Temporal Granularity and Temporal Tagging of Text

Inderjeet Mani and **George Wilson**
The MITRE Corporation, W640
11493 Sunset Hills Road
Reston, Virginia 22090
{imani,gwilson}@mitre.org

Abstract

We discuss the design of a scheme for temporal annotation of text. In contrast to previous temporal annotation schemes, our emphasis is on recognizing as well as resolving temporal references. We provide an overview of the design of our temporal annotation scheme, and illustrate the granularity issues which arise in the course of temporal tagging.

Introduction

The temporal distinctions made by natural languages pose many interesting challenges for knowledge representation in general, and for a theory of granularity in particular. Our interest here is in the information extraction problem of automatic tagging of temporal references in text. We have developed a temporal annotation scheme that has been used to annotate a corpus of documents, for use as a resource in training and testing temporal taggers. In contrast to previous temporal annotation schemes, e.g., (MUC-7 98), our emphasis is on recognizing as well as resolving temporal references. In what follows, we first introduce some notions of granularity in temporal expressions, and then discuss the specific case of indexicals. We then describe our approach to specific temporal phenomena in natural language, in terms of their representation in the annotation scheme. This short paper provides an overview of the design of our temporal annotation scheme, and illustrates the granularity issues which arise in the course of temporal tagging.

Temporal expressions reflect two well-known aspects of granularity. The first aspect has to do with fuzzy boundaries, also called "vague" or "non-crisp" boundaries in the spatial granularity literature, e.g., (Cohn and Gotts 96). Thus, just like "Southern England" is fuzzy as to its geographical boundaries, "Fall" is fuzzy as to when it starts and ends. "Saturday morning" is also fuzzy, for example, as to whether 2 am on Saturday is to be included. Likewise, in "now that we walk upright, we have lots of back problems" (Perry 00), "now" references a fuzzy interval of time extending from some period when man emerged to the time of the speech. (Note that this fuzziness is somewhat different from the issue of whether "Saturday morning" should be viewed as denoting an open or closed interval).

A second aspect of granularity is the fact that there can be different levels of "zoom". Zooming involves some notion of approximation, i.e., movement on a scale from more precise (or concrete or specific) to less precise (or abstract or general), or vice versa. Zooming in involves including more details in the representation, whereas zooming out abstracts away and ignores some details. Thus, an instant on one scale becomes an interval after zooming in. For example, a day may be broken up into 24 hours, an hour into sixty minutes, etc. Likewise, intervals may be zoomed out into instants.

There have been many accounts of zooming operations. For example, (Pianesi and Varzi 96) have characterized zooming in terms of refinement relations, in particular representing the interval to instant zoom in terms of a "minimal divisor" on structures corresponding to sets of events, where temporal differences within the divisor are neglected. (Becher et al. 98) represent changes in granularity based on a containment relation between intervals. (Euzenat 95) has investigated a number of fundamental algebraic properties of granularity operators, which are defined as mappings on ontological relations. (Mani 98), extending the abstraction theories of (Hobbs 84) and (Giunchiglia and Walsh 92), has characterized zooming operations as abstractions over the orderings induced by subset and containment relations.

Indexicals

In recognizing and resolving a variety of time expressios, we are especially interested in indexical time expressions, which designate times that are dependent on the speaker and some 'reference" time. Examples of indexicals include "now", "today", "yesterday", "tomorrow", "next Tuesday", "the following Tuesday", "two weeks ago", "20 mins after the next hour", etc. These indexicals include expressions which information extraction practitioners have termed "relative times". These include the times in examples like "On Wednesday June 1st, she gave birth to twins. The following Tuesday, she won the New York Lottery.", where the Wednesday is relative to the year specified somewhere earlier in the text, and where the Tuesday is relative to the Wednesday.

Annotation Scheme

Any annotation scheme should aim to be simple enough to be executed efficiently and accurately by humans, and yet precise enough for use in various natural language processing tasks. As in the (KSL-Time 99) ontology, time points are represented on a time line, and intervals (which are treated as sets of more than one time point) have start and end points on the time line. In our case, however, the time line is a calendric one (rather than the real number line), with points on this calendric line characterized in terms of the ISO standard CC:YY:MM:DD:HH:XX:SS, with an optional time zone (ISO-8601 97). Since the goal of the tagging is to resolve (rather than just recognize) temporal references, we map temporal expressions to points or intervals in this calendric time line. In doing so, durative expressions like "from 3 pm to 6 pm" are tagged as intervals. So are non-durative expressions which map to sets of more than one point on the calendric line (e.g., "last week", "the first three weeks in November"). Other expressions which designate a point on the calendric line are tagged as points. The ISO time spec introduces several levels of "zoom" granularity, corresponding to each of the 7 units in the time spec.

Each temporal expression is tagged with a TIMEX tag, of type TIME or DATE (the latter is inherited from the MUC-7 spec). The tag has also a VAL attribute, which indicates either starting time point in case of an interval, or the time point otherwise. For example, "yesterday" will have as its VAL the previous day's date with respect to the reference time. The VAL1 attribute references the end point in the case of an interval. Both VAL and VAL1 are expressed in the ISO time specification. The optional attribute OFFSET, which also uses the ISO time specification, is used when an indexical is expressed with an unknown reference time: e.g., "the next day" may have an offset of 00:00:00:01 which can be useful even if the reference time isn't known. When OFFSET is specified, DIRECTION (+ or -) is also specified.

It is worth noting that there are several kinds of temporal expressions that are not to be tagged, and that other expressions tagged as a time expression are not assigned a value, because doing so would violate the simplicity and preciseness requirements. We do not tag sets of intervals, like "the first three days of every month", unanchored intervals, such as "half an hour (long)", and event-dependent expressions, like "five days after he came back". Non-specific time expressions like generics, e.g., "April is usually wet", or "today's youth", and indefinites, e.g., "a Tuesday", are tagged without a value, as are gapped intervals, e.g., "every Tuesday in 1999", and expressions with vague quantifiers, e.g., "about three years ago". Expressions that do refer to a specific time whose boundaries are indeterminate, e.g., "midday", "Fall 1999", are tagged as time expressions but are not assigned a value. Finally, expressions which are ambiguous without a strongly preferred reading are left untagged.

To convey a flavor of the annotation, here is an example tagged sentence:

<s>The Foreign Minister told Thailand's Nation Newspaper <TIMEX TYPE="DATE" val="19980104">Sunday</TIMEX>, Pol Pot had left Cambodia but was not in Thailand, ending credence to a claim <TIMEX TYPE="DATE" val="19971228" val1="19980103">last week</TIMEX> the aged and ailing former Khmer Rouge leader had fled to China.</s>

Granularity Patterns

In a given text, indexical time expressions can exhibit many different patterns of zooming. In the case of indexicals where a specific time is being referred to and where the reference time is explicit, the level of approximation is roughly the least significant unit used in the indexical. Thus, "Today/yesterday/next Tuesday, March 2nd" refers to a specific time at day granularity, whereas "next month/nine months ago" refers to a specific time at month granularity. When indexicals are composed together, the head indexical usually determines the granularity based on the above least significant unit principle. Thus, in "one year ago today/yesterday/next Tuesday", we have day granularity, whereas in "one year ago next week", we have week granularity. There are exceptions to this principle, such as cases where "next year" is used to refer to New Year's Day. In addition, granularity zooming appears to involve rounding to the nearest "attractor" time, e.g., "twenty to nine" or "8:40" rather than "8:39 p.m.", "a quarter to four" rather than "3:42". Although English doesn't lexicalize time units in between hours and minutes, there appears to be a preferred granularity in describing time of day of 15 minutes or 5 minutes.

Interestingly, tense distinctions can span various time granularities. For example, some languages express a 5-way past tense distinction, by means of tense auxiliarie s, between a few hours ago/ one day ago/ a few weeks ago/ a few months ago/ distant past (Comrie 85). This means that the language provides specific levels of approximation in terms of indexical descriptions of when events occurred (i.e., with respect to distance from a reference time).

The zooming appears to serve a variety of functions: it may be used to further specify a time, e.g., "last week" followed by a zoom in to "on Monday", or to shift the focus, e.g., subsequently zooming out to "the week before". (These zooming principles are not stated in the annotation guidelines, as they assumed to be part of the annotator's understanding of the language.)

For many temporal expressions, the granularity may be underspecified. For example, "he graduated/bought a car three years ago" could mean year, week, day etc. In such a case, the annotation scheme explicitly calls for year granularity, since this is the least granularity committed to in the text. In contrast, "he graduated/bought a car three years ago today" specifies the granularity (day). Of course, even when the granularity is specified, the range of the interval may be ambiguous, e.g., "last year" (when uttered in the year 2000) could mean the year 1999, or the previous fiscal year (which runs from October 1, 1998 to September 30, 1999), while "the last year" can mean the final element in a series of years previously referred to, or to an interval of 12 months ending NOW. This sort of ambiguity is to be distinguished from granularity and indexicality.

Status

The annotation scheme has been used to annotate a corpus of news documents. We have so far annotated 80,000 words of text, consisting of print news from the New York Times and broadcast news (Voice of America) from the TDT2 (TDT2 99) corpus. We are also currently engaged in an effort within the information extraction community to annotate a larger sample of the TDT2 collection, as well as a collection of scheduling dialogs, and to conduct an inter-annotator reliability study. A preliminary time tagging program has been built, to tag and resolve time expressions according to this specification. Like the work of (Wiebe et al. 98), this tagger resolves time expressions. However, it starts from a part-of-speech representation rather than a semantic one, and thus has much fewer knowledge requirements; it also tracks shifts in reference times (used to resolve references such as "Thursday, Dominique Strauss-Kahn, France's finance minister, said: 'Today is clearly a historic day... '"). We have also taken the first steps towards extending the temporal tagging to events, providing a basis for the construction of event chronologies. These aspects, along with program performance evaluation, will be discussed in longer papers.

Conclusion

In conclusion, we would like to emphasize two points. First, the creation of text corpora with temporal annotations at the semantic level makes available a wealth of empirical NL data related to temporal granularity that can be valuable in theory construction as well as system development. Annotated text corpora have, after all, revolutionized progress on practical tasks such as information extraction, part-of-speech tagging, and parsing, among other NLP areas. Second, the extraction of temporal information from text, as carried out by time taggers, can be used to instantiate much richer knowledge bases designed to represent and reason with granularity. For example, the (KSL-Time 99) ontology specifies precedence and equality relations based on time line location for particular granularities (second /minute /hour /day /month /year/infinite), as well as across granularities when there isn't an overlap. In future, we will be exploring a variety of methods for extracting event chronologies for text. Being able to reason with granularities in such chronologies will be very useful.

Acknowledgments

We are grateful to Lisa Ferro, Chung Hee Hwang, and Leo Obrst for comments on the annotation scheme.

References

Becher, G., Clerin-Debart, F. and Enjalbert, P. 1998. A Model for Time Granularity in Natural Language. In *Proceedings of the International Workshop on Temporal Representation and Reasoning*, Sanibel Island, Florida, 16-17 May.

Cohn, A.G. and Gotts, N. M. 1996. *The Egg-Yolk Representation of Regions with Indeterminate Boundaries*. In Burrough, P. A. and Frank, A. U., eds. *Geographic Objects with Indeterminate Boundaries*, 171–188. Bristol, PA: Taylor & Francis Inc.

Comrie, B. 1985. *Tense*. Cambridge: Cambridge University Press.

Euzenat, J. An Algebraic Approach to Granularity in Qualitative Time and Space Representation. In *Proceedings of IJCAI95*.

Giunchiglia, F. and Walsh, T. 1992. A Theory of Abstraction. *Artificial Intelligence*, 57, 2-3, 323–390.

Hobbs, J. 1984. Granularity. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'84)*.

ISO-8601. 1997. ftp://ftp.qsl.net/pub/g1smd/8601v03.pdf

KSL Time Ontology. 1999. http://www.ksl.Stanford.EDU/ontologies/time/

Mani, I. 1998. A Theory of Granularity and its Application to Problems of Polysemy and Underspecification of Meaning, *Proceedings of KR'98*, Trento, Italy.

MUC-7. *Proceedings of the Seventh Message Understanding Conference*. Defense Advanced Research Projects Agency (DARPA), 29 April to 1 May, 1998.

Perry, J. 2000. Indexicals and Demonstratives. In Hale, R. and Wright, C., eds. *Companion to the Philosophy of Language*. Oxford: Blackwells.

Pianesi, F. and Varzi, A. 1996. Refining Temporal Reference in Event Structures. *Notre Dame Journal of Formal Logic*, 37, 1, 71–83.

TDT2. 1999. http://morph.ldc.upenn.edu/Catalog /LDC99T37.html

Wiebe, J., O'Hara, T.P., Ohrstrom-Sandgren, T. and McKeever, K. J. 1998. An Empirical Approach to Temporal Reference Resolution. *Journal of Artificial Intelligence Research*, 9, 247–293.