Linking Wikipedia Events to Past News

Arunav Mishra Dragan Milchevski Klaus Berberich
Max Planck Institute for Informatics
Saarbrücken, Germany
{amishra, dmilchev, kberberi}@mpi-inf.mpg.de

ABSTRACT

We consider the task of linking Wikipedia events to relevant news articles from the past. Descriptions of events are abundant in Wikipedia and systematically curated in year, decade, and century articles. To address this task, we develop a two-stage cascade approach that builds a query model from temporal expressions in a set of initially retrieved documents. As baselines we consider several methods that integrate publication dates and/or temporal expressions into a language modeling approach. Our experimental evaluation on 50 randomly sampled Wikipedia events with crowd-sourced relevance assessments shows that the two-stage cascade approach outperforms the baselines. Our experimental testbed of queries and relevance assessments is made publicly available.

Categories and Subject Descriptors

H.3.3 [**Information Search & Retrieval**]: Search process

Keywords

Temporal Information Retrieval, Linking, Wikipedia

1. INTRODUCTION

The free encyclopedia Wikipedia and online new articles have emerged as prominent yet orthogonal sources of information on events of global or local importance. While Wikipedia articles summarize past events and often abstract from fine-grained details that mattered when the events happened, news articles (on the Web) provide contemporary reports on the events.

One class of articles in Wikipedia, which are of particular interest to us in this work, are those discussing events that happened in a specific year, decade, or century. These articles contain a list of events, each consisting of a textual description of what happened and a date indicating when it happened. Consider, as a concrete example, the article `http://en.wikipedia.org/wiki/1987`, which lists the following as two of the seminal events in that year:

January 3 1987 : Aretha Franklin becomes the first woman inducted into the Rock and Roll Hall of Fame

October 11 1987 : The first National Coming Out Day is held in celebration of the second National March on Washington for Lesbian and Gay Rights.

What is often missing, though, are links to (contemporary) news articles that elaborate on the specific past event. With such links in place, it would be easy for users to browse past events, see how they were perceived when they happened, and contrast this to their coverage in Wikipedia. The issue here is not that such news articles are unavailable. News articles have been archived for a long time, and newspaper archives available today go back decades if not centuries – the archives of The New York Times, as a concrete instance, contain articles from as early as 1851.

In this paper, we address the problem of automatically linking Wikipedia events, like the ones above, to news articles from the past. We cast this problem into a novel retrieval task: Given a Wikipedia event consisting of a short textual description and a date, our objective is to retrieve news articles providing details on the event, which could be provided as a background reference.

Challenges of this task include that the textual descriptions of Wikipedia events are typically verbose, so that standard retrieval models are prone to topic drift. Also, they are not able to leverage the dates provided with Wikipedia events and relate them to *publication dates* and *temporal expressions* that come with news articles.

Contributions that we make in this work are: **(i)** a novel retrieval task that aims at linking Wikipedia events to news articles from the past; **(ii)** a two-stage cascade approach to address the task, which builds a temporal query model for reranking from textually relevant documents; **(iii)** an experimental comparison of our method against several baselines that make use of publication dates and/or temporal expressions conducted on **(iv)** a testbed consisting of 50 randomly sampled Wikipedia events from 1987 to 2007 for which we collected crowd-sourced relevance assessments made publicly available.

Organization. We discuss related work in the following Section 2. Baselines and our two-stage cascade approach are described in detail in Section 3. Section 4 introduces our experimental testbed and the results obtained on it. We conclude and point out next steps in Section 5.

TAIA'14 July 11, 2014, Gold Coast, Australia.

2. RELATED WORK

Linking of different kinds of document collections is a first line of prior research related to our work. Henzinger et al. [4], as the earliest work, identify news articles related to what's currently discussed on TV based on the embedded closed captions. Bron et al. [3], more recently, aim at linking textually rich news articles to sparsely annotated historic videos. Ikeda et al. [5] and Tsagkias et al. [10] look into connecting news articles to social media and vice versa, for instance, retrieving relevant blog posts for a newspaper article at hand. The focus in all of these works has been on coping with the disparate quantities of text in the source and target document collection and bridging the vocabulary gap between them. Publication dates have only been used for filtering. In contrast, our approach leverages both publication dates and temporal expressions from documents' contents to improve linking effectiveness.

Temporal Information Retrieval, as a second line of related work, has looked into making use of temporal information such as publication dates and temporal expressions to improve effectiveness in ad-hoc retrieval. Li and Croft [7] introduce a time-dependent exponential document prior in a language modeling approach. Other document priors motivated by findings from cognitive science have been considered more recently by Peetz et al. [8]. Berberich et al. [2] made use of temporal expressions in documents to better deal with explicitly temporal queries. Other work such as Jones and Diaz [6] and Peetz et al. [9] has used temporal information for query modeling in temporal query classification and ad-hoc retrieval, respectively. Both aforementioned approaches analyze the distribution of publication dates. In the former work to derive features for classification, in the latter to estimate a refined query model for a second round of retrieval. Our approach, besides targeting a different task, also considers temporal expressions in initially retrieved documents for query modeling.

3. LINKING WIKIPEDIA EVENTS TO PAST NEWS

In this section, we describe different approaches to address the linking task at hand. These include existing methods that leverage publication dates and/or temporal expressions as well as a novel two-stage cascade approach that builds a query model from the temporal expressions in pseudo-relevant documents. As a foundation, we begin by describing our formal notation.

3.1 Model

Let D denote our document collection. A document $d^t \in D$ comes with a publication date $t \in T$, where T is our time domain. The document d^t consists of a textual part d_{text} and a temporal part d_{time}. The textual part d_{text} is a bag of words drawn from a vocabulary V. The temporal part d_{time} is a bag of temporal expressions.

Times (chronons) in T are assumed to reflect the time passed (to pass) since (until) a reference date such as the UNIX epoch. In this work, we model temporal expressions simply as time intervals $[b, e] \in T \times T$ and assume $b \leq e$. For time intervals $[b, e]$ with coinciding boundaries (i.e., $b = e$) we use $[b]$ as a shorthand. When mapping temporal expressions contained in a document to this representation, we map onto the largest plausible time interval. Thus, in May 2014, as a concrete example would be mapped onto [2014.05.01, 2014.05.31]. Likewise, in the evening of March 31st 2011 would be mapped onto [2011.03.31].

Sometimes it will be convenient to treat the entire document collection as a single (coalesced) document. We use D_{text} and D_{time} to refer to the corresponding textual part and temporal part, respectively.

Given a Wikipedia event like the ones above, we derive a query q from it as follows: The textual part q_{text} is obtained directly from the description of the event. The temporal part q_{time} is obtained from the indicated date. Note that, for the scope of this work, we only consider Wikipedia events that indicate a specific day as their date. Hence, while $q_{time} \in T \times T$, we refer to q_{time} as a single time when convenient.

3.2 Time-Aware Language Models

We now develop several baseline methods to tackle the linking task, building on a unigram language model with Dirichlet smoothing [11] as a foundation.

Text Only. When considering only the textual parts of the query and documents, we use the following query likelihood to rank documents

$$P(q \mid d^t) = \prod_{v \in q_{text}} \frac{P(v \mid d_{text}) + \mu \cdot P(v \mid D_{text})}{|d_{text}| + \mu} \qquad (1)$$

with $P(v \mid d_{text})$ and $P(v \mid D_{text})$ as maximum-likelihood estimates and μ as a smoothing parameter.

Publication Dates. Intuitively, documents published around the time when the query event happened are more likely to discuss the event in detail. Our second methods implements this idea by relating documents' publication dates to the temporal part of the query q_{time}. It thus ranks documents according to the probability

$$P(q \mid d^t) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid t) \; . \qquad (2)$$

Here, the first factor is estimated according to Equation (1), the second factor is defined as

$$P(q_{time} \mid t) = \frac{1}{1 + e^{\,r|q_{time} - t|}} \; . \qquad (3)$$

With this sigmoid function, we thus favor documents published shortly before or after the time q_{time} when the query event at hand took place.

Temporal Expressions can be another strong indicator for whether a document discusses the query event at hand. Thus, if many of a document's temporal expressions refer to the time period when the event happened, chances are that the document discusses the event. Our third method ranks documents according to

$$P(q \mid d^t) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid d_{time}) \; . \qquad (4)$$

Again, the first factor follows Equation (1), the second is estimated, based on a simplified variant of [2], as

$$P(q_{time} \mid d_{time}) = \frac{1}{|d_{time}|} \sum_{[b,\, e] \in d_{time}} \frac{\mathbb{1}(q_{time} \in [b, e])}{e - b + 1} \; . \qquad (5)$$

To avoid zero probabilities, if none of a document's temporal expressions includes the query time q_{time}, we smooth this estimate by interpolating with a small constant. According to the above, a document yields a high probability $P(q_{time} \mid d_{time})$ if many of its temporal expressions, or more

precisely the corresponding time intervals, are at a fine temporal granularity and include the query time q_{time}.

Publication Dates + Temporal Expressions. As a final baseline method, we combine the above approaches and rank documents according to

$$P(q \mid d^t) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid t) \cdot P(q_{time} \mid d_{time}) \quad (6)$$

with factors as defined in Equations (1), (3), and (5).

3.3 Two-Stage Cascade Approach

Having established our baselines, we now describe our *two-stage cascade approach*. As a key difference to the baselines, our novel approach performs an initial round of retrieval, using the unigram language model with Dirichlet smoothing described in Equation (1), and estimates a *temporal query model* from the top-k documents retrieved, thus treating them as pseudo-relevant. It then re-ranks the top-K documents (with $k \leq K$) from the initial round of retrieval taking into account their divergence from the temporal query model, their publication date, and their fit to the textual description of the query event.

Intuitively, by estimating a temporal query model from pseudo-relevant documents, we cope with an overly specific query time q_{time} and instead consider salient temporal expressions related to the query event. This is expected to be particularly helpful for events that did not receive extensive coverage when they happened or whose ramifications linger on for a long time.

Stage 1. Let $R_k \subseteq D$ denote the set of top-k documents retrieved based on $P(q_{text} \mid d_{text})$. From those pseudo-relevant documents, we estimate a temporal query model Q_{time} as

$$P(\tau|Q_{time}) = \sum_{d \in R_k} \frac{P(q_{text} \mid d_{text})}{\sum_{d' \in R_k} P(q_{text} \mid d'_{text})} \cdot P(\tau|d_{time}) \; . \quad (7)$$

Akin to Equation (5), the second factor is estimated as

$$P(\tau \mid d_{time}) = \frac{1}{|d_{time}|} \sum_{[b,\, e] \in d_{time}} \frac{1}{|\tau|} \cdot \frac{\mathbb{1}(\tau \in [b,\, e])}{e - b + 1} \quad (8)$$

and corresponds to a temporal document model. Equations (7) and (8) describe probability distributions over times $\tau \in T$. The temporal document model assigns higher probability to times that are mentioned more often or through more specific temporal expressions in the document. Our temporal query model then aggregates these probabilities from the pseudo-relevant documents in R_k, taking their initial query likelihoods into account. As a result, our temporal query model assigns high probability to times mentioned often in textually relevant documents through specific temporal expressions.

Stage 2. In the second stage of our cascade approach, documents from R_K, the set of top-K documents retrieved initially, are re-ranked. As a first factor for re-ranking, we take the Kullback-Leibler divergence between the temporal query model estimated in Stage 1 and a smoothed temporal document model into account, i.e.:

$$D(Q_{time} \| d_{time}) = \sum_{\tau \in Q_{time}} P(\tau|Q_{time}) \cdot \log \frac{P(\tau \mid Q_{time})}{P'(\tau \mid d_{time})} \; . \quad (9)$$

Smoothing with the document collection as

$$P'(\tau \mid d_{time}) = \lambda \cdot P(\tau \mid d_{time}) + (1 - \lambda) \cdot P(\tau \mid D_{time}) \; . \quad (10)$$

is again required to avoid zero probabilities.

As a second factor, we consider the publication dates of documents. To this end, we determine the Kullback-Leibler divergence between q_{time} and the publication date t, in analogy to Equation (3), which in our specific setting, where q_{time} refers to a single time, simplifies to

$$D(q_{time} \parallel t) = \log(1 + e^{\,r|q_{time} - t|}) \; . \quad (11)$$

Finally, as a third factor, we also take the textual parts of the query and the document into account. To this end, we consider the Kullback-Leibler divergence $D(q_{text} \parallel d_{text})$ between their Dirichlet smoothed unigram language models estimated in analogy to Equation (1).

Putting it together, we re-rank documents according to

$$\alpha \cdot D(Q_{time} \| d_{time}) + \beta \cdot D(q_{time} \| t) + \gamma \cdot D(q_{text} \| d_{text})$$

where α, β, and γ are tunable parameters.

4. EXPERIMENTS

Next, we provide details on the setup and results of our experimental evaluation.

4.1 Setup

Methods under comparison are: (**a**) *LM* as a unigram language model with Dirichlet smoothing according to Equation (1) (using μ = 1,000); (**b**) *LM+P* as the method integrating documents' publication dates according to Equation (3) (using $r = 0.015$ [9]), (**c**) *LM+T* as the method integrating temporal expressions alongside unigrams according to Equation (5), (**d**) *LM+PT* as the approach considering both publication dates and temporal expressions according to Equation (6), and (**e**) *CM* as our two-stage cascade approach. Here, we found that building the query model from the top-10 initially retrieved documents and using it to re-rank the top-30 initially retrieved documents gives the best results. We set the mixing parameters as $\alpha = 0.20$, $\beta = 0.60$, and $\gamma = 0.20$. The smoothing parameter is set as $\lambda = 0.85$.

Implementation. All methods were implemented in Java. We used Stanford CoreNLP to tokenize documents and annotate temporal expressions in their contents.

Document Collection. We use The New York Times Annotated Corpus [1], which contains about 2 million documents published between 1987 and 2007.

Queries. We use *The English Wikipedia* dump released on February 3rd 2014 and randomly sample 50 Wikipedia events as queries from the articles for the years `1987` to `2007`.

Relevance Assessments were collected using the CrowdFlower platform. We pooled top-10 results for the methods under comparison, which resulted in 1,297 unique query-document pairs. We asked assessors to judge whether the document was **(0)** irrelevant, **(1)** somewhat relevant, or **(2)** highly relevant to the given query. Our instructions said that a document only be considered highly relevant if its main topic was the event given as a query. Each query-document pairs was judged by at least three assessor. We paid \$0.03 per batch of five query-document pairs.

For the sake of reproducibility, our queries and relevance assessments are made publicly available at:

`http://www.mpi-inf.mpg.de/~amishra/TAIA2014/`.

	LM	*LM +P*	*LM +T*	*LM +PT*	*CM*
MAP	0.35	0.43	0.40	0.42	**0.45**
P@5	0.55	0.63	0.61	0.61	**0.66**
P@10	0.48	0.57	0.54	0.54	**0.58**
NDCG@5	0.53	0.58	0.58	0.60	**0.60**
NDCG@10	0.54	0.62	0.60	0.62	**0.63**

Table 1: Retrieval effectiveness

4.2 Results

We measure retrieval effectiveness using mean average precision (MAP) as well as Precision (P) and Normalized Discounted Cumulative Gain (NDCG) at cut-offs 5 and 10. For MAP and P we consider a document relevant to a query if the majority of assessors judged it with label **(1)** or **(2)**. For NDCG we plug in the mean label assigned by assessors.

Effectiveness Results. Table 1 lists retrieval effectiveness measures for our five methods under comparison. It can be seen that *CM* consistently outperforms the baselines across all effectiveness measures, proving to be the most effective approach for the linking task. Comparing the baselines, we see that both *LM+P* and *LM+T* yield improvements over the text-only approach *LM*. Thus, considering either kind of temporal information helps with our linking task. Unfortunately, their effects are not additive, as can be seen from the performance of *LM+PT*, which does not yield a consistent improvement over the two and sometimes even performs worse than *LM+P*.

Gains & Losses. To get some insight into where *CM*'s improvements come from, we perform a gain/loss analysis based on P@10, ranking queries according to the difference between *CM* and the best performing baseline. The biggest gain of +0.2 (with *LM+P* and *LM+TP* as the best performing competitors) is observed for the query

March 19 2002 : US war in Afghanistan: Operation Anaconda ends after killing 500 Taliban and Al-Qaeda fighters, with 11 allied troop fatalities.

For this relatively verbose query, the two-stage cascade approach yields a P@10 of 0.8. Here, the temporal query model, shifts focus from the specific day to March 2002, which for this query (related to a relatively short operation) turns out beneficial.

We observe the biggest loss of −0.3 for the query

February 27 1991 : President Bush declares victory over Iraq and orders a cease-fire.

Interestingly, the best performing baseline in this case is *LM+PT*, which achieves a perfect P@10 of 1.0. In contrast to that, the *LM* achieves only a P@10 of 0.2. Here, the text-only baseline suffers from the ambiguity of the query (i.e., multiple presidents called Bush and multiple wars in Iraq) and is unable to focus its results on the right time period. As a building block, it also negatively affects the performance of our two-stage cascade approach *CM*, which through its re-ranking still achieves an acceptable P@10 of 0.7.

Easy & Hard Query Events. Finally, we identify easy and hard query events in our testbed. The easiest one, having the highest minimum P@10 across all methods, is

August 4 1993 : A federal judge sentences Los Angeles Police Department officers Stacey Koon and Laurence Powell to 30 months in prison for violating motorist Rodney King's civil rights.

LM, as the worst performing method, still achieves a P@10 of 0.8 for this query event. Likewise, we identify

May 3 1989 : Cold War - Perestroika - The first McDonald's restaurant in the USSR begins construction in Moscow. It will open on 31 January 1990.

as the hardest query event in our testbed, for which all methods under consideration fail to report any relevant document and achieve a P@10 of 0.0.

5. CONCLUSION AND FUTURE WORK

In this work, we have considered the novel linking task of identifying news articles from the past relevant to a given Wikipedia event. We made first strides at addressing it and observed that our two-stage cascade approach outperformed simpler baselines.

Future Work. Our focus in this work has been on exploiting temporal information. When inspecting Wikipedia events, it is evident that they often contain additional semantics that can be exploited. Thus, most of them mention at least one named entity, and many of them mention the geographic location where the event happened. Making the retrieval model aware of named entities and geographic locations is a promising direction for future work. Also, all Wikipedia events that we considered came with a date and happened during the time period covered by our news archive. It will be interesting to see how retrieval performance changes once we look at longer events (e.g., wars lasting a couple of years) and/or consider events that happened in the far past (e.g., in the 18th century).

6. REFERENCES

[1] The New York Times Annotated Corpus `http://corpus.nytimes.com`.

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

[3] M. Bron, B. Huurnink, and M. de Rijke. Linking archives using document enrichment and term selection. In *TPDL*, 2011.

[4] M. R. Henzinger, B.-W. Chang, B. Milch, and S. Brin. Query-free news search. *World Wide Web*, 8(2):101–126, 2005.

[5] D. Ikeda, T. Fujiki, and M. Okumura. Automatically linking news articles to blog entries. In *AAAI Spring Symposium*, 2006.

[6] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25, 2007.

[7] X. Li and W. B. Croft. Time-based language models. In *CIKM*, 2003.

[8] M.-H. Peetz and M. de Rijke. Cognitive temporal document priors. In *ECIR*, 2013.

[9] M.-H. Peetz, E. Meij, and M. de Rijke. Using temporal bursts for query modeling. *Information Retrieval*, 17, 2014.

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

[11] C. Zhai. Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1, 2008.