Korean TimeML and Korean TimeBank

Young-Seob Jeong†, Won-Tae Joo‡, Hyun-Woo Do§, Chae-Gyun Lim‡, Key-Sun Choi‡, and Ho-Jin Choi†

‡NAVER LABS, NAVER Corp.
Seongnam, 463-867, South Korea
pinode.waider@navercorp.com
†Korea Advanced Institute of Science and Technology
291 Daehak-ro, Yuseong-gu, Daejeon 305-701, South Korea
{izazu1725,realstorm103,rayote,kschoi,hojinc}@kaist.ac.kr

1. Introduction

Due to the exponentially increasing number of documents available on the Web and from other sources, it has become important to develop methods to automatically extract knowledge from unstructured, natural language documents. The knowledge extracted as such is useful for various applications in the areas of information retrieval (IR), trend analysis (TA), and question answering (QA) systems. Among the many aspects of extracting knowledge from documents, the extraction of temporal information has recently drawn attention. There are two well-known annotation languages of temporal information, Time Markup Language (TimeML) (Pustejovsky et al., 2003) and ISO-TimeML. Although these annotation languages define many tags and attributes for representing various types of temporal information, they do not incorporate language diversity. For example, they assume that annotation is performed in the token level. However, Korean is an agglutinative language whose words are formed by joining morphemes together, so it can not be annotated properly in the token level. As an annotation language for Korean, the Korean TimeML (KTimeML) was proposed (Im et al., 2009), and its contributions can be summarized as follows: (1) it employs a morpheme-level standoff annotation scheme, (2) it takes a surface-based annotation scheme, (3) it suggests to cancel the head-only markup policy of TimeML, (4) it addresses several Korean-specific issues (e.g., the usage of signal tag for only temporal connectives), and (5) it introduces the TARSQI Toolkit for the annotation process following the KTimeML. In this paper, we argue that the KTimeML has some limitations, and propose a revised version of the KTimeML. For example, the previous KTimeML did not consider some characteritics of Korean (e.g., a lunar calendar), and the morpheme-level annotation of the KTimeML makes it difficult to share the dataset. Our new KTimeML overcomes such limitations, and we also introduce the Korean TimeBank constructed using the new KTimeML.

The rest of this paper is organized as follows. Section 2 presents details of the Korean TimeML. Section 3 introduces the Korean TimeBank, and Section 4 concludes the paper.

2. Korean TimeML

2.1. Limitations of the Previous Korean TimeML

We argue that the previously proposed KTimeML has five limitations. First, although it was proposed as an annotation language for Korean, it misses some characteristics of Korean. Temporal expressions based on the lunar calendar appear often, where the normalized value of such temporal expressions cannot be represented using the Gregorian calendar. For example, for the sentence “어머니 생신은 4 월4일이다” (Mother’s birthday is on the 4th day of the 4th month in the lunar calendar), the normalized value of the 4th day of the 4th month will be different on different years in the Gregorian calendar (e.g., ‘2015-05-21’ for the year 2015, ‘2014-05-02’ for the year 2014). Moreover, there are some temporal expressions conveying vague temporal information that appear often in Korean. For example, ‘초 중반[cho-joong-ban]’ represents the beginning or middle phase of a period, and ‘중후반[joong-hoo-ban]’ represents the middle or ending phase of a period. There is no way to annotate these expressions using the previous KTimeML.

Second, there are temporal expressions conveying periodic patterns that can not be annotated using the previous KTimeML. For the sentence “I visit there twice every week, each of which takes one day”, there is no way to annotate the expression ‘every week’ because the attribute freq of
**timex3** tag can not represent ‘twice’ and ‘one day’ simultaneously. The reason for this limitation is the inconsistent usage of the attribute freq. That is, the freq is used to annotate not only a periodic frequency (e.g., ‘twice’), but also a periodic duration (e.g., ‘one day’). When these two periodic patterns appear simultaneously, then the temporal expressions will not be annotated properly using the previous KTimeML.

Third, the previous KTimeML takes a morpheme-level annotation. From a linguistic point of view, morpheme-level annotation seems perfect because the smallest meaningful unit of Korean is the morpheme. However, from a practical point of view, morpheme-level annotation makes it difficult to distribute or share the dataset. The reason is that there are multiple tag-sets of morphemes, so the datasets using different tag-sets will not be consistent with each other. Even if all the datasets are commonly based on a single tag-set, they will not be consistent unless they use the same morphological analyzer. Because the essential purpose of annotation language is to help to distribute or share the dataset, morpheme-level annotation must be avoided.

Fourth, different attribute names are used to denote the IDs of different tags. For example, tid is used for **timex3** tags, and lid is used for **tlink** tags. One may argue that using the different attribute names to denote IDs will make the various kinds of tags easier to recognize. However, in terms of further applications that make use of temporal information, it is not necessary to use various attribute names to denote tag IDs the kind of tag is already known when its attributes are parsed. Rather, using different names to denote IDs makes it complex to implement programs to parse the tag attributes.

Fifth, similar to ISO-TimeML, an event tag plays two roles: the role of an event token and the role of an event instance. Given the sentence “Kevin taught English yesterday and today”, there will be two event tags as follows.

```
<EVENT eid="e1" morph="m1" pred="TEACH"
  class="OCCURRENCE" tense="PAST"
  polarity="POS"/>
<EVENT eid="e2" pred="TEACH"
  class="OCCURRENCE" tense="PAST"
  polarity="POS"/>
```

The first event tag has the two roles (e.g., a role of an event token and a role of an event instance), while the second event tag has only the role of an event instance. From a practical point of view, this inconsistent functionality of event tags may cause difficulty in parsing of the annotated event tags, which would result in the inefficiency of further applications. In other words, it would be necessary to implement a program to recognize the role of the event tag, which would slow down the applications.

### 2.2 Modified Korean TimeML

To address the limitations of the previous KTimeML, we revise the KTimeML by introducing some additional attributes and modifying some existing attributes. In terms of the first limitation, we add an attribute calendar of **timex3** tag to denote the calendar types, where its value can be LU-NAR, JULIAN, or any other types of calendar. The default value of calendar is GREGORIAN when it is not explicitly clarified. To annotate expressions conveying vague temporal information, we introduce two values START_MID and MID_END to the attribute mod of **timex3** tag. The value START_MID represents the beginning or middle phase of a period (e.g., ‘초중반[cho-joong-ban]’) and the value MID_END represents the middle or ending phase of a period (e.g., ‘중후반[joong-hoo-ban]’).

To address the second limitation, we introduce an additional attribute prd of **timex3** tag, where it is used to represent periodic duration based on ISO-8601. The existing attribute freq of **timex3** is also modified so that it is used to represent only the periodic frequency. This role separation between freq and prd makes it possible to annotate temporal expressions that could not be annotated using the previous KTimeML. For example, given the sentence “I visit there twice every week, each of which takes three hours”, the **timex3** tag of ‘every week’ will have freq=’2X’ and prd=’PT3H’.

To address the third limitation, we propose taking a character-level annotation. Character-level annotation will make the dataset independent of morpheme tag-set and morphological analysis, which in turn makes it easy to distribute or share the dataset. To realize character-level annotation, we replace the attribute morph with five attributes e_begin, e_end, begin, end, and text. The attributes e_begin and e_end indicate token indices of the extent, while begin and end indicate character (letter) indices of the extent. For example, given the sentence “I work today” in Fig. 1, contains one **timex3** tag whose text is ‘today’, where e_begin=2, e_end=2, begin=0, and end=4. One may argue that the e_begin and e_end represent the token-level information, so it seems that the proposed annotation does not take the character-level annotation. It is true that these two attributes may seem unnecessary, because the other two attributes begin and end carry the character-level information. However, it is important to notice that the annotation language is used by not only computers, but also by human. Using only begin and end will be enough for the computers to work, but not for human annotators. The usage of e_begin and e_end helps human to easily annotate new corpus or to check an annotated corpus. For example, given a sentence “There were many works I had to do, but I left it yesterday”, human annotators have to annotate ‘yesterday’ with **timex3** tag. Without using e_begin and e_end, it will make the annotators hard to annotate, because the annotators have to count the number of preceding characters. Furthermore, it will also be more difficult to read or check whether the annotated **timex3** tag is correct or not, due to the same reason. Thus, the two attributes are necessary to help human annotators.
To address the fourth limitation, we just use the same attribute name id for every tag, as ISO-TimeML does. To address the fifth limitation, we employ a makeinstance tag, which is also adopted by TempEval shared tasks (Verhagen et al., 2009; Verhagen et al., 2010; UzZaman et al., 2013). The makeinstance tag takes the role of an event instances, while the event tag has only the role of an event token. This clear separation of the two roles will help the further applications to easily analyze event tags. As there is at least one instance for each event token, the number of event tags is always smaller than or equal to the number of makeinstance tags.

3. Korean TimeBank

There are some existing Korean datasets of temporal information. A Korean dataset constructed using timex2 was introduced (Jang et al., 2004), where timex2 is the former version of timex3. The first Korean dataset using timex3 appeared in TempEval-2, which provides datasets of six languages: Chinese, English, French, Italian, Spanish, and Korean. However, the Korean dataset of TempEval-2 is small in size (e.g., totally 26 documents) and has many annotation errors. There are some missing values of timex3 tags, and there are some tags that must be merged into one. An example of the errors can be found at the 11th sentence of the 2nd training document within the TempEval-2 Korean dataset. Moreover, it is annotated in the morpheme level, which implies that it will not be consistent with other datasets. Thus, we introduce a new Korean dataset, namely Korean TimeBank, which is annotated in the character level.

The source of the Korean TimeBank includes Wikipedia documents and hundreds of manually generated question-answer pairs. The domains of the Wikipedia documents are personage, music, university, and history. The annotation is performed by two well-trained annotators majoring in computer science and examined by a supervisor. The statistics of the Korean TimeBank are summarized in Table 1, and the Korean TimeBank will be extended regularly. The Kappa coefficient κ is described in Table 2.

Table 1: The statistics of Korean TimeBank.

| Item   | The number of items |
|--------|--------------------|
| document | 1078               |
| sentence | 4053               |
| timex3   | 2552               |
| event    | 11522              |
| makeinstance | 11577          |
| tlink    | 3985               |

Table 2: Kappa coefficient of Korean TimeBank.

| Tag     | Kappa coefficient |
|---------|-------------------|
| timex3  | 0.9983            |
| event   | 0.9889            |
| makeinstance | 0.9930      |
| tlink   | 0.9284            |

Figure 2: Sample annotated sentences of Korean TimeBank.
sentences, while timeAnnotation contains pairs of an original sentence and annotated tags. This stand-off scheme allows the original sentences to be kept unharmed. Each annotationInfo of timeAnnotation contains the pair of an original sentence and tags within a sentence, where sentence_id is an index of the sentence. The text of annotationInfo is the original sentence, and tag contains the annotated tags.

4. Conclusion

As there are several limitations of the previous Korean TimeML (KTimeML), we proposed a new modified version of KTimeML and introduced a Korean TimeBank constructed using a part of the new KTimeML. We believe that the Korean TimeBank will be widely used for many Korean-based studies and applications related to temporal information, because this is the first high-quality Korean dataset that is independent to any tag-sets or morphological analysis tools because it is annotated in the character level.

5. Acknowledgements

This work was supported by ICT R&D program of MSIP/I-ITP. [R0101-15-0062, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services]

6. Bibliographical References

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, Suntec, Singapore.

Jang, S. B., Baldwin, J., and Mani, I. (2004). Automatic timex2 tagging of korean news. ACM Transactions on Asian Language Information Processing, 3(1):51–65.

Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., and Katz, G. (2003). Timeml: Robust specification of event and temporal expressions in text. In New Directions in Question Answering, pages 28–34, Stanford, USA.

UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J., and Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In Proceedings of the Seventh International Workshop on Semantic Evaluation, pages 1–9, Atlanta, Georgia, USA.

Verhagen, M., Gaizauskas, R. J., Schilder, F., Hepple, M., Moszkowicz, J., and Pustejovsky, J. (2009). The tempeval challenge: Identifying temporal relations in text. Language Resources and Evaluation, 43(2):161–179.

Verhagen, M., Sauri, R., Caselli, T., and Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In Proceedings of the Fifth International Workshop on Semantic Evaluation, pages 57–62, Uppsala, Sweden.