시소러스와 술어 패턴을 이용한 의미역 부착 한국어 하위범주화 사전의 구축

(Constructing a Korean Subcategorization Dictionary with Semantic Roles using Thesaurus and Predicate Patterns)

양 승 현 † 김 영 섬 † 우 요 섭 †† 윤 덕 호 †††

(Seung Hyun Yang) (Young-Sum Kim) (Yo-Sub Woo) (Deok Ho Yoon)

요 약 하위범주화는 보어의 어휘 개념이 명시된 술어와 보어간 의존 관계를 정의하는 언어 정보로서 구문 및 의미 분석 등에 폭넓게 활용될 수 있는 기반 언어 자원이라는 데에 그 중요성이 있다. 본 논문에서는 표층문에서 통상 격표지로 표현되는 구문적 의존 관계뿐만 아니라, 보어가 갖는 의미역 정보가 부착되어 있으며 시소러스 개념 분류 체계와 연동 가능한 한국어 술어의 하위범주화 사전의 구축에 대해 설명하고 있다. 본 논문에서는 하위범주화 사전의 의미역 표현을 위해 총 25개의 의미역을 설정하고 있다. 이 의미역은 표층 격표지와 직접 연관되어 있기 때문에 통사적인 분석으로부터 직접 의미역 정보를 추출해서 의미 구조의 해석에 이용하는 것이 가능하다. 또한 명사 보어가 갖는 개념의 표현을 위해 상·하위어 관계를 갖는 12만 어휘 규모의 시소러스를 이용하고 있으며, 술어의 의존 관계 표현을 위해 동사, 형용사에 대해 각각 47, 17 개의 하위범주화 패턴을 이용하고 있다. 실용적 규모의 시소러스를 이용함으로써 문장에 나타난 명사의 시소러스 개념을 그대로 하위범주화 사전에 적용시켜 의미 정합 여부를 판단할 수 있는 실질적인 선택제약 체계를 구성할 수 있었고, 표층 격표지에 기초한 표준화된 술어 패턴을 이용함으로써 의미역의 결정 등에서 야기될 수 있는 비일관성을 방지하고 구축에 드는 비용을 절감할 수 있었다. 이상과 같은 방법으로 말뭉치에서 추출한 고빈도 술어 13,000 여개에 대해 하위범주화 사전을 구축하였으며, 적용 범위 평가 실험에 의하면 이 하위범주화 사전은 말뭉치에서 발견된 술어의 72.7%에 대해 하위범주화 정보를 제공할 수 있음을 확인하였다.

Abstract Subcategorization, defining dependency relation between predicates and their complements, is an important source of knowledge for resolving syntactic and semantic ambiguities arising in analyzing sentences. This paper describes a Korean subcategorization dictionary, particularly annotated with semantic roles of complements coupled with thesaural semantic hierarchy as well as syntactic dependencies. For annotating roles, we defined 25 semantic roles associated with surface case markers that can be used to derive semantic structures directly from syntactic ones. In addition, we used more than 120,000 entries of thesaurus to specify concept markers of noun complements, and also used 47 and 17 predicate patterns for verbs and adjectives, respectively, to express dependency relation between predicates and their complements. Using a full-fledged thesaurus for specifying concept markers makes it possible to build an effective selectional restriction mechanism coupled with the subcategorization dictionary, and using the standard predicate patterns for specifying dependency relations makes it possible to avoid inconsistency in the results and to reduce the costs for constructing the dictionary. On the bases of these, we built a Korean subcategorization dictionary for frequently used 13,000 predicates found in corpora with the aid of a tool specially designed to support this task. An experimental result shows that this dictionary can provide 72.7% of predicates in corpora with appropriate subcategorization information.

· 본 논문은 정보통신부 대학기초연구지원사업에서 연구비를 일부 지원받았음

† 종신회원 : (주)코난테크놀로지 이사
yan@konantech.co.kr
yskim@konantech.co.kr

†† 정 회 원 : 인천대학교 정보통신공학과 교수
yswooo@lion.inchon.ac.kr

††† 종신회원 : 한남대학교 정보통신공학과 교수
dhyoon@eve.hannam.ac.kr

논문접수 : 1998년 10월 15일
심사완료 : 2000년 1월 11일

1. 서 론

자연언어를 처리하는 데 있어 문장의 구조를 파악하는 구문 분석에는 필연적으로 많은 중의성이 내재되어 있으며, 특히 한국어는 상황 중심 언어라 불릴 만큼 어순 등과 같은 통사적 특성보다는 상황, 의미, 문맥이 문장의 분석에 더욱 중요한 역할을 한다. 따라서 한국어의 구문 분석은 CFG 형태의 단순 규칙기반 접근 방식으로는 중의성 해소에 한계가 있다고 인정되고 있으며, 이에 따라 지식기반의 언어 분석 방법에서는 이러한 중의성 해결을 위해 술어와 필수 문장 성분의 의존 관계를 나타내는 하위범주화 등의 언어 정보를 대량으로 구축하는 연구들이 진행되어 왔다[1-4]. 하위범주화는 해당 술어가 어떤 필수 성분을 지배하는가를 나타내는 것으로, 영어의 경우에는 성분성(constituency)을 기반으로 한 필수 성분의 절 내에서의 위치 및 전치사 등의 격표지로 결정되며 한국어의 경우는 격조사를 중심으로 한 술어와의 관계로 결정된다. 또한 이 정보에는 명사항과 술어 사이에 의미적 연어 관계에 기초한 선택 제약이 존재하므로, 이를 이용하여 문장의 통사적 모호성 해소에 유용한 정보로 활용할 수도 있다.

하위범주화 사전의 구축은 잘 훈련된 언어적 판단력과 풍부한 경험이 필요하고 대상 술어의 개수가 증가함에 따라 구축 규모도 크게 커지기 때문에, COMLEX, ACQUILEX 등 영어를 대상으로 한 범용 대형 사전 구축 연구의 예에서도 알 수 있듯이 부분적인 자동화를 통해 작업량을 절감하거나[5-8], 혹은 일반적 용도가 아닌 특수한 분야에서 제한적으로 사용할 것을 가정하여 작업량을 절감하기도 한다[9]. 그러나 비록 부분적으로 자동화된 도구를 이용한다 해도, 정확성이 요구되는 많은 부분은 숙련된 편찬자의 수작업에 의존해야 하는 것이 일반적이다[6]. 근래 한국어에 대해서도 다량의 데이타에 기초하여 하위범주화에 대한 특성과 문형을 정리한 통사적 수준의 하위범주화 사전이 발표되고 있으며[2-4], 이의 활용 측면에서의 연구도 있다[10]. 그러나 이러한 연구들의 경우, 예를 들어 보어 성분의 격 표시에 있어서 표층 격조사 수준의 정보만 제공하고 있으므로, 문장 구성 성분의 의미적인 역할의 파악 등 좀더 심층적 수준의 분석에 직접 이용하기는 어려운 실정이다. 또한, 시소러스를 이용한 완전한 개념 분류 체계에 기초하여 구축된 것이 아니고 보어 성분의 개념에 속하는 어휘의 예를 열거하는 방식이기 때문에 명사 개념 사전과 연동되어 사용하기가 어렵다는 문제점이 있다.

한국어의 경우 표층 조사만으로 의미역을 정확하게 결정하는 것은 불가능하지만 의미 분석의 단계를 비교적 통사적인 분석 결과와 가까운 층위로 설정하면 조사와 의미역 간의 직접 사상에 이용할 수 있는 가치있는 정보의 구축이 가능하다는 관점 하에서, 본 논문에서는 기존의 한국어 하위범주화에 관한 연구 결과를 바탕으로 하여 구문적 의존 관계 외에도 의미역(semantic role)이 부착되고 시소러스와 연동이 가능한 실용적인 한국어 하위범주화 사전을 구축하고자 한다.

2. 하위범주화 사전의 정보 및 표현

2.1 관련 연구

하위범주화 사전 구축에 관한 연구는, 용도에 따라 크게 특정 분야나 응용을 지향하여 영역 의존적으로 구축하는 경우[9-10]와 범용의 사전을 지향하여 영역 무관하게 구축하는 경우[2-3, 6-8, 11]로 구분할 수 있다. 후자의 경우는 단기간에 실용성 있는 사전을 구축하는 것이 힘들지만, 일단 구축되고 나면 특정 영역이나 응용에 적합하도록 적응시키는 과정을 통해 여러 가지로 활용하는 것이 가능하기 때문에 본 논문에서는 범용의 한국어 하위범주화 사전을 구축하는 것을 목표로 하였다.

일반적으로 하위범주화 사전에 수록되는 정보는 그림 1과 같이 선택제약에 사용될 수 있는 통사적 의존 관계와 보어의 개념 표지가 일반적이지만[2-3, 10-11], 이러한 정보는 언어 분석에 활용되어야 할 최소한의 것으로서 표층 수준의 의존 관계 분석에 이용하는 것은 가능하지만, 문장 속에서 보어 성분이 담당하는 의미역 등 좀더 깊은 층위의 분석에는 충분하지 않다는 문제가 있으며, 이 의미역 정보는 더 나아가서는 어휘 의미 중의성 해소(word sense disambiguation)를 통한 클러스터링, 또 의미 해석을 위한 성분의 의미 역할 파악 등에 활용할 수도 있는 매우 유용한 정보이므로 본 논문에서는 의미역을 포함하는 하위범주화 사전을 구축하고자 한다.

물론 최근에 하위범주화 사전 구축에 있어서 의미역 정보를 포함해야 할 필요성이 제기되고는 있지만[1, 4, 12], 의미역의 설정이 지나치게 심층적이어서 대용량의 하위범주화 사전을 구축하는 것이 용이하지 않거나[2] 일정 규모 이하의 초기 단계에서 진행되고 있는 것이 현재 실정이다[12]. 일반적으로 의미역은 세분화되고 또한 심층적인 형태로 표현되는 것이 의미 분석 등의 활용에 용이할 수도 있겠지만, 기존의 연구 결과[1]를 보면 지나치게 심층적인 의미역의 구분은 실제 사전 구축이 어려워 진다는 문제가 있다. 한편 한국어의 경우에는

사전구축자가 하위범주 패턴을 결정하거나 실제 구문 또는 의미 분석 시스템에서 사전을 활용할 때도 격조사를 참조하는 것이 필수적이므로, 본 논문에서는 표층 조사와 의미역이 직접 연관이 될 수 있도록 의미역 정보를 구축하였다.

또 한가지 중요한 부분으로, 이 하위범주화 정보를 효과적으로 활용하기 위해서는 하위범주화에 부착된 개념 표지를 실용적인 규모의 개념 분류 체계와 연동시키는 것이 필수적인데, 기존의 하위범주화 사전 연구[2-3, 10-11]를 보면 시소러스를 이용한 완전한 개념 분류 체계에 기초하여 구축된 것이 아니고 보어 개념의 외연, 즉 몇몇 어휘 예를 열거하는 방식이기 때문에 명사 의미 사전과의 연관성이 부족한 경우가 많았다. 본 연구에서는 12만 어휘규모의 명사 개념 계층 시소러스[4]를 의미 표지로 이용하여 하위범주화 보어에 부착될 개념과 연동시켰다. 대규모의 하위범주화 사전과 시소러스를 하나의 체계에서 유기적으로 연동함으로써 각각의 언어 자료가 갖는 정보의 단순 합 이상으로 정보 가치를 극대화하고 이를 토대로 각 응용에 좀더 실질적인 활용이 가능할 수 있게 된다..

```
※ 시스템공학연구소의 하위범주화사전[17]
  57-2. 만나다
  [분류]  : 동사
  [문형]  : 1이 2와 만나다
  [하위범주] 1 : 사람명사(철수, 영희), 동물명사(호랑이, 사자)
             2 : 사람명사(어머니, 친구, 영수), 동물명사(호랑이,
                 사자)
  [의미정보] 다의 :
             부사 : 우연히, 자주, 가끔, 마침내, 드디어
  [문법정보] 1 : 사동┌장 - 만나게 하다
                     └단 -
             2 : 피동┌장 - 만나 지다
                     └단 -
  [구분되어야 할 용언] 1. 동음 :
                       2. 동의 : 만나다
                                 상봉하다
                                 마주치다
※ 홍재성 동사 구문 사전[2]
  가까이하다(자/타)  1.  N1(가)  N2 (을+와)  V
                          N1 :인물  N2:인물
          ...
```

그림 1 기존 하위범주화 사전의 예

2.2 하위범주화 사전의 표현

구문의 관점에서 볼 때, 술어는 보어 성분과 이에 관련된 조사를 수반하므로, 일차적으로는 술어에 정합되는 조사와 보어의 개념이 하위범주화 사전에 수록되어야 한다. 또한 기존의 하위범주화 사전과는 달리, 본 논문에서는 의미역을 제공함을 목적으로 하고 있으므로 각 보어 성분에 대한 의미역 정보도 아울러 수록되어야 한다.

본 논문에서 구축한 하위범주화 사전의 일부가 그림 2에 예시되어 있다[1]. 여기서 [문형] 항목은 해당 술어의 표준 패턴인데, 이 예는 4개의 보어 성분을 가지고 있다. 각 보어 성분에는 격표지로서 여러 가지 조사가 명시되어 있으며, 그 외에도 의미역이 부착되어 있음을 알 수 있다. 그리고 [개념] 항목을 보면, 각 보어에 대해 시소러스 사전의 개념 계층에서 얻어진 하나 이상의 의미 표지가 부여되어 있음을 알 수 있다.

```
[편집번호] : 353  [용언] : 운반하다  [품사] : 동사
[문형] : 1 이[AGT]  2 을[ACC]  3 에서(부터)/(으)(로)부터[SRC]
         4 으로[GOL]
[개념]     1. 인간    교통시설
           2. 인간    물건
           3. 장소
           4. 장소
```

그림 2 하위범주화 인쇄 사전의 예

2.3 의미역 분류

보어 성분이 절 속에서 담당하는 의미적 역할을 정의하는 의미역을 그림 2에서와 같이 사전에 부착하기 위해서는 먼저 구축 목적이나 방법에 부합되는 수준의 의미역을 설정할 필요가 있다. 의미역의 설정에 대해서는, 일반적으로 의미역 정보의 활용 범위와 응용 목적에 따라 다르겠지만 국내외에서도 일부 연구 결과[1]가 제시되고 있다. 본 논문에서 하위범주화 사전의 구축을 위해 정의한 의미역은 표 1과 같으며, 문장의 표층 조사와 심층 역할 간의 연관 관계를 고려하여 의미역을 정의하고 있으므로 정의된 의미역은 표층 조사와 직접 연관성을 갖게 된다. 따라서 본 하위범주화 사전에 부착된 이 의미역 정보를 이용하면 언어의 통사적인 해석으로부터 직접 의미역 정보를 추출해서 의미 구조의 해석에 이용할 수 있다는 장점이 있다.

다음으로, 분류의 세분화 정도에 따라 본 연구의 의미역 분류 체계를 살펴보겠다. 일반적으로 성분이 갖는 의미적 역할을 명확히 구분하기 위해서는 의미역을 세분화할 필요가 있다. 그러나 하위범주화 사전의 구축에서 의미역이 지나치게 세분되어 있으면 의미역의 구분이 어려워져 사전의 일관성, 정확도에 문제가 생길 수도 있고 사전 구축에 소요되는 시간과 비용 역시 문제가 될 수 있다. 따라서 현실적으로는 적정선에서 의미역의

1) 실제로는 관계 데이타베이스로 구축되어 있지만 가독성을 위해 인쇄 형식으로 예시되어 있다.

표 1 하위범주화 사전의 의미역 구분

	의 미 역	설 명	표층 조사 예
1	AGT (agent)	행위자	이/가
2	CHD (characterized)	비행위자	
3	EXS (existent)	존재하는 대상	
4	PSS (possession)	소유된 대상	
5	TRR (transforming result)	변화의 결과	
6	TRS (transforming source)	변화의 출발	
7	FCS (focus)	형용사의 화제	
8	CNT (content)	술어와 필연적인 명사구	
9	ACC (accusative)	술어의 대상이나 목적격	을/를
10	PSS (possession)	소유된 대상	
11	RNG (range)	행동의 범위	
12	TRR (transforming result)	변화의 결과	
13	COM (commitative)	행동의 동반자	
14	SRC (source)	시작점이 되는 대상이나 장소	
15	CNT (content)	술어와 필연적인 목적어	
16	COM (commitative)	동반격	와/과
17	CSE (cause)	원인, 이유	에/에서/에서부터, 로/으로/으로부터
18	GOL (goal)	목표가 되는 장소, 방향, 대상	
19	ELM (element)	요소, 원소	
20	MEN (means)	방법, 수단	
21	INS (instrument)	도구, 기구	
22	STS (status)	자격	
23	TRR (transforming result)	변화의 결과	
24	CNT (content)	술어와 필연적인 내포절	
25	PTH (path)	중간 경유지	
26	RCP (recipient)	수혜자	
27	MGL (mental goal)	정신적인 대상	
28	SCP (scope)	문제의 범위, 대상 영역	
29	SRC (source)	시작점이 되는 대상이나 장소	
30	LOC (location)	공간적인 장소	
31	CHA (characterizing)	주어 성격 규명	
32	QNT (quantity)	수량	

세분화 수준을 결정하는 것이 필요한데, 본 논문에서는 이를 표층 연관 의미역의 수준에서 결정하고 있다. 예를 들어, 본 논문의 행위자격인 AGT는 넓은 의미에서 움직임을 갖는 것, 즉 자발적 또는 피동적인 움직임을 모두 포함하며 소유, 수여 등도 행위로 인정하고 있다. 물론, 아주 심층적으로 구분해 들어간다면 AGT 하나에 대해서도 표 2에 예시된 연구 결과와 같은 세분류가 가능하겠지만[1], 이는 본 하위범주화 사전에서 목적으로 하고 있는 표층 연관 의미역으로는 수용이 어려울뿐더러 결과의 정확도와 일관성에서도 문제가 될 수 있으므로 본 논문에서는 표 1과 같은 중간 층위의 중분류 체계를 택하고 있다. 표 1을 보면, 본 논문에서는 현재 32개의 표층 연관 의미역(의미역 자체만 놓고 보면 25개의 의미역)을 사용하고 있는데, 표층 조사에 따라 10개 내외의 역할 구분만 하면 되므로 비교적 상세하면서도 구분하기도 용이한 의미역 체계를 갖고 있음을 알 수 있다. 또한, 이 의미역은 표준 하위범주화 패턴에 부착되게 되어 사전구축자에게 제공되므로 작업하기가 용이하게 된다.

표 2 행위자격에 대한 심층 분류 연구의 예

의 미 역	설 명	예 문
AGT (agent)	행동주	일꾼이 망치질을 하였다.
EXP (experiencer)	경험자	아기가 잔다.
MAG (mental agent)	정신활동의 주체	나는 등산보다 산보를 좋아한다.
POS (possessor)	소유자	아이들이 폭탄을 가지고 있다.
SPK (speaker)	말하는 사람	존이 영희에게 가자고 제안하였다.
NFO (natural force)	자연력 행위주체	천둥이 아기를 깨웠다.
MVD (moved entity)	이동되는 주체	돌이 언덕 아래로 굴렀다.
AFF (affected)	행위받는 대상	그녀가 병으로 죽었다.
BEN (benefactive)	수혜자	철수가 책을 받았다.

3. 하위범주화 사전 구축

3.1 사전 구축

하위범주화 사전의 구축과 관련하여, 데이타를 획득하는 과정에서 일부 자동화된 도구를 활용하기도 하지만[5-8], 핵심적인 많은 부분은 수작업에 의존하는 것이 일반적[2-4, 6]이다. 예를 들어, 트리부착 말뭉치(tree-tagged corpus)를 활용하여 하위범주 정보를 추출하는 연구[10]도 있지만, 미리 구축된 대량의 트리부착 말뭉치가 없는 한 하위범주화 사전을 만들기 위해 오히려 더 큰 규모의 작업이 될 수도 있는 트리부착 말뭉치를 먼저 구축하는 것이 힘든 만큼 하위범주화 사전 구축의 자동화는 어려운 부분이 많다. 따라서 보조 도구를 활용하여 노동량을 절감하고 핵심적인 부분은 사람의 판단에 의존하여 수작업으로 진행하는 것이 가장 현실적인 방법이 될 것이다. 본 논문에서도 도구 프로그램을 이용하여 반자동적인 방법으로 하위범주화 사전을 구축하였다. 물론, 이러한 방법으로 대용량의 하위범주화 사전을 구축하게 되면 일단 사전구축자의 개인적인 경험 및 지식에 상당 부분을 의존할 수밖에 없다는 문

제점이 있고 여러 명의 사전구축자의 협업 과정에서 비일관성의 문제가 생길 수도 있다. 이러한 문제를 해결하기 위해서는 표준적인 하위범주 패턴을 미리 정의하는 것이 필요하며 이를 하위범주화 사전 구축 도구에서 제시함으로서 일관성 있는 결과를 얻을 수 있다.

실제 구축 시에는 [4]의 한국어 품사 부착 말뭉치를 기준으로 13,000여 개의 술어를 선정하여 하위범주 사전을 구축하였다. 출현 빈도가 높고, 중의성이 많은 어휘를 우선적으로 선별하였고, 피동형이나 사역형의 술어는 활용도가 특별히 높다고 판단되지 않는 한 배제하였다. 피동이나 사동 등의 변형이 일어나게 되면 하위범주 성분이나 의미역이 바뀌게 되지만 패턴에 따라 일정한 것이 일반적이므로 다음에 설명할 패턴변형 규칙으로 처리하는 것이 가능하기 때문이다. 또한 보조용언, 계사(copula) 혹은 전성 접미사 (예를 들어, '-하다, -되다') 등에 의한 파생 술어도 본 연구에서 제외하였다.

3.2 개념 및 시소러스

각 보어 성분에 해당하는 명사의 개념을 할당하는 데 필요한 주요 언어 지식원은 시소러스 사전이다. 한국어 명사의 계층적 개념 분류를 가지고 있는 한국어 시소러스는 국내 일부 연구에서 소규모로 구축된 바 있으며, 일본의 유의어신사전과 같은 계층 분류 사전이나 미 프린스턴 대학의 WordNet과 같은 동의어 계층 사전 등을 한국어에 적용하는 연구도 최근에 수행되고 있다[13-15]. WordNet은 최상위 노드부터 하부 노드쪽으로 작업자의 직관에 의해 구축해 가는 "하향식" 방법으로 구축되었는데, 일반적으로 이 방법은 작업자의 어휘 지식의 보유량의 차이와 언어적 직관에 따른 선택의 임의성이라는 문제가 있고, 많은 부분을 수작업에 의존할 수밖에 없는 방법이므로 막대한 시간과 비용이 소요되는 단점이 있다. 이와 반대로 사전 뜻풀이 문에서 상하위어 관계를 추출하여 점진적으로 계층 구조를 만들어 나가는 "상향식" 방법에서는 비일관성, 순환, 비적합성 등이 문제로 지적되고 있다[13-14]. 본 논문에서 이용하는 시소러스는 이러한 문제를 해결하기 위해, 하향식으로 초기 계층을 구성하고 뜻풀이 문을 적용하여 시소러스를 확장하는 상향식 단계로 이루어지는 "다단계" 시소러스 구축 방법[13]을 사용해 구축된 것으로서 어휘(=노드) 수 약 12만, 최대 깊이 9, 평균 깊이 5의 계층 구조를 가진 시소러스이다[4]. 또한 이 시소러스는 표 3에 예시된 바와 같이 개념간 거리를 계산하기 쉬운 접두어식 개념 계층 구조로 되어 있으며 상위어 노드를 여러 개 가질 수 있는 그래프 형태로 되어 있다

이 시소러스를 이용하면 문장에 나타난 명사의 시소러스 개념과 하위범주화 사전에 부여된 개념이 정합되는가를 판단할 수 있는 선택제약 체계를 구성할 수 있으므로 의미적 선택을 가능하게 하는 유익한 정보원을 구성할 수 있다. 이를 위해 시소러스의 개념 체계와 하위범주화 사전에 부착된 개념 정보를 사전 구축 과정에서 연동시키는 것이 중요함은 물론이다. 이러한 관점에서 볼 때, 잘 정의된 시소러스가 가용한지 여부는 향후 하위범주화 사전의 활용을 고려해 볼 때 하위범주화 사전의 구축과 별개로 매우 중요한 문제가 됨을 알 수 있다. 이러한 목적을 위해서는 시소러스가 상·하위 개념 정합을 위해 계층구조로 구성되어 있어야 함은 물론이고, 상·하위어 관계뿐만 아니라 개념간 거리를 판단하기 쉬운 형태로 기술되어야 한다. 본 논문에서 이용하는 시소러스의 접두어식 개념 계층 표기 방법은 이러한 목적에 잘 부합된다고 할 수 있다. 예를 들어, 'A00A00A00B00G00C'는 '물(water)'에 대한 개념 표지인데, 표 3에서 알 수 있듯이 최초 A는 '구조근원'(root) 노드를 의미하고, 다음부터 3자리씩 계층 구조상의 자식 노드를 가리키게 된다. 이를 통해, 'A00A00A00B00G00C'를 해독해 보면 '구조근원-구체물-무생물-물체-자연-물'의 계층 구조를 가짐을 알 수 있다. 따라서 이 접두어식 의미 태그만으로도 상위어-하위어 관계와 계층 구조상에서의 깊이를 모두 해독할 수 있으므로 하위범주화의 개념과 문장에 나타난 명사의 시소러스 개념의 상·하위어 정합을 위한 수단을 제공하고 있음을 알 수 있다.

표 3 접두어식 개념 계층 시소러스의 예

표제어	개념 표지
구조근원	A
구체물	A00A (구조근원-구체물)
무생물	A00A00A (구조근원-구체물-무생물)
생물	A00A00B (구조근원-구체물-생물)
물체	A00A00A00B (구조근원-구체물-무생물-물체)
자연	A00A00A00B00G (구조근원-구체물-무생물-물체-자연)
물	A00A00A00A00F (구조근원-구체물-무생물-물건-물)
물	A00A00A00A00M00J00A (구조근원-구체물-무생물-물건-속성물-사물-물)
물	A00A00A00B00G00C (구조근원-구체물-무생물-물체-자연-물)
물	A00C00I00I00b00A00D (구조근원-추상물-속성-성질-색채-빛깔-물)
물	A00C00I00R00G00P(구조근원-추상물-속성-질-정도-물)

표 4 동사의 하위범주 패턴

V1	(이 AGT)	V25	(이 AGT) (을 ACC) (에 GOL)
V2	(이 AGT) (에서 LOC)	V26	(이 AGT) (을 ACC) (에게 RCP)
V3	(이 EXS) (에/에서 LOC)	V27	(이 AGT) (을 ACC) (에게<서>/에서/로부터 SRC)
V4	(이 AGT) (에 MGL)	V28	(이 AGT) (을 ACC) (에 CSE)
V5	(이 AGT) (로/에/때문에 CSE)	V29	(이 AGT) (을/로 TRR)
V6	(이 CHD) (에 GOL)	V30	(이 AGT) (을 ACC) (으로 QNT)
V7	(이 CHD) (에 LOC)	V31	(이 AGT) (을 ACC) (으로 STS)
V8	(이 CHD) (로 ELM)	V32	(이 AGT) (에게 RCP) (에대하여 MGL)
V9	(이 AGT)(에서 SRC)(으로<해서> PTH)(에/으로 GOL)	V33	(이 AGT) (에게 RCP) (-고 CNT)
V10	(이 TRS) (이/으로 TRR)	V34	(이 AGT) (에게 RCP) (-다고/마고 CNT)
V11	(이 CHD) (이 CNT)	V35	(이 AGT) (와/에게 ACC) (-기로 CNT)
V12	(이 AGT) (을 ACC)	V36	(이 AGT) (-기로 CNT)
V13	(이 AGT) (을 ACC) (으로 INS)	V37	(이 ACC) (-다고 CNT)
V14	(이 AGT) (을 ACC) (으로 MEN)	V38	(이 AGT) (에게/을 RCP) (-라고/하라고/게/도록 CNT)
V15	(이 AGT) (을 CNT)	V39	(이 AGT) (에게 RCP) (-자고 CNT)
V16	(이 AGT) (을 CNT) (에 MGL)	V40	(이 AGT) (에게 RCP) (-느냐고/는가를/는지를 CNT)
V17	(이 AGT) (을 RNG)	V41	(이 AGT) (을 ACC) (에서<부터>/로부터 RC) (으로 GOL)
V18	(이 AGT) (을 SRC)	V42	(이 CHD) (와 COM) (이 CNT)
V19	(이 AGT) (을 PSS)	V43	(이 CHD) (에게 RCP) (이 CNT)
V20	(이 AGT) (을 COM)	V44	(이 CHD) (로 ELM)
V21	(이 AGT) (와 COM)	V45	(이 AGT) (을 ACC) (을 CNT)
V22	(이 CHD) (을 COM)	V46	(이 AGT) (을 ACC) (으로 ELM)
V23	(이 CHD) (와 COM)	V47	(이 AGT) (을 CNT) (에대하여 MGL)
V24	(이 AGT) (을 ACC) (와 COM)		

표 5 형용사의 하위범주 패턴

A1	(이 CHD) (에게 RCP)	A10	(이 PSS)
A2	(이 CHD) (에서 SRC)	A11	(이 EXS) (로 STS) (에 LOC)
A3	(이 CHD) (에 MGL)	A12	(이 EXS) (이 QNT)
A4	(이 CHD) (이 QNT)	A13	(이 AGT) (-을까/을지/은지/은가 CNT)
A5	(이 CHD) (이 FCS)	A14	(이 AGT) (이 CHD) (-기에/기가/하기가 CNT)
A6	(이 CHD) (와 COM) (에서 SCP)	A15	(-기가/하기가 CNT)
A7	(이 CHD) (와 COM) (이 FCS)	A16	(이 CHD) (=이다/이=아니다 CHA)
A8	(이 EXS) (에 LOC)	A17	(이 CHD) (이 CHA) (에서 SCP)
A9	(이 EXS) (에게 RCP)		

3.3 술어 패턴

술어의 하위범주화 정보를 구축할 때 어떤 성분을 보어로 해야 하는지를 결정하는 것 역시 어려운 일이다. 또한 보어 성분이 술어에 대해 어떤 의미역을 갖는지를 결정하는 것도 간단하지가 않으므로 비일관성의 한 원인이 될 수 있다. 이렇게 사전구축자에 따라 임의로 하위범주 보어 성분을 결정함으로써 야기되는 비일관성의 문제와, 아울러 유사한 판단을 반복해서 해야 하는 작업

량의 과부하를 방지하기 위해서는 한국어 술어가 가질 수 있는 하위범주화 패턴을 미리 정의하여 표준화하는 것이 일관성과 경제성 측면에서 바람직하다. 물론 한국어 술어가 갖는 모든 공통 패턴을 미리 정의하는 것이 원칙적으로 어려운 일이지만, 이 표준 패턴은 하위범주화 사전의 작성에 있어 상기한 바와 같은 문제점을 해소해 주는 중요한 수단이기 때문에 실제 구축 작업에서는 일상적인 언어에서 빈번히 쓰이는 유형들에 대해서만 정의할 수 있어도 매우 유용하게 활용될 수 있다. 본 논문에서는 동사 41개, 형용사 17개로 총 58개의 하위범주 패턴을 정의하여 사용하고 있다. 영어의 경우, Hornby 사전의 술어 패턴 분류[16]를 보면 실용적으로 27개 정도의 술어의 패턴만을 정의하고 있으며 어휘형을 포함하고 있는 패턴을 모두 집계하더라도 53개의 술어 패턴에 불과하므로 본 논문의 패턴 분류도 각 하위범주 패턴을 표현하기에 충분한 정도의 세분류임을 알 수 있다.

표 4와 표 5는 본 논문에서 정의한 동사와 형용사의 하위범주화 패턴의 예이다. 각 하위범주화 패턴은 해당 패턴의 술어가 가질 수 있는 보어에 대해 대표 표층 조사와 표층 연관 의미역이 명시되어 있어서 실제 구축될 하위범주화 데이타의 원형이 된다. 술어가 가지는 하위범주 성분은 최대 4개까지를 허용하였다.

일반적으로 하위범주화 사전은 구문 분석 등의 응용에 적용되므로, 하위범주화 사전과 정합되는 술어는 파생형이 아니라 기본형이 된다. 따라서, 피동형이나 사역형의 술어는 출현 빈도가 매우 높은 경우를 제외하고는 하위범주화 구축 대상에서 제외하고 있지만, 실제 말뭉치를 조사해 보면 피동이나 사역형 술어도 자주 출현함을 알 수 있다. 이러한 경우에 변형 술어가 지배하게 되는 보어의 표층 격표지도 달라지게 되므로 기본 하위범주화 패턴의 활용이 문제가 된다. 예를 들어 (1)에서 기술한 바와 같이 기본형과 피동형의 하위범주화 패턴을 비교해 보면, 의미역은 동일하지만 대응하는 표층 조사가 달라지는 것을 알 수 있다. 본 논문에서는 표 4, 5에서 정의한 하위범주 패턴에 대응하는 피동형, 사역형의 패턴 변형 규칙을 도입하는 것이 하위범주화 사전에 피동형, 사역형을 모두 등록하는 것보다 효율적이라고 판단하여, (2)와 같은 패턴 변형 규칙을 구축하였다. (2)에서 CSA는 사역 주어를 나타낸다.

(1) 기본형: 사랑하다 (이 AGT) (을 ACC)
예) 철수가 영희를 사랑한다.
피동형: 사랑받다 (에게 AGT) (이 ACC)
예) 영희가 철수에게 사랑받는다.

(2) 기본형: V12 (이 AGT) (을 ACC)
피동형: $V12^{P}$ (에게 AGT) (이 ACC)
사역형: $V12^{C}$ (이 CSA) (에게 AGT) (을 ACC)

4. 구축 및 평가

이상에서 논의한 시소러스, 의미역, 하위범주화 패턴을 바탕으로 해서 한국어 술어의 하위범주화 사전을 구축하는 과정에 대해 설명하겠다. 기존의 하위범주 연구 결과를 참고하고, 한국어 인쇄 사전과 말뭉치를 활용하여 하위범주화 사전을 구축하였다.

4.1 구축 작업

사전 구축에 소요되는 작업량을 줄이고, 특히 작업자간의 일관성을 유지하기 위해서는 사전 구축을 위한 도구의 개발이 필수적이다. 본 논문에서는 사전 구축 도구를 구현하고 이를 하위범주화 사전 데이타베이스 구축 과정에 활용하였다. 하위범주화 사전 구축 및 편집을 위한 도구는 술어의 표제어 입력이나 선택 시에 이미 구축된 하위범주화 사전의 데이타를 보여주고 이와 가장 유사하다고 판단되는 표준 패턴을 사전구축자에게 제시해 줄 수 있도록 설계되어야 한다. 또한 한국어의 경우에는 중의성이 있는 어휘나 여러 개의 패턴을 가지는 술어가 대부분이므로 이미 입력되어 있는 동일 표제어의 어휘 정보를 표시하여 중복 입력과 같은 문제를 방지하고 새로운 데이타 입력에 참고가 되도록 해야 한다. 실제 도구에서는 하위범주화 사전을 비롯하여 기타 정보들을 저장하기 위해 데이타베이스를 이용하였다. 이 도구는 대표 조사, 확장 조사, 의미역, 패턴 등을 가지고 있어서 작업자가 제시된 패턴 중 하나를 선택하면 이에 따라 하위범주 레코드의 기본 정보를 채워 작업자에게 제시해 주도록 하였다. 작업자는 시소러스를 이용하여 개념을 결정해서 하위범주화 데이타베이스에 입력하면 된다. 이와 같이 도구를 활용하는 반자동의 방식으로 13,000개의 고빈도 술어에 대해 하위범주화 사전을 구축하였다.

4.2 하위범주화 사전의 평가 및 응용

본 논문에서 구축한 결과물인 하위정보화 사전이 언어처리를 위한 자원으로서 유용한가의 여부는 이 결과물이 실제 말뭉치에서 어느 정도의 적용 범위를 갖는가를 조사함으로써 알 수 있다. 이를 위해 5,000여 문장으로 구성된 품사부착 말뭉치의 술어에 대해 하위범주화 사전을 활용하여 문장의 절과 구를 분리해 내는 의존구조 분석을 하고 이 결과를 수작업으로 확인하여 적용

범위를 조사하였다.

평가 대상 문장 집합은 21,923개의 술어를 포함하고 있으며, 이 중 보조용언, 계사나 접미사로 인한 파생 용언을 제외한 동사와 형용사는 14,372개이다. 조사 결과, 이 중에서 10,596 (10,596/14,372 = 73.7%) 개의 술어가 하위범주화 사전에 등록되어 있었으며, 하나의 보어 성분이라도 하위범주화 사전과 정합된 술어를 조사하면 그 개수가 7,705 (7,705/10,596 = 72.7%) 개이었다. 이 경우가 하위범주화 사전이 적용된 범위라고 평가할 수 있다. 나머지 2,891 (2,891/10,596 = 27.3%) 개의 술어는 하위범주화 사전에는 등록되어 있지만 단 하나의 보어 성분도 정합되지 않은 경우이다. 여기에는 "예쁜 순희가 ..."와 같이 아무 하위 성분도 없이 관형절을 이루는 술어의 경우와 개념 표지의 정합이 실패한 경우가 있는데, 수작업으로 확인해 본 결과 각각 2,172 (2,172/2,891 = 75.1%) 개, 719 (719/2,891 = 24.9%) 개 정도인 것으로 파악되었다. 후자의 경우가 실제 본 논문의 하위범주화 사전이 실패한 경우이며 시소러스와의 연계성을 높이는 작업을 통해 보완되어야 할 부분이다.

한편, (3)에서 알 수 있듯이 표층 문장에서 동일 성분의 생략으로 인해 한 명사가 여러 개의 술어에 정합이 가능하거나 두 명사가 모두 한 술어의 보어가 될 수 있으므로, 하위범주 데이타 자체만으로는 의존 관계를 찾아내기가 곤란한 경우가 많다. 이러한 문제는 하위범주화 정보가 전체 문장이 아닌 절 단위의 정보라는 점과 또한 절 단위의 의존 관계를 분석하기 위한 충분 조건이 아닌 필요 조건이라는 점을 고려해 볼 때, 하위범주화 사전에서 더 많은 양의 정보를 제공한다고 해서 해결될 수 있는 문제가 아니므로, 생략, 삽입, 이동 등의 변형 현상을 좀 더 심층적으로 다룰 수 있는 수준의 분석 방법 및 언어 자원이 요구되는 부분이다. 수작업으로 검토해 본 결과 이러한 경우가 정합에 성공한 것 중에서 약 15% 정도 됨을 확인하였다.

(3) 가. 철수가 갔다고 말했다.
나. 철수가 영희가 갔다고 말했다.
다. 철수가 영희가 갔다고 말한 적이 있는지 몰랐다.

평가를 수행한 결과, 구축된 하위범주화 사전이 말뭉치에서 발견된 술어의 상당수에 대해 하위범주화 정보를 제공해 주고 있음을 알 수 있었다. 또한 구조 분석 등의 응용에 유용한 정보원으로 이용될 수 있음을 확인하였으며, 더 나아가서는 보조사의 기능적 모호성 해소 및 의미 관계 결정에 이용할 수도 있다. 그러나 구조적 모호성 해소를 위해서는 등위/종속 접속으로 인한 성분의 생략 및 공유 등 구문 구조의 변형에 대처할 수 있어야 하므로, 하위범주화 사전 정보의 구축 및 유연한 적용과 함께 시소러스나 기타 분석에 필요한 사전 그리고 분석 방법론 등에 대한 모색이 동시에 진행되어야 한다.

5. 결 론

이상에서 한국어 술어의 하위범주화 사전 구축에 대해 논의하였다. 하위범주화는 술어와 보어 간의 의존 관계를 정의하는 언어 정보로서 구문 및 의미 분석 등에 폭넓게 활용될 수 있는 기반 언어 자원이라는 데에 그 중요성이 있다. 본 논문에서는 표층 격표지로 표현되는 구문적 의존 관계뿐만 아니라 특히 보어가 갖는 의미역 정보도 부착된 한국어 술어의 하위범주화 사전을 구축하였다.

명사 보어가 갖는 개념의 표현을 위해 상·하위어 관계를 갖는 12만 어휘 규모의 개념 계층 시소러스를 이용하였고, 의미역 표현을 위해 총 25개의 의미역을 설정하였다. 또한, 여러 사전구축자가 공동으로 하위범주화 사전을 구축하면 비일관성 문제가 발생할 수 있으므로, 동사와 형용사에 대해 각각 47, 17개의 표층 조사를 중심으로 한 표준 패턴을 정의하였다. 이때 각 보어 성분에 대해 의미역 정보도 함께 부여하여, 하위범주 패턴이 구문적으로는 물론 의미적인 의존 관계를 파악하는 데에도 적용될 수 있도록 하였다. 이상을 기반으로 고빈도 술어 13,000여 개에 대해 하위범주화 사전을 구축하였다. 적용 범위의 평가 실험에 의하면, 이 하위범주화 사전은 말뭉치에서 발견된 술어의 72.7%에 대해 하위범주화 정보를 제공할 수 있음을 확인하였다.

한편, 실제 활용 측면에서 볼 때 하위범주화 사전에 통계적인 값을 부여하는 것이 필요하다는 견해[10]도 있다. 다시 말해서, 하위범주 패턴에 대한 빈도 정보, 연어 정보, 의미역-술어의 공기 정보 등도 추출하여 하위범주화 사전에서 제공하는 정보를 확장하는 것도 하위범주화 사전의 실용화를 위해 향후 시도할 만한 연구가 될 것이다. 또한 하위범주화 사전의 규모가 커질수록 사전 편찬자의 언어 직관이나 경험에 의존하여 하위범주화 정보를 획득하는 것이 더 어려운 문제가 된다. 이 사전을 활용하여 구문 태깅 및 의미 태깅(sense tagging)을 수행하고, 태깅 결과를 이용하여 하위범주화 사전의 확장이나 추가적인 보완에 사용한다면 부분적으로 자동화된 통계적인 정보의 획득 및 부가도 가능할 것이다.

참 고 문 헌

[1] 장석진외, *자연언어처리의 기초연구*, 한국과학재단 보고서, 1989

[2] 홍재성, *현대 한국어 동사 구문 사전*, 두산 동아, 1997

[3] 김봉모, *한국어 문장 분석을 위한 하위범주화사전*, 국어공학센터/시스템공학연구소 보고서, 1996

[4] 서영훈외, *토큰 기반 한국어 분석기 개발- 한국어 의미 분석 사전 및 하위범주화 사전구축*, 한국전자통신연구원 보고서, 1998

[5] W. Peters, "Corpus-based Conceptual Characterisation of Verbal Predicate Structures," *Proc. of the Computational Linguistics in the Netherlands*, Antwerpen 1996.

[6] Ralph Grishman, Catherine Macleod, Adam Meyers, "Comlex Syntax: Building a Computational Lexicon," *Proc. of COLING-94*, pp. 268-272, 1994.

[7] Antonio Sanfilippo, Vector Poznanski, "The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources," *Proc. of ANLP-92*, pp. 80-87, 1992.

[8] Ted Briscoe, John Carroll, "Automatic Extraction of Subcategorization from Corpora," *Proc. of ANLP-97*, 1997

[9] Hideki Tanaka, "Verbal Case Frame Acquisition from a Bilingual Corpus: Gradual Knowledge Acquisition," *Proc. of COLING-94*, pp. 727-736, 1994.

[10] 박재득외, *국어정보처리기술 개발- 한글 언어처리 기반기술*, 한국전자통신연구원 부설 시스템공학연구소, 과학기술처 보고서, 1997

[11] 박동인외, *국어정보처리기술 개발- 지능형 처리기 개발*, 한국전자통신연구원 부설 시스템공학연구소, 과학기술처 보고서, 1997

[12] 홍재성외, *21세기 세종계획- 전자사전 개발*, 문화관광부 보고서, 1998

[13] 이종인, 한광록, 양승현, 김영섬, "한국어 명사의 시소러스 구축을 위한 시스템 설계 및 구현", *한국정보처리학회 논문지*, 제6권, 2호, 1999.

[14] 조평옥, 옥철영, "의미속성에 기반한 한국어 명사 의미 체계", *한국정보과학회 논문지 (B)*, 제26권, 4호, 1999.

[15] 문유진, "한국어 명사를 위한 WordNet의 설계와 구현", *한국정보과학회 논문지(B)*, 제23권, 4호, pp. 437-444, 1996.

[16] A. S. Hornby, *Guide to Patterns and Usage in English*, 2nd Edition, Oxford University Press, 1975.

[17] Rebecca Bruce, Janyce Wiebe, "Word-Sense Disambiguation Using Decomposable Models," *Proc. of ACL-94*, pp. 139-145, Jun 1994

양 승 현

1990년 서울대 공과대학 컴퓨터공학과 학사. 1992년 서울대 대학원 컴퓨터공학과 석사. 1997년 서울대 대학원 컴퓨터공학과 박사. 1997년 ~ 1999년 한국전자통신연구원 선임연구원. 1999년 ~ 현재 (주)코난테크놀로지 이사 겸 부설 멀티미디어정보기술연구소장. 관심분야는 교차언어 정보검색, 멀티미디어 정보검색, 문서 분류, 문서 요약, 기계번역, 자연언어처리, 진화 알고리즘

김 영 섬

1983년 한양대학교 전자통신공학과 공학사. 1985년 한양대학교 공과대학원 공학석사. 1989년 한양대학교 공과대학원 공학박사. 1989년 ~ 1999년 한국전자통신연구원 선임연구원. 1992년 ~ 1993년 미국 Bellcore Basic Research Center 객원연구원. 1999년 ~ 현재 (주)코난테크놀로지 대표이사. 관심분야는 디지털 아카이브, 멀티미디어 정보검색, 교차언어 정보검색, 기계번역, 자연언어처리

우 요 섭

1986년 한양대학교 공과대학 전자통신공학과 학사. 1988년 한양대학교 대학원 전자통신공학과 석사. 1992년 한양대학교 대학원 전자통신공학과 박사. 1992년 ~ 현재 인천대학교 정보통신공학과 부교수. 관심분야는 자연언어처리, 의미 태깅, 멀티미디어 정보검색, 기계번역

윤 덕 호

1985년 서울대 공과대학 컴퓨터공학과 학사. 1987년 서울대 대학원 계산통계학과 석사. 1993년 서울대 대학원 컴퓨터공학과 박사. 1998년 ~ 1999년 CMU 객원연구원. 1989년 ~ 현재 한남대학교 정보통신공학과 부교수. 관심분야는 문서 분류 및 요약, 기계번역, 자연언어처리, 멀티미디어 정보검색