의미속성에 기반한 한국어 명사 의미 TAG에 관한 연구

이수광, 조평옥, 안미정, 옥철영
울산대학교 전자계산학과
울산광역시 남구 무거동 산29, 680-749
okcy@uou.ulsan.ac.kr

박재득, 박동인
한국전자통신연구원 컴퓨터 소프트웨어기술연구소
대전광역시 유성우체국 사서함 106, 305-600
{dipark,jdpark}@etri.re.kr

A Study on A Korean Noun Semantic TAG based on Semantic Features

S. Lee, P. Cho, M. Ahn, C. Ock
Dept. of Computer Science, UOU

J. Park, D. Park
Computer Software Technology Institute, ETRI

요 약[1)]

의미 TAG는 한국어 기초어휘에 대한 개념지식을 구축하는 데 기본이 될 뿐만 아니라, 문장 분석시의 구조적 모호성과 단어 의미 모호성을 해소하는 중요한 단서를 제공할 수 있다.

이러한 의미 TAG가 실용적으로 여러 응용 시스템에서 사용되기 위해서는 광범위하고 타당한 자료를 바탕으로 하여 객관적인 방법으로 설정되어야 한다. 국어사전의 뜻풀이말에서의 상위개념을 표제어의 상위어로 선정하는 bottom-up 방식으로 구축하였던 한국어 명사의미체계는 근본적으로 사전편찬자의 비일관적인 뜻풀이말의 기술에 따른 여러 문제점이 있었다.

본 연구에서는 이러한 문제점들을 해결하기 위해서 사전 뜻풀이말에서 상위개념을 수식하는 어절과 용언의 의미호응관계에서 상위개념의 의미속성을 추출하고, 이들 의미속성에 의한 명사의미체계를 구축하여 이를 바탕으로 명사의미 TAG를 설정할 수 있도록 하였다.

1. 서론

자연어 처리의 각 단계에서는 자연언어가 지닌 중의성으로 인하여 여러 유형의 모호성(ambiguity)이 발생한다. 결국 자연언어 처리의 많은 연구들은 각 처리 단계에서 나타나는 중의성을 어떻게 효율적으로 해소할 것인지에 대한 연구로 집약된다고 볼 수 있다.

자연언어 처리 과정에서의 중의성을 해소하기 위한 방법론으로 크게 규칙기반 접근방법(Rule-Based Approach)과 통계기반 접근방법(Statistical Approach)으로 나누어 연구되었다.[13,14] 규칙기반 접근방법은 자연언어에 적용되는 공통의 원리나 규칙을 찾아내어 이를 이용하여 자연언어의 중의성 문제를 결정적(deterministic)으로 해결하는 방법으로 지식기반 방법(Knowledge Based Approach) 또는 제약기반 접근방법(Constraint-Based Approach)이다. 통계기반 접근방법은 자연언어가 실세계에서 사용되는 용례들과 부속 정보를 포함하는 다량의 코퍼스를 분석하여 자연언어에 대한 통계정보를 추출하여 중의성 문제를 확률적으로 해결하는 방법이다.

최근에 한국어에 대한 다량의 코퍼스[27]가 구축되어짐에 따라 이를 이용한 연구들에서 좋은 성과가 나타나고 있으며, 두 접근 방법의 문제점을 상호보완하는 입장에서 두가지 접근 방법을 통합함으로써 광범위한 데이터 처리가 가능하고 높은 정확도를 갖는 태깅 시스템 개발을 위한 연구가 활발히 진행되고 있다.[13,28]

현재까지의 한국어 처리에 대한 연구들은 주로 형태소분석과 구문분석 단계에서의 중의성 해소에 촛점이 맞추어져 연구되어 왔으며, 의미해석에도 이러한 방법론들을 적용하여 의미해석시의 중의성을 해소하려는 연구가 시도되고 있다.[6-9,12,18,19] 그러나 의미해석시에 이러한 방법론을 적용하기 위해서는 의미 TAG가 부착된 코퍼스가 필요하며, 근본적으로 의미 TAG에 대한 연구가 선행되어야 한다. 여기서 의미 TAG는 하나의 어휘가 나타내고자 하는 의미와 관련된 정보로서 의미 중의성을 지닌 다의어(Polysemy)나 동형이의어(Synonym)의 의미해석시에 해당 어휘의 의미를 구분하는 역할을 한다. 이러한 의미 TAG는 한국어 기초어휘에 대한 개념지식을 구축하는 데 기본이 될 뿐만 아니라, 문장 분석시의 구조적 모호성과 단어 의미 모호성을 해소하는 중요한 단서를 제공할 수 있기 때문에 기계번역, 정보검색, 문장 교정 등의 여러 응용에서 정확한 결과를 얻기 위한 매우 중요한 역할을 한다.[22,23]

2. 의미 TAG에 관한 연구현황

의미 TAG는 응용 분야나 의미해석이 요구하는 정밀도에 따라서 그의 복잡도나 의미 TAG

1) 본 연구는 한국전자통신연구원의 위탁과제 "우리말 개념망 명사 데이터 구축"로 수행된 결과임.

SET의 분류 체계가 달라질 수 있다. 의미 TAG와 관련한 국내외 기술개발 현황은 다음과 같다.

(1) WordNet[1,2,5]

WordNet은 미국의 Princeton 대학에서 수년간 개발했던 시스템으로, 기존의 전통적인 언어정보 구축과 현대 컴퓨터 처리기술을 잘 융합한 것으로 효과적인 언어처리용 사전을 구축하는 것을 목적으로 개발되었다. WordNet은 일종의 의미네트워크로 언어심리학적 원리에 기반을 둔 online lexical 데이터베이스로, 영어의 명사, 동사, 형용사, 부사에 대해서 의미적 연관관계로 연결된 유의어 집합으로 나누고 각각의 개념들로 표현하였다. WordNet은 약 120,000 어휘를 90,000 의미 TAG로 분류하고 단어의 의미, 동의어, 반의어, 상의어/하의어, 부분/전체 관계 등을 규명하였다.

(2) CYC의 지식베이스[3]

CYC는 인간 생활에 전반적으로 통용되는 일반 상식을 대규모의 지식베이스로 구축하고, ontology를 활용한 추론을 기반으로 정보검색과 의미해석분야에서 활용하고 있다. CYC 지식베이스는 약 100,000 개의 common sense concept definition과 1,000,000 개의 common sense concept assertion으로 분류하였다.

(3) EDR의 기계번역시스템을 위한 언어 자료베이스[4]

EDR은 일영/영일 기계번역을 위해서 일본 정부기관과 8개의 컴퓨터 관련 업체가 참여하여 개발한 시스템으로, 컴퓨터에게 인간의 언어 판단 능력을 부여하기 위한 대규모 언어 자료베이스가 구축되었다. EDR 시스템은 단어와 개념사이의 관계 및 단어의 문법적 특성과 주어진 의미를 표시하여 컴퓨터에게 형태소 및 구문처리를 할 수 있게 하는 단어사전(Word Dictionary), 모든 개념을 분류하여 상하위 관계를 표현하고 개념사이의 유사도를 계산하는 개념분류사전(Concept Classification Dictionary)과 개념사이의 의미적 공기관계를 결정할 수 있게 하는 개념기술사전(Concept Description Dictionay) 등으로 구성되어 있다. EDR에는 약 400,000 개의 개념과 400,000 개의 개념기술항을 분류하였다.

(4) 국내 기술개발현황

한국어의 어휘 의미에 관한 연구로는 국어학에서 한국어의 의미를 쓰임새나 형태에 따라서 명사, 기능어(조사), 동사를 분류하는 형태로 진행되어 왔다.[25,26] 그러나, 국어학에서 연구되어진 대부분의 결과들은 대량의 자료를 대상으로 하지 않았기 때문에 한국어 처리분야에서 직접적으로 쓰이기에는 자의적이고 부적합한 면이 많다.

한편 전산학적인 입장에서의 한국어의 어휘 의미에 대한 연구는 형태소해석이나 구문구조 분석시의 중의성 해소와 같은 특정 목적을 위한 것으로, 응용 분야에 제한적이며 연구 개발된 체계가 다른 시스템에 적용(portable)하기에는 어려움이 있고 의미해석이나 담화분석용으로 이용하기에는 많은 문제점이 있었다.[12,18,19]

최근에는 미국에서 개발된 WordNet을 한국어의 어휘에 적용하려는 시도[16,17]와 국어사전의 뜻풀이말을 이용하여 명사들의 상하의어 관계를 규명[11,15]함으로써 우리말의 의미계층체계를 구축하려는 연구가 시도되고 있으며, 의미격을 분류하거나 신경회로망을 이용한 의미 중의성 해소 방법들이 연구되고 있다.[11,12,20,21] 그러나 이러한 연구들은 연구실 수준의 실험적인 연구에 제한되어 광범위한 한국어 의미해석이나 여러 유형의 중의성 해소를 위한 체계적인 의미 TAG를 구축하는 데는 미흡한 면이 많다.

3. 의미속성에 기반한 명사의미체계

의미 TAG가 실용적이면서도 여러 응용 시스템에서 사용되기 위해서는 광범위하고 타당한 자료를 바탕으로 하여 객관적인 방법으로 설정되어야 할 것이다. 국어사전의 뜻풀이말에서의 상위개념을 표제어의 상위어로 선정하는 bottom-up 방식으로 구축하였던 한국어 명사의미체계는 연구자의 주관이 개입되지 않으므로써 비교적 객관적인 의미체계를 구축할 수 있는 장점이 있으나 다음과 같은 문제점이 있었다.[15]

첫째로, 사전편찬자의 비일관적인 뜻풀이말의 기술과 사전마다 다른 뜻풀이말로 인해 구축된 한국어 명사의미체계가 일관성을 지니지 못하였다.

둘째로, 단순히 사전의 뜻풀이말만을 이용하여 구축하였기 때문에 서로 다른 계층으로 분류되어야 할 명사들이 같은 계층으로 분류되어 있거나, 같은 계층으로 분류되어야 할 계층의 명사들이 다른 계층으로 분류되는 경우가 있었다.

셋째로, top-down 방식으로 구축된 명사의미체계와 비교할 때 의미체계의 depth와 breadth가 많은 차이가 있다. 특히 최상위계층은 전혀 다른 형태로 분류되었다.

따라서 본 연구에서는 이러한 문제점들을 해결하기 위해서 사전 뜻풀이말에서 상위개념(중심어)을 수식하는 어절들에서 상위개념과 용언의 의미호응관계에서 의미속성을 추출하여 이들 의미속

성에 의한 명사의미체계를 구축하고, 명사의미체계의 응용분야에서 요구하는 정밀도의 정도에 따라서 명사의미 TAG를 설정하도록 하였다.

3.1 상위개념의 의미속성(semantic feature)

Bottom_up 방식으로 구축한 기본적인 의미체계에서 "현상"을 상위어로 가지는 어휘(표제어)는 다음 <표 2>의 모든 어휘들이다. 이러한 의미체계에서 단순히 사전편찬자의 주관이나 오류 문제를 고려하지 않더라도, "현상"의 하위의미 관계의 어휘들간의 어떠한 연관성, Clustering, 혹은 하위개념 분류를 위한 정밀한 근거를 제시하기 어렵다. 그러나, 뜻풀이말에서 **중심어를 수식하는 어절까지를 표제어의 상위개념을 설정하는 기준**으로 삼으면 보다 정밀하고 객관적인 하위개념을 분류할 수 있다. 예를 들면, "현상"을 상위개념으로 가지는 표제어들의 뜻풀이말에서 "현상"을 수식하는 어절들을 세분하면 다음 <표 1>과 같다.

표 1. 수식어구에 따른 하위개념 분류

현상	되는	:	49	현상 나타내는	: 8
현상	**변하는**	:	27	현상 **줄어드는**	: 8
현상	**보이는**	:	25	현상 발음되는	: 6
현상	**일어나는**	:	21	현상 **나가는**	: 6
현상	**일으키는**	:	17	현상 **말하는**	: 6
현상	**바뀌는**	:	16	현상 못하는	: 6
현상	**생기는**	:	14	현상 자연	: 6
현상	**가는**	:	14	현상 방출하는	: 6
현상	있는	:	13	현상 나는	: 6
현상	**나타나는**	:	13	현상 늙어나는	: 6
현상	**변화하는**	:	12	현상 따는	: 5
현상	**이동하는**	:	10	현상 증대하는	: 5
현상		:	10	현상 움직이는	: 5
현상	**오르는**	:	10	현상 이루는	: 5
현상	**달라지는**	:	9	현상 탈락하는	: 4
현상	**흐르는**	:	8	현상 아니하는	: 4
현상	하는	:	8	현상 방사하는	: 4

위 <표 1>는 사전 뜻풀이말에서 "현상"을 중심으로 데이터를 추출하되 "현상"의 바로 앞 어절의 출현 빈도를 내림차순으로 정렬하였다(숫자는 사전의 뜻풀이말에서 발견된 횟수).

<표 1>에서 밑줄친 어절들은 "현상"을 보다 구체적으로 설명하는 것으로, "현상" 개념의 하위개념의 분류 기준으로 설정될 수 있는 **의미속성(semantic feature)**으로 볼 수 있다. 따라서 상위 개념 "현상"에 대해 각 의미속성을 지닌 하위어(표제어)를 추출하면 다음 <표 2>와 같다.

3.2 의미속성에 의한 명사의 Clustering

상위개념에 대한 하위개념의 의미를 보다 구체적으로 설명하는 의미속성을 위와 같이 설정하고, 다음 <표 3>와 같이 의미속성 다음에 출현하는 어절들을 살펴 보자.

표 2. 의미속성에 따른 하위개념 분류

되는 현상 (49)	광역 광영 구변수차 구변천문학 델린저현상 델타 매설 백화현상 백회 병 복빙 복사 빙화 빙환 사주 심적포화 심적현상 암순응 암숙 연상 연상라 연차 연차 월식 월액 융합 융합반응 융해 융해열 ...
변하는 현상 (27)	고조 고조 노화 노환 단풍 대비현상 대빈 레이놀즈 현상 레이더 매월상동 매음 맥놀이 맥동 모음변화 모음보표 변재 변천 산패 산패유 승화 액화 액화석유가스 음편 음표 증발 핵반응 핵변
보이는 현상 (25)	가현운동 가형 보색잔상 보생불 복시 복시 비안개 비압축성 신기루 아지랑이 아침노을 연주운동 연주자 운동잔상 운동장 일주운동 원충 적조 적조 ....
일어나는 현상 (21)	범화 범휘 벼락 상기 상기 색청 색체 자기유도 전자기유도 전자기장 접촉전기 접촉전염 정전기유도 정전렌즈 지상 지상 피에조전기 피에타 회전편광 회전포물면
일으키는 현상 (17)	경련 경련 광란성 광탐 기폭 부식 부식강상 생물발전 생물시대 아나필락시 아낙 약전 약전 집단히스테리 집달 첨단방전 첨단비대증
바뀌는 현상 (16)	구개음화 구거 기화 뒤소리되기 뒤시옷 르불규칙활용 르불탕법 변태 성전환 성절 알파 붕괴 알파선 움라우트 움막 핵변환 핵병기
생기는 현상 (14)	결정분화작용 결정수 다배현상 다번 분극 분극전류 열확산 열활물 침강반응 침강소 플레어 플레어 필링 필마

표 3. 동일한 의미속성을 지닌 개념

변하는	**현상** :27	보이는	**현상** :31	생기는	병 :114	바꾸는	**일** :46
변하는	**일** :21	보이는	것 :24	생기는	염증 : 63	바꾸는	장치 :30
변하는	**모양** : 5	보이는	**일** :16	생기는	**일** : 19	바꾸는	기재 : 3
변하는	것 : 3	보이는	별 :12	생기는	암 : 16	바꾸는	기술 : 2
변하는	활용 : 3	보이는	달 :12	생기는	**현상** : 14	바꾸는	**모양** : 2
변하는	뒤집 : 2	보이는	**모양** :11	생기는	급성 : 13	바꾸는	변환 : 1
변하는	운동 : 2	보이는	병 : 8	생기는	선기 : 10	바뀌는	**현상** :16
변하는	전파 : 2	보이는	사람 : 8	생기는	질환 : 10	바뀌는	**일** :15
변하는	작용 : 2	보이는	상태 : 8	생기는	열 : 9	바뀌는	작용 : 4
변하는	비율 : 2	보이는	수 : 6	생기는	부스럼 : 9	바뀌는	바람 : 4
변하는	별 : 2	보이는	부분 : 6	생기는	종기 : 8	바뀌는	확용 : 3
변하는	설정 : 2	보이는	곳 : 6	생기는	반성 : 8	바뀌는	점 : 2
변하는	정도 : 2	보이는	시험 : 5	생기는	암종 : 8	바뀌는	밸브 : 2
변하는	능력 : 2	보이는	생김새 : 4	생기는	발진 : 6	바뀌는	이치 : 2

위의 <표 3>에서 보는 것과 같이, 많은 상위개념들이 동일한 의미속성을 공유하는 것을 알 수 있다. 이는 다른 개념들간에도 동일한 의미속성에 의해 clustering될 수 있으며, 나아가서는 의미속성의 공유 정도에 따라서 개념들간의 상/하위 개념을 설정할 수 있는 가능성을 보여준다.

따라서 사전에서 변별가능한 의미속성을 추출하여 이를 바탕으로 의미체계를 구축한다면, 용언의 의미호응 관계를 고려한 명사의미체계를 설정할 수 있으며, 추후 새로운 명사가 추가되더라도 전체 의미체계는 큰 변화가 일어나지 않을 것이다.

3.3 용언의 의미호응관계에 의한 의미체계의 세분화

하나의 어휘가 여러 의미를 지닌 다의어이거나 동형이의어일 때, 해당 어휘의 의미는 함께 사용된 다른 어휘에 의해서 해당 어휘가 표현하고자 하는 의미가 명확하게 정의된다. 이러한 현상을 품사간의 의미 호응/제약이라 한다.

의미속성에 의해 구축된 명사의미체계를 이용하여 대상 어휘에 대해서 1차적으로 해당 개념으로 사상(mapping)시킨 다음, 용언간의 의미호응

제약 조건을 적용하여 명사나 용언의 의미체계를 점진적으로 확장시킨다.

예를 들어, '칼', '단도', '창', '총', '활', '미사일', '폭탄' 등의 명사들은 무기류의 의미범주/개념에 속하나, '칼', '창', '단도' 등의 명사는 동사 '**찌르다**', '**갈다**'류의 동사 의미와 호응하며, '총', '활', '미사일', '폭탄' 등의 명사는 '**갈기다**', '**쏘다**'의 동사 의미와 호응한다. 또한 '미사일', '폭탄' 등의 명사는 '**터지다**' 동사의 의미와 호응한다. 따라서, 무기류 명사는 그와 함께 사용되는 동사에 따라서 더욱 세분될 수 있다.

이와 같은 방식은 명사와 용언이 사용되어진 용례에 기초하기 때문에 연구자의 주관이나 사전 편찬자의 주관에 크게 영향을 받지 않아 실용적인 개념체계를 구축할 수 있다. 또한, 용언의 의미 격틀 구축작업을 병행할 수 있으며, 결과적으로 용언의 의미 격틀에 따른 의미 공기조건을 동시에 구축할 수 있다.

4. 명사의미체계 구축

4.1 기초 개념어휘 선정

명사의 의미체계구축을 위한 기초 개념어휘는 단순히 초등학교 교과서나 CORPUS에서 빈도수에 의해 추출한 기초 어휘와는 구별되어야 한다.[29] 이는 사전의 뜻풀이말에서 명사의 개념을 설명하는 어휘와 일반적으로 사용되는 어휘의 빈도와는 다를 수 있기 때문이다.

따라서, 본 연구에서는 금성출판사 MRD 사전에서 순수명사(표제어 100,782개)를 대상으로 다음과 같은 과정을 거쳐 기초 개념을 추출하였다.

첫째로, 표제어 100,782개의 뜻풀이말에서 사용된 중심어(표제어의 상위개념)를 추출하였다. 추출된 빈도수별 표제어의 상위개념은 <표 4>와 같다.

표 4. 뜻풀이말에서 사용된 상위개념의 빈도수

말	(3650)	준말	(3183)	사람	(2943)	하나	(2876)
일	(2797)	총칭	(1150)	곳	(943)	것	(808)
부분	(731)	물건	(725)	가지	(624)	병	(476)
이름	(433)	집	(431)	방법	(394)	성질	(392)
장치	(385)	기구	(371)	뜻	(369)	마음	(363)
상태	(363)	기관	(355)	현상	(322)	돈	(315)
학문	(301)	자리	(295)	책	(287)	비유	(286)
힘	(279)	길	(272)	글	(264)	땅	(264)
모양	(257)	나무	(250)	날	(246)	물질	(245)
제도	(242)	뜻	(241)	교목	(239)	곤충	(225)
새	(224)	방	(221)	없음	(220)	나라	(219)
벼슬	(219)	음식	(217)	그릇	(212)	여자	(212)
소리	(210)	옷	(210)	귀리	(206)	종이	(205)
기계	(202)	배	(200)	주의	(188)	숲	(181)
동물	(178)	수	(177)	줄	(174)	돌	(173)
기운	(171)	식물	(168)	때	(165)	집	(164)
수도	(163)	운동	(163)	방식	(161)	눈	(160)
중세	(159)	정도	(158)	관목	(155)	....	

<표 4>에서 뜻풀이말에서 "일"을 상위개념으로 가지는 표제어가 2,797개이다. 이렇게 추출된 개별 상위개념은 총 18,398개 였으며, 빈도별로 상위개념은 다음과 같다.

- 50회 이상 : 332(누적 빈도수)
- 10회 이상 : 1,589(누적 빈도수)
- 5회 이상 : 3,004(누적 빈도수)
- 3회 이상 : 4,970(누적 빈도수)

본 연구에서는, 추출된 개별 상위개념 18,398개 중에서 1~2회 사용된 상위개념은 표제어와 주로 유사어나 관련어 관계에 있어, 3회 이상 사용된 상위개념 어휘를 의미계층구조 구축의 기초 어휘로 선정하였다.

4.2 기초 상위개념 어휘 중 동형이의어 구분

동형이의어의 상위개념 어휘는 다른 의미체계에 속하는 하위어가 동일한 상위개념으로 연결되기 때문에, 순수명사의 뜻풀이말에서 사용된 상위개념의 빈도에 의해 선정된 어휘들 중에서 동형이의어를 구분하여야 할 필요가 있다. 따라서, 본 연구에서는 기초 어휘로 선정된 상위개념 어휘 중에서 동형이의어의를 조사(표 5 참조)하였고, 이러한 동형이의어의 상위개념을 사전의 뜻풀이말을 조사하여 뜻풀이말에서 다른 의미로 사용된 동형이의어의 상위개념들을 구분하였고 표제어에서도 구분하여 tagging 하였다.

표 5. 상위개념 어휘 중 동형이의어 목록

말	3650 (7)	:짐승 :총칭 :패 :그릇 :기호 :나무 :준말
일	2797 (2)	:것 :하루
가지	624 (5)	:줄기 :풀 :말 :있음 :느낌
병	476 (3)	:셋째 :현상 :그릇
집	431 (2)	:건물 :말
장치	385 (2)	:빛 :탈항
기구	371 (6)	:글귀 :구 :주머니 :도구 :기구 :구조
상태	363 (2)	:모양 :형편
기관	355 (7)	:관직 :파이프 :장치 :풍경 :관 :부분 :장치
현상	322 (4)	:상태 :사실 :생각 :준말

이러한 과정에서 실제로 "말"은 사전에서 7개의 동형이의어로 구분되나, 3,650개의 표제어 뜻풀이말에서는 실제 다음 4가지의 용법으로만 사용되었다.

- 말1(語) : 3,510
- 말2(馬) : 130
- 말3(단위) : 3
- 말4(장기) : 7

이렇게 동형이의어의 상위개념 어휘를 tag로 구분하여 명사의미계층을 구축하면, 다음 그림과 같이 다른 개념의 동형이의어 상위개념들은 다른 계층으로 분류될 것이다.

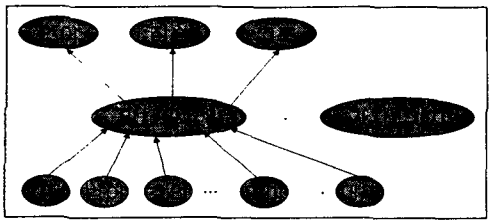

그림 1. 상위개념 어휘 중 동형이의어 구분

4.3 뜻풀이말의 상위개념으로 구축한 명사의미체계

상기의 과정으로 구축된 명사의 의미계층구조는 아래 <표 6>과 같다.

<표 6>에서 중간노드의 "물체a(27)"에서 27은 해당 노드의 하위개념의 수를 나타내며, 마지막 숫자는 단말노드에서 최상위노드까지의 depth를 나타낸다. 이렇게 사전 뜻풀이말로 구축된 명사의미계층은 사물(6), 공간(7), 관계(4), 모양(62), 것(17), 기호(15), 동안(71), 때(67), 방법(41), 양(20), 힘(72), 정신(8), 조건(5) 등의 13개의 최상위 노드와 최대 11개의 depth로 구성되었다. 그러나 뜻풀이말은 표제어에 대한 단순한 사전적인 정의로, 백과사전적인 의미정보가 필요한 개념망을 위해서는 그대로 사용하기에는 많은 문제점이 있다.

4.4 의미속성에 의한 명사의미계층구조 규칙

<표 6>에서 사람(684), 물건(263), 상태(108), 기구(103), 말(92), 음식(51) 등의 중간노드는 너무 많은 하위 중간노드를 가진다. 특히 물건(263)의 중간노드는 "간판"과 같이 직접 단말노드의 상위노드로 사용되거나, "가마니"와 같이 주머니(13)의 하위중간노드를 가지는 계층구조상에서 불일치한 모습을 보이고 있다. 또한 동안(71), 때(67), 방법(41), 양(20), 힘(72) 등의 최상위 루트노드가 바로 다음 하위중간노드를 너무 많이 가지는 기형적인 모습을 보이고 있다.

명사의미체계가 실용적이면서도 여러 응용 시스템에서 사용되기 위해서는 광범위하고 타당한 자료를 기초로 보다 체계적인 접근방법에 의해서 구축되어야 한다.

본 연구에서 구축하고자 하는 명사의미체계는 한국어 처리과정시의 어휘의 의미 모호성으로 인한 구문구조 해석의 중의성과 의미해석을 위한 기본적인 의미정보를 제공하는 것을 일차적인 목적으로 한다. 따라서, 본 연구에서는 한국어 처리의 여러 응용으로의 뛰어난 적응성(portability)과 견고성(robustness)을 제공할 수 있도록 명사의미계층구조에 대한 기본적인 원칙을 다음 <표 7>과 같이 설정하였다.

표 7. 명사의미계층구조 설정에 대한 기본 원칙

원칙 1. 단말노드에서 동일개념의 중간개념노드까지의 depth 차이가 2를 넘지 않도록 한다.
원칙 2. depth I의 중간노드의 하위중간노드 I+1의 개수는 최소한 2개 이상으로 유지한다.
원칙 3. 모든 단말노드에서 최상위 루트노드까지 depth 차이가 3 이하가 되도록 한다.
원칙 4. 단말노드에서 최상위루트노드까지의 depth를 최소한 5가 되도록 한다.

원칙 1에서 단말노드에서 동일개념의 중간개념노드까지의 depth 차이가 2를 넘지 않도록 함으로써, 동일 중간개념을 지닌 어휘들간의 의미유사도 범위를 일정 수준으로 유지할 수 있다.

원칙 2에서 depth I의 중간노드의 하위중간노드 I+1의 개수는 최소한 2개 이상으로 유지하도록 하는 것은, 하나의 중간개념노드가 다음 하위중간개념노드를 가지기 위해서는 최소한 2개 이상의 차별화된 의미속성을 가질 때만 분화시킨다. 차별화된 의미속성은 3.1절과 3.2절에서 제안한 뜻풀이말에서 중심어를 수식하는 어절를 이용하여 하위명사들을 clustering하고, 3.3절에서의 용언의 의미호응관계에 의해 하위 clustering된 명사들을 세분한다. 본 연구에서는 용언과의 의미호응관계를 추출하기 위해서 약 2,100개의 타동사의 의미와 용례를 조사하였다.

원칙 3에서 모든 단말노드에서 최상위 루트노드까지 depth 차이가 3 이하가 되도록 함으로써, 전체 명사의미체계의 균형이 유지되도록 한다.

원칙 4에서 단말노드에서 루트노드까지의

표 6. 뜻풀이말의 상위개념으로 구축한 명사의미체계

```
가곡a - 노래a(22) - 말c(92) - 소리a(58) - 청각a(1) - 감각a(13) - 느낌a(37) - 감정a(12) - 정신a(8) : 9
가마d - 탈것a(7) - 도구a(29) - 기구a(103) - 물건a(263) - 무생물체a(1) - 물체a(27) - 사물b(6) : 8
가마니a - 자루a(4) - 주머니a(13) - 물건a(263) - 무생물체a(1) - 물체a(27) - 사물b(6) : 7
가물a - 날씨a(3) - 상태a(108) - 형편a(23) - 모양a(62) : 5
가정부a - 여자a(48) - 사람a(684) - 생물체a(10) - 물체a(27) - 사물b(6) : 6
가족a(3) - 집a(8) - 켜a(4) - 부분a(51) - 것a(17) : 5
가지a(4) - 줄기a(15) - 식물부분a(3) - 부분a(51) - 것a(17) : 5
간판a - 물건a(263) - 무생물체a(1) - 물체a(27) - 사물b(6) : 5
간호사a - 사람a(684) - 생물체a(10) - 물체a(27) - 사물b(6) : 5
갈대a - 여러해살이풀a(51) - 풀b(5) - 식물a(22) - 생물체a(10) - 물체a(27) - 사물b(6) : 7
감a(2) - 음식a(51) - 물건a(263) - 무생물체a(1) - 물체a(27) - 사물b(6) : 6
.....
```

depth를 최소 5가 되게 함으로써, 의미체계가 여러 응용시스템에 적용할 때 최소한 4단계의 정밀도를 지닌 명사의미체계를 구성하도록 하였다.

4.5 의미속성에 기반한 명사의미체계와 의미 TAG

의미속성에 의한 수정된 최종 명사의미체계는 **구상물, 현상계, 추상물**의 3개의 최상위 노드를 가지며 최대 7개의 depth로 구성되었다. 이러한 명사의미계층구조의 상위 3단계까지의 명사의미계층구조는 <표 8>과 같다.[30]

표 8. 상위 3단계까지의 명사의미체계

1. 구상물:생물,무생물;
   1.1. 생물:사람,동물,식물;
   1.2. 무생물:자연물,인공물,음식물,구조물;
2. 현상계:감각적현상,심리적현상,생리적현상,자연적현상;
   2.1. 감각적현상:피부,눈,귀,냄새,맛;
   2.2. 심리적현상:생각,마음현상;
   2.3. 생리적현상:생리현상,병리적현상;
   2.4. 자연적현상:자연현상,온도,시간;
3. 추상물:시간,공간,일,힘,학문,돈,언어,기준,상태,태도,의견,행동,추상물(기타);
   3.1. 시간3:시기,시간단위,일,순간,시기;
   3.2. 공간:위치,장소1,장소2,방향(공간),경계,공간(기타);
   3.3. 일:계획,행사,직무,사건,일(기타);
   3.4. 힘:조직,작용;
   3.5. 학문:학문1,지식,문화;
   3.6. 돈:돈1,자금,재산,돈단위;
   3.7. 언어:글언어,말언어,정보;
   3.8. 기준:기준(기준),측량단위,규칙;
   3.9. 상태:일(상태),형상(상태),사람(상태),상태(통칭);
   3.10. 태도:대사람태도,대자신태도,예의;
   3.11. 의견:의견1,결정,이론,권리;
   3.12. 행동:양상,행위,대사람행동,대물건행동,일(행동);

최종 구축된 명사의미체계는 전체적으로 1,431개의 중간개념노드로 구성되어 있으며, 이를 계층별로 보면 <표 9>과 같다.

또한, 구축된 명사의미체계에서 단말노드에서 그의 상위개념 연결 리스트 형태는 다음 <표 10>과 같다.(표 6과 비교)

이렇게 구축된 명사의미체계는 그의 응용분야에서 요구하는 정밀도의 정도에 따라서 명사의미체계의 상위 단계에서부터 필요한 명사의미 TAG가 결정될 수 있을 것이다.

표 9. 명사의미체계의 계층별 개념수

level / root	1 level	2 level	3 level	4 level	5 level	6 level	계
구상물	1	2	7	60	213	156	439
현상계	1	4	12	31	88	38	174
추상물	1	13	56	190	433	125	818
level 합계	3	19	75	281	734	319	1,431

5. 결론

명사의미 TAG는 한국어 기초어휘에 대한 개념지식을 구축하는 데 기본이 될 뿐만 아니라, 문장 분석시의 구조적 모호성과 단어 의미 모호성을 해소하는 중요한 단서를 제공할 수 있기 때문에 자연언어 처리의 여러 응용에서 정확한 결과를 얻기 위한 매우 중요한 역할을 한다.

명사의미 TAG가 실용적이면서도 여러 응용시스템에서 사용되기 위해서는 광범위하고 타당한 자료를 바탕으로 하여 객관적인 방법으로 설정되어야 한다. 국어사전의 뜻풀이말에서의 상위개념을 표제어의 상위어로 선정하는 bottom-up 방식으로 구축하였던 한국어 명사의미체계는 연구자의 주관이 개입되지 않으므로써 비교적 객관적인 의미체계를 구축할 수 있는 장점이 있으나, 사전편찬시의 비일관적인 뜻풀이말의 기술에 따른 여러 문제점이 있었다.

본 연구에서는 이러한 문제점들을 해결하기 위해서 사전 뜻풀이말에서 상위개념을 수식하는 어절과 용언의 의미호응관계에서 상위개념의 의미속성을 추출하고 이들 의미속성에 의한 명사의미체계를 구축하고, 명사의미체계의 응용분야에서 요구하는 정밀도의 정도에 따라서 명사의미 TAG를 설정하도록 하였다. 이러한 방식으로 설정된 명사의미체계는 구상물, 현상계, 추상물의 3개의 최상위 노드에서 최대 7개의 depth로 구성되었고, 전체적으로 약 1,430개의 중간개념노드로 구성되어 있다.

이러한 의미 TAG는 기계번역에서 역어 선택

표 10. 의미속성에 의한 수정된 명사의미체계

가곡 - 음악(7) - 예능(4) - 문화(3) - 학문(3) - 추상물(13) : [6]
가마 - 차2(7) - 차(2) - 교통수단(3) - 인공물(29) - 무생물(4) - 구상물(2) : [7]
가마니 - 보관물류4(2) - 보관물류(7) - 인공물(29) - 무생물(4) - 구상물(2) : [6]
가뭄 - 날씨(3) - 기상(7) - 자연현상(4) - 자연적현상(3) - 현상계(4) : [6]
가정부 - 고용직2(5) - 고용직(4) - 직업(7) - 사람(9) - 생물(3) - 구상물(2) : [7]
가죽 - 옷감2(2) - 옷감(3) - 재료물(6) - 인공물(29) - 무생물(4) - 구상물(2) : [7]
가지 - 식물부분3(6) - 식물부분(4) - 식물(5) - 생물(3) - 구상물(2) : [6]
간판 - 표지물1(4) - 표지물(7) - 인공물(29) - 무생물(4) - 구상물(2) : [6]
간호사 - 서비스직2(3) - 서비스직(2) - 직업(7) - 사람(9) - 생물(3) - 구상물(2) : [7]
갈대 - 풀1(7) - 풀(3) - 식물(5) - 생물(3) - 구상물(2) : [6]
감 - 과일류(11) - 농산물(4) - 생산물(3) - 음식물(6) - 무생물(4) - 구상물(2) : [7]
.....

시에 문장 내에서 명사와의 의미호응관계에 의한 적합한 영어 동사 선택하거나, 문장 분석 시의 구조적 모호성과 단어 의미의 모호성을 해소하는 데에 효과적으로 사용될 수 있다. 또한, 문장 생성 시의 문맥 및 사용자 수준에 적합한 어휘 선택에 활용될 수 있으며, 정보검색에서 키워드와 유사한 개념을 가진 어휘에 대한 정보까지 검색해 주어 검색의 재현율을 높이는 데 활용될 수 있으며, 문장 검사/교정 시스템에서 문장 내에서 철자나 문법의 오류가 아닌 문맥의 오류에 의해 틀린 부분을 검사하고 교정하는 것을 가능하게 한다.

앞으로는 명사의미체계를 이용한 의미유사도를 결정하는 방법에 대한 연구, 명사뿐만 아니라 용언과 부사에 대한 의미 TAG에 대한 연구와 의미중의성 문제를 해결하기 위한 의미 Tagging 시스템에 대한 연구가 진행되어야 할 것이다.

참 고 문 헌

[1] G.A.Miller 외2, "Introduction to WordNet : An On-line Lexical Database", 1993

[2] http://www.cogsci.princeton.edu/~wn

[3] http://www.cyc.com

[4] http://www.iijnet.or.jp/EDR

[5] R.Beckwith 외2, "Design and Implementation of the WordNet Lexical Database and Searching Software"

[6] P.F.Brown, "Word Sense Disambiguation using Statistical Methods", Proc. of 29th Meeting of the ACL, 1991

[7] Gale, William, K.Church and D.Yarowsky, "A Method for Disambiguating Word Senses in a Large Copus", Computers and Humanities, 1992

[8] D.Yarowsky, "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, " Proc. of the 15th Int'l Conf. on Computaional Linguistics, 1992

[9] M.Sanderson, "Word Sense Disambiguation and Information Retrieval", Proc. of SIGIR, 1994

[10] E.M.Voorhees, "Using WordNet to disambiguate word sense for text retrival", Proc. of ACM SIGIR Conference, 1993

[11] 박영자, 송만석, "사전에서 추출한 의미 속성에 기반한 명사 의미 클러스터링", 정보과학회논문지(B), 25권 3호, 1998. 3

[12] 강원석 외2, "영한 기계번역에서의 전치사구 처리를 위한 격의미 체계와 의미속성 집합", 한글 및 한국어 정보처리 학술대회, 1995

[13] "자연언어처리 특집", 정보과학회지, 14권 7호, 1997

[14] "한글과 한국어 처리 특집", 전자공학회지, 24권 9호, 1997

[15] 조평옥, 옥철영, "한국어 명사 의미계층구조 구축", 한글 및 한국어 정보처리 학술대회, 1997

[16] 윤평현, 국어 명사의 의미관계에 대한 연구, 한국과학재단 연구결과보고서, 94-0100-11-01-1, 1995

[17] 문유진, "한국어 명사를 위한 WordNet의 설계와 구현", 한국어정보과학회, 1996

[18] 김현진 외3인, "용언의 구문관계를 이용한 명사 분류", 한글 및 한국어 정보처리 학술대회, 1997

[19] 류법모 외4, "구문구조부착 말뭉치를 이용한 술어의 하위범주화 정보 구축", 한글 및 한국어 정보처리 학술대회, 1997

[20] 정연수 외2, "개념분류 기법을 이용한 한국어 명사 분류", 한글 및 한국어 정보처리 학술대회, 1995

[21] 조정미, 김길창, "분포 정보를 이용한 의미 중의성을 지닌 한국어 동사의 의미 분별", 한글 및 한국어 정보처리 학술대회, 1995

[22] 조정미, 김길창, "한국어 의미 해석시 중의성 해소에 대한 연구", 정보과학회지, 1996

[23] 김길창, 한국어 이해에 나타나는 중의성 문제 처리 모델에 관한 연구, 한국과학재단, 연구결과 중간보고서, 94-0100-01-04-3

[24] 박영자, 송만석, "자연언어처리를 위한 한국어 동사 명사의 개념 분류", 한글 및 한국어 정보처리 학술대회, 1992

[25] 심재기 외2, "의미론", 집문당, 1994

[26] 천시권, 김종택, "국어의미론", 형설출판사, 1975

[27] "국어정보 데이터베이스", 한국과학기술원, 시스템공학센터, 1997

[28] 임희석, 김진동, 임해창, "언어지식과 통계정보의 보완적 특성을 이용한 품사태깅", 한글 및 한국어 정보처리 학술대회, 1997

[29] "국어 어휘의 분류 목록에 대한 연구", 국립국어연구원, 1993

[30] 옥철영, 우리말 개념망 명사데이터 구축, 한국전자통신연구원, 위탁과제 최종보고서, 98-57-M20912