한국어 TimeML
—텍스트의 사건 및 시간 정보 연구

유현조*, 장하연, 조유미, 김윤신
서울대학교 신라대학교
남승호, 신효필†
서울대학교

Hyun-Jo You, Hayeon Jang, Yu-Mi Jo, Yoon-shin Kim, Seungho Nam, Hyopil Shin. 2011. The Korean TimeML: A Study of Event and Temporal Information in Korean Text. *Language and Information 15.1*, 31–62. TimeML is a markup language for events and temporal expressions in natural language, proposed in Pustejovsky et al. (2003) and latter standardized as ISO-TimeML (ISO 24617-1:2009). In this paper, we propose the further specification of ISO-TimeML for the Korean language with the concrete and thorough examination of real world texts. Since Korean differs significantly from English, which is the first and almost only extensively tested language with TimeML, one continuously run into theoretical and practical difficulties in the application of TimeML to Korean. We focus on the discussion for the consistent and efficent application of TimeML: how to consistently apply TimeML in accordance with Korean specificity and what to be annotated and what not to be, i.e. which information is meaningful in the temporal interpretation of Korean text, for efficient application of TimeML. (**Seoul National University**)

Key words: 사건, 시간표현, 한국어, 말뭉치, 의미주석, TimeML, TimeBank, Korean language, corpus, semantic annotation

1. 서론

본 연구는 텍스트 수준의 시간성을 분석하기 위한 연구의 하나이다. 한국어 사건과 시간표현에 대한 어휘의미론적인 연구는 많이 이루어져 있으나 텍스트 수준의 분석에 관한 연구는 충분히 이루어지지 않았다. 다음은 어느 도서관의 공지사항에 나타난 한 문장이다.

* 주저자. E-mail: youhyunjo@snu.ac.kr
† 교신저자. E-mail: hpshin@snu.ac.kr

(1) 어린이 도서관 무료 개방 시간은 4월 20일부터 5월 25일까지 매주 일요일 오후 2시부터 오후 7시까집니다.

이 문장에 나타난 '4월 20일', '5월 25일', '매주', '일요일', '오후', '2시부터', '7시까지' 등 각각의 시간표현에 대한 어휘의미론적 해석은 큰 어려움을 요하지 않는다. 그러나 위 공지가 2008년에 작성된 것이라는 정보가 주어졌을 때 달력에서 2008년 4월 20일에서 5월 25일 사이의 일요일이 2008년 4월 20일과 27일, 그리고 5월의 4일, 11일, 18일, 25일이라는 것을 계산하고 이들 6개 날짜에 대해 각각 오후 2시부터 7시까지 도서관이 무료로 개방한다는 사실을 추론하는 것은 단순한 과제가 아니다. 이러한 시간 의미 해석과 추론이 가능해야 다음과 같은 질문에 '예'라고 답할 수 있다.

(2) 2008년 5월 4일 오후 5시는 어린이 도서관 무료 개방 시간인가?

더 나아가 질문 시점이 2008년 5월 23일 때 다음과 같은 질문에 대답하기 위해서는 '다음 주 일요일'을 2008년 6월 1일로 해석할 수 있어야 '아니오'라는 답을 이끌어낼 수 있다.

(3) 다음 주 일요일에도 어린이 도서관 무료로 개방하나요?

이렇게 시간표현 각각의 어휘 의미는 단순하지만 텍스트 내에서 그 시간 의미를 해석하는 것은 현재 많은 연구를 필요로 한다. 특히, 위와 같은 질문에 답하기 위한 질의응답 시스템에서 시간 추론은 필수적임에도 불구하고 시간 관계에 대한 해석을 필요로 하는 질의는 현재의 질의응답 시스템에서는 제대로 풀지 못하고 있다(Moldovan, Clark, and Harabagiu, 2005).

위의 예시에서 볼 수 있는 바와 같이 본 연구는 컴퓨터언어학적 응용을 염두에 두고 있다. 사건 및 시간 정보 분석은 정보 검색, 질의응답 시스템, 자동 요약과 같은 자연언어처리 응용 분야의 발전을 위해 중요한 요소로 인식되어 왔으며 사건의 시간 정보는 자연언어 텍스트의 의미 내용에서 핵심적인 부분을 구성하고 있다. 텍스트 수준의 사건의 해석의 예로 어느 병원의 수술 후 치료 과정에 관한 설명을 보자.

(4) 수술 후 일주일간은 매일 나오셔서 치료를 받으시고 이후로는 1주일마다 2번씩 2개월간 치료 받으시고 이후에는 3개월간 처방 받은 약을 매일 1회 복용하시고 이후에는 3개월 간격으로 방문하시어 검사 받으시면 됩니다.

이 텍스트에는 '수술', '치료', '복용', '방문', '검사' 등의 사건이 어떤 순서로 어느 기간 동안 어떻게 반복되는지 나타나있다. 이 텍스트의 정보를 이용하여 누군가 2009년 7월 8일에 수술을 받았을 때 그 사람이 받아야할 치료과정의 일정을 표 1

[표 1] 시간 정보 추출의 예

시간	사건
2009년 7월 8일	수술
2009년 7월 9일	치료
2009년 7월 10일	치료
...	
2009년 7월 15일	치료
2009년 7월 16일 – 22일	치료
2009년 7월 23일 – 29일	치료
...	
2009년 9월 10일 – 16일	치료
2009년 9월 17일	약 복용
2009년 9월 18일	약 복용
...	

과 같이 작성할 수 있을 것이다. 이렇게 시간 정보를 자동으로 추출, 요약, 추론하는 것은 텍스트 이해에서 매우 중요한 요소임에도 불구하고 다른 측면의 의미 해석에 비해 충분히 주목을 받지 못했다(Setzer, Galzauskas, and Hepple, 2005, p.234).

이러한 분석을 위해서는 자연언어처리를 염두에 둔 모든 분야에서 그러하듯이 가장 우선적으로 필요한 것이 정교하게 다듬어진 고급 주석 말뭉치의 구축이다. 주석 말뭉치는 자동화 시스템 발전의 원동력이 되고 상이한 시스템들을 평가 비교하는 기준과 기계 학습 기반 접근법을 위한 트레이닝 자료의 역할을 하기 때문에 그 필요성은 이루 말할 수 없다(Setzer, Galzauskas, and Hepple, 2005, p.234).

이렇게 텍스트에 나타나는 사건 및 시간 표현을 분석하기 위하여 제안된 형식이 바로 Pustejovsky et al. (2003)의 TimeML이라는 마크업 언어이며 이에 바탕하여 언어 자원의 의미 분석을 위한 국제 표준의 하나로 지정된 것이 ISO-TimeML이다. TimeML은 텍스트의 시간 정보를 주석하기 위한 형식으로 이를 이용하여 시간 정보 주석 말뭉치인 TimeBank가 구축될 수 있으며 TimeBank는 텍스트의 시간 정보 처리를 위한 연구에 이용될 수 있다.

2. 선행연구

2.1 한국어 사건 및 시간 정보 처리 연구

국내 연구에서는 사건 및 시간 정보 처리를 위한 논의가 많이 존재하지 않으나 그 중요성에 대한 공감대가 형성되면서 시간 정보 처리에서 가장 기초가 되는 사건 및

시간 표현 추출과 그 의미 해석에 대한 연구가 근래에 시작되었다. 관련된 공학적인 연구들의 초기 성과로 간단한 자연언어처리 기술을 이용한 시간 표현 추출 및 정규화(김평, 성기윤, 맹성현, 2003), 자동 요약에 사용될 수 있는 사건 탐지 기법(정영미 · 김용광, 2008) 등이 있다. 이론적 연구로 (남지순, 2008)에서는 인터넷 신문의 특정 유형 기사 텍스트에서 실현되는 시간 관련 표현을 추출하여 그 유형과 특징을 언어학적으로 기술하려는 시도를 하였다.

개별적인 두 사건의 시간 관계에 대한 연구에 더해서 본 연구에서 목표로 하는 것의 하나는 텍스트 전체의 시간 구조와 시간 관계 추론에 대한 이론적 연구이다. 시간에 따라 전개되는 이야기로서 자연 언어 텍스트가 주어지면 사건과 시간 정보를 통해 독자는 시간의 순서를 이해한다. 예를 들어, 철수의 하루 일과에 대한 텍스트를 읽고 "철수가 영희를 만나기 전에 무엇을 했는가?" 라는 질문에 대답할 수 있다.

이러한 연구는 시간 관계 분석 자동화의 기초가 되나 관련 선행 연구는 많지 않다. 이에 대한 연구로 '시간 닫힘(temporal closure)'이라는 개념을 도입하여 시간 관계 추론을 연구하려는 시도가 존재한다. 시간 닫힘은 텍스트에서 이미 알려진 시간 관계를 취하여 새로운 시간 관계를 도출하는 것이다. 즉, 암시적으로 존재하던 시간 관계를 명시적으로 만들어주는 것이다(Verhagen, 2005).

시간 논리 연구에는 시간 구조를 형식적으로 표상하기 위한 틀에 대한 연구, 시간 관계 추론에 관한 연구 등이 있다. 이러한 연구들은 언어 보편적인 성격을 가지므로 한국어 텍스트의 시간 정보를 보편적인 논리 구조로 표상한 후에는 언어보편적인 논리적 추론 규칙에 따라 분석할 수 있게 된다.

2.2 시간 관계 표상을 위한 연구

텍스트 수준의 시간 정보 분석에서 가장 핵심적인 문제는 텍스트에 나타나는 사건 및 시간표현들 사이의 관계를 분석하는 것이다. 이러한 관계를 분석하기 위해서는 형식적인 틀이 필요하다. 이러한 시간 관계를 표상하기 위한 형식화는 근래에 와서야 활발히 논의되기 시작하였다.

두 사건의 시간 관계를 분석하는 틀로 가장 널리 받아들여지는 것의 하나는 시간 관계를 순수하게 논리적인 차원에서 13가지로 구분하고 있는 Allen (1984)의 'interval algebra'이다. 본고에서 다루고 있는 TimeML에서도 기본적으로 Allen (1984)의 형식적 틀을 수용하고 있다.

자연 언어 텍스트에서 시간 관계를 분석하고 주석하기 위한 틀의 하나로 Katz and Arosio (2001)는 문장 내 선후 관계와 동사들 사이의 포함 관계를 분석하기 위한 단순한 시간 정보 주석 언어를 제안했다. 이 주석 언어는 $<$와 $>$로 선후 관계를 표시하고 $\subset$와 $\supset$로 포함 관계를 표시한다. Setzer (2001)는 STAG (Sheffield Temporal Annotation Guidelines)라는 주석 지침을 만들고 주석 언어를 정의했다. 여기에서는

사건과 시간이 5가지 관계, 즉 'before', 'after', 'includes', 'included', 'simultaneous'로 연결된다. 그리고 주석자 사이의 일치를 높이기 위하여 연역적 닫힘(deductive closure)을 사용하였다.

이러한 연구 성과를 통합하여 Pustejovsky et al. (2003)에서 XML 형식으로 시간 정보를 주석화하기 위한 마크업 언어인 TimeML을 제안하였으며 이를 바탕으로 실제 자연언어 텍스트를 주석한 TimeBank를 구축하는 실증적 작업이 이루어졌다.

2.3 ISO-TimeML

텍스트의 시간 정보를 표상하기 위한 마크업 언어 TimeML은 2007년부터 국제 표준으로 수용하기 위한 논의가 시작되어 2009년 ISO-TimeML이 제정되었다. 이러한 ISO-TimeML의 제정 논의에 영어, 스페인어, 중국어, 이태리어 등의 연구자들과 함께 국내 학자들도 적극적으로 참여하였다(Lee, Pustejovsky, and Boguraev, 2006).

ISO-TimeML에서 다루는 분석 대상은 각각 하나의 태그로 표상된다. 다음과 같이 사건(⟨EVENT⟩), 시간 표현(⟨TIMEX3⟩), 신호(⟨SIGNAL⟩)의 기본 태그 3가지와 상적 연결(⟨ALINK⟩), 종속적 연결(⟨SLINK⟩), 시간적 연결(⟨TLINK⟩)의 연결 태그 3가지가 그것이다.

(5) 기본 요소
   a. ⟨EVENT⟩: occurence, state, aspectual, perception, reporting, intensional action, intensional state
   b. ⟨TIMEX3⟩: time, date, duration, set
   c. ⟨SIGNAL⟩:

(6) 연결 요소
   a. ⟨ALINK⟩: initiation, culmination, termination, continuation, reinitiation
   b. ⟨SLINK⟩: modal, factive, evidential, conditional, counterfactive, negative evidential
   c. ⟨TLINK⟩: identity, simultaneous, before, after, immediate after, immediate before, includes, is-included, during, during-inverse, begins, begun-by, ends, ended-by

기본 요소는 텍스트에서 시간 정보를 제공하는 기본 단위를 분석하기 위한 것이다. 사건(⟨EVENT⟩)과 시간 표현(⟨TIMEX3⟩)이 그 기본 단위가 된다. 신호(⟨SIGNAL⟩)는 사건 또는 시간 표현 사이의 관계를 표현하는 언어 요소로서 전치사/후치사, 접속사 등이 이에 해당된다.

연결 요소는 시간 관계를 분석하기 위한 것이다. 상적 연결은 한 사건이 다른 사건에 상적 의미를 부여하는 관계를, 종속적 연결은 한 사건이 다른 사건을 논항 또는 종속절로 취하여 양상, 사실성, 증거성 등의 의미 기능을 하는 관계를 분석하기 위한 것이다. 시간적 연결은 사건과 사건, 사건과 시간 표현, 시간 표현과 시간 표현 사이에 존재하는 시간적 일치, 선후, 겹침 관계를 분석하기 위한 것이다.

2.4 한국어 TimeML

ISO-TimeML은 영어 텍스트의 시간 정보 분석에서 시작된 Pustejovsky et al. (2003)의 TimeML을 기초로 하여 다른 언어도 주석할 수 있도록 보편적인 틀을 제시하고 있다. 이러한 이유로 ISO-TimeML을 개별 언어에 적용하기 위해서는 해당 언어에 맞도록 명세화를 할 필요가 있다.

한국어를 위한 TimeML은 ISO-TimeML이 확정되기 이전부터 논의가 진행되었으며 이기용 (2008), Im and Saurí (2008), 임서현 외 (2009) 등에서 그 모습이 제시된 바 있다[1]. 이러한 논의들의 일부는 ISO-TimeML 제정 과정에 반영되기도 하였으나 ISO-TimeML과 완전히 일치하지는 않는다. 본고에서는 이러한 선행 연구의 성과를 기반으로 하여 한국어 TimeML을 명세화하고 실제 한국어 분석에서 생기는 문제점을 논의할 것이다.

한국어 TimeML에 대한 선행연구와 더불어 본고에서도 유지하고 있는 원칙을 요약하면 다음과 같다.

첫째, 한국어 TimeML은 ISO-TimeML과의 호환성을 우선으로 한다. 본 연구는 ISO-TimeML이 제정되기 이전부터 진행되었으며 별도로 발전되었으나 ISO-TimeML의 논의를 최대한 수용하여 반영하였다.

둘째, 한국어 구조를 정확히 반영할 수 있도록 한다. 한국어 TimeML의 명세화는 한국어 구조에 대한 이론적 연구의 결과를 정확히 반영하여 형식적으로 체계화되어야 한다. 한국어 TimeML은 단순히 기계적인 표현 방법이 아니라 한국어 사건 및 시간 표현에 대한 분석을 체계적으로 형식화하여 표상하기 위한 도구이다. 따라서 한국어 사건과 시간 표현의 구조와 특성에 대한 이론언어학적 연구가 엄밀하게 이루어져야 그에 대응되는 정확한 TimeML의 형식화가 이루어질 것이다.

셋째, 자연언어처리를 고려하여 형식화를 한다. 한국어 TimeML의 태그와 속성들은 이론적으로 한국어의 구조를 표상하는 것이면서 동시에 자연언어처리 도구에 의해 자동으로 처리될 대상이기도 하다. 전산화 가능성을 고려하여 지나치게 추상적이거나 지나치게 산만한 정보가 되지 않도록 형식화의 수준을 조절할 필요가 있다. 또한, 전산화를 위해서는 언어학적 이론 연구와 무관하게 형식적인 논의가 필요한 부분도 있다. 태그의 표기 방식에 대한 논의나 이론적 분석과 무관하게 데이터 관리를 위한 정보들의 명세화에 관한 논의가 그에 해당한다.

[1] 이 자리를 빌어 본 연구를 위한 세미나에서 ISO-TimeML에 대해 자세하게 설명해 주시고 한국어 TimeML 명세화와 한국어 TimeBank 구축에 많은 조언을 주신 이기용 선생님께 감사의 말씀을 드린다.

3. 주석 방식과 주석 단위

한국어의 사건 및 시간 정보에 대한 논의에 앞서 실질적인 말뭉치 구축을 위해서는 다음과 같은 주석 방식에 대한 논의가 필요하다.

(7) a. 행내(inline) 주석 방식과 격리(standoff) 주석 방식
b. 어절 단위 주석과 형태소 단위 주석
c. 토큰(token) 주석과 범위(span) 주석

3.1 행내 주석과 격리 주석

본 연구의 초기에는 ISO-TimeML의 명세화가 아직 결정되지 않은 단계로 텍스트 내에 주석을 바로 하는 행내(inline) 주석 방식과 텍스트와 떨어뜨려 주석을 다는 격리(standoff) 주석 방식에 대한 논의가 진행 중이었다. 기존의 TimeML은 행내 주석 방식을 따르고 있었으나 여러 논의 끝에 ISO-TimeML은 격리 주석 방식을 채택하였다. 두 주석 방식의 차이를 예시하면 다음과 같다.

(8) (행내 주석)

```
<TIMEX id="t1" type="DATE" ...>어제</TIMEX>
철수 가
<EVENT id="e1" class="OCCURENCE" ...>왔다.</EVENT>
```

(9) (격리 주석)

```
어제 철수 가 왔다.
<TIMEX target="#token1" id="t1" type="DATE" ... />
<EVENT target="#token3" id="e1" class="OCCURENCE" ... />
```

초기에 TimeML은 행내 주석 방식을 취했으며 영어 TimeBank 1.2까지 그러한 방식으로 구축되었다. 한국어의 경우에는 Im et al. (2009)에서 격리 주석을 제안한 바 있다. 행내 주석 방식은 간결하고 직관적이지만 원 텍스트를 변경하지 않는다는 대원칙에 부합하지 않는다. 격리 주석은 원 텍스트에 변경을 가하지 않는다는 장점이 있으나 주석을 원문에 연결하기 위하여 식별자(xml:id)를 필요로 한다. 원 텍스트의 언어 요소들에 식별자를 부여하는 것은 ISO 표준인 MAF, LAF를 통해 이루어질 수 있다. 구체적인 예는 Pustejovsky et al. (2010)에서 찾아볼 수 있다.

3.2 어절 단위 주석과 형태소 단위 주석

주석의 기본 단위를 어절과 형태소 중 결정하는 것은 실질적인 말뭉치 구축 작업에 있어서 중요한 문제이다. 형태소를 주석의 기본 단위로 할 경우 원시 말뭉치를 형태소 분석을 한 후에야 TimeML 주석을 수행할 수 있다. 형태소 분석기의 성능이

좋다고 하여도 오류가 상당수 존재하므로 형태소 분석 오류를 수정하는 과정을 거쳐야 한다. 어절 단위 주석은 이러한 전처리를 필요로 하지 않는다는 점에서 장점을 가진다.

본 연구에서 실제로 시험적인 주석을 수행한 결과 형태소 단위의 TimeML 주석 작업이 어절 단위 주석에 비해 능률이 매우 떨어지는 것이 확인되었다. 가능하다면 단순하고 전처리를 필요로 하지 않는 어절 단위 주석이 선호되므로 형태소 단위 주석이 필수적으로 요구되는지에 대한 검토가 필요하다.

우선, 형태소 단위 주석은 한 어절에 두 개 이상의 TimeML 태그를 주석해야할 경우에 요구된다. 다음과 같은 예는 하나의 어절에 두 개의 사건이 표현되는 경우로 형태소 단위 주석의 필요성을 보여준다.

(10) 어제 철수-가 왔다$_{e1}$-ㄴ다$_{e2}$ (형태소 단위 주석)

```
<EVENT id="e1" class="OCCURENCE" ... />
<EVENT id="e2" class="REPORTING" ... />
```

(11) 어제 철수가 왔단다$_{e1}$ (어절 단위 주석)

```
<EVENT id="e1" class="OCCURENCE" ... />
```

이와 같은 인용 구성은 접속절과 문장종결에서 '왔다면', '오겠지만', '왔다면서', '오겠다는데', '온다는구나' 등 매우 다양하게 나타난다. 특히 '왔단다'의 경우 '왔다'라는 사건의 시제는 과거이고 이를 인용하는 'ㄴ다'라는 사건의 시제는 과거가 아니라는 점에서 두 개의 독립적인 사건임이 분명하다. 어절 단위 주석을 할 경우 '왔단다'에 어떠한 시제 속성을 줄 것인지 문제가 된다.

다음으로 문법형태소를 제외하고 주석하는 것이 적절하다고 판단되는 경우들을 살펴볼 필요가 있다. 시간 표현의 경우 조사가 붙는 경우가 많이 있다. 조사를 제외하고 시간 표현에 주석을 하고자 한다면 형태소 단위로 주석할 필요가 있다. 아래 예에서는 주석할 부분에 밑줄을 그어 표시하였다.

(12) a. 2011년 3월 19일 9시부터 (어절 단위 주석)

b. 2011년 3월 19일 9시 부터 (형태소 단위 주석)

한국어 조사들은 위와 같이 그 어휘적 의미가 강한 경우들이 존재한다. 위와 같은 경우 '부터'를 제외하고 '2011년 3월 19일 9시'를 시간 표현으로 주석하는 것이 적절하다고 판단된다.

마지막으로 실질적인 말뭉치 구축에서 생기는 문제로서 띄어쓰기의 비일관성을 고려할 필요가 있다. 본 연구의 시험적인 주석에 따르면 원 텍스트의 띄어쓰기의

일관성을 기대할 수 없을 뿐만 아니라 정서법에 전혀 맞지 않는 오류들도 상당수 존재한다. 합성용언, 복합명사 등은 동일한 표현에 대해 띄어쓰기가 다르게 나타난다. 예를 들면 다음과 같다.

(13) a. '작전개시' 또는. '작전 개시'

b. '연구해왔다' 또는 '연구해 왔다'

이러한 경우 원 텍스트의 띄어쓰기를 그대로 수용하여 어절 단위 주석을 할 경우 동일한 표현이 경우에 따라 '작전개시' 하나의 사건으로 주석되거나 '작전'과 '개시'라는 두 개의 사건으로 주석될 수 있다. 어절이라는 것은 문법적 또는 언어학적 단위라기보다는 관습적인 단위인데 이를 기준으로 주석을 달리하는 것은 적절하지 않을 것이다. 이에 비해 언어학적으로 더 엄밀하게 정의할 수 있는 형태소를 단위로 할 경우 주석의 일관성을 보장할 수 있다.

3.3 토큰 주석과 범위 주석

주석의 기본 단위를 정하였다면 그것이 무엇인가에 관계없이 TimeML에서는 하나의 주석 단위가 토큰으로 다루어진다. 형태소 단위 주석을 하기로 결정되었다면 형태소가 하나의 토큰이 될 것이다.

이제 결정해야할 문제는 하나의 토큰에 TimeML 태그를 주석할 것인가, 여러 개의 토큰의 묶음에 하나의 TimeML 태그를 주석할 것인가의 문제이다. 영어 TimeML 1.2의 경우 다음과 같이 주석이 이루어졌다.

- ⟨EVENT⟩ 는 하나의 토큰에 주석
- ⟨TIMEX3⟩ 는 여러 토큰으로 이루어진 범위에 주석

영어에서 사건을 하나의 토큰에만 주석하도록 한 것은 행내 주석 방식을 채택했기 때문에 생긴 어쩔 수 없는 결과이다. 여러 개의 토큰으로 이루어진 사건이 분명 존재하나 그것이 텍스트 상에서 불연속적으로 나타나는 경우가 있기 때문에 핵어(head)에만 주석하기로 결정한 것이다. 이에 비해 시간 표현은 연속적인 형태로 나타나므로 범위(span)에 주석을 하였다.

한국어의 경우 형태소 단위 주석, 어절 단위 주석에 따라 토큰에 주석하는 경우와 범위에 주석하는 경우를 예시하면 다음과 같다.

(14) 어절 단위 주석
a. 먹고 있었다 (토큰에 주석)
b. 먹고 있었다 (범위에 주석)

(15) 형태소 단위 주석
a. 먹-고 있-었-다 (토큰에 주석)

b. 먹-고 있-었-다 (범위에 주석)

본 연구를 위하여 어절 단위 토큰 주석과 형태소 단위 토큰 주석을 시험적으로 수행하여 보았다. 어절 단위 토큰 주석은 대부분의 경우 문제가 없었으나 앞서 논의한 바와 같은 형태들이 텍스트에 출현하여 문제가 되었다. 형태소 단위 토큰 주석은 사건의 속성을 담고 있는 문법형태소들이 주석에서 제외되는 데에도 불구하고 그 속성을 태그에는 넣어주어야 한다는 점에서 주석자들에게 크게 인지적인 부담을 주었다.

본 연구의 시험적 주석에서는 미국 브랜다이스 대학에서 개발한 온라인 TimeML 주석 도구인 BAT[2]를 이용하였다. 이 시스템은 단어 단위 행내 주석 방식을 채택한 영어 TimeBank 구축에 적합한 인터페이스를 가지고 있어 형태소를 기본 단위로 하는 시험적 주석에는 사용의 어려움이 있다. 형태소 단위의 격리 주석 방식을 채택한 한국어 TimeML 주석을 위하여 새로운 주석 시스템이 개발 중에 있다.

4. 사건

사건(⟨EVENT⟩)은 TimeML의 기본 요소로서 본 연구에서 가장 중요한 것이다. 동시에 개별 언어마다 TimeML 명세화가 크게 달라질 수 있는 요소이다. 한국어 TimeML의 ⟨EVENT⟩ 태그를 위한 속성(attribute)과 값(value)들에 대해서 Im and Saurí (2008) , 임서현 외 (2009) 등에서 기본틀이 제시된 바 있다. 이 절에서는 한국어 텍스트의 사건을 TimeML의 틀에 따라 분석하는 데에 있어 고려해야할 주요한 사항들에 대해 논의하고자 한다.

4.1 사건의 판별

본 연구에서 어떤 것을 사건으로 분석하고 어떤 것은 사건으로 분석하지 않는지를 판단하는 것은 가장 기본적으로 논의되어야할 문제이다. 이에 대한 논의가 필요한 것은 본 연구의 목적이 텍스트 수준의 사건을 분석하는 데에 있어 일반적으로 널리 연구가 이루어진 어휘의미론적 관점의 사건 연구와는 방향을 달리 하기 때문이다. 이 절에서는 어떤 표현을 사건으로 판별하는가의 문제를 어휘의미론적 부류와 대비하여 차이점을 드러내어 논의하고자 한다.

사건은 기본적으로 무엇인가 일어나거나 발생하는 것을 지칭한다(Saurí et al., 2009). 문법적으로는 동사로 나타나는 것이 전형적인 사건에 해당하며 '회의', '조사', '연구' 등과 같은 행위성 명사도 사건을 표현할 수 있다. 사건은 다음 예 (16)과 같이 순간적이거나 예 (17)과 같이 지속적인 것일 수 있다.

(16) a. 정 전명예회장 일행에 앞서 오전 9시 47분 기자들이 먼저 중감위 회의

[2] http://www.timeml.org/site/bat/

실에 도착하였다.

b. 정 전명예회장을 제외한 나머지 현대 수행원들도 오전 10시 4분께 모두 북측 지역에 들어갔다.

(17) a. 대한적십자사 전방 사무소 , 군 관계자 등과 함께 방북 절차 등을 논의하는 등 분주히 움직였다.

b. 30여명의 북측 관계자들 가운데는 10여명의 기자들이 포함되어 있었다.

사건은 순간적이거나 일정 기간 동안 지속될 수 있는 것으로 일상적인 의미에서 '어떤 사건이 발생하는 것'뿐만 아니라 상태와 주변상황을 서술하는 것도 사건으로 파악한다. 그러나 모든 상태성 서술어를 ⟨EVENT⟩로 표지하는 것은 아니다(Saurí et al., 2006). 다음 예 (18)과 같이 텍스트의 흐름 속에서 뚜렷한 시간성을 가지는 상태 또는 상황 또한 사건에 해당한다.

(18) a. 이대로 가다가는 9월부터는 직원 월급조차 주기 어렵다.

b. 새해 첫 회의였지만 내용은 부실했다.

위의 예들은 어휘의미론적으로도 전형적인 사건에 해당하는 표현에 해당한다. 그러나 이러한 표현들이라 하여도 텍스트의 흐름 속에서 시간성을 가지지 않는 것은 사건으로 판별하지 않는다. ISO-TimeML에서 이것은 총칭적인 표현을 주석하지 않는다는 원칙으로 제시된다. 총칭적 표현을 주석하지 않는다는 것은 시간성을 가지지 않는 상태, 시간 표현과 직접 관련이 없는 사건, 텍스트의 흐름에 따른 변화를 확인할 수 없는 사건 등을 주석하지 않는다는 것을 뜻한다. 예를 들어 다음과 같은 예들이 어휘의미론적으로는 사건이지만 텍스트 수준에서는 사건으로 분석되지 않는 것에 해당된다.

(19) (형용사를 사건으로 판별하지 않는 예)

a. 수학에서 '불가능하다거나 존재하지 않음'을 보이는 것은 일반적으로 매우 어렵다.

b. 인터넷에 들어가 보면 정보는 넘쳐나지만 내용이 부실하다.

(20) (동사를 사건으로 판별하지 않는 예)

a. 시속 30킬로미터가 될 때까지는 100% 전기모터로만 움직입니다.

b. 프로그램 매수로 현물 매수물량이 늘어나면 당연히 주가는 오른다.

보조용언, 경동사, 형식용언 등은 다른 술어를 논항으로 취하는 구성의 경우에는 각각을 별개의 사건으로 판별한다. ISO-TimeML에서는 이것을 최소 덩어리(minimal chunk)의 원칙으로 제시하고 있다. 이러한 구성은 해당 사건의 상적인 속성 또는 사실성(factivity)과 증거성(evidentiality)에 영향을 미치므로 분리하여 분석하고자 하는 것이다.

(21) a. 그러나 둘째주 들어서면서부터 OPEC 회원국들은 기다렸다는 듯이 감산합의 연장의사를 밝히었고 유가는 반등하기 시작하였다.

b. 남은 문제는 감산합의가 언제까지 가느냐이지만 현재로선 OPEC 회원국들이 감산합의를 쉽게 풀 것 같지 않아 보인다 .

위의 예 (21a)에서 '시작하였다'는 '반등하기'라는 사건에 상적인 의미를 부여하는 사건이다. 예 (21b)에서 '같지 않아'는 '풀'로 표현된 사건에 부정적 증거(negative evidential)가 되고 '보인다'는 '같지 않아'에 증거성(evidential)를 부여한다. '풀 것 같지 않아 보인다'를 하나의 사건으로 분석하지 않고 이렇게 개별 요소로 나누어 분석하는 것은 여기에서 얻어진 형태-통사적 정보를 사건의 시간 의미 해석에 이용하고자 하는 의도를 가지고 있기 때문이다.

지금까지 살펴본 동사와 형용사의 사건 판별에 비해 명사의 사건 판별은 더 복잡한 양상을 보인다. 특히 한국어는 사건이 명사로 표현되는 경우가 매우 빈번하여 실제 말뭉치 분석에 있어 다양한 문제를 야기한다.

앞서 논의한 바와 마찬가지로 본 연구는 어휘의미론적인 사건에 대한 연구가 아니라 텍스트의 흐름에서 뚜렷한 시간성을 가지는 사건에 대한 연구이므로 한 명사가 어휘적으로 어떤 사건을 지칭하는 것이 분명해도 텍스트 내에서 시간성이 확보되지 않은 것은 사건으로 분석하지 않는다.

(22) (텍스트 내에서 시간성이 확보된 사건 명사)

a. 67년 8월 해병 5대대 25중대원으로 베트남전에 참전했던 작가 황석영씨도 "5중대는 해병대원들 사이에 유명했다"고 회상했다.

b. 국내 한반도 문제 전문가들은 대체로 올해 남북 정상회담이 열릴 것이라고 예측했다.

(23) (텍스트 내에서 시간성이 확보되지 않은 사건 명사)

a. 베트남전의 민간인 학살 문제는 참전군인들이 감당하기에는 너무나 무거운 짐이다.

b. 우리가 이에 대해 정략적 차원에서 대응한다면 남북 정상회담의 수혜자는 북한이 될 수밖에 없다.

위의 예 (22)에서 '베트남전'은 '참전했던'이라는 사건을 통해 '67년 8월'이라는 시간 표현에 묶여 있으며 '남북 정상회담'은 '열릴'이라는 사건을 통해 '올해'라는 시간 표현에 묶여 있다. 이에 비해 (23)의 '베트남전'과 '남북 정상회담'은 그것이 어휘론적으로는 어떤 사건을 가리키는 것이 분명함에도 텍스트 내에서 어떠한 시점 또는 기간에 묶여있지 않다. 이러한 차이에 의해 TimeML에서는 (22)은 사건으로 판별하지만 (23)은 사건으로 판별하지 않는다.

물론 우리는 (23)의 '베트남전'이라는 사건이 언제 발생한 것인지 알 수 있으나 그러한 정보는 텍스트 내에서는 얻어질 수 없다. 텍스트 외부에 있는 실세계에 대한 백과사전적 지식을 참조할 때에만 그러한 정보를 얻을 수 있다. 본 연구는 텍스트에 내재된 시간 정보에 대한 연구로서 그러한 실세계에 대한 지식의 참조는 배제되며 따라서 그 시간성이 확보되지 않는 '베트남전'은 사건으로서의 의미가 없고 분석 대상이 되지 않는다.

이러한 특정 사건을 지칭하는 명사 표현과 함께 일반적으로 행위를 가리키는 명사들도 사건으로 분석한다. 앞서 논의한 형용사, 동사의 판별과 동일하게 총칭적인 표현은 주석하지 않는다는 원칙을 따른다. 이러한 원칙 아래 한 가지 논의를 필요로 하는 것은 명사가 연쇄된 경우의 사건 판별의 문제이다. 한국어는 명사 연쇄가 빈번하다는 특성을 가지며 특히 행위성 명사가 연쇄되는 경우도 많다. 다음 예를 보자.

(24) 일부 전문가들은 OPEC의 감산 연장 확정에 따라 올 겨울 국제 유가가 배럴당 30달러선을 넘어설 가능성이 있다고 관측했다.

여기에서 '감산', '연장', '확정'을 개별 사건으로 분석한 것은 '감산하는 것을 연장하는 것을 확정하다'와 평행하게 최소의 사건을 개별적으로 분석한 것이다. 그러나 한국어 텍스트에서 이러한 사건의 판별은 지나치게 많은 사건을 과생성할 수 있다는 문제를 가지고 있다. 다음 예는 하나의 길지 않은 문장에 12개의 가능한 사건 표현이 나타나있다.

(25) 이날 노조는 KEC자본 규탄$_1$ 및 체포$_2$ 강행$_3$ 경찰책임자 처벌$_4$ 촉구$_5$ 결의$_6$ 대회$_7$에 참가한$_8$ 뒤 G20$_9$ 규탄$_{10}$ 집회$_{11}$에 함께 한다$_{12}$.

본 연구에서는 시험적인 말뭉치 분석에서 원칙대로 가능한 위와 같이 사건을 판별하였다. 이와 달리 'KEC자본 ... 결의 대회'를 더 이상 분석할 수 없는 하나의 사건을 지칭하는 명사로 파악하고 'G20 규탄 집회' 또한 하나의 사건으로 파악한다면 다음과 같이 4개의 사건만이 구별된다.

(26) 이날 노조는 KEC 자본 규탄 및 체포 강행 경찰책임자 처벌 촉구 결의 대회$_1$에 참가한$_2$ 뒤 G20 규탄 집회$_3$에 함께 한다$_4$.

위의 예 (25)와 같이 명사 연쇄의 가능한 모든 사건 표현을 사건으로 판별하는 것과 (26)과 같이 명사 연쇄로 이루어진 표현을 하나의 사건으로 판별하는 것은 이론적인 논의로만 결정될 수 있는 문제가 아니다. TimeML은 실제로 분석 말뭉치를 구축하여 텍스트 분석에 이용하기 위한 목적을 가지고 있는데 (25)와 같이 많은 비용을 필요로 하는 분석이 유용한 정보를 제공하지 않다면 의미가 없을 것이다. 본 연구에서는 '책임자 처벌 촉구 결의 대회'와 같은 연쇄를 개별 사건으로 분석하는 것이 '처벌을 촉구하다' 등의 구성과 마찬가지로 '처벌' 사건에 대해 '촉구' 사건이 부여하는 양상의 의미에 대한 정보를 제공한다는 점에서 유용하다고 판단하여 (25)와 같은 입장을 취하였다.

사건이나 행위를 지칭하지 않더라도 상태나 속성을 나타내는 명사는 형용사와 유사한 기준으로 사건으로 판별될 수 있다. 전형적으로는 '이다, 아니다, 있다, 없다' 등의 술어와 함께 쓰이는 경우에 시간성이 분명하게 드러난다. 다음 예에서 볼 수 있듯이 동일한 어휘가 맥락에 따라 달리 판별될 수 있다.

(27) (텍스트 내에서 시간성이 드러나는 상태 또는 속성)

a. 그러나 대선에서 '반김영삼 정서'가 팽배한 TK(대구 · 경북) 민심을 돌리기 위해서도 사면을 해야한다는 것이 당내 전반적인 분위기 이다.

b. 한국의 채권 신용도가 가까운 시일 안에 상향될 가능성 은 거의 없다.

(28) (텍스트 내에서 시간성이 드러나지 않는 상태 또는 속성)

a. 그뒤 3대째 의사인 가정 분위기 에 따라 치과의사의 길을 걷게 됐지만 음악에의 열정만은 포기하지 않았다

b. 경력이나 돈은 부족하지만 발전 가능성 을 보고 젊은 사람을 키워준다는 의미가 돼야 한다.

어휘의미론적으로 사건을 나타내는 것이 아니더라도 텍스트 내에서 시간적인 속성을 분명히 가지고 있을 때는 사건으로 분석해야 한다. 구체 명사의 경우에도 그것이 분명한 시간성을 드러내는 경우에는 사건으로 분석해야 한다.

(29) a. 4년 전 노무현 후보가 대통령 에 취임했을 때 보수세력은 그를 "나이는 50대 후반이지만 의식은 여전히 386 운동권"이라며 매우 못마땅해했다.

b. 윤 장관은 이날 오후 콜린 파월 미 국무장관과 한-미 외무장관회담을 하기 직전, 파월 장관과 함께 백악관을 방문해 20분간 부시 대통령을 만났다.

위의 예 (29a)의 '대통령'은 '4년 전' 일어난 상태의 변화를 서술하는 명사이므로 사건으로 판별되지만 (29b)의 '대통령'은 특정 인물을 지칭하는 것이므로 사건이 아니다.

단위 명사를 포함하여 어떠한 양을 표현하는 경우에도 맥락에 따라 판단하여야 한다. 단순히 양을 나타내는 것은 사건이 아니지만 특정 시점의 결과로서 상태를 의미할 때에는 사건으로서의 의미가 있다.

(30) (특정 시점의 상태를 나타내는 양적 표현)

a. 지난해 순익 전망치도 기존의 주당 1.98달러에서 2.12달러로 상향했다.

b. 23일 서울 외환시장에서 원달러 환율이 1154원을 기록 중이다.

(31) (단순히 양을 나타내는 표현)

a. 연말 할리데이 시즌 매출이 전년 동기에 비해 17퍼센트 증가했다고 밝혔다.

b. 원달러 환율은 22일 오전 10시 23분 전날보다 16.9원이 급등했다.

위의 예 (30)에서 '1.98달러', '2.12달러', '1154원' 등은 어떤 시점에서 상태를 나타내는 것이지만 (31)에서 '17퍼센트', '16.9원'은 변화의 양을 나타내는 것일 뿐이다. 따라서 (30)은 사건으로서의 의미가 있지만 (31)은 사건으로 파악될 수 없다.

4.2 사건의 속성

사건으로 판별된 표현을 TimeML의 틀에 따라 분석하기 위하여 사건의 속성을 주석하기 위한 체계를 마련해야 한다. TimeML에서 사건의 속성은 분석 대상 개별 언어의 문법적인 체계를 따르는 것을 기본으로 한다. ISO-TimeML에서 사건을 위한 기본적인 틀을 제시하고 있기는 하나 언어마다 문법 체계가 다르므로 ⟨EVENT⟩ 태그의 속성(attribute)과 속성값(value)은 개별 언어마다 크게 달라진다. 이 절에서는 한국어 TimeML을 위한 사건의 속성과 속성값을 논의하고자 한다.

한국어 TimeML의 사건의 속성 명세화에 있어서 가장 문제가 되는 것은 이론적으로 한국어 시제, 상, 법의 문법적 체계에 대한 단일한 합의가 존재하지 않는다는 점이다. 본 연구의 목적은 기존의 연구와 다른 새로운 시상 체계를 제시하는 데에 있지 않으며 선행 연구들의 성과를 최대한 반영하여 실질적인 언어 자원을 구축하기

```
attribute ::= id  pos class type [aspect] tense [modality]
              [mood] vForm [sTyep]
id        ::= IDREF
IDREF     ::= e<INTEGER>
pred      ::= CDATA
class     ::= 'OCCURRENCE' | 'ASPECTUAL' | 'STATE' | 'PERCEPTION' |
              'REPORTING' | 'I_STATE' | 'I_ACTION'
pos       ::= 'ADJECTIVE' | 'NOUN' | 'VERB' | 'OTHER'
polarity  ::= 'NEG' | 'POS'
type      ::= 'STATE' | 'PROCESS' | 'TRANSITION'
aspect    ::= 'PROGRESSIVE' | 'PERFECTIVE' | 'DURATIVE'
tense     ::= 'PAST' | 'NONE'
modality  ::= 'CONJECTURAL'
mood      ::= 'RETROSPECTIVE'
vform     ::= 'SENTENCE_FINAL' | 'CONNECTIVE' | 'ADNOMINAL' |
              'NOMINALIZED' | 'NONE'
stype     ::= 'DECLARATIVE' | 'INTERROGATIVE' | 'IMPERATIVE' |
              'PROPOSITIVE'
```

[그림 1] 한국어 ⟨EVENT⟩ 태그의 속성과 속성값

위한 틀을 확립하는 데에 있다. 따라서 시제, 상, 법(서법), 양상(양태)에 대한 선행 연구들을 종합하여 선행 연구들의 교집합을 이루면서 단순한 체계가 되도록 하였다.

개별 언어의 특성을 포착하기 위하여 ISO-TimeML은 기본적으로 정의된 속성 이외에 필요한 속성을 도입할 수 있도록 허용하고 있다. 한국어 TimeML을 위하여 사건의 속성으로 시제(tense), 상(aspect), 법(mood), 양상(modality), 부류(class), 품사(pos), 동사의 형태(vform), 문장의 유형(stype), 극성(polarity)을 도입하였다. 사건 태그의 속성은 그림 1에 정리하여 제시하였다. ISO-TimeML에는 유형(type)이라는 속성이 존재하여 한국어 TimeML에도 수용하였으나 본 연구에서는 다루지 않았다. 본 연구가 진행 중일 때에는 그 속성이 아직 명확하게 정의되지 않은 상태였고 다른 언어의 TimeML과 TimeBank에서 이러한 속성이 분석된 바 없었기 때문이다. 사건 태그를 예시하면 다음과 같다.

(32) 한나라당의 계파 공천 갈등이 더욱 날카로워지고 있다$_{e1}$.

```
<EVENT id="e1" tense="NONE" aspect="PROGRESSIVE"
       class="OCCURENCE" pos="VERB"
       vform="SENTENCE_FINAL" stype="DECLARATIVE"
       polarity="POS" />
```

4.2.1 시제와 상. 사건 분석에 있어서 TimeML은 표면형만 고려하여 주석한다. 시제는 문법적인 시간을 분석하는 것으로서 실제 물리적인 시간의 해석은 시제와 일치하지 않는다. ISO-TimeML에서는 기본적인 시제의 속성값으로 'NONE'과 함께 'FUTURE', 'PAST', 'PRESENT', 'IMPERFECT'를 제시하고 있다. 한국어 TimeML에서는 표면형만을 고려하며 형태론적으로 시제 체계를 파악한다는 입장에서 '과거'와 '비과거'의 시제 체계를 도입하여 형태소 '-었-'이 나타난 경우에는 과거, 그렇지 않은 경우에는 비과거로 분석하였다. 관련 국어학의 전반적인 논의는 최동주 (1998)를 참조할 수 있다.

시제가 발화 중심 또는 어떤 기준점에 사건의 시간을 연결시키는 것이라면, 문법적인 상은 화자가 사건의 구조를 표상할 수 있게 해주는 것이다. ISO-TimeML에서는 상의 속성값으로는 'NONE'과 함께 'PROGRESSIVE', 'PERFECTIVE', 'IMPERFECTIVE', 'PERFECTIVE PROGRESSIVE', 'IMPERFECTIVE PROGRESSIVE'를 기본으로 제시하고 있으며 이외의 값들은 개별언어에 따라 기술할 수 있도록 자유를 부여하고 있다. 한국어에서는 '-어 있-'으로 표현되는 지속상, '-고 있-'으로 표현되는 진행상, '-었었-'으로 표현되는 완료상을 도입하였다. 예시하면 다음과 같다.

(33) 조지 W.부시 미국 대통령은 7일(현지시간) 북한 핵 문제해법에 대해 " 모든 선택이 열려 있다$_{e1}$"고 말해$_{e2}$ 어떤 조치도 배제하지 않고 있음$_{e3}$을 시사했다$_{e4}$.

```
<EVENT id="e1" tense="NONE" aspect="DURATIVE" ... />
<EVENT id="e2" tense="NONE" aspect="NONE" ... />
<EVENT id="e3" tense="NONE" aspect="PROGRESSIVE" ... />
<EVENT id="e4" tense="PAST" aspect="NONE" ... />
```

(34) 지난해의 경우 국내 주식에 신규로 4조7천억원, 해외 주식에 3천억원을 투자했었다$_{e1}$.

```
<EVENT id="e1" tense="PAST" aspect="PERFECTIVE" ... />
```

4.2.2 법과 양상. 명제에 대한 화자의 인지적 태도를 범주화하는 데에 사용되는 법, 서법, 양상, 양태 등은 논의마다 용어와 개념에 있어서 차이를 보이고 있다(장경희 (1998) 참조). 본 연구에서는 TimeML의 'mood'를 '법'으로, 'modality'를 '양상'으로 지칭하였다. 법과 양상은 언어에 따라 크게 차이가 나므로 TimeML에서는 개별 언어에 따라 필요한 속성과 값을 도입하여 사용할 수 있도록 하고 있다. 한국어 TimeML에서는 '-더-'로 표현되는 회상법을 도입하였고 '-겠-'으로 표현되는 추측의 양상을 도입하였다. 관련된 국어학의 전반적인 논의는 장경희 (1998)를 참조할 수 있다.

(35) 김 원장은 원상복구 방법에 대해 "기존의 감사에서 실시했던$_{e1}$ 것과 같은 자료와 방법을 사용해 직불금 부당수령 관련 전산자료를 복구하겠다$_{e2}$"고 설명했다.

```
<EVENT id="e1" tense="PAST" aspect="PERFECTIVE"
                mood="RETROSPECTIVE" ... />
<EVENT id="e2" tense="NONE" modality="CONJECTURAL" ... />
```

4.2.3 사건의 부류. TimeML의 ⟨EVENT⟩ 태그는 부류(class)라는 속성을 가지고 있다. 이것은 사건 표현의 부류를 나누어 시간 관계 분석에 유용한 정보를 제공하기 위한 목적을 가지고 있다. 이 사건 부류는 어휘론적인 의미 부류의 관점에서는 적절하게 파악될 수 없다. 이 부류는 두 사건이 통사적으로 관계를 맺는 구성에서 시간 관계의 유형에 대한 부류이다.

사건 부류는 모두 6가지로 나누어진다. 대부분의 사건은 발생을 나타내는 부류와 상태를 나타내는 부류에 해당한다. 다른 사건을 논항으로 취하여 시간적 의미를 부여하는 경우에는 상적 부류에 속한다. 다른 사건을 보거나 듣는 등의 사건은 지각 부류에 속하고 어떠한 정보를 말하거나 알리는 등의 사건은 보고 부류에 속한다. 다른 사건을 논항으로 취하여 가능 세계를 도입하는 사건은 내포 행위 또는 내포 상태 부류에 속한다. 각 부류에 속할 수 있는 어휘의 예를 들면 다음과 같다.

(36) a. 발생(OCCURRENCE) 부류: 오르다, 빗나가다, 곤두박질치다, 와해되다
b. 상태(STATE) 부류: 다르다, 의미하다, 높다; 사실, 부족, 불가능
c. 상적(ASPECTUAL) 부류: 계속하다, 재개하다, 끝나다, 시작하다
d. 지각(PERCEPTION) 부류: 지켜보다, 보이다
e. 보고(REPORTING) 부류: 말하다, 주장하다, 전하다, 밝히다
f. 내포 상태(I_STATE) 부류: 모르다, 바라다, 생각하다
g. 내포 행위(I_ACTION) 부류: 요구하다, 해소하다, 극복하다, 철회하다

위의 예시는 해당 부류에 해당하는 일반적인 어휘를 예시한 것일뿐 반드시 위와 같이 부류가 결정되는 것은 아니다.TimeML은 어휘의미론적인 분석이 아니라 텍스트 수준의 분석이므로 어휘에 따라 부류가 결정되지 않는다. 동일한 어휘가 맥락에 따라 다른 부류가 될 수 있다. 다음 예들은 '보다' 동사가 맥락에 따라 단순히 보는 행위의 발생(occurrence)로 분석될 수도 있으며 다른 사건에 대한 지각(perception)이나 내포 상태(intensional state)로 분석될 수도 있음을 보이고 있다.

(37) a. 휠체어를 타고 공연을 보았다는 한 환자는 "오랜 병상생활 동안 모처럼 즐겁고 신나는 기분을 느꼈어요 ... (OCCURENCE).

b. 그런데 국방부는 방위사업청이 개청되고 나서 효율성이 얼마나 저하되었다는 것인지, 한 번도 근거를 제시하는 것을 보지 못했다. (PERCEPTION)

c. 그런데도 세계의 시각은 우리나라의 민주정치가 완전하다고 보았다. (I_STATE)

형태-통사론적으로 유사한 구성을 하여도 다른 부류가 될 수 있다. 다음 예시에서 볼 수 있듯이 '-다고' 구성을 취한다고 하여 항상 보고(reporting) 부류가 되는 것은 아니다.

(38) a. 이를 엠지에이에 넘겼으므로 저작권을 침해했다고 고소했다. (OCCURRENCE)

b. 하는 글을 지속적으로 올려 5·18 유공자들의 명예를 훼손했다고 주장했다. (REPORTING)

위의 예에서 '고소했다'는 '저작권을 침해했다'는 정보를 알리는 사건이 아니다. '저작권을 침해했다'는 사건은 '고소했다'는 사건의 이유로 제시된 것이다. 따라서 '고소했다'는 사건은 단순한 발생 부류에 해당한다. 이에 비해 '주장했다'는 '명예를 훼손했다'는 정보를 알리는 사건이므로 보고 부류가 된다.

발생 부류와 상태 부류는 텍스트에 사건이 홀로 드러난 경우에 해당되는 부류로서 전자에는 전형적인 동사가 해당되며 후자에는 전형적인 형용사가 해당된다. 지각 부류와 보고 부류는 절 또는 구로 표현된 다른 사건을 논항으로 취하는 구성에서 나타나는 부류로서 전자에는 '보다', '듣다', '느끼다' 등의 소수의 어휘가 속하며 후자에는 발화 행위를 나타내는 대부분의 동사가 속한다.

상적 부류는 다른 사건을 논항으로 취하여 상적 단면을 기술하는 사건으로 시간 관계 분석에서 상적 연결(⟨ALINK⟩)로 분석될 대상이 된다. 그 상적의미에 따라 '시작하다', '재시작하다', '끝내다', '완료하다', '계속하다'로 나눌 수 있다.

내포 행위(I_ACTION)와 내포 상태(I_STATE)는 다른 사건을 논항으로 취하여 가능 세계를 도입하는 행위 또는 상태 사건의 부류이다. '약속하다', '거부하다'와 같은 명시적인 수행 술어가 내포 행위의 전형적인 예에 해당하며 '알다', '모르다', '믿다', '확신하다' 등 인지 또는 사고에 관련된 동사들은 전형적인 내포 상태에 해당한다.

(39) 정부는 주민이 원하지 않으면 원자력 관련 시설을 짓지 않겠다고$_{e1}$ 약속하였으나$_{e2}$, 울진군 주민 다수가 처분시설 유치$_{e3}$를 원하는$_{e4}$ 경우에도 이를 못하게$_{e5}$ 가로막는$_{e6}$ 것은 주민의 의사에 반하여 불이익을 주는 것으로, 합당하지 않다고$_{e7}$ 판단된다$_{e8}$.

```
<EVENT id="e1" class="OCCURRENCE" ... />
<EVENT id="e2" class="I_ACTION" ... />
<EVENT id="e3" class="OCCURRENCE" ... />
<EVENT id="e4" class="I_STATE" ... />
```

```
<EVENT id="e5" class="OCCURRENCE" ... >/
<EVENT id="e6" class="I_ACTION" ... />
<EVENT id="e7" class="STATE" ... />
<EVENT id="e8" class="I_STATE" ... />
```

4.2.4 기타 문법적 속성. 사건은 시간적 속성과 밀접한 관련이 있는 시제, 상, 법(서법), 양상(양태) 이외에도 품사(pos), 동사의 형태(vform), 문장의 유형(stype), 극성(polarity) 등의 문법적 속성을 가지고 있다. 한국어 TimeML에서 이러한 속성들이 가질 수 있는 값은 앞의 그림 1에 제시하였다.

이들 중 극성(polarity)은 부정(NEG) 또는 긍정(POS)을 지칭하는 것으로 사건의 의미 해석에 있어서 중요한 역할을 한다.

품사(pos) 속성은 사건으로 표지된 구(phrase)의 통사 범주를 구별하기 위한 것이다. 한국어 TimeML에서 사건의 품사로는 동사, 명사, 형용사를 구별하고 있으며 이외의 품사로 사건이 표현되었을 때에는 기타(OTHER)로 분석하도록 하였다.

동사와 형용사의 경우 다양한 어미와 결합하여 평서문, 의문문, 청유문, 명령문과 같은 문장의 유형을 결정하기도 하고 주절, 접속절, 명사절, 관형절, 부사절 등과 같은 유형을 결정하기도 한다. 동사 형태(vform) 속성은 필수 속성으로 문장종결어미, 연결어미, 명사형어미, 부사형어미의 결합에 따라 그 값을 결정하고 문장유형(stype) 속성은 평서문, 의문문, 명령문, 청유문을 구별하여 값을 부여한다.

5. 시간 표현

시간 표현은 사건과 함께 TimeML의 기본을 이루는 요소이다. 사건의 경우와 달리 시간 표현은 개별 언어의 문법적인 특성에는 크게 좌우되지 않는다. 한국어 TimeML에서도 ISO-TimeML의 시간 표현 ⟨TIMEX3⟩ 태그를 그대로 따를 수 있어 새로운 속성과 속성값의 도입을 필요로 하지는 않는다. 시간 표현 태그의 형식은 그림 2에 제시되어 있다(Pustejovsky et al. (2003) 참조). 그러나 한국어 연구에 있어 사건에 대한 선행 연구는 풍부한 데에 비해 시간 표현에 대한 연구는 깊이 이루어지지 않아 ISO-TimeML의 틀를 그대로 한국어에 적용하기 위해서도 기초적인 논의를 필요로 한다.

5.1 시간 값의 해석

시간 표현의 분석에서 가장 중요한 것은 그 시간 값(value)을 해석하는 문제이다. 시간 값이란 물리적인 시간 상의 위치를 말하는 것이며 TimeML에서 형식적으로는 그레고리력에 따른 날짜와 시간의 표기에 관한 국제 표준 규격인 ISO 8601:2004로 표현한다. 예를 들어, 2011년 3월 31일은 '2011-03-31'로 표현한다. 문제는 텍스트에 나타나는 시간 표현이 언제나 이렇게 명시적으로 시간 값을 알려주지 않는다는

```
attributes ::= tid type [functionInDocument] [temporalFunction]
               (value | valueFromFunction)
               [mod] [anchorTimeID | anchorEventID]
tid      ::= TimeID
TimeID   ::= t<integer>
type     ::= 'DATE' | 'TIME' | 'DURATION' | 'SET'
functionInDocument  ::= 'CREATION_TIME' | 'EXPIRATION_TIME' |
                        'MODIFICATION_TIME' | 'PUBLICATION_TIME' |
                        'RELEASE_TIME' | 'RECEPTION_TIME' | 'NONE'
temporalFunction  ::= 'true' | 'false'
{temporalFunction  ::= boolean}
value  ::= CDATA
{value  ::= duration | dateTime | time | date | gYearMonth | gYear |
           gMonthDay | gDay | gMonth}
valueFromFunction    ::= IDREF
{valueFromFunction  ::= TemporalFunctionID
TemporalFunctionID   ::= tf<integer>}
mod ::= 'BEFORE' | 'AFTER' | 'ON_OR_BEFORE' | 'ON_OR_AFTER' |
        'LESS_THAN' | 'MORE_THAN' | 'EQUAL_OR_LESS' | 'EQUAL_OR_MORE' |
        'START' | 'MID' | 'END' | 'APPROX'
anchorTimeID    ::= TimeID
anchorEventID ::= EventID
```

[그림 2] 시간 표현의 속성과 값

데에 있다. 다음 예는 2000년 1월 16일 한겨레신문의 한 온라인 기사의 첫 번째 단락이다.

(40) 국제 원유가격이 새해 벽두부터 가파르게 뛰고 있다. 뉴욕시장 원유가는 14일 배럴당 28달러선을 넘어섰다. 91년 걸프전 이후 가장 높은 수준이다. 뉴욕시장 유가는 지난 한 주 사이에 배럴당 $3 \cdot 35$달러나 올랐다. 이런 추세라면 국제 유가가 곧 30달러를 돌파할 것이란 전망도 만만치 않다. (2000-01-16)

이 기사의 작성시간 2000년 1월 16일을 참조하여 '새해 벽두'가 2000년 초임을 알 수 있고 '14일'이 2000년 1월 14일임을 알 수 있다. 우리는 '91년'은 일반적으로 '1991년'의 줄임 표현이라는 것도 잘 알고 있다. 달력 정보를 이용하여 '지난 한 주'는 2000년 1월 9일에서 2000년 1월 15일에 걸친 주임을 알 수 있다. 문서 작성 시점 정보를 통해 시간 표현들의 값을 해석할 수 있다. 동일한 신문 기사의 이어지는 두 번째 단락을 보자.

(41) 유가가 뛰는 이유는 한 가지다. 석유수출국기구(OPEC) 회원국들이 오는 3월말로 끝나는 감산합의를 연장하려 하기 때문이다. 실제로 유가는 새 해 첫주엔 겨울철 임에도 불구하고 계속 내렸다.

'오는 3월말'은 미래의 시간으로 '2000년 3월말'임을 알 수 있다. 맥락 상 2001년 또는 그 이후의 해의 3월말로 해석되지는 않을 것이다. 그러한 경우에는 '내년 3월말' 등의 표현이 사용되었을 것이다. '새 해 첫주'는 과거의 시간으로 '1월 2일에서 1월 8일'에 걸친 주를 지칭하는 것임을 알 수 있다. 만약 이 텍스트가 2000년 12월말에 작성된 것이고 '새 해 첫주에는 ... 할 것이다'라는 문장 속에 사용되었다면 2001년의 첫 번째 주로 해석될 것이다. 또한 '겨울철'이 '새해 첫주'와 동일한 시간을 지칭하는 것임도 이해할 수 있다. 이와 같이 시간 표현의 해석을 위해서는 통사적인 정보를 포함하여 맥락에 대한 이해가 필요하다.

위의 예를 통해 시간 표현의 시간 값의 해석 과정이 단순하지 않음을 보았다. 문제가 되는 것은 주어진 시간 표현에 온전한 시간 정보가 드러나지 않은 경우임을 알 수 있었다. 즉, 시간 값을 처리하는 관점에서 시간 표현을 다음과 같이 크게 두 가지로 나누어 다룰 수 있다.

(42) 완전 명세된 시간 표현
2000년, 1987년 6월, 1941년 정월 11일

(43) 미명세된 시간 표현
지난해 12월, 지난해 여름, 작년 7월, 20년 전, 어제밤, 연초

완전 명세된 시간 표현은 시점에 대한 정보가 독립적으로 명확하게 표현된 경우로 시점값이 그대로 정해진다. 미명세된 시간 표현은 현재 시점에서 상대적인 위치를 통해 표현된 것으로 그 시점값을 알기 위해서는 추론이 필요하다.

주어진 시간 표현이 완전 명세된 것인가, 미명세된 것인가 구별하는 것은 시간 값을 해석하는 데에 있어 중요한 역할을 한다. 특히 미명세된 시간 표현의 의미 해석은 자연언어처리에서 필수적으로 요구되는 것 중의 하나이며 TimeML이 그 연구 기초를 제공하고자 하는 과제의 하나이다.

시간 값의 해석에서 또 문제가 되는 것은 언어-문화적인 특성이 반영된 표현들을 해석할 때 생기는 문제이다. 일반적인 시간 표현으로 '몇년 몇월 몇일 몇시 몇분'의 형태로 표현되는 물리적인 시간값은 언어-문화에 종속적이지 않으므로 ISO-TimeML의 틀을 그대로 사용하여 분석하는 데에 문제가 없다. 그러나 일부 한국어의 특징적인 시간 표현에 대한 속성값(value) 연구가 필요하다. 예를 들어, 하루 중의 시점을 나타내는 '아침', '점심', '저녁', '밤', '새벽' 등의 표현은 언어마다 차이가 있기 때문에 적절한 속성값이 새로 도입되어야할 여지가 있다. 한 달 중의 기간을

[표 2] 시간 표현의 유형

type	한국어의 예
TIME	하루 중 시간 정오, 오후 3시, 저녁 때
DATE	다양한 단위의 날짜 표현 일 (어제, 2010년 1월 8일, 지난 주 금요일, ...) 주 (다음 주, 10월 세째 주, ...), 월 (두 달 후, 1973년 11월), 계절 또는 분기 (지난 봄, 3사분기, ...) 연도 (1967, 이듬 해, ...) 세기, ...
DURATION	기간 두 달, 5시간
SET	집합 매주 금요일, 매달 첫째 일요일

나타내는 '초순', '중순', '하순' 등도 영어를 기초로 한 ISO-TimeML에서는 고려되지 않아 새로운 표현 기제가 필요하다.

5.2 시간 표현의 유형과 속성

시간 표현을 주석하기 위한 태그의 이름 ⟨TIMEX3⟩는 시간 표현(time expression)의 3번째 버전이라는 의미를 가지고 있다. TimeML이 제안되기 이전에 텍스트에 나타나는 시간 표현들을 주석하기 위한 연구가 존재하였다. 그러한 연구 성과 중의 하나인 ⟨TIMEX2⟩를 TimeML에서 수용하여 ⟨TIMEX3⟩로 발전시킨 것이다. ⟨TIMEX3⟩로 분석될 대상은 시간, 날짜, 기간 등의 명시적인 시간 표현이다.

ISO-TimeML에서는 시간 표현의 유형을 표 2에 제시된 바와 같이 크게 4가지로 나누고 있다. 특정 시점을 가리키는 시간 표현은 하루 중의 시간을 나타내는 것으로 시간(TIME) 유형, 하루 이상의 단위를 가리키는 날짜(DATE) 유형으로 나누어진다. 쉽게는 시간 유형은 시계를 통해 정보가 제공되는 것, 날짜 유형은 달력을 통해 정보가 제공되는 것으로 이해할 수 있다. 시간의 범위를 나타내는 시간 표현은 기간(DURATION) 유형에 해당한다. 마지막으로 집합(SET) 유형은 전형적으로는 반복되는 시점 또는 기간의 묶음을 지칭하는 것으로 단일 시점 또는 기간이 아닌 복합적인 시간 표현을 위한 유형이다.

우선, 단순한 예로 날짜(DATE) 유형과 기간(DURATION)에 속하는 시간 표현의

예를 살펴보자. 하루 이상의 단위는 모두 날짜 유형에 속한다. 아래 예에서 볼 수 있듯이 value 속성에 시간값을 명시하며 형식은 ISO 8601:2004를 따른다.

(44) 이에 학생들은 2008년 3월$_{t1}$ 복학해 학교를 마저 다녔고 3명은 졸업까지 했으나, 학교가 2009년 4월$_{t2}$ 뒤늦게 이들이 법정 공방으로 학교를 다니지 못한 2년$_{t3}$에 대해 '무기정학' 처분을 하겠다고 통보해오자 소송을 냈다.

```
<TIMEX3 id="t1" type="DATE" value="2008-03" />
<TIMEX3 id="t2" type="DATE" value="2009-04" />
<TIMEX3 id="t3" type="DURATION" value="P2Y" />
```

하루 이내의 시간 단위가 명시된 경우에는 시간(TIME) 유형으로 분석한다. '몇 시 몇 분'이라고 시간값이 명시된 경우에는 다음과 같이 해당 시간값을 주석한다.

(45) ... 한국미술협회 자문위원인 조각가 연제동 씨가 제작해 제56주년 광복절인 2001년 8월 15일 오전 11시$_{t1}$에 제막했다.

```
<TIMEX3 id="t1" type="TIME" value="2001-08-15T11:00" />
```

시간이 정확히 제시되지는 않았더라도 '아침, 점심, 저녁, 낮, 밤, 정오, 자정' 등 하루의 일부를 나타내는 시간 표현에 대해서는 지정된 기호를 이용하여 분석한다. 다음 예는 '밤'이라는 시간 표현을 ISO-TimeML에 따라 'NI'로 주석하는 예와 함께 미명세된 시간 표현의 분석을 예시하고 있다. 미명세된 시간 표현도 인간 주석자가 그 값을 해석하여 정확한 값을 주석한다.

(46) 유홍준 문화재청장이 11일 밤$_{t1}$ 불에 타버린 숭례문을 둘러본 뒤 현장을 떠나고 있다. (기사등록 : 2008-02-11 오후 09:15:49$_{t0}$ 기사수정 : 2008-02-12 오후 05:13:36)

```
<TIMEX3 id="t0" type="TIME" value="2008-02-11T21:15:49"
        functionInDocument="CREATION_TIME" />
<TIMEX3 id="t1" type="TIME"
        value="2008-02-11TNI" temporalFunction="true" />
```

이 예에서 '11일 밤'이라는 표현만으로는 그 시간 값을 알 수 없으나 문서 생성 시간, 발행 시간, 변경 시간 등의 정보를 이용하여 그 값을 계산할 수 있다. TimeML 주석 시에 인간 주석자가 텍스트의 의미를 이해한 후 시간 값을 분석하여 주석하고 temporalFunction 속성을 참 값으로 주어 미명세된 표현의 시간 값을 계산할 필요가 있음을 명시한다. 이러한 정보는 추후에 기계 학습 등에 이용될 수 있다.

집합(SET) 유형은 여러 시점 또는 기간의 집합을 표현하는 것으로 '매일, 매주, 달마다, 해마다' 등 반복 시간 표현이 전형적으로 이에 속한다. 이러한 반복 시간 표현을 분석하기 위한 기제로 양화(quant) 속성이 존재한다.

(47) 해마다 10월$_{t1}$ 이면 과학 담당 기자들이 한 자리에 모여 벌이는 익숙한 풍경이 되풀이된다.

```
<TIMEX3 id="t1" type="SET" value="XXXX-10" quant="EVERY" />
```

집합 유형은 양화(quant) 속성 이외에 빈도(freq) 속성과 다양한 수식(mod) 속성을 이용하여 추가 정보를 분석할 수 있다. 다음 예에서 '두 번'은 빈도로 분석되었으며 '정도'는 APPROX 속성값으로 분석되었다.

(48) 공동체 구성원들은 한 달에 두 번 정도$_{t1}$ 공동체 회관에서 전체회의를 열고, 매주 수요일과 금요일엔 회관에 모두 모여 저녁식사를 함께 해오면서

```
<TIMEX3 id="t1" type="SET" value="P1M" freq="2X" mod="APPROX" />
```

5.3 시간 표현의 판별

시간 표현도 사건과 마찬가지로 텍스트에서 시간 정보로서의 의미가 있는지를 판별할 필요가 있다. 어휘의미론적으로 시간적인 의미를 가지고 있다고 하여 모두 시간 표현으로 분석하는 것은 텍스트 수준의 시간 정보를 분석하는 데에는 의미가 없다.

문맥에 따라 달라질 수 있으나 대개 어휘적으로 '주석 가능(markable)' 표현과 '주석 불가능(non-markable)' 표현을 구별할 수 있다. 주석 가능 표현은 명시적인 시간 정보를 드러내는 경우를 지칭한다. 다음과 같은 어휘들이 그에 해당한다.

(49) 주석 가능 표현
- a. 분, 시, 오후, 자정, 낮, 주말, 주일, 월, 여름, 분기, 년, 년대, 세기, 학기, 미래, 과거, 시간, 기간, 종일, 주간
- b. 월요일, 크리스마스, 석가탄신일, 단오
- c. 8:00, 2010/11/17, 1985
- d. 최근의, 이전의, 현재의, 미래의, 과거의
- e. 매일, 매달, 매월, 매시, 매년, 매해
- f. 지금, 오늘, 어제, 내일

주석 불가능 표현은 그것이 개념적으로는 시간적인 의미를 가지는 어휘라고 하더라도 명시적인 시간값을 가질 수 없는 경우를 말한다. 주석 불가능 표현을 몇 가지 유형으로 나누어 살펴보자.

우선, 사건의 연쇄 또는 순서를 표현하는 '그 동안에, 마침내, 결국, 이어서, 그 후에, 미리, 머지 않아' 등은 시간 표현으로서 의미가 없다. 이들은 사건과 사건의 관계를 표지하는 요소이기는 하나 그 자체가 어떤 시점에 대한 정보를 제공하는 것은 아니기 때문이다. 아래 예시된 밑줄 친 표현이 그러한 예이다.

(50) a. ... 김경준씨와 결별했다"며, 그 이후에 본격적으로 이뤄진 주가조작과는 무관하다고 주장해왔기 때문이다.

b. 선거는 앞으로 2년이나 남았기 때문에 민주당의 대선 후보 구도를 섣불리 예측할 수는 없다. 그 동안에 무명 인사들이 나타나 대통령에 당선할 가능성도 충분하다는 게 미국의 정치 분석가들의 전망이다.

c. 정부가 마이크로소프트(MS) 환경에서만 작동해 스마트폰 등 모바일에서는 무용지물이던 액티브엑스(X)를 마침내 걷어내기로 했다.

시간적 의미를 가지는 양태 부사 '즉시, 바로, 잠시, 잠깐' 등도 시간 표현으로서 의미가 없다. 역시 특정 시점 또는 특정 기간에 대한 정보를 제공하는 표현이 아니기 때문이다. 아래 예시된 표현들은 TimeML에서 시간 표현으로 분석하지 않는다.

(51) a. 또 모든 상임위를 즉시 가동해 법안을 심의하고 오는 18일부터 예산결산특위를 열어 지난해 결산 심사를 진행하기로 했다.

b. 최근 들어 텔레비전에 잠깐 출연했다가 잠재력을 인정받아 연예계로 데뷔하는 경우가 늘고 있다.

이와 함께 명시적인 시간 표현인 것처럼 보이더라도 정량화되지 않은 '다년간의, 수년간, 몇 년 동안, 오랫동안' 등 역시 '잠시, 잠깐' 등과 동일한 이유에서 시간 표현으로 분석하지 않는다.

(52) a. 우리는 앞으로도 오랫동안 그를 생각해야 할 것 같습니다.

b. 최근 수년간의 자연재해와 화폐개혁 실패, 외부의 식량지원 중단, 작황 부진 등으로 평양에서도 식량 배급이 불안정할 ...

또 다른 시간 부사로 명시적인 시간 정보를 제공하지 않는 빈도 부사 '자주, 빈번히, 항상, 평소에, 늘, 종종' 등도 정량적인 정보를 제공하지 않으므로 시간 표현으로 분석할 수 없다.

(53) a. 조사결과 기형가축이 빈번히 발생했다는 충격적인 사실이 밝혀졌다.

b. 평소에 잘 놀고 건강하던 아이들도 오랜만에 장거리를 이동해 여러 친척들을 만나 함께 어울리고 뛰어놀다 보면 ...

이외에 '시간'이라는 표현이 상태나 상황을 지칭하는 경우 또는 명시적인 시간 표현이지만 그것이 고유 명사의 일부로 사용되는 경우 역시 TimeML의 분석 대상이 될 수 없다.

(54) a. 이들은 개런티는 물론 교통비조차 받지 않고 어려운 시간을 쪼개 연주를 들려준다.

b. 팔월의 크리스마스 (영화 제목)

6. 시간 관계

본고의 주제인 TimeML의 최종적인 목적은 시간 관계를 분석하는 데에 있다. 시간 정보의 기본 요소인 사건과 시간 표현을 분석한 후에야 텍스트에 나타나는 시간 관계를 분석할 수 있으므로 사건과 시간 표현에 대한 연구를 요하는 것이다.

시간 관계는 기본적으로는 시간적 선후 관계를 말하는 것이다. TimeML에서는 Allen (1983), Allen (1984)의 'interval algebra'에서 제시된 다음과 같은 13가지 관계에 의거하여 시간 관계를 분석하고 있다.

	$X = Y$	
$X < Y$	(X before Y)	$Y > X$
X m Y	(X meets Y)	Y mi X
X o Y	(X overlaps Y)	Y oi X
X s Y	(X starts Y)	Y si X
X d Y	(X during Y)	Y di X
X f Y	(X finishes Y)	Y fi X

여기에서 $X = Y$는 두 사건 X, Y가 동일한 시간을 가짐을 뜻한다. 사건 X가 Y에 선행하는 $X < Y$는 Y가 X에 후행하는 $Y > X$와 동일하므로 선행과 후행은 짝을 이룬다. 다음으로 X가 Y 직전인 X m Y는 Y가 X의 직후인 Y mi X와 동일하여 직전(m)과 직후(mi)는 짝을 이룬다. 나머지 관계들도 마찬가지이다. 여기에서 'i'는 역(inverse)을 의미한다.

이러한 시간 관계의 분석은 논리적인 체계에 따른 것으로 언어보편적이다. 따라서 어떤 개별 언어를 위한 TimeML이라고 하여도 ISO-TimeML과 달라질 수 없다. 본 연구의 한국어 TimeML에서도 시간 관계를 위한 태그는 ISO-TimeML의 명세를 그대로 수용하였다.

그러나 시간 관계 연구가 논리적인 분석에만 국한된 것은 아니다. TimeML은 자연 언어의 텍스트에서 시간 관계를 파악하는 것이므로 시간 관계의 분석은 개별 언어의 특성에 대한 분석을 필요로 한다. 예를 들어, 한국어에서는 사건 사이를 연결하는 '-면서', '-자마자', '-을 때', '-기 전에' '-고 나서' 등의 접속어미 또는 접속기능구성이 담고 있는 정보를 시간 관계 해석에 이용할 수 있도록 TimeML에서 주석해야할 것이다. 이러한 요소를 주석하기 위한 것이 ⟨SIGNAL⟩ 태그이다. 이 태그는 시간 표현과 사건 사이에 존재하는 관계를 명시적으로 나타내는 기능적인 언어요소를 주석한다.

TimeML에서 분석하고자 하는 것이 동시, 선후, 겹침, 포함 등의 물리적인 시간 관계임이 분명하나 이러한 시간 관계만 분석하는 것은 아니다. TimeML에서 이러한 시간 관계는 시간적 연결 태그인 ⟨TLINK⟩로 주석이 되며 이외에도 사건 사이의

관계를 주석하기 위한 상적 연결 태그 ⟨ALINK⟩와 종속적 연결 태그 ⟨SLINK⟩가 존재한다. 이는 TimeML이 물리적인 시간 관계 또는 순수하게 논리적인 시간 관계를 분석하기 위한 것이 아니라 자연 언어 텍스트에 나타나는 시간 관계를 분석하기 위한 것이기 때문이다. 상적 연결과 종속적 연결은 언어적인 시간 관계를 나타내며 최종적으로는 시간적 연결로 해석되기 위한 중간 장치이다.

6.1 상적 연결

사건 사이의 관계 중에서 가장 쉽게 분석될 수 있는 상적 연결(aspectual link)은 상사건(aspectual event)과 그 논항 사건 사이의 관계를 나타낸다(Pustejovsky et al., 2003, pp.8–9). ISO-TimeML에서는 한 사건이 다른 사건에 부여하는 상적인 의미를 5가지로 나누고 있다. 전형적인 한국어 어휘를 예로 들어 제시하면 다음과 같다.

(55) a. INITIATES: 시작하다
b. CONTINUES: 계속하다
c. TERMINATES: 멈추다, 중단하다
d. CULMINATES: 끝내다, 완료하다
e. REINITIATES: 재개하다

한 사건의 부류(class)가 ASPECTUAL인 경우 전형적인 상적 연결의 관계를 가지게 된다. 한국어에 존재하는 상적 의미를 부여하는 다양한 보조동사들이 이에 해당한다. 그러나 반드시 보조동사 구성으로 분석이 제한이 되는 것은 아니다. 사건 명사를 논항으로 취하여 상적 의미를 부여하는 경우 또한 상적 연결로 분석이 된다. 아래 상적 연결의 예를 보였다.

(56) 북한이 개성 공단과 금강산 관광 지구의 출입을 차단한지 하루 만인 10일 출입$_{e1}$을 재개했다$_{e2}$

```
<EVENT id="e1" class="OCCURENCE" tense="NONE" ... />
<EVENT id="e2" class="ASPECTUAL" tense="PAST" ... />
<ALINK eid="e2" relatedToEvent="e1" relType="REINITIATES" />
```

(57) 북한은 정확도가 향상된 미사일 개발$_{e1}$을 계속하고 있다$_{e2}$.

```
<EVENT id="e1" class="OCCURENCE" tense="NONE" ... />
<EVENT id="e2" class="ASPECTUAL" tense="NONE"
       aspect="PROGRESSIVE" ... />
<ALINK eid="e2" relatedToEvent="e1" relType="CONTINUES" />
```

6.2 종속적 연결

상적 연결 이외에 한 사건이 다른 사건을 논항으로 취하는 경우 두 사건의 관계는 종속적 연결(subordination link)로 분석된다. 자연 언어에서 이러한 구성은 어휘

요소가 다른 사건의 시간성에 영향을 주는 구성으로 다양한 양상(modality) 의미를 부여한다. 특히 논항 사건의 사실성(factuality) 또는 증거성(evidentiality)에 영향을 주는 구성에 대한 연구는 정보 추출, 텍스트 이해, 질의-응답 시스템 등의 발전에 필수적인 과제 중의 하나이다. 다음 예를 보자.

(58) a. 진주는 작년에 미스코리아 선발대회에 나간 것을 후회한다.

b. 그는 진주가 작년에 미스코리아 선발대회에 나갔다고 믿는다.

위의 예에서 '후회한다'는 종속절의 사건에 사실성을 부여하는 데에 비해 '믿는다'가 부여하는 양상의 의미는 그렇지 않다. 위와 같은 정보가 주어졌을 때 다음 질문에 대한 답은 달라질 것이다.

(59) 진주는 작년에 미스코리아 선발대회에 나갔는가?

사건의 사실성 정도에 관한 정보 없이 이와 같은 질문에 답하는 질의-응답 시스템은 불가능하다. 언어학적인 측면에서도 종속적 연결의 연구는 사건의 사실성 정도와 관련된, 양태, 사실성/반사실성, 증거성 등의 의미론, 화용론적 연구를 다루므로 중요하다.

종속적 연결은 이러한 의미를 분석하기 위한 것이다. TimeML에서는 종속적 연결의 유형을 다음과 같이 나누고 있다. 한국어의 예시와 함께 보이면 다음과 같다. 어휘가 아닌 문법 표지들도 종속적 연결의 대상이 될 수 있다.

(60) a. MODAL: 믿다, 원하다, 가능하다
b. EVIDENTIAL: 말하다, 시인하다
c. NEG_EVIDENTIAL: 부인하다, 부정하다
d. FACTIVE: 후회하다
e. COUNTER_FACTIVE: 금지하다
f. CONDITIONAL: '-면'

위 예에서 전형적인 술어의 예를 들었으나 한국어의 경우에는 종속절의 문법 형태에 따라 의미가 달라지므로 반드시 위와 같이 어휘적으로 유형이 결정되는 것은 아니다.

종속적 연결의 분석 예를 보이면 다음과 같다. 한국어 텍스트는 여러 개의 절로 이루어진 긴 문장이 빈번하게 이용되며 따라서 다음과 같이 종속적 연결 또한 빈번하게 나타난다.

(61) 최재천 의원은 "통외통위가 미군기지 반환 문제를 짚어보는 비공개 청문회를 의결해 놓았다$_{e1}$"며$_{e2}$ "여야 합동으로 이 청문회를 열어 미군기지 환경오염 문제를 따지는$_{e3}$ 것이 가능하다$_{e4}$"고 설명했다$_{e5}$.

```
<EVENT id="e1" class="ASPECTUAL" tense="PAST" ... />
<EVENT id="e2" class="REPORTING" tense="NONE" ... />
<SLINK eid="e2" subordinateEvent="e1" relType="EVIDENTIAL" />
<EVENT id="e3" class="OCCURRENCE" tense="NONE" ... />
<EVENT id="e4" class="I_STATE" tense="NONE" ... />
<EVENT id="e5" class="REPORTING" tense="PAST" ... />
<SLINK eid="e4" subordinateEvent="e3" relType="MODAL" />
<SLINK eid="e5" subordinateEvent="e4" relType="EVIDENTIAL" />
```

6.3 시간적 연결

상적 연결과 종속적 연결은 사건과 사건 사이의 관계에 한정된 관계인 데에 비해 시간적 연결(temporal link)는 사건과 사건, 사건과 시간 표현, 시간 표현과 시간 표현 사이에 존재하는 시간적 관계를 나타낸다. 또, 상적 연결과 종속적 연결이 자연 언어에 표현되는 시간 의미를 분석하기 위한 장치인 데에 비해 시간적 연결은 물리적인 또는 논리적인 시간 관계를 분석하기 위한 장치이다. TimeML의 시간적 연결은 기본적으로 Allen (1983)이 제시한 체계를 따르고 있다. 시간들 사이의 관계에는 기본적으로 7가지 유형이 존재하며 그 중 동일 관계를 제외한 나머지 6가지 유형은 짝을 이룬다. 따라서 모두 13가지 관계가 가능하다. 이러한 유형의 분류를 기초로 하여 시간 관계들이 가지는 속성이 연구되었다(Verhagen, 2005, 8). ⟨TLINK⟩의 시간 관계 분석 체계는 다음과 같다.

(62) a. IDENTITY
b. SIMULTANEOUS
c. BEFORE, AFTER
d. IAFTER, IBEFORE
e. INCLUDES, IS_INCLUDED
f. DURING, DURING_INV
g. BEGINS, BEGUN_BY
h. ENDS, ENDED_BY

이 중 'IDENTITY'는 텍스트 내에서 동일 사건을 지칭하는 다른 표현들 사이의 관계를 나타내는 것으로 서로 다른 두 사건이 동시에 일어나는 'SIMULTANEOUS'와는 구별된다.

텍스트에 나타나는 모든 사건과 시간표현 사이의 시간 관계를 분석하고 시간적 연결을 주석하는 것은 기하급수적으로 많은 수의 연결을 생성하게 되므로 현실적으로 불가능하다. 따라서 한 문장 내에서는 구조적으로 위계 관계에 있는 사건 사이에 시간적 연결을 주석하고 문장과 문장 사이에는 주절의 사건들 사이에만 시간 관계를 분석한다. 한 문장 내에서의 사건 사이의 시간 관계 분석을 예시하면 다음과 같다.

(63) 검찰도 이점을 미리 계산해$_{e1}$ 사전구속영장을 청구하기$_{e2}$ 전에$_{s1}$ 2차례에 걸쳐 공개소환요구를 했다$_{e3}$.

```
<EVENT id="e1" class="OCCURRENCE" tense="NONE" ... />
<EVENT id="e2" class="OCCURRENCE" tense="NONE" ... />
<SIGNAL id="s1" />
<EVENT id="e3" class="OCCURENCE" tense="PAST" ... />
<TLINK lid="l1" eid="e3" relatedEvent="e1"
       relType="AFTER />"
<TLINK lid="l2" eid="e3" relatedEvent="e2" signalID="s1"
       relType="BEFORE" />
```

위 예에서 '계산해'는 구조적으로 '공개소환요구를 했다'에 걸리므로 두 사건 사이의 관계를 ⟨TLINK⟩로 분석하였고 '청구하기 전에'와 '공개소환요구를 했다' 사이의 관계 역시 ⟨TLINK⟩로 분석하였다. 반면에 '계산해'와 '청구하기' 사이에는 위계적 관계가 없으므로 시간 관계를 분석하지 않았다. 그러나 분석된 시간 관계를 통해 '계산 < 공개소환요구 < 청구'의 시간 순서를 추론할 수 있다.

7. 결론

본 연구에서는 한국어 텍스트의 사건 및 시간 정보를 분석하기 위한 기본 틀로서 TimeML을 논의하였다. 한국어 TimeML을 위한 세부적인 명세화에 대한 논의와 함께 실제 한국어 텍스트를 분석하는 과정에서 제기될 수 있는 문제점을 논의하였다. 이 논문에서 다루어진 문제점들은 단지 머릿속에서 개념적으로 제기된 것이 아니라 여러 명의 연구자들이 실제로 시험적인 말뭉치 분석을 수행하는 작업을 통해 실질적으로 의미가 있다고 판단된 문제들이라는 점에서 의미가 있다.

이 논문에서는 자세히 다루지 않았으나 본 연구는 실제 텍스트에 TimeML을 주석하여 시간 정보 분석 말뭉치인 TimeBank를 구축하고 이를 이용하여 사건 및 시간 표현을 자동으로 분석하고 시간 추론을 할 수 있는 자연언어처리 도구를 개발하는 큰 연구의 일부이다. 후속 연구를 통하여 TimeBank의 구축과 분석을 수행하고 이를 이용하여 자연언어처리 도구를 개발하는 연구를 진행할 것이다.

< 참고문헌>

Allen, J. F. 1984. Towards a general theory of action and time. *Artificial Intelligence*, 23.2: 123–154.

Allen, James F. 1983. Maintaing knowledge about temporal intervals. *Communications of the ACM*, 26.11: 832–843.

Im, Seohyun and Roser Saurí. 2008. Multilingual Challenge to TimeML: A Korean Case Study. *Proceedings of the 18th Conference of International Linguists.*

Im, Seohyun, Hyunjo You, Hayun Jang, Seungho Nam, and Hyopil Shin. 2009. KTimeML: specification of temporal and event expressions in Korean text. In *Proceedings of the 7th Workshop on Asian Language Resources*, pp. 115–122, Suntec, Singapore. Association for Computational Linguistics.

Katz, G. and F. Arosio. 2001. The Annotation of Temporal Information in Natural Language Sentences. *In: Proceedings of ACL-EACL 2001, Workshop for Temporal and Spatial Information Processing.*

Lee, K., J. Pustejovsky, and B. Boguraev. 2006. Towards an international standard for annotating temporal information. In *Third International Conference on Terminology, Standardization and Technology Transfer, Beijing, China, ISO TC/37 and SC (August.*

Moldovan, Dan, Christine Clark, and Sanda Harabagiu. 2005. Temporal Context Representation and Reasoning. *International Joint Conferences on Artificial Intelligence, 1099-1105.*

Pustejovsky, J., José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. *IWCS-5, Fifth International Workshop on Computational Semantics.*

Pustejovsky, James, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *IREC 2010.*

Saurí, Roser, Lotus Goldberg, Marc Verhagen, and James Pustejovsky, 2009. *Annotating Events in English: TimeML Annotation Guidelines.*

Saurí, Roser, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky, 2006. *TimeML Annotation Guidelines.*

Setzer, Andrea. 2001. *Temporal Information in Newwire Articles: an Annotation Scheme and Corpus Study.* Ph.D. thesis, University of Shefield, UK.

Setzer, Andrea, Robert Galzauskas, and Mark Hepple. 2005. The Role of Inference in the Temporal Annotation and Analysis. *Language Resources and Evaluation (2005) 39: 234-265.*

Verhagen, Marc. 2005. Temporal Closure in an Annotation Environment. *Language Resources and Evaluation.*

김평, 성기윤, 맹성현. 2003. 사건 탐지/추적을 위한 시간 정보 추출. *제15회 한글 및 한국어정보처리 학술대회*에서, 22–29쪽.

남지순. 2008. 시간 표현 부사어구에 대한 유한 그래프문법. *한국어학*, 42: 155–191.

이기용. 2008. A Compositional Interval Semantics for Temporal Annotation. In *The 2008 PNU International Conference on Language & Knowledge Proceeding.* 부산대학교 인문한국 초청강연(2008년 4월 7일).

임서현, 김윤신, 조유미, 장하연, 고민수, 남승호, 신효필. 2009. KTARSQI: 한국어 텍스트의 시간 및 사건 표현. *제21회 한글 및 한국어 정보처리 학술대회*에서.

장경희. 1998. 서법과 양태. *문법 연구와 자료*에서. 태학사, 261–303쪽.

정영미, 김용광. 2008. 사건 중심 뉴스기사 자동요약을 위한 사건탐지 기법에 관한 연구. *정보관리학회지.*

최동주. 1998. 시제와 상. *문법 연구와 자료*에서. 태학사, 227–260쪽.

접수 일자: 2011년 5월 4일
게재 결정: 2011년 5월 30일