Korean TimeBank Including Relative Temporal Information

Chae-Gyun Lim[†], **Young-Seob Jeong**[§], **Ho-Jin Choi**[†]
[†]School of Computing, KAIST
291 Daehak-ro, Yuseong-gu, Daejeon 34141, South Korea
{rayote, hojinc}@kaist.ac.kr
[§]Department of BigData Engineering, Soonchunhyang University
22 Soonchunhyang-ro, Asan-si, Chungcheongnam-do 31538, South Korea
bytecell@sch.ac.kr

Abstract

Since most documents have temporal information that can be a basis for understanding the context, the importance of temporal information extraction researches is steadily growing. Although various attempts have been made to extract temporal information from researchers internationally, it is difficult to apply them to other languages because they are usually targeted at specific languages such as English. Several annotation languages and datasets had been proposed for the studies of temporal information extraction on Korean documents, however, the representation of relative temporal information is not enough to maintain it explicitly. In this paper, we propose a concept of relative temporal information and supplement a Korean annotation language to represent new relative expressions, and extend an annotated dataset, Korean TimeBank, through the revised language. We expect that it is possible to utilize potential features from the Korean corpus by clearly annotating relative temporal relationships and to use well-refined Korean TimeBank in future studies.

Keywords: Relative temporal information, Korean TimeML, Relative temporal relationship, Korean TimeBank

1. Introduction

Temporal information extraction is one of the important research fields in natural language processing, and it is necessary to understand the temporal context in real-world applications such like question answering or conversation system providing a good quality of service. As an international activity related to temporal information extraction research, there is a shared task called TempEval which is part of SemEval (UzZaman et al., 2012). According to the description of TempEval, the temporal information extraction consists of three separate processes as follows—*timex3*, *event*, and *tlink* extraction. TempEval has provided annotated datasets for performance comparison by adopting TimeML (Pustejovsky et al., 2003), which is one of well-known markup languages for annotating temporal information. However, since the dataset is basically constructed underlying English documents, it is not applicable to the studies of temporal information extraction in other languages. Preceding the construction of a corpus consisting of the documents in a particular language is very important because the linguistic characteristics inherent in the language have a significant impact to discover temporal relationships (Jeong, 2016).

For extracting temporal information in Korean, there are several previous researches. The Korean TimeML (Im et al., 2009), which is adopt morpheme-level stand-off annotation scheme and addressed some language-specific issues on Korean, was proposed. In 2015, the revised version of Korean TimeML was proposed to overcome several limitations such as applying the lunar calendar and character-level annotation scheme (Jeong et al., 2015). Moreover, they also published a corpus, namely Korean TimeBank, including a bunch of Korean documents annotated by the revised Korean TimeML.

In this paper, we propose a concept to find relative temporal information from Korean documents and complement the previous version of Korean TimeML to be able to clearly annotate that relative information. Additionally, we annotate many Korean documents to increase the size of the Korean TimeBank including the relative temporal information, and refine the annotated files in the corpus to improve the quality. Since the relative relationships between temporal entities are potentially useful information, we believe that this extended version of the Korean TimeBank will contribute to broad areas of research.

The rest of this paper is organized as follows. Section 2 explains a concept of relative temporal information and its annotation. Section 3 describes the extended version of Korean TimeBank in detail, and Section 4 concludes.

2. Relative Temporal Information

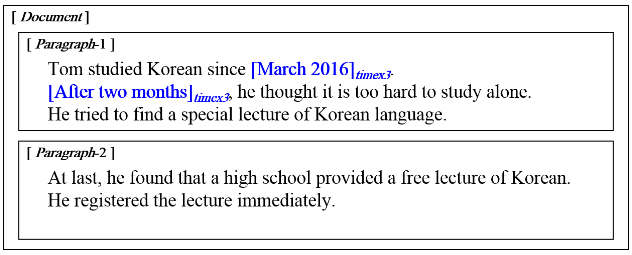
2.1. Concept of Annotating the Relative Temporal Information

The relative linkage between temporal entities—time expressions and events—is potentially important information when determining the temporal context. However, the *timex3* tag, which is treated as a target representation of annotation work in Korean TimeML, must store the exact value of date and time into '*value*' attribute. At this time, the date or time information is always converted to the absolute value (i.e., normalization) to store the *value* attribute even though that expression has a differential value from a specific reference date or time. For example, let us suppose there is a time expression 'a day ago' and its reference date is '2017-09-30'. We can obviously know that there are two *timex3* tags as follows.

```
<timex3 id="t1" value="2017-09-30"/>
<timex3 id="t2" text="a day ago"
value="2017-09-29"/>
```

Table 1: Examples of *relValue* attribute

Example in Korean	Meaning in English	relVaule
5년 4개월 후	after 5 years and 4 months	+P5Y4M
2주 전	2 weeks ago	-P2W
1시간 30분 25초 전	an hour, 30 minutes and 25 seconds ago	-PT1H30M25S
2년 10개월 15일 10시간 20분 30초 후	after 2 years 10 months 15 days 10 hours 20 minutes 30 seconds	+P2Y10M15DT10H20M30S
이번 달	this month	P0M


[*Document*]

[*Paragraph*-1]

Tom studied Korean since [March 2016]$_{timex3}$.
[After two months]$_{timex3}$, he thought it is too hard to study alone.
He tried to find a special lecture of Korean language.

[*Paragraph*-2]

At last, he found that a high school provided a free lecture of Korean.
He registered the lecture immediately.

Figure 1: An example of two *timex3* entities and a relation of *RT2* instead of other types.

According to the existing Korean TimeML, the *timex3* tag 't2' has normalized value from the reference tag 't1', however, it could not maintain that the meaning of 'a day ago' is 'one day earlier than the reference date'. We need to prevent this kind of information loss by separating the relative value into additional attribute.

In addition, we divide the types of relative temporal relationships among entities into four levels as follows.

- *RT1 (In-Sentence)*: a relative temporal relation of two entities which are existed in a 'single sentence'.
- *RT2 (Between-Sentences)*: a relative temporal relation of two entities between 'two consecutive sentences' where an entity is existed in a sentence first and another is in the next sentence.
- *RT3 (In-Paragraph)*: a relative temporal relation of two entities in a 'paragraph' that consists of three or more sentences.
- *RT4 (In-Document)*: a relative temporal relation of two entities in a 'whole document' that consists of two or more paragraphs.

When a temporal relation is annotated by one of the four levels above, a relationship type having a small tolerance range has a higher priority than others. For instance, let us consider a given document including only two short paragraphs as shown in Figure 1. In the first paragraph of this example, there are three sentences and two *timex3* entities—'March 2016' and 'after two months'. They obviously have a temporal relationship over two consecutive sentences. Therefore, the relation should be *RT2* instead of *RT3* or *RT4* due to the relation's range.

2.2. Changes in the Annotation Language

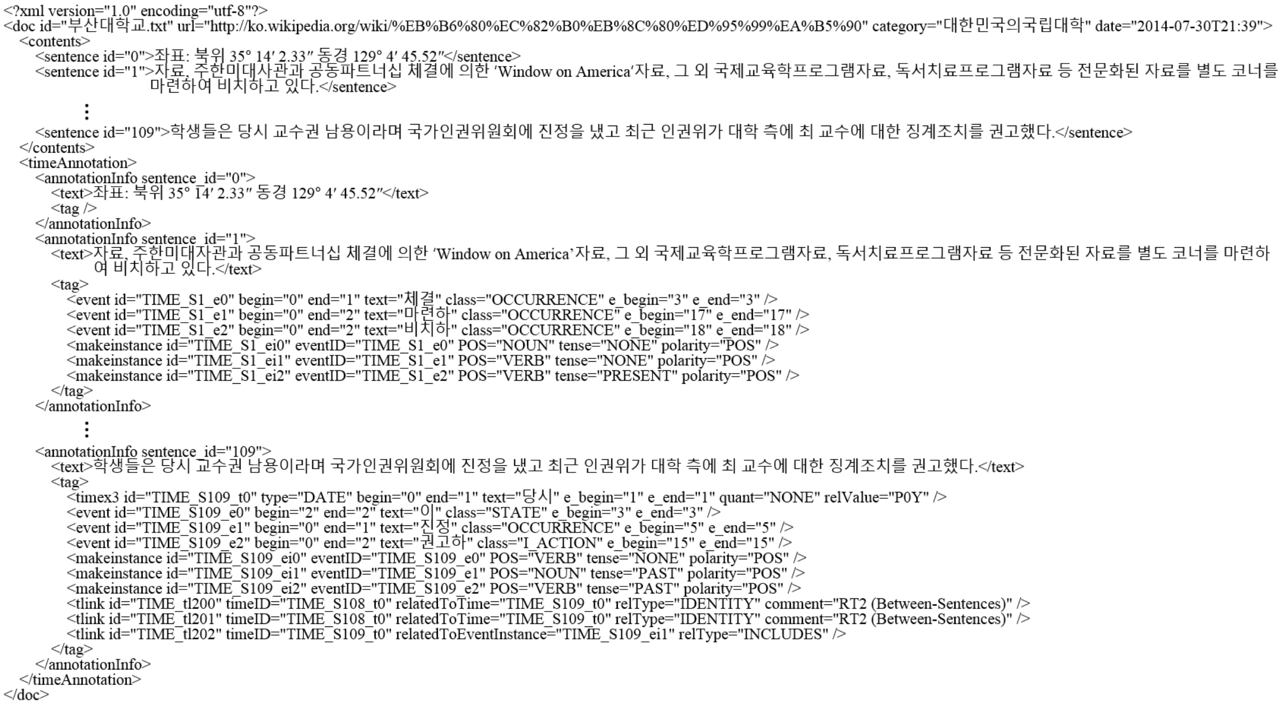
We basically annotated the Korean temporal information according to the structure of the existing Korean TimeML (Jeong et al., 2015). To accurately represent the differential value of relative temporal information, we add a new attribute '*relValue*' to the *timex3* tag in this annotation language. Table 1 shows some examples of the *relValue* attribute. The *relValue* refers the ISO-8601 standard to express an amount of the difference—similar to the value of 'DURATION' type of the *timex3*. And there is a prefixed symbol either '+' or '-' depends on the direction of the relative relationships between temporal entities. In other words, the *relValue* should be start with '+' symbol if the meaning of text is later than the reference date, and vice versa. However, in a special case such as the last example in Table 1, there is no symbol to prefix because the meaning of text exactly pointed the specific date or time.

We believe that this work will contribute to reasoning of temporal relationship according to the relative temporal information. (Gennari and Vittorini, 2016) explained a reasoning system based on a service-oriented architecture (SOA), which is called SOA-based Qualitative Temporal Reasoner (SQTR). SQTR had applied knowledge representation techniques and tools to improve the performance of temporal reasoning on the annotated data. Similar to this work, in the perspective of knowledge representation, our work also can help to grasp relative relationships among entities.

3. Extended Korean TimeBank

Korean TimeBank v2.0 is an extended version of the previous Korean TimeBank (Jeong et al., 2016), which had been introduced on the last LREC 2016. In this version of dataset, we adopt new concept of relative temporal information by adding new *relValue* attribute of *timex3* tags and annotating relative relationships among them. Also, we continuously refined the annotated documents to improve the quality of corpus.

The statistics of Korean TimeBank v2.0 is summarized in Table 2. Compared with the previous version of Korean TimeBank, the total number of documents and sentences increased by about 121.9% and 52.7%. In addition, the number of *timex3*, *event*, *makeinstance*, *tlink* tags increased by about 35.7%, 52.3%, 51.9% and 23.8%, respectively. The annotation work for the corpus was performed by two annotators and one supervisor, and the supervisor mediated and led the annotators' consent when they had different opinions for an annotation result.



```
<?xml version="1.0" encoding="utf-8"?>
<doc id="부산대학교.txt" url="http://ko.wikipedia.org/wiki/%EB%B6%80%EC%82%B0%EB%8C%80%ED%95%99%EA%B5%90" category="대한민국의국립대학" date="2014-07-30T21:39">
  <contents>
    <sentence id="0">좌표: 북위 35° 14′ 2.33″ 동경 129° 4′ 45.52″</sentence>
    <sentence id="1">자료, 주한미대사관과 공동파트너십 체결에 의한 'Window on America'자료, 그 외 국제교육학프로그램자료, 독서치료프로그램자료 등 전문화된 자료를 별도 코너를
                    마련하여 비치하고 있다.</sentence>
              ⋮
    <sentence id="109">학생들은 당시 교수권 남용이라며 국가인권위원회에 진정을 냈고 최근 인권위가 대학 측에 최 교수에 대한 징계조치를 권고했다.</sentence>
  </contents>
  <timeAnnotation>
    <annotationInfo sentence_id="0">
      <text>좌표: 북위 35° 14′ 2.33″ 동경 129° 4′ 45.52″</text>
      <tag />
    </annotationInfo>
    <annotationInfo sentence_id="1">
      <text>자료, 주한미대자관과 공동파트너십 체결에 의한 'Window on America'자료, 그 외 국제교육학프로그램자료, 독서치료프로그램자료 등 전문화된 자료를 별도 코너를 마련하
           여 비치하고 있다.</text>
      <tag>
        <event id="TIME_S1_e0" begin="0" end="1" text="체결" class="OCCURRENCE" e_begin="3" e_end="3" />
        <event id="TIME_S1_e1" begin="0" end="2" text="마련하" class="OCCURRENCE" e_begin="17" e_end="17" />
        <event id="TIME_S1_e2" begin="0" end="2" text="비치하" class="OCCURRENCE" e_begin="18" e_end="18" />
        <makeinstance id="TIME_S1_ei0" eventID="TIME_S1_e0" POS="NOUN" tense="NONE" polarity="POS" />
        <makeinstance id="TIME_S1_ei1" eventID="TIME_S1_e1" POS="VERB" tense="NONE" polarity="POS" />
        <makeinstance id="TIME_S1_ei2" eventID="TIME_S1_e2" POS="VERB" tense="PRESENT" polarity="POS" />
      </tag>
    </annotationInfo>
              ⋮
    <annotationInfo sentence_id="109">
      <text>학생들은 당시 교수권 남용이라며 국가인권위원회에 진정을 냈고 최근 인권위가 대학 측에 최 교수에 대한 징계조치를 권고했다.</text>
      <tag>
        <timex3 id="TIME_S109_t0" type="DATE" begin="0" end="1" text="당시" e_begin="1" e_end="1" quant="NONE" relValue="P0Y" />
        <event id="TIME_S109_e0" begin="2" end="2" text="이" class="STATE" e_begin="3" e_end="3" />
        <event id="TIME_S109_e1" begin="0" end="1" text="진정" class="OCCURRENCE" e_begin="5" e_end="5" />
        <event id="TIME_S109_e2" begin="0" end="2" text="권고하" class="I_ACTION" e_begin="15" e_end="15" />
        <makeinstance id="TIME_S109_ei0" eventID="TIME_S109_e0" POS="VERB" tense="NONE" polarity="POS" />
        <makeinstance id="TIME_S109_ei1" eventID="TIME_S109_e1" POS="NOUN" tense="PAST" polarity="POS" />
        <makeinstance id="TIME_S109_ei2" eventID="TIME_S109_e2" POS="VERB" tense="PAST" polarity="POS" />
        <tlink id="TIME_tl200" timeID="TIME_S108_t0" relatedToTime="TIME_S109_t0" relType="IDENTITY" comment="RT2 (Between-Sentences)" />
        <tlink id="TIME_tl201" timeID="TIME_S108_t0" relatedToTime="TIME_S109_t0" relType="IDENTITY" comment="RT2 (Between-Sentences)" />
        <tlink id="TIME_tl202" timeID="TIME_S109_t0" relatedToEventInstance="TIME_S109_ei1" relType="INCLUDES" />
      </tag>
    </annotationInfo>
  </timeAnnotation>
</doc>
```

Figure 2: An example of annotated document in the Korean TimeBank v2.0.

In this version of dataset, there are 184 *timex3* tags which are specified the *relValue* attribute. And there are 333 *tlink* tags which are directly connected to those *timex3* tags including *relValue*. Table 3 summarizes the statistics of relative temporal relations in the Korean TimeBank.

As an example of annotation work, Figure 2 shows a part of a sample annotated document in the Korean TimeBank. We used stand-off scheme that annotated results of a document write into a separated file by XML format. Figure 3 shows the XML schema to store the annotated document of Korean TimeBank. Each block means an XML node where the header text is the name of node and the list of items is attributes of the node. An arrowed line means a connection from a parent node to its child node. Small text nearby the arrow, either '1' or '*' symbol, is a cardinality of the node where '1' means a single child node must be appeared and '*' means any number of child nodes is allowed. Since we create an annotated document for a corresponding document individually, only one *doc* node is existed as a root node. For each sentence of the given document, *sentence* node is created and it will be a child of *contents* node. Also, *annotationInfo* nodes are stored corresponding to the sentences. At the bottom of this diagram, there are four types of tags we used to—i.e., *timex3*, *event*, *makeinstance*, and *tlink*. The temporal information is stored in the annotation document by these four kinds of tag nodes.

Table 2: Summary of the Korean TimeBank v2.0

Name	Count
# of documents	2,393
# of sentences	6,189
# of empty sentences (no tags)	112
# of words	78,327
Avg. # of words per sentence	12.65584
# of morps	188,687
Avg. # of morps per sentence	30.48748
# of *timex3* tags	3,462
# of *event* tags	17,543
# of *makeinstance* tags	17,583
# of *tlink* tags	4,933

Table 3: Relative Temporal Relations in the Korean TimeBank v2.0

Relation Type	Count	Proportion
RT1 (In-Sentence)	242	72.67%
RT2 (Between-Sentences)	34	10.21%
RT3 (In-Paragraph)	12	3.60%
RT4 (In-Document)	45	13.51%
Total	333	100%

4. Conclusion

In this paper, we supplemented the annotation language to reflect relative temporal information and presented an extended dataset—Korean TimeBank v2.0. Not only the number of documents and sentences in the corpus has significantly increased, but the concept of relative temporal information has also been annotated additionally, so we expect this Korean TimeBank to be useful in various applications of temporal information extraction.

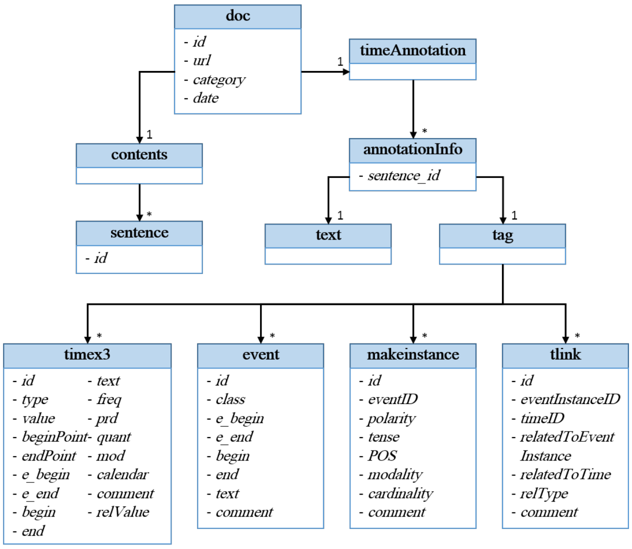


Figure 3: XML schema of annotated document.

5. Acknowledgements

This work was supported by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean Government (MSIP) (No.2013-0-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services; and No.2017-0-00868, Development of Conversational Solution for Intelligent Chat Services Based on Pragmatic and Context Analysis of Dialogues).

6. Bibliographical References

Gennari, R. and Vittorini, P. (2016). Qualitative temporal reasoning can improve on temporal annotation quality: How and why. *Applied Artificial Intelligence*, 30(7):690–719.

Im, S., You, H., Jang, H., Nam, S., and Shin, H. (2009). Ktimeml: specification of temporal and event expressions in korean text. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 115–122. Association for Computational Linguistics.

Jeong, Y.-S., Kim, Z. M., Do, H.-W., Lim, C.-G., and Choi, H.-J. (2015). Temporal information extraction from korean texts. In *Proceedings of the 19th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 279–288.

Jeong, Y.-S., Joo, W.-T., Do, H.-W., Lim, C.-G., Choi, K.-S., and Choi, H.-J. (2016). Korean timeml and korean timebank. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*, pages 356–359.

Jeong, Y.-S. (2016). *Temporal Information Extraction from Korean Texts*. Doctoral thesis, School of Computing, Korea Advanced Institute of Science Technology (KAIST).

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003). Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

UzZaman, N., Llorens, H., Allen, J., Derczynski, L., Verhagen, M., and Pustejovsky, J. (2012). Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333*.