한국어 PropBank 프레임 파일 확장 도구 설계[1)]

이정국○, 김유섭
한림대학교 컴퓨터공학과
percussive@hallym.ac.kr, yskim01@hallym.ac.kr

A Design of Frame File Extension Tool for Korean PropBank

Jung-Kuk Lee○, Yu-Seop Kim
Dept of Computer Engineering, Hallym University

요 약

본 논문에서는 한국어 PropBank의 구축을 위한 동사의 프레임 파일 확장 및 구축에 대한 연구를 논한다. 문장 단위의 의미 분석에 있어서 가장 중요하다고 볼 수 있는 의미 역 결정을 위해서 필요한 언어자원중, PropBank는 동사의 술어-논항 구조를 태그해 놓은 말뭉치로써 가장 널리 쓰이는 언어자원 중 하나이다. PropBank는 크게 술어-논항 구조를 태그한 말뭉치와 개별 동사들의 논항 구조를 기술한 프레임 파일로 이루어져 있다. 한국어 PropBank 구축을 위해서는 구문 표지 부착 말뭉치에 술어-논항 구조의 표지 부착 작업 및 한국어 동사의 프레임 파일의 구축 및 확장이 이루어져야 하는데, 본 논문에서는 세종 계획에서 발표한 용언 격틀 파일을 사용하여 기존의 한국어 PropBank 프레임 파일을 확장하는 도구를 설계하였다.

주제어: 세종 동사 격틀 파일, Proposition Bank, 표지 부착 말뭉치

1. 서론

한국어의 경우 형태소 분석 및 구문 분석에서는 매우 많은 연구들이 진행되어 왔고 이들 분석과 관련하여 다양한 표지 부착된 한국어 말뭉치가 개발되었다. 이러한 말뭉치를 활용한 기계 학습 및 통계 기반 알고리즘의 개발은 한국어 정보처리 기술을 매우 극적으로 발전시켜왔으나, 의미역 결정과 같은 문장 단위의 의미 분석은 관련 한국어 말뭉치가 없어 난항을 겪고 있다. 이런 측면에서 의미역 결정을 위한 한국어 말뭉치의 구축은 한국어 의미 분석 관련 연구에 있어 중요한 연구 가치를 가지고 있다고 할 수 있다.

현재 의미 분석을 위한 말뭉치로써 해외에서 가장 널리 쓰이는 것은 PropBank이다[1]. PropBank는 동사의 술어-논항 구조를 태그해 놓은 말뭉치로써 의미역 결정 관련 연구에 큰 영향을 미치고 있다.

PropBank는 구문 표지 부착 말뭉치에 의미 표지를 부착한 말뭉치와 하나의 동사에 대하여 모든 격틀 집합을 모아서 하나의 파일로 저장한 격틀 파일 (frame file)로 이루어진다. 한국어 PropBank는 University of Pennsylvania의 Linguistic Data Consortium[2)]에서 일부 구축되었으나 말뭉치의 크기가 실용화에 큰 도움이 되기 어려울 정도로 제한이 있어 본 연구진은 이를 직접 구축하는 작업을 진행중이다. 본 논문에서는 PropBank 구축을 위해 말뭉치의 의미 태그를 부착하는 중에 이미 구축된 프레임 파일의 규모에 한계를 느끼고 이를 확장하고자 하였다. 예를 들어 세종 계획에서 발표된 동사 격틀 파일의 크기는 한국어 PropBank의 프레임 파일보다 5배나 크다. 따라서 본 논문에서는 세종 격틀 파일을 변형하여 한국어 PropBank의 프레임 파일을 확장하는 도구를 개발하고자 하였다.

이를 위해서는 먼저, 세종 격틀 파일과 한국어 PropBank의 프레임 파일을 비교 분석하여 자동으로 변환할 수 있는 영역을 찾고, 음차 표기되어 있는 한국어 PropBank의 프레임 파일의 이름을 대응되는 표준 한국어 표기로 변환하였다. 그리고 유의어 분석을 통하여 이미 프레임 파일에 기술된 정보가 이중 등재되는 것을 필터링하였고, 마지막으로 자동 변환이 어려운 부분을 수작업을 통하여 변환하기 위한 사용자 인터페이스를 개발하였다.

2. 세종 격틀 파일과 PropBank 격틀 파일

세종 격틀 파일과 PropBank 격틀 파일은 모두 xml 파일 형식을 가지고 있다. 혼란을 방지하기 위해 세종 용언 사전의 격틀 파일은 격틀파일, PropBank의 격틀 파일은 프레임파일이라 하겠다. 먼저 세종 계획에서 발표된 격틀 구조에서 병합에 필요한 부분은 표 1과 같다.

<sense> 태그는 동사의 의미를 구분하여 격틀을 구성한다. <trans> 태그의 값은 동사의 의미를 영단어로 표기하고 있다. <frame>의 값은 격틀을 나타내는데 주어, 목

1) 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2010-0010612)
2) http://www.ldc.upenn.edu

적어 등이 어떻게 쓰이는가를 기술하고 있다. 표 1의 '가열하다' 동사의 경우 'X이 Y을 가열한다.' 라고 되어 있다. <sel_rst> 태그의 속성 부분에서 X와 Y가 무엇인지를 설명하고 있는데 X는 AGT(agent), 인간을 의미

표 1 격틀 파일 '가열하다' 내용 일부

태그	속성	값
<sense>	n="01"	
<trans>		boil
<frame>		X=N0-이 Y=N1-을 V
<sel_rst>	arg="X" tht="AGT"	인간
<sel_rst>	arg="Y" tht="THM"	구체인공물 (냄비|밥솥|후라이팬|시험관|용광로) |액체
<eg>		철수가 액체가 든 시험관을 가열하여 기체를 채취하고 있다.

하며 Y는 THM(theme), 구체인공물을 의미하고 있다. <eg> 태그는 해당 격틀 쓰임에 대한 예문을 나타낸다. 이 태그를 가지고 프레임 파일을 확장하게 된다.

프레임 파일에서 "가열"의 사례는 표 2와 같다.

표 2 프레임 파일 '가열' 의 내용 일부

태그	속성	값
<frameset>		
<lemma>		가열
<edef>		heat
<role>	argnum= "0" argrole= "agent, causer"	
<role>	argnum= "1" argrole= "thing heated"	
<mapitem>	src= "sbj" trg= "arg0"	
<mapitem>	src= "obj" trg= "arg1"	
<text>		텔레비전 디너(오븐에 가열해 곧 먹을 수 있는 알루미늄 용기 사용 냉동음식).....
<Arg>	n="1"	
<term>		알루미늄 용기 사용 냉동음식

<frameset> 태그는 프레임의 집합을 나타내는데 의미로 구분되어진다. 그 의미는 <edef>태그의 값을 통해 영단어로 표기되어 뜻을 나타내고 있다.

하나의 동사의 각 의미에는 여러 역할들이 정의되어 있는데 <role> 태그에서 그것을 찾을 수 있다. 이 역할들의 집합을 <roleset>이라 한다. '가열'의 경우 arg0은 agent 혹은 causer, arg1은 thing heated 로 나타내어져 있다. <mapitem>태그의 속성 값은 role의 argnum에 해당하는 주어, 목적어 등을 나타낸다. 표 2에서 arg0은 주어, arg1은 목적어임을 알 수 있다. <text>는 프레임 사용의 예문을 나타낸다. 예문의 관계를 나타내는 <arg> 태그는 <text> 값인 예문에 argnum 번호를 표기하고 자식인 <term> 태그의 값을 통해 '알루미늄 용기 사용 냉동음식'이 목적어인 arg1으로써 가열되는 대상임을 알 수 있게 구성되어 있다.

PropBank 프레임 파일은 세종 격틀 파일에 비해 더 많은 태그를 가지고 있기 때문에 확장하는 것이 간단하지만은 않다. 하지만 격틀 구성을 동사의 의미별로 분류한다는 점, 의미를 영단어로 표기한다는 점에서 확장의 가능성을 보여주고 있다.

3. PropBank 프레임 파일 확장

프레임 파일 확장 구조는 그림 1과 같다. 사용자가 검색어를 입력하면 격틀 파일을 검색하여 내용을 사용자에게 보여주고 프레임 파일을 검색하게 된다. 프레임 파일의 존재 여부에 따라 새로운 파일을 생성하여 확장 작업을 하거나, 기존 파일을 불러와서 확장하게 된다.

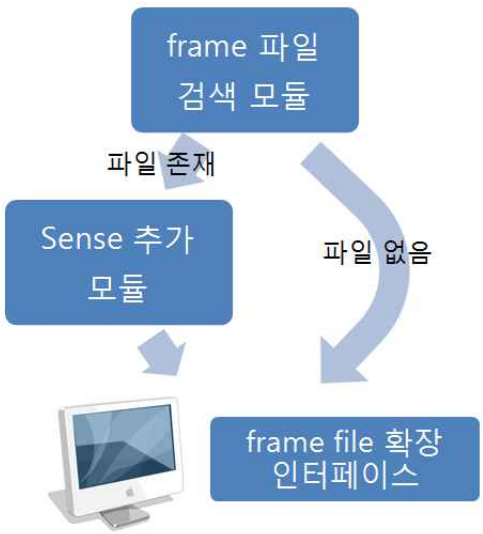

그림 1 프레임 파일 확장 구조

3.1. PropBank 프레임 파일 검색

프레임 파일의 존재 여부를 파악하기 위해 파일을 검색할 때, 세종 격틀 파일과 달리 그림 2에서처럼 프레임

파일은 파일명이 음차 표기된 알파벳으로 되어있기 때문에 검색하는 것이 용이하지 않다. 예를 들어 프레임 파일에서 '가열'을 찾으려면 'ka-yeol.kor.xm' 파일을 열어야 하는데, 사용자가 이 파일을 찾는 것이 쉽지 않다.

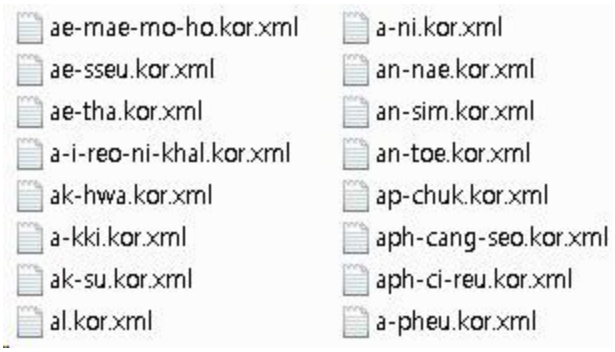

그림 2 PropBank 격틀 파일명

따라서 본 논문에서는 각 프레임 파일의 lemma를 이용하여 모든 파일명을 lemma 태그의 내용으로 임시로 변경하여 활용한다. 하지만 이렇게 변경을 하더라도 프레임 파일의 lemma는 '다', '하다', '되다'가 빠져있기 때문에 격틀 파일에서 검색할 때 동사에서 '다', '하다', '되다'를 제거하여 프레임 파일을 검색한다. 실제 프로그램에서 파일을 검색할 때는 격틀 파일 기준으로 '가열하다' 를 검색하게 되면 '하다'가 제거된 '가열'로 검색이 되게 된다.

3.2. Sense 추가 모듈

프레임 파일이 존재 할 경우 격틀 파일을 불러와서 수정한다. 이미 격틀 파일과 동일한 의미를 가진 격틀이 프레임 파일에 존재 한다면 확장할 필요가 없고, 다른 의미를 가질 때 확장하게 된다.

격틀 파일은 <sense> 태그로 그 의미가 구분되어 있다. 그 의미는 <trans> 태그에 영단어로 나타나 있다. 프레임 파일은 의미별로 <frameset>이 추가되고 <edef> 태그에 그 의미가 영단어로 표시된다.

격틀 파일과 프레임 파일의 의미를 파악하기 위해 각각의 <trans> 와 <edef> 의 내용을 비교하였다. 영단어로 표기되어 있다 하더라도 격틀 파일 '가열하다'의 <trans> 부분은 'boil', 프레임 격틀 파일'가열'의 <edef> 부분은 'heat' 로 되어있는 것을 볼 수 있는데 이 경우 두 의미가 같다고 볼 수 있다. 본 논문에서는 이러한 유의어를 판별하는 방법으로 VerbNet v3.1을 사용한다[2].

VerbNet은 콜로라도 대학에서 진행한 SemLink 프로젝트의 일부이다. VerbNet에서 cooking-45.3.xml 파일을 열면 표 3에서와 같이 <MEMBER> 태그로 유의어 목록을 나타낸다.

표 3 VerbNet 의 cooking-45.3.xml 파일 내용 일부

```
<MEMBER name="boil" wn="boil%2:30:01
        boil%2:30:00" grouping="boil.01 boil.02"/>
<MEMBER name="heat" wn="heat%2:30:01 heat%
                2:30:00" grouping="heat.01"/>
```

표 3에서처럼 cooking-45.3.xml 파일은 유의어로 boil, heat 등을 가지는 것을 알 수 있다. 본 논문에서는 단어가 서로 유의어인지를 판별하기 위한 것이므로 모든 xml 문서의 <MEMBER> 태그의 name 속성값을 모두 가져와서 역색인 구조를 새로 구성하여 표 4처럼 하나의 파일에 정보를 저장하였다.

표 4 MEMBER 태그만 통합된 유의어 사전

```
<verb lemma="boil">cooking-45.3.xml</lemma>
<verb lemma="heat">cooking-45.3.xml</lemma>
```

따라서 boil, heat 두 단어를 가지고 검색을 하게 되면 cooking-45.3.xml 이라는 동일한 파일명이 검색되므로 유의어로 판별할 수 있게 된다.

3.3. 사전 통합 인터페이스 설계

표 1의 격틀 파일의 태그를 모두 활용하여야 프레임을 확장할 수 있다. 매핑 되는 태그는 표 5에 나타나있다.

표 5 확장시 매핑 되는 태그

격틀 파일	프레임 파일
<sense>	<frameset>
<trans>	<edef>
<frame>	<mapitem>
<sel_rst>	<role>
<eg>	<text>

세종 격틀 파일의 <trans>와 PropBank 프레임 파일의 <edef>의 값으로 VerbNet을 사용해 유의어를 판별하고, 유의어가 없을 경우 세종 격틀 파일의 sense 태그 하나당 프레임 파일의 frameset 이 추가된다. 격틀 파일 검색어에서 '다', '하다', '되다' 가 제외된 단어를 프레임 파일의 <lemma> 태그에 매핑하고 <trans> 의 값을

<edef>에 그대로 가져오게 된다.

표 1에서 frame 추가의 경우 '가열하다' 는 격틀 파일의 frame이 'X=N0-이 Y=N1-을 V '으로 나타남을 볼 수 있는데, <sel_rst> 태그에서 X는 AGT(agent), Y는 THM(theme)임을 알 수 있다. 이것을 프레임 파일의 role로 옮기면 arg0 = agent, arg1=theme이 되는 규칙을 발견할 수 있고 mapitem 에 arg0를 subject, arg1을 object로 추정하는 것이 가능하다. 이 부분은 확실하게 모든 경우가 확장이 가능하다고 할 수 없으므로 사용자가 수정할 수 있도록 설계한다.

격틀 파일의 <eg> 내용을 프레임 파일의 <example>의 <text>로 매핑한다. <relation> 태그의 내용은 격틀 파일에 있는 것이 아니므로 주어, 동사의 term을 판별하여 사용자가 직접 작성하게끔 구성한다.

4. 구현

본 논문에서 구현된 프레임 파일 확장 도구의 외형은 그림 2와 같다. 먼저 사용자가 그림 2의 왼편에 있는 세종 격틀 파일의 검색 창에서 격틀 파일을 검색하면 하단에 트리 형태로 격틀 파일의 내용 중 확장에 필요한 부분이 정리되어 표기된다. 그 후에 자동으로 세종 격틀 파일의 검색어 중에서 '다', '하다', '되다'를 제거한 검색어로 프레임 파일을 검색하여 그림 2의 오른편 하단의 트리에 필요한 부분만 정리하여 표기한다. 만약 프레임 파일이 없을 경우에는 임의로 빈 파일을 생성하여 격틀을 추가할 수 있도록 한다.

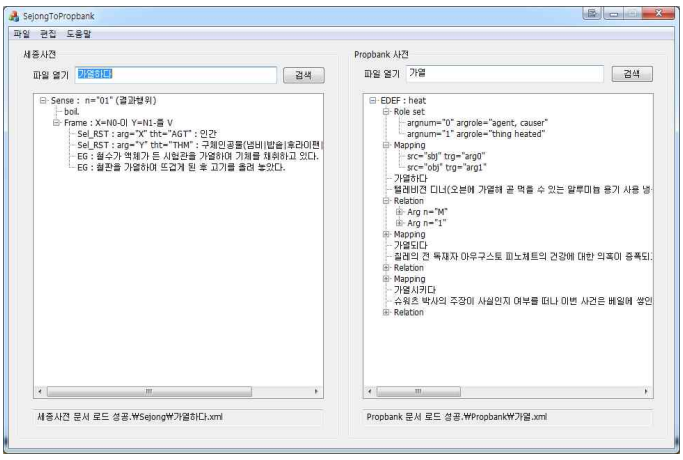

그림 4 격틀 파일 확장 도구

격틀 파일 확장 도구는 자동으로 격틀을 확장하는 것 외에도 말뭉치의 의미 태그를 부착할 때 격틀 파일을 쉽게 참조할 수 있도록 구성되었다. 격틀을 확장할 때는 편집 - 병합 메뉴를 선택한다. 확장하는 부분은 그림 4처럼 좌측에는 세종 격틀 파일을 sense별로 다시 표기하였고 오른쪽에는 해당하는 의미의 PropBank 프레임 파일의 구조를 표기하였다. 좌측 상단의 리스트 박스는 세종 격틀 파일이 가지고 있는 <sense>별로 구분되어 확장하고자 하는 <trans>를 선택할 수 있는데 프레임 파일에서 <edef>값을 모두 가져와 유의어를 비교해 의미가 겹치는 <sense>는 확장할 필요가 없음을 나타내었다.

추가해야 할 격틀을 선택 후 우측의 컴포넌트를 통해 사용자가 직접 프레임 파일의 격틀 정보를 추가하거나, 프로그램이 추천하는 메뉴를 그대로 추가하여 확장할 수 있다.

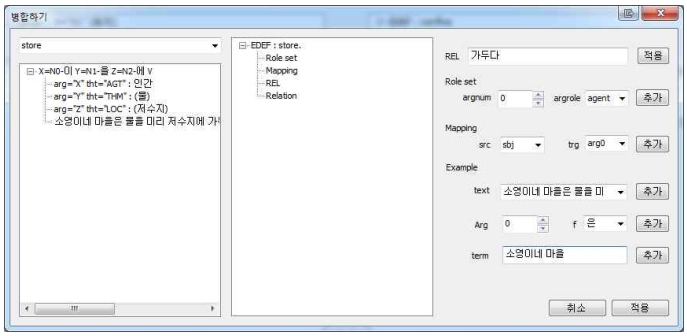

그림 5 의미별 격틀 확장 메뉴

5. 결론

본 논문에서 개발한 도구는 세종 격틀 파일을 통해 주어, 목적어를 판별하고 프레임 파일에 적절하게 매핑하고 있다. 하지만 격틀 파일의 정보 자체가 프레임 파일에 비해 작고, 확장하는 내용을 자동으로 추천을 하더라도 최종 결정은 사용자가 일일이 확인해야 하는 만큼 작업 속도는 빠르지 않아 자동화 영역을 확장해야 한다.

참고문헌

[1] Palmer, M., P. Kingsbury, and D. Gildea, "The Proposition Bank: An Annotated Corpus of Semantic Roles," Computational Linguistics, Vol. 31, No. 1, pp.71-106, 2005.

[2] Linguistic Data Consortium, "http://verbs.colorado.edu/~mpalmer/projects/verbnet.html"