Linking Archives Using Document Enrichment and Term Selection

Marc Bron, Bouke Huurnink, and Maarten de Rijke

ISLA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam
m.m.bron@uva.nl, b.huurnink@uva.nl, derijke@uva.nl

Abstract. News, multimedia and cultural heritage archives are increasingly offering opportunities to create connections between their collections. We consider the task of linking archives: connecting an item in one archive to one or more items in other, often complementary archives. We focus on a specific instance of the task: linking items with a rich textual representation in a news archive to items with sparse annotations in a multimedia archive, where items should be linked if they describe the same or a related event. We find that the difference in textual richness of annotations presents a challenge and investigate two approaches: (i) to enrich sparsely annotated items with textually rich content; and (ii) to reduce rich news archive items using term selection. We demonstrate the positive impact of both approaches on linking to same events and linking to related events.

1 Introduction

News, multimedia, and cultural heritage archives are opening up and publishing their content online, enabling users to search for items of interest across multiple archives. With the general public gaining access to archive content, an increasing number of users can be expected to exhibit exploratory behavior [2], rather than directed search typical of professional users [10]. In order to make archives accessible to the general public, modes of access supporting exploratory behavior should be examined.

One way to enable exploration over (multiple) archives is to create links between individual items. On the web a common method to enable exploration is to create hyperlinks between documents allowing a user to wander from one document to the next and gradually explore a topic of interest. In an archival setting the creation of links has received little attention, likely due to the focus within archives on annotation and preservation of individual items, rather than on supporting browsing behavior.

We examine the linking problem in an archival setting, focusing on events. Here we aim to connect an item from one archive to items in another archive that discuss the same or related events. Links to items describing the same event allow users to access different views of the same event, while links to items describing related events allow users to explore interconnected relationships between events. Such event-based links are particularly valuable for exploring news archives. We focus on a specific instance of the task: linking items from a newspaper archive with a rich textual representation to items from a multimedia archive that tend to have sparse annotations, see Figure 1.

Our scenario is characterized by two somewhat complementary challenges. First, the targets of our linking task—archived multimedia items—are relatively sparsely annotated which leads to recall problems. Second, the source of a link is a news article in

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 360–371, 2011.

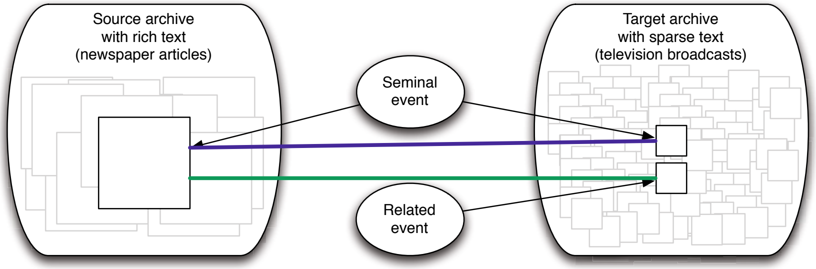


Fig. 1. Illustration of the event based linking task between archives

a news archive; such articles may be long and discuss issues that are only indirectly related to the seminal event that triggered the article, thereby potentially giving rise to a precision problem. These two problems motivate our primary research question: *when linking from an archive with rich textual content to an archive with sparsely annotated items, how can we compensate for the relative richness and sparsity of the representations to improve linking performance?* We divide this into two sub-questions:

RQ1. Does expanding sparse item representations with text from other sources improve linking performance?

RQ2. What effect does modeling reduced versions of the original richly represented source item have on linking performance?

We approach the linking task as a retrieval problem: given a source item, retrieve target items it should be linked to. To address our first research question we improve the representation of target items by enriching their sparse annotations with text from the target archive, the source archive, and Wikipedia. To address our second research question we reduce the representation of the source item by selecting a subset of terms from the source article, from manually created annotations or from automatically created ones.

The contributions of this paper are three-fold: (i) we define and motivate a new task, i.e., linking archives based on events, and identify future directions; (ii) we report on a set of experiments investigating the effect of item representation along two dimensions; and (iii) we demonstrate the effectiveness of linking based on enrichment of sparsely annotated items with content from a different archive.

We discuss related work in §2. Then in §3 we describe our item enrichment and linking approaches. In §4 we describe our data sets and experimental setup. In §5 we report our results and provide an analysis. We conclude in §6.

2 Related Work

We discuss work related to our task of event-based linking from a textually rich to a textually sparse archive. We consider links at the item level, i.e., linking from one item to another item, rather than linking phrases and words from an item representation such as is done for example in Wikipedia linking [15, 16] and hypertext linking [4].

One example of linking between items is the *alignment* task: identifying items in comparable collections that discuss the same person, entity, or concept. This task is addressed by [11], who use a retrieval engine to align entries from four encyclopedias. They find that using document structure such as the title and body of an encyclopedia entry is an important component of achieving successful alignment. Our work differs in three respects: we link from an archive with textually rich items to an archive with sparsely annotated items; we include two types of linking; and we enrich sparsely annotated items.

Another area to cover the linking task is *topic tracking*, in which items are connected when they discuss the same seminal event or related events [1]. Commonly, this is done within a collection consisting of either a single news source [7] or a collection of multiple textual news services [17, 21]. Work on topic detection and tracking includes work on detecting novelty and redundancy using language models [21] and new event detection using an adaptation of the vector space model with named entities [12]. These methods use techniques from information retrieval to find link targets, based on similarity. We will also follow a retrieval-based approach, but as we are dealing with archives with disparate quantities of text, we focus on the effect of using document expansion and term selection techniques for linking.

The news context has also seen work into linking news articles to blog post entries that discuss them. Using the structure of news articles (title, lead, body, etc.) to model a query can help in linking to correct blog posts [20]. An early paper on the topic of cross-media linking investigates generating connections between news photos, videos, and text on the basis of dates and named entities present in texts associated with the items [3]. [14] investigated cross-media news content retrieval to provide complementary news information. This was done on the basis of news articles and closed captions from news broadcasts, and focused on differences in topic structure in the captions to find complementary news articles for broadcasts. Also relevant is work on linking passages from the closed captioning of television news broadcasts to online news articles [9]. Here, the focus was on the time-based aspect of identifying articles about the news subject being discussed at any particular point in time. An interesting finding was that term selection was valuable in identifying the correct relevant articles. We also apply term selection to our source items, but link to sparse content and additionally experiment with expanding target items.

The problem of text sparsity in the linking task has seldom been addressed. However, in the field of of information retrieval, it has been found that when faced with very short documents (i.e., documents with sparse text), *document expansion* can help improve retrieval performance [19]. Document expansion refers to combining text from related documents with the text of an original document. For a more in-depth study of expansion in the context of traditional document search, see [5].

3 Approach

We formally define the variants of the linking archives tasks addressed in this paper: *same event linking* and *related event linking*. Given an event e, described by a source item s from a source archive A_s with rich text representations, create links to target items $T = \{t_1, \ldots, t_n\}$ in a target archive A_t, where the event described by each $t_i \in T$ is the *same* as e. In the second variant, each $t_i \in T$ is *related* to e. The notions of same and related event will be defined in §4 below.

An item is *textually rich* when it contains, on top of human-annotated metadata, textual content. *Sparse* representations only contain human-annotated metadata. In the specific setting in which we are working, source items are news articles, hence textually rich, and target items are sparsely represented items in a video catalog (see below).

Linking model. We model the task of linking archives by finding a ranked list of target items t whose representation is most similar to the representation of the source item s. We use the vector space model [18] as our similarity function:

$$\mathrm{sim}(s,t) = \frac{\boldsymbol{V}(s) \cdot \boldsymbol{V}(t)}{|\boldsymbol{V}(s)||\boldsymbol{V}(t)|}, \tag{1}$$

where $\boldsymbol{V}(s)$ and $\boldsymbol{V}(t)$ are vector representations of s and t, respectively, the numerator is the dot product of the vectors and $|\boldsymbol{V}|$ is the length of $\boldsymbol{V}$.

We compute the similarity between a fixed source item s and every potential target item t in the target archive and rank each t according to its similarity. The resulting ranked list is cut off at some rank n so as to yield a list of link targets.

Document expansion. To address sparseness of the representation of a target item t we use other items x for expanding the representation of t. Below we consider multiple sources A_x for the expansion items x: the source archive A_s, the target archive A_t or even an external archive A_e. To obtain expansion items x we compute the similarity between $t \in A_t$ and each item in expansion archive A_x and rank its items by similarity, as in (1). The resulting ranked list is cut off at some rank m to yield a list of expansion items; these are then concatenated to t to form an expanded representation of t.

Selecting representative terms. Recall that our source items are textually rich. To address the potential of topic drift that may result from textual richness, we investigate the effect of automatically selecting a small number of terms from the text associated with a source item (instead of using all terms) when ranking candidate target terms. To select terms, we take the top $k\%$ terms from s ranked by their TFIDF score, which is defined as:

$$\mathit{TFIDF}(a) = \frac{c(a,d)}{|d|} \cdot \log\left(\frac{|D|}{|\{d \in D : d \text{ contains } a\}|}\right),$$

where $c(a,d)$ gives the count of term a in document d, $|d|$ is the length of a document and $|D|$ is the number of documents in the collection. As is well-known, TFIDF assigns more importance to terms that have a high frequency in a few representations, rewarding terms that are discriminative for a specific representation.

Selecting representative entities. We refine the selection of representative terms by only considering named entities. Named entities are a special type of term found to be important in identifying related events [12]. To select entities we apply a named entity recognizer [6] based on conditional random fields to the content of all source archive items. We then select the top $k\%$ entities based on their TFIDF value as with the terms.

Date filter. Finally, we also examine the use of the date field present in the metadata of both source and target items. Not only is the date field one of the most consistently filled fields in archival data, but it has also been shown that dates are useful when detecting same events [13]. We use a simple date filter that only allows a link from a source item s to a target item t if t's date is within an N day window around the date of s.

4 Experimental Setup

In this section we describe our experimental setup. We start by describing the collection used for evaluating the linking rich-to-sparse archive task, and follow with a description of our experiments with document expansion and term selection.

4.1 Evaluation Collection

Our evaluation collection consists of a source archive containing textually rich newspaper articles and a target archive of textually sparse television news broadcasts to which we want to link. We single out a set of source items as our test cases for linking, and for each test source item, we have a set of relevance judgments indicating which items in the target archive refer to (i) the same seminal event, and (ii) related events.

Source archive. Our source archive consists of 346,559 newspaper articles published by a Dutch newspaper, the NRC Handelsblad,[1] from 3 Jan. 2005 to 8 Jun. 2010. Each article consists of the article text (article title and body) and a series of metadata fields created by archivists at the newspaper. These metadata fields comprise of *persons*, *locations*, *organizations*, *events* and *keywords* that are the subject of an article. Rather than exploiting the detailed specifics of NRC's archive, we combine all of the data from the metadata fields for an article together; we refer to this aggregated set of items as the *metadata* for s. We refer to the article text as the *content* of s. On average, source item content has 409 terms and metadata has 8 terms, for a total of 417 terms per item.

Target archive. Our target archive in this paper consists of 73,666 television news stories obtained from the Netherlands Institute for Sound and Vision, the Dutch national audiovisual broadcast archive.[2] We restrict the target archive to news stories broadcast during a period that encompassed the period of the source article collection, (1 Jan. 2005–20 Dec. 2010). We limited the target archive to news stories as other program categories, e.g., game shows and soap operas, are unlikely to yield suitable link targets for news articles. Each news story is manually described by professional archivists, with free-text *description* and *summary* fields and structured fields describing *persons*, *locations*, *keywords* and *other names* that are the subject of the news story. Once again, rather than considering the text of all these fields individually, we combine them to form the *metadata* for a given target item t. On average, target item metadata consists of 13 terms, illustrating the relative sparsity of text as compared to the source archive.

Events. We use the definition of event used at the Topic Detection and Tracking (TDT) campaign which makes a distinction between *seminal events*, i.e., high impact news events that generate follow-up events, and *related events* that are caused or predict the seminal event but are not seminal events by themselves.[3]

Test source items. In order to evaluate our linking approaches, we select a set of source items to use as test items to be linked. We use two requirements for our selection: the

[1] http://www.nrc.nl/

[2] http://instituut.beeldengeluid.nl/

[3] http://projects.ldc.upenn.edu/TDT5/Annotation/TDT2004V1.2.pdf

selected item should contain a clear seminal event (to facilitate judgments of system-generated links) and there should be at least one item in the target archive that covers the same or a related event. To satisfy the first requirement, we randomly select news events from Wikipedia listings of important events per month[4] and manually search the source archive to identify a newspaper item describing the event. To satisfy the second requirement, we search in the target archive to make sure that there is at least one television broadcast that describes the same or a related event. If so, the item is selected as a test source item. In total we selected 50 test source items, describing a range of events such as *16 May 2007: Nicolas Sarkozy is sworn in as the new president of France*; and *7 July 2009: A memorial service is held in the Staples Center in Los Angeles for the deceased pop icon Michael Jackson.*

Relevance judgments. We create relevance judgments using the pooling method adopted by TREC [8], the de-facto standard for creating relevance judgments for small to moderately sized test collections. We performed pooling on the basis of the sets of results produced by different linking systems. For each system and source item, the top 20 ranked documents were selected for inclusion in the pool. These results were then merged and duplicated documents were removed. The merged lists of results were then shown to human assessors, with results for each individual source item being judged by the same assessor to ensure consistency of results. The assessors were instructed to make a distinction between target items that describe the same event as the source item and targets that describe related events. The assessors' instructions were taken from the TDT assessor manual.[4] Due to the nature of related events, for each source item the average number of target items describing the same event is much lower than the number describing related events (2.4 vs. 11.8).

4.2 Experiments

Recall the two main research questions from § 1. (RQ1) Does expanding sparse item representations with text from other sources improve linking performance? (RQ2) What effect does term selection on the original textually rich source item have on linking performance? Below we list the experiments we conduct in order to answer these questions. All experiments are performed on the two tasks: *same event linking*, i.e., linking to items that describe the same event, and *related event linking*, i.e., linking to items that describe a related event.

Baseline. As our baseline we perform linking using all of the source item's content and metadata without term selection to representations of target items without expansion.

Expanding sparse text representations. In order to answer our first research question we evaluate the effect of increasing the number of documents used to expand target items on linking performance. An overview of the experiments is given in Table 1. We experiment with three sources of information for document expansion: the target archive itself, expanding target items with representations from other items in the archive; Wikipedia, the online encyclopedia; and the richly represented, news-focused items in the source archive.

[4] See e.g., `http://nl.wikipedia.org/wiki/Januari_2009`

Table 1. Description of the expansion models. In all cases the original sparse target metadata is concatenated with n expansion documents to form the expanded item representation.

Exp. model	A_x	**Description**
baseline	–	no expansion
n target docs	A_t	add $n \in \{1, \ldots, 10\}$ documents from target archive
n Wikipedia docs	A_e	add $n \in \{1, \ldots, 10\}$ documents from the Wikipedia encyclopedia
n source docs	A_s	add $n \in \{1, \ldots, 10\}$ documents from source archive

Table 2. Descriptions of the term selection models evaluated in the term selection experiments. In all experiments the number of terms in the *source* item representation (content and metadata) is reduced. Target items consist of their original metadata without expansion.

TS model	**Description**
baseline	all text associated with s, including content text and metadata text
content	s content text only
metadata	s metadata text only
title	s title
lead	first 2 sentences of content s
$x\%$ terms	select top $x\%$ terms from *content* s, using TFIDF ($x \in \{10, 20 \ldots, 100\}$)
$y\%$ ne	select top $y\%$ entities in *content* s, using TFIDF ($y \in \{10, 20 \ldots, 100\}$)
combined	combine metadata s with optimal $x\%$ *terms* and $y\%$ *ne* from content s

Term selection for rich text representations. In order to answer our second research question we evaluate the effect of reducing the amount of text in a source item on linking performance. An overview of our term selection experiments is given in Table 2. First, we experiment with using newspaper article structure to reduce the source item representation, following the framework presented in [20]. We then experiment with using only the most representative unique terms and named entities in the source item. We also investigate using only the manual annotations, i.e., metadata, of a source item. Finally, we experiment with using the optimal combination of these options.

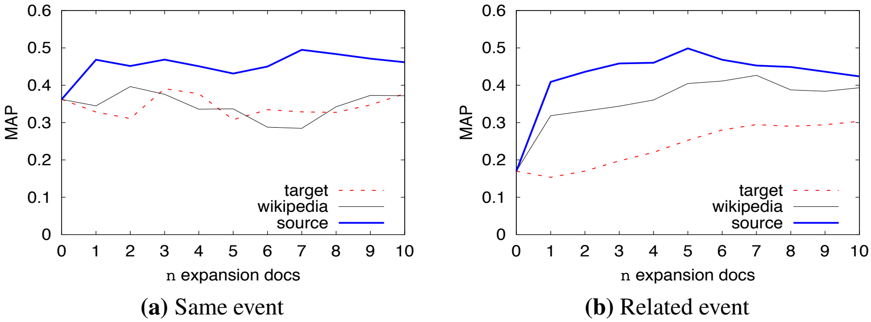
Evaluation measures and significance testing. We use three evaluation metrics for evaluating linking performance. Mean Average Precision (MAP), the average of the Average Precision (AP) scores over all test items. This metric evaluates the number of correct link targets in a list (of length 100 in our case), where correct targets higher in the list are assigned more importance. Precision at rank five (P@5) only considers link targets in the top five. A perfect score of 1.0 indicates that all five targets at the top are correct. When less correct targets exist the maximum score will be lower. Mean Reciprocal Rank (MRR) is the average of the Reciprocal Rank (RR) for each source item. The RR is the inverse of the first correct answer and indicates at which rank of the list of target items the first correct target is found. We use a standard paired t-test to determine significant differences between results. We use $^{\triangle}$ or $^{\triangledown}$ ($^{\blacktriangle}$, $^{\blacktriangledown}$) to indicate whether a score is significantly higher or lower than the baseline with a significance level of $\alpha < .05$ ($\alpha < .01$).

5 Results

5.1 Document Expansion

We first contrast different archives for expansion, i.e., the source archive, target archive, and Wikipedia. Figure 2a shows the MAP scores for *same event linking* using different expansion archives. Expanding with documents from archives other than the source archive does little to improve over the baseline even with the optimal number of expansion documents. Figure 2b shows that for *related event linking* expansion with documents from all three archives improves over the baseline. Again expanding with source archive documents achieves best performance. The performance scores for both event linking tasks, with the optimal number of expansion documents, are given in Table 3. The optimal number of documents to expand with from the source archive is seven for *same event linking* and five for *related event linking*; both yield a significant improvement over the baseline. We note that although optimized for MAP, the other early precision metrics follow the same trend in that the optimal number of documents for MAP is also the optimal number for the other metrics. The P5 scores for *same event linking* do improve (by 40.9%), but remain relatively low; this is due to the small number of relevant target items per test item (on average 2.4).

Let us examine the source item that benefits most from document expansion in the *same event linking* task. The title of the source item is "Openness expenses Dutch Royal Family." The description of the target item is: "Prime Minister Balkenende promises the House of Representatives transparency in the expenses of the Royal Family." The underlying event of the source and target item is the same, i.e., a parliamentary discussion about transparency with respect to the expenses of the Dutch royal house. However, the viewpoint of the event is described from a different angle in each item: the source item focuses on a request for more transparency from the house of representatives, while the target item focuses on the prime minister promising this transparency. Document



(a) Same event **(b)** Related event

Fig. 2. Document expansion with n documents, from the target archive, the Wikipedia encyclopedia, and the source archive. Here 0 indicates no expansion.

Table 3. Results of document expansion; significance is tested against the baseline

	Same event				Related event			
Exp. Model	*detail*	*MAP*	*P5*	*MRR*	*detail*	*MAP*	*P5*	*MRR*
baseline	–	.3623	.2000	.4819	–	.1699	.2732	.5082
n target docs	$n = 3$	.3907	.2227	.4654	$n = 10$	.3036$^{\blacktriangle}$	.3854	.5705
n wikipedia docs	$n = 2$	.3964	.2136	.4425	$n = 7$	.4266$^{\blacktriangle}$	.4537$^{\blacktriangle}$	.6988$^{\vartriangle}$
n source docs	$n = 7$	.4949$^{\vartriangle}$	.2818	.5435	$n = 5$	.4988$^{\blacktriangle}$	.4829$^{\blacktriangle}$	.6864$^{\vartriangle}$

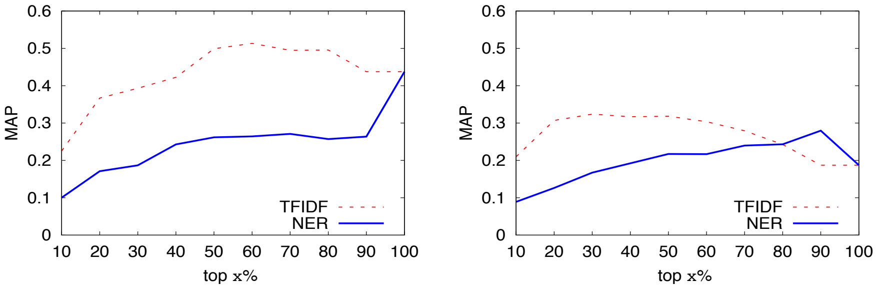
Table 4. Results of the term selection experiments; significance tested against the baseline

	Same event				Related event			
TS Model	*detail*	*MAP*	*P5*	*MRR*	*detail*	*MAP*	*P5*	*MRR*
baseline	–	.3623	.2227	.4820	–	.1699	.2732	.5083
content	–	.3582	.1955	.4800	–	.1583	.2634	.4838
metadata	–	.1636$^{\blacktriangledown}$	.0636$^{\blacktriangledown}$	.1863$^{\blacktriangledown}$	–	.1768	.2000	.2887$^{\blacktriangledown}$
title	–	.4157	.2227	.4597	–	.2264	.2829	.4300
lead	–	.4428	.2318	.5386	–	.2681	.3366	.5294
x% terms	$x = 60\%$	.5133$^{\vartriangle}$	.2682	.6390	$x = 30\%$	.3229	.3268	.4799
y% ne	$y = 100\%$	.4374	.2091	.5592	$y = 90\%$	.2796	.2829	.4724
combined	x=60%, y=100%	.4660	.2409	.5849	x=30%, y=90%	.3387	.3317	.4459

expansion works as it adds text from multiple news articles about the parliamentary discussion on transparency to the target item, compensating for different views. Similarly, for *related event linking*, document expansion increases the number of viewpoints of a seminal event covered in a source item to improve linking performance.

5.2 Term Selection

On the *source* item side we experiment with different term selection techniques. In this section, we link to the original unexpanded target items. Table 4 shows that using only terms from a specific field, e.g., lead or title, improves over using the whole document in terms of absolute scores for both *same event linking* and *related event linking*, but not significantly so. We also select terms and named entities from the content of the *source* item based on their TFIDF score. Figure 3a shows the MAP score for *same event linking* while using only the top $x\%$ of the terms (dotted line) or named entities (solid line). We observe that removing any named entities decreases performance. For selecting terms there is an optimum when only 60% of the terms (ranked by TFIDF) are selected. Table 4 shows that *same event linking* with the optimum of 60% of the terms selected from the source item, a significant improvement over the baseline is achieved. When linking to related events, selecting terms from the source item does not lead to significant improvements over the baseline; this is not surprising as *related event linking* is more recall oriented and benefits from having a source item description that covers



(a) Same event **(b)** Related event

Fig. 3. Term selection with the top $x\%$ (ranked by TFIDF) of the original terms and named entities from the original source item retained

all aspects of a seminal event. When less terms from the source item are selected, less aspects of an event are covered making linking to related events more difficult.

We take a closer look at the source item that benefits most from term selection on the *same event linking* task. The title of the source item is "Ukrainian president dissolves parliament." One of the target items is described by: "President Yushchenko of Ukraine dissolves parliament and issues new elections." The source item content consists of 342 words and mentions various aspects of the event, e.g., comments of the opposition leader and protests leading to dissolution of parliament. Each aspect potentially matches with the description of a target item. As the political situation in the Ukraine was unstable for a number of years, many target items cover aspects of this topic. By only selecting a small number of terms specific to the seminal event, term selection prevents a drift in topic towards other aspects of the source item description.

5.3 Further Improving Linking Performance

In order to see how far we can push linking performance we conduct two additional experiments. In the first we combine the best models, i.e., the best term selection is used to find targets and the target items have been expanded with the optimal number of documents. The combination achieves a MAP of .4801 on the *same event linking* task, which does not improve over using document expansion (.4949) or term selection (.5133) by itself. We find similar results for *related event linking*. We find that for items where document expansion helps, term selection has relatively poor performance, and vice versa. This fits the intuition that term selection and expansion have opposite effects: one makes an item's event description more specific, while the other broadens the description. Depending on the source item only one of the effects may be desired.

Our second experiment is with a date filter that restricts target items to a period of 14 days around the date of the source item. On the *same event linking* task this results in a baseline MAP score of .5689 and scores of .7263 and .7397 MAP for the best document expansion and term selection models, respectively. Scores for all models go up, including the baseline, but the same significant differences in performance remain between the baseline and the best models. On the *related event linking* task using a date filter decreases performance, from .1699 to .0883 MAP for the baseline and to .1487

MAP and .2077 MAP for the optimal document expansion and term selection settings, respectively. That filtering on dates improves performance on *same event linking* is unsurprising: this is a high precision-oriented task and a news broadcast and article about the same event are published around the same day. This effect, however, is specific to linking news based on same events. Related events, which may be distributed over a long period, do not demonstrate the same effect.

6 Conclusions

With archives opening up their content online and enabling interconnectivity, the challenge of linking between items in different archives arises. We consider the task of linking archives based on events. We investigate two variants of this task: *same event linking* and *related event linking*. We use a retrieval approach to link items from a news paper archive with very rich text descriptions to videos in a multimedia archive with relatively sparse annotations. This mismatch between the representations in both archives gives rise to our two research questions: (i) does expanding sparse item representations with text from other sources improve linking performance; and (ii) what effect does term selection on a textually rich source item have on linking performance?

In answer to (i), we find that expanding *target* items with documents from other sources improves performance for both *same event linking* and *related event linking*. Using expansion documents from the source archive, however, is most effective as the content has the same focus as the target archive. In answer to (ii), we find that reducing the number of terms in the *source* item representation is most effective for *same event linking*. The reduced items are more robust to topic drift and form a better match for the short event descriptions in the target archive. *Related event linking* also improves but not as much as with target item expansion. Related events benefit more from rich descriptions (as obtained through expansion) that cover all aspects of an event.

Turning to directions for future work, our combination of document expansion and term selection techniques would benefit from further investigation. We combined the best document expansion model with the best term selection settings, but this naive approach did not outperform the individual methods. Another direction is investigating other settings where rich and sparse archives need to be linked; our document expansion and term selection techniques are general enough to be applied in non-news settings, e.g., linking art encyclopedias (with rich text) to museum collections (with sparse text).

Acknowledgements. This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the PROMISE Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191, the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, and under COMMIT project Infiniti.

References

[1] Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y., et al.: Topic detection and tracking pilot study: Final report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 194–218 (1998)

[2] Bron, M., van Gorp, J., Nack, F., de Rijke, M.: Exploratory search in an audio-visual archive: Evaluating a professional search tool for non-professional users. In: EuroHCIR 2011: 1st European Workshop on Human-Computer Interaction and Information Retrieval (July 2011)

[3] Carrick, C., Watters, C.: Automatic association of news items. Information Processing & Management 33(5), 615–632 (1997)

[4] Cohn, D., Hofmann, T.: The missing link-a probabilistic model of document content and hypertext connectivity. In: NIPS 2001, pp. 430–436 (2001)

[5] Diaz, F., Metzler, D.: Improving the estimation of relevance models using large external corpora. In: SIGIF 2006, pp. 154–161. ACM, New York (2006)

[6] Finkel, J., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: ACL 2005, pp. 363–370. ACL (2005)

[7] Franz, M., Ward, T., McCarley, J., Zhu, W.: Unsupervised and supervised clustering for topic tracking. In: SIGIR 2001, pp. 310–317. ACM, New York (2001)

[8] Harman, D.K.: The TREC test collections. In: Voorhees, E.M., Harman, D.K. (eds.) TREC: Experiment and Evaluation in Information Retrieval. MIT, Cambridge (2005)

[9] Henzinger, M., Chang, B.-W., Milch, B., Brin, S.: Query-free news search. In: World Wide Web, vol. 8, pp. 101–126 (2005)

[10] Huurnink, B., Hollink, L., van den Heuvel, W., de Rijke, M.: Search behavior of media professionals at an audiovisual archive: A transaction log analysis. J. American Soc. Information Science and Technology 61(6), 1180–1197 (2010)

[11] Kern, R., Granitzer, M.: German encyclopedia alignment based on information retrieval techniques. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 315–326. Springer, Heidelberg (2010)

[12] Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: SIGIR 2004, pp. 297–304. ACM, New York (2004)

[13] Li, Z., Wang, B., Li, M., Ma, W.: A probabilistic model for retrospective news event detection. In: SIGIR 2005, pp. 106–113. ACM, New York (2005)

[14] Ma, Q., Nadamoto, A., Tanaka, K.: Complementary information retrieval for cross-media news content. Information Systems 31(7), 659–678 (2006)

[15] Meij, E., Bron, M., Hollink, L., Huurnink, B., de Rijke, M.: Learning semantic query suggestions. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 424–440. Springer, Heidelberg (2009)

[16] Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: CIKM 2007, vol. 7, pp. 233–242 (2007)

[17] Radev, D., Otterbacher, J., Winkel, A., Blair-Goldensohn, S.: NewsInEssence: summarizing online news topics. Comm. of the ACM 48(10), 95–98 (2005)

[18] Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Comm. of the ACM 18(11), 613–620 (1975)

[19] Tao, T., Wang, X., Mei, Q., Zhai, C.: Language model information retrieval with document expansion. In: HLT-NAACL 2006, pp. 407–414 (2006)

[20] Tsagkias, M., de Rijke, M., Weerkamp, W.: Linking online news and social media. In: WSDM 2011, pp. 565–574. ACM, New York (2011)

[21] Zhang, Y., Callan, J., Minka, T.: Novelty and redundancy detection in adaptive filtering. In: SIGIR 2002, pp. 81–88. ACM, New York (2002)