Multilinguality in Temporal Annotation: A Case of Korean

Kiyong Lee

Korea University
Department of Linguistics, Seoul 136-701, South Korea
klee@korea.ac.kr

Abstract. The aim of this paper is to apply TimeML, an annotation scheme for events and temporal expressions, to the annotation of Korean in an attempt to test its multilingual extendability. TimeML has been well validated by the successful annotation of a corpus of 186 news articles and some other documents in English. One of its remaining tasks, however, is multilingual extension. This paper aims at contributing to this task and also at promoting TimeML as an international standard.

1 Aim and Scope

At the time when this paper is orally presented at PACLIC20, Wuhan in China, a meeting will have taken place initiating a standardization effort, as is discussed in [1], on temporal annotation at Brandeis University in October, 2006. The aim of this paper is to make some contributions to this effort by focusing on the task of multilingual extension in temporal annotation and, in particular by extending the application of TimeML, a well-established temporal annotation scheme, to Korean. The scope of the present discussion, however, remains at the purely descriptive level. To gain some explanatory validity, it will be supported with relevant linguistic observations from Korean.

2 Pervasiveness of Temporal Information

Humans live on temporal information and carry on their daily life accordingly. Many people get up early in the morning with the sun rising. They eat meals at fixed time of the day and spend a certain amount of time working during the day. While working, they converse a lot with each other, simply chatting on trivial subjects or seriously exchanging information about daily business transactions. Their daily ordinary use of language is thus full of expressions related to time and events.

One of the routine activities is reading newspapers or browsing a web site. Consider the following headline of a news item copied from www.joins.com, which is claimed to be the first internet newspaper in Asia, run by Joong-Ang Ilbo in Korea:

(1) Sample Fragment
`9.11 theyre 5cwunyen maca CNN, tangsi sayngcwungyey caypangsong`
9/11 terror 5th anniversary meet CNN, that time live-relay rebroadcast

This headline roughly means that, at the 5th anniversary of the 9/11 terrorists' attacks, CNN plans to rebroadcast the on-site report that was relayed live at that time. The headline is then accompanied by three short paragraphs elaborating this rebroadcast plan of CNN.

Every word here except for the acronym CNN either refers to a time or to an event.

(2) Time or Event Referring Expressions
1. 9.11: 9/11
2. theyre: terror
3. 5cwunyen: 5th anniversary
4. maca: meet
5. CNN
6. tangsi: that time
7. sayngcwunggyey: relaying live coverage
8. caypangsong: rebroadcast

(1), (3) and (6) are all temporal expressions, referring to the date 9/11 (September 11), 5th anniversary, and that time, namely 9/11, respectively. Each of the expressions (4), (7) and (8), on the other hand, refers to an event of some sort, which can be a state like the occasion of remembering the 5th anniversary or an activity or occurrence like relaying live coverage or rebroadcasting. This simple example illustrates how the small world that surrounds us is filled with expressions referring to time and events.

3 TimeML

TimeML was developed over the past four or so years and its most recent annotation guidelines version 1.2.1 is dated January 2006. Temporal annotation creates precious language resources out of raw corpora for language technology applications such as information extraction from text and summarization as well as question answering.[1]

As specified in [2] and [3], TimeML annotates for temporal relations with four TimeML tags <EVENT>, <TIMEX3>, <SIGNAL>, <MAKEINSTANCE> and also with three link tags <TLINK>, <SLINK>, and <ALINK>. It practices principled separation between events and temporal expressions. And it develops a typology of temporal links, while looking forward to formalization by a temporal ontology without adhering to any particular ontology.

[1] TimeML, for instance, was developed as part of the TERQAS (Temporal Expression Recognition for Question Answering Systems) colloquium, supported by the ARDA-funded project program AQUAINT, to improve the performance of question answering systems.

While TIMEX3, a component of TimeML, annotates time referring expressions, TimeML in general concentrates on annotating events and temporal relations over times and events. Events comprise everything that happens or is situated in time. In TimeML, events are subtyped into: (1) occurrence ('attach', 'crash'), (2) state ('love', 'kidnapped'), (3) reporting ('report', 'say'), (4) i-action ('try', 'promise'), (5) i-state ('believe', 'want'), (6) aspectual ('begin', 'stop'), and (7) perceptual ('see', 'hear'). As stated in [4: 492], this classification is justified on some distinctive evidence of temporal inferences.

Here is a sample annotation, taken from [5]:

(3) Sample TimeML Annotation

```
John left 2 days before the attack.

John
<EVENT eid="e1" class="OCCURRENCE">
left </EVENT>
<MAKEINSTANCE eiid="ei1" eventID=e1 tense="PAST" aspect="NONE"/>
<TIMEX3 tid="t1" type="DURATION" value="P2D" temporalFunction="false">
2 days </TIMEX3>
<SIGNAL sid="s1">
before </SIGNAL>
the
<EVENT eid="e2" class="OCCURRENCE">
attack </EVENT>
<MAKEINSTANCE eiid="ei2" eventID="e2" tense="NONE"
aspect="NONE"/>
```

The two event instances, `ei1` and `ei2`, are linked by the temporal relation BEFORE as in:

(4) Temporally Linking Events

```
<TLINK eventInstanceID="eil" signalID="s1"
relatedToEventInstance="ei2" relType="BEFORE">
```

This relates the event of John's leaving to the event of the attack such that the former occurs BEFORE the latter.

4 Korean Extension

Perhaps Jang and others' work in [6] was the first to annotate temporal information in Korean text. This work adopted TIMEX2 as its annotation scheme with its coverage restricted to time referring expressions only, as introduced in [7]. The work presented in this paper, however, adopts the most recent version 1.2.1 of TimeML as its annotation scheme with its annotation scope ranging from time to event expressions as well as their temporal relations. This section shows how TimeML applies to Korean fragments.

4.1 Annotating a News Headline

Newspaper headlines or segmented phrases are good samples for temporal annotation. Unlike formal semantics systems, annotanda need not be full sentences or phrases, but any parts of a text that carry relevant information. Especially for a language like Korean which freely allows the deletion of non-verbal parts like Subjects, Objects and other Complements in sentences, the mechanism of annotation is an excellent tool for doing semantics. Newspaper headlines without any full sentential structure are thus good samples for annotation. Here is the TimeML-compliant annotation of a Korean newspaper headline for illustration:

(5) Korean News Headline Annotated

```
1. Joong-Ang Ilbo
2. <TIMEX3 tid="t0" type="DATE" value="2006-08-27"
   functionInDocument="CREATION_TIME">
   2006.08.27 </TIMEX3>
3. <TIMEX3 tid="t1" type="DATE" value="2001-09-11"
   temporalFunction="TRUE" anchorTimeID="t2"/>
   9.11 (9/11, September 11)
4. <EVENT eid="e1" class="OCCURRENCE">
   theyre (terrorists' attacks) </EVENT>
   <MAKEINSTANCE eventInstanceID="ei1" eventID="e1" pos="NOUN"
   tense="NONE" aspect="NONE"/>
5. <TIMEX3 tid="t2" type="DATE" value="2006-09-11"
   temporalFunction="TRUE" anchorTimeID="t0">
   5cwynyen (5th anniversary) </TIMEX3>
6. <EVENT eid="e2" class="OCCURRENCE">
   maca (meet) </EVENT>
   <MAKEINSTANCE eventInstanceID="ei2" eventID="e2" pos="NOUN"
   tense="NONE" aspect="NONE"/>
7. CNN
8. <TIMEX3 tid="t3" type="DATE" value="2001-09-11"
   temporalFunction="TRUE" anchorTimeID="t1">
   tangsi (that time) </TIMEX3>
9. <EVENT eid="e3" class="OCCURRENCE">
   sayngcwunggyey (live news coverage relay)</EVENT>
   <MAKEINSTANCE eventInstanceID="ei3" eventID="e3" pos="NOUN"
   tense="NONE" aspect="NONE"/>
10. <EVENT eid="e4" class="OCCURRENCE">
    caypangsong (rebroadcast) </EVENT>
    <MAKEINSTANCE eventInstanceID="ei4" eventID="e4" pos="NOUN"
    tense="NONE" aspect="NONE"/>
11. <TLINK tid="t1" relatedToTime="t2" relType="BEFORE"/>
12. <TLINK tid="t3" relatedToTime="t1" relType="IDENTITY"/>
13. <TLINK eventInstanceID="ei1" relatedToTime="t1"
    relType="IS_INCLUDED"/>
14. <TLINK eventInstanceID="ei2" relatedToEventInstance="ei1"
    relatedToTime="t2" relType="IS_INCLUDED"/>
```

```
15. <TLINK eventInstanceID="ei3" relatedToTime="t1"
    relType="IS_INCLUDED"/>
16. <TLINK eventInstanceID="ei4" relatedToTime="t2"
    relType="IS_INCLUDED"/>
```

Here, `9.11` is analyzed as a date referring to 9/11 (September 9) and thus its representation can be normalized as `"XXXX-09-11"`.[2] The unknown year `XXXX` is supplied later by the value of the 5th anniversary, `"2006-09-11"`, where the information about the year 2006 is only relevant. Note also that the expression 'anniversary' is understood as referring to a date, 'a day which is an exact year or number of years after something has happened'. Like any other festivals or days of remembrance, an anniversary can also be treated as an event, since it can be celebrated. But it seems to refer both to a date and to an associated event (instance). That is why `relatedToEventInstance="ei1"` is inserted for the `TLINK` of `eventInstance="ei2"`.[3]

4.2 Discussions

Word Segmentation Like Japanese, Korean is an agglutinative language. Both verbal and nominal stems can each have a long sequence of verbal endings or nominal particles. The following would be simple examples:

(6) Word Segmentation

```
1. sahultongan cassta (for 3 days slept)
   3 day for sleep-PAST-DECL
2. nonmwun makam 9wol 15ilkkaci (paper due by September 15)
   paper due 9 month 15 day by
```

As pointed out in [6], the segmentation of word forms into stems and endings, particles or suffixes may be necessary to capture and annotate the distinct meaning of each temporal expression, for each of them carries its own meaning. According to TimeML, each of these temporal expressions is easily annotated as shown below:

(7) Time Phrases Annotated

```
a. sahultongan cassta (for 3 days slept)
1. <TIMEX3 tid="t1" type="DURATION" value="P3D">
   sahul (3 days) </TIMEX3>
2. <SIGNAL sid="s1">
   tongan (for) </SIGNAL>
3. <EVENT eid="e1" class="STATE">
   cassta (slept) </EVENT>
4. <MAKEINSTANCE eiid="ei1" eventID="e1" pos="VERB"
   tense="PAST" aspect="NONE"/>
```

[2] Dates, times, and durations should be represented according to the W3C-defined profile of ISO 8601.

[3] TimeML 1.2.1 should be relaxed to allow the simultaneous occurrence of both `relatedToEventInstance` and `relatedToTime` simultaneously for `TLINK`.

```
   5. <TLINK eventInstanceID="ei1" signalID="s1"
      relatedToTime="t1" relType="DURING" />
  b. nonmun makam 9wol 15ilkkaci (paper due by 9/15)
   1. nonmun (paper)
   2. <EVENT eid="e2" class="OCCURRENCE">
      makam (due) </EVENT>
   3. <MAKEINSTANCE eiid="ei2" eventID="e2" pos="NOUN" />
   4. <TIMEX3 tid="t2" type="DATE" value="XXXX-09-15">
      9wol 15il ( 9/15) </TIMEX3>
   5. <SIGNAL sid="s2">
      kkaci (by or till) </SIGNAL>
   6. <TLINK eventInstanceID="ei2" signalID="s1" relatedToTime="t2"
      relType="IBEFORE" />
```

Note here that the value "IBEFORE" of relType is understood to be "Inclusive_BEFORE", namely texttt"ON_OR_BEFORE".

The Korean lexicon contains a very large number of words or phrases derived from Chinese. In this case, finer-grained segmentation is required. For example, 'sahul-tongan', 'sahul-kan' and 'samil-kan' have the same meaning and should be annotated identically, where the suffixes 'tongan' and 'kan' are duration signals.

Marking Tense Korean is a verb-final language. This means that, unlike the SVO language like English, the sentence pattern conforms to SOV in an ordinary situation. In a coordinate structure there can be a series of verbs as in the following:

(8) Tense Marking

```
   1. eceyspam nanun (Last night I)
      last night I-TOP
   2. shawolul hako (took a shower and)
      shower-ACC do
   3. wainul han can masiko (drank a glass of wine and)
      wine-ACC one glass drink
   4. cam cassta (slept)
      sleep sleep-PAST-DECL
```

But, unlike English, tense is marked on the sentence-final verb only. The question now is how to tell that all these verbs denote the events which occurred in the past and also sequentially one after the other.

(9) Linking Events Denoted by Verbs with No Tense Marking

```
   1. <TIMEX3 tid="t0" type="DATE" value="2006-09-11"
      functionInDocument="CREATION_TIME" />
   2. <TIMEX3 tid="t1" type="TIME" value="2006-09-10TNT"
      temporalFunction="TRUE" anchorTime="t0"
      comment:"TNT stands for night">
      ecyespam (last night) </TIMEX3>
   3. nanun (I)
```

```
4. <EVENT eid="e1" class="OCCURRENCE">
   shawolul hako (do/take shower)
   <MAKEINSTANCE eiid="ei1" eventID="e1" pos="VERB"
   tense="NONE" aspect="NONE" />
5. wine han canul (wine one glass)
6. <EVENT eid="e2" class="OCCURRENCE">
   masiko (drink)
   <MAKEINSTANCE eiid="ei2" eventID="e2" pos="VERB"
   tense="NONE" aspect="NONE" />
7. cam (sleep)
8. <EVENT eid="e3" class="OCCURRENCE">
   cassta (slept)
   <MAKEINSTANCE eiid="ei3" eventID="e3" pos="VERB"
   tense="PAST" aspect="NONE" />
9. <TLINK eventInstanceID="ei3" relatedToTime="t1"
   relType="IS_INCLUDED" />
```

According to this annotation, `TLINK` anchors only the sleeping event instance `ei3` to last night, `t1, 2006-09-10TNT`, while the other two event instances are not assigned `tense value` explicitly by textttTLINK. Nevertheless, all the event instances denoted by these verbs are understood to have occurred in the past because of their temporal ordering linked by `TLINK`:

(10) Temporal Ordering among Events

```
1. <TLINK eventInstanceID="ei1" relatedToEventInstance="ei2"
   relType="BEFORE" />
2. <TLINK eventInstanceID="ei2" relatedToEventInstance="ei3"
   relType="BEFORE" />
```

Assuming the transitivity of `BEFORE`, `ei1` occurred before `ei2`, and both `ei1` and ei2 before `ei3`. By inference, all occurred last night.

5 Concluding Remarks

With its detailed annotation guidelines, TimeML 1.2.1 has nicely applied, without any modification, to the temporal annotation of Korean samples for testing. TimeML has been found most suitable to the temporal semantic annotation of Korean at least for two reasons. First, because of its focus on various types of events and their temporal relations, TimeML is judged to be an excellent annotation scheme for a language like Korean whose sentences often consist of verbs or event-denoting nominals only, without any Subject, Objects or Complements. Secondly, because of its linking mechanisms, TimeML can annotate the temporal ordering and relations over events and event instances of verbs and event-denoting expressions in a language like Korean where their tense and aspect are often underspecified, especially in coordinate structures with a series of verbs. `TLINK`, for instance, annotates the temporal ordering of all these verbs in a sequence and makes it possible to infer the temporal anchoring of the event instances denoted by the tenseless verbs.

TimeML is a highly expressive specification language for events and temporal expressions. Along with the development of a TimeML-compliant automatic ancillary annotator for Korean, a future task is to annotate a larger portion of Korean, a corpus containing more complex expressions for testing the full power of TimeML.

Acknowledgments. This is an unreviewed paper without going through the scrutinizing process of review. For inviting me to present a paper at PACLIC20, I would like to thank Maosong Sun, Tingting He, Po Hu and other Organizing, Programming, and Local Committee members, and also Chu-Ren Huang and other Steering Committee members of PACLIC. I owe many thanks to James Pustejovsky and his PhD student Seohyun Im at Brandeis University for helping me annotate the sample fragments in this paper and also for useful discussions, and Bran Boguraev and Suk-Jin Chang for making detailed constructive comments on the draft at various stages of its development.

References

1. Lee, K., Pustejovsky, J., Boguraev, B.: Towards an International Standard for Annotating Temporal Information. Proceedings of the International Conference on Terminology, Standardization and Technology Transfer (TSTT'2006 Beijing) (2006) 25-35
2. Boguraev, B., Castanõ, J., Gaizauskas, R., Ingria, B., Katz, G., Knippen, B., Littman, J., Mani, I., Pustejovsky, J., Sanfilippo, A., Setzer, A., Sauri, R., Stubbs, A., Sundheim, B., Symonenko, S., Verhagen, M.: TimeML 1.2.1: A Formal Specification Language for Events and Temporal Expressions. Available from http://www.timeml.org/site/publications (2005)
3. Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., Pustejovsky, J.: TimeML Annotation Guidelines Version 1.2.1. Available from http://www.timeml.org/site/publications (2006)
4. Mani, I., Pustejovsky, J., Gaizauskas, R. (eds.): The Language of Time: A Reader. Oxford University Press Oxford (2005)
5. Pustejovsky, J., Ingria, R., Sauri, R., Castanõ, J., Littman, J., Gaizauskas, R., Setzer, A., Kats, G., Mani, I.: The Specification Language TimeML. In: Mani, I., Pustejovsky, J., Gaizauskas, R. (eds.): The Language of Time: A Reader. Oxford University Press Oxford (2005)
6. Jang, S.B., Baldwin, J., Mani, I.: Automatic TIMEX2 Tagging of Korean News. ACM Transactions on Asian Language Information Processing, Vol. 3.1. (2004) 51-65
7. Ferro, L., Gerber, L., Mani, I., Sundheim, B., Wilson, G.: Instruction Manual for the Annotation of Temporal Expressions. MITRE, Washington C3 Center, McLean, Virginia (2002)