation which are based on MRW and incorporate additional information about documents, eras, topics and topic correlation. Our experiments are performed on 7 Wikipedia category datasets containing 3 cities datasets and 4 persons datasets with the results demonstrating high 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 their typical histories.

（ 2） We propose 2 different models to discover typical histories of entities utilizing information about sentences, eras, topics and topic correlation between sentences. 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 about cities and persons.

2. Related Work

Our research is related to the following types of tasks:

Multi-Document Summarization. Multi-document summarization is the process of creating a summary that retains the most important information from many documents. Summarization methods can be coarsely divided into extractive summarization or abstractive summarization techniques. As example of extractive methods, the centroid-based method MEAD [2] scores sentences based on sentence-level and inter-sentence features including cluster centroid, position, and TF-IDF, etc. Graph-based ranking methods, such as LexRank [4], 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 [6] is commonly used. Wan *et al.* have improved the graph-ranking algorithm by utilizing sentence-to-sentence and sentence-to-topic relationships [7]. In contrast, abstractive methods create summary containing words not explicitly present in the original. In this process, information fusion [8], sentence compression [9] and reformulation [10] may be needed.

Timeline Summarization. Timeline Summarization defined as the summarization of sequences of documents (typically news articles about the same event), have been actively studied in the recent years. In [11], 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 [12]. His method summarizes timestamped biographies to a set of event classes along with the typical times in a person's life when those events occur.

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

3. Problem Statement

3.1 Input

The input to the summarization task are document containing histories of entities within the same category. Each history-related document spans over a certain range of time

and each sentence refers to some historical event. The dates of events can be either explicitly mentioned in the sentence or can be estimated based on surrounding sentences.

We note that naturally, sometimes categories can consist of entities with very diverse histories. 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 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. Each event in the summary is represented by i 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 category history;

（2） **diverse**: events contained in the summary should be both diverse in their content and in terms of its occurring time to cover more content and time span;

（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* [18] and *exemplar view* [19]. The first one suggests that a category be represented by a constructed prototype (sometimes called centroid), such that entities closer to the prototype are considered to be better examples of the associated 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 summaries 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 uses a few typical representative instances to describe the whole category.

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 [13] was widely utilized for automatically learning the meaning behind the words and the relationships between the words based on neural networks. We use the distributed vector representation to represent terms and events. The vector representation of an event is a weighted combination of vectors of terms contained in the normalized sentence. The corresponding weight for a term is its TF-IDF value calculated from the original corpus.

5. Prototype Summary Generation

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 of 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_b^i of the era. Formally, let $\Theta = (T_1, T_2, ..., T_m)$, and $T_i = [\tau_b^i, \tau_e^i | \tau_b^i \in \xi, \tau_e^i \in \xi]$. 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.*

The above hypotheses are the reason for the two step process of era detection. We discuss both the steps below:

Chi-Square Test. The initial set of unit segments of the category history is $\xi = (t_1, t_2, ..., t_n)$, where each time unit t_i represents a year. A chi-square test of independence is a significance test widely used to determine a significant association between two categorical variables. 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 value for each initial time unit is calculated by the following equation:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (1)$$

The observed value O is the number of events in an initial segment, while the expected value E is found by calculating the average of O in the two adjacent time units. 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.

Optimization. The second 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 initial time units are combined. We next use an optimization formula to determine the final eras based on hypothesis 2. Given the present number of final periods, m, every possible combination Θ of segments from the second set will be explored iteratively. Formally, let $\Theta = (T_1, T_2, ..., T_m)$, and $T_i = [\tau_b^i, \tau_e^i | \tau_b^i \in \varsigma, \tau_e^i \in \varsigma]$. 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.

$$\begin{aligned}\Theta \equiv & \text{argmax}[\omega_1 \sum_{i=1}^{m} IntraSimilarity(T_i) \\ & - \omega_2 \sum_{i=1}^{m-1} InterSimilarity(T_i, T_{i+1}) \\ & + (1 - \omega_1 - \omega_2) Abundancy(T_i)]\end{aligned} \tag{2}$$

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

$$IntraSimilarity(T_i) = \sum_{e_i \in T_i} \sum_{e_j \in T_i} \frac{Sim_{cosine}(e_i, e_j)}{|T_i|^2} \tag{3}$$

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

$$InterSimilarity(T_i, T_{i+1}) = \sum_{e_i \in T_i} \sum_{e_j \in T_{i+1}} \frac{Sim_{cosine}(e_i, e_j)}{|T_i| \cdot |T_{i+1}|} \tag{4}$$

The abundancy of a era indicates how many category instances have at least one of their events located in this era. Let d_i be the instance entity that an event e_i is located in, and D be the set of documents in D. The abundancy is formed as follows:

$$Abundancy(T_i) = |D| \tag{5}$$

Finally, the era combination that has the highest score by applying Eq. (2) is adopted.(注1)

5.2 Topic Detection

Different entities may share similar historical events which are not confined to the same eras. For instance, many Japanese cities once have been hit by earthquakes in different years. Thus, in addition to detecting eras we also conduct topic detection for better representing event importance. Note that the topics are usually not equally important. Events referring to important topics will be deemed more salient than events referring to trivial ones.

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) allows for soft topic to event association, yet, LDA does not explicitly computes topic-topic association which could constitute another useful signal for topic importance estimation.

We construct topics with Correlated Topic Model (CTM) [16], which captures both topic-event relations and 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 assumes a set of latent topics Z=[z_1, z_2, ..., z_K] (K is pre-specified). Each document d_j is viewed 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 $P(Z|d_j)$, the per-topic word distributions $P(W|z_k)$ and the topic-topic correlations $Corr(z_i,z_j)$. We then incorporate this information to event importance calculation, 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. As discussed previously, the underlying correlated topic themes as well as detected 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 less important era or belonging to trivial topics. Thus in order to calculate event importance with the prototype-view based MRW, we state four hypotheses for determining important events:

Hypothesis 3 *An important event is similar to the important era that the event is contained in.*

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 correlated to other important events.*

(注1): We experimentally set weights for ω_1, ω_2 and ω_3 in Eq. (2) to be 0.4, 0.4 and 0.2, respectively.

Formally, let $G = (V_s,\ V_t,\ V_z,\ Q_{ss},\ Q_{st},\ Q_{sz})$ be a graph with three types of sets of vertices $\{V_s,\ V_t,\ V_z\}$ and three types of edge sets $\{Q_{ss},\ Q_{st},\ Q_{sz}\}$. Let $V_s{=}V{=}\{v_i\}$, $V_t{=}T{=}\{t_j\}$, $V_z{=}Z{=}\{z_k\}$ denote the sets of events, detected eras and that of detected topics, respectively. Let $Q_{ss} = \{q_{ij}|v_i, v_j \in V_s\}$, $Q_{st} = \{q_{ij}|v_i \in V_s, t_j \in V_t$ and $t_j = era(v_i)\}$, $Q_{st} = \{q_{ik}|v_i \in V_s, z_k \in V_z\}$ represent the sets of links between events, links between an event and an era, links between an event and a topic, respectively.

Graph G contains two layers. The upper layer consists of era vertices V_t and topic vertices V_z, while the lower layer represents event vertices V_s. Below we are going to explain the way to assign initial scores to vertices and the way to compute edge weights.

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

$$I(t_i) = Sim(t_i, D) \tag{6}$$

$$Sim(t_i, D) = \sum_{v_i \in t_i} \sum_{v_j \in D} \frac{Sim_{cosine}(v_i, v_j)}{|t_i| \cdot |D|} \tag{7}$$

where D is the document set. On the other hand, $I(z_i)$ is an importance of a topic z_i:

$$I(z_i) = \frac{\sum_{d \in D} P(z_i|d)}{|D|} \tag{8}$$

$Sim(v_i, t_j)$ denotes the similarity between an event v_i and the era t_j that v_i is contained in, it is denoted as follows:

$$Sim(v_i, t_j) = \frac{1}{|t_j|} \cdot Sim_{cosine}(v_i, v_j|v_j \in t_j) \tag{9}$$

$Sim(v_i, z_k)$ denotes the similarity between an event v_i and a topic z_k, it is denoted as follows:

$$Sim(v_i, z_k) = \frac{\sum_{w \in v_i} P(w|z_k)}{|v_i|} \tag{10}$$

Let $Score(v_i)$ be the score of a vertex v_i. We then compute the initial score $Score^0(v_i)$ of vertices as follows:

$$Score^0(v_i) = I(t_i) \cdot Sim(v_i, t_i) \cdot \sum_{z_k \in Z} I(z_k) \cdot Sim(v_i, z_k) \tag{11}$$

Each edge q_{ij} in Q_{ss} is associated with an affinity weight w_{ij} between events v_i and v_j. Considering hypothesis 5 and hypothesis 6, the weight is computed using both the similarity $Sim(v_i, v_j)$ and correlation $Corr(v_i, v_j)$ between the two events:

$$w_{ij} = \alpha \cdot Sim_{cosine}(v_i, v_j) + (1 - \alpha) \cdot Corr(v_i, v_j) \tag{12}$$

$$Corr(v_i, v_j) = \sum_{z_i \in Z} \sum_{z_j \in Z} \frac{Sim(v_i, z_i) \cdot Sim(v_j, z_j) \cdot Corr(z_i, z_j)}{|Z|^2} \tag{13}$$

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

$$p_{ij} = \frac{w_{ij}}{\sum_{v_k \in V_s} w_{ik}} \tag{14}$$

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

$$Score(v_i) = (1 - d) + d \cdot \sum_{v_j \in V_s, v_j \neq v_i} p_{ji} \cdot Score(v_j) \tag{15}$$

where d is a damping factor set by default to 0.85. The computation ends when the difference between the scores computed at two successive iterations for any events is less than 0.0001. α in Eq. (12) is empirically set to be 0.6.

5.4 Exemplar-based MRW

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

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

Formally, let $G = (V,\ Q)$ be an undirected graph, where $V{=}\{v_i\}$ and $Q = \ q_{ij}|v_i, v_j \in V\}$ denote the sets of entities and the links between entities respectively.

Considering the above hypothesis regarding entity importance, the affinity weight w_{ij} of edge q_{ij} between the two entities v_i and v_j is computed using the similarity $Sim(v_i, v_j)$ between the histories of v_i and v_j.

Each entity history is a sequence of events. Since cosine similarity is not a proper similarity measure for temporal sequences, we propose to use Dynamic Time Warping (DTW) for measuring similarity between two entities' histories. DTW calculates an optimal match between two sequences. Hence, entities' histories can be "warped" nonlinearly in the time dimension so as their similar events are 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 are not judged to be identical.

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

The transition probability p_{ij} from v_i to v_j is computed using Eq. (14), and the importance score $Score(v_i)$ for an event v_i is deduced by iteratively computing Eq. (15) until convergence.

After entity importance scores are calculated, top m important entities are selected. Let the expected summary size be k events and the number of events in history of the *ith* ranked entity v_i be $size(v_i)$. m is then decided as follows:

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

We next merge 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 [4].

6. Post-Processing

6.1 Redundancy Removal

After historical events of a certain category are ranked by importance, we apply a modified version of MMR (Maximal Marginal Relevance) [6] denoted as T-MMR (Temporal-Maximal Marginal Relevance) to minimize redundancy. T-MMR 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. T-MMR allows extracting the events which have high importance score and are not similar to the already extracted ones.

$$\begin{aligned}\text{T-MMR} \equiv &\text{argmax}[\alpha score(s_i) - \beta \text{max} Sim_{cosine}(s_i, s_j) \\ &- (1-\alpha-\beta) min \frac{1}{|t_i - t_j| + 1}]\end{aligned} \tag{18}$$

Here, s_i denotes a sentence in the set of candidate sentences which have not been extracted, while s_j represents a sentence in the set of already extracted sentences. The values of α and β are experimentally assigned to be 0.5 and 0.4, respectively.

6.2 Generalization

Each event in the summarized timeline should be represented by a set of meaningful words. However, our models only produce summary in which each event is in the format of a certain sentence from the history of one entity. The sentence representation may in addition contain too specific details which are true only for the city from which the sentence has been extracted. For example, many cities in Japan have suffered from earthquakes, and the sentence "earthquake hit city" would be a good general description instead of longer sentences giving additionally detailed descriptions of particular damages in particular cities. Thus we choose to generalize those sentences in summary to a set of more representative words referring to the same event.

More concretely, for each sentence indicating an event in the summary, we seek for m most similar sentences in the corpus and build up a cluster of $m+1$ sentences. Sentences within each cluster are semantically similar and each cluster represents an event. Then we run Term Frequency-Inverse Cluster Frequency (TF-ICF) on clusters and extract a set of meaningful words for each cluster. Those sets of words are used as final representation of events in summarized timeline. We set the number of events used for building the event clusters to be 10.

$$tficf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot log \frac{|C|}{|c : c \ni t_i|} \tag{19}$$

7. Experiments

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

7.1 Datasets

In this study, we validate 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 use Wikipedia categories as a convenient data source.

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

For person categories, we utilize a dataset of 242,970 biographies publicly released by Bamman *et al.* [12]. Each biography consists of several life events represented by bag of unigrams with a timestamp in a way similar to ours. Contrastingly, the timestamp is measured as the difference between the observed date in the event and the date of birth of the entity, namely age. In other words, the timestamp of an event can be either absolute time or relative time. Both

types make sense in different cases.

Some basic statistics about our datasets is shown as follows.

Table 1 Summary of datasets (The time range of datasets C_1, C_2, C_3 are based on absolute time, while 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

7.2 Baselines

We describe 2 baselines as follows:

(1) ***LexRank + MMR(LexRank)*** LexRank methods has been widely adopted in multi-document summarization task. It constructs a sentence connectivity matrix and computes sentence importance based on the algorithm similar to PageRank. We also use MMR to remove redundancy. Finally, selected events are generalized from sentences to the sets of words.

(2) ***K-Means Clustering(K-Means)*** K-Means clustering is a popular method for cluster analysis in data mining. 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 experimentally pick up 10 sentences closest to the cluster centroid to build up an event cluster. Then TF-ICF is applied to extract meaningful words for each event cluster.

7.3 Experiment Settings

We set the parameters as follows:

(1) size of summary: we experimentally set the summary size of city datasets to be 20 events, and of person datasets to be 10 events considering the size of categories.

(2) parameters in the prototype-based methods: we empirically let the number of eras for the city dataset to be 10, and for the person datasets to be 5 considering the length of the span of category history. In addition, the number of topics are experimentally set to be identical to the size of summary.

7.4 Evaluation Criteria

Manually creating summaries of typical histories of certain categories is a difficult task. Thus to text our methods we conduct a qualitative evaluation based on the following five criteria.

Each event in the summary is graded in terms of:

Saliency. It measures how sound and important the extracted events are.

Comprehensibility. It measures how easily the bag of words representing the event can be associated with real historical events.

Besides, each summary is graded in terms of:

Diversity. It measures how varying or diverse the events in the summary are, both semantically diverse and temporally diverse should be taken into consideration.

Coverage. It measures the extent to which important events in category history are included in summary.

Interestingness. It measures how interesting the results are and, implicitly, it represents the degree to which the extracted events were novel to annotators.

We have 4 methods to be tested (2 proposed methods and 2 baseline methods). The annotators were asked to evaluate the 28 different summary (4 methods, each with 7 datasets). We have 5 annotators in total (4 males, 1 female) who have significant interest in history. Each summary is ensured to be evaluated by 3 annotators. During the assessment, the annotators were allowed to utilize any external resources, such as 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 annotation scores have been completed we average saliency and comprehensibility scores per each summary. Lastly, we average individual scores given by annotators to obtain final scores per each summary.

7.5 Evaluation Results

Below we discuss the key experimental results.

Average results. Fig. 1 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 criteria. On average, our proposed methods outperform ***LexRank*** by 12.0% and ***K-Means*** by 15.8% across all the metrics. Especially, in terms of saliency, the proposed methods are better than ***LexRank*** by 24.1% and than ***K-Means*** by 23.4%. This proves that incorporating importance of eras, topics and entities all help to improve the saliency of events contained in summary.

7.6 Additional Observations

we have some additional observations as follows.

Diversity. We find that the proposed methods work better on the city datasets than on the person datasets. It may be because city histories have longer time span hence their events may be characterized by higher diversity.

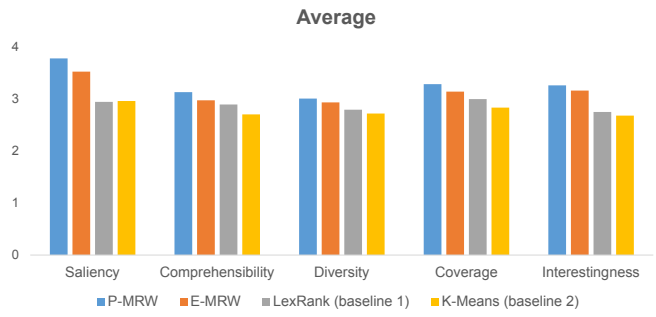

Figure 1 Average Results of All Datasets.

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

8. 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 multiple entity histories. We then proposed 4 methods for this task which utilize diverse kinds of information such as information about documents, eras, topics and correlation between events and incorporate them into graph-ranking models. The output summary is in the form of key representative events represented by sets of meaningful words and approximate event dates. The effectiveness of our models has been demonstrated by experiments on 7 Wikipedia category datasets.

In the future, we plan to incorporate abstractive summarization strategies for improving 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 by [17]. We will also investigate differences of entity types in terms of their impact on the summary quality.

References

[1] Bairi, R. B., Carman, M., & Ramakrishnan, G. (2015, April). On the Evolution of Wikipedia: Dynamics of Categories and Articles. In Ninth International AAAI Conference on Web and Social Media.

[2] Radev, D. R., Jing, H., Sty, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing & Management, 40(6), 919-938.

[3] Lovsz, L. (1993). Random walks on graphs. Combinatorics, Paul erdos is eighty, 2, 1-46.

[4] Erkan, Gnes, and Dragomir R. Radev. "LexRank: Graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research 22 (2004): 457-479.

[5] E. Rosch. On the internal structure of perceptual and semantic categories. Cognitive Development and Acquisition of Language, pages 111144, 1973.

[6] Carbonell, Jaime, and Jade 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.

[7] Wan, Xiaojun, and Jianwu 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.

[8] Barzilay, Regina, Kathleen R. McKeown, and Michael 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.

[9] Knight, Kevin, and Daniel Marcu. "Summarization beyond sentence extraction: A probabilistic approach to sentence compression." Artificial Intelligence 139.1 (2002): 91-107.

[10] McKeown, Kathleen, et al. "Towards multidocument summarization by reformulation: Progress and prospects." AAAI/IAAI. 1999.

[11] Yan, Rui, 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.

[12] Bamman, David, and Noah A. Smith. "Unsupervised discovery of biographical structure from text." Transactions of the Association for Computational Linguistics 2 (2014): 363-376.

[13] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[14] Berndt, Donald J., and James Clifford. "Using Dynamic Time Warping to Find Patterns in Time Series." KDD workshop. Vol. 10. No. 16. 1994.

[15] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[16] Blei, David, and John Lafferty. "Correlated topic models." Advances in neural information processing systems 18 (2006): 147.

[17] Jatowt, A., Au Yeung, C. M., & Tanaka, K. (2013, October). Estimating document focus time. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (pp. 2273-2278). ACM.

[18] E. Rosch. Cognitive representations of semantic categories. Journal of Experimental Psychology: General, 104:192233, 1975.

[19] L. R. Brooks. Nonanalytic concept formation and memory for instances. In Cognition and categorization, 1973. Hillsdale.