Event Digest: A Holistic View on Past Events

Arunav Mishra*
*Max Planck Institute for Informatics
Saarbrücken, Germany
{amishra,kberberi}@mpi-inf.mpg.de

Klaus Berberich*,†
†htw saar
Saarbrücken, Germany
klaus.berberich@htwsaar.de

ABSTRACT

For a general user, easy access to vast amounts of online information available on past events has made retrospection much harder. We propose a problem of automatic event digest generation to aid effective and efficient retrospection. For this, in addition to text, a digest should maximize the reportage of time, geolocations, and entities to present a holistic view on the past event of interest.

We propose a novel divergence-based framework that selects excerpts from an initial set of pseudo-relevant documents, such that the overall relevance is maximized, while avoiding redundancy in text, time, geolocations, and named entities, by treating them as independent dimensions of an event. Our method formulates the problem as an Integer Linear Program (ILP) for global inference to diversify across the event dimensions. Relevance and redundancy measures are defined based on JS-divergence between independent query and excerpt models estimated for each event dimension. Elaborate experiments on three real-world datasets are conducted to compare our methods against the state-of-the-art from the literature. Using Wikipedia articles as gold standard summaries in our evaluation, we find that the most holistic digest of an event is generated with our method that integrates all event dimensions. We compare all methods using standard Rouge-1, -2, and -SU4 along with Rouge-NP, and a novel weighted variant of Rouge.

CCS Concepts

•**Information systems → Retrieval tasks and goals; Summarization;** *Probabilistic retrieval models; Information retrieval diversity;*

Keywords

Linking; Event digest; Diversification; Semantic annotations

1. INTRODUCTION

Today, in this era of digitization, the World Wide Web plays an integral role as an effective and efficient digital medium for providing information on events of global as well as local importance. Large volumes of online news data are generated by media houses and other independent providers as they report eagerly on current events or those that have happened in the past. Contributing to the volume, variety, and velocity of the data, social media is also proving to be a new popular medium of news propagation across the globe. On one hand, this change from traditional print media to publishing online news has given rise to less polarizing and more democratic journalism. On the other hand, from the perspective of a general user, vast amounts of information with a high degree of redundancy have made it difficult to connect the dots and get a holistic understanding on past events with large ramifications.

State-of-the-art vertical news search engines, like Google News, are among the first choices of a general user when seeking information on past events. However, these search engines are keyword based and retrieve a large ranked list of news articles, all of which are temporally biased to the query issuing time. It is hard for a user to sift through all retrieved news articles so as to get a holistic view on a past event. For such an information need, it would be useful if a system could automatically generate an *event digest* by extracting text from retrieved news articles. With such a digest given, the user can first get a broader view on the event and then, if desired, refer to individual documents to get necessary details. A concrete example of an event digest is given in Table 1. For further motivation consider this scenario: A journalist, Laura Lang, wants to quickly get a holistic view on the event of East Timor's independence, illustrated in Table 1. She uses Google News and issues the keyword query {*East*, *Timor*, *votes*, *independence*, *Indonesia*, *referendum*}. Not to her surprise, she finds that the system retrieves numerous (more than one thousand) news articles published by different news agencies. To get a good understanding of the event, she tediously sifts through many articles, most of which contain redundant information. However, with a concise digest given, she can first get an overview of the event, and then jump into news articles connected to the *excerpts* in the digest to get necessary details.

One plausible solution to the information overloading problem is to link orthogonal sources of information on past events [20, 22, 25]. With a similar goal, in a recent work [19, 20], we investigated a linking task that leveraged semantic annotations to identify relevant news articles that can be linked to excerpts from Wikipedia articles. We motivate that Wikipedia articles summarize past events by often abstracting from fine-grained details, and on the other hand, online news are published as the events happen and cover all angles with necessary details. Individually, they both fall short in providing a full picture due to context missing from news articles, and fine-grained details omitted from Wikipedia articles. However, connections can facilitate navigation between them and help in getting a larger picture. One drawback of such a linking task is that for an excerpt from Wikipedia that represents an event with long ramifications, like the one in our example, many news articles get linked

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2911526

Figure 1: Example of an event digest.

Event Query

Description: East Timor votes for independence from Indonesia in a referendum
Time: August 30, 1999
Geolocations: 1643084; 1623843; 7289708;
Entities: East_Timor; Indonesia; 1999_East_Timorese_crisis;

Event Digest (with chronological ordering on publication dates)

• **Publication Date:** *July 20, 1999* **Source Link:** *http://goo.gl/rJYDiZ*
(1) Indonesia is preparing to relinquish control of East Timor after 23 years of occupation and it believes that independence advocates are highly likely to win a referendum next month says an authentic internal government report that has been made available to reporters by advocates of independence. **(2)** Late next month estimated 400 000 East Timorese are to choose between broad autonomy within Indonesia option 1 or independence option 2.

• **Publication Date:** *August 29, 1999* **Source Link:** *http://goo.gl/Cz6Jkk*
(3) Former president Jimmy Carter whose human rights and diplomacy organization the Carter Center is monitoring the referendum here said this this month some top representatives of the government of Indonesia have failed to fulfill their main obligations with regard to public order and security.

• **Publication Date:** *November 21, 1999* **Source Link:** *http://goo.gl/hdqYm8*
(4) The last time it was East Timor which voted for independence from Indonesia in August only to be plunged into a spasm of violence that required an Australian led international military force to quell it. **(5)** Acehs latest push for independence began with the fall of President Suharto in May 1998 and accelerated after the East Timor referendum.

• **Publication Date:** *September 24, 2000* **Source Link:** *http://goo.gl/AijWVY*
(6) East Timor has been under a transitional United Nations administration since the Aug. 30 independence vote last year. **(7)** The groups pillaged East Timor after last year's independence vote which freed the territory from military control.

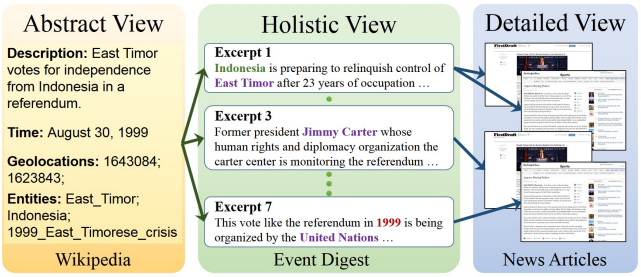
• **Publication Date:** *August 24, 2001* **Source Link:** *http://goo.gl/EAGBxC*
(8) This vote like the referendum in 1999 is being organized by the United Nations which has continued to administer East Timor a former Portuguese colony annexed by Indonesia as it struggles to its feet economically and politically.

to it. An event digest in such a case becomes an intermediate level of linking that presents a *holistic view*. Excerpts from Wikipedia, an *abstract view*, are connected to excerpts in the digest which are in-turn connected to news articles that give a *detailed view* as illustrated in Figure 2. As other use cases, since event digests are generated with a fixed length, smaller digests can be considered as *sneak peaks* (snippets) into news articles, retrieved as a search result. Longer digests can be treated as automatically generated reports for deeper analysis of an event's ramifications.

In this paper, we propose to address the following problem: *given an event from Wikipedia along with a time interval indicating its occurrence period, automatically generate a digest that presents a holistic view on the event.* As input, we consider: **1)** an *event query* that comes with a short textual description, and a time interval indicating when the event happened; and **2)** a set of textually *pseudo-relevant documents* retrieved using a standard retrieval model with a keyword query generated from the event description. As output, our goal is to return a diverse set of excerpts from news articles to compose an *event digest* with its total length under a given length budget such that it presents a holistic view on the event in the query.

Traditional multi-document extractive summarization tasks [4, 12, 14, 16, 18, 21, 31] focus on generating textual summaries from filtered relevant documents such that they are as close as possible to a manually created summary. Unsupervised methods in this realm, consider only text to maximize relevance and reduce redundancy in the generated summaries. However, we define an event to be a joint distribution over independent *text*, *time*, *geolocation*, and *entity* dimensions, indicating the time period, geographic locations, and entities affected by its ramifications. To present a holistic view on an event, we motivate that relevant information along all four dimensions has to be diversified. For the example in Figure 1, diversifying across time will result in information on *causality* (excerpt 1 and 2), *effects during happening* (excerpts 3 to 5), and *after-math* (excerpts 6 to 8) of the event. Similarly, diversifying across geolocations will give information on the entire geographical scope, and diversifying across entities will give information on all persons, places, and organizations involved. We refer to such a view as an *event digest* that contrasts from a traditional notion of a summary.

Figure 2: Different views on a past event.



Challenges. Leveraging text, time, geolocation, and entity dimensions of an event to automatically generate an event digest becomes a challenge. Further, we note that the event descriptions are verbose. Thus, it becomes a challenge to deal with verbosity to select relevant excerpts into the digest.

Contributions made by this paper are the following: **1)** we propose the new problem of event digest creation. **2)** We present a novel method that uses a *divergence-based framework*, and formulates the problem as an integer linear program (ILP) to perform global inference for the event digest creation. To the best of our knowledge, we are the first to present a *unified method* to explicitly diversify across text, time, geolocations, and entities using query modeling approaches. **3)** We present an experimental evaluation on three real-world datasets by treating Wikipedia articles central to an event query as a gold standard.

Organization. In Section 2 we review the literature; Section 3 gives details of our approach. Conducted experiments and their results are described in Section 4. We conclude in Section 5.

2. RELATED WORK

We contrast the event digest generation problem defined in this paper from prior works along the following five lines.

Extractive Summarization focuses on selecting sentences from a single or multiple input documents to create a summary. This line of research has received much attention in the past [4, 12, 14, 16, 18, 21, 31]. It was also investigated at the Document Understanding (DUC) and Text Analysis Conferences (TAC). From various subclasses of extractive summarization, we identify three that seem to be most related to our task: *multi-document summarization*, *query focused multi-document summarization*, and *timeline generation*. In the realm of unsupervised summarization techniques, MMR [4] stands as the most popular approach that defines an objective function rewarding relevance and penalizing redundancy. McDonald et al. [18] proposed an ILP formulation with a slight change to the original MMR objective function. Litvak and Last [16], and Riedhammer et al. [21] proposed to use key phrases to summarize news articles and meetings. Gillick et al. [8] maximized the coverage of

the salient terms in input documents to generate summaries. However, in the problems investigated by all above mentioned works, there is no notion of a user query. This stands as a difference to our problem where we have to generate a digest for a given event query. Further, we incorporate additional semantics to identify informational excerpts for the digest. However, in our approach, we incorporate the formulation given by Riedhammer et al. [21] to develop our text-only method.

Query-Focused Multi-Document Summarization takes into consideration a topic that is input as a user query to generate a topic-focused summary. For this task, supervised approaches have recently proved to be effective [12, 14, 31]. However, they require labeled data for training. Firstly, these approaches focus on short queries, like TREC adhoc topics used in TAC, whereas in our problem, event queries are verbose textual descriptions of events. Secondly, we present an unsupervised approach that formulates an ILP for event digest generation.

Timeline Generation as a subclass that focuses on events, has also received attention [1, 5, 28]. The main goal is to generate a time-stamped list of updates as sentences, key phrases, etc., covering different facets of an event. As an early approach, Allan et al. [1] proposed clustering-based approaches on entities and noun phrases to generate a timeline for a given event. Chieu et al. [5] leveraged burstiness as a ranking metric to identify sentences to be included into a timeline. McCreadie et al. [17] proposed an incremental update summarization task and presented supervised methods to address it. Recently, in a different direction, Shahaf et al. [22] addressed the information overloading problem by presenting a map of connected news articles that captures the story development of a given event. Timeline generation and incremental update summarization tasks aim at presenting a concise ordered summary of events. This is different from our task in two ways: firstly, we do not focus on the ordering of the excerpts in a digest; and secondly, we focus on generating a holistic view by explicitly diversifying across different dimensions of a past event to aid retrospection.

Search Result Diversification problem originally aimed at identifying documents from a relevant set that catered to different information needs of a user query. Further, we look into prior novelty-based strategies to diversify search results. MRR [4] is among the first formulations that penalized documents based on redundancy. This was extended by Zhai et al. [30] for language models and they proposed a risk minimization framework to diversify search results. Wang et al. [26] proposed a mean-variance analysis (MVA) diversification objective. A recent work that becomes interesting is presented by Dou et al. [6] as they attempt to diversify across multiple implicit sub topics by treating them as dimensions of the query. All the methods above cater only to text, and extending them to time, geolocations, and entity dimensions is not straightforward.

Passage Retrieval tasks have been well studied in the past. Systems retrieving passages have been proven to be effective for IR tasks when the documents are long or contain diverse topics. One popular way to define a passage is based on the document structure [3, 9, 23]. Another example of passages are windows consisting of a fixed number of words. These can be further classified into overlapping [3] or non-overlapping windows [13]. The traditional passage retrieval tasks do not take diversity of the passages into consideration. However, we find that the definitions of the passages to be complementary to our excerpts.

3. APPROACH

We present our approach to create a concise digest for a given event that presents a holistic view by describing as many aspects as possible. Intuitively, while selecting excerpts from the input documents, if the reportage of time, geolocations, and entities associated with the input event is maximized then a holistic view can be developed. We propose a novel *divergence-based framework* for event digest creation. Under this framework, our method estimates independent query and excerpt models, and maximizes the relevance while avoiding inter-excerpt redundancy based on the KL-divergence between the models. We define an event as a joint distribution over text, time, geolocation, and entity dimensions. Our method extends the divergence-based retrieval framework, and formulates a single *unified* linear problem to perform global inference across the event dimensions. We design an ILP with appropriate binary indicator variables and constraints.

3.1 Definitions

We begin by defining our notations and representations.

Event Query q is derived from a given Wikipedia event that comes with a short textual description and a time interval indicating its occurrence period. We assume that an event is a joint distribution over four independent dimensions: text, time, geolocation, and entity. Thus, from a given query we derive the following four parts from the textual description: query-text part q_{text} as a bag of textual terms; query-time part q_{time} as a bag of explicit temporal expressions; query-space part q_{space} as a bag of geolocations; and query-entity part q_{entity} as a bag of entity mentions.

Excerpt ε is a single unit of an input document that gives information on an event. In this work, we fix an excerpt as a single sentence, however other definitions may be adopted depending on the application. Analogous to the query, each excerpt has four parts: text ε_{text}, time ε_{time}, geolocation ε_{space}, and entity ε_{entity} part.

In our method, we sometimes use the entire collection as a single coalesced document and refer to its corresponding parts as C_{text}, C_{time}, C_{space}, and C_{entity}.

Time dimension is modeled as a two-dimensional space $T \times T$, as proposed by Berberich et al. [2]. We normalize a temporal expression to an interval $[tb, te]$ with begin time tb and end time te. Further, each interval is described as a quadruple $[tb_l, tb_u, te_l, te_u]$ where tb_l gives the lower bound, and tb_u gives the upper bound for the begin time tb of the interval. Analogously, te_l and te_u give the bounds for the end time te. A *time unit* or *chronon* t indicates the time passed (to pass) since (until) a reference date such as the UNIX epoch. We fix the granularity of a time unit to a single day.

Geolocation dimension is modeled using the geodetic system in terms of $latitude \times longitude$. A geolocation s is represented by its minimum bounding rectangle (MBR). Each MBR is described as quadruple $[tp, lt, bt, rt]$, where point (tp, lt) is the top-left corner and (bt, rt) is the bottom-right corner of the MBR. A geolocation unit g refers to a geographical point in our two-dimensional grid. Further, we empirically set a minimum resolution $resol_{lat} \times resol_{long}$ of the grid to smooth out noisy annotations at a very high granularity (like streets and avenues).

Entity e refers to a location, person, or organization. Our entity dimension represents all entities in the YAGO2 [10] knowledge base. We use the YAGO URI of an entity, as its unique identifier while estimating query- and excerpt-entity models.

3.2 Query and Excerpt Models

The divergence-based framework with the independence assumption between the dimensions allows us to estimate the corresponding query models uniquely. For this, we first expand the original parts of a query (other than text) with the given input set of the documents, thus treating them as pseudo-relevant. Intuitively, by expanding the query parts, we cope with overly specific annotations in the original query. We refer to our prior work [20] for more detailed explanation. We estimate the independent query models from corresponding expanded parts as follows.

Query-Text Model Q_{text}. Query modeling for text is a well-studied problem. The main intuition is that the query-text model should capture the true intent of the query in the text dimension. In our approach, we treat the set of excerpts R in the input documents as pseudo-relevant, and estimate a feedback model. We then combine the feedback model with the empirical query model, estimated from q_{text}, to boost salient terms for the event in the query. Since the best way combining is through linear interpolation [29], we define the generative probability of a term $\mathcal{W}$ as,

$$P(\mathcal{W} \mid Q_{text}) = (1-\theta) \cdot P(\mathcal{W}|q_{text}) + \theta \cdot \sum_{\varepsilon \in R} P(\mathcal{W}|\mathcal{E}_{text}). \quad (1)$$

A term $\mathcal{W}$ is generated from the feedback model with θ probability and from the original query with $(1-\theta)$ probability. Since we use a subset of the available terms, we finally re-normalize as,

$$\hat{P}(\mathcal{W} \mid Q_{text}) = \frac{P(\mathcal{W} \mid Q_{text})}{\sum_{\mathcal{W}' \in V} P(\mathcal{W}' \mid Q_{text})} . \quad (2)$$

Query-Time Model Q_{time} can be understood as a probability distribution in our time domain $T \times T$ that captures the true temporal scope of an input event. We assume that the time part q_{time} of a query is sampled from Q_{time}. The generative probability of a time unit t from the query-time model Q_{time} is estimated by iterating over all the time intervals $[tb, te] \in q_{time}$ as,

$$P(t \mid Q_{time}) = \sum_{[tb,te] \in q_{time}} \frac{\mathbb{1}(t \in [tb_l, tb_u, te_l, te_u])}{|[tb_l, tb_u, te_l, te_u]|} \quad (3)$$

where $\mathbb{1}(\cdot)$ function returns 1 if there is an overlap between a time unit t and interval $[tb_l, tb_u, te_l, te_u]$. If a time unit overlaps with a time interval then we add a probability mass proportional to the inverse of the interval's area in the space. Intuitively, this assigns higher probability to time units that overlap with a layer of specific (smaller area) intervals in q_{time}. For computation of areas and intersections of temporal intervals we refer to [2]. To handle *near misses* [20] perform an additional two-dimensional Gaussian smoothing that blurs the boundaries of Q_{time} by spilling some probability mass to adjacent time units. With this, the new generative probability is estimated as,

$$\hat{P}(t \mid Q_{time}) = \sum_{t \in T \times T} G_\sigma(t) \cdot P(t \mid Q_{time}) \quad (4)$$

where G_σ denotes a 2-D Gaussian kernel that is defined as,

$$G_\sigma(t) = \frac{1}{2\pi\sigma^2} \cdot exp\left(-\frac{(tb_l, tb_u)^2 + (te_l, te_u)^2}{2\sigma^2}\right) .$$

Finally, we re-normalize similar to Equation 2.

Query-Space Model Q_{space}**.** Analogous to time, the query-space model is a probability distribution from which the geolocation part of a given query is sampled. It captures the true geographical scope of the event described in the query. The generative probability of a geolocation unit g from the query-space model Q_{space} by iterating over all $[tp, lt, bt, rt] \in q_{space}$ is estimated as,

$$P(g \mid Q_{space}) = \sum_{(tp,lt,bt,rt) \in q_{space}} \frac{\mathbb{1}(g \in [tp, lt, bt, rt])}{|[tp, lt, bt, rt]|} . \quad (5)$$

The $\mathbb{1}(\cdot)$ functions indicates an overlap between a space unit l and a MBR $[tp, lt, bt, rt]$. Since we normalize with the area of the MBR, a geolocation unit gets higher probability if it overlaps with many specific geolocations (MBR with smaller area) in q_{space}. Area of a MBR, $|[tp, lt, bt, rt]|$ can easily be computed as $(rt - lt + resol_{lat}) * (tp - bt + resol_{long})$. To avoid the issue of near misses we estimate $\hat{P}(l|Q_{Space})$ with additional Gaussian smoothing as described in Equation 4, and finally re-normalize as per Equation 2.

Query-Entity Model Q_{entity} is a probability distribution over our entity space and captures the entities that are salient to the event in a given query. To estimate the Q_{entity} from q_{entity}, we follow a similar process as described for the query-text model, by combining the empirical entity model with a feedback model estimated from the pseudo-relevant excerpt set R. The generative probability of an entity e is estimated as,

$$P(e|Q_{entity}) = (1-\theta) \cdot P(e|q_{entity}) + \theta \cdot \sum_{\varepsilon \in R} P(e|\mathcal{E}_{entity}) \quad (6)$$

where $P(e \mid q_{entity})$ and $P(e \mid \mathcal{E}_{entity})$ are the likelihoods of generating the entity from the original query and pseudo-relevant excerpts $\varepsilon \in R$ respectively. We finally re-normalize as in Equation 2.

Excerpt Model in each dimension is estimated by following a similar methodology as for the query modeling. However, we additionally add Dirichlet smoothing [29] to the excerpt models with the collection C as a background model. For the text dimension, the excerpt-text model $\mathcal{E}_{text}$ is formally estimated as,

$$P(\mathcal{W} \mid \mathcal{E}_{text}) = \frac{\hat{P}(\mathcal{W} \mid \mathcal{E}_{text}) + \mu \cdot P(\mathcal{W} \mid C_{text})}{\mid \mathcal{E}_{text} \mid + \mu} \quad (7)$$

where $\hat{P}(\mathcal{W}|\mathcal{E}_{text})$ is computed according to Equation 1 and μ is set as the average excerpt length of our collection [29]. Similarly, for time, geolocations, and entity models we follow Equation 3, 5, and 6 respectively. However for estimating $\mathcal{E}_{time}$ and $\mathcal{E}_{space}$, we do not employ the Gaussian smoothing (Equation 4) as this tends to introduce additional information into the excerpts artificially.

3.3 ILP Formulation

After estimating necessary query and excerpt models, we next describe our ILP designed for the event digest generation. With our assumptions in mind, we first specify our exact requirements. A digest should portray the following characteristics:

i) contain relevant excerpts to a given event query;
ii) avoid redundancy;
iii) maximize the reportage of the temporal scope, geolocations, and entities;
iv) length in words should not be more than a given budget L.

To design an ILP, we define the following binary indicator variables: S_i indicates if a candidate excerpt ε_i is finally selected into the digest; for a given excerpt ε_i, M_{ij} indicates the single most redundant excerpt ε_j that is already selected into the digest; T_{it} indicates if there is an overlap with $t \in Q_{time}$; G_{ig} indicates if there is an overlap with $g \in Q_{space}$; E_{ie} indicates if there is an overlap with $e \in Q_{entity}$. Using the above definitions, we can now precisely formulate our ILP as illustrated in Algorithm 1.

Algorithm 1 ILP to generate an event digest.

Maximize:

$$\sum_i \Big[\alpha\Big(\lambda \cdot rel_i S_i - (1-\lambda) \cdot \sum_{j \neq i} red_{ij} M_{ij}\Big) + \frac{\beta}{N_t} \sum_{t \in Q_{time}} w_{it} T_{it} + \frac{\gamma}{N_g} \sum_{g \in Q_{space}} w_{ig} G_{ig} + \frac{\psi}{N_e} \sum_{e \in Q_{entity}} w_{ie} E_{ie} \Big]$$

Subject to:

Constraints on text:

1) $\sum_j M_{ij} = S_i \quad \forall\ i$

2) $M_{ij} \leqslant S_i \quad \forall\ i$

3) $M_{ij} \leqslant S_j \quad \forall\ j$

4) $M_{ik} \geqslant S_k - (1 - S_i) - \sum_{j: red_{ij} \geqslant red_{ik}} S_j \quad \forall i \neq k$

Constraints on time:

5) $\sum_i T_{it} \geqslant 1 \quad \forall\ t \in Q_{time}$

6) $T_{it} \geqslant S_i \cdot O_{it} \quad \forall\ i,\ t \in Q_{time}$

7) $T_{it} \leqslant S_i \quad \forall\ i,\ t \in Q_{time}$

Constraints on geolocation:

8) $\sum_i G_{ig} \geqslant 1 \quad \forall\ g \in Q_{space}$

9) $G_{ig} \geqslant S_i \cdot O_{ig} \quad \forall\ i,\ g \in Q_{space}$

10) $G_{ig} \leqslant S_i \quad \forall\ i,\ g \in Q_{space}$

Constraints on entity:

11) $\sum_i E_{ie} \geqslant 1 \quad \forall\ e \in Q_{entity}$

12) $E_{ie} \geqslant S_i \cdot O_{ie} \quad \forall\ i,\ e \in Q_{entity}$

13) $E_{ie} \leqslant S_i \quad \forall\ i,\ e \in Q_{entity}$

Objective Function can be explained as four parts: text, time, geolocation, and entity. In the text part, rel_i function computes the relevance between an excerpt ε_i and query q. Each excerpt is penalized with the maximum textual redundancy score red_{ij} with the already selected excerpts into the digest. The parameter λ balances textual relevance and redundancy estimates.

To explain the rest of the formulation, let us consider the time part in isolation. For each excerpt, the following steps are followed: first, identify the time units that overlap with the given Q_{time}. Second, weights w_{it} of the time units are summed and assigned as temporal scores. The rest of the parts, i.e., space and entity, are handled similar to time. To specify the global importance of the dimensions, four parameters, α, β, γ, and ψ are introduced into the objective function. Finally, we normalize the time, geolocations, and entity parts with the size of their corresponding query models denoted as N_t, N_g, and N_e respectively.

Constraints defined for our ILP are categorized into the four parts corresponding to the objective function. Constraints on text are defined on the binary indicator variable M_{ij}. Constraints 1-3 enforce that exactly one excerpt in the digest is selected for consideration as most redundant. Constraint 4 is to ensure that if $M_{ik} = 1$ then ε_j is most redundant to ε_i and they both are selected into the final digest. Constraints 5, 8, and 11 ensure that each unit in the query model is covered by at least one excerpt in the digest, thus maximizing total coverage of the query units in the digest. Constraints 6, 9, and 12 specify that if an excerpt is selected then overlapping units across the dimensions are covered. For these constraints, we introduce an additional binary variable O_{ik} that indicates if there is an overlap between a kth unit in query with excerpt ε_i model. Finally, constraints 7, 10, and 13 are required as sanity check that a unit can be covered only if an overlapping excerpt is selected.

Textual relevance rel between a given query q and excerpt ε in the objective function is estimated by computing the KL-divergence KLD between their language models, denoted as Q_{text} and $\mathcal{E}_{text}$ respectively. Formally, this is given as,

$$rel_i = -KLD(Q_{text} || \mathcal{E}_{i\ text}) \ . \tag{8}$$

Textual Redundancy red between any two excerpts can be simply interpreted as the similarity between them. For this, we compute the Jensen-Shannon divergence JSD which is the symmetric variant of the KLD and a popularly used distance metric. In this case, lower divergence indicates higher redundancy between the excerpts. Formally, this is defined as,

$$red_{ij} = -JSD(\mathcal{E}_{i\ text} || \mathcal{E}_{j\ text}) \tag{9}$$

Weights w_{it}, w_{ig}, and w_{ie} specify the importance of the time t, geolocation g, and entity unit e respectively for an excerpt ε_i. Under our divergence-based framework, we define the weights as the negative KL-divergence between the generative probability of a unit from the query and excerpt models. Formally, we define weights for all dimensions analogous to time as,

$$w_{it} = -KLD\big(P(t|Q_{time}) || P(t|\mathcal{E}_{i\ time})\big) \ . \tag{10}$$

For a single dimension considered in isolation, summing over the weights gives the overall divergence between excerpt and query models in that dimension. For example, in the time dimension the divergence $KLD(Q_{time} || \mathcal{E}_{i\ time})$ can be computed by summing over query time units as $\sum_{t \in Q_{time}} w_{it} T_{it}$. Intuitively, objective is maximized by globally minimizing the overall divergence of a query with entire digest. However, since the KL-divergence scores are not bounded, we normalize the weights across all excerpts as,

$$\frac{KLD - KLD_{min}}{KLD_{max} - KLD_{min}} \ .$$

In our divergence framework, the redundancies in time, geolocations, and entity dimensions are minimized implicitly by maximizing the coverage of the units in query models. However, at the same time, the relevance of excerpts in these dimensions are also considered. The ILP solver first selects excerpts that cover most important units (receiving high probability in query model) with the lowest divergence scores as indicated by their weights.

4. EXPERIMENTS

Next, we give details on the conducted experiments. For reproducibility, we make the experimental data publicly available[1].

4.1 Setup

We begin by describing our test collections, query set, gold standards, and measures used in the experimental setup.

Test Collections. We perform experiments on three real-world datasets: **1)** The New York Times Annotated Corpus (NYT) with about 2 million news articles published between 1987 and 2007; **2)** English Gigaword corpus with about 9 million news articles published between 1991 and 2010; and **3)** ClueWeb12-B13 (CW12) corpus with about 50 million web pages crawled in 2012. We process the queries in our test set with a standard query likelihood document retrieval model. Top-10 retrieved documents from each dataset are considered pseudo-relevant and input into our methods.

[1] http://resources.mpi-inf.mpg.de/d5/eventDigest/

Test Queries are generated from the *timeline of modern history*[2] in Wikipedia that contains the most prominent news events in the past. We randomly sample 100 events occurring between 1987 and 2007 as test queries. Each query comes with a short textual description and a time interval indicating when the event happened. Further, we automatically annotate each query with time, geolocations, and entities by disambiguating mentions in their descriptions.

Gold Standard. We consider a Wikipedia article that describes the specific event in a query as its human-generated or *gold standard digest* created by Wikipedians. Since these articles on past events are elaborate and cover most of the important aspects, they are apt for evaluating our task. For this, we manually identify Wikipedia articles that are central to an event query.

Measures. We use the following measures:

• *Rouge-1*, *Rouge-2*, and *Rouge-SU4* measures [15] are well established for evaluating method-generated against gold standard (human-generated) summaries.

• *Rouge-NP*: We introduce a new measure that takes into consideration only the noun phrases overlap. Generally, noun phrases represent the key concepts in a gold standard, and a larger overlap indicates better information coverage in the digest. This is further motivated by Taneva et al. [24] in their experimental evaluation.

• *weighted-Rouge (w-Rouge)*: The above rouge measures evaluate how close a method-generated digest is to the gold standard. However, due to the disparate quantity of text between a method-generated digest and the gold standard, these measures are not indicative of the diversity of excerpts in a digest. Thus, we introduce w-Rouge that computes Rouge-1 score of a method-generated digest S with each paragraph p of the corresponding gold standard GS. The individual Rouge-1 scores are weighted with the normalized length of each paragraph $\frac{|p|}{|GS|}$. To get the final score for a method, we average over all queries q in query set QS. Formally,

$$\textit{w-Rouge} = \frac{1}{|QS|} \sum_{q \in QS} \sum_{p \in GS} \frac{|p|}{|GS|} \text{Rouge-1}(S, p) \qquad (11)$$

Additionally, we also report the mean variance ($MVar$) of the w-Rouge across all the queries. Formally, this is given as,

$$MVar = \frac{1}{|QS|} \sum_{q \in QS} \frac{1}{N} \sum_{p \in GS} \left[\textit{w-Rouge}(S, p) - \textit{w-Rouge}(S, GS)\right]^2$$

where N is the number of paragraphs in GS. We assume that in a long Wikipedia article, each paragraph describes an aspect of the central event. Thus, a method-generated summary that gives diverse information should show overlap with more number of paragraphs in the gold standard Wikipedia article. A method that generates a digest that is closer to the gold standard by covering more aspects of the given event should have a higher mean F1 score and mean variance of w-Rouge.

Implementation. All the methods are implemented in Java. For the temporal annotation, we use Stanford CoreNLP toolkit [3]. To annotate geolocations, we use an open-source gazetteer-based tool[4] that extracts locations and maps them to the GeoNames[5] knowledge base. For entity annotations, we use the AIDA [11] system. We use the Gurobi ILP solver[6] in our experiments.

[2]https://en.wikipedia.org/wiki/Timeline_of_modern_history
[3]http://stanfordnlp.github.io/CoreNLP/
[4]https://github.com/geoparser/geolocator
[5]http://www.geonames.org/
[6]http://www.gurobi.com

4.2 Methods

We next describe the different methods that are compared in our experiments. We distinguish three frameworks that use integer linear programs for global inference: **1)** maximum marginal relevance [4, 18, 21], **2)** coverage-based [7, 8, 27], and **3)** divergence-based methods. While the first two are derived from literature as state-of-the-art frameworks for unsupervised methods, the third divergence-based framework is proposed in this paper. We extend the frameworks to incorporate time, geolocations, and entities, and design methods that leverage different combination of the dimensions under each framework. All our methods formulate the event digest problem as solvable ILPs.

Maximum Marginal Relevance [4] is arguably the most popular unsupervised framework for generating document summaries. We compare the following two methods that fall under this framework:

• *Mcd*: As first method, we consider the summarizer presented by McDonald et al. [18] that uses an ILP for global inference in summarization. Though they follow the MMR [4] style formulation, they make a slight change to the global objective function by introducing linear approximation. This results in candidates being penalized with the average redundancy to the already selected excerpts. Their objective is defined as,

$$Maximize : \sum_i \left[\lambda \cdot rel_i S_i - (1-\lambda) \cdot \sum_{j \neq i} red_{ij} S_{ij}\right]. \qquad (12)$$

We refer to [18] for the full set of constraints. The generalized ILP framework allows us to define the rel and red functions using language modeling methods as described in Equation 8 and 9.

• *Rdh:* More recently, Riedhammer et al. [21] propose an ILP formulation that got rid of the linear approximation in the global objective function of *Mcd*, thus giving an optimal solution. In their formulation, they introduce an additional binary variable M_{ij} to indicate the maximum redundancy of an excerpt to the already selected excerpts in the digest. Further, they have additional constraints that are defined on this variable which leads to efficient convergence to the optimal value. Their global objective function is defined as,

$$Maximize : \sum_i \left[\lambda \cdot rel_i S_i - (1-\lambda) \cdot \sum_{j \neq i} red_{ij} M_{ij}\right]. \qquad (13)$$

We refer to [21] for full set of constraints. Similar to *Mcd*, we use the definitions for rel and red as given in Equations 8 and 9.

Coverage-Based Framework [8] is also popular in the summarization community as an unsupervised global inference method. It follows the idea of implicitly reducing the redundancy in the final summary by maximizing the coverage of textual units. Prior works [7, 27] propose various definitions for such units in context of different tasks. This framework remains state of the art, and approaches have shown to work well in comparison to other unsupervised global inference methods. At large, the framework is general enough and can easily be extended to our event dimensions. Their global objective function is defined as,

$$Maximize : \sum_i w_i \cdot C_i \,, \qquad (14)$$

where w_i is defined as $P(c \mid Q_{text})$ probability of generating a term from query-text model, and C_i is a binary indicator variable that marks the occurrence of a term c in an excerpt ε_i. Using this framework, we design the following methods that consider different subsets of the four event dimensions:

• *Cov-txtEM* and *Cov-txtQM*: As text-only methods, *Cov-txt* maximizes the coverage of the salient terms associated with an event. In their original work, Gillick et al. [8] relied heavily on preprocessing the documents to be summarized including key-phrase extraction. For this work, we do not do any preprocessing. In contrast, we make use of a query-text model Q_{text} to capture salient textual terms for the event in the query. We motivate that this makes their method more query-focused and stronger as a baseline. To demonstrate the advantage of incorporating a query model, we compare two methods: *Cov-txtEM* that uses only the empirical terms (after stop words removal), and *Cov-txtQM* that estimates a query model by expanding the query with terms from the pseudo-relevant documents as shown in Equation 1.

• *Cov-T*, *Cov-S*, *Cov-E*, *Cov-ST*, and *Cov-EST*: In principle, the coverage-based method can easily be extended to time, geolocations, and entity dimensions. In the time dimension, we adapt the objective function in Equation 14, such that it selects excerpts by maximizing the global coverage of all time units $\tau \in Q_{time}$. We label this method as *Cov-T*. Similarly, *Cov-S*, and *Cov-E* maximize the coverage of geolocation units $l \in Q_{space}$ and entities $e \in Q_{entity}$ respectively. We motivate that *Cov-E* method is similar to the original approach proposed by Gillick et al. that maximizes concepts, which in our case are entities. It is not hard to think of methods that maximize a combination of the dimensions in their objective functions. *Cov-ST* maximizes the coverage of geolocations and time, and *Cov-EST* additionally combines entities.

In all the above methods, weights w_i as in Equation 14, are generative probabilities from query models described in Section 3.2.

Divergence-Based Framework, discussed in Section 3, takes into consideration the divergence between query and excerpt models in all the dimensions. We note that the text-only method under this framework is equivalent to *Rdh* that defines its rel and red functions in Equation 13 based on the KL-divergence between corresponding text models. We design the following methods:

• *Div-T*, *Div-S*, *Div-E*, *Div-ST*, *Div-txtST*, *Div-EST*, and *Div-txtEST*: In Section 3, we present a unified divergence-based framework that maximizes the textual relevance and minimizes the redundancy across text, time, geolocations, and entities. We label this as the *Div-txtEST* method. However, one can think of methods that consider only a subset of the dimension. We thus design *Div-T*, *Div-S*, *Div-E*, *Div-ST*, *Div-txtST*, and *Div-EST* methods that leverage a combination of text (*txt*), time (*T*), space (*S*), and entities (*E*) as indicated by the suffixes in their labels.

Random. Finally, we consider the *Rand* method that selects excerpts at uniformly random from the input top-10 pseudo-relevant documents until the length constraint is satisfied.

Parameters of the methods were tuned by varying one parameter at a time while keeping the others fixed to observe the change in the overall result quality as described in [28]. We have two groups of parameters, first denoted by λ that balances the relevance rel with redundancy red. Second, that specifies the importance of text, time, geolocations, and entities with parameters, α, β, γ, and ψ respectively. For NYT, Gigaword, and CW12 datasets, we set λ to 0.85, 0.90 , and 0.95 respectively. In the second group of parameters, we tune β, γ, and ψ while fixing $\alpha = 1 - (\beta + \gamma + \psi)$. We empirically observe that setting the three parameters too high leads to a deterioration of the results. For NYT, Gigaword, and CW12, we set $\beta = [0.10, 0.10, 0.10]$, $\gamma = [0.05, 0.01, 0.01]$, and $\psi = [0.01, 0.30, 0.01]$. Finally, for the query models in Equation 1, 3, 5, and 6, we use settings described in [20].

Table 1: Results on The New York Times dataset.

Methods	Rouge 1			Rouge 2			Rouge NP			Rouge SU4		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Rand	0.618	0.073	0.113	0.155	0.014	0.024	0.084	0.007	0.012	0.209	0.025	0.039
Mcd	0.652	0.078	0.121	0.192	0.019	0.030	0.086	0.008	0.013	0.205	0.026	0.039
Rdh	0.662	0.078	0.121	0.203	0.019	0.031	0.088	0.008	0.013	0.210	0.026	0.040
Cov-txtEM	0.652	0.051	0.085	0.180	0.011	0.020	0.091	0.005	0.009	0.219	0.017	0.029
Cov-txtQM	0.646	0.079	0.122	0.190	0.018	0.030	0.086	0.009	0.014	0.204	0.026	0.039
Cov-T	0.544	0.023	0.041	0.188	0.007	0.012	0.079	0.003	0.006	0.182	0.008	0.014
Cov-S	0.464	0.027	0.043	0.154	0.006	0.011	0.058	0.003	0.005	0.149	0.009	0.014
Cov-E	0.666	0.069	0.110	0.215	0.018	0.031	0.076	0.007	0.011	0.213	0.023	0.036
Cov-ST	0.647	0.038	0.062	0.214	0.010	0.017	0.087	0.005	0.008	0.212	0.013	0.021
Cov-EST	0.666	0.073	0.115	0.214	0.019	0.032	0.078	0.008	0.012	0.214	0.024	0.038
Div-T	0.647	0.068	0.110	0.195	0.017	0.028	0.088	0.008	0.013	0.211	0.023	0.037
Div-S	0.653	0.072	0.113	0.199	0.018	0.030	0.090	0.008	0.013	0.212	0.025	0.038
Div-E	0.652	0.077	0.120	0.195	0.019	0.031	0.091	0.008	0.014	0.213	0.026	0.041
Div-ST	0.662	0.081	0.124	0.210	0.020	0.033	0.090	0.008	0.014	0.215	0.027	0.041
Div-txtST	0.667	0.082	0.125	0.214	0.021	0.034	0.090	0.009	0.014	0.214	0.027	0.041
Div-EST	0.649	0.080	0.122	0.196	0.020	0.031	0.087	0.008	0.013	0.211	0.027	0.041
Div-txtEST	0.675	0.084	**0.127**	0.219	0.022	**0.035**	0.089	0.010	**0.016**	0.219	0.028	**0.042**

Table 2: Results on the Gigaword dataset.

Methods	Rouge 1			Rouge 2			Rouge NP			Rouge SU4		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Rand	0.643	0.069	0.109	0.191	0.016	0.027	0.163	0.017	0.028	0.242	0.026	0.041
Mcd	0.663	0.073	0.113	0.204	0.019	0.030	0.191	0.020	0.032	0.255	0.028	0.043
Rdh	0.654	0.075	0.119	0.201	0.019	0.031	0.190	0.021	0.034	0.247	0.028	0.045
Cov-txtEM	0.676	0.052	0.086	0.204	0.013	0.022	0.184	0.012	0.021	0.257	0.020	0.033
Cov-txtQM	0.652	0.078	0.122	0.197	0.019	0.031	0.178	0.021	0.033	0.242	0.029	0.045
Cov-T	0.372	0.015	0.027	0.120	0.004	0.007	0.127	0.005	0.008	0.150	0.006	0.010
Cov-S	0.509	0.021	0.036	0.167	0.005	0.008	0.164	0.006	0.011	0.206	0.008	0.013
Cov-E	0.664	0.070	0.111	0.212	0.018	0.030	0.173	0.017	0.028	0.255	0.027	0.042
Cov-ST	0.553	0.028	0.048	0.169	0.007	0.012	0.180	0.008	0.014	0.214	0.011	0.018
Cov-EST	0.665	0.071	0.112	0.214	0.018	0.030	0.174	0.017	0.028	0.257	0.027	0.043
Div-T	0.650	0.077	0.120	0.198	0.019	0.031	0.191	0.022	0.035	0.244	0.029	0.045
Div-S	0.647	0.071	0.112	0.201	0.017	0.028	0.191	0.020	0.033	0.245	0.026	0.042
Div-E	0.646	0.077	0.121	0.190	0.018	0.029	0.189	0.022	0.035	0.241	0.029	0.045
Div-ST	0.668	0.045	0.074	0.212	0.011	0.019	0.201	0.013	0.021	0.260	0.017	0.028
Div-txtST	0.655	0.081	0.125	0.201	0.020	0.033	0.192	0.023	0.036	0.248	0.031	0.047
Div-EST	0.619	0.074	0.116	0.189	0.018	0.030	0.182	0.022	0.034	0.233	0.028	0.044
Div-txtEST	0.655	0.081	**0.126**	0.205	0.021	**0.034**	0.192	0.022	**0.036**	0.248	0.031	**0.048**

Table 3: Results on the ClueWeb12-B13 dataset.

Methods	Rouge 1			Rouge 2			Rouge NP			Rouge SU4		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Rand	0.604	0.058	0.095	0.173	0.013	0.022	0.135	0.011	0.018	0.248	0.021	0.035
Mcd	0.642	0.058	0.095	0.197	0.014	0.024	0.157	0.011	0.019	0.269	0.021	0.036
Rdh	0.630	0.063	0.101	0.193	0.015	0.026	0.156	0.012	0.020	0.261	0.023	0.038
Cov-txtEM	0.649	0.052	0.084	0.209	0.012	0.022	0.167	0.011	0.018	0.278	0.019	0.032
Cov-txtQM	0.650	0.065	0.105	0.206	0.016	0.028	0.164	0.014	0.022	0.277	0.024	0.040
Cov-T	0.410	0.011	0.020	0.155	0.004	0.007	0.140	0.003	0.006	0.186	0.005	0.009
Cov-S	0.375	0.014	0.026	0.131	0.004	0.008	0.112	0.004	0.007	0.168	0.006	0.010
Cov-E	0.628	0.060	0.097	0.199	0.015	0.026	0.140	0.012	0.020	0.260	0.022	0.036
Cov-ST	0.532	0.019	0.036	0.191	0.006	0.011	0.170	0.005	0.010	0.237	0.008	0.015
Cov-EST	0.634	0.060	0.097	0.203	0.015	0.026	0.146	0.012	0.020	0.265	0.022	0.037
Div-T	0.608	0.070	0.110	0.173	0.016	0.026	0.124	0.011	0.018	0.248	0.024	0.039
Div-S	0.568	0.055	0.092	0.163	0.013	0.022	0.117	0.009	0.015	0.235	0.019	0.033
Div-E	0.617	0.073	0.114	0.174	0.016	0.026	0.119	0.011	0.017	0.253	0.025	0.040
Div-ST	0.654	0.045	0.069	0.215	0.010	0.017	0.171	0.008	0.013	0.284	0.016	0.025
Div-txtST	0.614	0.072	0.113	0.172	0.017	0.027	0.122	0.011	0.019	0.249	0.025	0.041
Div-EST	0.613	0.073	0.113	0.170	0.016	0.026	0.115	0.011	0.017	0.248	0.025	0.041
Div-txtEST	0.632	0.075	**0.117**	0.192	0.018	**0.030**	0.152	0.014	**0.023**	0.264	0.027	**0.044**

4.3 Results

We compare the quality of event digests generated from the three different datasets. We also compare the variance of weighted-Rouge measure that we propose to highlight the diversification effect of each method.

Rouge score analysis. Table 1, Table 2, and Table 3 show the results of generating an event digest of length 250 words from NYT, Gigaword, CW12 documents respectively. We compare the digest generated by different methods against Wikipedia articles as gold standards, and report Rouge-1, Rouge-2, Rouge-NP, and Rouge-SU4 scores. We find that across all the three datasets the best quality digest is generated by the *Div-txtEST* method.

Firstly, we note that the random method *Rand* already achieves a decent F1 score. Across the three datasets, selecting excerpts randomly from the top-10 input pseudo-relevant NYT and Giga-

Table 4: Statistical significance test with two tailed paired t test at three α levels: 0.01 (▲/▼), 0.05 (△/▽), and 0.10(∧/∨); of all against the text-only methods. Three symbols in each cell denote increase (up), decrease (down), or no change (–) in Rouge-1, -2, and -SU4, respectively.

Datasets	Methods	Mcd	Rdh	Cov–txtQM	Cov–T	Cov–S	Cov–E	Cov–ST	Cov–EST	Div–T	Div–S	Div–E	Div–ST	Div–txtST	Div–EST	Div–txtEST
NYT	Mcd	– – –	▲ △ △	– – –	▼ ▼ ▼	▼ ▼ –	▼ – ▽	▼ ▼ –	▼ ∧ –	▼ ▼ –	▼ – –	– – ∧	– ▲ –	▲ ▲ ▲	– – ▲	▲ ▲ ▲
	Rdh	▼ ▽ ▽	– – –	▲ ▽ –	▼ ▼ ▼	▼ ▼ ▼	▼ – ▽	▼ ▼ ▼	▼ – ▽	▼ ▼ ▽	▼ ∨ ▽	▼ ∨ ▲	△ – ▲	– ▲ ∧	▲ – ▲	– ▲ ▲
	Cov–txtQM	– – –	▼ △ –	– – –	▼ ▼ ∨	▼ ▼ –	▼ – ∨	▼ ▼ –	▼ △ –	▼ ▽ –	∨ – –	– – △	– ▲ ∧	▲ ▲ ▲	– – △	▲ ▲ ▲
Giga	Mcd	– – –	▲ ▲ △	▲ △ △	▼ ▼ ▼	▼ ▼ ▼	– – –	▼ ▼ ▼	– – –	▲ – –	– – –	▲ – –	▼ ▼ ▼	▲ ▲ ▲	▲ – –	▲ ▲ ▲
	Rdh	▼ ▼ ▽	– – –	– – –	▼ ▼ ▼	▼ ▼ ▼	▼ – ▼	▼ ▼ ▼	▼ – ▼	– – –	– – –	– – –	▼ ▼ ▼	▲ – △	– – –	▲ ▲ ▲
	Cov–txtQM	▼ ▽ ▽	– – –	– – –	▼ ▼ ▼	▼ ▼ ▼	▼ ▽ ▼	▼ ▼ ▼	▼ – ▼	– – –	▽ ∨ –	– ▽ –	▼ ▼ ▼	– – △	– – –	– ∧ ∧
CW12	Mcd	– – –	▲ ▲ –	▲ ▲ ∧	▼ ▼ ▼	▼ ▼ ▼	– – –	▼ ▼ ▼	– ∧ –	▲ ∧ ▲	∨ – –	▲ ▲ ▲	▼ ▼ ▼	▲ △ ▲	▲ △ ▲	▲ ▲ ▲
	Rdh	▼ ▼ –	– – –	– – –	▼ ▼ ▼	▼ ▼ ▼	▽ – –	▼ ▼ ▼	∨ – –	– – –	– – –	△ – –	▼ ▼ ▼	△ – –	△ – –	▲ ▲ –
	Cov–txtQM	▼ ▼ ∨	– – –	– – –	▼ ▼ ▼	▼ ▼ ▼	∨ ∨ –	▼ ▼ ▼	∨ – –	△ – –	– ∨ –	▲ – –	▼ ▼ ▼	▲ – –	▲ – –	▲ △ ∧

Table 5: Comparison of methods in Weighted-Rouge1.

	NYT			GIGA			CW12		
	F1	**MVar**	**MSD**	**F1**	**MVar**	**MSD**	**F1**	**MVar**	**MSD**
Rand	0.022	1.87E-03	0.032	0.031	1.90E-03	0.030	0.025	1.51E-03	0.026
Mcd	0.023	2.10E-03	0.034	0.033	2.27E-03	0.031	0.026	1.47E-03	0.026
Rdh	0.023	2.07E-03	0.034	0.032	2.34E-03	0.033	0.026	1.58E-03	0.027
Cov-txtQM	0.023	2.19E-03	0.034	0.033	2.30E-03	0.033	0.029	1.72E-03	0.028
Cov-T	0.009	5.55E-04	0.012	0.008	3.95E-04	0.008	0.005	1.93E-04	0.005
Cov-S	0.010	7.29E-04	0.013	0.011	5.52E-04	0.011	0.006	2.59E-04	0.007
Cov-E	0.022	1.93E-03	0.032	0.031	1.97E-03	0.030	0.026	1.48E-03	0.026
Cov-ST	0.013	8.95E-04	0.018	0.014	8.28E-04	0.014	0.008	3.90E-04	0.010
Cov-EST	0.023	2.06E-03	0.033	0.031	2.00E-03	0.030	0.027	1.44E-03	0.026
Div-T	0.021	1.89E-03	0.032	0.033	2.47E-03	0.033	0.029	1.93E-03	0.030
Div-S	0.022	1.92E-03	0.032	0.029	2.28E-03	0.031	0.020	1.33E-03	0.024
Div-E	0.023	2.05E-03	0.034	0.033	2.36E-03	0.032	0.030	2.00E-03	0.030
Div-ST	0.023	2.22E-03	0.035	0.025	1.39E-03	0.021	0.022	1.28E-03	0.019
Div-txtST	0.023	2.16E-03	0.035	0.034	2.47E-03	0.033	0.030	1.99E-03	0.030
Div-EST	0.023	2.09E-03	0.034	0.031	2.40E-03	0.031	0.030	1.99E-03	0.030
Div-txtEST	**0.024**	**2.30E-03**	**0.036**	**0.034**	**2.49E-03**	**0.033**	**0.031**	**2.07E-03**	**0.031**

word news articles generate better digests as compared to CW12 web pages. The text-only methods, *Mcd*, *Rdh*, and *Cov-txtQM* perform significantly better than *Rand*, as expected. Among the text-only methods, *Rdh* and *Cov-txtQM* show significant improvements over *Mcd* in terms of Rouge scores. At the same time, *Rdh* and *Cov-txtQM* follow different frameworks and prove to have overall similar performance. However, *Cov-txtQM* method proves to be better for Gigaword, and CW12 datasets while for NYT, *Cov-txtQM* gets significantly lower Rouge-2 F1 score. *Cov-txtEM* that uses only the empirical query terms proves to be the worst text-only method, thus highlighting the advantage of incorporating our query-text model. Next, we look at the methods that extend the coverage based framework into the different dimensions. *Cov-T* and *Cov-S* do not prove to be effective in any dataset. The *Cov-E* gets significantly higher score from its contemporary methods considering only time and geolocations. This method gets the highest gain over its contemporaries in the CW12 dataset. However, it always performs significantly worse than the text-only methods as shown in Table 4. The coverage-based methods in the time, geolocations, and entity dimensions get worse scores than the *Rand* due to relatively shorter digest generated. Later in the section, we discuss this in detail. Our proposed divergence-based methods perform better that the coverage-based methods across all three datasets. This is because *Div-T*, *Div-S*, *Div-E*, *Div-ST*, and *Div-STE* always perform better than the *Cov-T*, *Cov-S*, *Cov-E*, *Cov-ST*, and *Cov-STE* methods. Next we analyze the different combinations of dimensions in the divergence framework. The text-only method under this framework is equivalent to *Rdh* that performs significantly better than *Div-T*, *Div-S*, and *Div-E*. However, different combinations of the time, geolocations, and entities as *Div-ST* and *Div-STE* perform better than the text-only method. Finally, the *Div-txtSTE* with highest score proves to be the best method for our task.

Variance Analysis. From our experiments so far what is clear is that event digests generated by the *Div-txtEST* method are closest to the gold standard Wikipedia articles. However, we perform an extended evaluation using the proposed w-Rouge measure that computes the mean Rouge-1 F1 score with respect to individual paragraphs in the Wikipedia articles. Firstly, we assume that each paragraph describes some aspect of the event. Thus, a larger mean with high variance in the weighted F1 score indicates higher coverage of the Wikipedia paragraphs, and hence better diversity in the generated digest. As shown in Table 5, all methods get higher mean and variance from *Rand* across all three datasets. We find that the *Div-txtEST* method proves to be the method that generates most diversified digest by achieving the highest mean and MVar scores. The next best method is *Div-txtST*.

Varying Digest Length. Next we analyze the effect on the quality of the digest by varying the digest *length budget*. Table 6 compares the *Div-txtEST* method with the text-only methods, *Cov-txtQM* and *Rdh*, in terms of Rouge-2. What we find is that for smaller length budget, *Cov-txtQM* performs better than other methods. We also find that both *Cov-txtQM* and *Div-txtEST* perform better than the *Rdh* across all the length budgets in Gigaword and CW12 datasets. The poor performance of the *Rdh* as compared to *Cov-txtQM* can easily be understood by analyzing their formulation. For very small budget of only 50 words, on average only two excerpts are selected into the digest by all the methods. *Rdh* first selects the most relevant excerpt, and then as the next it selects the one that is least redundant from the first. This causes the Precision to fall and an overall decrease in the F1 score. On the other hand, *Cov-txtQM* attempts to cover as many important terms with high probability in the Q_{text} to maximize the coverage. This generates a better digest for the smaller length budget. However, in the NYT dataset, for larger length budgets, *Cov-txtQM* suffers due to the lesser redundancy in the news articles. Thus, as it tries to maximize term coverage, the Precision falls resulting in lower F1 scores as compared to *Rdh*. The effect of diversifying across time, geolocations, and entities in the *Div-txtEST* method is evident when the length budget is larger than 100.

Discussion. As the first point, we discuss the poor performance of the coverage framework in time, geolocation, and entity dimensions. The coverage-based framework selects excerpts such that the maximum number of *units* (of time, geolocations, or entities) in a given query are covered. The associated weights, as described in Equation 14, denote the importance of the units for a given query, thus forcing the ILP solver to cover more important units first. As a drawback, this framework does not take into consideration the importance of the units for the excerpts. This causes selection of excerpts that may not be relevant. Moreover, in the dimensions other than text, excerpts that do not come with explicit annotations are

Table 6: Varying the length budget of methods.

Datasets	Length	Rdh			Cov-txtQM			Div-txtEST		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
NYT	50	0.166	0.003	0.006	0.231	0.005	**0.009**	0.236	0.004	0.008
	100	0.222	0.009	**0.016**	0.211	0.008	0.015	0.223	0.008	0.015
	200	0.208	0.016	0.027	0.194	0.015	0.026	0.213	0.016	**0.027**
	300	0.199	0.022	0.036	0.184	0.021	0.034	0.205	0.023	**0.036**
	400	0.194	0.029	0.044	0.176	0.026	0.040	0.199	0.029	**0.044**
	500	0.189	0.034	0.050	0.169	0.032	0.046	0.194	0.035	**0.051**
Giga	50	0.146	0.003	0.005	0.268	0.006	**0.010**	0.242	0.005	0.009
	100	0.225	0.009	0.016	0.236	0.010	**0.018**	0.227	0.009	0.017
	200	0.209	0.016	0.027	0.206	0.016	0.027	0.209	0.017	**0.028**
	300	0.196	0.022	0.036	0.193	0.022	0.035	0.200	0.024	**0.037**
	400	0.189	0.028	0.043	0.186	0.028	0.043	0.191	0.030	**0.044**
	500	0.184	0.033	0.049	0.182	0.034	0.049	0.186	0.035	**0.051**
CW12	50	0.192	0.004	0.007	0.265	0.006	**0.010**	0.198	0.005	0.009
	100	0.189	0.007	0.013	0.223	0.009	**0.016**	0.187	0.009	0.016
	200	0.184	0.012	0.021	0.191	0.016	0.026	0.185	0.016	**0.027**
	300	0.187	0.017	0.028	0.176	0.021	0.032	0.176	0.022	**0.035**
	400	0.185	0.021	0.032	0.168	0.026	0.039	0.172	0.028	**0.041**
	500	0.183	0.023	0.035	0.164	0.030	0.044	0.169	0.033	**0.046**

automatically disregarded. Since a single temporal or geographical expression can represent a large time interval (e.g., a century) or geographic area (e.g., a continent) respectively, the coverage of query units are easily maximized by selecting few excerpts. For example, the entire temporal scope of query in Figure 1 is covered by excerpt 8. This causes the *Cov-T* and *Cov-S* methods to generate digests with fewer excerpts. Hence, they receive a worse Rouge score than the *Rand* which simply benefits from generating longer digest. On the other hand, the divergence-based framework additionally regards the importance (higher generative probability) of a unit for the individual excerpts. While generating the digest, the ILP solver first selects excerpts which cover important query units with higher probability, thus lowering the overall divergence of the digest to the query. Moreover, since each excerpt is associated with an independent smoothed model for each dimension, no excerpt is disregarded for digest generation.

Next, we discuss the diversification of excerpts in the digest achieved by each method. We note that Wikipedia articles as gold standard digests are textually larger than the system generated digests. We assume that Wikipedia articles cover most aspects of a given event as a query. To get a better insight into the diversification of the excerpts, we compare the methods using w-Rouge, proposed by us. The *Div-txtEST* gets the highest mean and variance scores, and proves that it achieves the best diversification.

We next discuss the individual dimensions. Text is clearly the most important dimension in all our methods. However, we find that the text-only methods heavily rely on the query modeling techniques. Using only the empirical query terms leads to worse performance as shown by the *Cov-txtEM* method. The *Cov-E* method uses only entities instead of all terms, as motivated by Gillick et al., is not able to beat text-only methods in Rouge-2 scores. Time and geolocations are important indicators of identifying event-related excerpts and work well as a combination. Individually, we run into sparsity problems with very few annotations in excerpts. This is more pronounced in the CW12 dataset. Combination of text, time, and geolocations as the *Div-txtST* proves to be the second best method in the news datasets where we get comparatively more annotations. This is also due to the fact that every excerpt from a news article is also annotated with the publication date of the source article. Entities are more prominent in the CW12 documents and help to reinforce the text model. However, across all the datasets, combination of all four dimensions proves to be most effective.

Gain/Loss analysis. Rouge measures assume that higher n-gram overlap with gold standard implies more relevant excerpts in the digest. Thus, to get insights into the overall quality of digests, we manually identify relevant excerpts (highlighted in green) by referring to their source documents. We look at queries for which *Div-txtEST* shows the highest gain and worst loss in w-Rouge scores, when compared to the best among the text-only methods.

It achieves the highest gain in w-Rouge from the best text-only method for the query:

January 26, 2001: An earthquake hits Gujarat, India, killing almost 20,000.

For this query, the best text-only method proves to be *Cov-txtQM*. Let us compare the digest generated by both methods:

Event Digest by the *Div-txtEST* method:
• In 1988 a devastating earthquake struck northern Armenia, killing 25,000 people, and in 1990 50,000 were killed by an earthquake that struck Rasht, Iran. • AP, 10405 2005 Oct 8, A 7.6-magnitude earthquake hit Kashmir near the Pakistan-India border reaching to Afghanistan. • Reuters, 725052005 Jul 24, A 7.2 earthquake hit Indias southern Andaman and Nicobar Islands and part of Indonesia. • **Shobha De, The Week, February 18, 2001 The story of the devastating earthquake in Gujarat is the story of women.** • **Scientists are already working to prepare earthquake probability map.** • **Skyscrapers need special construction to make them earthquake resistant.** • **This earthquake was not felt in Amreli, Junagarh or Porbander districts.** • **Leaders called for greater cooperation within the region to deal with the aftermath of disasters like the Kashmir earthquake and last year's devastating tsunami.** • **Additional Info The earthquake was centred 4.5 kms E of Gandhidham Gujarat, India.** • The Herald,India – was affected by the December 26, 2004 earthquake and subsequent tsunamis. • **Exactly a month ago, 18,122 people were killed in a deadly earthquake in the Kutch district.** • **The earthquake was felt strongly in parts of east-central Kachchh near the towns of Bachau and Vondh.** • Quito, now the capital of Ecuador, was shaken by an earthquake in 1797, and more than 40,000 people died. • **Almost four weeks after the earthquake, Gujarat is still coming to terms with what was and what is.** • The newly reopened Peace Bridge linking the Indian and Pakistani portions of disputed Kashmir nearly collapsed during the earthquake.

Event Digest by the *Cov-txtQM* method:
• 543 AD - Disastrous earthquakes shaked much of the world; 803 AD - Fierce storms lashed the west coast of Ireland, killing close to 1,000; 851 AD - Rome had a violent earthquake that damaged Pope Leos 4-year-old Leonine Wall and further destroyed the Colosseum; 856 AD - an earthquake at Corinth killed an estimated 45,000 Greeks; 856 AD - an earthquake at Damghan, Iran killed an estimated 200,000; 893 AD - an earthquake at Ardabil, Iran killed about 150,000 people1138 AD - an earthquake at Aleppo, Syria claimed lives of aprox 230,000 people; 1290 AD - an earthquake at Chihli, China killed about 100,000 people; 1319 AD - an Armenian earthquake shatters the city of Ani. • By Dear Anonymous As far as policymakers go, it is an ardous task to make them understand the dangers of exploiting all our...By Scientists suggest local-level mapping as India upgrades its seismic map. • The earthquake that hit Sikkim on September 18, killing some 150 people and devastating the Himalayan state, was unexpected. • 1976 AD - an earthquake and tidal wave hit Mindanao, Philippines; 1976 AD - an earthquake hit Guatemala; 1978 AD - an earthquake destroyed Tabas a city in eastern Iran; 1985 AD - a magnitude 8.1 earthquake devastated part of Mexico City and three coastal states1; 985 AD - Nevada del Ruiz errupted, 85 mi northwest of Bogot. • In North America, the San Francisco earthquake of 1906 caused extensive damage and claimed about 700 lives. • **Almost four weeks after the earthquake, Gujarat is still coming to terms with what was and what is.**

We note that the *Div-txtEST* method selects more relevant excerpts. The *Cov-txtQM* method maximizes the coverage of the textual terms in the query-text model which leads to selection of irrelevant excerpts that enlist past earthquakes. These are however not selected by *Div-txtEST* due to their large divergence with query-time and -space models. Thus, due to combined diversification across the time, geolocations, and entities, *Div-txtEST* generates better quality digest with more relevant excerpts and a wider coverage.

The *Div-txtEST* method gets the worst loss in w-Rouge from the best text-only method, *Cov-txtQM*, for query:

November 22, 1995: Rosemary West is sentenced to life for killing 10 women and girls, including her daughter and stepdaughter, after the jury returns a guilty verdict at Winchester Crown Court. The trial judge recommends that she should never be released from prison, making her only the second woman in British legal history to be subjected to a whole life tariff (the other is Myra Hindley).

Let us compare the digest generated by both methods:

Event Digest by the *Div-txtEST* method:
• An Orange County jury found Miller guilty in November, and on Nov. 27 see voted 11-1 to recommend the death sentence. • A Pasco County jury found Partin guilty of 1st-degree murder in March. • In July, a jury found Ballard, 67, white, guilty of the 1st-degree murder of Traub. • The same jury found Abdool guilty of first-degree murder last week for the Feb. 25, 2006 death of 17-year-old Amelia Sookdeo. • Last month, on March 5 see, a jury voted 7-5 to recommend a death sentence for Smith. • The sentence followed an 11-1 jury recommendation that Wade receive the death penalty. • The judge may override the jury decision. • Albright Gregory Murphy, 45, was convicted of the 1st-degree premeditated killing in a March jury trial. • The points in Welchs appeal, he challenged his sentence on the basis of jury selection. • The jury voted 7 to 5 to recommend the killer receive a death sentence. • The jury voted 10 to 2. • **She was found guilty of murder and sentenced to life imprisonment.** • The jury recommended by a 9-3 vote that Ballard be executed. • Judges typically follow jury recommendations, so Teppers ruling was somewhat surprising. • On Sept. 26 the same jury that found him guilty, unanimously recommended the death penalty. • After a bench trial, JudgeBurge found Kovarbasich guilty of voluntary manslaughter on April 29, 2010. • A life sentence was recommended by Florida prosecutors. • Watson vacated the sentence in 2005. • The sentence is determined by a jury.

Event Digest by the *Cov-txtQM* method:
• An Orange County jury found Miller guilty in November, and on Nov. 27 see voted 11-1 to recommend the death sentence. • It is a very difficult case, but it is the verdict of the jury, he said. • The sentence is determined by a jury. • Albright Gregory Murphy, 45, was convicted of the 1st-degree premeditated killing in a March jury trial. • As he had during the verdict, Riley showed little emotion as his sentence was imposed. • On Sept. 26 the same jury that found him guilty, unanimously recommended the death penalty. • But next week vital and previously withheld testimony from one of her children could overturn the sentence By Tracy McVeigh • **A battered wife serving life imprisonment for killing her husband may soon be freed following evidence from her traumatize daughter which was held back from the original jury.** • **The jury was unanimous West was guilty of ten murders, and the judge, Mr Justice Mantell, sentenced her to life imprisonment with a recommendation that she should serve at least 25 years.** • The jury voted 10 to 2. • Miscarriage of justice and womens groups have been campaigning on Donna's behalf since the verdict two years ago. • After a three-day trial and only two and a half hours of deliberation, a jury of five men and one woman convicted Wade of second-degree murder. • Circuit Judge Lynn Tepper sentenced Smith, 30, to life in prison for killing Robert Crawford in 1999, breaking with a split jurys recommendation that Smith be sentenced to death.

Both methods select few relevant excerpts into the digest. However, the text-only method selects excerpts based on term matching that leads to better Rouge scores. *Div-txtEST* method suffers due to the sparsity of annotations in the event dimensions.

5. CONCLUSION

We proposed the problem of generating a digest that presents a holistic view on a given Wikipedia event. We proposed a novel divergence-based framework for selecting excerpts from pseudo-relevant input news articles such that the global divergence between the digest and the query is minimized. The problem was formulated as an ILP for global inference to maximize the overall relevance of the digest to the input event query, while reducing inter-excerpt redundancies in text, time, geolocations, and entity dimensions. In experimental evaluation, we compared several methods, and found that our divergence-based method that considers all dimensions of an event proves to be most appropriate for event digest generation.

References

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. *SIGIR* 2001.

[2] K. Berberich, S. J. Bedathur, O. Alonso, and G. Weikum. A language modeling approach for temporal information needs. *ECIR* 2010.

[3] J. P. Callan. Passage-level evidence in document retrieval. *SIGIR* 1994.

[4] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. *SIGIR* 1998.

[5] H. L. Chieu and Y. K. Lee. Query based event extraction along a timeline. *SIGIR* 2004.

[6] Z. Dou, S. Hu, K. Chen, R. Song, and J. R. Wen. Multi-dimensional search result diversification. *WSDM* 2009.

[7] E. Filatova. Event-based extractive summarization. *ACL Workshop on Summarization* 2004.

[8] D. Gillick and B. Favre. A scalable global model for summarization. *ILP* 2009.

[9] M. A. Hearst and C. Plaunt. Subtopic structuring for full-length document access. *SIGIR* 1993.

[10] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *AAAI Press* 2013.

[11] J. Hoffart. Discovering and Disambiguating Named Entities in Text. *SIGMOD/PODS Ph.D. Symposium* 2013.

[12] P. Hu, D. H. Ji, H. Wang, and C. Teng. Query-focused multi-document summarization using co-training based semi-supervised learning. *PACLIC* 2009.

[13] M. Kaszkiel and J. Zobel. Effective ranking with arbitrary passages. *Journal of the American Society for Information Science and Technology*, 52(4) 2001.

[14] Y. Li and S. Li. Query-focused multi-document summarization: Combining a topic model with graph-based semi-supervised learning. *COLING* 2014.

[15] C. Y. Lin. Rouge: A package for automatic evaluation of summaries. *ACL Workshop* 2004.

[16] M. Litvak and M. Last. Graph-based keyword extraction for single-document summarization. *ACL* 2008.

[17] R. McCreadie, C. Macdonald, and I. Ounis. Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. *CIKM* 2014.

[18] R. McDonald. A Study of Global Inference Algorithms in Multi-document Summarization. *Springer Berlin-Heidelberg* 2007.

[19] A. Mishra and K. Berberich. Linking Wikipedia Events to Past News. *TAIA* 2014.

[20] A. Mishra and K. Berberich. Leveraging Semantic Annotations to Link Wikipedia and News Archives. *ECIR* 2016.

[21] K. Riedhammer, B. Favre, and D. Hakkani-Tür. Long story short–global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52(10) 2010.

[22] D. Shahaf, C. Guestrin, E. Horvitz, J. Leskovec. Effective ranking with arbitrary passages. *Communications of the ACM*, 58(11) 2015.

[23] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. *SIGIR* 1993.

[24] B. Taneva , G .Weikum. Gem-based entity-knowledge maintenance. *ECIR* 2013.

[25] M. Tsagkias, M. de Rijke, and W. Weerkamp. Linking online news and social media. *WSDM* 2011.

[26] X. Wang, H. Fang, and C. Zhai. A study of methods for negative relevance feedback. *SIGIR* 2008.

[27] K. Woodsend and M. Lapata. Multiple aspect summarization using integer linear programming. *EMNLP-CoNLL* 2012.

[28] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. *SIGIR* 2011.

[29] C. Zhai and J. D. Lafferty. Two-stage language models for information retrieval. *SIGIR* 2002.

[30] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. *SIGIR* 2003.

[31] S. H. Zhong, Y. Liu, B. Li, and J. Long. Query-oriented unsupervised multi-document summarization via deep learning model. *Expert Systems with Applications*, 42(21) 2015.