ABSTRACT

It is often difficult to ground text to precise time intervals due to the inherent uncertainty arising from either missing or multiple expressions at year, month, and day time granularities. We address the problem of estimating an excerpt-time model capturing the temporal scope of a given news article excerpt as a probability distribution over chronons. For this, we propose a semi-supervised distribution propagation framework that leverages redundancy in the data to improve the quality of estimated time models. Our method generates an event graph with excerpts as nodes and models various inter-excerpt relations as edges. It then propagates empirical excerpt-time models estimated for temporally annotated excerpts, to those that are strongly related but miss annotations. In our experiments, we first generate a test query set by randomly sampling 100 Wikipedia events as queries. For each query, making use of a standard text retrieval model, we then obtain top-10 documents with an average of 150 excerpts. From these, each temporally annotated excerpt is considered as gold standard. The evaluation measures are first computed for each gold standard excerpt for a single query, by comparing the estimated model with our method to the empirical model from the original expressions. Final scores are reported by averaging over all the test queries. Experiments on the English Gigaword corpus show that our method estimates significantly better time models than several baselines taken from the literature.

CCS Concepts

- Information systems → Content analysis and feature selection; Probabilistic retrieval models;

Keywords

excerpt-time model; temporal scoping; distribution propagation; temporal content analysis; probabilistic models; sparsity reduction;

1. INTRODUCTION

Time is considered an integral dimension for understanding and retrospecting on past events. Temporal information associated with a past event, not only indicates its occurrence period but also helps to understand its causality, evolution, and ramifications. In this big data era, with ever increasing amounts of information on past events made available on the World Wide Web, time becomes an important indicator to organize and search relevant information. Typical sources of such temporal information are digital news archives, blogs, and online encyclopedias like Wikipedia. With this conception, time as a dimension has thus been leveraged to improve the effectiveness of various tasks in temporal information extraction, retrieval, and multi-document summarization.

Temporal information extraction [34] focuses on recognizing and normalizing temporal expressions embedded in text into precise time points or intervals. The normalized time intervals are usually represented in a standard format like TIMEX3 [34] of the TimeML markup language. Commonly, temporal expressions are categorized into four types: explicit, relative, implicit, and free-text. Most of the approaches [1, 20, 32] operate only on individual terms or phrases, and associating temporal information to larger textural units (like sentences or paragraphs) is out of scope. For such larger textual units describing events with multiple aspects, it is difficult to tag them with precise time intervals.

In temporal information retrieval, most approaches make use of the meta-data like creation time or publication dates associated with documents to identify their temporal scope. To estimate the relevance of a document for a given query, approaches [23, 30] model the scope of documents by applying an exponential decay function to smooth their publication dates. There are a few approaches [2, 5, 17, 28, 31] that leverage expressions embedded in the text of the documents to estimate their temporal scopes. In a prior work [27], we present an approach to estimate query-time models which are combined with language models to improve document retrieval effectiveness. However, most of these approaches that rely on explicit temporal expressions are not easy to extend to sentences or passage retrieval due to the high sparsity of annotations.

In extractive summarization, time has been leveraged to order sentences in textual summaries. For a summary generated in context of an event, it is intuitive to present a chronological ordering of the sentences extracted from the news articles. Some approaches [3] leverage the publication dates of the source news articles as a proxy to model the temporal scope of sentences extracted from them. This strategy however assumes that the documents are highly coherent and focus on a single time period indicated by their publication dates. Other approaches [6, 29] combine chronological with several strategies while ordering multiple sentences from the same document. In another recent work [26], we estimate time models for sentences by considering temporal expressions in a set of pseudo-relevant documents. However, in their approach, all sentences from a single document that do not come with any temporal expression end up having similar time models. Though this
### Table 1: Example of redundant news article excerpts describing the same event, along with their temporal annotations.

<table>
<thead>
<tr>
<th>Excerpts (Events)</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_1$: The electronic producers Skrillex and Diplo, who under the name Jack Ü had one of the biggest hits last year with “Where Are Ü Now”, featuring Justin Bieber, won both best dance recording for that song and best electronic/dance album.</td>
<td>$T_1$: “last year” = [01-01-2015, 31-12-2015]</td>
</tr>
<tr>
<td>$\varepsilon_2$: On Monday, Skrillex and Diplo won the best dance/electronic album for “Skrillex and Diplo Present Jack and best dance recording for Where Are Ü Now”.</td>
<td>$T_2$: “Monday” = [02-02-2016, 03-02-2016]</td>
</tr>
<tr>
<td>$\varepsilon_3$: Diplo and Skrillex took home the gold twice at the Grammy Awards night winning Best Dance Recording and Best Dance/Electronic Album.</td>
<td></td>
</tr>
</tbody>
</table>

Approach proves to be more effective than text-only approaches, it can be further improved with more accurate and discriminative time models estimated for sentences with missing annotations during summary generation.

It can be noted that approaches addressing the above mentioned tasks severely suffer due to high sparsity of temporal information when extended to finer granularity of textual units like passages or sentences. We refer to the finer textual units simply as excerpts. To deal with sparsity, we propose to estimate an excerpt-time model that captures the temporal scope of a given excerpt describing an event. An excerpt-time model can be understood as a probability distribution over time units or chronons such as days, months, or years. Thus, the temporal scope of excerpts is represented as a probabilistic time model instead of precise intervals on a time scale. To further explain this, consider the first example excerpt $\varepsilon_1$ from Table 1. The single excerpt gives information on two independent time periods. First, the song “Where Are Ü Now” becoming a hit in the year “2015” and second, Skrillex and Diplo winning the Grammy award which took place on “February 2, 2016”. Temporally tagging $\varepsilon_1$, will associate the entire excerpt to the year “2015”. However, the desideratum is that the excerpt is associated with a probabilistic model in a time domain where the salient time units pertaining to the year “2015” and the day “February 2, 2016” receive higher probabilities. Also, even though $\varepsilon_3$ does not mention any expression, it still conveys temporal information with a scope that needs to be estimated as its time model.

One plausible approach to estimate a more accurate time model for a given excerpt is to leverage the redundancy in the data. For example, given both $\varepsilon_1$ and $\varepsilon_2$ from Table 1, one can note that the second part of $\varepsilon_1$ highly overlaps with the event described in $\varepsilon_2$. Since $\varepsilon_1$ comes with an expression “Monday” that is normalized to “February 2, 2016” with a standard temporal tagger, we can use this redundancy to estimate a more accurate time model for $\varepsilon_1$. Similarly, the time model of $\varepsilon_3$ that does not come with any temporal expression can be estimated by combining the time models of $\varepsilon_2$ and $\varepsilon_3$. This is because it contains overlap with both the excerpts.

Motivated from the above examples, in this paper we address the following problem: from a given set of excerpts describing events, for each excerpt automatically estimate an excerpt-time model capturing its temporal scope. As input, we consider: 1) a set of excerpts along with their source documents; 2) a set of normalized temporal expressions extracted from the input set of excerpts. As output our method automatically estimates a probabilistic time model for each excerpt capturing its temporal scope.

**Challenges** include estimating an excerpt-time model from a given set of salient temporal expressions for an excerpt. In addition, leveraging redundancy in text to improve the quality of estimated excerpt-time models is a challenge.

**Applications.** Excerpt-time models can be leveraged in various applications. In temporal information extraction, these models can lead to improvements in accuracy of free text temporal expression normalization. For example, $\varepsilon_3$ contains the expression “Grammy Awards night” whose normalization can be improved by combining the other two excerpts in Table 1. For information retrieval, accurate excerpt-time models can be used to estimate better query-time models [27]. A direct application in extractive summarization is to leverage the time models to improve the summary quality [26], and chronologically ordering excerpts. Several tasks in timeline generation, question answering, and event detection also benefit from the time models.

**Contributions** made by this paper are as follows: 1) We propose the problem of temporally scoping excerpts by estimating probabilistic excerpt-time models; 2) we propose a distribution propagation framework that models several inter-excerpt relations, and then estimates excerpt-time models by propagating information from related excerpts that come with temporal expressions; 3) finally, we propose two new measures, Model Quality and Generative Power, to evaluate our method on a real-world data set.

**Organization.** In Section 2, we review related work from the literature and give technical background; Section 3 gives details of our approach. Conducted experiments and their results are described in Section 4. Finally, we conclude in Section 6.
Some attention has also been given to annotating text segments that do not contain any expressions. Chasin et al. [9] temporally anchor events to a time period between the prior time stamp and subsequent time stamp occurring in text. However, they themselves consider this as a naïve algorithm which is also acknowledged by Gung et al. [14] as a good baseline. Gung et al. use a clustering-based method to annotate textualy similar sentences with the same time-stamps. Though this seems to be a reasonable assumption, hard clustering may lead to overly-specific or generic temporal scope for specific incidents of larger events. More complex approaches [7, 21, 36] make use of additional linguistic features to anchor time. Bramsen et al. [7] break complex sentences into clauses and look for temporal markers like after, in next year, etc., to detect breaks in temporal continuity. Jatowt et al. [15] and Kotsakos et al. [19] attempt to estimate the precise focus time of text documents in a news corpus at the year granularity by leveraging temporal language models. More recently, Laparra et al. [21] use event detection algorithms to propagate time stamps.

In the context of our problem, recognizing and normalizing temporal expressions is not sufficient to estimate excerpt-time models. This is because the majority of excerpts extracted from documents do not come with expressions. Also, the expressions often suffer from uncertainty [5]. Other approaches mentioned above rely on corpus statistics and temporal relations. Firstly, these methods are not comparable as they aim at annotating text segments with a single precise time point or interval as opposed to a distribution like in our problem. Secondly, there exists a difference in their notion for “events” which makes them incomparable. Most commonly in prior works, events are described as topics with few keywords while we consider a sentence from a news article. We use the SU-Time toolkit to annotate the temporal expressions in the excerpts that are extracted from news articles.

Temporal Information Retrieval is another research area where leveraging temporal information associated with documents has received much attention. Among the first works, Li and Croft [23] proposed a method to combine a document’s language model with a time dependent prior indicating its temporal scope. This approach was extended by Peetz et al. [30] by investigating different priors based on cognitive models. Berberich et al. [5] was the first to present a two dimensional representation to model the time domain, and design a query-likelihood based method to rank documents by combining text and time. Recently, Efron et al. [12] presented a non-parametric kernel density estimation method to incorporate temporal relevance feedback from users to improve retrieval quality. As a recent work in this direction, Mishra et al. [28] present a method to link Wikipedia events to news articles by treating a given event as a query and estimating query-time models from a set of pseudo-relevant documents. Most of the approaches mentioned above rely only on the meta-data associated with the documents like their creation times. In addition, they focus on document retrieval and may not extend to excerpts due to high sparsity.

In this work, we adopt the two-dimensional representation of time as presented by Berberich et al. [5]. This allows us to estimate two-dimensional time models by capturing the uncertainty in time. Mishra et al. [28] proposed a method to estimate a query time model that captures the temporal scope of the query. In this work, we adopt a methodology proposed by Mishra et al. [28] to estimate excerpt-time models from a set of temporal expressions. However, our work in this paper differs in two ways. First, we aim to estimate time models for excerpts, instead of entire documents, which may not come with expressions. Second, we propose a distribution propagation framework to estimate the time-models more accurately instead of relying on pseudo-relevant documents.

Extractive Summarization. Many of the studies that leveraged time, either used document publication date as a proxy for estimating sentence temporal scope [3, 6, 29] or created handcrafted rules to identify temporal annotations for relations [13, 24, 25]. Later works [14] leveraged rule-based taggers to tag and normalize temporal expression. However, as an issue pointed out by Mani et al. [24] temporal expressions only constitute about 25% of the temporal information in a typical news corpus and are insufficient for any learning-based method. Most recently, Mishra et al. [26] address the problem of event digest generation from an input set of news articles. They presented a framework to diversify across text, time, geolocations, and named entities for the digest generation.

In this work, we adopt the concept of excerpt-time models introduced by Mishra et al. [26]. In their work, they attempt to estimate a probabilistic time model for each excerpt that capture the temporal scope of the events described by them. However, the focus of their problem was different. Moreover, they relied on the temporal expressions that are mentioned in the source documents of the excerpt to estimate the time models. In their approach, excerpts taken from the same documents that do not come with any explicit expression get associated with similar time models that are no longer discriminative. This stands as a contrast to this work where we focus on estimating an accurate time model for each excerpt by leveraging the redundancy of information in a set of documents.

Semi-Supervised Label Propagation [4, 18, 37–39] aims at labeling data by propagating class labels from labeled data. For this, the geometry of the data is leveraged by first generating a data graph where the nodes represent data points, and an edge between two nodes represents similarity between them. This paradigm was introduced by Zhu and Ghahramani [38]. In their later work [39], they proposed another formulation of the propagation algorithm that models it as a Harmonic Gaussian Field. A similar propagation algorithm was proposed by [37] that introduces self-loops so that during each iteration, nodes also receive a small contribution from their initial values. Recently, Karasuyama et al. [18] proposed an algorithm to efficiently identify the manifold structure of a data graph and at the same time learn the hyper-parameters by a novel feature propagation method.

In this work, among the several approaches, we choose to adopt the algorithm proposed by Zhu and Ghahramani [38] for designing our distribution propagation framework. However, there are a few fundamental differences. Firstly, their algorithm is designed for data points in Euclidean space. However, in the text space, this type of similarity (distance function) is known to not perform well. So we incorporate JS-Divergence as our distance metric. Secondly, in their original objective function, they introduce a hyper-parameter \( \sigma \) as an importance parameter for each dimension. This is learned independently for each dimension. However, we linearly combine all the inter-excerpt relation weights and have a single scaling parameter that is set to the average of the total weight for simplicity.

3. APPROACH

We design a distribution propagation algorithm that is based on label propagation proposed by Zhu et al. [38]. In our method, from the input set of pseudo-relevant excerpts \( R \), the subset of excerpts that come with temporal expressions are treated as a seed set \( R_{\text{seed}} \). For each excerpt in \( R_{\text{seed}} \), we estimate an empirical excerpt-time model from their original expressions. The time models for excerpts with missing expressions are initialized to a uniform distribution. An event graph is generated by considering all excerpts in \( R \) as nodes, and models their relationships as weighted edges. Finally, as an iterative process, the empirical time models from ex-
3.1 Definitions

We first define the notations and representations used to design our method. Table 2 summarizes our notations.

<table>
<thead>
<tr>
<th>Notations</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d$, $C$, $V$</td>
<td>Document, Collection, and Vocabulary</td>
</tr>
<tr>
<td>$d_{event}$, $d_{context}$</td>
<td>Wikipedia Current Events portal, and source document</td>
</tr>
<tr>
<td>$\varepsilon$, $\varepsilon_{text}$, $\varepsilon_{time}$</td>
<td>Excerpt, excerpt text part, excerpt time part</td>
</tr>
<tr>
<td>$E_{text}$, $E_{time}$</td>
<td>Excerpt-text, and excerpt-time model</td>
</tr>
<tr>
<td>$t$, $\tau$</td>
<td>Temporal expression, time unit</td>
</tr>
<tr>
<td>$[tb, te]$, $tb$, $te$</td>
<td>Time interval, begin time point, and end time point</td>
</tr>
<tr>
<td>$tb_{b}$, $te_{b}$</td>
<td>Upper bound on begin and end time points</td>
</tr>
<tr>
<td>$R$, $R_{seed}$, $R_{text}$</td>
<td>Input set, seed set with annotated excerpts, and test set with a single excerpt whose time model is to be estimated</td>
</tr>
<tr>
<td>$\min_{ij}$</td>
<td>Earliest time unit in $R_{seed}$</td>
</tr>
<tr>
<td>$\max_{ij}$</td>
<td>Latest time unit in $R_{seed}$</td>
</tr>
<tr>
<td>$G_{event}$, $w_{ij}$</td>
<td>Event graph, and edge weight between two nodes</td>
</tr>
<tr>
<td>$\delta_{ij}$</td>
<td>Total similarity between two excerpts</td>
</tr>
<tr>
<td>$\delta_{ij}^{text}$, $\delta_{ij}^{time}$, $\delta_{ij}^{context}$</td>
<td>Textual, positional, conceptual, and contextual similarity</td>
</tr>
</tbody>
</table>

Table 2: List of notations

Excerpts with expressions are propagated to those that are strongly related but do not come with any temporal expression.

A time model for an excerpt can be understood as a probability distribution in our time domain that captures the true temporal scope of the excerpt. We leverage redundancy in data to estimate a time model for excerpts that do not come with explicit information. We follow the idea that strongly related excerpts refer to the same event and hence share the same temporal focus.

3.2 Excerpt Models

With our notations and representations defined, we now adopt query-modeling techniques described by Mishra et al. [27].

**Excerpt-Text Model** $\varepsilon_{text}$ refers to a unigram language model estimated from $\varepsilon_{text}$ that captures the event described in the excerpt. It can be observed that any $\varepsilon_{text}$ comes with two types of terms: first, that convey background information, and second, that describe a specific event. To stress on the terms salient to the described event, we combine the empirical excerpt-text model with a background model estimated from: 1) the textual content of the source news article, $d_{context}$; and 2) the textual descriptions of events from Wikipedia Current Events portal coalesced into a single document, $d_{event}$. The $d_{context}$ background model puts emphasis on the contextual terms, like (Skrillex, Diplo, Grammy, album). The $d_{event}$ background model puts emphasis on terms like (won, award) that are discriminative for the specific event at hand.

We combine the excerpt-text model with a background model by linear interpolation. The generative probability of a word $w$ from the excerpt-text model $\varepsilon_{text}$ is estimated as,

$$P(w \mid \varepsilon_{text}) = (1 - \lambda) \cdot P(w \mid \varepsilon_{text}) + \lambda \cdot \beta \cdot P(w \mid d_{event}) + (1 - \beta) \cdot P(w \mid \varepsilon_{context})$$

A term $w$ is generated from the background model with probability $\lambda$ and from the original excerpt with probability $1 - \lambda$. Since we use a subset of the available terms, we finally re-normalize the excerpt-text model. The new probability $P(w \mid \varepsilon_{text})$ is estimated as,

$$\hat{P}(w \mid \varepsilon_{text}) = \frac{P(w \mid \varepsilon_{text})}{\sum_{w' \in V} P(w' \mid \varepsilon_{text})}$$

The **Excerpt-Time Model** $\varepsilon_{time}$ can be understood as probability distribution that captures the salient periods for an event described in the excerpt. Further, we assume that a temporal expression $\tau \in \varepsilon_{time}$ is sampled from the excerpt-time model $\varepsilon_{time}$. The generative probability of any time unit $\tau$ from the time model $\varepsilon_{time}$ is estimated by iterating over all the temporal expressions $\tau = [tb, te]$ in $\varepsilon_{time}$ as,

$$P(\tau \mid \varepsilon_{time}) = \sum_{[tb, te] \in \varepsilon_{time}} \frac{I(\tau \in [tb, te])}{\|[tb, te] \in \varepsilon_{time}||}$$

where the $I(\cdot)$ function returns 1 if there is an overlap between a time unit $\tau$ and an interval $[tb, te]$. The denominator computes the area of the temporal expression in $T \times T$. For any given temporal expression, we can compute its area and its intersection with other expressions as described in [5]. Intuitively, the above equation assigns higher probability to time units that overlap with a larger number of specific (smaller area) intervals in $\varepsilon_{time}$. Finally, we re-normalize as per Equation 2.

3.3 Inter-Excerpt Relations

Edge weights in an event graph denote the relationship between two excerpts. Larger weights between two excerpts indicate closer relation with more informational overlap, and hence point to the fact that excerpts may focus on the same time periods. In our method, the edge weights are computed by enforcing an exponential function over the total similarity $\delta_{ij}$ between two excerpts $\varepsilon_i$ and $\varepsilon_j$ thus allowing propagation from similar excerpts more freely. Formally the edge weights are computed as,

$$w_{ij} = exp \left( \frac{\delta_{ij}}{\sigma} \right)$$

where $\sigma$ is the scaling parameter and is set to the average similarity between the excerpts. We model $\delta_{ij}$ as a linear combination of five
similarity scores capturing different relationships as,
\[ \delta_{ij} = \delta_{text\ ij} + \delta_{exp\ ij} + \delta_{pos\ ij} + \delta_{ctx\ ij}. \]  

In the above, each of the factors are normalized across the excerpts using min-max normalization. Formally, this is computed as,
\[ \hat{\delta}_{ij} = \frac{\delta_{ij} - \min(\delta_{ij})}{\max(\delta_{ij}) - \min(\delta_{ij})}. \]

Finally, as motivated by Zhu et al. [38], we additionally smooth the weight matrix with a uniform transition probability matrix \( U \) where \( U_{ij} = 1/|\mathcal{U}| \) to compute \( W \) as,
\[ W = \gamma \cdot U + (1 - \gamma) \cdot W. \]  

We compute the following inter-excerpt relations by leveraging the excerpt models estimated in Section 3.2.

Text similarity \( \delta_{text\ ij} \) between two excerpts can be a strong indicator of their information overlap. Leveraging this idea, we state the following hypothesis:

If two excerpts are textually similar, then they most likely discuss the same event and time period.

In order to estimate the text similarity between any two excerpts \( \varepsilon_i \) and \( \varepsilon_j \), we compute Jensen-Shannon divergence between their excerpt-text models which is estimated as per Equation 1. Formally, this is given as
\[ \delta_{text\ ij} = -JSD(\varepsilon_i\ text || \varepsilon_j\ text). \]  

Positional similarity \( \delta_{pos\ ij} \). It can be assumed that the source news articles from which the excerpts have been extracted exhibit a coherent structure. This observation to a certain degree can be generalized to the temporal dimension of the articles. With this assumption, we state the following hypothesis:

If two excerpts occur in the same document, and have higher positional proximity in the document, then they most likely discuss the same event and time period.

Unlike the two prior similarities for calculating the positional similarity, we compute the absolute distance between the sentence-positions of two excerpts, and apply an exponential decay function over it as shown by Tao et al. [33]. Formally this is,
\[ \delta_{pos\ ij} = \begin{cases} \log(a + \exp(-\text{Dist}(\varepsilon_i, \varepsilon_j))) & \text{if } \varepsilon_i, \varepsilon_j \in d \\ 0 & \text{otherwise} \end{cases} \]  

where \( a \) is set to 0.3 [33], and \( \text{Dist()} \) function returns the difference in positions as
\[ \max(\text{pos}(\varepsilon_i), \text{pos}(\varepsilon_j)) - \min(\max(\text{pos}(\varepsilon_i), \text{pos}(\varepsilon_j))). \]

Conceptual similarity \( \delta_{exp\ ij} \) between two excerpts can be computed as the noun phrase overlap between them. Intuitively, noun phrase extraction can be considered as a soft form of entity recognition. Higher similarity between excerpt-noun-phrase models indicates a stronger relationship. We state the following hypothesis:

If two excerpts contain similar noun phrases, then they most likely discuss the same event and time period.

To estimate \( \delta_{exp\ ij} \), we first run it through an open source part-of-speech based noun phrase extractor. Then for an excerpt \( \varepsilon_i \), we estimate its excerpt-noun-phrase model \( E_{\varepsilon_i\ np} \) according to Equation 1 by simply treating each noun phrase as a distinct term. Finally, conceptual similarity between any two excerpts \( \varepsilon_i \) and \( \varepsilon_j \) is computed as Jensen-Shannon divergence between their excerpt-noun-phrase models. Formally, that is,
\[ \delta_{exp\ ij} = -JSD(E_{\varepsilon_i\ np} || E_{\varepsilon_j\ np}). \]  

\[ \text{Algorithm 1 \ Temporal Distribution Propagation} \]
\[ \text{Construct set } R \text{ from input documents} \]
\[ \text{Initialize the } min_t \text{ and } max_t \text{ as earliest and latest day in } R, \text{ respectively} \]
\[ \text{for excerpt } \varepsilon_i \in R \text{ do} \]
\[ \text{if } |\varepsilon_i\ time| > 0 \text{ then} \]
\[ \text{Estimate } \hat{E}_{\varepsilon_i\ time} \text{ (in Section 3.2)} \]
\[ Y_i \leftarrow Y_i \cup \hat{E}_{\varepsilon_i\ time} \]
\[ \text{else} \]
\[ \text{Initialize } P(\gamma|\varepsilon_i\ time) \]
\[ Y_u \leftarrow Y_u \cup \mathcal{E}_{\varepsilon_i\ time} \]
\[ \text{end if} \]
\[ \text{end for} \]
\[ \text{Generate an event graph } G \text{ (in Section 3.1)} \]
\[ \text{Compute the affinity matrix } W \text{ (in Section 3.3)} \]
\[ \text{Compute diagonal degree matrix } D_{\varepsilon_i} = \sum_j w_{ij} \]
\[ \text{Initialize } Y(0) \leftarrow (Y_i, Y_u) \]
\[ \text{while not convergence to } Y(x) \text{ do} \]
\[ Y(t+1) \leftarrow D^{-1} \cdot W \cdot Y(t) \]
\[ Y_u(t+1) \leftarrow Y(t) \]
\[ \text{end while} \]
\[ \text{Final } \mathcal{E}_{\varepsilon_i\ time} \text{ for } \varepsilon_i \text{ is then obtained from } Y(x) \]

Contextual similarity \( \delta_{ctx\ ij} \) becomes an important indicator of the relationship between two excerpts if their textual description is sparse. It may also indicate if two events are part of a common larger event and should have happened in a similar time period. Considering this idea, we state the following hypothesis:

If two excerpts have higher contextual similarity, then most likely they are a part of the same event and time period.

In order to estimate \( \delta_{ctx\ ij} \), we leverage the lead (first) paragraphs of the source news articles of \( \varepsilon_i \) and \( \varepsilon_j \). For each excerpt, we first estimate an excerpt-context model \( E_{\varepsilon_i\ ctx} \) from the lead paragraph of their source news articles. This is done similar to Equation 1. With this the contextual similarity can simply be defined as the Jensen-Shannon divergence between the excerpt-context models as,
\[ \delta_{ctx\ ij} = -JSD(E_{\varepsilon_i\ ctx} || E_{\varepsilon_j\ ctx}). \]  

### 3.4 Distribution Propagation

Our distribution propagation algorithm is based on label propagation as proposed by Zhu et al. [38]. Intuitively, we leverage the idea that if two excerpts have a strong inter-excerpt relationships between them, then they may refer to the same event, and hence have a similar temporal scope.

Pseudo-code of our method is illustrated in Algorithm 1. As the first step, we extract a set of excerpts \( R \) from a given set of input documents. We then extract the earliest and latest time unit at a fixed time granularity from \( R \). This fixes the total scope of our time domain \( T \times T \). In the next step, we estimate the excerpt-time model for each excerpt that comes with temporal expression in the seed set \( R_{seed} \). We add these excerpts to the labeled class \( Y_i \). For the rest of the excerpts, we assume that the generative probability of any time unit \( \tau \) is uniform, and we add these to the unlabeled class \( Y_u \). Given all excerpts in \( R \), we then construct an event graph \( G \) with excerpts as nodes. The weight matrix \( W \) is computed that models the relationships between every excerpt-pair and treats them as weighted edges. The algorithm then performs the following steps until convergence: first, all excerpts propagate their time models. Second, instead of letting the generative probabilities in excerpt-time models in \( Y \) getting readjusted, we reinitialize their original distributions. Finally, we retrieve the excerpt-time models of the unlabeled excerpts from \( Y(x) \). With the weight matrix \( W \) row normalized, the algorithm is guaranteed to converge as proven by Zhu et al. [38].
4. EXPERIMENTS

In this section, we describe the details of the conducted experiments. We make all our experimental data publicly available.

4.1 Setup

We note that there is no ready-to-use ground truth for our task. In order to generate a test set of excerpts, we adopt a query-driven methodology where we randomly select a set of Wikipedia events and treat them as a user query. We then retrieve top-K documents relevant to this event and treat them as a set of pseudo-relevant excerpts. This methodology has two effects. First, the search space for our distribution propagation algorithm is reduced. Second, this method can be considered as filtering out noisy temporal expressions. However, it is worthwhile to highlight that our method is not dependent on this filtering step and can be generalized to any set of excerpts. This method, however, enables us to systematically evaluate the effectiveness of our method as compared to other baselines.

Test Collections. We perform experiments on the English Gigaword corpus with about 9 million news articles published between 1991 and 2010. We process the queries in our test set with a standard query-likelihood document retrieval model. Top-10 retrieved documents with an average of 150 excerpts are considered pseudo-relevant and input into our methods.

Test Queries are generated from the timeline of modern history in Wikipedia that enlists the most prominent news events in the 20th and the 21st centuries. We randomly sample 100 events that took place between 1987 and 2007, and treat them as test queries. Each query comes with a short textual description and a time interval indicating the occurrence of the event. For the experiments, we ignore the time interval, and only leverage the textual expression as a keyword query to retrieve the pseudo-relevant documents.

Ground Truth. We rely on the empirical temporal expressions that originally occur in the excerpts to evaluate our methods. In our evaluation methodology, we perform leave-one-out cross validation by randomly selecting excerpts that come with temporal expressions into ground truth. The evaluation measure is computed based on how well the estimated time model for an excerpt describes the empirical model that is estimated from the original expressions in textual description of the excerpt. For a use-case experiment, we make use of the time interval associated with each query as ground truth. We describe the use-case experiment in Section 5.

Implementation. All our methods are implemented in Java. For the temporal annotation, and noun-phrase extraction we use Stanford SUTime toolkit [8]. We additionally use the Weka toolkit to implement the cross-validation framework for the evaluation.

4.2 Evaluation Method, Goals and Measures

Excerpt-time model estimated by our distribution propagation method can simply be understood as a two-dimensional probability distribution that captures the true temporal scope of the excerpt. In this distribution, the time units with high probability represent the temporal focus of the event in the excerpt. To evaluate the quality of the time models estimated for the excerpts, we propose the following steps: 1) From the set of excerpts retrieved for each query, we first distinguish those that originally come with an expression. As described in Section 3, these are then considered as an initial seed set for the distribution propagation. 2) We then perform leave-one-out fold cross-validation by randomly selecting one of these excerpts and ignoring temporal expressions occurring in them. This is considered as set $R_{seed}$. For each fold, we run our algorithm in Section 3 by considering the rest of the excerpts as seed set $R_{seed}$.

3) We truncate and re-normalize the estimated model based on the scope of the empirical model to make them comparable, and then compute the evaluation measures. Final scores are reported by first averaging across the folds for a single query and then taking the mean across all the queries in the test set.

In addition to evaluating the excerpt-time model quality, we also investigate the effects of varying the time granularity for the modeling. Thus, we perform experiments at three fixed time granularities: day, month, and year. For this, we fix the granularity of time and represent the original expression in that granularity. The finest granularity in our experiments is the day level. For experiments at the month granularity, we additionally perform a preprocessing step where we relax an expression originally in the day granularity into its month. For example, “February 2, 2016” is relaxed to “February 2016”. Similarly, for the experiments at year granularity, we relax the original expression occurring at the day and the month granularity into its year. For example, “February 2, 2016” is relaxed to “2016”.

In our experiments we aim to evaluate: 1) the quality of estimated excerpt-time models; 2) the importance of the inter-excerpt relations; 3) the quality of excerpt-time models at different time-granularities. In order to achieve our evaluation goals, we define the following measures to compare our methods:

Model Quality $MQ$. We compute how close an excerpt-time model $E_{time}$ estimated after propagation is to the empirical model $E_{time}$. We define model quality as the KL-divergence between the two. Formally,

$$MQ = \frac{1}{|R_{test}|} \sum_{\epsilon \in R_{test}} -KL(D(\epsilon_{time}||E_{time})) .$$

Intuitively, a method estimating $E_{time}$ more accurately should have a smaller divergence to $E_{time}$ and hence higher $MQ$. As a relative measure, the methods can be ranked according to the model quality.

Generative Power $GP$. With the assumption that the empirical time part of an excerpt is generated from the excerpt-time model, we define the generative power as the likelihood of generating the original time intervals in the time part $E_{time}$ from the estimated time model $E_{time}$ after propagation. To compute $GP$, we adopt the approach to estimate the generative probability of time intervals proposed by Berberich et al. [5], and compute the generative probability as,

$$GP = \sum_{\epsilon \in R_{test}} \left( \sum_{\epsilon_{time}} P(\epsilon||E_{time}) \right) .$$

Intuitively, the better the estimate of $E_{time}$, the higher the likelihood of generating the empirical time part $E_{time}$.

Precision $P$. We test how well our estimated model can predict the empirical time-part of an excerpt. For this, we note that temporal expressions associated to excerpts in the $R_{test}$ come at a certain granularity, i.e., year, month or day. For each excerpt, we generate a ranked list $R_{\epsilon}$ of temporal intervals at a fixed granularity, where the ranking is based on their generative probabilities from the estimated excerpt-time models. Finally, we define a notion of relevance for the time intervals and compute time-precision indicating the quality of the estimated model.

A generated interval is considered relevant if it overlaps with the empirical time-part of the excerpt. Formally, we define the binary relevance $Rel()$ between an empirical time interval $t$ and an esti-
mated time interval \( \hat{t} \) as,
\[
\text{Rel}(t, \hat{t}) = \begin{cases} 
1 & \text{if } \min(t_b, t_e) - \max(t_b, \hat{t}_b) > 0 \\
0 & \text{otherwise}
\end{cases}
\] (13)

where \( t_b \) and \( t_e \) are begin and end time points of \( t \). This formulation is also used in [15]. Using this notion for relevance we formally define precision score for each excerpt as,
\[
P = \frac{1}{|R_t|} \sum_{t \in t_{time}} \sum_{t \in R_t} \text{Rel}(t, \hat{t})
\] (14)

where \(|R_t|\) is the total number of time units in \( R_t \). Using the same notion for relevance, we additionally compute measures like recall, mean average precision \( \text{MAP} \), and normalized discounted cumulative gain \( \text{NDCG} \) scores at 5 and 10 cut off levels which are standard for information retrieval tasks.

4.3 Methods

We categorize our methods into three major types: local time (\( LT \)) based, nearest-neighbor (\( NN \)) based, and distribution propagation (\( DP \)) based methods. The \( LT \) methods take into account only the information in the source documents to estimate excerpt-time models. We consider two methods that use the publication date (\( pd \)), and the surrounding temporal expressions (\( S \)) as described later. On the other hand, both \( NN \) and \( DP \) methods estimate excerpt-time models by leveraging the event graph generated as described in Section 3.2. We define several variants of the methods that consider different combinations of the inter-excerpt relations, i.e., text (\( T \)), position (\( P \)), conceptual (\( N \)), and contextual (\( X \)) similarities as indicated by their suffixes. Finally, we compare a state-of-the-art method proposed by Jatowt et al. [15] that uses term-time associations to predict focus time period of an excerpt. Next, we describe the different methods that we compare in detail.

\( LT-pd \) method assumes that each excerpt taken from a news article describes an event from the time period indicated by its publication date as motivated by several works [3, 6, 23, 29, 30].

\( LT-pdS \) method in addition to the publication date takes into consideration the surrounding temporal expressions of a given excerpt in the source document. Intuitively, temporal expressions denote temporal contextual change. Thus, the closest mentioned temporal expression prior to a given excerpt, and the closest subsequent mentioned expression can indicate its temporal scope. This method however assumes that all excerpts in a document are temporally coherent. This method is motivated from [9, 14].

\( NN-T \) method estimates excerpt-time model from time models of textually similar excerpts. This can be understood as a two-step method where in the first step, for a given excerpt, we compute its textual similarity (as described in Section 3.3) to the other excerpts. Then in the second step, we estimate the excerpt-time model by interpolating empirical models of excerpts weighted by their textual similarity. Similar methods have been used in the past for estimating query-time models. The two-step method is simply realized by iterating once in our distribution propagation algorithm (Algorithm 1) on an event graph that models on the textual relationship between the excerpts. This is motivated from the pseudo-relevance based method presented in [26, 28].

\( NN-TPNX \) method is analogous to the simpler \( NN-T \) however, as an extended method, it leverages all the relationships described in Section 3.3 between the excerpts.

\( DP-TPNX \) method implements the distribution algorithm described in Section 1 by leveraging all the relations described in Section 3.3 to generate the event graph. We additionally compare the \( DP-T \), \( DP-P \), \( DP-N \), \( DP-X \), \( DP-TPN \), \( DP-TPX \), and \( DP-TNX \) variants of this method that leverage different combinations of inter-excerpt relations as indicated by their suffixes.

\( ADJ \) refers to the method proposed by Jatowt et al. [15] for estimating the temporal focus of documents. As a baseline, we extend their approach to estimate focus time for excerpts from news articles. Briefly, this method first generates an undirected weighted graph \( G(V, E) \) where the \( V \) denotes the unique words, and \( E \) denotes word relationships. This graph is then used to estimate word-time association scores. Using these scores, a temporal weight is estimated for each word which is then used to estimate the focus time of each excerpt. From several variants of method proposed by Jatowt et al., we select their best performing method on Web data the uses the \( A_{dir}(w, t), \omega_{w}^{timesK}, \) and \( S_{TF}(\varepsilon, t) \). We refer to their prior work [15] for a full description of the method. We first use their method to predict the focus time interval of an excerpt in the year, month, and the day time granularities. We then estimate an excerpt-time model with the predicted time interval as described in Section 3.2.

\( Rand \) method randomly sets the edge weights in an event graph. This method highlights the quality of the temporal expressions in the seed set.

Parameters. We set the following parameters for our methods. In Equation 1, we set \( \lambda = 0.85 \) and \( \beta = 0.5 \) as motivated by [28]. In Equation 4, \( \sigma \) is set to the average inter-excerpt relation weight estimated for each query. Finally, for smoothing the estimated weight matrix in Equation 6, we set \( \gamma = 0.0005 \) as motivated by [38]. In all the \( DP \) methods, we set the maximum iteration to 15.

4.4 Results

We compare the different methods and report the results in terms of the various measures introduced earlier.

Overall Results from all our methods are shown in Table 3. We find the distribution propagation method \( DP-TPNX \) proves to be the most effective method for estimating time models for excerpts across all measures. The difference in overall result quality of all the methods, in terms of GP, are found to be statistically significant with student’s t test at \( \alpha < 0.001 \). However, the difference between \( NN-T \) and \( NN-TPNX \) is found to be insignificant.

Firstly, we find that the \( ADJ \) proves to be the weakest method to estimate temporal models for excerpts across all granularities. The simplest \( LT-pd \) method is the second weakest at the day and year granularities across all metrics. However, it performs significantly better than the \( ADJ \) method at all three granularities. Leveraging the closest prior and subsequent temporal expressions around an excerpt along with the publication date for estimation of its time model as the \( LT-pdS \) method shows significant improvement over the simpler \( LT-pd \) method. We observe a gain of 10% over \( LT-pdS \) in GP at the day granularity. We also observe a similar improvement across other metrics. The nearest-neighbor methods \( NN-T \) and \( NN-TPNX \) perform significantly better than the \( LT \) methods at the day and year granularities. The more complex \( NN-TPNX \) shows marginal improvement over the \( NN-T \) method at all granularities in terms of GP. Finally, the distribution propagation method \( DP-TPNX \) outperforms other methods across all the time granularities and measures. We find more than 35% improvement from the \( NN \) methods in the day granularity in terms of GP. The method also shows similar significant improvements at the month, and year granularities over the other methods. The \( Rand \) method gives the worst results in the month granularity as compared to the year or day. This is indicative of sparsity of annotations in the input set at this granularity and may not be generalizable.
inter-excerpt relations as leveraged by contextual, conceptual, and text similarity to model the relations in terms of MQ at all granularities. However, a combination of the above approaches can only be estimated between excerpts from a single document. This relation makes strong coherence assumption on the structure of news articles which may not hold for all news articles. Moreover, considering lower ranked documents, we add more irrelevant excerpts to the event graph for a given query. They often do not pose strong relation to the other excerpts, and hence have a negative impact on the overall result quality.

Gain/Loss Analysis. To get insights into the individual queries for which our \textit{DP-TPNX} shows the highest gain and worst loss in terms of P@5 against the best baseline at all three granularities.

At the day granularity, \textit{DP-TPNX} shows the highest gain of +0.63 with a score of 0.88 over the \textit{LT-pd} method which proves to be the best performing baseline with 0.25, for the following query:

\begin{itemize}
  \item \textbf{q1:} Rwandan President Juvénal Habyarimana and Burundi President Cyprien Ntaryamira die when a missile shoots down their jet near Kigali, Rwanda. This is taken as a pretext to begin the Rwandan Genocide.
\end{itemize}

It suffers the worst loss of −0.05, with a score equal to 0.88 while the best performing method for this query with a 0.93 score is \textit{LT-pd} method, for the following query:

\begin{itemize}
  \item \textbf{q2:} Ten-Day War: Fighting breaks out when the Yugoslav People’s Army attacks secessionists in Slovenia.
\end{itemize}

The \textit{NN-T} method is the next best baseline with score of 0.42.

At the month granularity, the query for which the \textit{DP-TPNX} shows the largest gain of +0.43 with a score of 0.70 is as follows:

\begin{itemize}
  \item \textbf{q3:} Troops of Laurent Kabila march into Kinshasa. Zaire is officially renamed Democratic Republic of the Congo.
\end{itemize}

The best baseline for this query is the \textit{LT-pdS} with 0.26 and the next best method is \textit{NN-T} with 0.21 P@5 scores. \textit{DP-TPNX} suffers the worst loss of −0.23 with a score of 0.11 for the following query:

\begin{itemize}
  \item \textbf{q4:} Cold War: The leaders of the Yemen Arab Republic and the People’s Democratic Republic of Yemen announce the unification of their countries as the Republic of Yemen.
\end{itemize}

We find that for this query, the \textit{LT-pdS} method that gets a score of 0.33 and become the best method.

At the year granularity, \textit{DP-TPNX} gets the most gain +0.17 with a score of 0.83 for the following query:

\begin{itemize}
  \item \textbf{q5:} Several explosions at a military dump in Lagos, Nigeria kill more than 1,000.
\end{itemize}

We find that all the other baselines receive an equal score of 0.67 for this query. Our method suffers the worst loss of −0.89 with score 0 for the following query:

\begin{itemize}
  \item \textbf{q6:} Dissolution of Czechoslovakia: The Czech Republic and Slovakia separate in the so-called Velvet Divorce.
\end{itemize}

For this query, we find that the \textit{LT-pdS} is the best method with a score of 0.91 and \textit{LT-pdS} is the second best with a score of 0.89.

### Table 3: Comparison of results from all methods over 100 queries.

<table>
<thead>
<tr>
<th>Method</th>
<th>MQ</th>
<th>GP</th>
<th>P@5</th>
<th>P@10</th>
<th>MAP</th>
<th>Recall</th>
<th>ndcg@5</th>
<th>ndcg@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{DP-TPNX}</td>
<td>-2.227</td>
<td>0.754</td>
<td>0.828</td>
<td>0.736</td>
<td>0.836</td>
<td>0.902</td>
<td>0.847</td>
<td>0.839</td>
</tr>
<tr>
<td>\textit{NN-TPNX}</td>
<td>-14.082</td>
<td>0.395</td>
<td>0.408</td>
<td>0.356</td>
<td>0.182</td>
<td>0.404</td>
<td>0.406</td>
<td>0.377</td>
</tr>
<tr>
<td>\textit{NN-T}</td>
<td>-14.120</td>
<td>0.389</td>
<td>0.407</td>
<td>0.355</td>
<td>0.182</td>
<td>0.404</td>
<td>0.404</td>
<td>0.376</td>
</tr>
<tr>
<td>\textit{Rand}</td>
<td>-14.212</td>
<td>0.388</td>
<td>0.407</td>
<td>0.354</td>
<td>0.182</td>
<td>0.404</td>
<td>0.405</td>
<td>0.375</td>
</tr>
<tr>
<td>\textit{LT-pdS}</td>
<td>-14.614</td>
<td>0.371</td>
<td>0.472</td>
<td>0.449</td>
<td>0.286</td>
<td>0.332</td>
<td>0.467</td>
<td>0.249</td>
</tr>
<tr>
<td>\textit{LT-pd}</td>
<td>-17.807</td>
<td>0.368</td>
<td>0.368</td>
<td>0.185</td>
<td>0.192</td>
<td>0.362</td>
<td>0.310</td>
<td></td>
</tr>
<tr>
<td>\textit{ADJ}</td>
<td>-23.120</td>
<td>0.033</td>
<td>0.033</td>
<td>0.016</td>
<td>0.017</td>
<td>0.032</td>
<td>0.028</td>
<td></td>
</tr>
<tr>
<td>\textit{DP-N}</td>
<td>-2.278</td>
<td>0.477</td>
<td>0.499</td>
<td>0.486</td>
<td>0.503</td>
<td>0.518</td>
<td>0.504</td>
<td>0.493</td>
</tr>
<tr>
<td>\textit{NN-TPNX}</td>
<td>-10.326</td>
<td>0.328</td>
<td>0.337</td>
<td>0.330</td>
<td>0.269</td>
<td>0.309</td>
<td>0.333</td>
<td>0.320</td>
</tr>
<tr>
<td>\textit{NN-T}</td>
<td>-10.327</td>
<td>0.327</td>
<td>0.336</td>
<td>0.330</td>
<td>0.269</td>
<td>0.309</td>
<td>0.333</td>
<td>0.320</td>
</tr>
<tr>
<td>\textit{Rand}</td>
<td>-10.355</td>
<td>0.323</td>
<td>0.334</td>
<td>0.327</td>
<td>0.268</td>
<td>0.309</td>
<td>0.331</td>
<td>0.318</td>
</tr>
<tr>
<td>\textit{LT-pdS}</td>
<td>-10.438</td>
<td>0.332</td>
<td>0.334</td>
<td>0.333</td>
<td>0.284</td>
<td>0.300</td>
<td>0.330</td>
<td>0.325</td>
</tr>
<tr>
<td>\textit{LT-pd}</td>
<td>-13.104</td>
<td>0.280</td>
<td>0.280</td>
<td>0.281</td>
<td>0.219</td>
<td>0.219</td>
<td>0.277</td>
<td>0.267</td>
</tr>
<tr>
<td>\textit{ADJ}</td>
<td>-27.889</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.017</td>
<td>0.017</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>\textit{DP-N}</td>
<td>-1.497</td>
<td>0.751</td>
<td>0.722</td>
<td>0.687</td>
<td>0.741</td>
<td>0.747</td>
<td>0.731</td>
<td>0.699</td>
</tr>
<tr>
<td>\textit{NN-TPNX}</td>
<td>-2.384</td>
<td>0.714</td>
<td>0.691</td>
<td>0.663</td>
<td>0.696</td>
<td>0.696</td>
<td>0.695</td>
<td>0.667</td>
</tr>
<tr>
<td>\textit{NN-T}</td>
<td>-2.415</td>
<td>0.707</td>
<td>0.686</td>
<td>0.652</td>
<td>0.690</td>
<td>0.696</td>
<td>0.691</td>
<td>0.658</td>
</tr>
<tr>
<td>\textit{Rand}</td>
<td>-2.449</td>
<td>0.698</td>
<td>0.675</td>
<td>0.635</td>
<td>0.689</td>
<td>0.696</td>
<td>0.684</td>
<td>0.645</td>
</tr>
<tr>
<td>\textit{LT-pdS}</td>
<td>-3.064</td>
<td>0.739</td>
<td>0.730</td>
<td>0.724</td>
<td>0.694</td>
<td>0.696</td>
<td>0.723</td>
<td>0.709</td>
</tr>
<tr>
<td>\textit{LT-pd}</td>
<td>-4.266</td>
<td>0.675</td>
<td>0.674</td>
<td>0.676</td>
<td>0.620</td>
<td>0.620</td>
<td>0.662</td>
<td>0.651</td>
</tr>
<tr>
<td>\textit{ADJ}</td>
<td>-24.462</td>
<td>0.159</td>
<td>0.160</td>
<td>0.159</td>
<td>0.119</td>
<td>0.122</td>
<td>0.138</td>
<td>0.134</td>
</tr>
</tbody>
</table>

![Figure 2: The effect of increasing the size of the event graph on \textit{DP-TPNX} (blue) and \textit{NN-T} (yellow) methods.](image)
Discussion. First, we discuss the performance of the different methods. The ADJ method turns out to be the least successful method for the time model estimation. The single temporal expression predicted by this method often does not overlap with the expressions in the ground truth. Since our evaluation metrics rely on the overlap, this method receives very low scores at all granularities. The publication date-based LT-pd method is observed to be less effective for estimating excerpt-time models. Similar to the ADJ, this method also relies on a single expression to estimate time models and suffers from sparsity. Further, the assumption that the news articles present information only on events occurring around its publication is repudiated. The LT-pdS method which is a simple extension to the LT-pd estimates much better excerpt-time model. The Rand method benefits over the LT methods from more number of temporal expressions randomly selected. Among the methods that leverage the event graph, the NN-T is motivated from the popular pseudo-relevance feedback models [26] for estimating excerpt models. As expected, this method estimates more accurate time models as compared to the simpler publication date-based methods in terms of MQ. Finally, the distribution propagation improves the estimates of the excerpt-time model through multiple iterations (ideally until convergence).

Next, we discuss the different granularities. At the day granularity, larger interval are represented as a set of days with uniform probability (as described in Section 4.2). Similarly, at year all temporal expressions are relaxed. However, in the month granularity, year-level expression are expanded while the days-level expressions are relaxed. Due to this mixed effect, we find that on average quality, year-level expression are expanded while the days-level expressions are relaxed. However, in the month granularity, larger interval are represented as a set of days with uniform probability. Due to this mixed effect, we find that the results are more pronounced at the day as compared to the year granularity. At the year, due to the relaxation we find that the results are more pronounced at the day as compared to the year granularity. We design a two-stage approach. In Stage 1, leveraging $q_{text}$ we make use of a standard KL-divergence based retrieval model, $R_{text}$, to retrieve a set of top-100 documents. We then generate a set of excerpts $e$ from these documents be fixing an excerpt to a single sentence. In Stage 2, we then generate an event graph $G_{event}(e,r)$ as described in Section 4.1. However, as a slight variation to the previous method, we inject the $q_{text}$ as a special node. Next, we run our distribution propagation algorithm as described in Section 3 to estimate a query-time model $Q_{time}$. We compare our distribution propagation method $DP-TNX$ that takes into consideration the textual (T), conceptual (N), and contextual (X) similarity. Positional similarity becomes inapplicable in this setting. Further, we compare the query text against the lead paragraph of the source document for an excerpt to estimate their contextual similarity. As baselines, we consider the nearest neighbor NN-T method, and the method proposed by Jatowt et al. [15].

Analogous to an excerpt-time model, a query-time model $Q_{time}$ can be understood as a probability distribution over the time units that captures the temporal scope of the event in the query. Intuitively, the units that occur with high probability indicate the salient time period associated with the event. From an estimated query-time model, we generate a rank list of time units with decreasing probabilities in $Q_{time}$. However, unlike previous experiments, we do not truncate but consider the full estimated query-time model where the temporal scope of the query is set to $[\min_{t}, \max_{t}]$. As ground truth, we make leverage the time part $q_{time}$ in the original query. We compare the methods at three time granularity, year, month, and day using the measures define in Section 4.2.

Table 5 shows the results of the experiment. The results are found to be statistically significant with students t test at $\alpha = 0.05$. The $DP-TNX$ method is able to best estimate the occurrence time period at all time granularities across all measures. The result quality is lower than compared to the main experiments because of mainly two reasons. Firstly, we note that most queries come with a single specific temporal expression usually at the day granularity which is considered as ground truth. This is in contrast to the previous setting where pseudo-relevant excerpts may come with multiple expressions at different granularities. Secondly, we find that the accuracy of query focus time is strongly dependent on the quality of the input pseudo-relevant excerpts. Topical drifts in the input excerpts will result in a query focus time that does not match with the ground truth. This makes the use-case problem even harder.

<table>
<thead>
<tr>
<th>Method</th>
<th>P@1</th>
<th>P@5</th>
<th>P@10</th>
<th>MAP</th>
<th>ndcg@5</th>
<th>ndcg@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DP-TNX$</td>
<td>0.46</td>
<td>0.13</td>
<td>0.07</td>
<td>0.52</td>
<td>0.54</td>
<td>0.55</td>
</tr>
<tr>
<td>NN-T</td>
<td>0.36</td>
<td>0.13</td>
<td>0.08</td>
<td>0.46</td>
<td>0.49</td>
<td>0.51</td>
</tr>
<tr>
<td>ADJ</td>
<td>0.18</td>
<td>0.09</td>
<td>0.06</td>
<td>0.30</td>
<td>0.32</td>
<td>0.36</td>
</tr>
<tr>
<td>$DP-TNX$</td>
<td>0.28</td>
<td>0.09</td>
<td>0.05</td>
<td>0.31</td>
<td>0.31</td>
<td>0.32</td>
</tr>
<tr>
<td>NN-T</td>
<td>0.22</td>
<td>0.08</td>
<td>0.05</td>
<td>0.27</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>ADJ</td>
<td>0.09</td>
<td>0.05</td>
<td>0.04</td>
<td>0.16</td>
<td>0.16</td>
<td>0.19</td>
</tr>
<tr>
<td>$DP-TNX$</td>
<td>0.19</td>
<td>0.16</td>
<td>0.12</td>
<td>0.19</td>
<td>0.17</td>
<td>0.21</td>
</tr>
<tr>
<td>NN-T</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.15</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td>ADJ</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>
6. CONCLUSION AND FUTURE WORK

We proposed a novel problem to estimate time models for excerpts that are extracted from news articles. To address this problem, we proposed a distribution framework that extends a popular semi-supervised label propagation algorithm to propagate time models from excerpts that come with temporal expressions to those that are strongly related but missing annotations. In our experiments, we found that our method estimates most accurate time models as compared to several baselines. We compare the methods in terms of existing precision, recall, NDCG measures, and two new Model Quality and Generative Power measures.

One interesting perspective on the time models can come from the direction of generating temporal embeddings for excerpts. Recently, leveraging word embeddings [22] have shown significant improvements in various text-based applications. The time model estimation presented may be considered as the first step of embedding excerpts into a space that models the time dimension of information content. As a future direction, we plan to design advanced methods to identify better source of temporal information than relying on blind feedback, and output dense representation of excerpts to better capture its temporal semantics.

References