

Acquiring Vague Temporal Information from the Web

Steven Schockaert
Ghent University
Krijgslaan 281 - S9
9000 Gent, Belgium
Steven.Schockaert@UGent.be

Martine De Cock
Ghent University
Krijgslaan 281 - S9
9000 Gent, Belgium
Martine.DeCock@UGent.be

Etienne E. Kerre
Ghent University
Krijgslaan 281 - S9
9000 Gent, Belgium
Etienne.Kerre@UGent.be

Abstract

Many real-world information needs are naturally formulated as queries with temporal constraints. However, the structured temporal background information needed to support such constraints is usually not available to information retrieval systems. As an alternative, we automatically compile temporal knowledge bases from web documents, combining whatever quantitative and qualitative temporal information we can find about events of interest. By using simple heuristic techniques for temporal information extraction, 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.

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, 3, 14, 16]. 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 such as Google Maps¹. 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 are bound to fail more often. To find relevant temporal relations, we therefore utilize redundancy-based heuristics that yield a much higher recall than what we could expect from linguistic processing. A second, more fundamental problem is related to the ill-defined nature of the time spans of most historical events and time periods (e.g., the Renaissance, the Dotcom Bubble, World War II). As a consequence, time spans can often not be modelled as crisp intervals, and partial inconsistencies will occur among extracted temporal relations. For example, while most people would consider World War II to be completely before the Cold War, historians tend to point at 1917 as the real beginning of the Cold War, thus suggesting that World War II happened during the Cold War. To address these issues, we propose a framework based on fuzzy temporal reasoning [11, 12], allowing us to model time periods as fuzzy sets [10], as well as partial inconsistencies between temporal relations. The result of this fuzzy temporal reasoning phase is a more reliable, and logically consistent knowledge base.

2 Related Work

There is a large body of work on extracting temporal information from texts, although most of it is focused on news stories. For example, [6] is concerned with resolving temporal expressions such as *today*, *last week*, or *in April*. In [4], an attempt is made to automatically assign time stamps to event-clauses, while [5] deals with learning temporal relations between different clauses of a sentence. To facilitate machine learning approaches to temporal information ex-

¹<http://maps.google.com>

traction, the TimeML markup language has been conceived [8]. In [2], for instance, TimeBank, a TimeML annotated corpus, is used to train a system that recognizes events and temporal relations. In [7], temporal reasoning is used to support question answering, based on a temporal information extraction module 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 (e.g., [3, 9, 16]). Finally, in [14], co-occurrences of dates and place names in historical documents are used to identify significant events.

The techniques we propose to extract temporal relations are conceptually simpler than the aforementioned approaches, because we need to deal, in principle, with arbitrary web documents (different genres, writing styles, etc.), rather than only news stories or other collections of time-stamped documents. To the best of our knowledge, this is the first paper experimenting with fuzzy temporal reasoning to improve temporal information extraction.

3 Collecting Temporal Information

3.1 Fuzzy Time Spans

The beginning and ending dates of well-known events 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 Siege of Antwerp took place, we can submit queries such as “the Siege of Antwerp began on” and “the Siege of Antwerp 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 $\langle date \rangle$ ”. Note that the use of patterns to find appropriate entities is a standard technique in QA systems [15]. 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. The most significant beginning and ending dates that are found for such vague events can be aggregated to a fuzzy time span, i.e., a mapping from the universe of dates to the unit interval $[0, 1]$. If A is the fuzzy time span of an event e , for each date d , $A(d)$ represents the degree to which d is generally considered to be during e . We refer to [10] for a detailed discussion on the construction of such fuzzy time

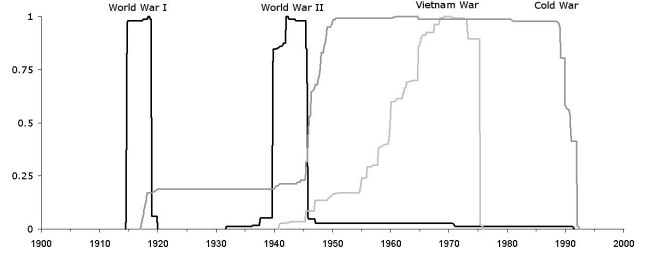


Figure 1. Fuzzy time spans of World War I, World War II, the Vietnam War and the Cold War.

spans. Figure 1 depicts the fuzzy time spans we found for World War I, World War II, the Vietnam War and the Cold War.

3.2 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 3.1 to fail. However, the actual time spans are usually not required in an IR setting: all we need to establish is whether or not two given events satisfy a given temporal relation. One way of checking temporal relations is by comparing the (fuzzy) time spans of both events, but we can also try to find evidence for temporal relations directly. In this section, we will 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.

3.2.1 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. For example, assume that we need to check whether a *before* relation holds between two events a and b . Let $D^a = \{d_1^a, d_2^a, \dots, d_n^a\}$ be dates that were found to occur within 200 characters of event a in web documents and let f_i^a be the number of times date d_i^a was thus found. Similarly, let $D^b = \{d_1^b, d_2^b, \dots, d_m^b\}$ be the dates that were found for event b , and let f_j^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*):

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

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

$$c(a, b) = 2 \cdot \max(0, \frac{pos(a, b)}{pos(a, b) + neg(a, b)} - 0.5) \quad (1)$$

provided that $pos(a, b) + neg(a, b) > 0$; otherwise, we define $c(a, b) = 0$. Note that a factor 2 is introduced to obtain a confidence score in $[0, 1]$. It is easy to see that $c(a, b) = 1$ iff all dates in D^a are strictly before all dates in D^b . Because the confidence score $c(a, b)$ becomes more reliable as the sizes of D^a and D^b increase, we sometimes require that $n \geq 5$ and $m \geq 5$. For *during* relations, a similar technique is used; we omit the details.

3.2.2 Document structure

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., [13]). Similarly, although historical documents have a tendency to digress, thereby linking events from the main linear narrative to earlier or later events [14], we can still expect the order of occurrence to be chronological more often than not.

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*. If n_1 is significantly higher than n_2 , this is a strong indication for *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. 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.

3.3 Fuzzy Temporal Reasoning

Using the techniques from Sections 3.1 and 3.2, we can construct a knowledge base (KB) containing (fuzzy) time

spans and qualitative temporal relations involving a given collection of events. While we can expect most of the information in this KB to be correct, it will inevitably contain some mistakes as well. In part this can be explained by the heuristic nature of the adopted techniques, although other causes apply as well (e.g., using the web as a source for data acquisition). An even more fundamental problem is related to the vagueness of event boundaries. As an example, consider the vague boundaries of World War II from Figure 1. Depending on the point of view taken, the Japanese Invasion of China (1937) is either before or during World War II. Accordingly, the techniques outlined above yield conflicting information. The most appropriate solution, in this case, is not to ignore either of the viewpoints, but to model that both relations are satisfied to a certain degree between 0 and 1. Using a fuzzy temporal reasoner, it is possible to find a consistent interpretation, assigning degrees of membership to the relations from the initial KB. We refer to [11, 12] for a detailed discussion on the semantics of such fuzzy temporal relations, as well as sound and complete reasoning algorithms. Note that the purpose of fuzzy temporal reasoning in this context is twofold: inferring new temporal relations between the events in the KB and increasing its reliability by detecting and repairing (partial) inconsistencies.

4 Experimental Results

Our focus is on the automatic acquisition of (fuzzy) temporal information from the web, given a collection of events of interest. To eliminate the need for deep linguistic processing from the present analysis, we focus on a class of events whose occurrences can be straightforwardly recognized in texts: named military conflicts (e.g., the Battle of the Bulge, the Vietnam War).

To generate a ground truth for our event-based retrieval task, we extracted lists of military conflicts, mostly battles, that are considered to be during various wars according to Wikipedia². For the 25 wars from Table 1, this led to a total number of 1674 events. In our evaluation, we look at how well four different systems succeed in deciding which of these events were during World War II, which were during World War I, etc. The first system, *B1* (Baseline 1), only uses (fuzzy) time spans and the strategy from Section 3.2.1. Similarly, *B2* (Baseline 2) only uses the strategy from Section 3.2.2. Next, *B3* (Baseline 3) combines both strategies, but does not apply any fuzzy reasoning. Finally, *FR* (Fuzzy Reasoning) additionally uses fuzzy temporal reasoning, as explained in Section 3.3.

For each war *W*, a different knowledge base was compiled, based on approximately 4000 documents related to each war that were found using Google and Yahoo!. Next,

²[http://en.wikipedia.org/wiki/Category:](http://en.wikipedia.org/wiki/Category:Battles_by_war)

Battles_by_war, accessed October 29, 2007.

Table 1. Comparison of the different systems in terms of average precision.

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

the 4 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. The average precision of the rankings for all 25 wars is shown in Table 1. A particularly interesting observation is that the performance of *B1* is largely complementary to the performance of *B2*, explaining the large improvement of *B3* over both *B1* and *B2*. Next, as the results for *FR* 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).

5 Concluding remarks

In this paper, we have proposed a novel strategy 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 events.

Acknowledgments

The first author was funded as a research assistant of the Research Foundation – Flanders.

References

- [1] O. Alonso, M. Gertz, and R. Baeza-Yates. On the value of temporal information in information retrieval. *ACM SIGIR Forum*, 41(2):35–41, 2007.
- [2] B. Boguraev and R. Ando. TimeML–compliant text analysis for temporal reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 997–1003, 2005.
- [3] H. Chieu and Y. Lee. Query based event extraction along a timeline. In *Proceedings of the 27th ACM SIGIR Conference*, pages 425–432, 2004.
- [4] E. Filatova and E. Hovy. Assigning time-stamps to event-clauses. In *Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing*, pages 88–95, 2001.
- [5] M. Lapata and A. Lascarides. Learning sentence–internal temporal relations. *Journal of Artificial Intelligence Research*, 27:85–117, 2006.
- [6] I. Mani and G. Wilson. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 69–76, 2000.
- [7] D. Moldovan, C. Clark, and S. Harabagiu. Temporal context representation and reasoning. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 2005.
- [8] J. Pustejovsky, R. Knippen, J. Littman, and R. Saurí. Temporal and event information in natural language text. *Language Resources and Evaluation*, 39:123–164, 2005.
- [9] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *Proceedings of the 30th ACM SIGIR Conference*, pages 103–110, 2007.
- [10] S. Schockaert. Construction of membership functions for fuzzy time periods. In *Proceedings of the ESSLLI 2005 Student Session*, pages 297–305, 2005.
- [11] S. Schockaert and M. De Cock. Temporal reasoning about fuzzy intervals. *Artificial Intelligence*, 172:1158–1193, 2008.
- [12] S. Schockaert, M. De Cock, and E. Kerre. Fuzzifying Allen’s temporal interval relations. *IEEE Transactions on Fuzzy Systems*, 16(2):517–533, 2008.
- [13] C. Schokkenbroek. News stories: structure, time and evaluation. *Time & Society*, 8(1):59–98, 1999.
- [14] D. Smith. Detecting and browsing events in unstructured text. In *Proceedings of the 25th ACM SIGIR Conference*, pages 73–80, 2002.
- [15] M. Soubbotin and S. Soubbotin. Patterns of potential answer expressions as clues to the right answer. In *Proceedings of the TREC-10 Conference*, pages 175–182, 2001.
- [16] R. Swan and J. Allan. Automatic generation of overview timelines. In *Proceedings of the 23rd ACM SIGIR Conference*, pages 49–56, 2000.