KTARSQI: 한국어 텍스트의 시간 및 사건 표현 주석

임서현○ 김윤신 조유미, 장하연, 고민수, 남승호, 신효필

브랜다이스 대학교 신라대학교 서울대학교

컴퓨터학과 국어교육과 언어학과

ish97@cs.brandeis.edu yoonshin@silla.ac.kr {jmocy84, hyan05, ryan0802, nam, hpshin}@snu.ac.kr

KTARSQI: The Annotation of Temporal and Event Expressions in Korean Text

Seohyun Im○ Yoon-shin Kim, Yoomi Jo, Hayun Jang, Minsoo Ko, Seungho Nam, Hyopil Shin

Department of Computer Science Brandeis University USA

Department of Korean Language Education Silla University Korea

Department of Linguistics Seoul National University Korea

요 약

정보추출(information extraction), 질의-응답 시스템(Question-Answering system) 등의 자연언어처리 응용분야에서 시간과 사건에 관련한 정보를 추출하는 것은 중요한 부분이다. 그럼에도 불구하고, 한국어의 자연언어처리 응용분야에서는 아직까지 이 연구가 본격화되지 않았다. 미국 TARSQI 프로젝트의 연구결과를 바탕으로 하여 한국어 텍스트에서 시간 및 사건 표현의 주석, 추출, 추론을 위한 명세 언어(KTimeML), 주석 말뭉치(KTimeBank), 자동 태깅 시스템(KTarsqi Toolkit: KTTK)의 개발을 목표로 2008년 KTARSQI 프로젝트가 시작되었다. 이 논문에서는 KTARSQI 프로젝트의 목표와 과제에 대한 전반적인 소개와 함께, 현재까지 진행된 작업의 결과로서 사건 태그의 명세와 주석에 관한 논의를 덧붙인다.

주제어: 시간 표현, 사건 표현, 주석, KTARSQI, KTimeML, KTimeBank

1. 서 론

시간(time)과 사건(event)표현의 주석과 그 정보의 추출은 질의-응답 시스템(Question-Answering system)을 비롯한 자연언어처리 응용분야에 있어 중요한 역할을 한다. 예를 들어 '김대중 전 대통령이 사망한 해'에 대한 질문에 질의-응답시스템이 적절한 응답을 하기 위해서는 '사망'이라는 **사건**이 발생한 **시간**정보의 추출이 매우 중요한 관건이다. 이 논문은 한국어 텍스트에서 시간과 사건표현의 주석, 추출과 추론을 위한 KTARSQI(Korean Temporal Awareness and Reasoning System for Question Interpretation) 프로젝트의 소개와 그 첫 단계인 사건표현의 주석에 관한 논의를 목적으로 한다.

KTARSQI 프로젝트는 미국 브랜다이스 대학교의 James Pustejovsky가 주도한 TARSQI[1] 프로젝트의 결과를 기반으로 한국어 텍스트에서 시간 및 사건표현 주석, 추출과 추론을 위한 마크업언어(KTimeML) 개발, 주석 말뭉치(KTimeBank)와 자동 태깅 시스템(KTTK)의 개발을 목표로 한다. 2008년 프로젝트가 시작된 이후 현재까지 KTimeML의 명세(specification) 구성과 형태소 분석기 등의 전처리 작업, 사건 태깅 시스템 개발, KTimeBank를 위한 주석 작업 등을 진행해 오고 있다.

이 논문의 구성은 다음과 같다. 다음 절에서는 TimeML에 관한 개략적인 소개를 한다. 3절에서는 KTimeML, KTimeBank, KTTK를 비롯한 KTARSQI에 관한 개관을 한다. 4절에서는 프로젝트의 첫 단계로서, 사건 표현 주석의 언어학적 기초와 주석 실험 결과의 분석을 설명한다. 마지막으로 결론과 향후 프로젝트 과제에 대하여 요약한다.

2. TimeML

TimeML은 텍스트에서 사건과 시간표현의 주석을

1) TARSQI 프로젝트는 미국 Brandeis 대학 컴퓨터학과의 James Pustejovsky가 이끄는 프로젝트로 시간 및 사건표현 마크업 언어(TimeML[1]), 주석 말뭉치(TimeBank[2]), 자동 태깅 시스템(Tarsqi Toolkit[3])의 개발을 하여 왔다. TARSQI에 대한 자세한 정보는 www.timeml.org 참조.

위한 명세 언어로서 일차적으로 영어를 그 자원으로 하여 개발되었다. 2008년 시간 및 사건표현 주석을 위한 ISO 표준(ISO-TimeML)으로 채택되었으며(ISO/DIS 24617-1: 2008)[2], 이태리어, 스페인어, 중국어, 프랑스어 등을 비롯한 여러 언어들에 TimeML을 적용하고자 하는 노력이 확산되고 있다. 2010년에 Semeval-2010의 일부로서 열리게 될 TempEval-2([4])에는 위의 언어들과 한국어가 함께 참여할 예정이다.

TimeML 주석의 기본 방침은 첫째, 단어를 주석의 기본단위로 하며 인라인(in-line) 주석방식을 채택한다. 둘째, 표층기반 주석(surface-based annotation)을 원칙으로 한다. 즉, 의미론적 해석에 따른 주석이 아니라 표층에 나타난 문법 범주 그대로 주석하는 것을 원칙으로 한다. 예를 들어 영어의 현재시제 표현이 미래를 지시하는 경우에도 현재시제로 표시한다.(예: *John meets his advisor tomorrow* -> tense="PRESENT") 셋째, 주석은 통사론적으로 해당 구(phrase)의 머리(head)만 주석한다.(head-only principle; 예를 들어 *John's* singing *of the song*에서는 *singing*만 주석한다.) 한국어의 경우에는 영어와 다른 유형론적 특성 때문에 이러한 TimeML의 주석 원칙을 그대로 적용할 수 없다. 이에 대한 구체적인 논의는 3절에서 다룬다.

TimeML은 주요 데이터 구조로서 **EVENT, TIMEX3, SIGNAL, LINK** 이 네 가지를 포함한다. EVENT 태그는 사건표현, TIMEX3는 시간표현 (날짜, 시간, 기간, 등)을 주석하는 데 쓰인다. SIGNAL 태그는 시간 표현들이 갖는 관계를 표시하는데 쓰인다(예: *when, after* 등). LINK 태그는 다시 TLINK(temporal link), ALINK(aspectual link), SLINK(subordinate link)로 나뉜다. TLINK는 사건-시간, 사건-사건, 시간-시간 간의 시간적 순서와 관계를 주석하는 태그이다. ALINK는 특별히 상동사(예: *begin, finish* 등)와 그 사건 지시 논항의 관계를 나타내는 데 쓰이며, SLINK는 증거성(evidentiality), 사실성(factivity), 내포성(intensionality) 등의 의미 정보를 전달하는 종속 관계(subordination)를 주석하기 위한 태그이다. TimeML에 대한 더 자세한 논의는 [1]과 [5]를 참조할 수 있으며, 여기서는 TimeML의 태그들이 고루 나타나는 예를 보이는 것으로 설명을 대신한다.

(1) John said$_{e1}$ that Mary began$_{e2}$ to work$_{e3}$

```
John
<EVENT id="e1" class="REPORTING" tense="PAST"
  aspect="NONE" polarity="POS">
said </EVENT>
that Mary
<EVENT id="e2" class="ASPECTUAL" tense="PAST"
  aspect="NONE" polarity="POS">
began </EVENT>
to
<EVENT id="e3" class="OCCURRENCE"
  tense="NONE" aspect="NONE" polarity="POS">
work </EVENT>

<TLINK eventID="e1" relatedToEvent="e2"
  relType="AFTER" />
<ALINK eventID="e2" relatedToEvent="e3"
  relType="INITIATES" />
<SLINK eventID="e1" subordinatedEvent ="e2"
  relType="EVIDENTIAL" />
```

위의 예문은 3개의 사건표현(*said, began, work*)이 포함되어 있는 예이다. *said*와 *began*의 증거성 관계를 TimeML은 SLINK로 나타낸다. ALINK는 상동사 *begin*과 *work*의 관계를 나타내며, 이 사건들 사이의 시간적 순서 혹은 관계를 TLINK를 통해 나타낼 수 있다.

3. Korean TARSQI (KTARSQI)

한국어의 자연언어처리 응용에 관련된 연구 분야에서는 시간 및 사건 표현 주석과 정보추출에 관한 연구가 거의 없다가, 최근 들어 한국어 시간, 사건 정보추출에 관한 연구가 시작되었다. 특히 [6]은 한국어 텍스트에서 Timex2를 이용한 시간표현 자동 태깅 시스템을 개발하였고[3], [7]은 TimeML의 한국어 적용에 관한 의미론적 해석을 시도하고 있으며, [8]은 TimeML을 한국어에 적용함에 있어서의 문제점과 개선방향에 관하여 논의하고 있다. 이러한 논의에 힘입어, 한국어 텍스트에서 시간 및 사건 표현 주석과 추출, 추론을 위한 통합적 시스템의 구축을 목표로 KTARSQI 프로젝트가 시작되었다. 이제 KTARSQI 프로젝트의 세 가지 과제에 대하여 간략히 설명한다. 3.1에서는 TimeML을 한국어에 이식하는 데 있어서의 문제점과 EVENT 태그를 중심으로 한 KTimeML의 명세를 간략히 기술한다. 3.2는 KTimeML을 증명하기 위한 골드 스탠더드(Gold-standard)로서 시간 및 사건표현 주석 말뭉치(KTimeBank)에 대하여 기술하고, 3.3에서는 KTimeML에 따른 시간 및 사건표현 자동 태깅 시스템의 개발을 소개한다.

3.1 KTimeML

한국어는 TimeML이 대상으로 하고 있는 영어와는 달리 교착어라는 유형론적 특성 때문에 형태소의 분석이 대부분의 자연언어처리 및 응용분야에 있어서 중요한 과제가 된다. 한국어를 위한 TARSQI 시스템의 적용에 있어서도 형태소와 관련된 주석 방식의 논의가 필수적이다. 다음의 예를 보자.

2) KTARSQI 프로젝트 팀과 ISO-TimeML의 책임자인 고려대학교 이기용 선생님이 KTimeML과 ISO-TimeML의 한국어 부분에 대하여 하나의 통일된 명세를 제시하기 위하여 논의하여 왔으며, 이러한 공동작업의 결과로서 KTimeML과 ISO-TimeML 한국어 부분은 동일한 명세를 갖고 있다.

3) Timex2는 TimeML 이전에 개발된 시간표현 주석을 위한 기술 언어이다. TimeML은 Timex2를 발전시킨 Timex3를 쓰고 있다. KTimeML은 Timex3를 따른다.

(2) John said$_{e1}$ that Mary came$_{e2}$

```
<SLINK eid="e1" subordinatedEvent="e2"
  relType="EVIDENTIAL" />
```

영어의 인용문은 주로 인용동사가 인용절을 이끄는 형태로 표현되며, 인용(예: *said*)과 인용대상(예: *came*)이 되는 두 개의 사건으로 표층에 표현된다. 따라서 TimeML은 이 두 사건의 관계를 SLINK를 통해 주석한다. 그러나 형태소의 결합이 의미 확장을 가져오는 한국어의 경우 인용문 구성이 형태소의 결합으로 나타나는 경우가 많다.

(3) 철수가 영희가 왔다면서 기뻐했다.

(3)의 '왔다면서' 에서 인용의 의미는 형태소 '면서' 가 대체한다. 따라서 단어를 기본 단위로 하면 인용과 인용대상 사건을 따로 나누어 주석할 수 없게 된다. 이러한 문제를 해결하기 위해서 KTimeML은 형태소를 주석의 기본단위로 취급한다. 또한 본래의 텍스트를 유지하고, 형태소묶음을 하나의 KTimeML 주석 단위로 취급하기 위하여 인라인(in-line)이 아닌 스탠드오프(stand-off) 주석을 지향한다. 다음은 형태소 기반의 스탠드오프(stand-off) 주석을 보여준다.

(4) 오-았-다-면서
1 2 3 4

```
<EVENT eid="e1" markable="1 2 3" tense="PAST"
  aspect="NONE" polarity="POS" />
<EVENT eid="e2" markable="4" tense="NONE"
  aspect="NONE" polairity="NONE" />

<SLINK eid="e2" subordinatedEvent="e1"
  relType="EVIDENTIAL" />
```

위의 예를 통해 한국어의 경우 형태소 기반의 주석이 KTimeML이 의도하는 시간 및 사건표현 주석의 목적에 비추어 타당함을 알 수 있다[4]. 주석자간 일치도를 높이기 위한 표층기반 주석과 상당부분 실용적 측면에서 고안된 머리만 주석하는 원칙은 KTimeML에서도 수용하고 있다.

요약하면 한국어 시간 및 사건표현 마크업 언어인 KTimeML의 주석은 형태소 기반 stand-off 주석, 표층기반 주석, 머리만 주석하는 원칙을 따른다[9]. 이 외에도 한국어의 특성을 고려하여 KTimeML은 EVENT 태그의 속성으로 TimeML보다는 ISO-TimeML을 따른다. 아래에 EVENT 태그의 BNF를 소개한다.

```
attributes ::= eid pred markable class pos tense
               [aspect] [mood] [sType] modality
               [vForm]
eid ::= EventID
EventID ::= e<integer>
markable ::= MORPHIDs
pred ::= CDATA
class ::= 'OCCURRENCE'|'ASPECTUAL'|'STATE'|
          'PERCEPTION'|'REPORTING'|'I_STATE'|
          'I_ACTION'
pos ::= 'ADJECTIVE'|'NOUN'|'VERB'|'OTHER'
tense ::= 'PAST'|'NONE'
aspect ::= 'PROGRESSIVE'|'PERFECTIVE'|
           'DURATIVE'|'NONE'
mood ::= 'RETROSPECTIVE'|'NONE'
         {default, if absent, is 'NONE'}
sType ::= 'DECLARATIVE'|'INTERROGATIVE'|
          'IMPERATIVE'|'PROPOSITIVE'| 'NONE'
          {default, if absent, is 'NONE'}
modality ::= 'CONJECTUAL'|'NONE'
             {default, if absent, is 'NONE'}
vForm ::= 'sFINAL'|'CONNECTIVE'|'NOMINALIZED'|
          'ADNOMINAL'|'NONE'
          {default, if absent, is 'NONE'}
polarity ::= 'NEG'|'POS'
             {default, if absent, is 'POS'}
```

EVENT 태그의 속성값 중 한국어와 관련하여 TimeML과 다른 속성과 그 값들만 간단히 설명하면, 형태소 기반 스탠드오프(stand-off) 주석이므로 markable속성이 추가되고, 한국어 형태소 '-더-' 가 갖는 의미 속성을 표현하기 위해 RETROSPECTIVE mood가 추가되었다. 또한, 어미 변형으로 문장의 서법을 나타내므로 sType 속성이 추가되었고, '자고 있다' 와 같은 상태지속을 나타내기 위해 aspect 속성값으로 DURATIVE가 추가된다. 한국어의 '-겠-' 은 보통 양상형태소(modal marker)로 취급되므로 modality 속성값으로 CONJECTUAL이 도입된다. 다음은 한국어 주석의 예이다.

(5) 어제 서울-은 비-가 오-았-겠-더-라 .
1 2 3 4 5 6 7 8 9 10 11

```
<EVENT eid="e1" markable="6 7 8 9 10" pred="come"
  pos="VERB" class="OCCURRENCE" tense="PAST"
  aspect="NONE"mood="RETROSPECTIVE"
  modality="CONJECTUAL"vForm="sFinal"
  sType="DECLARATIVE" polarity="POS"/>
```

이 논문에서는 EVENT 태그의 BNF만 소개하였다. 현재 KTimeML의 명세와 주석 가이드라인은 계속 개발 중에 있다. 3.2에서는 시간 및 사건표현 주석 말뭉치인

4) 형태소 기반 주석에 관하여 형태통사 단위 혹은 철자단위 기반 주석과 같은 이견이 있을 수 있으나 현재 한국어 자연언어처리의 연구수준에 비추어 형태통사 단위 기반 주석이 쉽지 않을 것으로 보이며, 반면 형태소 분석기의 개발수준은 상당히 높은 점 등 을 감안하면 전처리 등의 문제가 보다 쉽게 해결되는 것으로 보인다. 따라서 당분간 KTimeML은 형태소 기반 주석을 원칙으로 한다.

KTimeBank의 구축에 대하여 간략히 소개한다.

3.2 KTimeBank

TimeML의 타당성을 입증하기 위한 방편으로 개발된 TimeBank는 주로 신문기사로 구성된 시간, 사건표현 주석 말뭉치로서 자동 태깅 시스템인 TTK의 평가를 위한 Gold-standard로 기능하며 동시에 그 시스템의 기계학습을 위한 자원으로도 쓰이고 있다. 이와 마찬가지로 한국어의 KTimeML의 입증과 KTTK를 위한 한국어 주석말뭉치의 개발도 KTARSQI의 통합 시스템 개발을 위해 필수적이다.

일차 작업을 위해 구축한 KTimeBank는 세종 프로젝트에서 만든 한겨레 신문 말뭉치로부터 추출한 약 200개의 문서로 구성되어 있으며, 주로 국제정치면의 기사가 포함되어 있다. 현재 KTimeML의 명세, 가이드라인 작성과 함께 주석 작업이 진행 중이다. 지금까지 사건 태그의 주석에 집중하였고, 그 결과는 4절에서 보다 자세히 논의한다. 최종 결과물은 일차 작업의 완료가 되는 시점에 KTimeBank 1.1로 공개될 예정이다. 다음 단계로 말뭉치 크기의 확장과 기사 종류의 다양성 확보를 목표로 하고 있다(현재는 국제면 기사만 포함되어 있다). 다음은 시간, 사건표현 자동 태깅 시스템인 KTTK(Korean Tarsqi Toolkit)을 소개한다.

3.3 KTTK

KTTK는 KTimeML에 따라 한국어 텍스트에 나타나는 시간 및 사건 표현을 자동으로 주석하기 위한 자동 태깅 도구이다. 영어의 자동 태깅 시스템인 TTK와 가장 큰 차이점은 형태소 분석기인 Pykts를 이용하여 전처리를 한다는 점이다[10]. 이를 통하여 형태소 분석과 여러 가지 표준화(normalization)를 거친 다음 사건, 시간 표현 주석을 하게 된다. 그 구성은 `EVENT` 태거, `TIMEX3` 태거, `LINK` 태거가 중심이다. KTTK는 세종 프로젝트에서 구축된 데이터 자료 - 사건 표현, 시간 표현 등의 목록 -을 많이 이용하기 때문에 통사론적 정보 혹은 규칙 기반 시스템과 함께 데이터베이스의 이용이 강화된 형태가 될 것이다. 다음은 KTTK의 시스템 구조이다.

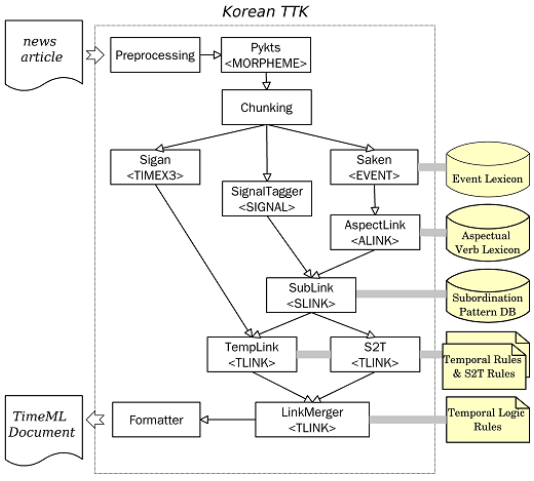

그림 1. KTTK 시스템 구조

현재 KTS 형태소 분석기를 이용한 전처리와 사건 주석 시스템인 Saken의 개발을 진행하고 있는 중이다. 이에 대한 보다 자세한 논의는 [11]을 참조할 수 있다. 4절에서는 KTimeBank의 일부로서 작업한 사건 태그 주석에 대한 논의를 설명한다.

4. EVENT 태그 주석 실험

4.1 주석

이번 주석 작업에서는 한국어 텍스트의 선택된 표현이 사건인가 아닌가를 판별하는 것에 초점을 맞추었으며, 사건의 속성값이나 다른 태그들에 대해서는 고려하지 않았다.

주석 말뭉치. KTimeBank를 위해 한겨레 말뭉치로부터 구축한 200개의 문서 중 100개의 문서를 대상으로 하였다. 형태소를 토큰으로 설정하였고, 약 10만개의 토큰이 포함된다.

전처리 사건 태그 주석을 위한 전처리로서 KTS를 이용한 형태소 분석을 하였고, 주석의 편의를 위한 청킹이나 오류 수정과 같은 과정은 거치지 않았다. 결과적으로 KTS의 형태소 분석 결과 중 문제가 되는 것은(1) 동사성 명사와 결합하는 '하다/되다'가 접미사, 경동사, 혹은 동사의 구분 없이 무조건 각각 개별 토큰으로 분리된다; (2) 하나의 사건임에도 불구하고 두 개 이상의 형태소로 구분되는 경우가 발생한다(예: 민주 화); (3) 띄어쓰기 오류 등으로 인하여 사건 주석이 어려운 경우가 있다. 이와 같은 문제점들은 이후 청킹의 도입과 전처리 결과의 자동 혹은 반자동 수정에 의하여 시정될 것이다.

주석자. 총 6명의 인원이 주석을 하였으며, 모두 언어학 배경을 가지고 (박사 1명, 박사과정 1명, 석사과정 4명), 주석 가이드라인 논의에 참여한 사람들로 구성되었다. 모두 주석 가이드라인 논의에 참여하였으므로 따로 주석 훈련 세션을 가지지는 않았다. 주석자간 일치(inter-annotator agreement) 측정을 위하여 2명이 동일한 문서를 각각 주석하였다.

주석도구. 주석을 위한 도구는 독자적으로 개발하지 않고, TempEval-2를 위하여 TARSQI팀에서 개발한 BAT (Brandeis Annotation Tool)[5]을 이용하였다. BAT에 입력으로 제공되는 단위는 형태소 단위이다.

4.2 사건 태그 주석의 언어학적 기초 및 가이드라인

TimeML과 마찬가지로 KTimeML은 사건을 행위, 시간과 관련된 일시적 상태 등을 포함하여 '발생하는 것' 혹은 '일어나는 것'으로 정의한다. 한국어에서 사건을 지시하는 표현은 일반적으로 동사, 일부 형용사, 동사성 명사구(서술명사), 명사화 표현(-음/기), 동사의 관형사형(-ㄴ/ㄹ) 등을 포함한다.

사건 주석 임무를 위하여 제시된 주석 가이드라인의 핵심 내용은 다음과 같다([12] 참조).

- **명사구 주석에 관한 지침**: 한국어 명사구는 조사가

5) BAT에 관한 자세한 논의는 www.timeml.org/site/bat 참조.

생략되고, 명사끼리 결합하는 복합 명사구가 많기 때문에 머리만 주석하는 원칙을 지키는 데 있어 상세한 가이드라인이 필요하다.

- **자격, 직위, 직업 등 명사에 관한 지침**: KTimeML은 함의(entailment) 관계를 처리할 기제가 없으나, 자격, 직위, 직업 등의 정보는 시간, 사건정보와 관련하여 중요한 정보이므로 주석자의 판단에 따라 주석이 필요하다고 생각되는 경우 주석하도록 하였다.
- 사건과 실체의 의미를 중의적으로 갖는 명사들은 사건으로 주석하도록 하였다.
- **동사성 명사와 경동사 '하다/되다'** : 한국어의 경우 경동사의 처리문제가 중요하다. 이번 작업에서는 경동사가 사건 태그 속성 값에 관한 정보를 모두 담고 있기 때문에 경동사를 따로 주석하는 TimeML의 방식을 따라 둘 다 주석하기로 하였다. 다만 현재 작업에서는 경동사의 범주에 '하다'와 '되다'만 포함한다. 이후 작업에서는 경동사성을 지닌 다른 동사들의 처리에 관하여 논의가 진행될 것이다.
- **동사 있다/없다**: 한국어의 동사 '있다'는 존재(exist)의 의미 외에도 영어의 BE 동사와 같은 기능도 포함한다. 따라서 주석할 때 존재의 의미를 지닌 경우를 제외하고는 주석을 하지 않는다.
- **보조동사군**(부정, 양태, 상 등): 한국어에는 부정, 상, 양태 등을 나타내는 보조동사군이 다양하게 분포되어 있다. 이번 작업에서는 이 보조동사군을 주석하지 않는다.
- **인용문**: 3.1에서 살펴본 것처럼 한국어에서는 인용을 형태소가 담당하는 경우가 많으므로, 인용동사가 없는 경우 인용표지 형태소에 따로 주석한다.
- **형용사**: 주석하지 않는 것을 원칙으로 하되 주석자의 판단에 따라 필요한 경우 한다.

4.3 주석 결과 분석

약 10만 개의 토큰(형태소) 중 사건은 20120개(약 20%)이다. TimeBank의 8%에 비하면 높은 비율인데, 이는 한국어의 특성도 관여하겠지만 형태소 단위를 기본으로 하는 것과 동사성 명사와 경동사를 모두 사건으로 주석하는 데에도 원인이 있는 것으로 보인다.

주석자간의 일치도 분석을 위해서는 precision, recall과 전통적인 주석자간 일치도 분석 통계 기법인 kappa[6]를 이용하였다. 주석자간 일치를 분석해 보면, 전체적으로 약 60%의 일치(kappa)를 보인다. 개별 주석자들 간의 일치는 표 1과 같다.

표 1. 주석자간 일치

주석자	precision	recall	kappa
HY-YM	0.69	0.90	0.76
MS-YM	0.46	0.95	0.55
HY-MS	0.77	0.45	0.52
MH-HY	0.71	0.80	0.70
HZ-HY	0.57	0.81	0.62
HY-JH	0.67	0.86	0.71

사건 태그 주석 실험은 위의 결과에서 보듯이 보통 정도의 주석자간 일치를 보이고 있다. 불일치의 원인으로는 첫째, 전처리를 통한 형태소 분석결과의 불완전성 둘째, 가이드라인의 불충분한 설명, 셋째, 주석자 훈련의 부재, 넷째, 한국어의 언어학적 특성에 기인한 혼란 등을 꼽을 수 있겠다. 앞의 세 가지 요인들은 기술적인 요인으로 앞으로 KTimeBank 구축을 위한 주석 작업에서 고려하여 개선하여야 할 요인들이다. 이 논문에서는 네 번째 언어학적 문제점에 대한 논의를 간략하게 한다.

첫째, 한국어의 경우 부정, 양태, 상 등을 나타내는 보조동사군이 많고, 이 보조동사들에 대하여 일관되게 사건으로 주석을 하거나 주석을 하지 않는 방식을 취할 수 없다는 점이 고려되어야 한다. 왜냐하면 보조동사 각각이 갖는 실질적 의미의 정도가 다르기 때문이다. 예를 들어 '않다'의 경우 부정의 의미 밖에 없지만 '~ 버리다'의 경우에는 그 의미를 단순히 문법적인 것으로만 보기가 쉽지 않다. 따라서 보조동사군에 대한 사건 주석의 원칙을 세워야 한다.

둘째, 한국어에서는 경동사가 중요한 언어학적 논의 중 하나인데, 사건 주석을 하는 데 있어 경동사의 범위를 어디까지 확대할 것인가가 문제일 수 있다. 예를 들어 '하다'나 '되다'는 일반적으로 경동사로 받아들여지는 데 반해 '받다', '주다', '저지르다' 등은 경동사로 볼 것인가 아닌가가 문제시될 수 있고, 그렇게 확대된다면 경동사를 사건으로 주석하는 것은 실질적으로 사건과 시간정보를 추출하는데 있어 무의미한 사건태그와 TLINK를 잉여적으로 생성하는 가능성을 배제할 수 없다.

셋째, 한국어는 명사 연쇄가 비교적 조사 결합 없이도 자유롭기 때문에 명사구 내에서의 사건 주석이 문제가 될 수 있다. 예를 들어 '민주화 정책'의 경우 '민주화'가 사건인 것은 분명하지만, '정책'을 사건으로 보아야 할지의 문제와 사건으로 본다면 둘 다 주석을 해야 할지도 문제가 될 수 있다.

넷째, 영어의 전치사(예: *through*)가 한국어에서는 동사로 표현되는 경우(예: ~를 통하여)가 많기 때문에 이런 예들을 사건으로 주석할 것인가 아니면 SIGNAL 등으로 주석할 것인가의 여부도 문제가 된다. 이와 같은 한국어의 언어학적 특성을 고려하여 주석 가이드라인을 만들기 위해서는 향후 언어학의 이론적 논의와 실용적 고려가 병행되어야 할 것이다.

6) 이 주석자간 일치 분석은 주석 도구인 BAT가 자동으로 처리한 결과이다.

5. 결론 및 향후 과제

지금까지 한국어 텍스트에서 시간 및 사건 정보 추출을 위한 주석 시스템의 개발에 대하여 논의하였다 시간 및 사건 표현의 주석은 정보추출 등의 자연언어처리 응용분야에 중요한 역할을 함에도 불구하고, 최근까지 한국어에서는 그러한 시스템이 개발되지 않았다 이 논문의 목적은 시간 및 사건표현 주석 정보 추출, 추론을 위한 KTimeML, KTimeBank, KTTK 개발과 구축의 필요성과 함께 현재 진행되고 있는 KTARSQI Project에 대한 전반적인 소개를 하고자 하는 것이라 할 수 있다 이에 더하여 구체적인 주석 실험으로 사건 주석과정과 분석결과를 보이며, 문제점과 향후 개선 방향을 언급하였다.

현재 KTARQI 프로젝트는 이제 막 시작한 단계에 있으나 앞으로 이 연구의 결과가 한국어 자연어처리응용분야에 도움이 되기를 기대하며 또한 도움이 되리라 예상한다. KTimeML의 명세와 주석 가이드라인의 개발, KTimeBank 구축을 위한 주석 작업, 그리고 KTTK의 세부 툴들을 개발하는 것이 향후 이 프로젝트의 전개 방향이며, 그 결과는 KTARSQI 웹페이지와 학회 논문, 저널 등을 통해 공개될 것이다. 이 자료들이 많은 도움이 되기를 바란다.

참고문헌

[1] James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizaukas, Andrea Setzer, and Graham Katz. TimeML: Robust Specifications of Event and Temporal Expressions in Text. IWCS-5. *Fifth International Workshop on Computational Semantics*. 2003.

[2] James Pustejovsky, Jessica Littman, Roser Saurí, Marc Verhagen. *TimeBank 1.2. Documentation*. 2006.

[3] Marc Verhagen and James Pustejovsky. Temporal Processing with the TARSQI Toolkit. In *Proceedings Coling 2008: Companion volume - Posters and Demonstrations*, Pages 189-192. 2008.

[4] Pustejovsky, J., M. Verhagen, X. Nianwen, R. Gaizauskas, M. Happle, F. Shilder, G. Katz, R. Saurí, E. Saquete, T. Caselli, N. Calzolari, K.-Y. Lee, and S.-H. Im. *TempEval2: Evaluating Events Time Expressions and Temporal Relations: SemEval Task Proposal*. 2008.

[5] Roser Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. TimeML Annotation Guidelines Version 1.2.1. 2006.

[6] Jang, Seok-Bae, Jennifer Baldwin, and Inderjeet Mani. Automatic TIMEX2 Tagging of Korean News. In *Proceedings of ACM Transactions on Asian Language Information Processing*. Vol. 3, No. 1, Pages 51-65. 2004.

[7] Kiyong Lee. Formal Semantics for Temporal Annotation, An invited plenary lecture for CIL 18. In *Proceedings of the 18th International Congress of Linguistics*, CIL 18, Seoul, Korea. 2008.

[8] Seohyun Im and Roser Saurí. TimeML Challenges for Morphological Languages: A Korean Case Study. In *Proceedings of CIL 18*, Seoul, Korea. 2008.

[9] Seohyun Im, Hyunjo You, Hayun Jang, Seungho Nam, Hyopil Shin. KTimeML: Specification of Temporal and Event Expressions in Korean Text. In *Proceedings of the 7th Workshop on Asian Language Resources in conjunction with ACL-IJCNLP 2009*, Suntec City, Singapore. 2009.

[10] 김재훈, 서정연. ms. *한국어 자연언어처리를 위한 품사 태그셋 버전* 1.0. KAIST. 1994.

[11] 유현조, 김문형, 준호 줄리아노, 남승호, 신효필. ms. *한국어 사건 인식 시스템* 2009.

[12] 임서현, 김윤신, 남승호. ms. *KTimeML EVENT 태그 주석 가이드라인*. 2009.