Noname manuscript No.
(will be inserted by the editor)

Reasoning about Fuzzy Temporal Information from the Web

Towards Retrieval of Historical Events

Steven Schockaert · Martine De Cock · Etienne E. Kerre

Received: date / Accepted: date

Abstract When searching for information about historical events, queries are naturally formulated using temporal constraints. However, the structured temporal information needed to support such constraints is usually not available to information retrieval systems. Furthermore, the temporal boundaries of most historical events are inherently ill–defined, calling for suitable extensions of classical temporal reasoning frameworks. In this paper, we propose a framework based on a fuzzification of Allen's Interval Algebra to cope with these issues. By using simple heuristic techniques to extract temporal information from web documents, we initially focus more on recall than on precision, relying on the subsequent application of a fuzzy temporal reasoner to improve the reliability of the extracted information, and to deal with conflicts that arise because of the vagueness of events. Experimental results indicate that a consistent and reliable knowledge base of fuzzy temporal relations can thus be obtained, which effectively allows us to target temporally constrained retrieval tasks.

Keywords Temporal Reasoning · Fuzzy Set Theory · Event-based Retrieval

1 Introduction

As time is paramount in our perception of the world, much of the information users are looking for is subject to temporal constraints. Users may, for instance, be interested in pictures of the New York skyline before and after September 11, 2001, in facts and figures about the 1986 FIFA World Cup, or in news stories about the first manned moon landing. Accordingly, there is a growing interest in information retrieval (IR) systems that exhibit some form of temporal awareness [1]. We will refer to such systems as event–based, or temporally aware IR systems. For example, in the question answering (QA) community, there has recently been considerable attention devoted to answering temporally restricted questions such as *How many paintings did Piet Mondriaan make during his Amsterdam years* and *In what city did the Olympic Winter Games take*

S. Schockaert, M. De Cock, E.E. Kerre
Ghent University, Department of Applied Mathematics and Computer Science, Krijgslaan 281, B-9000 Gent, Belgium
E-mail: {steven.schockaert,martine.decock,etienne.kerre}@ugent.be

place before Salt Lake City[1] [2–6]. In the context of multi–document summarization, temporal information has, among others, been employed to obtain a chronological ordering of sentences from different documents [7–9], to summarize relevant information about events from a stream of news stories [10,11], and to automatically generate overview timelines containing the most important events from a news corpus [12–14]. Finally, in the context of historical digital libraries, some efforts have been made towards temporally aware query interfaces, allowing users to find documents about certain time periods or events [15,16,3,17].

Nonetheless, the capabilities of current IR systems to handle events and temporal information are still quite limited. This is in marked contrast to geographic IR systems [18] and local search services like Google Maps[2] or Yahoo! local[3], which can rely on a vast amount of structured, geographical background knowledge, predominantly in the form of gazetteers. The key problem in transferring results from the field of geographic IR, being conceptually very similar to event–based retrieval, is the fact that no reasonably comprehensive, structured repositories of temporal information are available. An appealing strategy may be to apply information extraction techniques to acquire temporal information about events automatically from large document collections. However, existing techniques for recognizing and grounding events in documents are very much focused on news stories, relying heavily on the fact that news stories tend to have an explicit time stamp and on language characteristics of the news genre.

When moving outside the realm of news stories, explicit temporal information becomes rare. Quantitative temporal information, i.e., dates and time spans of events, can often not be found, and linguistic techniques to obtain qualitative temporal relations, e.g., based on the tense and aspect of verbs, are bound to fail more often. The solution we propose is to use, in a first step, heuristic, redundancy–based techniques that result in a considerably higher recall, at the cost of slightly reduced precision. The underlying assumption of this strategy is that the subsequent use of temporal reasoning to enforce consistency in the extracted knowledge base can detect most of the erroneous information, resulting in a sufficiently high overall precision.

Second, and perhaps most fundamental, many events and time periods, such as the Renaissance, World War II or the recent Subprime Mortgage Crisis, are characterized by gradual, ill–defined beginnings and/or endings. To deal with these problems, we propose a novel framework for compiling temporal information about events from web documents, centered around a fuzzification of Allen's Interval Algebra [19] that was developped in [20–22]. While in principle only crisp information can be extracted from web documents — people do not usually talk in terms of membership degrees — fuzzy temporal information can be obtained by aggregating partially conflicting information from different sources. For large–scale events, enough information can usually be found to construct reliable (fuzzy) time spans, capturing their (imprecise) temporal boundaries. Examples of such fuzzy time spans are shown, for instance, in Figure 3. For lesser–known events, on the other hand, only qualitative information can typically be obtained (e.g., before and during relations). When interpreted as classical temporal relations, this qualitative information will inevitably be inconsistent. For example, regarding the temporal relation between the US Housing Bubble (HB) and the Subprime Mortgage Crisis (SMC), we find among others the following statements:

[1] Questions taken from the 2006 CLEF English–Dutch question set.

[2] `http://maps.google.com`

[3] `http://local.yahoo.com`

1. ... and a monumental housing bubble burst that spawned the subprime mortgage crisis that still plagues markets.[4]
2. The [subprime mortgage] crisis began with the bursting of the US housing bubble ... [5]
3. Following the collapse of the housing bubble after the subprime mortgage crisis in the US, ... [6]

An advanced temporal information extraction module could interpret the first statement as evidence that the ending of HB is strictly before the beginning of SMC. The second statement, on the other hand, seems to imply that the ending of HB coincides with the beginning of SMC, while the third statement indicates that the ending of HB was strictly after the beginning of SMC. An important hypothesis, lying at the root of our approach, is that such conflicts provide us with useful clues about the vagueness of event boundaries. Specifically, after temporal information has been extracted from the web, we apply a fuzzy temporal reasoning algorithm to obtain a consistent knowledge base. In this way, neither of the three statements above would be completely ignored, and the knowledge base would encode that the three corresponding temporal relations all hold to some degree.

The structure of the paper is as follows. In Section 2, we present an overview of related work about the use of temporal information in IR tasks and the automated extraction of temporal information from document collections. Next, Section 3 familiarizes the reader with some necessary preliminaries about fuzzy temporal relations. In Section 4, we discuss how fuzzy temporal reasoning can be used to detect and repair inconsistencies in information extracted from the web, leading to more reliable knowledge bases. Subsequently, in Section 5, we focus on the actual extraction of temporal information from the web. Specifically, we introduce techniques to construct (fuzzy) time spans for events, as well as two redundancy–based heuristics to find instances of before and during relations. Section 6 deals with the evaluation task we consider in this paper: retrieval of historical events. In particular, we introduce a number of techniques to select, from a set of candidates, those events that satisfy a given temporal constraint. Finally, Section 7 provides an experimental validation of our approach.

2 Related Work

There is a large body of work on extracting temporal information from news stories. For example, [23] is concerned with resolving temporal expressions such as *today*, *last week*, or *in April*. Problems include the disambiguation between specific and non–specific (e.g., February is usually cold) temporal expressions, and deciding which temporal expressions should be resolved w.r.t. the document time stamp and which should be resolved w.r.t. other reference dates. In [24], an attempt is made to automatically assign time stamps (intervals or points) to every event–clause in a news story, while [25] deals with learning which temporal relations may hold between the main and subordinate clauses of a sentence, starting from sentences where a temporal marker (e.g., before, while, until, ...) makes this relation explicit.

[4] `http://www.businessweek.com/investor/content/jan2008/pi20080125_322728.htm`, accessed September 16, 2008.

[5] `http://en.wikipedia.org/wiki/Subprime_mortgage_crisis`, accessed September 16, 2008.

[6] `http://www.socialistworker.co.uk/art.php?id=15121`, accessed September 16, 2008.

To facilitate machine learning approaches to temporal information extraction, the TimeML markup language has been conceived [4], which allows to annotate events and time expressions with semantic information, as well as temporal relations between events and between events and time expressions. In [26], for instance, TimeBank, a TimeML annotated corpus, is used to train a system that recognizes events and temporal relations between them. In [3], temporal reasoning is used to support question answering, based on temporal information extracted by a classifier which was trained on the TimeBank corpus.

Another relevant line of research tries to identify phrases that describe events in collections of time–stamped documents by looking at the distribution of the time stamps of the documents in which these phrases occur. In particular, to verify that a phrase e corresponds to an event with time span T, [14] proposes to count the number of documents whose time stamp is respectively during and outside T, and for each of these two groups, the number of documents which contain e and the number of documents which do not. Based on these frequency counts, a χ^2 test is then used to test whether e occurs significantly more during T than outside T. In [12], a similar approach is adopted, although events are represented as complete sentences, rather than phrases, and the log–likelihood ratio is used, rather than χ^2. A similar solution to this problem, based on naive scan methods, is suggested in [27], where Flickr[7] tags are used rather than time–stamped documents. In [28] and [29], statistical language models are used to account for variation in term usage over time. Finally, in [17], co-occurrences of dates and place names in historical documents are used to identify significant events.

Most of the techniques described above, fail to work when other types of documents than news stories are considered. The TimeBank corpus, for instance, consists entirely of news stories. Moreover, many types of documents are not time–stamped. Historical documents, for example, often cover a large time period, making document time stamps of little value [17]. Furthermore, while statistical techniques can be used to identify time segments (e.g., days, weeks, months or years) during which an event is talked about, and thus to provide an approximate location in time, they are not suitable to identify exact temporal boundaries of events. To find exact beginning and ending dates of events, surface patterns such as *<EVENT> began on <DATE>* can be used. The use of patterns to find appropriate entities is a standard technique in QA systems [30–32].

In this paper, we propose an alternative methodology for extracting temporal information from large document collections. As important differences with state–of–the–art techniques, our method does not rely on time stamps, and, since it does not require sophisticated linguistic processing, it is less tied to one particular genre. The main novelty lies in the combination of

1. naive, simple techniques to extract an initial knowledge base (ensuring a sufficiently high recall);
2. intelligent post–processing, in the form of fuzzy temporal reasoning, to make the extracted information more reliable (ensuring a sufficiently high precision).

Finally, note that this paper is an extended and revised version of [33]. A preliminary version of some of the ideas in this paper can also be found in [34].

[7] `http://www.flickr.com`

Table 1 Definition of qualitative temporal relations between the fuzzy time intervals A and B, and their correspondence with the classical definitions when $A = [a^-, a^+]$ and $B = [b^-, b^+]$ are crisp intervals.

Notation	Crisp	Fuzzy
$bb^{\ll}(A,B)$	$a^- < b^-$	$\sup_p T_W(A(p), \inf_q I_W(B(q), L^{\ll}(p,q)))$
$bb^{\preccurlyeq}(A,B)$	$a^- \leq b^-$	$\inf_q I_W(B(q), \sup_p T_W(A(p), L^{\preccurlyeq}(p,q)))$
$ee^{\ll}(A,B)$	$a^+ < b^+$	$\sup_q T_W(B(q), \inf_p I_W(A(p), L^{\ll}(p,q)))$
$ee^{\preccurlyeq}(A,B)$	$a^+ \leq b^+$	$\inf_p I_W(A(p), \sup_q T_W(B(q), L^{\preccurlyeq}(p,q)))$
$be^{\ll}(A,B)$	$a^- < b^+$	$\sup_p T_W(A(p), \sup_q T_W(B(q), L^{\ll}(p,q)))$
$be^{\preccurlyeq}(A,B)$	$a^- \leq b^+$	$\sup_p T_W(A(p), \sup_q T_W(B(q), L^{\preccurlyeq}(p,q)))$
$eb^{\ll}(A,B)$	$a^+ < b^-$	$\inf_p I_W(A(p), \inf_q I_W(B(q), L^{\ll}(p,q)))$
$eb^{\preccurlyeq}(A,B)$	$a^+ \leq b^-$	$\inf_p I_W(A(p), \inf_q I_W(B(q), L^{\preccurlyeq}(p,q)))$

3 Preliminaries

3.1 Fuzzy Temporal Relations

Time spans of vague events are naturally represented as fuzzy sets of real numbers, as illustrated in Figure 3. To ensure that only fuzzy sets are considered that are intuitively acceptable as time spans, some additional criteria are typically imposed.

Definition 1 A fuzzy (time) interval is a normalised, convex, upper semi–continuous fuzzy set in $\mathbb{R}$ with a bounded support.

Recall that a fuzzy set A in $\mathbb{R}$ is normalised if $A(p) = 1$ for some p in $\mathbb{R}$. Furthermore, a normalised fuzzy set A in $\mathbb{R}$ with a bounded support is convex and upper semi–continuous iff all α–level sets $A_\alpha = \{p | p \in \mathbb{R} \wedge A(p) \geq \alpha\}$ are closed intervals (or singletons) for $\alpha \in]0, 1]$.

Qualitative temporal relations between crisp intervals are usually defined as constraints on their boundary points. For example, it holds that $[a^-, a^+]$ is during $[b^-, b^+]$ iff $b^- < a^-$ and $a^+ < b^+$. Because beginnings and endings of fuzzy time intervals are gradual, a different approach is required when defining fuzzy temporal relations. Our definitions are inspired by the fact that such constraints on the boundary points of crisp intervals can equivalently be expressed using a first–order formulation which does not explicitly refer to these boundary points. For example, let $A = [a^-, a^+]$ and $B = [b^-, b^+]$. It holds that

$$a^- < b^- \Leftrightarrow (\exists p)(p \in A \wedge (\forall q)(q \in B \Rightarrow p < q)) \tag{1}$$

Let T_W, I_W and S_W respectively denote the Łukasiewicz t–norm, implicator and t–conorm defined for a and b in $[0, 1]$ by $T_W(a, b) = \max(0, a+b-1)$, $I_W(a, b) = \min(1, 1-a+b)$, $S_W(a, b) = \min(1, a+b)$. The right–hand side of (1) can straightforwardly be generalized using the Łukasiewicz connectives, i.e., we define the degree $bb^{\ll}(A, B)$ to which the beginning of a fuzzy time interval A is strictly before the beginning of a fuzzy time interval B as

$$bb^{\ll}(A,B) = \sup_{p \in \mathbb{R}} T_W(A(p), \inf_{q \in \mathbb{R}} I_W(B(q), L^{\ll}(p,q)))$$

where $L^{\ll}(p, q) = 1$ if $p < q$ and $L^{\ll}(p, q) = 0$ otherwise. In the same way, we can define other types of fuzzy temporal relations. These fuzzy relations are summarized in Table 1, where $L^{\preccurlyeq}$ is defined as $L^{\preccurlyeq}(p, q) = 1 - L^{\ll}(q, p)$ for all p and q in $\mathbb{R}$. Note

that the definitions of our fuzzy temporal relations coincide with the corresponding classical definitions when A and B are crisp intervals. For a detailed motivation on why we choose these particular definitions, we refer to [20–22].

3.2 Fuzzy Temporal Reasoning

In principle, 32 values in $[0, 1]$ are needed to completely express our knowledge about the fuzzy temporal relationship between two (unknown) fuzzy time intervals A and B, i.e., an upper bound and a lower bound for the values of $bb^{\preccurlyeq}(A, B)$, $bb^{\ll}(A, B)$, $bb^{\preccurlyeq}(B, A)$, $bb^{\ll}(B, A)$, and similar for $ee^{\preccurlyeq}$, $ee^{\ll}$, $eb^{\preccurlyeq}$, $eb^{\ll}$, $be^{\preccurlyeq}$ and $be^{\ll}$. However, it can easily be shown that fuzzy temporal relations such as $bb^{\ll}$ and $bb^{\preccurlyeq}$ are dual to each other, i.e.[20]:

$$bb^{\preccurlyeq}(A, B) = 1 - bb^{\ll}(B, A) \qquad ee^{\preccurlyeq}(A, B) = 1 - ee^{\ll}(B, A) \tag{2}$$

$$be^{\preccurlyeq}(A, B) = 1 - eb^{\ll}(B, A) \qquad eb^{\preccurlyeq}(A, B) = 1 - be^{\ll}(B, A) \tag{3}$$

Thus, in practice only 16 values in $[0, 1]$ are needed. Specifically, we write

$$C(A, B) = \langle [\alpha_1, \beta_1, \gamma_1, \delta_1, \alpha'_1, \beta'_1, \gamma'_1, \delta'_1], [\alpha_2, \beta_2, \gamma_2, \delta_2, \alpha'_2, \beta'_2, \gamma'_2, \delta'_2] \rangle \tag{4}$$

to denote the following set of lower bounds

$$\begin{array}{llll}
be^{\preccurlyeq}(A, B) \geq \alpha_1 & be^{\ll}(A, B) \geq \alpha'_1 & be^{\preccurlyeq}(B, A) \geq \alpha_2 & be^{\ll}(B, A) \geq \alpha'_2 \\
bb^{\preccurlyeq}(A, B) \geq \beta_1 & bb^{\ll}(A, B) \geq \beta'_1 & bb^{\preccurlyeq}(B, A) \geq \beta_2 & bb^{\ll}(B, A) \geq \beta'_2 \\
ee^{\preccurlyeq}(A, B) \geq \gamma_1 & ee^{\ll}(A, B) \geq \gamma'_1 & ee^{\preccurlyeq}(B, A) \geq \gamma_2 & ee^{\ll}(B, A) \geq \gamma'_2 \\
eb^{\preccurlyeq}(A, B) \geq \delta_1 & eb^{\ll}(A, B) \geq \delta'_1 & eb^{\preccurlyeq}(B, A) \geq \delta_2 & eb^{\ll}(B, A) \geq \delta'_2
\end{array}$$

We will furthermore write $C_1(A, B)$ (resp. $C_2(A, B)$) to denote the subset of $C(A, B)$ containing the lower bounds for the fuzzy temporal relations applied to (A, B) (resp. (B, A)). Both $C_1(A, B)$ and $C_2(A, B)$ can be represented by a list of 8 values; for the set $C(A, B)$ defined in (4), we write

$$C_1(A, B) = [\alpha_1, \beta_1, \gamma_1, \delta_1, \alpha'_1, \beta'_1, \gamma'_1, \delta'_1] \quad C_2(A, B) = [\alpha_2, \beta_2, \gamma_2, \delta_2, \alpha'_2, \beta'_2, \gamma'_2, \delta'_2] \tag{5}$$

Note that $C_1(A, B) = C_2(B, A)$ and $C_2(A, B) = C_1(B, A)$.

A knowledge base of fuzzy temporal relations then corresponds to a set Θ of constraints of the form (4), where A and B are treated as variables (unknown fuzzy time intervals). The most important reasoning task in this context is deciding whether such a Θ is satisfiable (or consistent), i.e., whether there exist fuzzy time intervals for each of the variables such that all constraints in Θ are satisfied. In general this problem is NP–complete [21], hence complete reasoners are not likely to be sufficiently scalable to cope with large sets of events. To cope with this, in [22] we introduced an approximate algorithm that runs in polynomial time. This algorithm, which is similar in spirit to the path–consistency based algorithms that are traditionally employed for temporal reasoning, is presented in Procedure `Closure`, where it is assumed that Θ contains information about the events x_1, x_2, ..., x_n. Note that we can assume, without loss of generality, that Θ contains a temporal relation $C(x_i, x_j)$ for each $x_i \neq x_j$. If nothing is known about the temporal relationship between x_i and x_j, this

relation is given by $\langle[0,0,0,0,0,0,0,0],[0,0,0,0,0,0,0,0]\rangle$. Procedure `Closure` makes calls to the functions `Normalise`, `Consistent` and `Compose`, which implement the behaviour of fuzzy temporal relations. Specifically, `Normalise` tries to strengthen the available bounds, by looking at elementary properties. For instance, it can be shown that $eb^{\preccurlyeq}(A,B) \leq bb^{\preccurlyeq}(A,B) \leq be^{\preccurlyeq}(A,B)$. Hence, in (4), if $\alpha_1 < \max(\beta_1,\delta_1)$, we can strengthen our bound on $be^{\preccurlyeq}(A,B)$ from α_1 to $\max(\beta_1,\delta_1)$. Next, `Consistent` is used to decide whether a given combination of values for $\alpha_1, \ldots, \delta_2'$ in (4) is consistent. For instance, if $\beta_1 + \beta_2' > 1$ this function will return false, as we know from (2) that $bb^{\preccurlyeq}(A,B) + bb^{\ll}(B,A) = 1$ for all fuzzy time spans A and B. Finally, `Compose` is used to derive information about the fuzzy temporal relationship between fuzzy time intervals A and C, given information about the fuzzy temporal relationship between A and a fuzzy time interval B, and between B and C. For the technical details regarding `Normalise`, `Consistent` and `Compose`, we refer to [22]. The operators $\cup$ and $\subset$ on lines 9 and 10 should be understood as set operations on the corresponding sets of lower bounds. Specifically, let $C_1(A,B)$ be given by (5), and let S be an arbitrary set of lower bounds, given by $S = [\alpha,\beta,\gamma,\delta,\alpha',\beta',\gamma',\delta']$, then we define

$$\begin{aligned} C_1(A,B) \cup S = [&\max(\alpha,\alpha_1), \max(\beta,\beta_1), \max(\gamma,\gamma_1), \max(\delta,\delta_1),\\ &\max(\alpha',\alpha_1'), \max(\beta',\beta_1'), \max(\gamma',\gamma_1'), \max(\delta',\delta_1')] \end{aligned}$$

$$C_1(A,B) \subset S \Leftrightarrow C_1(A,B) \cup S = S \wedge C_1(A,B) \neq S$$

Procedure `Closure`

```
1  for i ← 1 to n do
2      for j ← i + 1 to n do
3          Normalise(C(x_i, x_j))
4          if ¬Consistent(C(x_i, x_j)) then
5              return inconsistency found
6  todo ← {(i, j, k) | 1 ≤ i, j, k ≤ n ∧ i ≠ j ≠ k}
7  while todo ≠ ∅ do
8      Select and remove a triplet (i_0, j_0, k_0) from todo
9      S ← C_1(x_{i_0}, x_{k_0}) ∪ Compose(C_1(x_{i_0}, x_{j_0}), C_1(x_{j_0}, x_{k_0}))
10     if C_1(x_{i_0}, x_{k_0}) ⊂ S then
11         C_1(x_{i_0}, x_{k_0}) ← S
12         Normalise(C(x_{i_0}, x_{k_0}))
13         if Consistent(S) then
14             todo ← todo
15                     ∪ {(i_0, k_0, l) | 1 ≤ l ≤ n ∧ l ≠ i_0 ≠ k_0}
16                     ∪ {(l, i_0, k_0) | 1 ≤ l ≤ n ∧ l ≠ i_0 ≠ k_0}
17         else
18             return inconsistency found
```

4 Reasoning about Temporal Information from the Web

Throughout this section, we assume that we have a technique at our disposal to extract temporal information from the web. We assume that for some events, we know the corresponding fuzzy time span; such events are called grounded. Second, we assume

that some instances of before and during relations have been extracted, i.e., for some pairs of events (e_1, e_2), we know that e_1 happened before e_2, written $before(e_1, e_2)$, or that e_1 happened during e_2, written $during(e_1, e_2)$. For each of these relations, we assume that a confidence score in $[0, 1]$ is available; high confidence scores indicate that strong evidence for the relation was found on the web. In Section 5, we will discuss a number of techniques to actually extract such temporal information from web documents. The reasoning problem discussed here, however, is largely independent of the specific information extraction technique used.

Based on the kind of information we encounter in web documents, we cannot straightforwardly extract membership degrees for temporal relations; e.g., either there is reason to believe that e_1 is before e_2, or there is not, but web documents are not likely to contain evidence that e_1 is before e_2 to some degree. Therefore, we assume that the qualitative information we initially have is crisp. As illustrated in the introduction, however, this will easily lead to conflicts when some of the events involved are vague. Therefore, in a second step, we use a fuzzy temporal reasoner to weaken these initial interpretations, indicating that some of the before and during relations are only satisfied to a particular degree. Specifically, the algorithm proceeds by repeatedly detecting and repairing inconsistencies, until a consistent and more reliable knowledge base (KB) is obtained. An additional effect of applying a fuzzy temporal reasoner is that new information is inferred, based on transitivity properties of fuzzy temporal relations.

Specifically, we propose a variant of Procedure `Closure`, which is called Procedure `Closure-rev`. A first deviation from `Closure` is that the closure process in `Closure-rev` is not halted the moment an inconsistency is detected. Instead, all consequences which do not rely on inconsistent premises are derived. The second difference is that inconsistencies can now occur between temporal relations and groundings (i.e., fuzzy time intervals), in addition to inconsistencies amongst different temporal relations. To cope with this, reference to a function `Grounding-consistent` has been added which returns *true* iff the corresponding temporal relation is compatible with the available groundings. In particular, this function always returns *true* when either the first or the second argument refers to an ungrounded event. When both x_i and x_j correspond to grounded events, the exact temporal relationship between these events can easily be calculated. `Grounding-consistent`(x_i, x_j) then returns *true* if the derived temporal relation between x_i and x_j is compatible with this exact temporal relationship.

Procedure `Closure-rev`

1 $todo \leftarrow \{(i,j,k) | 1 \leq i,j,k \leq n \wedge i \neq j \neq k\}$
2 **while** $todo \neq \emptyset$ **do**
3 Select and remove a triplet (i_0, j_0, k_0) from $todo$
4 **if** `Consistent`$(C_1(x_{i_0}, x_{j_0}))$ *and* `Consistent`$(C_1(x_{j_0}, x_{k_0}))$ *and* `Grounding-consistent`(x_{i_0}, x_{j_0}) *and* `Grounding-consistent`(x_{j_0}, x_{k_0}) **then**
5 $S \leftarrow C_1(x_{i_0}, x_{k_0}) \cup$ `Compose`$(C_1(x_{i_0}, x_{j_0}), C_1(x_{j_0}, x_{k_0}))$
6 **if** $C_1(x_{i_0}, x_{k_0}) \subset S$ **then**
7 $C_1(x_{i_0}, x_{k_0}) \leftarrow S$
8 `Normalise`$(C(x_{i_0}, x_{k_0}))$
9 $todo \leftarrow todo \cup \{(i_0, k_0, l) | 1 \leq l \leq n \wedge l \neq i_0 \neq k_0\}$
10 $\cup \{(l, i_0, k_0) | 1 \leq l \leq n \wedge l \neq i_0 \neq k_0\}$

After `Closure-rev` has finished, an attempt is made to repair the detected inconsistencies. An inconsistency can be repaired by weakening one or more of the premises that have been used to obtain it. Initially, when $before(e_1, e_2)$ is added to the knowledge base, this is represented as the temporal relation $\langle[1,1,1,1,1,1,1,1],[0,0,0,0,0,0,0,0]\rangle$. In other words, it is imposed that the fuzzy time intervals E_1 and E_2 of e_1 and e_2 (which may or may not be known) should satisfy $eb^{\ll}(E_1, E_2) \geq 1$. This is a rather strict interpretation of $before(e_1, e_2)$ which can be weakened in various ways. In particular, for a fixed $\Delta = \frac{1}{\rho}$ (for some ρ in $\mathbb{N} \setminus \{0\}$), we consider the following chain of representations (in decreasing order of strength):

$$\begin{array}{c}
\langle[1,1,1,1,1,1,1,1],[0,0,0,0,0,0,0,0]\rangle \\
\langle[1,1,1,1-\Delta,1-\Delta,1-\Delta,1-\Delta,1-\Delta],[0,0,0,0,0,0,0,0]\rangle \\
\cdots \\
\langle[1,1,1,\Delta,\Delta,\Delta,\Delta,\Delta],[0,0,0,0,0,0,0,0]\rangle \\
\langle[1,1,1,0,0,0,0,0],[0,0,0,0,0,0,0,0]\rangle \\
\langle[1-\Delta,1-\Delta,1-\Delta,0,0,0,0,0],[0,0,0,0,0,0,0,0]\rangle \\
\cdots \\
\langle[\Delta,\Delta,\Delta,0,0,0,0,0],[0,0,0,0,0,0,0,0]\rangle \\
\langle[0,0,0,0,0,0,0,0],[0,0,0,0,0,0,0,0]\rangle
\end{array} \tag{6}$$

Similarly, $during(e_1, e_2)$ is initially represented as a constraint $bb^{\preccurlyeq}(E_2, E_1) \geq 1 \wedge ee^{\preccurlyeq}(E_1, E_2) \geq 1$ on the (possibly unknown) fuzzy time intervals of e_1 and e_2. Again, this representation can be gradually weakened:

$$\begin{array}{c}
\langle[1,0,1,0,0,0,0,0],[1,1,0,0,0,0,0,0]\rangle \\
\langle[1-\Delta,0,1-\Delta,0,0,0,0,0],[1-\Delta,1-\Delta,0,0,0,0,0,0]\rangle \\
\cdots \\
\langle[\Delta,0,\Delta,0,0,0,0,0],[\Delta,\Delta,0,0,0,0,0,0]\rangle \\
\langle[0,0,0,0,0,0,0,0],[0,0,0,0,0,0,0,0]\rangle
\end{array} \tag{7}$$

In our experiments, we use $\Delta = 0.25$, which balances expressivity and efficiency: smaller values of Δ lead to increased flexibility, but require more computation time. In principle, an inconsistency detected by Function `Grounding-consistent` can be repaired in two ways: by discarding at least one grounding or by weakening the representation of one or more temporal relations. In practice, however, we only apply the latter technique, i.e., inconsistencies are always repaired by weakening the representation of temporal relations. The main motivation is that the fuzzy time intervals tend to be much more reliable than the temporal relations, the latter typically being the result of inherently fallible techniques. To avoid over–sensitivity to small variations in the membership functions of fuzzy time intervals, inconsistencies with groundings are only repaired if the amount by which the inconsistent lower bound is too high, is at least $\frac{\Delta}{2}$. In other words, the actual definition of Function `Grounding-consistent` is given by

$$\begin{aligned}
\texttt{Grounding-consistent}(x_i, x_j) \equiv\ & be^{\preccurlyeq}(X_i, X_j) \geq \alpha - \frac{\Delta}{2} \wedge bb^{\preccurlyeq}(X_i, X_j) \geq \beta - \frac{\Delta}{2} \\
& \wedge ee^{\preccurlyeq}(X_i, X_j) \geq \gamma - \frac{\Delta}{2} \wedge eb^{\preccurlyeq}(X_i, X_j) \geq \delta - \frac{\Delta}{2} \\
& \wedge be^{\ll}(X_i, X_j) \geq \alpha' - \frac{\Delta}{2} \wedge bb^{\ll}(X_i, X_j) \geq \beta' - \frac{\Delta}{2} \\
& \wedge ee^{\ll}(X_i, X_j) \geq \gamma' - \frac{\Delta}{2} \wedge eb^{\ll}(X_i, X_j) \geq \delta' - \frac{\Delta}{2}
\end{aligned}$$

where X_i and X_j are the fuzzy time intervals of events x_i and x_j.

Initially, before Procedure `Closure-rev` is applied, every temporal relation in the knowledge base corresponds to the representation of an assertion of the form *before*(e_1, e_2) or *during*(e_1, e_2). We will refer to these temporal relations as the initial relations. After applying `Closure-rev`, a number of inconsistent temporal relations may have been derived. Each of these inconsistencies, however, can be traced back to its premises, i.e., a particular set of initial relations. By sufficiently weakening one or more of these premises, the cause of each inconsistency can be eliminated. To this end, also all previous updates to the knowledge base that were based on one of the weakened initial relations, have to be made undone. Finally, Procedure `Closure-rev` is applied a second time. If inconsistencies still occur, some initial relations are further weakened, and the whole process is repeated until no inconsistencies can be discovered anymore.

Thus, the process of inconsistency repairing is reduced to choosing which premises to weaken. To make this choice, our confidence in each of the individual temporal relations plays a central role. The lower the confidence score of a relation, the higher the chance that it is either incorrect, or that disagreement about its correctness exists due to vagueness. In addition to confidence scores, we can base our decision on the number of inconsistencies a certain premise participates in. If a given initial relation r is (partially) incorrect, it is likely that more than one inconsistency will be derived from it. In other words, the number of times w^- that a relation r occurs as the premise of an inconsistent relation provides useful information about the likelihood of its correctness. There also is a second reason why a high value of w^- serves as an indication that r should be weakened. In general, we are interested in finding a consistent knowledge base containing as much information as possible. A high value of w^- suggests that a lot of conflicts will be solved by only weakening r. If we decide not to weaken r, several other relations may have to be weakened to obtain the same effect, resulting in a less informative knowledge base.

Whereas inconsistent relations can provide evidence against the correctness of a particular initial relation, we can sometimes also establish evidence in favor. In particular, if a consistent relation q is derived between two grounded events e_i and e_j, we can be certain that it is correct (assuming the groundings are always correct). Hence, the number of times w^+ a relation r occurs as the premise of such a correct relation provides information about the likelihood of its correctness as well. In particular, an initial relation of the form *before*(e_1, e_2) is given a score $s^{bef}(e_1, e_2)$ defined by

$$s^{bef}(e_1, e_2) = \frac{1 + w^+}{1 + w^+ + w^-} c^{bef}(e_1, e_2)$$

where $c^{bef}(e_1, e_2)$ is our confidence in *before*(e_1, e_2). In the same way, an initial relation of the form *during*(e_1, e_2) is given a score $s^{dur}(e_1, e_2)$ defined by

$$s^{dur}(e_1, e_2) = \frac{1 + w^+}{1 + w^+ + w^-} c^{dur}(e_1, e_2)$$

where $c^{dur}(e_1, e_2)$ is our confidence in *during*(e_1, e_2). Among all the premises of an inconsistent relation q, the relation with the lowest score s_{min} is weakened. Furthermore, to increase the robustness of the approach, all premises of q whose score is close to s_{min} are weakened as well. Specifically, we weaken all premises whose score is less than $s_{min} + \lambda$. In our experiments, we used $\lambda = 0.1$; using a higher value results in increased robustness (incorrect relations are removed/weakened with a higher

Fig. 1 Fuzzy time span of the Great Depression

probability), while using a lower value results in a more informative knowledge base (correct relations are removed/weakened with a lower probability). When a relation is weakened, its representation is changed to the next representation in the chain (6) or (7).

5 Collecting Temporal Information

5.1 Fuzzy Time Spans

The beginning and ending dates of well–known events and time periods can usually be extracted from web documents relatively easily. When there is a high number of documents that contain information about an event, it is likely that at least some of these documents explicitly mention its temporal boundaries. For example, if we want to know when the Great Depression took place, we can submit queries such as "the Great Depression began on", "the Great Depression took place from" or "the Great Depression ended on" to a search engine. From the search results, we can subsequently extract the corresponding beginning and ending dates using patterns such as "$\langle EVENT \rangle$ took place from $\langle DATE \rangle$ until $\langle DATE \rangle$". For most events, however, a number of different possible beginning and ending dates are thus found. This can be because some documents contain incorrect information, or because the use of patterns leads to misinterpretation of some sentences. Most frequently, however, different dates are found because the exact beginning and ending dates of historical events are affected by vagueness. Therefore, we aggregate the most significant beginning and ending dates that are found for such vague events to a fuzzy time interval. We refer to [35] for a detailed discussion on the construction of such fuzzy time spans, where also underspecified and vague beginning/ending dates are considered (e.g., World War II began in early September 1939). Figure 1 illustrates the fuzzy time interval that was thus obtained for the Great Depression. Large increases (resp. decreases) in the membership degrees correspond to beginning (resp. ending) dates that are mentioned often in web documents, while small increases (resp. decreases) correspond to dates that are mentioned only a few times.

As a second example, Figure 2 illustrates the result for the Battle of Britain. While it may appear at first glance that military conflicts such as battles have well–defined beginnings and endings, more often than not, the opposite turns out to be true. When does, in general, a battle exactly start and end, for instance. From the moment that

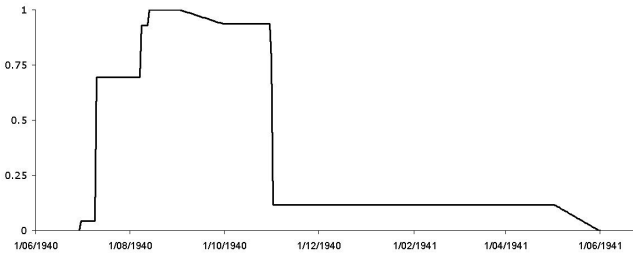


Fig. 2 Fuzzy time span of the Battle of Britain

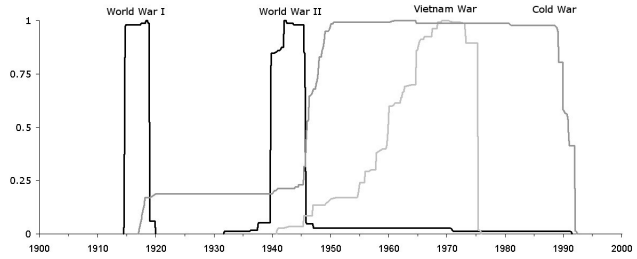


Fig. 3 Fuzzy time spans of World War I, World War II, the Vietnam War and the Cold War.

troops are moving in position? From the moment the first shot is fired, or the first bomb is dropped? Usually, the official time span of a battle reflects the period during which fighting is most intense, but this again is ill–defined, and to a large extent arbitrary. As a consequence, historians tend to disagree about the most appropriate time span of events such as battles, e.g.[8]:

> British historians date the battle from 10 July to 31 October 1940, which represented the most intense period of daylight bombing. German historians usually place the beginning of the battle in mid-August 1940 and end it in May 1941, ...

Note that the fuzzy time interval of the Battle of Britain in Figure 2 clearly reflects the two most commonly used beginning dates (mid-August and July 10, 1940). A similar observation can be made w.r.t. the ending, although October 31, 1940 is mentioned much more often than May 1941, and thereby has a much greater influence on the membership degrees. Finally, Figure 3 depicts the fuzzy time spans we found for World War I, World War II, the Vietnam War and the Cold War. To each fuzzy time span, we attach a confidence score, based on how many times a beginning and ending date have been found in web documents, and how much agreement there exists between these dates. The less agreement, the more supporting documents need to be found to obtain a high confidence score. We refer to [36] for further details.

[8] `http://en.wikipedia.org/wiki/Battle_of_Britain`, accessed October 24, 2007.

Table 2 Dates found in web documents within 200 characters from the Battle of the Somme

Date	Freq.	Date	Freq.
01/07/1916	21	28/11/1916	5
12/08/1916	3	11/08/2001	2
13/11/1916	20		

5.2 Qualitative Relations

For many lesser–known events, it is likely that no web document explicitly mentions a beginning or ending date, causing the approach outlined in Section 5.1 to fail. For example, explicit mentions of ending dates for battles are particularly rare. Moreover, beginning and ending dates are often presented in (textual or non–textual) forms which are very hard to recognize by automated methods. However, the actual time spans are usually not required in an IR setting: all we need to establish is whether or not an event satisfies a given temporal constraint. For example, to assess whether information about the Battle of the Somme is relevant to a query asking for information about "battles during World War I", we need to find out whether a during relation holds between the Battle of the Somme and World War I. One way to accomplish this is by comparing the (fuzzy) time spans of both events, but we can also try to find evidence for temporal relations directly, without the need for time spans. Unfortunately, explicit mentions of temporal relations between historical events appear to be very rare in web documents, making linguistic and pattern–based approaches often of limited value. In this section, we will therefore focus on two heuristic techniques, which are complementary to existing, more linguistically oriented approaches, and offer a much higher recall at the cost of slightly reduced precision. Moreover, as will become clear below, in all but a few cases, errors introduced by our heuristic techniques can be detected and eliminated afterwards by the fuzzy temporal reasoner.

5.2.1 Co-occurring dates

A first heuristic technique is inspired by the observation that dates which often occur near an event name are usually related to it, typically corresponding to beginning or ending dates, or dates of important sub–events. Therefore, for each event of interest, we retrieve the first 50 documents returned by Yahoo!, using its name as query. Subsequently, we extract from these documents all dates which occur within 200 characters from the event name. Note, however, that although a high number of dates may thus be found for each event, it is usually not possible to construct a reliable (fuzzy) time span from these dates. For example, for the Battle of the Somme, we have found the dates presented in Table 2. Although most of the dates that were found are during the Battle of the Somme, the most commonly used ending date (November 18, 1916) is actually missing. However, using such co–occurring dates, we can derive useful information about the likelihood that some temporal relation holds between two given events.

First, consider the temporal relation *before* between two events a and b. Let the dates that were found for event a be given by $D^a = \{d_1^a, d_2^a, \ldots, d_n^a\}$, and let f_i^a be the number of times date d_i^a was found. Similarly, let $D^b = \{d_1^b, d_2^b, \ldots, d_m^b\}$ be the dates that were found for event b, and let f_i^b be the corresponding frequency. Every pair of dates (d_i^a, d_j^b) such that $d_i^a < d_j^b$ (i.e., date d_i^a comes strictly before date d_j^b in

time) serves as evidence for $before(a,b)$, whereas every pair of dates (d_i^a, d_j^b) such that $d_i^a \geq d_j^b$ serves as evidence against $before(a,b)$:

$$pos^{bef}(a,b) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ d_i^a < d_j^b}}^{m} f_i^a f_j^b \qquad neg^{bef}(a,b) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ d_i^a \geq d_j^b}}^{m} f_i^a f_j^b$$

As soon as $pos^{bef}(a,b) > neg^{bef}(a,b)$, or equivalently $\frac{pos^{bef}(a,b)}{pos^{bef}(a,b)+neg^{bef}(a,b)} > 0.5$, there is reason to believe that $before(a,b)$ holds. This leads to the following confidence score

$$c_1^{bef}(a,b) = 2 \cdot \max(0, \frac{pos^{bef}(a,b)}{pos^{bef}(a,b) + neg^{bef}(a,b)} - 0.5) \tag{8}$$

provided that $pos^{bef}(a,b) + neg^{bef}(a,b) > 0$; otherwise, we define $c_1^{bef}(a,b) = 0$. Note that a factor 2 is introduced to obtain a confidence score in $[0,1]$. It is easy to see that $c_1^{bef}(a,b) = 1$ iff all dates in D^a are strictly before all dates in D^b. Note that (8) does not take into account how many dates were found for a and b. Because the confidence score $c_1^{bef}(a,b)$ becomes more reliable as the sizes of D^a and D^b increase, we sometimes require that $n \geq 5$ and $m \geq 5$.

Note that if a fuzzy time interval is available for both event a and event b, instead of using (8), the measures from Table 1 could be used to assess to what degree a is before b. Next, if a fuzzy time interval A is known for event a, but no fuzzy time interval is known for b, we can count positive and negative evidence by looking at how many dates were found for b that are after A. Note that a crisp date d is a special case of a fuzzy time interval, hence we can use the measure $eb^{\ll}$ to define $pos^{bef}(a,b)$ and $neg^{bef}(a,b)$. We obtain

$$pos^{bef}(a,b) = \sum_{j=1}^{m} eb^{\ll}(A, d_j^b) \cdot f_j^b$$

$$neg^{bef}(a,b) = \sum_{j=1}^{m} (1 - eb^{\ll}(A, d_j^b)) \cdot f_j^b = \sum_{j=1}^{m} be^{\preccurlyeq}(d_j^b, A) \cdot f_j^b$$

and $c_1^{bef}(a,b)$ is again given by (8). Finally, if a fuzzy time interval is available for event b, but not for event a, we can proceed in an entirely similar way.

To check whether event a happened during event b, written $during(a,b)$, we can proceed in a similar way, defining the confidence score $c_1^{dur}(a,b)$ based on a comparison of the dates in D^a and D^b. In the experiments, below, however, $during(a,b)$ typically needs to be verified in the case that b is a large–scale event, for which we can expect a reliable (fuzzy) time span B to be available. In that case, a more reliable confidence score can be defined as follows.

$$pos^{dur}(a,b) = \sum_{i=1}^{n} B(d_i^a) f_i^a \qquad neg^{dur}(a,b) = \sum_{i=1}^{n} (1 - B(d_i^a)) f_i^a$$

and $c_1^{dur}(a,b)$ is defined in terms of $pos^{dur}(a,b)$ and $neg^{dur}(a,b)$ like $c_1^{bef}(a,b)$ in (8). Note that $pos^{dur}(a,b)$ and $neg^{dur}(a,b)$ essentially correspond to the number of dates co–occurring with event a that are during b and not during b respectively. As B is a fuzzy time span, however, a date d can be during B to some degree in $[0,1]$.

5.2.2 Document structure

Our second heuristic is based on the structure of event occurrences in web documents. Specifically, let n_1 be the number of times we find (the first occurrence of) a before (the first occurrence of) b in sections of web documents, lists on web pages, and in titles of sections within the same level; let n_2 be the number of times we find b before a. Furthermore, let m_1 be the number of times event a occurs in the body of a section whose title refers to event b and let m_2 be the number of times event b occurs in the body of a section whose title refers to event a. As will be discussed in detail below, the values of n_1, n_2, m_1 and m_2 can be used to check whether $\mathit{before}(a, b)$ and $\mathit{during}(a, b)$ are likely to hold. To obtain these values, we retrieve relevant documents using the Yahoo! search engine. However, if we use a query such as "Battle of the Somme", all top ranked documents will be specifically about this battle, which heavily biases the resulting values. Therefore, we exclude all documents whose title explicitly refers to the Battle of the Somme.

There are many reasons why the order of occurrence of events in a narrative may be different from their chronological ordering. News stories, for instance, tend to start with the most recent events, after which they might go into detail about relevant background information from the past. Nonetheless, linguistic analyses have demonstrated that the event order in news stories is — albeit not completely — to a large extent chronological (e.g., [37]). Similarly, although historical documents have a tendency to digress, thereby linking events from the main linear narrative to earlier or later events [17], we can still expect the order of occurrence to be chronological more often than not. Hence, n_1 being significantly higher than n_2 is a strong indication for $\mathit{before}(a, b)$. To test whether the difference between n_1 and n_2 is greater than could be expected by chance, we employ a binomial test:

$$p_2^{bef}(a, b) = \sum_{k=n_1}^{n_1+n_2} \binom{n_1 + n_2}{k} 0.5^k (1 - 0.5)^{n_1+n_2-k}$$

If $p_2^{bef}(a, b)$ is sufficiently small (e.g., $p_2^{bef}(a, b) < 0.05$), $\mathit{before}(a, b)$ is assumed.

Instances of during relations can be found in a similar way, by looking at section titles containing the name of an event. For instance, if the title of a section refers to World War I and its body contains a reference to the Battle of the Somme, there is some reason to believe that the Battle of the Somme happened during World War I. Note, however, that also the opposite might occur: a section about the Battle of the Somme referring to World War I in its body. In other words, if m_1 is sufficiently high, it is very likely that either $\mathit{during}(a, b)$ or $\mathit{during}(b, a)$ holds. In many cases, $\mathit{during}(b, a)$ can be excluded a priori using background information. For example, knowing that battles can be part of a war but not vice versa, we can exclude the case that World War I is a part of the Battle of the Somme. Our confidence in $\mathit{during}(a, b)$ can then be expressed by any increasing function of $m_1 + m_2$ in $[0, 1]$, e.g.:

$$c_2^{dur}(a, b) = \frac{m_1 + m_2}{c + m_1 + m_2}$$

for some constant $c > 0$. If neither $\mathit{during}(a, b)$ nor $\mathit{during}(b, a)$ can be excluded a priori, a high value of $m_1 + m_2$ is not sufficient to conclude $\mathit{during}(a, b)$. To decide whether $\mathit{during}(a, b)$ holds, in this case, we compare the values of m_1 and m_2. In particular, when $\mathit{during}(a, b)$ holds, it is likely that m_1 is significantly higher than m_2.

For example, we can expect to find more section titles referring to World War I than section titles referring to the Battle of the Somme. Again, we can use a binomial test to determine the significance.

6 Event retrieval

To perform event–based IR, we typically need to find those objects (e.g., documents, people, events) that satisfy a given temporal constraint. This temporal constraint may contain explicit time references. A user may, for instance, be interested in documents from a historical digital library about painters from the 18th century, while question answering systems need to deal with questions such as *Who was prime minister of Belgium in the 1950s.* Another possibility, however, is that the temporal constraint itself already refers to an event: images of Italian paintings from the Renaissance period, blog entries about the Subprime Mortgage Crisis, etc. It is this latter case in which we are primarily interested. By far the most frequently occurring temporal relation in such constraints is the during relation. Therefore, we will focus the discussion below on during relations, although other types of temporal relations can be treated entirely analogously (e.g., before and between relations).

6.1 Compiling a Knowledge Base

Our focus is on the automatic acquisition of (fuzzy) temporal information from the web, given a collection of events of interest. A related, but largely orthogonal problem is finding occurrences of (significant) events in texts and recognizing which occurrences refer to the same event. Especially for events that are not named, this problem is highly non–trivial, often requiring deep linguistic processing (e.g., [2]). To avoid such problems in the present analysis, we focus on named events which are easy to recognize in texts: military conflicts such as the Battle of the Bulge or the Vietnam War. Specifically, we have focused in our evaluation on the wars from Table 3.

Rather than constructing one large KB, containing information about all of the 25 wars, a separate KB was constructed for each war to ensure scalability. To acquire a knowledge base for a war, we first identify a set of possibly related events by looking for phrases that frequently co-occur with the war. In the specific case of World War II, for example, the following five queries are sent to Google:

1. `allintitle:World War II`
2. `"World War II"`
3. `"World War II" events`
4. `"World War II" battle`
5. `"World War II" timeline`

The first query asks for documents which have the terms “World” “War” and “II” in their title, while the second asks for documents containing the exact phrase “World War II”. The last three queries ask for documents that additionally contain the terms “events”, “battle” or “timeline”, which tends to increase the likelihood of finding relevant event names in the returned web documents. Next, these five queries are also sent to Yahoo!, replacing the first query by `intitle:World War II` to conform to its syntax. For each query, at most 1000 documents were retrieved, which leads to a maximum of

Table 3 For every war, a separate knowledge base was constructed by retrieving a number of web documents (2nd column) from which events of interest were extracted. Next, a number of before and during relations was identified (3rd and 4th column), as well as a number of (fuzzy) time spans (5th column). Evaluation is based on a number of battles which are known to have taken place during the respective wars (6th column).

Name	# Documents	# Before	# During	# Time Spans	# Battles
American Civil War (1861–1865)	3904	3938	420	34	338
American Revolutionary War (1775–1783)	3359	7329	762	44	132
Chinese Civil War (1927–1950)	1247	7179	264	41	51
Continuation War (1941–1944)	3545	6233	833	39	8
Falklands War (1982)	1262	6314	372	42	29
Finnish War (1808–1809)	1281	6252	566	42	10
First Boer War (1880–1881)	1087	5826	520	53	4
First Chechen War (1994–1996)	1260	2653	89	25	6
Gulf War (1990–1991)	4489	7694	189	42	7
Korean War (1950–1953)	4357	8841	150	49	26
Napoleonic Wars (1803–1815)	1326	4881	380	39	191
Philippine–American War (1899–1902)	1209	5551	261	44	16
Polish September Campaign (1939)	974	3393	442	25	28
Polish–Soviet War (1919–1921)	1209	3848	395	31	15
Russo–Japanese War (1904–1905)	1281	4772	410	40	17
Second Boer War (1899–1902)	1246	3827	225	41	15
Second Chechen War (1999–2007)	1261	2612	88	26	15
Second Sino–Japanese War (1937–1945)	1072	5113	514	29	64
Spanish Civil War (1936–1939)	4260	8592	461	57	24
Spanish–American War (1898)	1224	2516	194	32	21
Vietnam War (1959–1975)	4513	9185	341	47	115
War of the Pacific (1879–1884)	1305	4040	254	34	10
World War I (1914–1919)	3723	8652	517	56	198
World War II (1939–1945)	4073	8876	673	47	329
Yom Kippur War (1973)	3730	7243	350	42	4

10000 documents. In practice, however, there often is overlap between the result lists of the different queries. The actual number of documents retrieved for each war is shown in the second column of Table 3.

In a second step, a part–of–speech (POS) tagger is used to extract noun phrases (NPs) occurring in these web documents[9]. From these NPs, we subsequently selected those that likely refer to a military conflict using a number of simple heuristic rules. A simple NP, which does not contain any prepositions, is selected if it satisfies the following requirements:

1. it contains a capitalized word different from "The";
2. it contains a reference to some kind of military conflict, i.e., a word such as battle, siege, attack, offensive, war, operation, campaign, ...;
3. it does not start with a number of selected words, including a, an, his, her, this, most, some, every, any, ...

Examples of noun phrases satisfying these requirements are "World War II", "the Pearl Harbor attack" or "Operation Desert Storm"; examples of noun phrases which violate at least one requirement are "operation desert storm", "D-day" and "most World War II battles". In addition to simple NPs, also noun phrases of the form "$\langle NP_1 \rangle$ IN $\langle NP_2 \rangle$" are allowed, where $\langle NP_1 \rangle$ and $\langle NP_2 \rangle$ are simple noun phrases and IN denotes an arbitrary preposition, provided the following requirements are satisfied:

1. $\langle NP_2 \rangle$ contains a capitalized word different from "The";
2. $\langle NP_1 \rangle$ contains a reference to some kind of military conflict;
3. $\langle NP_2 \rangle$ does not contain a reference to some kind of military conflict;
4. $\langle NP_1 \rangle$ does not start with a number of selected words.

Examples of noun phrases satisfying these requirements are "the Battle of the Bulge", "the Attack on Pearl Harbor" and "the Battle for Leyte Gulf". Next, the number of occurrences of each event are counted, ignoring case, as well as a possible starting "the", e.g. "The Battle of the Bulge" and "battle of the Bulge" would be treated as the same event name. As an example, Table 4 displays the most frequently occurring event names in the set of documents that was retrieved for World War II. This table contains many famous World War II battles and operations, although many other military conflicts are found as well (e.g., the Vietnam War, World War I, ...). Note that not all names actually refer to events: national world war ii memorial, defense dept., war information, ... However, it is unlikely that temporal relations will be found involving these names. Therefore, we can expect that most of these non–events will be excluded from the final knowledge base.

Our aim is to obtain a reliable knowledge base, containing temporal information about the most important World War II events (and similar for the other 24 knowledge bases). In contrast to our strategy in extracting the temporal relations from the web, the focus in the construction process of the KB is more on accuracy (precision) than on completeness (recall). To ensure that sufficient information about each event in the knowledge base can be found, the construction of the knowledge base is restricted to the 250 most frequently occurring event names. Note that this has an additional advantage of efficiency. For each of these events, we try to construct a (fuzzy) time interval from the web using the technique from Section 5.1; if its confidence score is sufficiently high, it is added to the knowledge base. Furthermore, for each of the 250×249 event–pairs,

[9] We used the POS–tagger from the Stanford NLP Group, available from `http://nlp.stanford.edu/`.

Table 4 The most frequently occurring event names in documents about World War II.

Name	Freq.	Name	Freq.	Name	Freq.
world war ii	12032	national world war ii memorial	99	revolutionary war	54
second world war	1300	operation market garden	97	german attack	54
world war	1175	world war 2 t-shirts and gifts	94	invasion of normandy	53
cold war	682	operation barbarossa	92	world war ii timeline	53
world war i	530	operation overlord	87	american battle monuments commission	53
world war 2	501	battle of the coral sea	87	russian revolution	51
battle of the bulge	482	war department	86	operation sealion	51
battle of britain	430	battle of normandy	84	japanese attack	50
world war two	429	pearl harbor attack	80	soviet offensive	50
civil war	417	operation torch	79	battle of france	49
war on germany	349	winter war	78	world war ii casualties	47
battle of midway	336	north african campaign	77	second battle of el alamein	46
vietnam war	305	battle of kursk	75	world war ii era	46
war ii	285	battle of leyte gulf	74	battle of guadalcanal	46
pacific war	277	world war ii history	73	world war iii	45
korean war	276	world war ii world war ii	73	invasion of france	43
first world war	251	phony war	70	world war ii combat	41
attack on pearl harbor	223	german invasion of poland	67	french revolution	41
war on japan	217	italian campaign	67	doolittle raid	41
great war	185	american civil war	65	japanese war crimes	41
battle of the atlantic	167	european war	64	persian gulf war	40
spanish civil war	146	bombing of pearl harbor	64	german war machine	39
world war i.	137	d-day invasion	63	german invasion of the soviet union	39
total war	133	great patriotic war	63	invasion of russia	38
world war ii memorial	128	spanish-american war	61	world war iv	38
normandy invasion	127	siege of leningrad	61	world war one	37
iraq war	122	german offensive	61	war bonds	37
war on the united states	120	world war ii online	60	battle of berlin	36
japanese attack on pearl harbor	118	air war	59	battle of okinawa	36
battle of stalingrad	115	battle of iwo jima	57	first battle of el alamein	35
american revolution	106	war information	56	operation bagration	35
defense dept.	105	invasion of sicily	55	invasion of italy	35
invasion of poland	102	world war ii posters	55	battle of the river plate	34
gulf war	102	german invasion	54	war front	33

we check whether a before or during relation is likely to hold, using the two heuristic techniques from Section 5.2. In particular, a before or during relation is added to the knowledge base if the evidence found by at least one of both techniques is deemed significant and the corresponding confidence score is at least 0.8. In that case, a new confidence score is assigned to the relation which is a weighted sum of the confidence scores assigned by both techniques. Note that in this way, a higher confidence is given to relations that are found by both techniques. Table 3 summarizes the number of before and during relations which are thus added to each of the knowledge bases, as well as the number of fuzzy time intervals. Note that the number of grounded events is typically between 25 and 50, i.e., between 10% and 20%. Furthermore, note that the number of before relations that is found is much greater than the number of during relations.

Next, we apply the fuzzy temporal reasoning from Section 4 to repair inconsistencies. Most of the inconsistencies in the World War II knowledge base are due to events that are erroneously assumed to be during World War II, i.e., the errors from Table 4. After the fuzzy temporal reasoning, only the events from Table 5 are still considered to be during World War II, to some extent. Comparing this table to Table 4, it is clear that almost all non–event names have been removed from the knowledge base. The only exceptions are "world war ii commemorative series" and "world war ii letters", which do not refer to events at all, and "world war ii world war ii" and "war ii", which are the result of incorrect HTML parsing or POS tagging. Furthermore, all real errors — events that did not happen during World War II — have been completely removed, e.g., world war i, cold war, vietnam war, korean war, ... Also note that the degree to which each of the events is assumed to be during World War II (also shown in Table 5) provides useful information. In particular, low membership degrees (0.25) often occur with vague and ambiguous events such as "german offensive", "war on finland", "war on bulgaria". Finally, note that the knowledge base still contains the most significant World War II events, such as "battle of midway", "battle of britain", "battle of france", "battle of normandy", "attack on pearl harbor", ...

6.2 Retrieving Events

The result of the fuzzy temporal reasoning phase is a highly reliable, consistent KB of fuzzy temporal relations and (fuzzy) time spans. While this KB is likely to contain the most important events of interest, many others will be missing. Looking at the events from Table 5, for instance, we see many large–scale events (e.g., phony war, italian campaign, operation barbarossa) and some famous battles (e.g., battle of midway, battle of iwo jima, battle of the bulge), but most of the hundreds of World War II battles are missing. However, as explained below, even if an event e is missing, the KB can play a key role in deciding whether or not e happened during an event from the KB. For example, considering the Battle of Crete (BC), the following scores can be used to decide whether $during(BC, WW2)$ holds:

$$c_1^{dur}(BC, WW2) \qquad\qquad c_2^{dur}(BC, WW2) \tag{9}$$

If both scores are 0, it may be that $during(BC, WW2)$ is false, but it is also possible that nothing can be established about the temporal relation between the Battle of Crete and World War II using our heuristic techniques. In this latter case, we can often solve the dilemma using the KB. For example, knowing from the KB that the

Table 5 Events during World War II after repairing all inconsistencies in the corresponding knowledge base.

Name	Deg.	Name	Deg.	Name	Deg.
north african campaign	1.0	japanese invasion	0.75	war on britain and france	0.75
operation market-garden	1.0	phoney war	0.75	normandy invasion	1.0
operation barbarossa	0.75	dieppe raid	0.75	d-day invasion	1.0
battle of the atlantic	0.75	attack on pearl harbor	0.75	battle of el alamein	1.0
battle of the philippine sea	1.0	allied invasion of normandy	1.0	invasion of france	0.75
german invasion of the soviet union	0.25	allied strategic bombing	1.0	invasion of sicily	1.0
doolittle raid	1.0	operation torch	1.0	battle of moscow	0.75
italian campaign	1.0	ardennes offensive	0.75	battle of monte cassino	0.75
winter war	0.75	invasion of italy	1.0	war on germany and italy	1.0
world war ii	1.0	allied invasion	1.0	invasion of britain	0.75
coral sea battle	0.25	battle of berlin	1.0	battle of coral sea	0.75
battle of france	0.25	world war ii world war ii	0.5	battle of stalingrad	1.0
great patriotic war	0.75	battle of normandy	1.0	european war	0.75
world war 2	0.75	battle of the coral sea	1.0	operation dragoon	1.0
second battle of kharkov	1.0	invasion of the soviet union	0.75	battle for okinawa	0.75
battle of midway	1.0	guadalcanal campaign	1.0	second battle of el alamein	1.0
german offensive	0.25	hurtgen forest battle	1.0	invasion of poland	0.75
german war effort	1.0	allied war effort	1.0	continuation war	0.75
war on finland	0.25	war on japan	0.75	battle of kursk	1.0
climactic battle	1.0	war on the soviet union	0.75	attack on poland	0.75
operation dynamo	0.75	pearl harbor attack	0.75	world war two	0.75
battle of iwo jima	1.0	battle of the bulge	1.0	world war ii letters	1.0
first battle of el alamein	1.0	invasion of normandy	1.0	operation overlord	1.0
second world war	0.75	siege of leningrad	1.0	battle of okinawa	0.75
pacific war	0.75	battle for stalingrad	1.0	operation compass	0.75
japanese attack on pearl harbor	1.0	1945 world war ii	0.75	japanese attack	0.75
war on bulgaria	0.25	naval battle of guadalcanal	1.0	war ii	0.75
world war ii commemorative series	0.75	bombing of pearl harbor	0.75	battle of guadalcanal	1.0
invasion of midway island	0.75	battle of britain	0.75		

Battle of Britain (BB) and the Normandy Invasion (NI) are during World War II, we can derive that $during(BC, WW2)$ holds if we can establish that both $before(BB, BC)$ and $before(BC, NI)$ are the case. To verify the latter relations, the following scores can be used

$$c_1^{bef}(BB, BC) \cdot c_1^{bef}(BC, NI)$$
$$(1 - p_2^{bef}(BB, BC)) \cdot (1 - p_2^{bef}(BC, BI))$$

Similarly, knowing from the KB that Operation Barbarossa is during World War II, it is sufficient to derive that the Battle of Kiev (BK) happened during Operation Barbarossa to conclude $during(BK, WW2)$. In general, to check whether $during(e_1, e_2)$ holds, we can

1. try to establish directly that $during(e_1, e_2)$ holds using the heuristic techniques from Section 5.2;
2. try to establish that e_1 took place during an event e, which is contained in the knowledge base and is known to be during e_2 to a large degree;
3. try to establish that e_1 took place between the events e and e', both being contained in the knowledge base and known to be during e_2 to a large degree.

The latter two strategies can be implemented using the following scores:

$$c_3^{dur}(e_1, e_2; \lambda) = \max\{c_1^{dur}(e_1, e) | \Theta \models during(e, e_2) \geq \lambda\}$$
$$c_4^{dur}(e_1, e_2; \lambda) = \max\{c_2^{dur}(e_1, e) | \Theta \models during(e, e_2) \geq \lambda\}$$
$$c_5^{dur}(e_1, e_2; \lambda) = \max\{c_1^{bef}(e, e_1) \cdot c_1^{bef}(e_1, e') | \Theta \models \{during(e, e_2) \geq \lambda, during(e', e_2) \geq \lambda\}\}$$
$$c_6^{dur}(e_1, e_2; \lambda) = \max\{(1 - p_2^{bef}(e, e_1)) \cdot (1 - p_2^{bef}(e_1, e')) | \Theta \models \{during(e, e_2) \geq \lambda, during(e', e_2) \geq \lambda\}\}$$

where Θ is the knowledge base corresponding to event e_2 and $\lambda \in [0, 1]$. Note that although we only need to establish during relations, before relations from the KB are still useful, e.g., to implement $c_5^{dur}(e_1, e_2; \lambda)$ and $c_6^{dur}(e_1, e_2; \lambda)$. Using these additional scores when both $c_1^{dur}(e_1, e_2) = 0$ and $c_2^{dur}(e_1, e_2) = 0$ helps to disambiguate between situations where $during(e_1, e_2)$ is false and situations in which $during(e_1, e_2)$ could not be established due to a lack of information. Another way of tackling this problem is to check if either $before(e_1, e_2)$ or $before(e_2, e_1)$ can be derived, in which case we can conclude that $during(e_1, e_2)$ is false. The corresponding scores are defined analogously, and are denoted by $c_1^{ndur}(e_1, e_2)$, $c_2^{ndur}(e_1, e_2)$, $c_3^{ndur}(e_1, e_2; \lambda)$, $c_4^{ndur}(e_1, e_2; \lambda)$, $c_5^{ndur}(e_1, e_2; \lambda)$ and $c_6^{ndur}(e_1, e_2; \lambda)$.

In this way, to find events that are during a given event e_2, a large number of scores are at hand, which need to be combined to produce a meaningful ranking of events. Ideally, the events about which we are confident they are during e_2 are ranked first, followed by the events about which nothing could be derived, and finally, the events about which we are confident they are not during e_2. First note that in the scenario we are envisioning, scores $c_3^{dur}(e_1, e_2; \lambda)$, $c_5^{dur}(e_1, e_2; \lambda)$, $c_3^{ndur}(e_1, e_2; \lambda)$ and $c_5^{ndur}(e_1, e_2; \lambda)$ are of no use. The reason is that these scores are based on available dates for event e_1. If enough dates are available for e_1, however, the relationship between e_1 and e_2 could also be identified directly, using c_1^{dur} and c_1^{ndur}. This holds because in

this case, event e_2 is a large–scale event, for which we typically have a fuzzy time span at our disposal.

To combine the remaining scores, a statistical classifier could be trained which decides if an event e should be ranked higher or lower than event e', given the scores for both events. This, however, requires that a sufficient amount of training and test data is available. Other approaches, such as most voting mechanisms, rely on weights that are manually assigned to each scoring function. After initial experimentation with such techniques, we found that the performance of the overall system heavily depended on these weights, where different weights led to optimal performance for different events. As the robustness of the resulting systems is therefore questionable, we will rely on a simpler strategy, focusing on the principle, rather than trying to find an optimal way of combining the different scores. In particular, for each scoring function c, we define a classifier C for events e and e' as

$$C(e,e') = \begin{cases} 1 & \text{if } c(e,e_2) > c(e',e_2) \\ -1 & \text{if } c(e,e_2) < c(e',e_2) \\ 0 & \text{otherwise} \end{cases}$$

assuming that we are interested in events during e_2. Next, these classifiers are ranked according to their reliability. For example, assume that the classifiers C_1, C_2 and C_3 are used, and that C_1 is deemed more reliable than C_2, which is in turn deemed more reliable than C_3. In this case, event e is ranked before event e' if

$$C_1(e,e') > 0$$

or, if

$$C_1(e,e') = 0 \text{ and } C_2(e,e') > 0$$

or, if

$$C_1(e,e') = 0 \text{ and } C_2(e,e') = 0 \text{ and } C_3(e,e') > 0$$

If also $C_3(e,e') = 0$, the relative ranking of e and e' is arbitrary. We will denote this system by $[c_1, c_2, c_3]$, where c_i is the scoring function corresponding to classifier C_i. Note that this approach only relies on a meaningful ranking of classifiers according to their reliability, and no parameter tuning is required.

7 Experimental Results

An additional advantage of using military conflicts is that it facilitates the experimental set–up. In general, generating a ground truth for an event–based retrieval task is hard, because the time spans of events are usually not available in a structured form, and often not even well–defined. In the case of military conflicts, however, Wikipedia can be used to this end. Specifically, we extracted lists of military conflicts, mostly battles, that are considered to be during various wars according to Wikipedia[10]. For the 25 wars from Table 6, this led to a total number of 1674 events. In our evaluation, we look at how well different systems succeed in deciding which of these events were during

[10] `http://en.wikipedia.org/wiki/Category:Battles_by_war`, accessed October 29, 2007.

World War II, which were during World War I, etc. In particular, we have compared the performance of five different systems. The first system, $B1$ (Baseline 1), only uses (fuzzy) time spans and qualitative relations that have been obtained by comparing dates, i.e.

$$B1 = [{c_1}^{dur}]$$

Similarly, $B2$ (Baseline 2) only uses qualitative relations that have been obtained by looking at document structure:

$$B2 = [{c_2}^{dur}]$$

Next, $B3$ (Baseline 3) combines both strategies as follows:

$$B3 = [{c_2^*}^{dur}, {c_1}^{dur}, {c_2}^{dur}]$$

where

$${c_2^*}^{dur} = \begin{cases} {c_2}^{dur} & \text{if } m_1 + m_2 \geq k \\ 0 & \text{otherwise} \end{cases}$$

Note that m_1 and m_2 have been defined in Section 5.2.2. An optimal performance was found for $k = 2$. This means that when $m_1 + m_2 \geq 2$, Baseline 2 is more reliable than Baseline1, whereas Baseline 1 is more reliable when $m_1 + m_2 = 1$. The system $F1$ (Fuzzy Reasoning 1) uses the knowledge base to obtain a conclusion when the two heuristic techniques fail:

$$\begin{aligned} F1 = [&{c_2^*}^{dur}, {c_1}^{dur}, {c_6}^{dur}(.,.;1), {c_4}^{dur}(.,.;1), {c_2}^{dur}, \\ &{c_6}^{dur}(.,.;0.75), {c_4}^{dur}(.,.;0.75), {c_6}^{dur}(.,.;0.5), \\ &{c_4}^{dur}(.,.;0.5), {c_6}^{dur}(.,.;0.25), {c_4}^{dur}(.,.;0.25)] \end{aligned}$$

where in particular ${C_2^*}^{dur}$, ${C_1}^{dur}$, ${C_6}^{dur}(.,.;1)$ and ${C_4}^{dur}(.,.;1)$ are considered to be the most reliable classifiers. Finally, $F2$ (Fuzzy Reasoning 2) additionally considers negative information:

$$\begin{aligned} F2 = [&{c_2^*}^{dur}, {c_1}^{dur}, {c_6}^{dur}(.,.;1), {c_4}^{dur}(.,.;1), {c_2}^{dur}, {c_6}^{ndur}(.,.;1), \\ &{c_2}^{ndur}, {c_1}^{ndur}, {c_4}^{ndur}(.,.;1), {c_6}^{ndur}(.,.;0.5), {c_4}^{ndur}(.,.;0.5), \\ &{c_6}^{dur}(.,.;0.75), {c_4}^{dur}(.,.;0.75), {c_6}^{dur}(.,.;0.5), \\ &{c_4}^{dur}(.,.;0.5), {c_6}^{dur}(.,.;0.25), {c_4}^{dur}(.,.;0.25)] \end{aligned}$$

For each of the 25 considered wars W, the 5 systems were used to produce a ranking of the military conflicts from Wikipedia. Ideally, all conflicts that took place during W are found at the top of this ranking, followed by the other events. We evaluated the performance of each system in terms of precision and recall. The average precision of the rankings for all 25 wars is shown in Table 6. Note that both $B1$ and $B2$ achieve a decent performance. Especially the performance of $B2$ is somewhat surprising: while $B1$ is based on the fuzzy time spans of all 25 wars in addition to co–occurring dates for all events, in $B2$ only document structure is taken into account. Note, however, that the actual performance of the systems $B1$ and $B2$ is not our central concern: a variety of related heuristics may be used, which may result in (slightly) higher or lower MAP

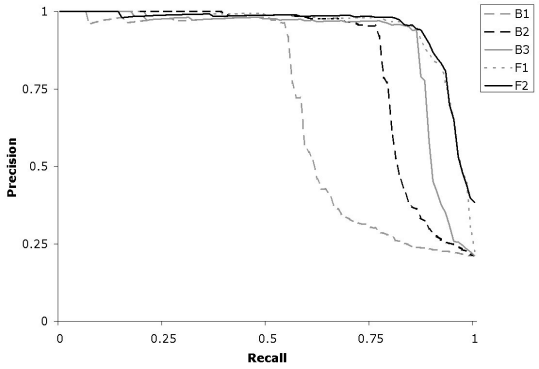
Table 6 Comparison of the different systems in terms of average precision.

Name	$B1$	$B2$	$B3$	$F1$	$F2$
Am. Civil War	0.865	0.285	0.872	0.895	0.919
Am. Revol. War	0.851	0.078	0.819	0.841	0.849
Chinese Civil War	0.551	0.623	0.837	0.918	0.963
Continuation War	0.420	0.131	0.451	0.452	0.476
Falklands War	0.431	0.917	0.994	1	1
Finnish War	0.013	0.013	0.013	0.012	0.020
First Boer War	1	0.002	1	1	1
First Chechen War	0.503	0.183	0.838	0.834	0.848
Gulf War	0.470	0.016	0.461	0.453	0.460
Korean War	0.413	0.871	0.932	0.934	0.936
Napoleonic Wars	0.068	0.125	0.068	0.068	0.065
Philippine–Am. War	0.763	0.754	0.816	0.913	0.920
Polish Sept. Camp.	0.277	0.307	0.505	0.738	0.775
Polish–Soviet War	0.410	0.787	0.853	0.915	0.934
Russo–Japanese War	0.658	0.770	0.943	0.943	0.944
Sec. Boer War	0.737	0.534	0.779	0.941	0.933
Sec. Chechen War	0.191	0.541	0.663	0.701	0.748
Sec. Sino–Jap. War	0.395	0.610	0.794	0.889	0.894
Spanish Civil War	0.676	0.595	0.877	1	1
Spanish–Am. War	0.582	0.148	0.514	0.481	0.512
Vietnam War	0.796	0.849	0.967	0.980	0.980
War of the Pacific	0.305	0.007	0.305	0.488	0.585
World War I	0.801	0.739	0.919	0.937	0.939
World War II	0.690	0.796	0.909	0.945	0.948
Yom Kippur War	0.510	1	1	1	1
MAP	0.535	0.467	0.725	0.771	0.786

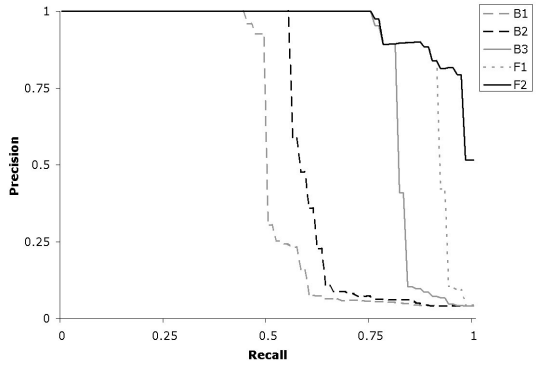
scores. Below, we therefore treat $B1$ and $B2$ as our baseline systems, mainly looking at the extent to which the performance of these systems can be improved.

A particularly interesting observation is that the performance of $B1$ is largely complementary to the performance of $B2$. For example, while $B1$ performs significantly better than $B2$ for the American Revolutionary War or the First Boer War, the opposite is true for the Falklands War or the Yom Kippur War. This is further illustrated by the results for $B3$, which improve greatly on the results of both $B1$ and $B2$. These results suggest that combining techniques based on quantitative information (e.g., date occurrences) with purely qualitative techniques (e.g., based on document structure, co–occurrence of event names, etc.) is paramount. Next, as the results for $F1$ reveal, applying fuzzy temporal reasoning has a clearly positive impact, which is substantial in several cases (e.g., Polish September Campaign, Second Boer War, War of the Pacific). Finally, the results of $F2$ show that introducing negative information (not during) consistently leads to (slightly) better performance. Note that for the best performing system, $F2$, an average precision of over 90% is achieved for 14 out of the 25 wars.

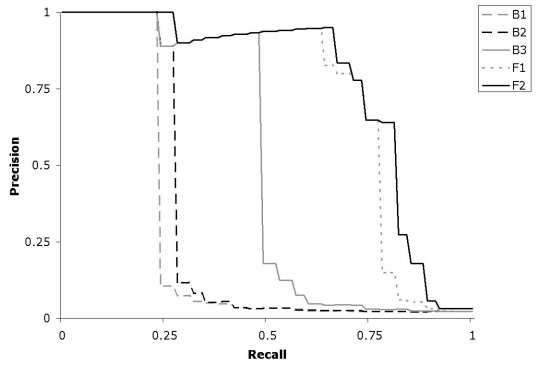
To gain a better understanding of why B_3, F_1 and F_2 yield increasingly better results, Figure 4 depicts a number of Precision–Recall graphs. Looking at Figure 4(a–c), we can see that $B1$ and $B2$ display an almost perfect behaviour at small recall levels, but precision very quickly drops to almost 0 from a particular point. This means that these systems are very strong in terms of precision: if evidence is found that e is during W, this is a reliable indication of $during(e, W)$. Their drawback, however, is a limited strength in terms of recall: for a large number of relevant events, no evidence can be found. By adding more sophisticated techniques, evidence for $during(e, W)$ can be



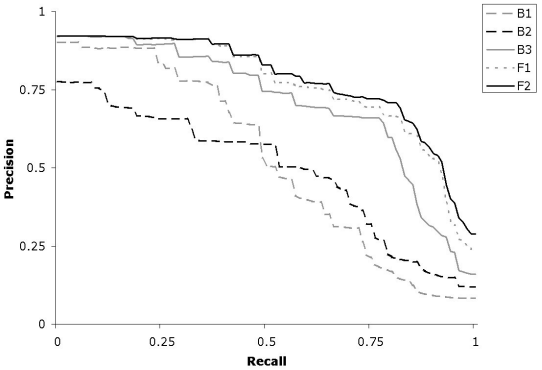
(a) World War II



(b) Chinese Civil War



(c) Polish September Campaign



(d) Average

Fig. 4 Comparison of the different systems by Precision–Recall graphs, obtained using 101 recall levels. The composite Precision–Recall graph in Figure 4(d) has been obtained by averaging the precision values at all 101 recall levels.

found for a larger group of events e. This observation essentially explains why the simple technique of combining different scoring functions by cascading classifiers works surprisingly well. First, we try to rank events according to classifiers with high precision and low recall; if this fails, increasingly less reliable classifiers are tried, characterized by an increasingly lower precision and higher recall. Figure 4(d) depicts the result of averaging the Precision–Recall graphs over all 25 wars. This again shows that $B3$ is consistently better than both $B1$ and $B2$, that $F1$ is consistently better than $B3$ and that $F2$ is consistently better than $F1$. However, neither of $B1$ and $B2$ is better than the other: $B1$ displays the best performance for recall levels up to 0.5 (on average), while $B2$ displays the best performance for higher recall levels.

8 Concluding remarks

In this paper, we have discussed a number of techniques to collect temporal information about events from the web. While for well–known events, (fuzzy) time spans can easily be extracted, explicit temporal information about lesser–known events can often not be found. To cope with this, we have introduced two heuristic techniques to acquire qualitative temporal relations as a surrogate for missing time spans. Furthermore, a fuzzy temporal reasoning algorithm is used to (partially) eliminate (partially) incorrect information from the extracted temporal relations. This leads to a highly reliable knowledge base, containing temporal information about a relatively small number of significant events, all related to a given event E. Using this knowledge base, we can identify events e that are during E (or before E, or after E) more easily, as e does not need to be linked to E directly: it suffices to link e to one or more events from the knowledge base.

Experimental results demonstrate that by mining qualitative temporal relations from the web, in addition to (fuzzy) time intervals, accurate results can be obtained. We stressed how the performance of both heuristic techniques is to a large extent complementary, which explains why surprisingly good results are obtained by combining both techniques. Next, using fuzzy temporal reasoning, the performance is substantially improved in a large number of cases. Additionally considering negative information (i.e., not during) leads to a further (slight) improvement.

Note that while we have exclusively dealt with military conflicts, the domain–independent nature of the techniques suggests that the same strategy can be applied in other domains as well. The exact nature of the heuristics from Section 5.2, however, might need to be adapted to the specific application domain. For example, military conflicts are often described in documents adopting a style which is reminiscent of encyclopedia articles, exhibiting a tendency to mention dates wherever possible. Although a similar pattern might be expected for other types of historical events, the relative impact of co–occurring dates and document structure might vary. Furthermore, it is not clear whether these heuristics would be useful at all in the context of, e.g., contemporary events. When moving to news events, for example, a significant contribution of linguistic techniques can be expected to arrive at meaningful temporal relations from news stories. These relations could be combined with temporal relations that are mined from blog posts, requiring even other (heuristic) techniques. In each case, however, we are likely to end up with a combination of reliable quantitative information (dates and fuzzy time spans) and qualitative relations, the latter being typically less reliable due to the heuristic nature of extraction techniques or the limitations of linguistic analysis.

Finally, note that to the extent possible, we have exclusively focused on the problem of constructing temporal knowledge bases using qualitative relations. Dealing with events in practical applications often involves a number of additional challenges, which are, however, mostly orthogonal to the problem described in this paper. These include co–reference of event names (e.g., "First World War" vs. "the Great War"), ambiguity of event names (e.g., Iraq War) and normalising time expressions (e.g., next Monday).

Acknowledgement

Steven Schockaert is funded as a postdoctoral fellow of the Research Foundation – Flanders.

References

1. Alonso, O., Gertz, M., Baeza-Yates, R.: On the value of temporal information in information retrieval. ACM SIGIR Forum **41**(2), 35–41 (2007)
2. Harabagiu, S., Bejan, C.: Question answering based on temporal inference. In: Proceedings of the AAAI-2005 Workshop on Inference for Textual Question Answering (2005)
3. Moldovan, D., Clark, C., Harabagiu, S.: Temporal context representation and reasoning. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence (2005)
4. Pustejovksy, J., Knippen, R., Littman, J., Saurí, R.: Temporal and event information in natural language text. Language Resources and Evaluation **39**, 123–164 (2005)
5. Saquete, E., Martínez-Barco, P., Muñoz, R., Vicedo, J.: Splitting complex temporal questions for question answering systems. In: Proceedings of the 42nd Annual Meeting of the ACL (2004)
6. Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peas, A., de Rijke, M., Sacaleanu, B., Santos, D., Sutcliffe, R.: Overview of the CLEF 2005 multilingual question answering track. In: Proceedings of CLEF 2005 (2005)
7. Barzilay, R., Elhadad, M., McKeown, K.: Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research **17**, 35–55 (2002)
8. Mani, I., Schiffman, B., Zhang, J.: Inferring temporal ordering of events in news. In: Proceedings of the 2003 HLT-NAACL Conference, pp. 55–57 (2003)
9. Okazaki, N., Matsuo, Y., Ishizuka, M.: Improving chronological sentence ordering by precedence relation. In: Proceedings of the 20th International Conference on Computational Linguistics (2004)
10. Allan, J., Gupta, R., Khandelwal, V.: Temporal summaries of news topics. In: Proceedings of the 24th ACM SIGIR Conference, pp. 10–18 (2001)
11. Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proceedings of the 21st ACM SIGIR Conference, pp. 37–45 (1998)
12. Chieu, H., Lee, Y.: Query based event extraction along a timeline. In: Proceedings of the 27th ACM SIGIR Conference, pp. 425–432 (2004)
13. Prabowo, R., Thelwall, M., Alexandrov, M.: Generating overview timelines for major events in an RSS corpus. Journal of Informatics **1**, 131–144 (2007)
14. Swan, R., Allan, J.: Automatic generation of overview timelines. In: Proceedings of the 23rd ACM SIGIR Conference, pp. 49–56 (2000)
15. Allen, R.: A query interface for an event gazeteer. In: Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, pp. 72–73 (2004)
16. Mckay, D., Cunningham, S.: Mining dates from historical documents. In: technical report, Department of Computer Science, University of Waikato (2000)
17. Smith, D.: Detecting and browsing events in unstructured text. In: Proceedings of the 25th ACM SIGIR Conference, pp. 73–80 (2002)
18. Jones, C., Alani, H., Tudhope, D.: Geographic information retrieval with ontologies of place. In: Proceedings of the International Conference on Spatial Information Theory: Foundations of Geographic Information Science

19. Allen, J.: Maintaining knowledge about temporal intervals. Communications of the ACM **26**(11), 832–843 (1983)
20. Schockaert, S., De Cock, M., Kerre, E.: Fuzzifying Allen's temporal interval relations. IEEE Transactions on Fuzzy Systems **16**(2), 517–533 (2008)
21. Schockaert, S., De Cock, M.: Temporal reasoning about fuzzy intervals. Artificial Intelligence **172**, 1158–1193 (2008)
22. Schockaert, S., De Cock, M.: Efficient algorithms for fuzzy qualitative temporal reasoning. IEEE Transactions on Fuzzy Systems (to appear)
23. Mani, I., Wilson, G.: Robust temporal processing of news. In: Proceedings of the 38th Annual Meeting of the ACL, pp. 69–76 (2000)
24. Filatova, E., Hovy, E.: Assigning time-stamps to event-clauses. In: Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing, pp. 88–95 (2001)
25. Lapata, M., Lascarides, A.: Learning sentence–internal temporal relations. Journal of Artificial Intelligence Research **27**, 85–117 (2006)
26. Boguraev, B., Ando, R.: TimeML–compliant text analysis for temporal reasoning. In: Proceedings of the International Joint Conference on Articifical Intelligence, pp. 997–1003 (2005)
27. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from Flickr tags. In: Proceedings of the 30th ACM SIGIR Conference, pp. 103–110 (2007)
28. de Jong, F., Rode, H., Hiemstra, D.: Temporal language models for the disclosure of historical text. In: Proceedings of the 16th International Conference of the Association for History and Computing, pp. 161–168 (2005)
29. Diaz, F., Jones, R.: Using temporal profiles of queries for precision prediction. In: Proceedings of the 27th ACM SIGIR Conference, pp. 18–24 (2004)
30. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.M., Shaked, T., Soderland, S., Weld, D., Yates, A.: Web–scale information extraction in knowitall. In: Proceedings of the 13th International Conference on World Wide Web, pp. 100–110 (2004)
31. Ravichandran, D., Hovy, E.: Learning surface text patterns for a question answering system. In: Proceedings of the 40th Annual Meeting of the ACL, pp. 41–47 (2002)
32. Soubbotin, M., Soubbotin, S.: Patterns of potential answer expressions as clues to the right answer. In: Proceedings of the TREC-10 Conference, pp. 175–182 (2001)
33. Schockaert, S., De Cock, M., Kerre, E.: Acquiring vague temporal information from the web. In: Proceedings of the International Workshop on Fuzzy Logic on the Web, at WI-IAT 2008, pp. 265–268 (2008)
34. Schockaert, S., Ahn, D., De Cock, M., Kerre, E.: Question answering with imperfect temporal information. In: Proceedings of the 7th International Conference on Flexible Query Answering Systems, LNAI 4027, pp. 647–658 (2006)
35. Schockaert, S.: Construction of membership functions for fuzzy time periods. In: Proceedings of the ESSLLI 2005 Student Session, pp. 297–305 (2005)
36. Schockaert, S.: Reasoning about fuzzy temporal and spatial information from the web. Ph.D. thesis, Ghent University (2008)
37. Schokkenbroek, C.: News stories: structure, time and evaluation. Time & Society **8**(1), 59–98 (1999)