한국어 형용사 의미계층의 전산적 추출

송 상 헌
고려대학교 언어학과

최 재 웅
고려대학교 언어학과

The Computational Extraction of Semantic Hierarchies for Korean Adjectives

Song Sanghoun
Dept. of Linguistics
Korea Univ.

Jae-Woong Choe
Dept. of Linguistics
Korea Univ.

요 약

자연 언어의 각 어휘는 서로 관계를 가지고 계층적·입체적 모델로 존재한다. 이러한 전제에서 출발한 연구 가운데 대표적인 것이 의미 계층이다. 본고에서는 한국어 형용사의 의미 계층을 추출하는 것을 목표로 하여, 형식적·객관적 방법론을 정립하고, 결과를 비교적 신속하고 정확하게 이끌어 낼 수 있는 전산적 처리 도입하였다. 우선 전체 구축에 필요한 절차를 세우고 각 단계에서 필요한 방법과 휴리스틱을 정리하였다. 이를 바탕으로 사전 뜻풀이말을 이용하여 반자동으로 작업하였으며, 일부 코퍼스를 활용하였다. 최종 알고리즘으로는 Top-Down 방식을 택하였다. 이렇게 추출된 한국어 형용사 의미 계층은 226개의 최상위어에서 시작하여 총 3,792개의 표제어를 망라한다. 또한 수직적 계열 관계만을 명시했을 경우 나타날 수 있는 한계를 보완하기 위해, 동의어·반의어와 같은 수평적 의미 관계와 공기 명사와 같은 결합 관계 등을 함께 기술하였다. 한편 표제항을 뜻풀이말의 공기 명사를 이용하여 의미별로 분류하고, 각 분류마다 별도의 의미 계층을 수립하였다.

1. 서론

일반적으로 어휘는 하나 또는 그 이상의 개념을 담고 있다. 이때, 이 어휘와 개념의 관계에서 [어휘, 개념]의 집합이 낱낱으로 존재하는가 아니면 개념의 속성에 따라 서로 연관 관계를 맺고 구조적으로 존재하는가 하는 근본적인 의문이 제기된다[14].

본고에서는 어휘는 각 의미적 체계를 지니고 계층적, 입체적 모델로 존재한다는 견해를 기본 전제로 삼고자 한다. 이러한 전제에 따르면 자연 언어에 존재하는 어휘를 일련의 개념 연속선상에서 순차적으로 나열할 수 있다. 그 나열된 계층 목록을 의미 계층이라고 한다.

이러한 의미 계층은 자연어처리에서 흔히 문제가 되는 중의성을 해소하는데 필요한 중요한 기재일 뿐만 아니라 정보검색에서 적절한 문서를 추출하는 데에도 응용 가능성이 큰 자원이다. 뿐만 아니라, 언어 연구 전반에 기여할 수 있다.

의미계층에 대한 선행연구로 대표적인 것이 WordNet ([16]참조)이다. 이는 영어의 명사, 동사, 형용사, 부사 등에 대하여 의미적 연결 관계를 구축한 동의어 집합으로 자연어처리 제 분야와 언어학 연구에서 폭넓게 사용되고 있다. [11]을 비롯한 국내 연구진도 의미 계층에 대한 자원 확보를 위해 노력한 바 있다.

종래의 이러한 의미 계층 연구는 주로 명사를 대상으로 이루어져 왔으며, 용언에 대해서는 주로 동사를 대상으로 하였다. 연구의 최종 목표는 한국어 용언 전체의 위계 구조를 이끌어 내는 것이어야 하겠으나, 본고에서는 우선 형용사를 대상으로 한다. [5]에서는 '크다, 작다, 많다, 적다' 의 형용사 어휘를 대상으로 논항 의미부류 표준화를 연구하는 데, 이들을 대상으로 선정한 이유에 대해 "형용사가 동사에 비해 논항구조가 단순하여 논항의 의미를 비교하기에 용이하다"고 하였다. 즉, 형용사의 구조를 밝히는 일이 동사의 구조를 밝히는 일보다 상대적으로 부담이 적다고 하겠다. 따라서 우선 형용사를 대상으로 하여 용언 의미 계층 구조를 구성하는 방법론을 모색하고, 여기서 검증된 방식을 바탕으로 전체 용언에 확대하면 순차적이고 안정적인 연구가 될 것으로 기대한다.

한편 한국어의 내용어 가운데 명사나 동사의 경우엔 앞선 연구에서 상당 부분 이루어 졌으나 그동안 상대적으로 소홀히 여겨져 온 것이 형용사이다. 본고는 형용사에 대한 의미계층을 도출해 내는 연구를 통해 한국어 어휘사이의 의미 관계에 대한 보다 종합적인 파악에 기여하고자 한다.

2. 선행 연구

2.1. 전산적 처리를 통한 의미 체계 연구

의미 계층과 종종 동의어처럼 여겨지는 의미망이란 말 그대로 단어들의 의미 관계를 Network의 형태로 가공한 것을 말한다. WordNet은 각 단어의 개념사이에 동의어, 반의어, 의미간의 상·하위, 부분·전체 등의 의미 관계들을 설정한다.

한국어에 적용된 의미망으로는 한국과학기술원 전문용어언어공학연구센터에서 개발한 카이스트 한국어 어휘의 미망과 울산대학교 한국어처리연구실에서 개발한 어휘지능망 U-WIN 등이 있다.

위에서 언급한 의미망을 비롯한 어휘 지식 베이스(Lexical Knowledge Base; 이하 LKB)는 최근 컴퓨터를 이용한 구축이라는 방법론을 택하고 있다. 그 이유는 컴퓨터 사양의 향상으로 인한 대용량 언어 자료 처리 가능성 증대, 수동 구축에 따르는 시간과 노력의 손실 최소화, 객관적 기준 적용의 필요성 등이 있다.

전산적 처리를 통한 의미 체계 연구에서 주로 활용되는 방법 중의 하나가 이른바 기계 가독형 사전(Machine Readable Dictionary; 이하 MRD)의 뜻풀이말을 이용하는 것이다. 사전 편찬자는 개발하고자 하는 사전의 여러 변수 등을 고려하여 사전을 설계하고 구축한다. 그렇게 구축된 사전은 언어 연구에 유용한 지식원을 담고 있을 뿐만 아니라 언어의 여러 현상을 폭넓게 반영하고 보여준다. 즉, 사전에 명시된 정보만을 잘 활용하여도 그 언어 어휘 의미의 기본적 구조를 이끌어 낼 수 있다.

[3]에서는 단어와 단어 사이를 연결해 주는 의미 관계 LKB를 구축하였다. 이를 위하여 MRD 형태인 온라인 사전을 가공하여 구축된 LDB(Lexical DataBase)를 재가공하는 방식을 택하였다. 즉, 사전 뜻풀이말의 의존 구조를 분석하고, 그에 의존 구조 패턴을 적용하여 'is-a' 관계를 도출하였다.

[10]에서는 명사 어휘의미망을 구축하기 위해 MRD의 뜻풀이말을 이용하였다. 사전 뜻풀이말의 첫 번째 문장에서 가장 뒷부분에 오는 명사를 상위어로 간주하고, A의 하나', 'A의 일부', 'A의 한 갈래' 등의 패턴일 때는 A를 상위어로 간주하였다. 이 경우 의미 하나하나에 올바른 개념번호를 부착하기 위해 의미구분이 필요한데 이 과정은 수동으로 진행하였다.

[11]에서는 국어사전의 명사에 대한 뜻풀이말을 이용하여 Bottom-Up 방식으로 '한국어 명사 의미 계층 구조'를 구축하였다. 이는 트리가 43개, 중간 노드가 2,443개, 단말 노드가 10,347개이며, 깊이가 17인 하나의 포리스트(forest)를 이룬다.

[13]에서도 사전을 기반으로 한 한국어 의미망의 구축을 시도하였다. 국어학적인 의미 관계를 이용한 상하 관계를 기본으로 하되 보완적 측면에서 동의·유의 관계, 부분·전체 관계, 반의 관계 등을 추가로 정의하였다.

2.2. 한국어 형용사

한국어 형용사에 대한 논쟁은 주로 과연 한국어에 형용사가 있는가 하는 문제이다. 일부에서는 한국어의 형용사는 동사의 하위 유형이라고 주장하며, 다른 학자들은 한국어 형용사는 동사와 엄연히 구분되는 분포적·통사적 특질을 지니고 있다고 주장한다. 이점은 형용사 의미 계층을 구성할 경우 어디까지를 형용사로 인정하여 다루어야할 것인가의 문제와 관련된다. 본고에서는 어떠한 표제어가 '동사다' 혹은 '형용사다'를 규정하기 보다는 그 어휘의 쓰임이 구체적으로 어떠하며 다른 어휘와 어떠한 관계를 맺고 있느냐를 중심으로 대상을 판별키로 한다.

한국어 형용사는 한편 동의[1]·반의[2]관계가 명사를 비롯한 다른 품사와 차이[3]가 있으며, 다의성이 매우 높다는 특성을 지닌다[6].

3. 방법론 설정

본고에서는 사전에 등재된 형용사 표제어 목록을 뜻풀이말을 이용하여 계층별로 정리하는 것을 1차적 목표로 한다. 표제어 목록을 나열하고 구성하는 방법론을 세우기 위해 기본적으로 고려해야할 사항을 정리하면 아래와 같다.

(1) 의미 계층은 두 가지 차원에서 구성한다. 하나는 전체 형용사 표제어를 대상으로 사전 의미 구분 없이 직접 분류를 시도하는 것이며, 다른 하나는 우선 각 의미 영역별로 어휘들을 구분한 뒤에 각 영역 내에서 계층 구조를 판별하는 것이다.

(2) 대상이 되는 표제어를 우선 선정한다. 이때, 표제어 선정은 형식적인 기준을 따르는 것으로 한다.

(3) 사전 뜻풀이말을 이용하여 상위어, 동의·반의어, 공기 명사 등을 추출하되, 필요할 경우 부족한 부분은 코퍼스를 통해 보완하도록 한다.

(4) 사전 뜻풀이말의 유형을 우선 정리하고 그 유형에 맞는 처리를 연구자의 직관 개입을 배제한 형식적 기준에 의해 시행한다. 즉, 선결된 유형별 정리 방법을 충실히 따르는 것으로 한다.

(5) 위 (4)에서 각 엔트리의 의미 구분은 수작업을 통해 하도록 한다. 이때, 작업을 원활히 수행하기 위해

---

1) 유의 관계에는 엄밀한 의미에서 동의어와 유의어로 구분된다. 동의어는 의미가 완전히 같아 모든 문맥에서 치환이 가능한 데 비해, 유의어는 개념적 유사성을 전제로 한다. 그러나 본고의 목적이 그 세밀한 구분에 있는 것이 아닌 만큼 이후 동의어는 유의어를 포괄하는 의미로 사용하기로 한다.

2) [8]에서는 형용사 의미 부류에서 고려해야할 주요한 사항은 반의관계의 이용이라는 점을 지적하였다.

3) WordNet에는 형용사에만 similarity 관계가 명시되어 있다. 이것은 단어 사이의 similarity가 아니라 유의어 집합들 사이의 similarity이다.

tool을 개발하여 활용한다.

(6) 보다 객관적인 계층 체계를 수립하기 위해 각 처리 과정에서 먼저 알고리즘을 설정하고 이에 따라 전산적인 처리를 한다.

(7) 위 알고리즘을 수행하기 위하여 적절한 휴리스틱[4]을 도입한다. 그러나 상식적으로 재고의 여지가 있는 부분에서는 무리하게 적용하지 않는다.

(8) 동의어·반의어는 그 개수가 많지 않을 것으로 예상된다([6]). 사전 뜻풀이말에도 설명상의 한계가 있을 것이고, 무엇보다 형용사의 특성이 정확한 의미의 동의어·반의어가 많지 않다는 점을 고려한다. 부족한 동의어·반의어 관계는 위 (7)에서 언급한 휴리스틱을 통해 확장하도록 한다.

(9) 어느 어휘의 의미는 공기하는 요소의 속성에 의해 결정된다는 '분포 가설[17]' 에 따라, 의미 분류별 계층 구조는 이 공기 명사를 기준으로 한다.

(10) 위 (9)의 의미 분류 시 기존의 분류 체계([1])를 재활용한다. 분류 체계를 새롭게 세우는 일은 많은 노력을 필요로 할 뿐만 아니라, 반드시 기존 분류 방식보다 우수하다는 보장을 할 수 없다.

(11) 기타 구체적인 방법은 [11]을 원용한다. 명사 의미 계층을 만든 선행 연구의 방식을 형용사의 의미 계층 추출에 있어 유사하게 적용해 본다. 명사와 형용사는 의미 구조상 차이가 있으나, 원칙에 따라 세밀한 분류체계를 제시한 기존 연구를 우선 적용하는 것이 시행착오를 줄이는 길이 될 것이다.

4. 추출

4.1. 형용사 의미 계층의 표상

어휘는 개념을 중심으로 서로 유기적인 관계를 맺고 존재한다. 여기서 관계는 상·하위 관계와 같은 수직적 구성과 동의·반의어 관계와 같은 수평적 구성으로 나뉜다.

보통의 의미 계층 혹은 의미 부류는 이 가운데 전자의 측면에 주로 관심을 두고 있다. 그러나 [4], [7], [8], [12], [14], [15] 등의 선행 연구를 검토해 본 결과, 의미 구조에 대한 올바른 표상은 수직적·수평적 관계를 모두 아우르는 형태가 바람직하다. 따라서 형용사 의미 계층의 표상은 다음과 같은 자료 구조로 설계한다.

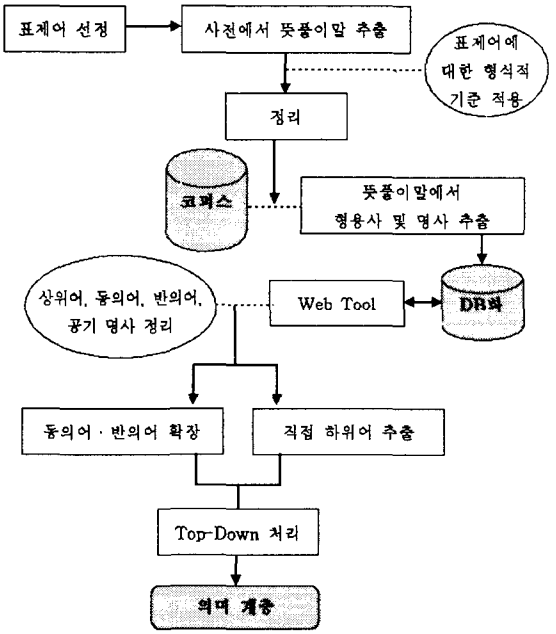
형용사 의미 계층의 자료 구조

A : 형용사 자료 구조의 집합
= {a_1, a_2, … a_n}

a_k : 특정 표제어의 자료 구조

[
표제어(ENTRY)
갈래번호1 … 격틀에 따른 구분
갈래번호2 … 의미에 따른 구분
갈래번호3 … 세부 의미에 따른 구분
하의어(HYPONYM) {a_i, a_j, … }
동의어(SYNONYM) {syn_1, syn_2… }
반의어(ANTONYM) {ant_1, ant_2, … }
공기 명사(NOUN) {$noun_1$, $noun_2$, … }
]

4.2. 전체 수행 절차



[그림 1] 전체 흐름도

4.3. 대상 선정

다루고자 하는 표제어는 2003년도 21세기 세종계획에서 작업한 전자사전에서 품사표지 'va' 가 부착된 1,759개를 대상으로 한다.

이것을 1차 대상으로 한 이유는 모든 형용사를 대상으로 하기에는 실질적인 어려움이 있어 빈도가 높은 집단으로 한정해야할 필요가 있기 때문이다. 2003년도 세종계획 전자사전 결과물은 우선 현대 한국어에서 비교적 출현빈도가 높은 어휘를 대상으로 하고 있어 이 점에서 대상으로 삼기에 타당하다. 또한 세종계획 전자사전에는 품사·통사·의미 등의 정보가 명시되어 있어 이후 다른 연구로의 확장이 용이하다. 뿐만 아니라 격틀 등의 정보나 그에 따른 관련 용례까지 정리하고 있어 후행연구의 가능성이 크다.

뜻풀이말을 추출하고자 하는 사전은 연세 한국어 사전의

4) 본고에서 말하는 휴리스틱이란 '결정과정의 단순화를 위한 한 지침'일체를 말한다. 이는 가장 이상적인 해답을 구하는 것이 아니라 가장 만족할 만한 수준에서 '그러할 것이라고 추측'되는 간편한 규칙을 의미한다. 단, 전산학에서 일반적으로 말하는 알고리즘 기법과는 약간의 차이가 있음을 밝힌다.

용언류로 한다. 연세 한국어 사전에서 용언은 표제어 기준 총 19,231개이며, 동음이의어, 다의어를 제외한 표면형으로 본다면 12,251개이다.

이 가운데 형용사로 품사표지를 지닌 것은 4,377개이며, 마찬가지로 표면형으로는 2,897개이다. 이 정도면 중규모 이상의 사전으로 판단되며, 기초 연구에는 무리가 없을 것으로 보인다.

이때, 세종계획 전자사전의 범위와 연세 한국어 사전의 범위에 차이가 있으므로 형용사 표제어에서 누락되는 것이 발생한다. 다시 말해, 연세 한국어 사전에 형용사로 등재되어 있으나 세종계획 전자 사전에서 추출한 1,759개의 표제어에 들지 않은 것들은 제외된다.

반면, 세종계획 전자사전에는 형용사로 등재되어 있으나 연세 한국어 사전에 동사로 등재되어 있는 표제어가 있다. 이 경우에는 별도의 선별이 필요하다. 그 유형에 따른 선별 방법은 아래와 같다.

(1) 동사가 형용사 표제어와 동음이의어인 경우에는 제외한다. (예; 싸다1:pack, 싸다2:cheap)

(2) 동형 표제어에서 유사한 의미를 동사와 형용사를 모두 지니고 있는 경우에는 이 동사를 대상에 포함시킨다. (예; 굵다, 밝다, 늦다, 크다, 흐리다 등)

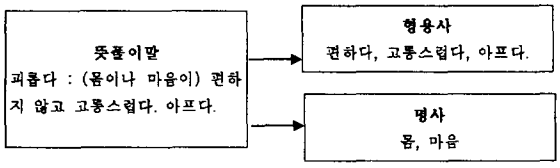
(3) 세종계획 전자사전 표제어에 있으나 해당 의미의 표제어가 연세 한국어 사전에 동사로만 표기되어 있을 경우 대상에 포함시킨다.(예; 여물다, 째지다 등)

(4) '의성/의태어 + 하다' 형은 대상에 포함시킨다. (예; 달랑달랑하다 등)

위와 같은 후처리까지 거쳐 정리를 하면, 총 2,514개의 표제항이 얻어진다. 이를 동음이의어와 다의어를 고려하지 않고 순수 글자로만 헤아렸을 때 총 1,528개의 표면형이 있다. 이후의 표제어 구별은 표제어에 각 갈래 번호를 부착하고 구분을 위해 그 사이에 '-' 를 삽입하는 방식을 취한다.

4.4. 사전 뜻풀이말에 나타난 형용사 및 명사 추출

다음 단계의 작업은 뜻풀이말에서 상위어, 동의어, 반의어, 공기 명사 등을 정리하는 일이겠으나 향후 구분 작업의 편이를 위해 뜻풀이말에서 형용사와 명사만을 우선 추출하는 작업을 먼저 하기로 한다. 뜻풀이말에서 형용사와 명사에 해당하는 어휘만을 따로 정리해두면 이들을 취사선택하여 상위어 등을 손쉽게 구현할 수 있다. 예를 들면 아래의 그림과 같다.



[그림 2] 뜻풀이말에서 '형용사', '명사' 추출

위와 같은 처리를 하는 데 있어서 사용할 수 있는 방법으로 우선 생각할 수 있는 것이 형태소 분석기의 이용이다. 그러나 이를 이용했을 경우 오히려 처리의 절차와 시간이 늘어나는 결과를 초래하게 된다.

그 다음으로 생각할 수 있는 것이 코퍼스의 이용이다. 21세기 세종계획에서 구축한 품사표지 코퍼스를 이용하여 각 어절에서 형용사 또는 명사의 기본형을 추출하는 일이 가능하다.

이를 위해 사용한 자원은 21세기 세종계획 천만 어절 코퍼스를 가지고 1차 예비 구축한 품사부착 코퍼스 DB이다. 이를 이용하여 각 표제항에 따라 뜻풀이말 내의 형용사와 명사 모두 추출할 수 있었다. 이때 다음과 같은 조건을 처리에 적용하여 정교화하였다.

(1) 형용사의 경우 표제항에 포함되지 않는 것은 제외한다.

(2) '수', '것' 등과 같이 지나치게 일반적인 명사는 제외한다.

코퍼스를 이용한 형용사 및 명사 추출

T : 대상이 되는 표제어 및 뜻풀이말 집합, 표제항
T_{adj} : 뜻풀이말에서 추출한 형용사
T_{noun} : 뜻풀이말에서 추출한 명사
D : T의 뜻풀이 말 = $\{d_1, d_2, \cdots d_n\}$, n = 2514
W : D의 어절 = $\{w_1, w_2, \cdots w_m\}$, m = 각 뜻풀이말 어절 수
C_{word} : 코퍼스 어절 테이블
$C_{morpheme}$: 코퍼스 형태소 테이블
C_{tag} : 코퍼스 품사 테이블

A. $C_{word}(w_j)$가 C_{tag}에서 [형용사]면 $C_{morpheme}(w_j)$에 '-다' 를 결합하여 T_{adj}에 추가한다.
B. $C_{word}(w_j)$가 C_{tag}에서 [어근]이면 $C_{morpheme}(w_j)$을 기본형 복원하여 T_{adj}에 추가한다.
C. $C_{word}(w_j)$가 C_{tag}에서 [일반명사]면 $C_{morpheme}(w_j)$을 T_{noun}에 추가한다.

4.5. 상위어, 동의어, 반의어, 공기 명사 기술

형용사의 경우 상위어 등의 구분은 [11]에서 제시한 뜻풀이말의 유형에 따른 기술 방식을 따르기로 한다. 기술에 있어서 [11]은 동음이의어와 다의어를 고려하지 않고 상위어 등을 정리하였을 경우, 헝클어진 계층 구조가 출현할 수 있음을 지적한다. 따라서 의미구분 기술은 새롭게 정리된 구분 표지를 그대로 사용[5]하기로 한다.

상위어의 선택은 뜻풀이말에서 특정 형용사의 의미가 보다 세분화되는 경우 이 형용사를 그 표제어의 상위어로 두는 것을 원칙으로 한다. 아래에서 각 유형에 따른 휴리스틱을 보도록 하자.

[유형1] - Adj다 : Adj를 동의어로 한다.
(예; 궁상맞다 : 꾀죄죄하다. 초라하다. → '꾀죄죄하다', '초라하다' 를 동의어로 한다.)

[유형2] - Adj지 않다 : Adj를 반의어로 한다.

---

5) 예를 들면 아래와 같다.
나쁘다-i-1-a : 마음에 들지 않다. 좋은 느낌이 아니다.
나쁘다-i-1-b : (사물이나 사정이) 좋은 상태가 아니다. 정상적이 아니다.
나쁘다-i-2- : (도덕적으로) 옳지 않다.
나쁘다-ii-0- : (건강에) 해롭다.

(예; 마땅찮다 : 적당하지 않다. → '적당하다' 를 반의어로 한다.)

[유형3] - Noun이 Adj다 : Adj를 상위어로 한다.
(예; 맛있다 : 맛이 좋다. → '좋다' 를 상위어로 한다.)

[유형4] - 정도부사 Adj다 : Adj를 상위어로 한다.
(예; 구차하다 : (살림살이가) 매우 가난하다 → '가난하다' 를 상위어로 한다.)

[유형5] - Adj 느낌이/데가/분위기가 있다 : Adj를 상위어로 한다.
(예; 뻔뻔스럽다 : 보기에 뻔뻔한 데가 있다. → '뻔뻔하다' 를 상위어로 한다.)

[유형6] - Adj 상태에 있다 : Adj를 상위어로 한다.
(예; 바쁘다 : (무엇이) 서둘러 해야 하거나 급한 상태에 있다. → '급하다' 를 상위어로 한다.)

[유형7] - Adj하게 보이다/있다 : Adj를 상위어로 한다.
(예; 허름하다 : (집이) 값이 싸게 보이다. → '싸다' 를 상위어로 한다.)

[유형8] - Adj1고 Adj2다 : Adj1, Adj2를 상위어로 한다.
(예; 가증스럽다 : 보기에 몹시 괘씸하고 밉다. → '괘씸하다' , '밉다' 를 상위어로 한다.)

[유형9] - Adj1면서 Adj2다 : Adj1, Adj2를 상위어로 한다.
(예; 괄괄하다 : (성격이) 호탕하면서 급하다. → '호탕하다' , '급하다' 를 상위어로 한다.)

[유형10] - Adj1(부사형) Adj2다 : Adj1, Adj2를 상위어로 한다.
(예; 굵다2 : (글씨가) 뚜렷하게 크다. → '뚜렷하다' , '크다' 를 상위어로 한다.)

[유형11] - Adj1 정도로 Adj2다 : Adj1, Adj2를 상위어로 한다.
(예; 달다5 : 기분이 좋을 정도로 마음에 흡족하다. → '좋다' , '흡족하다' 를 상위어로 한다.)

[유형12] - Adj1한 Noun이 Adj2다/서술어 : Adj1, Adj2를 상위어로 한다.
(예; 뭉클하다1 : 물렁물렁한 것이 닿는 느낌을 주다. → ' 물렁물렁하다' 를 상위어로 한다.)

[유형13] - Adj1지 못하고/않고 Adj2다 : Adj1를 반의어, Adj2를 상위어로 한다.
(예; 박하다 : (마음 씀씀이나 태도가) 너그럽지 못하고 쌀쌀하다. → '너그럽다' 를 반의어로 하고, '쌀쌀하다' 를 상위어로 한다.)

[유형14] - Adj1지도 Adj2지도 않다 : Adj1, Adj2를 반의어로 한다.
(예; 미지근하다 : 차지도 뜨겁지도 않다. → '차다' , '뜨겁다' 를 반의어로 한다.)

[유형15] - ☞ 'Adj' 의 속된말/비표준어/높임말/줄임말 : Adj을 동의어로 한다.
(예; 가엽다 : ☞가엾다. → '가엾다' 를 동의어로 한다.)

형용사 뜻풀이말에 나타난 바를 토대로 의미 관계를 기술하는 일은 다음과 같다.

(1) 사전 뜻풀이말을 살펴보고, 이것이 위 15개 유형 가운데 어디에 해당하는 지 파악한다.

(2) 추출된 형용사에 해당하는 표제어로 다시 검색하여 그 표제어에서 해당 의미와 관련된 갈래번호를 찾는다.

(3) 의미에 따른 구분이 되었으면 찾아낸 형용사를 그 구분표지까지 포함하여 유형별 휴리스틱에 따라 동의어, 반의어, 상위어로 분류하여 입력한다.

(4) 공기 명사는 위 4.4에서 검출된 것 가운데 오류가 없는지만 확인한다.

4.6. 구성

4.6.1. 동의어·반의어

형용사는 엄밀한 의미의 동의어, 반의어를 지니지 않는다([6]). 아울러 한정된 사전 뜻풀이말을 통해 추출하였기 때문에 추출된 동의어, 반의어의 개수가 상대적으로 부족하다. 따라서 동의어 및 반의어를 확장할 수 있는 방법이 모색되어야 한다. 동의·반의어어의 경우 아래와 같은 추론을 통해 확장한다.

동의·반의어 집합 구성을 위한 휴리스틱
A. 동의어의 동의어는 동의어로 한다.
B. 반의어의 반의어는 동의어로 한다.
C. 표제어㉮와 표제어㉯가 서로를 상위어로 하고 있으면 상위어로 하지 않고 각기 상대방의 동의어로 한다.
D. 동의어와 상위어가 중첩되는 것이 있으면 상위어로 한다. 즉, 동의어 집합에서 제외한다.
E. 동의어의 반의어는 반의어로 한다.
F. 반의어의 동의어는 반의어로 한다.

4.6.2. 직접 하의어

의미 계층을 구성하는 방법은 크게 Bottom-Up 방식과 Top-Down 방식으로 구분할 수 있다. Bottom-Up은 최하의어에서 시작하여 자신의 상위어를 순서대로 찾아나가 연결하는 방식이다. 이에 비해 Top-Down은 최상위어에서 시작하여 자신의 하의어로 차례차례 진행한다. [11]에서는 의미 계층을 Bottom-Up방식으로 추출하였다. 이러한 Bottom-Up 방식은 하나의 표제어가 여러 관점에서 하의어를 지닐 수 있는 형용사의 특성을 고려하면 적합하지 않다. 이에 본고에서는 Top-Down 방식을 이용하여 의미 계층을 추출하고자 한다.

Top-Down을 위해서는 주어진 표제항의 상위어 정보를 역으로 재구성하여 직접 하의어 정보를 도출해야 하며, 이를 바탕으로 의미 계층상에서 최상위어 목록을 뽑아야 한다.

4.6.3. 의미 계층

의미 계층은 최상위어 목록의 각각의 표제어에서 시작해 하단으로 차례로 내려가면서 자신의 하의어를 재귀적으로 도출하는 방식을 택한다. 이 경우 자신의 상위어에

다시 연결되는 순환 형태를 제거한다. 즉, 특정 표제어의 하의어에 자신의 상위어가 다시 출현하게 되면 무한 루프의 구조를 가지게 되므로 더 이상 연결되는 일이 없도록 끊어야 한다. 이는 아래의 ㉮에서 처리한다.

Top-Down 방식의 의미 계층 추출 알고리즘

```
TOP : 최상위어 목록
  = {t1, t1, … tn}
H : 직접 하의어 자료 구조
  = {h(상위어, 하의어 목록)1, h(상위어, 하의어 목록)2, …}
T : 대상이 되는 표제어 및 뜻풀이말 집합, 표제항
  = {t1, t2, … tk}, k = 2514
  Thyper : 상위어
  Tsyn : 동의어
  Tant : 반의어
  Tnoun : 공기 명사
HIERARCHY : 의미 계층
  = {hierarchy1, hierarchy2, …}

GetHyponym(entry, i):
  T(entry)을 hierarchyi에 추가
  if entry가 H에 있으면:
    for hyponym in h(entry, 하의어 목록):
      if hyponym이  hierarchyi에 없으면: … ㉮
        return GetHyponym(hyponym, i)
  else:
    return ∅

for i = 1 to n:
  GetHyponym(ti ,i)
```

4.6.4. 공기 명사를 이용한 의미 유형별 형용사 의미 계층 구성

앞의 4.4에서 뜻풀이말에 나타난 공기 명사군을 추출한 것은 이들이 각 형용사가 지니는 의미적 범주를 명확히 해주는 역할을 하기 때문이다. 즉, 이들을 활용하면 각 형용사마다의 의미적 차이를 알 수 있으며, 같은 형용사 안에서도 다의 관계에 따라 구분을 할 수 있다. 예컨대, '나쁘다-i-1-b'와 그 직접 하의어군의 공기명사를 보면 아래와 같다.

하의어	공기 명사
구태의연하다--0-	상태
극심하다--0-	일, 상태, 정도
끔찍하다--2-	사건, 상황, 진저리, 정도
둔하다--3-	머리, 능력
사납다--4-	사정, 현상
열악하다--0	질, 조건
조악하다--0-	제품, 질
해롭다--0-	영향

[표 1] '나쁘다-i-1-b'의 직접 하의어

즉, 뜻풀이말에서 나타나는 공기 명사를 토대로 형용사의 의미 분류를 하면 각 부류마다 계층 관계를 도출할 수 있을 것으로 기대한다.

형용사 의미 부류는 위에서 3절의 (10)에서 목표한 바와 같이, 기존의 의미 부류를 활용한다. 구체적으로는 [1]에서 나눠진 의미 부류를 쓰도록 한다. 의미 구분의 준거가 되는 공기 명사는 표제항에서 추출한 공기 명사 전체의 빈도수를 내어 총 5개 이상 출현한 것 가운데, [1]에서 구체적으로 언급된 것을 대상으로 한다.

여기서 의미 분류별 계층 구축의 기초가 되는 자원을 무엇으로 할 것인가의 문제가 제기되는데, 우선 위에서 구축된 전체 의미 계층을 사용하는 방법을 들 수 있다. 즉, 계층상에서 어느 표제어가 특정 공기 명사를 지니고 있으면 그 표제어와 같은 위계의 표제어 및 하위 표제어를 모두 같은 의미 부류로 취급하는 것이다. 그러나 이는 전체 의미 계층을 확인한 결과 옳지 않은 것으로 확인[6]되었다. 따라서 의미 분류별 계층을 위해 각 표제항의 공기 명사를 바탕으로 의미 분류별로 표제항을 재편하였다. 이후 각 의미 유형별 표제항 정보에 전체 의미 계층 내에서 각 표제어어의 직접 상위어와 직접 하의어를 결합하여 각각의 의미 유형마다 최상위어 집합을 도출한다. 각 의미 유형별 의미 계층은 이를 바탕으로 이루어지게 된다.

그 결과 각 의미 영역에 맞는 표제어들만 계층 관계에 포함되는 것을 관찰할 수 있었다.

5. 결과

위와 같은 과정을 거쳐 집계된 의미계층에는 상위어 2,638개, 동의어 179개, 반의어 430개, 공기 명사 3,618개가 포함된다.

5.1. 전체 형용사 의미 계층

전체 형용사 의미 계층은 226개의 최상위어에서 시작한다. 각각의 최상위어는 하의어로 계층을 확장해 가면서 하의어의 하의어를 재귀적으로 결합한다. 이렇게 이루어진 의미 계층에는 총 3,792개의 표제어가 망라된다. 하나의 최상위어가 평균 17개가량의 하의어를 지니는 것이다.

가장 계층이 깊은 것은 18단계이며, 가장 많은 표제어를 하의어로 포함한 최상위어는 '크다2--2-'로 총 414개의 하의어를 가진다.

의미 계층을 명시할 때에 각 표제어에 대한 동의어, 반의어, 공기명사를 함께 표시하도록 하였다. 활용 가치를 우선으로 하는 의미 계층이라면 수직적 상하관계 뿐만 아니라 수평적 의미 관계와 결합관계까지도 함께 기술할 수 있어야 하기 때문이다.

아래의 예를 보도록 하자.

6) 예컨대, '진하다2--2-' 경우 공기 명사를 '빛깔, 냄새, 안개, 화장' 으로 지니는 데, 하의어에 속하는 '붉다--0-', '검붉다--0-', '불그스레하다--0-', '빨갛다--0-' 등은 색채 형용사가 분명하지만 '짙다--5-', '걸쭉하다--1-', '탁하다--1-', '흐리다2--1-', '혼탁하다--1-' 등은 색채 형용사라고 보기에 무리가 따른다.

'나쁘다-i-1-a' 의 의미 계층
나쁘다-i-1-a [ANT|좋다-i-1- 좋다-i-2-] [NOUN|마음 느낌]
구태의연하다-0- [ANT|좋다-i-1- 좋다-i-2-] [NOUN|상태]
불길하다-0- [NOUN|일 느낌]
사위스럽다-0- [ANT|좋다-i-1- 좋다-i-2-] [NOUN|느낌 일]
흉하다-1- [NOUN|운]
흉하다-2- [NOUN|모습]
게걸스럽다-0- [NOUN|음식 물건 욕심 꼴]
징그럽다-0- [NOUN|소름 정도]
징글맞다-0- [NOUN|느낌]
추악하다-0- [NOUN|겉모습 마음씨]
추하다-1- [NOUN|외모]
흉측하다-0-

5.2. 의미 부류별 형용사 의미 계층

[1]의 의미 분류를 바탕으로 한 분류별 의미 계층에 관한 사항은 아래의 [표 2], [표 3]과 같다. 이를 분석하여 보면 '인성 형용사' 와 '색채 형용사' 의 경우 의미 계층의 폭이 넓다고 할 수 있다. 이는 다음과 같은 점에서 추론할 수 있다.

(1) 망라된 표제어가 비교적 크다

(2) 평균 하의어 개수가 크다

(3) 깊이가 깊다.

의미 계층의 '폭이 넓다' 는 점은 다의어의 발생이 풍부하고, 의미를 보다 세분화한 표제어가 많다는 의미가 될 것이다.

형용사 의미 구분	개수	비율 (해당 개수 / 표제항 개수)
청각 형용사	52	2.07%
후각 형용사	13	0.52%
미각 형용사	48	1.91%
몸감각 형용사	149	5.93%
정서 형용사	61	2.43%
인성 형용사	343	13.64%
물성·공통성 형용사	57	2.27%
태도 형용사	133	5.29%
색채 형용사	74	2.94%
모양 형용사	48	1.91%
도량 형용사	48	1.91%
자태 형용사	38	1.51%
평가 형용사	168	6.68%
빈도 형용사	1	0.04%
분포 형용사	12	0.48%
관계 형용사	24	0.95%

[표 2] 형용사 의미 분류에 따른 통계 정보

형용사 의미 구분	최상위어 개수	망라된 표제어	하의어 평균 개수	깊이
청각 형용사	34	116	3.41	4
후각 형용사	16	30	1.88	2
미각 형용사	42	126	3.00	4
몸감각 형용사	117	273	2.33	4
정서 형용사	56	203	3.63	4
인성 형용사	215	812	3.78	9
물성 공통성 형용사	52	104	2.00	4
태도 형용사	133	346	2.60	5
색채 형용사	47	161	3.43	7
모양 형용사	54	126	2.33	4
도량 형용사	28	64	2.29	3
자태 형용사	37	86	2.32	3
평가 형용사	142	395	2.78	5
빈도 형용사	1	2	2.00	2
분포 형용사	11	19	1.73	4
관계 형용사	23	42	1.83	2

[표 3] 형용사 의미 분류에 따른 의미 계층 정보

아래에서 구체적인 예를 보도록 하자.

[색채 형용사] 부류에서 '진하다2--2-' 의 위계[7]
진하다2--2- [NOUN| 빛깔 냄새 안개 화장]
붉다-0- [NOUN| 빛깔 피]
검붉다-0- [NOUN| 빛]
불그스레하다-0- [NOUN| 빛깔]
빨갛다-0- [NOUN| 색깔 사과]

5.3. 구축물 평가

[11]에서는 결과에서 최상위 노드가 지나치게 많다는 점을 문제점으로 지적하였다. 이와 유사한 측면은 본 연구에서도 최상위어 226개가 전체 2,514개의 9%에 해당한다. 한편 본 연구에서는 하위 노드가 매우 적은 표제어가 많이 발견된다. 전체 226개의 최상위어 가운데 하의어를 단 1개만 포괄하고 있는 것이 60개로서 이는 전체의 27% 가량에 해당한다.

그러나 위와 같은 문제점에도 불구하고 의미 계층의 일반적 구성은 안정적이라고 할 수 있다. 계층의 실제 결과를 관찰해 보면 상위어에서 하의어로 내려가는 흐름이 한국어의 어휘론적 관점에서 자연스러울 뿐만 아니라, 의미 영역이 서로 겹치는 경우도 크게 발견되지 않는다. 이를 통해 의미 계층이 유의미하게 도출되었음을 알 수 있다.

의미 유형별 의미 계층의 경우 공기명사를 바탕으로 하여 의미 유형별로 표제항을 재구성한 뒤, 각 표제항의

7) 형용사의 경우 일반적으로 다의성이 높은 특성을 지니는 관계로 하나의 표제어가 여러 의미 부류에 속할 수 있다. 예컨대 위에서 '진하다2--2-' 의 경우 '빛깔' 로 색채 형용사의 범주에 들지만, 동시에 '냄새' 로 후각 형용사의 범주에도 포함된다.

의미에 직접 상관이 있는 직접 상하위어만을 결합하여 구축하였는데, 이 경우 각 의미 유형별로 의미 계층이 유의미하게 정리되는 것을 확인할 수 있었다.

구축된 결과는 한국어 어휘 의미론적 시각에서 참조자료로 활용될 가치가 있으며, 추후 보완 및 발전을 통하여 자연언어처리 시스템에도 활용될 가능성을 보인다.

6. 결론 및 향후과제

본고에서는 한국어 형용사의 의미 계층의 필요성을 밝히고 그 표상 및 추출 방법론을 제시하였다. 아울러 의미 계층을 추출하는 데 있어서 연구자의 직관에 의한 한계를 극복하기 위해 형식적이고 객관적인 기준에서 작업을 수행하는 것은 가장 큰 원칙으로 하였다.

향후 과제로서는 전망할 수 있는 것은 첫째, 뜻풀이에 이용할 사전을 확대해야 한다. 본 연구의 특성상 대규모 한국어 사전이 필요하다. [5]에서는 '표준국어대사전', '연세 한국어 사전', '우리말 큰 사전'의 3종을 이용하였다. 완전한 계층 구조를 도출해 내기 위해서는 이 정도 자원은 있어야 한다.

둘째, 형용사에 국한하지 않고 용언 전체로 확대해야 한다. 이에 형용사의 의미 계층을 도출하는 방식을 연구한 본고는 그 출발점이 될 것이다.

셋째, 뜻풀이말을 유형별로 분석하여 의미 관계를 도출하는 데 있어 보다 자동화할 수 있는 방법을 모색해야 한다. 이는 향후 연구에서 가장 중점을 두어야할 부분이다.

참고문헌

[1] 김정남 (2001), 국어 형용사의 의미 구조. 한국어의미학 8. 한국어의미학회.

[2] 남지순 (2003), 한국어 형용사 전자사전 DECOS-ADJ의 어휘부 구축을 위한 몇가지 논의. 어학연구 39-1. 서울대학교 언어교육원.

[3] 문준혁 (1999), 어휘 지식 베이스를 이용한 단어 사이의 의미 관계 결정, 서강대학교 대학원 컴퓨터공학과 석사학위 논문.

[4] 박동호 (2003), 의미부류 체계의 구축과 적용. 어학연구 39-1. 서울대학교 언어교육원.

[5] 박철우·남승호 (2004), 형용사 논항 의미부류 표준화를 위한 기초 연구 - '크다, 작다, 많다, 적다'를 중심으로 -. 언어학 38. 한국언어학회.

[6] 유명희·최석두 (2002), 형용사 시소러스 설계에 관한 연구. 제9회 한국정보관리학회 학술대회 논문집. 한국정보관리학회.

[7] 유현경·강현화 (2001), 한국어 학습사전에 있어서의 유의어에 관한 연구. 제2차 한국어세계화 국제학술대회 발표집. 한국어세계화추진위원회.

[8] 이재윤·김태수 (1999), WordNet과 시소러스: 언어정보의 탐구 1. 연세대학교 출판부.

[9] 이창기·이근배 (1999), WordNet을 이용한 한국어 시소러스 자동 구축. 제 11회 한글 및 한국어 정보처리 학술대회 논문집.

[10] 한국과학기술원 전문용어언어공학연구센터 (2003), KAIST 다국어 어휘의미망.

[11] 조평옥·안미정·옥철영·이수동 (1999), 사전 뜻풀이말에서 구축한 한국어 명사 의미 계층구조. 한국 인지과학회 논문지 10-4. 한국인지과학회.

[12] 차재은·강범모 (2002), 다의 설정의 방법에 대하여. 한국어학 15. 한국어학회.

[13] 최호섭·옥철영·장문수·장명길 (2002), 사전을 기반으로 한 한국어 의미망 구축과 활용. 2002년도 한국정보과학회 봄 학술발표논문집. 한국정보과학회.

[14] 한정한·도원영 (2005), 한국어 동사 의미망 구축을 위한 어휘의미관계 유형. 한국어학 28. 한국어학회.

[15] 홍재성 (1998), 동사·형용사의 사전적 처리. 새국어생활 8-1. 국립국어원.

[16] Fellbaum, C (1998), WordNet: An Electronic Lexical Database--Language, Speech, and Communication. KMIT PRESS.

[17] Harris, Z (1964), Distributional Structure: The Structure of Language. Prentice-Hall.