ISO-TimeML: An International Standard for Semantic Annotation

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary

Brandeis University, Korea Unviersity, Tilburg University, INRIA & HUB-IDSL
Waltham, MA USA, Seoul, South Korea, Tilburg, The Netherlands, Paris, France
jamesp@cs.brandeis.edu, klee@korea.ac.kr, harry.bunt@uvt.nl, laurent.romary@inria.fr

Abstract

In this paper, we present ISO-TimeML, a revised and interoperable version of the temporal markup language, TimeML. We describe the changes and enrichments made, while framing the effort in a more general methodology of semantic annotation. In particular, we assume a principled distinction between the annotation of an expression and the representation which that annotation denotes. This involves not only the specification of an annotation language for a particular phenomenon, but also the development of a meta-model that allows one to interpret the syntactic expressions of the specification semantically.

1. Introduction

In order to fully interpret a natural language utterance, it is necessary to understand the temporal information conveyed in the text, including all event and temporal expressions, as well as their anchoring and ordering relations. TimeML (Pustejovsky et al. 2005a, 2005b) (www.timeml.org) is an annotation scheme specifically designed for the markup of events, times, and their temporal relations in text. The TimeML scheme annotates all expressions having temporal import, broadly categorized as temporal expressions and eventualities (situations, events, states, and activities). The EVENT tag specifies various attributes, including the class of event, tense, grammatical aspect, polarity (negative or positive), any modal operators which govern the event being tagged, and cardinality of the event if its mentioned more than once. Likewise, time expressions are flagged and their values normalized, based on TIMEX3, an extension of the ACE (2004) (tern.mitre.org) TIMEX2 annotation scheme. Temporal expressions and events participate in temporal relationships (e.g., “before”, “simultaneous”), subordinating relationships (e.g., “intensional”, “factive”), and aspectual relationships (e.g., “initiates”, “continues”). TimeML provides an additional expressive capability of capturing and representing the complexities of these relationships.

Within the context of the ISO TC 37/SC4 Semantic Annotation Framework (SemAF), TimeML has been adopted as the foundation for a a formal specification language called *ISO-TimeML*, for temporal information markup in natural language.

Unlike prior event annotation schemes, ISO-TimeML’s somewhat unique definition of an event does not limit the standard’s applicability to other natural language genres. An ISO-TimeML event is simply something that can be related to another event or temporal expression using an ISO-TimeML relationship — thus an ISO-TimeML-compliant representation can be adapted (derived) from the full standard specification, appropriate to different genres, styles, domain, and applications. Future work will involve applying the standard in such different contexts, and formulating guidelines and principles for appropriate use of ISO-TimeML in a variety of language engineering environments.

2. Semantic Annotation and Interoperability

Following ISO CD 24612 (*Language resource management - Linguistic annotation framework*) and Ide & et al. (2003), we assume a fundamental distinction between the concepts of *annotation* and *representation*). The term ‘annotation’ is used to refer to the process of adding information to segments of language data, or to refer to that information itself. This notion is independent of the format in which this information is represented. The term ‘representation’ is used to refer to the format in which an annotation is rendered, for instance in XML, independent of its content. According to the proposed internation standard (LAF), *annotations* are the proper level of standardization, not representations. Hence, ISO-TimeML defines a markup language for annotating documents with information about time and events at the level of annotations.

The distinction between annotations and representations is reflected in the specification of ISO-TimeML, which makes a distinction between an *abstract syntax* as well as a *concrete syntax*. The abstract syntax specifies the elements making up the information in annotations, and how these elements may be combined to form complex annotation structures; these combinations are defined as set-theoretical structures, independent of any particular representation format. There are infinitely many ways in which these structures can be represented. In line with other ISO TC 37/SC 4 proposals, an XML-based concrete syntax is defined for representing ISO-TimeML annotations. Any other representation that is a faithful rendering of the abstract syntax of ISO-TimeML can readily be converted into this XML representation and vice versa. ISO-TimeML has a semantics associated with its abstract syntax, which defines the meanings of ISO-TimeML annotation structures. The fact that this semantics is associated with the *abstract* syntax, rather than with a particular concrete syntax, explains why all concrete representations of ISO-TimeML annotations are semantically equivalent.

3. ISO-TimeML

3.1. Introduction

When talking about the semantics of events and temporal entities generally, there are three things that need to be accounted for. Assuming both events and times are interval-like, these are:

(1) a. The position of the interval relative to others (ORDER):
b. The size of the interval (MEASURE):;
c. The number of intervals (QUANTITY):.

Currently, the ISO-TimeML framework adequately handled positional information (order), captured generally by Allen-like interval relations.

(2) a. John taught on Tuesday.
b. John taught before Mary arrived.

The TLINK mechanism provides temporal ordering and anchoring of event predicates interpreted as intervals. We introduce a function, τ, which interprets an event as an interval.

(3) a. teach= e_1, tuesday= t_2
b. $\exists e_1 \exists t_2 [teach(e_1) \wedge tuesday(t_2) \wedge \tau(e_1) \subseteq t_2]$

The anchoring relation of the teaching event to the TIMEX3 *Tuesday* is represented by a TLINK relation, as shown below:

```
<TLINK eventID="e1" relatedToTime="t2"
  signalID="s1" relType="IS_INCLUDED"/>
```

In (2a), the event e_1 is anchored within the temporal expression t_2. Similarly, in (2b), e_1 is ordered before the event e_2. TLINK handles both adequately.

(4) a. teach= e_1, arrive = e_2
b. $\exists e_1 \exists e_2 [teach(e_1) \wedge arrive(e_2) \wedge \tau(e_1) < \tau(e_2)]$

```
<TLINK eventID="e1" relatedToEvent="e2"
  signalID="s2" relType="BEFORE"/>
```

In the discussion that follows, we will indicate how measure and quantity can be represented within ISO-TimeML. First, however, we turn to the distinction between an abstract syntax and an annotation specification.

3.2. Properties

The specification of ISO-TimeML consists of three components, mirroring the LAF distinction of abstract annotations and concrete representations: (1) an abstract syntax of ISO-TimeML annotations; (2) a format for representing these annotations in XML (a concrete syntax); and (3) a semantics of ISO-TimeML.

The abstract syntax of ISO-TimeML defines the set-theoretical structures that constitute the information about time and events that may be contained in annotations. The definition of the abstract syntax consists of two parts:

(5) a. a specification of the elements from which these structures are built up, called a 'conceptual inventory'; and
b. a set of syntax rules which describe the possible combinations of these elements.

What these combinations mean, i.e. which information they capture, is specified by the semantics associated with the abstract syntax.

A concrete syntax consists of the specification of names for the various sets forming the conceptual vocabulary, plus a listing of specific named elements of these sets, and a specification of how to represent ISO-TimeML annotation structures defined by the syntax rules of the abstract syntax mentioned above. A particular XML-based syntax for temporal annotation has been defined in the TimeML effort (Pustejovsky et al., 2003; 2007) and is largely adopted by ISO-TimeML, modulo the stand-off character of ISO-TimeML annotations, illustrated in the next section.

The final component of ISO-TimeML consists of a specified semantic interpretation of the XML representations provided by the concrete syntax. There are currently two semantic fragments: one using Interval Temporal Logic, a first-order logic for reasoning about time; the other one uses a DRT-like, event-based semantics for the abstract ISO-TimeML syntax.

3.3. Standoff Annotation

There are several changes to TimeML, introduced by the ISO-TimeML specification. Perhaps the most significant structural change is the move from in-line to stand-off annotation. This is in accordance with the general methodology to create interoperable annotation languages that do not modify the text being annotated.

ISO-TimeML conforms to the following three ISO standards: ISO 24610-1:2006 FSR (jointly developed with the TEI Consortium), ISO DIS 24611 MAF, and ISO DIS 24612 LAF. A proper management of stand-off annotation requires dealing with identifiers (xml:id) and pointers in conformance to most recent XML technologies and articulate these mechanisms with the XML elements provided by the other ISO standards for linguistic annotation. For instance, MAF species how a text is segmented into tokens and how these tokens are represented in XML (element ⟨token⟩). In turn, ISO-TimeML annotations may point to such tokens as illustrated in the example sentence below.

(6) Mia visited Seoul to look me up yesterday.

This data is now segmented into word forms, as follows:

(7) TOKENIZATION:

```
<maf xmlns:"http://www.iso.org/maf">
  <seg type="token" xml:id="token1">Mia</seg>
   <seg type="token" xml:id="token2">visited
    </seg>
   <seg type="token" xml:id="token3">Seoul
    </seg>
   <seg type="token" xml:id="token4">to</seg>
   <seg type="token" xml:id="token5">look
    </seg>
   <seg type="token" xml:id="token6">me</seg>
   <seg type="token" xml:id="token7">up</seg>
   <seg type="token" xml:id="token8">yesterday
    </seg>
   <pc>.</pc>
       </maf>
```

As is specied in LAF, this inline segmentation may also be replaced by an offline identification of tokens through spans based, for instance, on character shifts: e.g., <seg ... form="Mia"/> is replaced by <seg ... from 0 to 3/>. Note here that the complex verb "looked ... up" is treated as a single word segment, consisting of two discontinuous tokens, "looked" and "up". On the basis of the segmented text, ISO-TimeML can now annotate the given text in a standoff manner, as represented below:

(8) STANDOFF ANNOTATION:

```
<isoTimeML
    xmlns:"http://www.iso.org./isoTimeML">
<TIMEX3 xml:id="t0" type="DATE"
   value="2009-10-20"
functionInDocument="CREATION_TIME"/>
<EVENT xml:id="e1" target="#token2"
   class="OCCURRENCE" tense="PAST"/>
<EVENT xml:id="e2"
   target="#range(#token5,#token7)"
   class="OCCURRENCE"
   tense="NONE" vForm="INFINITIVE"/>
<TIMEX3 xml:id="t1" type="DATE"
   value="2009-10-19"/>
<TLINK target="#range(#e1,#t0)"
   relType="BEFORE"/>
<TLINK target="#range(#e1,#t1)"
   relType="ON_OR_BEFORE"/>
<TLINK target="#range(#e2,#t1)"
   relType="IS_INCLUDED"/>
</isoTimeML>
<tei-isoFSR xmlns:
   "http://www.iso.org./tei-isoFSR">
<fs xml:id="t0">
 <f name="Type" value="2009-10-20"/>
</fs>
</tei-isoFSR>
```

Note that the temporal expression "yesterday" is interpreted as referring to the date "2009-10-19" on the assumption that the creation time for the text is 2009-10-20. Further, the event time of Mia's visiting Seoul is understood as taking place in the past, "yesterday" or earlier.

3.4. Measuring Events

Another significant change introduced by ISO-TimeML is in the treatment of temporal durations. The TimeML `DURATION` type is based on the TIMEX2 treatment of durations, which is interpreted as a contiguous temporal interval. Consider, for example the durative events below:

(9) a. John slept for 2 hours.
b. a three-day vacation

It was assumed that the interpretation in such event readings situated the event completely within a specific and named interval. For this reason, it was thought adequate to treat such cases with a TLINK relation; namely, the SIMULTANEOUS relType, as shown below:

```
<EVENT id="e1"   pred="SLEEP"/>
<TIMEX3 id="t1" type="DURATION" value="P2H"/>
<TLINK eventID="e1" relatedToTime="t1"
   relType="SIMULTANEOUS" />
```

This is inadequate, however, on two accounts. First, it is descriptively incomplete, in that this is not always the desired interpretation for a duration phrase. For example, consider the sentence below.

(10) John taught for three hours on Tuesday.

In this case, the interpretation is ambiguous. Did John teach without stopping for three hours sometime during the day or did he teach for an hour, take a break, teach again, and so forth? Either interpretation is possible, so it would be incorrect to commit the interpretation to the contiguous (convex hull) interval reading. The second problem with this treatment is that it fails to characterize the temporal expression as a measurement of the event, as expressed in the abstract syntax for the language (as mentioned above).
To deal with this problem, ISO-TimeML reifies the role that certain expressions in the language play in measuring over a domain; that is, a new link is introduced for measuring out events, called MLINK, with the inherent relation type of `MEASURE`. A temporal expression such as *3 hours* is expressed as a TIMEX3 of type `DURATION`, with the interpretation of a "time amount" (Bunt and Pustejovsky, 2010). This can be used in either non-contiguous or contiguous interpretations. A measure is equal to the sum of all times that add up the desired period of time (ex. *P3H* = $\forall i[\Sigma i = P3H]$). This reflects more transparently the abstract syntax specified within ISO-TimeML, where the distinction is made between an interval and the measure of an interval. The annotation fragment is illustrated below.

```
<EVENT id="e1"   pred="TEACH"/>
<TIMEX3 id="t2" type="DURATION" value="P3H"/>
<MLINK eventID="e1" relatedToTime="t2" />
```

Formally, we assume that a measure function, μ, such as introduced in Bunt (1985), can be used interpret this relation, as represented as below. The details of this proposal are more fully presented in Bunt and Pustejovsky (2010).

(11) a. teach= e_1, tuesday= t_2, m= 1 hour
b. $\exists e_1 \exists t_2[teach(e_1) \wedge \mu(\tau(e_1)) = v \wedge v = 1\_hour \wedge tuesday(t_2) \wedge \tau(e_1) \subseteq t_2]$

3.5. Counting Events

Anchoring and ordering relations in ISO-TimeML intrinsically quantify the event participating in the relation. But as has been pointed out, there is no clear way to embed an event within a temporal quantifier expression (Pratt-Hartmann, 2007, Bunt and Pustejovsky, 2010). Consider again the sentence mentioned above:

(12) John taught on Tuesday.

Within TimeML, the translation between the distinct elements are given below:

(13) a. EVENT tag introduces a quantified event expression $\Longrightarrow \exists e_1[teach(e_1)]$;
b. TIMEX3 tag introduces the temporal expression $\Longrightarrow \exists t_2[tuesday(t_2)]$;
c. TLINK introduces the ordering relation $\Longrightarrow \lambda y \lambda x[\tau(x) \subseteq y]$.

Assuming approaches to the semantics of TimeML as taken in Pratt-Hartmann (2007) and Katz (2007), the resulting semantics of the sentence is a conjunction of these relations:

(14) b. $\exists e_1 \exists t_2[teach(e_1) \wedge tuesday(t_2) \wedge \tau(e_1) \subseteq t_2]$

Now, what happens if we have a quantified expression? The TimeML representation is not really very clear in how it interprets sentences such as (15) below.

(15) John taught every Monday in November.

As before, the translation between the distinct elements in this sentence would be given as follows:

(16) a. EVENT tag introduces a quantified event expression $\Longrightarrow \exists e_1[teach(e_1)]$;
b. TIMEX3 tag introduces the temporal expression $\Longrightarrow \exists t_1[monday(t_1)]$;
c. TIMEX3 tag introduces the temporal expression $\Longrightarrow \exists t_2[november(t_2)]$;
d. TLINK introduces an ordering relation $\Longrightarrow \lambda y \lambda x[\tau(x) \subseteq y]$;

But this does not give us the right scope and interpretation. This results in an interpretation where one event of teaching occurs over every Monday in November. Bunt and Pustejovsky (2010) explore the option of explicitly marking the distributive property (Bunt, 1985) of the quantification in the annotation directly. This would allow us to then scope the temporal expression over the event predicate, as illustrated below:

(17) b. $\forall t_1 \exists e_1 \exists t_2[(Monday(t_1) \wedge November(t_2) \wedge t_1 \subseteq t_2) \rightarrow (teach(e_1) \wedge \tau(e_1) \subseteq t_1)]$

The details of how quantification should best be expressed in the annotation specification are still being worked out; the abstract syntax of ISO-TimeML, however, does allow us to express such scope relations in the syntax directly.

4. Concluding Remarks

The primary purpose of constructing ISO-TimeML in ISO 24617-1 SemAF-Time is to produce sustainable language resources with annotation for practical applications. Any system that utilizes and processes such resources is expected to be robust and sustainable independent of syntactic well-formedness. Such sustainability can easily be surmised because ISO-TimeML only relies on the proper tokenization of text in compliance of MAF without requiring syntactic information in general.

As specified in ISO DIS 24617-1 SemAF-Time, ISO-TimeML is still being revised at this stage of writing this abstract, but is expected to be published as an international standard by ISO. It has been approved by ISO/TC 37/SC 4 for its submission to ISO/CS for publication. Some issues that relate to quantification and measurement, as mentioned above, still need to be fully implemented within ISO-TimeML. There is, however, general agreement on the approach adopted towards these issues.

References

Bunt, C. Harry (1985), *Mass Terms and Model-theoretic Semantics*, Cambridge University Press, Cambridge.

Bunt, Harry and James Pustejovsky (2010) “Annotating Temporal and Event Quantification”, in *Proceedings of 5th ISA Workshop*, Hong Kong, Jan. 15-17, 2010.

Burnard, Lou and Syd Bauman (2007), TEI P5:- Guidelines for Electronic Text Encoding and Interchange, TEI Consortium.

ISO-24610-1:2006 Language resource management: Feature structures, Part 1: Feature structure representation (FSR).

ISO-DIS 24611:2009 Language resource management: Morpho-syntactic annotation framework (MAF).

ISO-DIS 24612:2009 Language resource management: Linguistic annotation framework (LAF).

ISO-DIS 24617-2:2009 Language resource management: Semantic annotation framework Part 1: Time and events (SemAF-Time).

Pratt-Hartmann, Ian (2007) “From TimeML to TPL”, in Schilder, Frank, Graham Katz, and James Pustejovsky (eds.), *Annotating, Extracting and Reasoning About Time and Events*, Lecture Notes in Computer Science 4795, Springer, Berlin/Heidelberg.

Katz, Graham (2007) “Towards a Denotational Semantics for TimeML” in Schilder, Frank, Graham Katz, and James Pustejovsky (eds.), *Annotating, Extracting and Reasoning About Time and Events*, Lecture Notes in Computer Science 4795, Springer, Berlin/Heidelberg.

Mani, Interjeet, James Pustejovsky, and Rob Gaizauskas (eds.) (2005), The Language of: Time: A Reader, Oxford University Press, Oxford.

Pustejovsky, James, Robert Ingria, Roser Sauri, Jose Gastano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Christopher Habel (2005), “The Specification Language TimeML”, in Mani et al. (2005a).

Pustejovsky, James, Robert Knippen, Jessica Littman, and Roser Saurí”, (2005b) “Temporal and Event Information in Natural Language Text”, *Language Resources and Evaluation* 39, pg. 123-164, Springer.