Supporting temporal question answering: strategies for offline data collection

David Ahn

ISLA, University of Amsterdam
ahn@science.uva.nl

Steven Schockaert, Martine De Cock, and Etienne Kerre

Ghent University
Steven.Schockaert,Martine.DeCock,Etienne.Kerre@UGent.be

Abstract

We pursue two strategies for offline data collection for a temporal question answering system that uses both quantitative methods and fuzzy methods to reason about time and events. The first strategy extracts event descriptions from the structured year entries in the online encyclopedia Wikipedia, yielding clean quantitative temporal information about a range of events. The second strategy mines the web using patterns indicating temporal relations between events and times and between events. Web mining leverages the volume of data available on the web to find qualitative temporal relations between known events and new, related events and to build fuzzy time spans for events for which we lack crisp metric temporal information.

1 Introduction

Time structures our world, and the questions we ask reflect that. Not only do we want to know quantitative information—when did some event happen or how long did some state of affairs persist—but we also want qualitative information—what was going on before or during major events, or what happened afterwards. While the amount of information available to answer such questions continues to increase, the temporal information needed is not always fully specified. No information source is obliged to timestamp every referenced event, so while evidence for qualitative temporal relations abounds, there is often no quantitative information to verify it. Furthermore, many events, such as the Cold War or the Great Depression, are inherently vague—with gradual beginnings or endings—or ill-defined aggregations of smaller events.

In order for a temporal QA system to be able to make use of such limited, incomplete temporal information, careful consideration must be given both to

the extraction of temporal information it needs and to the temporal reasoning mechanisms it employs. We are presently at work on a temporal QA system that provides access to events extracted from Wikipedia and satellite events mined from the web and that models the time span of vague events as fuzzy sets and qualitative temporal relations as fuzzy relations. In this paper, we focus on the creation of the knowledge base of events and temporal information, which takes place offline, prior to any user interaction. A separate paper [10] describes the fuzzy reasoning mechanisms the system uses.

In §2 and §3, we introduce temporal questions and sketch the architecture of our temporal QA system. In §4, we describe event extraction from Wikipedia, an online encyclopedia. In §5, we describe web mining for fuzzy and qualitative temporal information. Note that the work described here is still in progress, so while our extraction methods are in use, we are still experimenting with them, and the QA system is not yet complete.

2 Temporal questions

There are a variety of question types that fall under the umbrella of *temporal questions*, including questions that ask for times as answers, those that ask for temporal relations, and those that ask for information restricted to a certain time period [6]. The degree of explicitness of temporal reference in a temporal question also varies significantly: some temporal questions refer explicitly to a date or time, while others refer to times only implicitly, by reference to events or states. Here, we focus on *temporally restricted questions*, and in particular, those restricted by events, such as these (from the CLEF 2005 QA track):

(1) Who played the role of Superman before he was paralyzed?

(2) What disease did many American soldiers get after the Gulf War?

The data collection we describe, though, can be used to support the answering of other types of temporal questions, as well.

Temporally restricted questions consist of two parts: the main clause, which indicates the information request, and the temporal restriction, which is a subordinate clause or PP headed by a temporal preposition or connective, such as *before, after, during*, etc., which we refer to as *temporal signals* [7]. The temporal signal that connects the two parts of a temporally restricted question indicates the *temporal relation* that must hold between the time spans of the restricting event and the requested events. Since much of the temporal information we have access to regarding events is vague and incomplete, we explore the use of *fuzzy temporal reasoning* for temporal QA, instead of the standard Allen algebra of temporal interval relations [1]. The model we use is a generalization of Allen's algebra that is suitable for vague events and relations. For crisp events, our reasoning algorithm is equivalent to Allen's path-consistency algorithm. For vague events, fuzzy relations can express that a given qualitative relation is only satisfied to a certain degree [11].

3 Architecture of a temporal QA system

Our temporal QA system follows a strategy of extracting information likely to be useful in answering questions—in our case, a knowledge base of events and temporal relations—in a pre-processing stage, before any questions are asked [3,4]. The system consists of several components: a question analysis module, the knowledge base, and an answer selection module.

Question analysis: Since we are focusing on temporal questions in which a temporal relation restricts the information being queried, our question analysis module must separate the non-temporal part of the question—the actual information request—from the temporal restriction. Our question analysis module parses the question and extracts phrases headed by temporal signals as potential temporal restrictions. It then uses standard pattern-based techniques to extract keyword queries and the expected answer type.

Knowledge base: The knowledge base (KB) consists of two parts: an XML database containing descriptions of individual events and a temporal relation network containing inclusion and before/after relations for events in the KB. Quantitative temporal information about events (i.e., starting and ending dates for crisp events and fuzzy sets for vague ones) is contained in the XML database. The rest of this paper describes the construction of the KB.

Answer selection: To answer a temporally restricted question, we must find events that match the non-temporal part of the question and filter out those that do not satisfy the restriction. We treat the problem of finding the events as a retrieval problem, using the keyword queries from question analysis, with event descriptions as target documents. Checking whether an event satisfies the restriction is a matter of inferring whether an appropriate qualitative temporal relation holds between the event and a time or event matching the restriction. We use IR techniques to find events matching the restriction, and we use both quantitative and fuzzy temporal reasoning to make the inference [10]. From the remaining event descriptions, we use standard techniques to extract an answer. Typically, the information request is mapped to a named entity type by the question analysis module, so appropriate named entities are harvested and scored and the top-scoring entity is returned.

4 Extracting events from Wikipedia

Wikipedia is a free, open-domain, web-based encyclopedia [12]. In addition to traditional encyclopedia entries, it also has entries for a variety of time periods, which contain lists of historical and/or current events. We extract events from the entries for years. The standardized formatting of year entries in Wikipedia, together with the wiki markup used for this formatting, makes extracting event descriptions straightforward. A typical year entry contains sections delimited by the second-level headings `==Events==`, `==Births==`, and `==Deaths==`. Each of these sections is optionally split into subsections de-

limited by third-level headings indicating months (e.g., `===May===`) or the lack of a date (`===Unknown date===`). Within these subsections are asterisk-delimited lists, each item of which corresponds to an event (or a date with a list of events). Event descriptions begin with a date or a date range, if known, and then continue with one or two sentences describing the event. This text contains phrases marked up as wiki links (pointers to Wikipedia entries):

(3) [[March 10]] - The [[New Hampshire]] primary is won by [[Henry Cabot Lodge]], Ambassador to [[South Vietnam]].

Sometimes, an event description begins with a wiki link, set off with a colon, indicating a larger event (which we call a *super-event*) of which it is a part:

(4) [[August 8]] - [[Watergate scandal]]: US President [[Richard Nixon]] announces his resignation (effective [[August 9]])

Given the structured nature of year entries, simple hand-built patterns can be used to perform what amounts to shallow semantic interpretation, extracting event descriptions from the entries, including temporal location information and limited participant and mereological information (via wiki links and super-events, when present). For each extracted event description, the date(s) and any embedded wiki links are extracted, using simple pattern-matching, and the text of the description is parsed. This information is added to our XML database as an `event` element with the following sub-elements: `date`/`start_date`/`end_date` (normalized dates; which one(s) depends on what is given in the entry), `super_event` (wiki link to super event, if present), `description` (text of the description), and `parse` (converted to XML).

From the entries for the years from 1600 to 2005, we have extracted about 33,000 events, somewhat over half (about 19,000) birth and death events.

5 Web mining for fuzzy and qualitative information

The basic idea behind web mining is that there is enough information on the web that if there is a significant connection between two events, we should be able to find this connection by searching for patterns that typically express it. We use web mining to build representations of the time span of vague events and to find both additional events related to events already in the KB and for qualitative temporal relations between events in the KB.

While most of the smaller-scale events extracted from Wikipedia come with quantitative temporal information, it is not always fully specified. Furthermore, many of the super-events from Wikipedia, as well as the new events we mine from the web, lack such information. We cope with this by searching the web for beginning and ending dates using a simple pattern-based approach—sending patterns to Google and extracting information from the returned snippets. The patterns we use include, e.g., ⟨*event*⟩ *began on* ⟨*date*⟩ and ⟨*event*⟩ *lasted until* ⟨*date*⟩. If there is sufficient agreement among different web pages about the beginning and ending date of an event, we represent the

time span of this event as an interval. If not, we use the techniques described in [9] to construct a suitable fuzzy set [13] to represent the time span, which is stored in the XML database as part of the event representation. Of course, for some events, we may fail to find sufficient information about beginning or ending dates, in which case they remain undated, or *ungrounded*, events.

We also use a pattern-based approach to mine the web both for new events and for temporal relations relating ungrounded events to grounded events. Because new events are only usable if they can be temporally connected to events already in the KB, we can use a uniform set of hand-crafted patterns that indicate a temporal relation between events. The patterns we use include, e.g., $\langle NP_1 \rangle$ *gave way to* $\langle NP_2 \rangle$ and $\langle NP_2 \rangle$ *took place after* $\langle NP_1 \rangle$, for before/after relations, and $\langle NP_1 \rangle$ *and other events during* $\langle NP_2 \rangle$ and $\langle NP_1 \rangle$ *took place during* $\langle NP_2 \rangle$, for inclusion relations. All our patterns relate NP descriptions of events, which means that they can only be used with known events that have NP descriptions. Fortunately, this includes all super-events and newly mined events, which make up most of the ungrounded events in the KB.

Since we can use the same patterns to mine for new events and for temporal relations for ungrounded events, we combine the tasks. Our basic procedure for mining with these patterns is as follows. Substitute either $\langle NP_1 \rangle$ or $\langle NP_2 \rangle$ with the NP description of a known event, send the resulting pattern to Google, and parse and tag the returned snippets with named entities. Extract NPs in the other NP position, discarding those tagged as `person`, `location`, or `organization`. For each remaining NP, if it refers to a known event in the KB, add to the temporal relation network a link between the original known event and the event referred to by the mined NP. Otherwise, add a new event for the mined NP and a link between the original event and the new event.

The hardest step is determining whether a mined NP refers to an event already in the KB. Coreference resolution is clearly necessary to temporal constraints on coreferring event descriptions, but to maintain consistency in the KB, we must be careful in asserting coreference. We do not try to solve the cross-document coreference task completely but instead split mined NPs into two groups. The first group, which includes all indefinite, demonstrative, quantificational, and pronominal NPs, is added to the temporal network but is never considered for coreference. The second group contains NPs that we are confident refer to a unique event and that are thus candidates for coreference.

To find these NPs, we are experimenting with heuristics to determine unique reference. Some heuristics are capitalization-based, while others are based on collocation measures using web hit counts [5,8], such as the ratio between the number of hits for the entire NP and the product of the hits for each of the individual words in the NP (similar to pointwise mutual information [2]). When an NP is determined to be uniquely referring, we use string matching to determine whether it is coreferent with an existing event description. If it is, the temporal relation mined is added to the KB for this existing event. Otherwise, a new event and relation are added to the KB.

6 Conclusion

We have described the construction of a knowledge base of events and temporal relations for a temporal QA system. We take advantage of a freely available, structured resource—Wikipedia—to obtain relatively accurate quantitative information about events. We also mine the web to build fuzzy sets for vague events and to find events for which we can only get qualitative information.

Acknowledgements The first author was supported by the Netherlands Organization for Scientific Research (NWO), under project number 612.066.302. The second author was supported by a PhD grant from the Research Foundation – Flanders.

References

[1] Allen, J., *Maintaining knowledge about temporal intervals*, Communications of the ACM **26** (1983), pp. 832–843.

[2] Church, K. et al., *Using statistics in lexical analysis*, in: *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Lawrence Erlbaum, 1991 .

[3] Fleischman, M., E. Hovy and A. Echihabi, *Offline strategies for online question answering: Answering questions before they are asked*, in: *ACL 2003*, 2003.

[4] Jijkoun, V., G. Mishne and M. de Rijke, *Preprocessing documents to answer Dutch questions*, in: *BNAIC'03*, 2003.

[5] Magnini, B., M. Negri, R. Prevete and H. Tanev, *Is it the right answer? Exploiting web redundancy for answer validation*, in: *ACL-02*, 2002.

[6] Pustejovsky, J. et al., *TERQAS final report*, `http://www.cs.brandeis.edu/~jamesp/arda/time/readings/TERQAS-FINAL-REPORT.pdf` (2002).

[7] Saurí, R. et al., *TimeML annotation guidelines*, `http://www.cs.brandeis.edu/~jamesp/arda/time/timeMLdocs/annguide12wp.pdf` (2004).

[8] Schlobach, S., D. Ahn, M. de Rijke and V. Jijkoun, *Data-driven type checking in open domain question answering*, Journal of Applied Logic (to appear).

[9] Schockaert, S., *Construction of membership functions for fuzzy time periods*, in: J. Gervain, editor, *ESSLLI 2005 Student Session*, 2005.

[10] Schockaert, S., D. Ahn, M. De Cock and E. E. Kerre, *Question answering with imperfect temporal information*, in: *FQAS-2006*, to appear.

[11] Schockaert, S., M. De Cock and E. E. Kerre, *Imprecise temporal interval relations*, in: *LNCS 3849*, Springer, 2006 .

[12] Wikipedia, *Wikipedia, the free encyclopedia*, `http://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=35397363`, [Accessed 16-January-2006].

[13] Zadeh, L. A., *Fuzzy sets*, Information and Control **8** (1965), pp. 338–353.