Leveraging Semantic Annotations to Link Wikipedia and News Archives

Arunav Mishra(✉) and Klaus Berberich

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

Abstract. The incomprehensible amount of information available online has made it difficult to retrospect on past events. We propose a novel linking problem to connect excerpts from Wikipedia summarizing events to online news articles elaborating on them. To address this linking problem, we cast it into an information retrieval task by treating a given excerpt as a user query with the goal to retrieve a ranked list of relevant news articles. We find that Wikipedia excerpts often come with additional semantics, in their textual descriptions, representing the time, geolocations, and named entities involved in the event. Our retrieval model leverages text and semantic annotations as different dimensions of an event by estimating independent query models to rank documents. In our experiments on two datasets, we compare methods that consider different combinations of dimensions and find that the approach that leverages all dimensions suits our problem best.

1 Introduction

Today in this digital age, the global news industry is going through a drastic shift with a substantial increase in online news consumption. With new affordable devices available, general users can easily and instantly access online digital news archives using broadband networks. As a side effect, this ease of access to overwhelming amounts of information makes it difficult to obtain a holistic view on past events. There is thus an increasing need for more meaningful and effective representations of online news data (typically collections of digitally published news articles).

The free encyclopedia Wikipedia has emerged as a prominent source of information on past events. Wikipedia articles tend to summarize past events by abstracting from fine-grained details that mattered when the event happened. Entity profiles in Wikipedia contain excerpts that describe events that are seminal to the entity. As a whole, they give contextual information and can help to build a good understanding of the causes and consequences of the events.

Online news articles are published contemporarily to the events and report fine-grained details by covering all angles. These articles have been preserved for a long time as part of our cultural heritage through initiatives taken by media houses, national libraries, or efforts such as the Internet Archive. The archives of The New York Times, as a concrete example, go back until 1851.

N. Ferro et al. (Eds.): ECIR 2016, LNCS 9626, pp. 30–42, 2016.
DOI: 10.1007/978-3-319-30671-1_3

Individually, both Wikipedia and news articles are ineffective in providing complete clarity on multi-faceted events. On one hand, brief summaries in Wikipedia that abstract from the fine-grained details, make it difficult to understand all dimensions of an event. On the other hand, news articles that report a single story from a larger event do not make its background and implications apparent. What is badly missing are connections between excerpts from Wikipedia articles summarizing events and news articles. With these connections in place, a Wikipedia reader can jump to news articles to get the missing details.

Table 1. Examples of Wikiexcerpts

No.	Wikiexcerpt
1	**Jaber Al-Ahmad Al-Sabah:** After much discussion of a border dispute between **Kuwait** and **Iraq**, **Iraq** invaded its smaller neighbor on *August 2, 1990* with the stated intent of annexing it. Apparently, task of the invading **Iraqi** army was to capture or kill Sheikh Jaber.
2	**Guam:** The United States returned and fought the Battle of **Guam** on *July 21, 1944*, to recapture the island from Japanese military occupation. More than 18,000 Japanese were killed as only 485 surrendered. Sergeant Shoichi Yokoi, who surrendered in *January 1972*, appears to have been the last confirmed Japanese holdout in Guam.

We propose the following linking problem: *Given an excerpt from Wikipedia, coined Wikiexcerpt, summarizing an event, how can we identify past news articles providing contemporary accounts?* We cast this research question into a query-based retrieval task: given a source text, as a user query, retrieve a ranked list of documents that should be linked to it. In this task, the user poses the Wikiexcerpt as a query and the goal is to retrieve relevant articles from a news collection. Two concrete examples of Wikiexcerpts are given in Table 1.

Standard document retrieval models for keyword queries rely on syntactic matching and are ineffective for our task. Due to the verbosity of the Wikiexcerpts, they are prone to topic drift and result in lower retrieval quality. The Wikiexcerpts also contain additional semantics like temporal expressions, geolocations, and named entities which can be leveraged to identify relevant documents. Making the retrieval model aware of these semantic annotations so as to identify contemporary and relevant documents is not straightforward.

Our approach integrates text, time, geolocations, and named entities in a principled manner, treating them as independent dimensions of event query and ranks documents by comparing them to the query event along these dimensions.

Contributions made in this work are as follows: **(1)** a novel *linking task* to connect Wikipedia excerpts to news articles; **(2)** novel query modeling techniques to estimate independent models for *text*, *time*, *geolocations*, and *named entities* in a query; **(3)** a framework to combine independent query models to rank documents.

Organization. In Sect. 2, we first introduce our notations. Then Sect. 3, gives details on how we estimate the independent query models. Conducted experiments and their results are described in Sect. 4. Section 5 puts our work in context with existing prior research. Finally, we conclude in Sect. 6.

2 Model

Each document d in our document collection C consists of a textual part d_{text}, a temporal part d_{time}, a geospatial part d_{space}, and an entity part d_{entity}. As a bag of words, d_{text} is drawn from a fixed vocabulary V derived from C. Similarly, d_{time}, d_{space} and d_{entity} are bags of temporal expressions, geolocations, and named-entity mentions respectively. We sometimes treat the entire collection C as a single coalesced document and refer to its corresponding parts as C_{text}, C_{time}, C_{space}, and C_{entity}. In our approach, we use the Wikipedia Current Events Portal[1] to distinguish event-specific terms by coalescing into a single document d_{event}. Time unit or *chronon* τ indicates the time passed (to pass) since (until) a reference date such as the UNIX epoch. A temporal expression t is an interval $[tb, te] \in T \times T$, in time domain T, with begin time tb and end time te. Moreover, a temporal expression t is described as a quadruple $[tb_l, tb_u, te_l, te_u]$ [5] where tb_l and tb_u gives the plausible bounds for begin time tb, and te_l and te_u give the bounds for end time te. A geospatial unit l refers to a geographic point that is represented in the geodetic system in terms of latitude (lat) and longitude ($long$). A geolocation s is represented by its minimum bounding rectangle (MBR) and is described as a quadruple $[tp, lt, bt, rt]$. The first point (tp, lt) specifies the top-left corner, and the second point (bt, rt) specifies the bottom-right corner of the MBR. We fix the smallest MBR by setting the resolution $[resol_{lat} \times resol_{long}]$ of space. A named entity e refers to a location, person, or organization from the YAGO [15] knowledge base. We use YAGO URIs to uniquely identify each entity in our approach. A query q is derived from a given Wikiexcerpt in the following way: the text part q_{text} is the full text, the temporal part q_{time} contains explicit temporal expressions that are normalized to time intervals, the geospatial part q_{space} contains the geolocations, and the entity part q_{entity} contains the named entities mentioned. To distinguish contextual terms, we use the textual content of the source Wikipedia article of a given Wikiexcerpt and refer to it as d_{wiki}.

3 Approach

In our approach, we design a *two-stage* cascade retrieval model. In the first stage, our approach performs an initial round of retrieval with the text part of the query to retrieve top-K documents. It then treats these documents as pseudo-relevant and expands the *temporal*, *geospatial*, and *entity* parts of the query. Then, in the second stage, our approach builds independent query models using the expanded query parts, and re-ranks the initially retrieved K documents based on their divergence from the final integrated query model. As output it then returns top-k documents ($k < K$). Intuitively, by using pseudo-relevance feedback to expand query parts, we cope with overly specific (and sparse) annotations in the original query and instead consider those that are salient to the query event for estimating the query models.

[1] http://en.wikipedia.org/wiki/Portal:Currentevents.

For our linking task, we extend the KL-divergence framework [27] to the text, time, geolocation, and entity dimensions of the query and compute an overall divergence score. This is done in two steps: First, we independently estimate a query model for each of the dimensions. Let Q_{text} be the unigram *query-text model*, Q_{time} be the *query-time model*, Q_{space} be the *query-space model*, and Q_{entity} be the *query-entity model*. Second, we represent the overall query model Q as a joint distribution over the dimensions and exploit the additive property of the KL-divergence to combine divergence scores for query models as,

$$\begin{aligned} KL(Q \,||\, D) = KL(Q_{text} \,||\, D_{text}) + KL(Q_{time} \,||\, D_{time}) \\ + KL(Q_{space} \,||\, D_{space}) + KL(Q_{entity} \,||\, D_{entity}). \end{aligned} \tag{1}$$

In the above equation, analogous to the query, the overall document model D is also represented as the joint distribution over D_{text}, D_{time}, D_{space}, and D_{entity} which are the independent document models for the dimensions.

The KL-divergence framework with the independence assumption gives us the flexibility of treating each dimension in isolation while estimating query models. This would include using different background models, expansion techniques with pseudo-relevance feedback, and smoothing. The problem thus reduces to estimating query models for each of the dimensions which we describe next.

Query-Text Model. Standard likelihood-based query modeling methods that rely on the empirical terms become ineffective for our task. As an illustration, consider the first example in Table 1. A likelihood-based model would put more stress on $\{$*Iraq*$\}$ that has the highest frequency, and suffer from topical drift due to the terms like $\{$*discussion*, *border*, *dispute*, *Iraq*$\}$. It is hence necessary to make use of a background model that emphasizes event-specific terms.

We observe that a given q_{text} contains two factors, first, terms that give background information, and second, terms that describe the event. To stress on the latter, we combine a query-text model with a background model estimated from: **(1)** the textual content of the source Wikipedia article d_{wiki}; and **(2)** the textual descriptions of events listed in the Wikipedia Current Events portal, d_{event}. The d_{wiki} background model puts emphasis on the contextual terms that are discriminative for the event, like $\{$*Kuwait*, *Iraq*, *Sheikh*, *Jaber*$\}$. On the other hand, the background model d_{event} puts emphasis on event-specific terms like $\{$*capture*, *kill*, *invading*$\}$. Similar approaches that combine multiple contextual models have shown significant improvement in result quality [24,25].

We combine the query model with a background model by linear interpolation [28]. The probability of a word w from the Q_{text} is estimated as,

$$P(w|Q_{text}) = (1-\lambda)\cdot P(w|q_{text}) + \lambda\cdot\big[\beta\cdot P(w|d_{event}) + (1-\beta)\cdot P(w|d_{wiki})\big] \ . \tag{2}$$

A term w is generated from the background model with probability λ and from the original query with probability $1 - \lambda$. Since we use a subset of the available terms, we finally re-normalize the query model as in [20]. The new generative probability $\hat{P}(w \mid Q_{text})$ is computed as,

$$\hat{P}(w \mid Q_{text}) = \frac{P(w \mid Q_{text})}{\sum_{w' \in V} P(w' \mid Q_{text})}. \tag{3}$$

Query-Time Model. We assume that a temporal expression $t \in q_{time}$ is sampled from the query-time model Q_{time} that captures the salient periods for an event in a given q. The generative probability of any time unit τ from the temporal query model Q_{time} is estimated by iterating over all the temporal expressions $t = [tb_l, tb_u, te_l, te_u]$ in q_{time} as,

$$P(\tau \mid Q_{time}) = \sum_{[tb,te] \in q_{time}} \frac{\mathbb{1}(\tau \in [tb_l, tb_u, te_l, te_u])}{|[tb_l, tb_u, te_l, te_u]|} \tag{4}$$

where the $\mathbb{1}(\cdot)$ function returns 1 if there is an overlap between a time unit τ and an interval $[tb_l, tb_u, te_l, te_u]$. 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 q_{time}.

The query-time model estimated so far has hard temporal boundaries and suffers from the issue of *near-misses*. For example, if the end boundary of the query-time model is “10 January 2014” then the expression “11 January 2014” in a document will be disregarded. To address this issue, we perform an additional *Gaussian smoothing*. The new probability is estimated as,

$$\hat{P}(\tau \mid Q_{time}) = \sum_{t \in T \times T} G_\sigma(t) \cdot P(\tau \mid Q_{time}) \tag{5}$$

where G_σ denotes a Gaussian kernel that is defined as,

$$G_\sigma(i) = \frac{1}{2\pi\sigma^2} \cdot exp\left(-\frac{(tb_l, tb_u)^2 + (te_l, te_u)^2}{2\sigma^2}\right) . \tag{6}$$

Gaussian smoothing computes a weighted average of adjacent units with a weight decreasing with the spatial distance to center position τ in two dimensional space. The σ in the kernel defines the neighborhood size and can be empirically set. As a result of the Gaussian smoothing, the temporal boundaries are blurred, spilling some probability mass to adjacent time units. Finally, since we use only a subset of temporal expressions we re-normalize similar to Eq. 3.

Query-Space Model. We assume that a user samples a geolocation s from query-space model Q_{space} to generate q_{space}. The query-space model captures salient geolocations for the event in a given Wikiexcerpt. The generative probability of any spatial unit l from the query-space model Q_{space} by iterating over all geolocations $[tp, lt, bt, rt] \in q_{space}$ is estimated as

$$P(l \mid Q_{space}) = \sum_{(tp,lt,bt,rt) \in q_{space}} \frac{\mathbb{1}(l \in [tp, lt, bt, rt])}{|[tp, lt, bt, rt]|} . \tag{7}$$

Analogous to the Eq. 4, the $\mathbb{1}(\cdot)$ function returns 1 if there is an overlap between a space unit l and a MBR as $[tp, lt, bt, rt]$. Intuitively, query-space model assigns higher probability to l if it overlaps with a larger number of more specific (MBR with smaller area) geolocations in q_{space}. As the denominator, it is easy to compute $|\,[tp, lt, bt, rt]\,|$ as $|s| = (rt - lt + resol_{lat}) * (tp - bt + resol_{long})$. Addition of the small constant ensures that for all s, $|s| \, > \, 0$.

Similar to the query-time model, to address the issue of near misses we estimate $\hat{P}(l|Q_{Space})$ that additionally smooths $P(l|Q_{space})$ using a Gaussian kernel as described in Eq. 5 and also re-normalize as per Eq. 3.

Query-Entity Model. The query-entity model Q_{entity} captures the entities that are salient to an event and builds a probability distribution over an entity space. To estimate Q_{entity} we make use of the initially retrieved pseudo-relevant documents to construct a background model that assigns higher probability to entities that are often associated with an event. Let D_R be the set of pseudo-relevant documents. The generative probability of entity e is estimated as,

$$P(e \mid Q_{entity}) = (1 - \lambda) \; \cdot \; P(e \mid q_{entity}) + \lambda \; \cdot \; \sum_{d \in D_R} P(e \mid d_{entity}) \tag{8}$$

where $P(e \mid q_{entity})$ and $P(e \mid d_{entity})$ are the likelihoods of generating the entity from the original query and a document $d \, \in \, D_R$ respectively.

Document Model. To estimate the document models for each dimension, we follow the same methodology as for the query with an additional step of Dirichlet smoothing [28]. This has two effects: First, it prevents undefined KL-Divergence scores. Second, it achieves an IDF-like effect by smoothing the probabilities of expressions that occur frequently in the C. The generative probability of a term w from document-text model D_{text} is estimated as,

$$P(w \mid D_{text}) = \frac{\hat{P}(w \mid D_{text}) + \mu P(w \mid C_{text})}{|D_{text}| + \mu} \tag{9}$$

where $\hat{P}(w \mid D_{text})$ is computed according to Eq. 3 and μ is set as the average document length of our collection [28]. Similarly, we estimate D_{time}, D_{space}, and D_{entity} with C_{time}, C_{space}, and C_{entity} as background models to tackle the above mentioned issues. To estimate D_{time} and D_{space}, we follow methods similar to Eqs. 4 and 7. However, we do not apply the Gaussian smoothing (as described in Eq. 5) as it tends to artificially introduce temporal and spatial information into the document content.

4 Experiments

Next, we describe our experiments to study the impact of the different components of our approach. We make our experimental data publicly available[2].

[2] http://resources.mpi-inf.mpg.de/d5/linkingWiki2News/.

Document Collection. For the first set of experiments, we use The New York Times[3] Annotated Corpus (NYT) which contains about two million news articles published between 1987 and 2007. For the second set, we use the ClueWeb12-B13 (CW12) corpus[4] with 50 million web pages crawled in 2012.

Test Queries. We use the English Wikipedia dump released on February 3rd 2014 to generate two independent sets of test queries: **(1)** *NYT-Queries*, contains 150 randomly sampled Wikiexcerpts targeting documents from the NYT corpus; **(2)** *CW-Queries* contains 150 randomly sampled Wikiexcerpts targeting web pages from CW12 corpus. *NYT-Queries* have 104 queries, out of 150, that come with at least one temporal expression, geolocation, and named-entity mention. In the remaining 46 test queries, 17 do not have any temporal expressions, 28 do not have any geolocations, and 27 do not have any entity mentions. We have 4 test queries where our taggers fail to identify any additional semantics. *CW-Queries* have 119 queries, out of 150, that come with at least one temporal expression, geolocation and entity mention. 19 queries do not mention any geolocation, and 26 do not have entity mentions.

Relevance Assessments were collected using the CrowdFlower platform[5]. We pooled top-10 results for the methods under comparison, and asked assessors to judge a document as **(0)** irrelevant, **(1)** somewhat relevant, or **(2)** highly relevant to a query. Our instructions said that a document can only be considered highly relevant if its main topic was the event given as the query. Each query-document pairs was judged by three assessors. Both experiments resulted in 1778 and 1961 unique query-document pairs, respectively. We paid $0.03 per batch of five query-document pairs for a single assessor.

Effectiveness Measures. As a strict effectiveness measure, we compare our methods based on mean reciprocal rank (MRR). We also compare our methods using normalized discounted cumulative gain (NDCG) and precision (P) at cutoff levels 5 and 10. We also report the mean average precision (MAP) across all queries. 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.

Methods. We compare the following methods: **(1)** *txt* considers only the query-text model that uses the background models estimated from the current events portal and the source Wikipedia article (Eq. 2); **(2)** *txtT* uses the query-text and query-time model (Eq. 4); **(3)** *txtS* uses the query-text and query-space model (Eq. 7); **(4)** *txtE* uses the query-text and query-entity model (Eq. 8); **(5)** *txtST* uses the query-text, query-time and query-space model; **(6)** *txtSTE* uses all four query models to rank documents.

Parameters. We set the values for the different parameters in query and document models for all the methods by following [27]. For the NYT corpus, we treat

[3] http://corpus.nytimes.com.

[4] http://www.lemurproject.org/clueweb12.php/.

[5] http://www.crowdflower.com/.

top-100 documents retrieved in the first stage as pseudo-relevant. For CW12 corpus with general web pages, we set this to top-500. The larger number of top-K documents for the CW12 corpus is due to the fact that web pages come with fewer annotations than news articles. In Eq. 2 for estimating the Q_{text}, we set $\beta = 0.5$ thus giving equal weights to the background models. For the interpolation parameters, we set $\lambda = 0.85$ in Eqs. 2 and 8. For the Gaussian smoothing in Eq. 6 we set $\sigma = 1$. The smallest possible MBRs in Eq. 7 is empirically set as $resol_{lat} \times resol_{long} = 0.1 \times 0.1$.

Implementation. All methods have been implemented in java. To annotate named entities in the test queries and documents from the NYT corpus, we use the AIDA [16] system. For the CW12 corpus, we use the annotations released as Freebase Annotations of the ClueWeb Corpora[6]. To annotate geolocations in the query and NYT corpus, we use an open-source gazetteer-based tool[7] that extracts locations and maps them to GeoNames[8] knowledge base. To get geolocations for CW12 corpus we filter entities by mapping them from Freebase to GeoNames ids. Finally, we run Stanford Core NLP[9]on the test queries, NYT corpus and CW12 corpus to get the temporal annotations.

Results. Tables 2 and 3 compare the different methods on our two datasets. Both tables have two parts: **(a)** results on the entire query set; and **(b)** results on a subset of queries with at least one temporal expression, geolocation, and entity mention. To denote the significance of the observed improvements to the *txt* method, we perform one-sided paired student's T test at two alpha-levels: 0.05 (‡), and 0.10 (†), on the MAP, P@5, and P@10 scores [8]. We find that the *txtSTE* method is most effective for the linking task.

In Table 2 we report results for the NYT-Queries. We find that the *txtSTE* method that combines information in all the dimensions achieves the best result across all metrics except P@5. The *txt* method that uses only the text already gets a high MRR score. The *txtS* method that adds geolocations to text is able to add minor improvements in NDCG@10 over the *txt* method. The *txtT* method achieves a considerable improvement over *txt*. This is consistent for both NYT-Queries (a) and NYT-Queries (b). The *txtE* method that uses named-entities along with text shows significant improvement in P@5 and marginal improvements across other metrics. The *txtST* method that combines time and geolocations achieves significant improvements over *txt*. Finally, the *txtSTE* method proves to be the best and shows significant improvements over the *txt*.

In Table 3, we report results for the CW-Queries. We find that the *txtSTE* method outperforms other methods across all the metrics. Similar to previous results, we find that the *txt* method already achieves high MRR score. However, in contrast, the *txtT* approach shows improvements in terms of P@5 and NDCG@5, with a marginal drop in P@10 and MAP. The geolocations improve

[6] http://lemurproject.org/clueweb12/FACC1/.

[7] https://github.com/geoparser/geolocator.

[8] http://www.geonames.org/.

[9] http://nlp.stanford.edu/software/corenlp.shtml.

Table 2. Results for NYT-Queries

	NYT-Queries (a) - 150 queries						**NYT-Queries (b)** - 104 queries					
Measures	*txt*	*txtT*	*txtS*	*txtE*	*txtST*	*txtSTE*	*txt*	*txtT*	*txtS*	*txtE*	*txtST*	*txtSTE*
MRR	0.898	0.897	0.898	0.898	0.898	**0.902**	0.921	0.936	0.921	0.921	0.936	**0.942**
P@5	0.711	0.716	0.709	0.716 ‡	**0.719**	0.717	0.715	0.740 ‡	0.715	0.723 ‡	**0.742** ‡	0.740 ‡
P@10	0.670	0.679	0.669	0.671	0.679	**0.682** †	0.682	0.692	0.681	0.684	0.692	**0.696** †
MAP	0.687	0.700	0.687	0.688	0.701	**0.704** †	0.679	0.702 †	0.679	0.682	0.703 †	**0.708** ‡
NDCG@5	0.683	0.696	0.682	0.685	0.697	**0.697**	0.686	0.721	0.686	0.689	0.721	**0.723**
NDCG@10	0.797	0.813	0.798	0.796	0.814	**0.815**	0.794	0.823	0.795	0.795	0.823	**0.825**

Table 3. Results for CW12-Queries

	CW-Queries (a) - 150 queries						**CW-Queries (b)** - 119 queries					
Measures	*txt*	*txtT*	*txtS*	*txtE*	*txtST*	*txtSTE*	*txt*	*txtT*	*txtS*	*txtE*	*txtST*	*txtSTE*
MRR	0.824	0.834	0.831	0.827	0.833	**0.836**	0.837	0.855	0.846	0.842	0.854	**0.855**
P@5	0.448	0.460	0.451	0.456 †	0.467 ‡	**0.475** ‡	0.456	0.468	0.459 †	0.466 ‡	0.478 ‡	**0.488** ‡
P@10	0.366	0.349	0.366	0.375 ‡	0.367	**0.375** †	0.377	0.358	0.378	0.390 ‡	0.377	**0.386** ‡
MAP	0.622	0.616	0.628	0.640 ‡	0.640 †	**0.653** ‡	0.623	0.616	0.631	0.647 ‡	0.642 †	**0.661** ‡
NDCG@5	0.644	0.657	0.651	0.654	0.666	**0.673**	0.655	0.675	0.664	0.669	0.684	**0.695**
NDCG@10	0.729	0.723	0.734	0.746	0.744	**0.755**	0.736	0.736	0.744	0.759	0.755	**0.769**

the quality of the results in terms of MAP and significantly improve P@10. Though individually time and geolocations show only marginal improvements, their combination as the *txtST* method shows significant increase in MAP. We find that the *txtE* method performs better than other dimensions with a significant improvement over *txt* across all metrics. Finally, the best performing method is *txtSTE* as it shows the highest improvement in the result quality.

Discussion. As a general conclusion of our experiments we find that leveraging semantic annotations like time, geolocations, and named entities along with text improves the effectiveness of the linking task. Because all our methods that utilize semantic annotations (*txtS*, *txtT*, *txtE*, *txtST*, and *txtSTE*) perform better than the text-only (*txt*) method. However, the simple *txt* method already achieves a decent MRR score in both experiments. This highlights the effectiveness of the event-specific background model in tackling the verbosity of the Wikiexcerpts. Time becomes an important indicator to identify relevant news articles but it is not very helpful when it comes to general web pages. This is because the temporal expressions in the news articles often describe the event time period accurately thus giving a good match to the queries while this is not seen with web pages. We find that geolocations and time together can better identify relevant documents when combined with text. Named entities in the queries are not always salient to the event but may represent the context of the event. For complex queries, it is hard to distinguish salient entities which reduces the overall performance due to topical drifts on a news corpus. However, they prove to be effective to identify relevant web pages which can contain more general information thus also mentioning the contextual entities. The improvement of our method over a simple text-based method is more pronounced for the ClueWeb corpus than the news corpus because of mainly two reasons: firstly, the news corpus is too narrow with much smaller number of articles; and secondly, it is slightly easier to retrieve relatively short, focused, and high quality news articles. This is supported by the fact that all methods achieve much higher MRR scores for the NYT-Queries.

Gain/Loss Analysis. To get some insights into where the improvements for the *txtEST* method comes from, we perform a gain/loss analysis based on NDCG@5. The *txtSTE* method shows biggest gain (+0.13) in NDCG@5 for the following query in *NYT-Queries*:

West Windsor Township, New Jersey: The West Windsor post office was found to be infected with anthrax during the anthrax terrorism scare back in 2001-2002.

The single temporal expression *2001-2002* refers to a time period when there were multiple anthrax attacks in New Jersey through the postal facilities. Due to the ambiguity, the *txtT* and *txtS* methods become ineffective for this query. Their combination, however, as the *txtST* method becomes the second best method achieving NDCG@5 of 0.7227. The *txtEST* combines the entity *Anthrax* and becomes the best method by achieving NDCG@5 of 0.8539. This method suffers worst in terms of NDCG@5 (−0.464) for the following query in *CW-Queries*:

Human Rights Party Malaysia: The Human Rights Party Malaysia is a Malaysian human rights-based political party founded on 19 July 2009, led by human rights activist P. Uthayakumar.

The two entities, *Human Rights Party Malaysia* and *P. Uthayakumar* and one geolocation, *Malaysia*, do prove to be discriminative for the event. Time becomes an important indicator to identify relevant documents as *txtT* becomes most effective by achieving NDCG@5 of 0.9003. However, a combination of text, time, geolocations and named entities as leveraged by *txtEST* achieves a lower NDCG@5 of 0.4704.

Easy and Hard Query Events. Finally, we identify the easiest and the hardest query events across both our testbeds. We find that the following query, in the *CW-Queries*, gets the highest minimum P@10 across all methods:

Primal Therapy: In 1989, Arthur Janov established the Janov Primal Center in Venice (later relocated to Santa Monica) with his second wife, France.

For this query even the simple *txt* method gets a perfect P@10 score of 1.0. Terms *Janov*, *Primal*, and *Center* retrieve documents that are pages from the center's website, and are marked relevant by the assessors. Likewise, we identify the hardest query as the following one from the *NYT-Queries* set:

Police aviation in the United Kingdom: In 1921, the British airship R33 was able to help the police in traffic control around the Epsom and Ascot horse-racing events.

For this query none of the methods were able to identify any relevant documents thus all getting a P@10 score equal to 0. This is simply because this relatively old event is not covered in the NYT corpus.

5 Related Work

In this section, we put our work in context with existing prior research. We review five lines of prior research related to our work.

First, we look into efforts to link different document collections. As the earliest work, Henzinger et al. [14] automatically suggested news article links for an ongoing TV news broadcast. Later works have looked into linking related text across multiple archives to improve their exploration [6]. Linking efforts also go towards enriching social media posts by connecting them to news articles [26]. Recently, Arapakis et al. [1] propose automatic linking system between news articles describing similar events.

Next, we identify works that use time to improve document retrieval quality [23]. To leverage time, prior works have proposed methods that are motivated from cognitive psychology [21]. Time has also been considered as a feature for query profiling and classification [17]. In the realm of document retrieval, Berberich et al. [5] exploit explicit temporal expressions contained in queries to improve result quality. As some of the latest work, Peetz et al. [20] detect temporal burstiness of query terms, and Mishra et al. [19] leverage explicit temporal expressions to estimate temporal query models. Efron et al. [11] present a kernel density estimation method to temporally match relevant tweets.

There have been many prior initiatives [7,13] to investigate geographical information retrieval. The GeoCLEF search task examined geographic search in text corpus [18]. More recent initiatives like the NTCIR-GeoTime task [12] evaluated adhoc retrieval with geographic and temporal constraints.

We look into prior research works that use entities for information retrieval. Earlier initiatives like INEX entity ranking track [10] and TREC entity track [3] focus on retrieving relevant entities for a given topic. More recently, INEX Linked Data track [4] aimed at evaluating approaches that additionally use text for entity ranking. As the most recent work, Dalton et al. [9] show significant improvement for document retrieval.

Divergence-based retrieval models for text have been well-studied in the past. In their study, Zhai et al. [27,28] compare techniques of combining backgrounds models to query and documents. To further improve the query model estimation, Shen et al. [24] exploit contextual information like query history and click through history. Bai et al. [2] combine query models estimated from multiple contextual factors.

6 Conclusion

We have addressed a novel linking problem with the goal of establishing connections between excerpts from Wikipedia, coined Wikiexcerpts, and news articles. For this, we cast the linking problem into an information retrieval task and present approaches that leverage additional semantics that come with a Wikiexcerpt. Comprehensive experiments on two large datasets with independent test query sets show that our approach that leverages time, geolocations, named entities, and text is most effective for the linking problem.

References

1. Arapakis, I., et al.: Automatically embedding newsworthy links to articles: From implementation to evaluation. JASIST **65**(1), 129–145 (2014)
2. Bai, J., et al.: Using query contexts in information retrieval. In: SIGIR.(2007)
3. Balog, K., et al.: Overview of the TREC 2010 entity track. In: DTIC.(2010)
4. Bellot, P., et al.: Report on INEX 2013. ACM SIGIR Forum **47**(2), 21–32 (2013)
5. Berberich, Klaus, Bedathur, Srikanta, Alonso, Omar, Weikum, Gerhard: A Language Modeling Approach for Temporal Information Needs. In: Gurrin, Cathal, He, Yulan, Kazai, Gabriella, Kruschwitz, Udo, Little, Suzanne, Roelleke, Thomas, Rüger, Stefan, van Rijsbergen, Keith (eds.) ECIR 2010. LNCS, vol. 5993, pp. 13–25. Springer, Heidelberg (2010)
6. Bron, Marc, Huurnink, Bouke, de Rijke, Maarten: Linking Archives Using Document Enrichment and Term Selection. In: Gradmann, Stefan, Borri, Francesca, Meghini, Carlo, Schuldt, Heiko (eds.) TPDL 2011. LNCS, vol. 6966, pp. 360–371. Springer, Heidelberg (2011)
7. Cozza, Vittoria, Messina, Antonio, Montesi, Danilo, Arietta, Luca, Magnani, Matteo: Spatio-Temporal Keyword Queries in Social Networks. In: Catania, Barbara, Guerrini, Giovanna, Pokorný, Jaroslav (eds.) ADBIS 2013. LNCS, vol. 8133, pp. 70–83. Springer, Heidelberg (2013)
8. Croft, B., et al.: Search Engines: Information Retrieval in Practice. Addison-Wesley, Reading.(2010)
9. Dalton, J., et al.: Entity query feature expansion using knowledge base links. In: SIGIR.(2014)
10. Demartini, Gianluca, Iofciu, Tereza, de Vries, Arjen P.: Overview of the INEX 2009 Entity Ranking Track. In: Geva, Shlomo, Kamps, Jaap, Trotman, Andrew (eds.) INEX 2009. LNCS, vol. 6203, pp. 254–264. Springer, Heidelberg (2010)
11. Efron, M., et al.: Temporal feedback for tweet search with non-parametric density estimation. In: SIGIR.(2014)
12. Gey, F., et al.: NTCIR-GeoTime overview: Evaluating geographic and temporal search. In: NTCIR.(2010)
13. Hariharan, R., et al.: Processing spatial-keyword (SK) queries in geographic information retrieval (GIR) systems. In: SSDBM.(2007)
14. Henzinger, M.R., et al.: Query-free news search. World Wide Web **8**, 101–126 (2005)
15. Hoffart, J., et al.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. In: IJCAI.(2013)
16. Hoffart, J., et al.: Robust Disambiguation of Named Entities in Text. In: EMNLP.(2011)
17. Kulkarni, A., et al.: Understanding temporal query dynamics. In: WSDM.(2011)
18. Mandl, Thomas, Gey, Fredric C., Di Nunzio, Giorgio Maria, Ferro, Nicola, Larson, Ray R., Sanderson, Mark, Santos, Diana, Womser-Hacker, Christa, Xie, Xing: GeoCLEF 2007: The CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, Carol, Jijkoun, Valentin, Mandl, Thomas, Müller, Henning, Oard, Douglas W., Peñas, Anselmo, Petras, Vivien, Santos, Diana (eds.) CLEF 2007. LNCS, vol. 5152, pp. 745–772. Springer, Heidelberg (2008)
19. Mishra, A., et al.: Linking wikipedia events to past news. In: TAIA.(2014)
20. Peetz, M., et al.: Using temporal bursts for query modeling. Inf. retrieval **17**(1), 74–108 (2014)

21. Peetz, Maria-Hendrike, de Rijke, Maarten: Cognitive Temporal Document Priors. In: Serdyukov, Pavel, Braslavski, Pavel, Kuznetsov, Sergei O., Kamps, Jaap, Rüger, Stefan, Agichtein, Eugene, Segalovich, Ilya, Yilmaz, Emine (eds.) ECIR 2013. LNCS, vol. 7814, pp. 318–330. Springer, Heidelberg (2013)
22. Perea-Ortega, José M., Ureña-López, LAlfonso: Geographic Expansion of Queries to Improve the Geographic Information Retrieval Task. In: Bouma, Gosse, Ittoo, Ashwin, Métais, Elisabeth, Wortmann, Hans (eds.) NLDB 2012. LNCS, vol. 7337, pp. 94–103. Springer, Heidelberg (2012)
23. Ricardo, C., et al.: Survey of temporal information retrieval and related applications. ACM Comput. Surv. (CSUR) **47**(2), 1–41 (2014)
24. Shen, X., et al.: Context-sensitive information retrieval using implicit feedback. In: SIGIR.(2005)
25. Tan, B., et al.: Mining long-term search history to improve search accuracy. In: KDD.(2006)
26. Tsagkias, M., et al.: Linking online news and social media. In: WSDM.(2011)
27. Zhai, C., et al.: Model-based feedback in the language modeling approach to information retrieval. In: CIKM.(2001)
28. Zhai, C., et al.: Two-stage language models for information retrieval. In: SIGIR.(2002)