Extracting Temporal Information from Text

Andrea Setzer
University of Sheffield

Abstract

The extraction of temporal information from text is an important and challenging task, yet it has been largely ignored in Information Extraction work so far. To address this problem, we aim to develop an annotation scheme, a theory of how temporal information is conveyed and heuristics to extract temporal information from text. Our approach will be introduced in this paper.

1 Introduction

A vast amount of textual information is available nowadays, from a large number of sources, many of them electronic. One way of dealing with this amount is to use Information Extraction (IE) systems and techniques. IE analyses documents in order to find information about prespecified events (or entities or relations) and to map them into predefined, structured representations, called templates. For an overview of IE see for example Gaizauskas and Wilks (1998) and Cowie and Lehnert (1996).

Although the extraction of temporal information from text is an important, interesting and challenging task, it has received relatively little attention, one of the few exceptions being the 7th Message Understanding Conference, where the extraction of limited temporal information was a part of one of the tasks.

To address this problem, our work aims to develop an annotation scheme for temporal information, a theory of how temporal information is conveyed and heuristics to extract temporal information from text. The approach is as general as possible and will attempt to locate all
events described in the text in time. Even for specific Information Extraction tasks, where only certain events are extracted, is it not clear whether these events can be temporally located independently of other 'non-IE' events. Our approach is a corpus-oriented one and will analyse `real' texts to find out how temporal information is conveyed and what information is necessary to place events in time. The chosen genre is newspaper articles, since large corpora are available and much work has been done on this genre.

The work is proceeding in the following three interrelated phases.

Phase one deals with the development of an annotation scheme. Annotations, explanatory notes about the text, can be used for many purposes in IE. Two important uses are that they are a way of validating the representation scheme and that an evaluation and training corpus can be provided by manually annotating texts. The results of an IE system can then be compared against this corpus. Two different kinds of annotations are necessary. A clause-level 1 annotation defines the events described in a clause, their arguments (participants, temporal expressions and so on), whereas an inter clausal-level annotation, which uses the output of analysing the sentence-level annotation, describes the temporal relations between clause-level events. Once this phase is finished, a small corpus of hand annotated text will be produced.

The next phase is to generalise from the data gathered in the phase above and develop a theory of how temporal information is conveyed in newspaper articles.

The output of the first two phases is a specification of an annotation scheme, a theory of how temporal information is conveyed and a number of annotated texts. In a third phase, based on the outputs of the first two

phases, heuristics will be developed, which when implemented would extract the temporal information from annotated texts. Or more precisely, extract all the information necessary to establish the temporal relations between the events described. A possible extension of my work would be to implement one or more of the heuristics developed and evaluate them against held-out annotated data, thus showing that is is possible to automate the process of analysing (manually) annotated text and establishing the temporal relations between events described.

We are currently at the end of phase one and this paper will describe our findings so far. Sections 2 and 3 give on overview of the kind of temporal information found in newswire articles and an introduction to the temporal ontology we propose to use. The representation of events and temporal relations between them is dealt with in section 4 and the annotation scheme and reasons for developing it are described in section 5. Details about the remaining future work are given in section 6.

It should be stressed that this work is still in early stages and that concrete results have not yet been obtained.

2 Temporal Information

To find out how events are located on a timeline, it is necessary to find out (a) what events are, what they comprise and how they are indicated in text and (b) what the temporal information conveyed in text is, what is explicit and what implicit and what has to be inferred.

One phenomenon that became apparent whilst analysing newswire articles is that we can identify three levels of events (although other views are possible). Example 1 illustrates this.

`(1) The plane was seen hitting the water shortly after 11 a.m. by a fisherman, who radioed the Coast Guard, said Petty Officer Jeff Fenn, a spokesman for the base at Governors Island in New York Harbor.`

*Events can be **reported** by someone, indicated by a `reporting' verb (report, say ...). There is usually no temporal information given about these events, although they happen after the reported event and before the article appears in the newspaper. When they happen is not as important as their role of identifying the source of information for the reported event. In example 1: ([...] said Petty Officer Jeff Fenn [...]). Events can be `**experienced**' or observed, the observation temporally coinciding with (part of) the observed event. It is not yet clear, how important events of this kind are. In example 1: (The plane was seen [...] by a fisherman [...]). And then events actually **occur**, the most frequent and most important type of event. In example 1: ([...] hitting the water [...]).*

When we come to the ways temporal information is conveyed in real text, then we face a wide variety of factors. The means can be explicit or implicit, locate events absolutely or relatively. Temporal prepositional phrases, temporal adverbial phrases, verbal tense and aspect (in the Vendlerian (1967) sense of aspect), temporal subordinate conjunctions and narrative sequence are amongst those means. The temporal information regarding one particular event is not necessarily restricted to one clause or sentence. Three main phenomena can be found.

(1) More details about an event are given in a later sentence, example:

(2) Two clauses connected by a temporal subordinate conjunction describe two events, which are placed in time relative to each other. Example: All 75 people on board the Aeroflot Airbus dies when it ploughed into a Siberian mountain in March 1994.

(3) Two events are located in time relative to each other, but the information is spread across more than one sentence. Example: The area is 55 miles from the site off Long Island where a TWA 747 crashed one week earlier.

This leads to another observation. Usually when performing IE tasks, events matching predefined templates are extracted from the text. Even if one wanted to locate only those `IE events' in time, it not clear that the temporal

extracted from the text. Even if one wanted to locate only those IE events in time, it not clear that the temporal information needed can only be (sufficiently) found in sentences containing IE events. This led to our more general approach of locating all events described in the text in time.

3 Temporal Ontology

So far we have been talking only about events and not about states. Events are seen here as complex structures including arguments of events (see below) or even subevents. Events normally take place over a period of time but can also be conceptually instantaneous. Events are not just seen as something being predicated over points or intervals, events themselves are taken as primitives. States, which occur less often, are conveyed by clauses taking generics as subjects (Altimeters are instruments for ...), that give information about class attributes (The plane which can carry four people...) or that describe intrinsic properties or temporally unbounded states (The area is 55 miles wide.), to name only a few examples. Such states, holding over the whole period of time in question, are of far less interest than events and are left out here for the sake of brevity. Our distinction between states and events is not a classical one, but one that is practical and adequate for this application and this genre.

As mentioned in section 2, we need to know the structure of events and how they are indicated in the text in order to extract them and order them in time. Events are indicated by finite verbs and also by certain kinds of nominalisations (like This morning's launch failure...). The latter have not been the target of our research do far and will not be discussed here further.

What comprises an event? This is dictated by the arguments the finite verb takes and the temporal information given. Logical subject, object and the finite verb itself are the core elements, a set than can easily be expanded should the application so require. The temporal information needed for this task can be found in the tense and aspect of the verb, in temporal prepositional and adverbial phrases as well as in temporal subordinate conjunctions.

4 Representation

Two levels or types of representation are needed, for the events themselves and for the temporal relations between them. The representation of events will reflect their structure, as described in section 3, and include the indicating finite verb, its logical subject and object and temporal expressions (amongst other arguments of the verb).

Vague temporal information and unclear temporal relations between events make it quite difficult to develop an adequate graphical representation (called Discourse Event Map or DEM). It would be ideal to be able to place all events on a timeline, but in attempting to do so we encounter two main problems. One is that not all events can be associated with a calendrical date and placing a `dateless' event at a certain distance from a `dated' event on a timeline creates the illusion that there is a certain period of time between those events, although this is not the case. A solution could be to not associate events with a calendrical date at all and to place all events relative to each other. But not all temporal relations are given in the text. If we know, for example, that events e1, e2 and e3 all happen before e4, then all three have to be placed somewhere to the left of e4, even though we do not know how they stand in relation to each other. So placing them can easily lead to false interpretations about their temporal relationships. To discuss more possibilities and their advantages and disadvantages is beyond the scope of this paper.

Another decision that has to be made regards the the different types of events, as described in section 2, namely report, experience and occurrence. As mentioned before, the main purpose of reporting events is to establish a source of information and our approach is to not place reporting events on the DEM but to associate the information source with the event itself. Occurrence events will certainly be inserted into the DEM but the case of experience events has still to be decided.

5 Why and what to Annotate?

Elaborating an annotation scheme helps to conceptualise what the temporal information conveyed is. In the

Elaborating an annotation scheme helps to conceptualise what the temporal information conveyed is. In the middle of the interrelated processes of developing a representation and finding out the structure of events, it keeps one close to the (real) text and thus forces one to look at what information is actually there. The actual annotation of a large number of texts gives information about the variety and distribution of the ways temporal information is expressed. An annotated corpus can serve as a basis for training and evaluating algorithms, it is a bench-mark, a target that should be matched.

The annotation scheme ties in with what was said about temporal information (section 2) and representation (section 4). Events and temporal relations between them are of interest and we need to annotate what indicates these in the text. Events are conveyed by clauses whereas the relations are conveyed by a variety of techniques, including narrative sequence, temporal prepositional phrases and temporal conjunctions. Hence we need clause level and inter clausal annotation. The clause level annotation reflects the structure of an event and includes the finite verb, its logical subject, object and indirect object, temporal expressions like prepositional and adverbial phrases and conjunctions, to name just the core. The inter clausal level annotation is not yet fully defined, but influencing factors are for example temporal anaphora, event anaphora and subeventness.

6 Future Work

We are about to finish phase one with finalising the annotation scheme and producing a small set of hand-annotated texts.

The next phase is to generalise from the data gathered in phase 1 and develop a theory of how temporal information is conveyed in newspaper articles. The answers to the following questions will help in identifying what factors make ordering events in time possible.

- *How important are aspectual classes in this particular genre? Can we do without analysing them?*
- *Does the order of the events in the text reflect the temporal order or are there many cases of events being reported in an inverted order?*
- *Do consecutive sentences usually describe different events or are there many occurrences of explanation or refinement?*
- *Is there much co-reference between events and how can we deal with it?*
- *How many surface cues are there and how much has to be inferred? How complex are the inferences?*
- *How much World Knowledge is needed? How far can we get without any?*

Based on the output of the first two phases we will develop heuristics, which when implemented would extract from the text the information necessary to order the events described in time. A possible extension would be to implement one or more of these heuristics and evaluate them against held-out annotated data. This would show that it is possible to automate the process of analysing annotated text and that we can establish the temporal relations between events.

Space and time do not allow discussion of the issues described here in more detail and many questions are still open. Other approaches, like Discourse Representation Theory (Kamp and Reyle (1993)) and ter Meulen's (ter Meulen (1995)) approach to representing time in natural language, are being analysed as to whether their integration into our work is possible and will be fruitful. Much work has been done on representing temporal information and is taken into account. One example is Moulin's Conceptual Graph approach (Moulin (1992)), which seems to be very appropriate to our work.

Amongst other ideas to be pursued is whether other relations (causal or subevents) than temporal ones would help in locating events in time.

References

Cowie, J. and W. Lehnert. 1996. Information extraction. Communications of the ACM, 39(1):80--91.

Gaizauskas, R. and Y. Wilks. 1998. Information Extraction: Beyond Document Retrieval. Journal of

Gaizauskas, R. and Y. Wilks. 1998. Information Extraction: Beyond Document Retrieval. Journal of Documentation, 54(1):70--105.

Kamp, H. and U. Reyle. 1993. From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language. Kluwer Academic.

Moulin, B. 1992. Conceptual-Graph Approach for the Representation of Temporal Information in Discourse. Knowledge-Based Sytems, 5:183--192.

ter Meulen, A. 1995. Representing Time in Natural Language. MIT Press. Cambridge, Massachussets.

Vendler, Z. 1967. Linguistics in Philosophy. Cornell University Press, Ithake, New York.