

Saken: 한국어 사건 인식 시스템

유현조, 김문형, 준호 줄리아노, 남승호, 신호필
서울대학교 언어학과

{youhyunjo, likerainsun, jjunho, nam, hpshin}@snu.ac.kr

Saken: A Korean Event Recognizer

Hyun-Jo You, Moonhyung Kim, Juliano P. Junho, Seunggho Nam, Hyopil Shin
Department of Linguistics, Seoul National University

요 약

한국어 자연언어 텍스트에서 사건을 자동으로 인식하기 위한 Saken 태거를 소개하고자 한다. Saken 태거는 한국어 사건 및 시간의 자동 인식을 위한 시스템인 한국어 TARSQI 툴킷을 구성하는 하나의 모듈로 개발된 것이나 독립적으로 사건 추출 도구로 사용될 수도 있다. Saken 태거는 미리 구축된 사건의 목록이나 특정 도메인으로 적용 대상을 제한하지 않고 보편적으로 사용될 수 있는 사건 분석기를 지향하고 있다. 이 논문에서는 사건 태깅을 위한 언어학적 배경과 Saken 태거를 구성하는 세부 모듈을 소개하고 신문 기사를 이용한 평가 실험 결과를 분석할 것이다.

주제어: 사건 인식, 사건 표현, 시간 정보, TimeML, TARSQI

1. 서 론

사건 및 시간 정보는 자연언어 텍스트의 의미에서 핵심적인 부분을 구성하며 그 정보를 자동으로 추출하는 것은 텍스트 이해 전체 작업에서 매우 중요한 요소임에도 불구하고 다른 측면의 의미 해석에 비해 충분히 주목을 받지 못했다[1]. 시간 추론이 질의응답 시스템에서 필수적임에도 불구하고 시간 관계를 필요로 하는 질의는 현재의 질의응답 시스템에서는 풀지 못하고 있다. 2001-2004년의 factoid TREC 질문 중 약 8.7%가 시간 맥락 추론을 필요로 하고 있다는 사실에서 시간 정보 처리의 중요성을 가늠해 볼 수 있다. 예를 들어, ‘Who was the president of South Korea when Bill Clinton was President of the US?’ 라는 질문에 답하기 위해서는 사건 탐지와 맥락 기반 추론을 필요로 한다[2]. 국내 연구들에서도 시간 정보가 정보 추출, 질의응답 시스템, 자동 요약과 같은 자연언어처리 응용분야에서 중요한 역할을 한다는 데에 인식을 같이 하고 있으나 언어학 기반의 이론적인 기초 연구는 거의 논의된 바 없다. 관련된 공학적인 연구들의 초기 성과로 간단한 자연언어처리 기술을 이용한 시간 표현 추출 및 정규화[3,4,5], 자동 요약에 사용될 수 있는 사건 탐지 기법[6] 등이 있다.

자연언어 텍스트의 사건 및 시간 정보 추출에 관한 본격적인 연구로 영어를 대상으로 하는 TARSQI (Temporal Awareness and Reasoning System for Question Interpretation) 프로젝트를 꼽을 수 있다. 이 프로젝트를 통해 사건 및 시간 정보를 표상하기 위한 마크업 언어 TimeML의 명세화가 이루어졌으며[7], 그에 따라 실제 분석된 영어 말뭉치 타임뱅크(TimeBank)가 구축되었고[8], 사건 및 시간 정보를 자동으로 분석해 주는 도구인 Tarsqi Tool이 개발되었으며[9], 여러 언어에서 이러한 도구들을 평가하기 위하여 SemEval의 일환으로

TempEval-2가 예정되어 있다[10]. 현재 영어 이외에 이탈리아어, 스페인어, 중국어, 프랑스어에 대해 관련 프로젝트가 진행 중이다. 한국어 TARSQI 프로젝트도 기초적인 이론적 논의의 성과[12,13]에 바탕을 두고 실질적인 작업으로서 한국어 타임뱅크 구축과 한국어 TARSQI 툴킷 개발을 위한 연구가 진행되고 있다[14,15].

사건 및 시간 정보의 자동 추출하는 데에 있어 가장 기본이 되는 것은 시간 표현의 자동 인식과 사건 표현의 자동 인식이다. 그것이 이루어진 후에야 사건들 사이의 시간 관계를 분석할 수 있으며 사건들과 시간 표현들 사이의 관계를 분석하여 시간 표현이 가리키고 있는 시간 상의 기준점들 사이에 사건들을 위치시킬 수 있다. 앞서 언급한 바와 같이 한국어의 사건 및 시간 정보의 자동 분석에 관한 연구들이 존재하나 TimeML과 같은 표준적인 형식[1]을 따르는 범용 도구에 대한 연구는 찾아보기 어렵다. 시간 표현에 대해서는 TimeML 개발 이전의 기술 언어인 Timex2를 따르는 한국어 자동 태깅 연구[16]가 있으나 TimeML로의 발전에 맞춘 후속 작업이 이루어지지 않았으며 사건 표현과 관련된 연구는 없는 실정이다.

이 논문에서는 한국어 자연언어 텍스트의 사건 및 시간 정보 자동 추출을 위한 한국어 TARSQI 툴킷이라는 큰 틀 속에서 가장 기본이 되는 모듈인 사건 자동 인식 시스템 Saken을 소개하고자 한다. 사건 태거인 Saken은 하나의 모듈로서 개발된 것이나 TARSQI 툴킷과 무관하게 독립적으로 실행될 수 있다. 이러한 독립성과 더불어 특정 도메인에 한정되지 않고 사건 표현을 추출하여 태깅할 수 있는 범용성을 중요시하여 개발되었으며 다른 목적을 가진 시스템에서도 컴포넌트로 도입하여 사용할 수 있는 가능성이 열려있다.

1) TimeML은 2008년 사건 및 시간표현 주석을 위한 ISO 표준으로 채택되었다(ISO/DIS 24617-1; 2008).

2. 언어학적 배경

2.1. TimeML

TimeML은 자연언어 텍스트에서 사건 및 시간표현을 표상하기 위한 마크업 언어로 ISO 표준으로 채택되었다. 한국어 TARSQI 프로젝트는 ISO-TimeML을 준수하며 한국어의 문법적 특성을 효율적으로 표상하기 위해 필요한 속성들을 명세화 하였으며[14,15], 실제 주석 작업의 결과를 통한 심도있는 후속 논의가 이루어지고 있다[17].

TimeML은 기본 요소 EVENT, TIMEX3, SIGNAL과 연결 요소 ALINK, SLINK, TLINK로 구성된다. EVENT 객체는 사건 표현, TIMEX3 객체는 시간 표현, SIGNAL 객체는 명시적으로 드러난 시간 관계 표현인 접속사/전치사/후치사를 표상하기 위한 것이다. ALINK는 상(aspect), SLINK는 종속(subordination)이라는 특수한 시간적 연결 관계를 표상하기 위한 것이며, TLINK는 TimeML 객체 사이의 일반적인 시간 관계를 표상하기 위한 것이다.

이러한 TimeML의 기본틀은 한국어에도 그대로 적용된다. ISO-TimeML에 따라 EVENT의 경우 언어에 따라 시제, 상, 범의 문법 구조가 다르므로 해당 언어에 적합하게 속성과 속성값을 설정할 수 있다. 이에 따라 한국어를 위한 EVENT 명세화 논의가 이루어지고 있다[14,15]. 현재 한국어 TARSQI 프로젝트에서 채택한 EVENT의 BNF는 다음과 같다.

```

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'}

```

2.2. 사건 주석

이러한 TimeML의 여러 요소 중 가장 기본이 되는 EVENT 객체의 주석에 대해 자세히 살펴보자.

한국어에서 사건을 나타내는 표현은 동사, 형용사, 서술성 명사가 주를 이룬다. 대부분의 동사와 행위성 명사가 사건으로 태깅되며, 제한적으로 형용사와 상태성 명사의 일부가 사건으로 태깅될 수 있다. 사건으로 태깅될 수 있는 이러한 전형적인 품사 이외에도 맥락에 따라 개체성 명사를 사건으로 태깅해야 하는 경우도 있다.

실제 텍스트에서 사건 주석을 위한 기준들 중 동사로 표현되는 전형적인 사건 표현은 쉽게 판별할 수 있으나 명사, 형용사, 형식적인 동사, 인용 표현을 사건으로 태깅할 때에 어려움이 발생하므로 상세한 가이드라인이 필요하다[15,17].

행위성 명사를 사건으로 판별하는 것 자체는 어렵지 않으나 명사가 연쇄되는 표현에서 사건을 태깅하는 것은 문제가 된다. 이때 머리에 해당하는 마지막 명사가 사건 표현일 때만 태깅하는 것이 원칙이다. 명사 연쇄 구성이 빈번하고 광범위하게 사용되는 한국어의 특성상 사건 표현이 명사로 나타나는 빈도가 높으며 명사 연쇄 내에 시간 관계가 표현되는 경우도 많다. 예를 들어, ‘이스라엘-팔레스타인 평화 협상 재개’와 같은 경우에는 ‘재개’ 뿐만 아니라 ‘협상’도 사건으로 태깅해야 한다.

행위성 명사나 상태성 명사가 아니더라도 자격 직업, 직업 등의 명사와 장소를 나타내는 명사는 사용된 맥락에 따라 사건으로 태깅될 수 있다.

형용사는 매우 엄격하게 주석한다. 형용사가 변화할 수 있는 속성을 의미할 때, 그리고 문장에서 그 속성이 발생하는 시간에 대한 정보를 분명히 추출할 수 있을 때만 태깅하고, 나머지 경우에는 태깅하지 않는 것을 원칙으로 한다. 일반 명사를 사건으로 태깅해야 하는 경우가 있는 것과 마찬가지로 맥락에 따라 시간 정보가 의미있다고 판단되는 경우만을 사건으로 태깅한다.

품사가 동사에 해당하는 것을 항상 사건으로 태깅하는 것은 아니다. 행위성 명사와 ‘하다/되다’가 결합한 경우에는 두 토큰에 모두 태깅하기로 하였다. 예를 들어, ‘회의+하다’의 경우 ‘회의’와 ‘하다’를 모두 사건으로 태깅한다. 이에 비해 ‘이다/아니다’는 태깅하지 않으며, ‘있다/없다’는 존재의 의미를 지니는 경우만을 선별적으로 사건으로 태깅하고, 다양한 보조동사들은 기본적으로 태깅하지 않는 것을 원칙으로 한다.

인용 표현이 명시적인 인용 동사로 나타나는 경우는 문제가 되지 않으나 ‘-냐, -다, -자, -라’에 다른 어미가 연결된 문법적 구성으로 표현된 경우 문법 형태소를 사건으로 태깅해야 한다. 예를 들어, 다음 용례

이란 최고지도자 아야톨라 알리 하메네이는 21일 ‘우라늄 농축 중단을 조건으로 협상을 벌이자’는 서방의 제안에 대해 “이런은 이미 결정을 내렸다. 핵과 관련해 신의 뜻대로 우리의 길을 계속 갈 것이다”라며 거부 의사를 명백히 했다.

에서 밑줄 친 ‘-라며’는 인용 표지이므로 사건으로 태깅해야 한다.

3. 사건 자동 인식 시스템 Saken

Saken은 한국어 TARSQI 프로젝트에서 개발한 사건 자동 인식 시스템이다. 한국어 TARSQI 프로젝트는 두 가지 주요한 연구를 수행하고 있다. 하나는 앞의 2절에서 살펴본 한국어를 위한 TimeML을 명세화와 타임뱅크 구축이고, 다른 하나는 자연언어 텍스트에서 사건 및 시간 정보의 분석을 위한 도구인 한국어 TARSQI 툴킷을 개발하는 것이다. Saken은 한국어 TARSQI 툴킷을 구성하는 하나의 모듈로서 텍스트에서 사건을 지시하는 표현을 찾아 태깅하는 역할을 한다.

사건 분석기는 사건을 지시하는 언어 표현을 찾아 EVENT 객체를 생성하는 컴포넌트이다. 사건 분석은 두 단계로 나뉠 수 있다. 첫 단계는 사건을 지시하는 언어 표현을 포착하여 그 경계를 찾아내는 EVENT 판별 단계이다. 두 번째 단계는 포착된 사건 표현에서 사건의 속성에 관련된 언어 정보를 찾아내어 해당 EVENT 객체의 속성들의 값을 채우는 것이다.

3.1. 전처리

사건 분석기 Saken을 비롯한 다른 모든 TARSQI 컴포넌터는 원시 텍스트의 전처리를 요구하며 전처리의 가장 중요한 요소는 형태소 분석과 청킹이다. 이와 더불어 타임뱅크 구축의 보조 도구로서 사용하기 위해 TimeML 메타데이터 추출이 필요하다. 그외에 URL 등 분석에 불필요한 요소를 삭제하여 정제하는 등의 기술적인 요소들도 필요하다.

형태소 분석은 원래 자동화 시스템에서 내부적으로 언어 정보를 얻기 위한 목적으로 필요한 것이나 한국어의 경우에는 타임뱅크 주석자들에게 제공될 기본 자료를 만드는 데에도 사용되고 있다. 한국어에서는 공백으로 분리되는 어절 단위를 TimeML 태깅의 최소 단위로 사용할 수 없기 때문에 형태소를 태깅의 최소 단위로 하고 있다. 앞서 2.2절에서 언급된 문법적 인용 표지, SIGNAL로 태깅되는 영어 등의 전치사에 대응하는 조사 등의 존재가 형태소 단위로 태깅해야 할 이유가 된다.

형태소 분석을 위하여 소스가 GPL로 공개되어 있는 KTS²⁾를 사용하였다. 이 형태소 분석기는 행위성 명사 상태성 명사, 시간 명사 등의 품사를 분류하고 있어 사건 및 시간 분석을 위해 중요한 정보를 기본적으로 제공한다. 장점이 있다.

청킹은 형태소들을 묶어 TimeML 객체를 생성하기 위한 기본 단위를 제공하는 단순한 수준에서 이루어진다. 하나의 청크에 둘 이상의 TimeML 객체가 대응될 가능성이 있을 경우 Saken을 비롯한 여러 컴포넌트의 처리 과정의 설계에 어려움이 발생한다. 따라서 청킹은, 예를 들어, ‘떡/pv 고/ecx 있/px 더/efp ㄴ데/ecs’와 같은 형태소의 연쇄를 한 덩어리로 묶어주는 수준까지만 필요로 한다.

3.2. 사건 표현 판별

사건 판별은 명사, 형용사, 동사로 형태소 분석된 것만을 후보로 이루어지며 각 품사에 따라 다른 방법을 사용한다. 동사는 대부분의 경우가 사건으로 태깅되며 명사와 형용사는 일부만이 사건으로 태깅된다.

3.2.1. 동사 청크에서 사건 표현 판별

동사 청크에서 사건 판별은 어휘 목록을 기반으로 한다. 동사는 기본적으로 사건으로 태깅한다. 그러나 ‘이다/아니다’, ‘있다/없다’, 경동사, 보조동사 등의 형식적인 요소들의 경우에는 일부를 제외하고는 사건으로 태깅하지 않는 것이 기본이 된다. 이러한 구분을 위하여 형식적인 동사의 목록이 작성되었다. ‘새는 난다’와 ‘지구는 돈다’와 같은 총칭적인 표현은 사건으로 태깅하지 않는 것이 원칙이나 형태-통사적 패턴 정보를 이용하여 이러한 경우를 걸러낼 장치는 마련되지 못했다.

3.2.2. 명사 청크에서 사건 표현 판별

명사 청크의 경우에는 어휘 목록과 통사 패턴 정보를 이용하여 판별한다. 행위성 명사의 목록을 작성하여 일차적인 판단 기준으로 이용하고 있다. 세종전자사전의 의미부류를 이용하여 추출한 목록과 ‘하다/되다’와 결합하여 동사로 파생되는 명사의 목록을 통합하여 ‘가담, 강연, 공격, 담합, 생각, 시작, 질문, 회의’ 등 8629개의 사건 후보 명사 목록을 작성하였다. 영어 사건 태거 Evita도 이와 유사한 방식으로 만들어졌으나 TimeBank 1.2를 이용한 학습을 통해 사건이 아닌 것을 필터링하는 단계를 추가적으로 수행한다[11]. 현재로서는 구축된 한국어 타임뱅크가 존재하지 않으므로 학습에 이용할 입력 자료가 없어 이러한 걸러내기가 불가능하다.

한국어에서는 명사 연쇄가 빈번하게 사용되어 이를 처리하는 것이 중요하다. ‘평화 협상 재개’와 같은 명사 연쇄는 가장 오른쪽 명사부터 시작하여 사건 여부를 판별하며 사건으로 판별될 경우 그 왼쪽 명사에 대한 사건 여부 판별이 연쇄적으로 진행된다. 사건이 아닌 명사가 나올 때 이 과정이 중단된다. ‘방학 시작’과 같은 연쇄의 경우 ‘방학’ 자체는 행위성 명사가 아니지만 ‘시작’이 사건을 논항으로 취하여 상적 의미를 부여하는 요소이므로 ‘방학’도 사건으로 태깅한다. 이를 위하여 상 표현 어휘의 목록을 이용하고 있다.

명사는 그 어휘의 수가 매우 많을 뿐만 아니라 특정 도메인에서 특수한 전문용어를 사용할 수 있으므로 어휘 목록만으로는 범용 태거로서의 기능을 발휘할 수 없다. 이를 위하여 단순한 통사 패턴 정보를 이용하여 미등록어를 사건으로 처리하고 있다. 예를 들어, ‘파싱 하다’와 같이 경동사와 결합하거나 ‘파싱 시작’과 같이 상 표현과 함께 나타나는 패턴 정보를 이용하여 ‘파싱’을 사건으로 태깅한다.

3.2.3. 형용사 청크에서 사건 표현 판별

2) <http://kldp.net/projects/kts>

형용사 칭크의 사건 판별은 가장 어려운 부분이다 영어 사건 태거 Evita의 경우 TimeBank 1.2에서 사건으로 태깅된 형용사의 목록을 이용하여 태깅할 뿐 다른 기제를 도입하지 않고 있다. Saken 태거는 ‘-었-’, ‘-겠-’, ‘-더-’ 등 시간 관련 문법 형태가 명시적으로 나타난 경우에는 사건으로 태깅하고 있다.

사건 태거는 사전 정보와 형태통사 패턴 정보를 조합하여 사용하고 있으며 한국어 타임뱅크가 구축됨에 따라 기계 학습을 이용한 필터링을 도입할 것이다. 이러한 세 가지 방법론의 조합으로 많은 부분을 해결할 수 있을 것이나 총칭적인 표현을 걸러내는 것은 어려울 것으로 보인다.

3.3. 사건 속성 분석

사건의 문법적 속성을 분석하는 단계도 품사에 따라 다른 방법을 사용한다. 한국어의 교착적 특성으로 인해 문법적 속성은 쉽게 뽑아낼 수 있으며, 어휘의미적 성격의 class 속성을 위해서는 별도의 처리를 필요로 한다.

사건으로 판별된 동사 칭크와 형용사 칭크의 경우에는 어휘에 결합되어 있는 어미들과 사건의 문법 속성 간에 일대일 대응 관계를 맺을 수 있으므로 간단한 규칙으로 속성을 추출할 수 있다. 개별 사건으로 태깅되지 않는 보조동사와 결합한 용언의 속성은 보조동사에 결합한 어미에서 추출하여 채워 넣는다.

명사 칭크의 경우에는 대부분의 속성값을 가지지 않으므로 별다른 분석을 필요로 하지 않는다. 다만 형식적인 용언과 결합하여 사건을 이루는 명사의 속성은 그 문법적 속성을 형식 용언에 결합한 어미들에서 추출하여 채워 넣는다.

사건의 속성 중 가장 기본적인 것은 클래스(class)이다. 클래스 속성은 형태-통사 패턴 정보를 통하여 알 수 없는 어휘의미적 속성이다. 클래스의 판정은 클래스 사전에 의존하여 이루어진다. 영어 사건 태거 Evita의 경우에는 수십여개로 이루어진 작은 클래스 사전을 이용하고 있으며 사전에 없는 것은 무조건 ‘OCCURENCE’ 속성값을 주고 있다. 한국어에서는 행위성 명사들이 영어에 비해 광범위하게 사용되고 행위성 명사는 어휘적으로 매우 분화되어 있기 때문에 영어에 비해 큰 규모의 클래스 사전을 필요로 한다. 현재는 Saken을 위한 클래스 사전은 충분히 다듬어지지 않은 상태로 ‘REPORTING’과 ‘PERCEPTION’ 클래스에 해당하는 어휘목록을 보완할 필요가 있다.

한국어의 특성상 사건의 속성을 분석하는 것은 간단한 절차로 가능하다. 사건 자체의 판별이 정확하게 이루어지기만 한다면 속성의 분석 단계는 효율적으로 수행될 수 있다.

4. 결과

사건 인식 시스템을 평가하기 위하여 현재 구축 중인 한

국어 타임뱅크의 일부 중간 결과물과 비교하였다. 이 말뭉치는 현재 약 10만 개의 형태소 분량의 신문 기사 텍스트로 이루어져 있으며, 각 기사마다 두 명의 주석자가 사건 표현을 선별하여 태깅하였다. 주석자 간의 일치도의 범위는 다음과 같다[17].

precision	0.46-0.77
recall	0.45-0.90
kappa	0.52-0.76

주석자 간의 일치도가 높지 않은 이유로는 전처리를 통한 형태소 분석의 불완전성 가이드라인의 불충분한 설명, 주석자 훈련의 부재, 한국어의 언어학적 특성에 기인한 혼란 등을 꼽을 수 있다[17].

주석자 간 일치도가 낮기 때문에 현재 구축 중인 한국어 타임뱅크는 태거의 평가를 위한 기준으로 삼기에는 완성도가 떨어진다. 이러한 점을 보완하기 위하여 주석 결과를 주석자 간 합집합과 교집합 두 세트로 작성하여 평가 기준으로 삼았다.

다음은 주석자 간 합집합, 즉 어느 한 주석자라도 사건으로 태깅한 토큰을 모두 모은 결과를 기준으로 삼아 Saken 태거를 평가한 결과이다.

	True	False
Positive	14136	7647
Negative	5154	66992

Precision : 64.89%
 Recall : 73.28%
 F-Measure : 68.83%,

한국어 사건 태거의 평가 결과는 영어 사건 태거 Evita의 정확도 74.03%, 재현율 87.31%, F-측도 80.12%에 비교하면 매우 낮은 편이다. 그러나 주석자 간 정확도와 재현율이 보이는 큰 편차에 비하면 이러한 자동 태깅 수행 성능은 충분히 안정적이라고 볼 수 있다. 현재 태거 성능을 향상시키기 위해 앞서 불완전한 주석자 간 일치도의 개선을 관찰해야 할 것으로 보인다. 주석자 간 일치도가 올라감에 따라 태거의 성능을 비례적으로 향상시킬 수 있느냐가 앞으로의 관건이 될 것이다.

다음은 주석자 간 교집합, 즉 주석자들 사이에 일치한 토큰만을 모은 결과를 기준으로 삼아 사건 태거를 평가한 결과이다.

	True	False
Positive	8069	13714
Negative	1734	70412

Precision : 37.04%
 Recall : 82.31%
 F-Measure : 51.09%,

정확도(precision)가 매우 낮은 것은 사건 명사 태깅과 더불어 형용사의 태깅에서 생겨난 차이로 보인다 영어에서 사건으로 태깅되는 토큰이 8%인데 비해 한국어에서는 20%가 사건으로 주석되며[17] 이 중 사건 명사가 큰 비중을 차지하고 있다. Saken 태거가 사건을 지칭할 가능성이 있는 서술성 명사를 광범위하게 사건으로 판정하는 것이 정확도가 낮아진 하나의 원인이다. 주석자들은 형용사를 사건으로 태깅하는 것을 매우 엄격하게 제한하고 있는 가이드라인을 따르고 있는 데에 비해 Saken 태거는 이것을 적절하게 판별하는 모듈을 가지고 있지 못하며 더 많은 형용사 토큰을 사건으로 태깅하는 경향이 있다. 영어 사건 태거인 Evita의 경우에는 TimeBank 1.2에서 형용사가 사건으로 태깅된 목록을 추출하여 자동 태깅의 기반으로 삼고 있으나 우리는 아직 그러한 리소스를 가지고 있지 못하다.

주석자 간 합집합을 기준으로 하였을 때 정확도가 약 65%인 데에 비해 주석자간 교집합을 기준으로 하였을 때 정확도가 약 37%로 떨어져 큰 차이를 보이고 있다. 태거의 성능을 개선하기 앞서 사건 태깅에 대한 합의와 그를 바탕으로 하는 명확한 가이드라인이 작성되어야 할 것으로 보인다. 그 결과에 따라 현재 Saken 태거의 정확도는 37-65% 사이의 어느 한 점으로 귀착될 것이며 서술성 명사와 형용사를 중심으로 주석 결과를 면밀히 분석하여 Saken 태거의 정확도를 높일 수 있을 것이다.

재현율은 평가 기준에 따라 73-82%이다. 정확도 60%대에서 이러한 재현율이 유지된다면 개발 초기 단계의 태거로서는 만족할 만한 수준이다. 그러나 평가 기준이 될 한국어 타임뱅크의 주석이 어떤 방향으로 안정되느냐에 따라 정확도가 크게 달라질 것이고 정확도가 크게 떨어진다면 현재의 재현율은 별다른 의미가 없을 것이다.

이번 실험을 통하여 한국어 사건 자동 태깅의 개선을 위해 초점을 맞추어야 할 것은 재현율 보다는 정확도임을 알 수 있었다. 사건으로 태깅될 가능성이 있는 표현이 동사, 형용사, 서술성 명사로 한정되어 있기 때문에 어휘-문법 정보를 이용한 규칙으로 재현율은 어렵지 않게 높일 수 있었다. 그러나 해당 후보들 중 사건으로 태깅되지 않는 것을 걸러내는 과정은 맥락 정보에 크게 의존하며 매뉴얼 태깅에서도 주석자들의 주관적인 판단을 요구하는 복잡한 과정이다 이것이 자동 태깅의 가장 큰 걸림돌이며 정확도를 낮추는 주요 원인이다. 대부분의 동사는 사건으로 태깅되나 형용사와 서술성 명사 중 많은 경우가 사건으로 태깅되지 않는다. 그 판단은 문맥을 이해한 후에 가능하며 현재로서는 이에 대한 형태, 통사, 어휘적인 패턴을 발견할 수 없다.

영어 사건 태거 Evita가 이미 구축된 TimeBank 1.2가 이용하고 학습하고 사건 후보를 필터링하여 문제를 해결하고 있는 데에 비해, 현재의 Saken 태거는 그러한 학습 자료를 확보하지 못한 상태에서 사전과 규칙만을 이용하여 태깅하고 있다. Evita와는 주어진 환경이 다르므로 우선은 어휘 정보와 형태-통사 패턴을 이용하여 사건 후보를 필터링하는 모듈을 최대한 개선하고 한국어 타임뱅크가 충분히 안정된 후에 학습 모듈을 도입하는 것을 Saken 태거의 개발 방향으로 결론 내린다.

5. 결론 및 향후 과제

한국어 사건 자동 인식 시스템인 Saken의 세부 구성을 소개하고 평가 결과를 분석하였다 최초의 사건 및 시간 분석 말뭉치 구축 프로젝트로서 연구의 초기 단계에 있는 까닭에 분석 지침이 완벽하지 않고 연구자들 사이에 충분한 합의가 이루어지지 않았으며 이에 따라 매뉴얼 태깅에서도 주석자 간의 일치도가 높지 않았다 이러한 말뭉치를 평가 기준으로 삼았기 때문에 평가 결과는 대략의 예측을 가능하게 해줄 뿐이다. 주석의 방향이 어느 쪽으로 안정화되는가에 따라 Saken 태거에 대한 평가 결과는 크게 달라질 수 있으나 전반적으로 재현율은 어느 정도 높은 편이고 정확도는 매우 떨어지는 상황이라고 판단된다.

Saken 태거를 개선하기 위하여 가장 절실하게 요구되는 것은 규모가 크지 않더라도 정제된 한국어 타임뱅크이다. 우선적으로는 현재 사용되고 있는 어휘 목록과 형태-통사 패턴 규칙의 확장을 통하여 Saken 태거를 개선시킬 수 있을 것이며, 한국어 타임뱅크가 모습을 갖추에 따라 기계 학습 방법을 도입하여 정확도를 높을 수 있을 것으로 기대된다.

현재 Saken 태거는 사건 판별뿐만 아니라 사건 속성도 분석할 수 있는 여건을 갖추고 있으나 한국어 타임뱅크에는 아직 사건의 속성에 대한 주석이 되어 있지 않아 이와 관련한 태거의 평가를 수행하지 못하였다 한국어 타임뱅크가 구축됨에 따라 평가가 조속히 이루어질 것이다.

참고문헌

- [1] Setzer, Andrea, Robert Galzauskas, and Mark Hepple. The Role of Inference in the Temporal Annotation and Analysis. *Language Resources and Evaluation* (2005) 39: 234-265. Springer Netherlands. 2006.
- [2] Moldovan, Dan, Christine Clark, Sanda Harabagiu. Temporal Context Representation and Reasoning. *International Joint Conferences on Artificial Intelligence*, 1099-1105. 2005.
- [3] 김평, 남덕윤, 최기석, 맹성현. 자동 추출된 시간 정보를 이용한 사건 클러스터링. *한국정보과학회 2005 한국컴퓨터종합학술대회 논문집(B)* 2005. 7. 466-468. 2005.
- [4] 김평, 맹성현. 사건 탐지 및 추적을 위해 신문기사에서 자동 추출된 시간 정보의 유용성 판단. *정보과학회논문지*. 2006.
- [5] 김평, 성기운, 맹성현. 사건 탐지/추적을 위한 시간 정보 추출. *제15회 한글 및 한국어정보처리 학술대회*. 22-29쪽. 2003.
- [6] 정영미, 김용광. 사건 중심 뉴스기사 자동요약을 위한 사건탐지 기법에 관한 연구. *정보관리학회지*. 2008.

[7] J. Pustejovsky, J. Castaño, R. Ingria, R. Saurí, R. Gaizaukas, A. Setzer, and G. Katz. "TimeML: Robust Specifications of Event and Temporal Expressions in Text." *IWCS-5. Fifth International Workshop on Computational Semantics*. 2003.

[8] Pustejovsky, J., J. Littman, R. Saurí, M. Verhagen. *TimeBank 1.2. Documentation*. 2006.

[9] M. Verhagen and J. Pustejovsky. "Temporal Processing with the TARSQI Toolkit". In *Proceedings Coling 2008: Companion volume - Posters and Demonstrations*, pp.189-192. 2008.

[10] J. Pustejovsky, M. Verhagen, X. Nianwen, R. Gaizauskas, M. Happle, F. Schilder, 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.

[11] R. Sauri, R. Knippen, M. Verhagen, and J. Pustejovsky. "Evita: a robust event recognizer for QA systems." *Proceedings of HLT/EMNLP 2005: 700-707*. 2005.

[12] 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.

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

[14] 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.

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

[16] Jang, Seok-Bae, J. Baldwin, and I. Mani. "Automatic TIMEX2 Tagging of Korean News." In *Proceedings of ACM Transactions on Asian Language Information Processing. Vol. 3, No. 1*, pp.51-65. 2004.

[17] 임서현, 김윤신, 조유미, 장하연, 고민수, 남승호, 신호필. KTARSQI: 한국어 텍스트의 시간 및 사건 표현 주석. 미발표. 2009.

<부록1> 품사별 정확도와 재현율 (주석자 간 합집합 기준)

POS	TP	FP	FN	TN ¹⁾	Precision ²⁾	Recall ³⁾
a	5	23	12	1179	0.18	0.29
ad	0	0	0	16	N/A	N/A
ajs	0	0	0	137	N/A	N/A
ajw	0	0	0	71	N/A	N/A
at	0	0	2	95	N/A	0.00
ecq	0	0	134	839	N/A	0.00
ecs	0	0	349	1332	N/A	0.00
ecx	0	0	7	2399	N/A	0.00
ef	0	0	21	2141	N/A	0.00
efp	0	0	2	2023	N/A	0.00
exm	0	0	23	5108	N/A	0.00
exn	0	0	0	502	N/A	N/A
f	0	0	4	1395	N/A	0.00
i	0	0	1	5	N/A	0.00
jc	0	0	3	5681	N/A	0.00
jca	0	0	61	3621	N/A	0.00
jcm	0	0	0	1973	N/A	N/A
jcp	0	0	12	1599	N/A	0.00
jcw	0	0	0	20	N/A	N/A
jj	0	0	10	918	N/A	0.00
jx	0	0	3	2779	N/A	0.00
m	0	0	0	99	N/A	N/A
md	0	1	7	442	0	0.00
mn	0	2	2	167	0	0.00
nb	0	4	135	1671	0	0.00
nbu	3	7	75	1496	0.3	0.04
nc	3431	3657	3978	14483	0.48	0.46
nca	3620	80	4	6	0.98	1.00
ncs	125	103	0	0	0.55	1.00
nct	0	0	16	1029	N/A	0.00
nn	0	1	24	395	0	0.00
nmn	0	17	19	1637	0	0.00
nnp	0	0	1	393	N/A	0.00
npp	0	0	2	271	N/A	0.00
nq	0	1	34	3560	0	0.00
pa	286	922	1	5	0.24	1.00
pad	5	25	0	0	0.17	1.00
pv	2849	1425	37	138	0.67	0.99
px	76	1051	66	369	0.07	0.54
s'	0	60	0	1402	0	N/A
s,	0	0	0	698	N/A	N/A
s-	0	0	0	129	N/A	N/A
s.	0	0	0	2176	N/A	N/A
s'	0	16	0	578	0	N/A
su	0	0	24	30	N/A	0.00
sy	0	0	0	376	N/A	N/A
xa	0	0	2	143	N/A	0.00
xn	0	0	19	1414	N/A	0.00
xpa	142	136	64	52	0.51	0.69
xpv	3594	116	0	0	0.97	1.00

1) **TP**: True Positive, **FP**: False Positive,
FN: False Negative, **TN**: True Negative
2) **Precision** = TP / (TP + FP)
3) **Recall** = TP / (TP + FN)