국어사전을 이용한 한국어 명사에 대한 상위어 자동 추출 및 WordNet의 프로토타입 개발

김 민 수† 김 태 연†† 노 봉 남†††

요 약

인간은 문장 안에 있는 명사를 인식할 때 그 명사의 상위 개념을 머리에 떠올린다. 컴퓨터에게 인간의 단어 인식작용을 시뮬레이션하기 위해서는 단어의 상위 개념(상위어)을 지식 베이스(WordNet)로 구축해야만 한다. 현재 한국에서는 많은 인력과 시간이 소요되기 때문에 WordNet의 작업을 시작하지 못하였으나, 컴퓨터의 성능이 급격히 향상되고 상용화된 MRD(Machine Readable Dictionary)가 이용가능하게 됨에 따라 자동으로 Wor dNet구축의 가능성을 보이고 있다. 본 논문에서는 한국어 MRD(Machine Readable Dictionary)의 명사의 정의(description)를 이용하여 자동으로 한국어 명사 WordNet을 구축하는 방법을 제안한다. 한국어 문장의 구조적인 특징을 분석하여 상위 개념 (상위어)를 추출하는 규칙을 제안한다. 그것은 중심적인 말이 보통 뒤에 나타난다는 것과 명사의 정의문은 특수한 구조를 갖는다는 것을 반영하였다. 또한, 이러한 규칙에 의해 만들어진 상위어들을 결합한 한국어 명사의 WordNet 프로토타입을 개발하였다. 약 2500개 표본 단어의 상위어를 추출한 결과 약 92여 퍼센트의 상위어가 옳게 추출되었다.

The Automatic Extraction of Hypernyms and the Development of WordNet Prototype for Korean Nouns using Korean MRD (Machine Readable Dictionary)

Min-Soo Kim,† Tae-Yeon Kim,†† Bong-Nam Noh†††

ABSTRACT

When a human recognizes nouns in a sentence, s/he associates them with the hyper concepts of nouns. For computer to simulate the human's word recognition, it should build the knowledge base (WordNet) for the hyper concepts of words. Until now, works for the WordNet haven't been performed in Korea, because they need lots of human efforts and time. But, as the power of computer is radically improved and common MRD becomes available, it is more feasible to automatically construct the WordNet. This paper proposes the method that automatically builds the WordNet of Korean nouns by using the description of nouns in Korean MRD, and it proposes the rules for extracting the hyper concepts (hypernyms) by analyzing structural characteristics of Korean. The rules effect such characteristics as a headword lies on the rear part of sentences and the descriptive sentences of nouns have special structure. In addition, the WordNet prototype of Korean Nouns is developed, which is made by combining the hypernyms produced by the rules mentioned above. It extracts the hypernyms of about 2,500 sample words, and the result shows that about 92 percents of hypernyms are correct.

*이 논문은 한국과학재단의 1994년도 특정 연구 신청 과제 연구비에 의하여 연구되었음
†준 회 원: 전남대학교 강사
††준 회 원: 광주예술전문대학 컴퓨터 그래픽 디자인학과 전임강사
†††종신회원: 전남대학교 전산학과 교수
논문접수 : 1995년 8월19일, 심사완료:1995년11월17일

1. 서 론

컴퓨터가 인간의 언어를 인식하고 그에 따라 동작할 수 있으려면 인간 언어의 특징을 분석하

여 컴퓨터가 알 수 있는 방법으로 변환하여야 한다. 인간은 문장 안에 있는 명사를 인식할 때 그 명사의 상위 개념을 머리에 떠올린다. 명사에 대한 상위 개념은 그 명사의 특성을 나타내고 명사를 분류할 수 있으며, 이것을 컴퓨터가 이용한다면 명사의 의미와 특징을 인식할 수 있을 것이다. 컴퓨터에게 인간의 단어 인식작용을 시뮬레이션하기 위해서는 단어의 상위 개념을 knowledge base(WordNet)로 구축해야만 한다[4, 7].

현재 한국에서는 많은 인력과 시간이 소요되기 때문에 WordNet의 작업을 시작하지 못하였다. 그러나, 컴퓨터의 성능이 급격히 향상되고 상용화된 MRD(Machine Readable Dictionary)가 이용가능하게 됨에 따라 자동으로 WordNet 구축의 가능성을 보이고 있다.

명사 상위 개념의 WordNet을 구축하기 위하여서는 이용가능한 MRD에서 명사를 선택한다. 일련의 알고리즘을 통해서 선정된 명사들의 상위 개념을 구할 수 있다. 이렇게 구해진 명사와 상위 개념의 관계를 중심으로 WordNet의 프로토타입을 개발한다. 명사의 상위 개념을 찾는데 수작업의 경우 모호성이 없지만 일관성(consistency)이 없고 비용이 많이 든다[3, 10]. 따라서, 여기에서는 MRD를 참고하여 컴퓨터가 선정된 명사의 정의(description)를 자동으로 분석함으로써 상위 개념을 지정하게 한다[1, 5]. 이렇게 알아낸 명사 상위 개념은 그것의 상위 개념을 가질 수 있으며, 이러한 관계를 WordNet의 프로토타입으로 개발하는 방법을 시도한다.

예를 들면, "공사비"는 "공사에 드는 비용"이므로 "공사비"의 상위 개념은 "비용"임을 자동으로 분석하도록 한다. 또한, "비용"은 "어떤 일을 하는데 드는 돈"이므로 "비용"의 상위 개념은 "돈"임을 자동으로 분석하도록 한다. "공사비" "비용" "돈"과 같이 명사의 상위 개념을 기반으로 WordNet프로토타입이 구축된다.

MRD에 기술된 명사의 정의는 일반 문장과는 달리 특수한 구조로 되어 있다. 예를 들면, "공사에 드는 비용"과 같이 문장에서 설명하고자하는 중심적인 단어 "비용"이 문장 뒷부분에 나타난다. 이것은 "공사비"와 "비용"의 상위 개념을 추출하는데 이용될 수 있다. 본 논문에서는 이와 같은 명사의 정의에 나타나는 문장 구조의 특징을 분석함으로써 상위 개념의 자동 추출이 가능하도록 하였다.

본 논문에서는 한국어 문장에서는 중심적인 말이 보통 뒤에 나타난다는 구조적인 특성과 한국어 MRD에서 명사 정의문의 특수한 구조적인 특징을 분석하여 상위어를 추출하는 규칙을 제안한다. 그리고, 한국어 MRD의 명사의 정의를 이용하여 자동으로 추출한 한국어 명사 상위 개념들을 결합하여 한국어 WordNet을 구축한다.

본 논문에서는 국어 품사중 명사를 선택하여 실험하였다. 일차적인 작업으로 정확한 상위 개념을 추출할 수 있도록 명사 정의 부분을 분석하여 10가지의 규칙을 만들어 냈다. 실험 결과를 분석함으로써 한국어 MRD에서의 명사 상위 개념 추출에 약 92여 퍼센트가 정확하다는 성과를 올렸고, 명사 정의의 구조적 특징을 확인할 수 있었다. 또한 추출된 명사와 상위 개념의 관계를 WordNet으로 구축할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 용어를 정의하였다. 3장에서는 국어 MRD의 구성을 살펴보았다. 4장에서는 상위어 추출 방법을 10가지 제시하였다. 5장에서는 실험 결과를 분석하였고, 마지막으로 6장에서는 결론과 향후 연구 방향을 제시하였다.

2. 용어 및 연구 동향

2.1 용어 정의

여기에서는 본 논문에 사용되는 상위어, 중심어, 구별어 등 용어 개념을 살펴본다.

【정의 1】 상위어(hypernym)

용어 B가 어떤 다른 용어 A 영역의 일부가 될 때 A를 B의 상위어라 한다. 다시 말해서, 용어 B의 집합이 용어 A의 집합에 포함되는 경우에 용어 A를 용어 B의 상위어라 하고 용어 B를 용어 A의 하위어라고 한다[6]. 이러한 정의에서 용어 A를 구성하는 의미적 특성 집합은 용어 B

를 구성하는 의미적 특성 집합을 포함한다는 추론을 할 수 있다. 즉, 용어 B는 용어 A의 의미적 특성을 상속받는다고 말할 수 있다. 이것은 의미적 특성의 상속이라는 관념에서 나온 것이다[10].

상위어는 서론에서 설명한 상위 개념에 해당된다. 상위어에 대한 수식적 표현은 {A는 B의 상위어이다 : $A(x) \supset B(x)$, if $\forall x$}로 한다. 이것에서 '모든 B는 A이다' 라는 개념을 추론할 수 있다. 예를 들면,

"미뉴에트"

{((음))3박자의 느리고 우미한 프랑스의 무도곡.}$에서 '무도곡은 미뉴에트의 상위어이다'라는 것을 알 수 있다. 따라서 '모든 미뉴에트는 무도곡이다' 라는 개념을 추론할 수 있다.

【정의 2】 중심어(headword)

명사의 정의에 있어서 가장 핵심이 되는 단어들이다. 다시 말해서, 명사의 정의 내용을 살펴보면 문장에서 설명하고자 하는 내용의 중심이 되고 가장 많은 수식을 받는 단어에 해당된다. 따라서 본 논문에 필요한 상위어 추출 방법은 이러한 중심어를 찾는 방법이라고 할 수 있다.

【정의 3】 구별어(differencia)

중심어에 의미론적 특성을 가미한 어구이다[10].

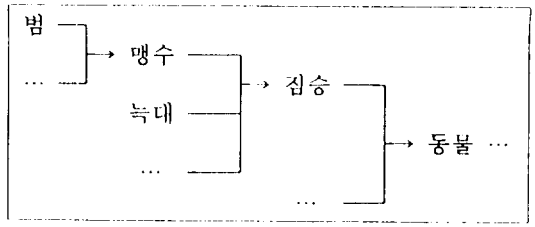
구별어는 일반적으로 수식어라고도 하며 중심어의 특성을 설명하고 중심어를 수식하는 역할을 하고 있다. 위의 예에서 보듯이 '미뉴에트'를 설명하고자 하는 중심적인 단어는 '무도곡'이며 '3박자의 느리고 우미한 프랑스의'라는 어구는 무도곡의 특징을 설명하는 부분으로 구별어이다.

【정의 3】 WordNet

정의 1에서 정의한 상위어들의 상관관계를 망으로 구성한 것이다.

$$\text{WordNet} = \bigcup_{x \in \text{Word}} PS(x),\ PS(x)$$
$$= \{(A(x),\ B(x)) \mid A(x) \supset B(x), \forall x\}$$

예를 들면, (그림 1)과 같이 '범'에 대한 상위어는 '맹수'이고 '맹수'에 대한 상위어는 '짐승', '짐승'에 대한 상위어는 '동물' 등처럼 구성한 것을 말한다.



〈그림 1〉 WordNet의 예
(Fig. 1) Example of WordNet

2.2 연구 동향

자연어 처리에 사전을 이용하는 방법은 영어의 경우에 활발히 진행되어 왔다. 영어의 경우는 현재까지 나와있는 MRD만 해도 수십 가지가 있다. 대표적으로 Longman Dictionary of Contemporary English (LDOCE), Webster's Seventh New Collegiate Dictionary (W7), Merrian-Webster Pocket Dictionary (MWPD), Oxford English Dictionary (OED), Collins Dictionary of English (COLLINS) 등이 있다. 이중에서 W7과 LDOCE가 가장 많이 이용되고 있다. Krovertz와 Croft는 MRD를 이용하여 단어 의미의 모호성을 해결하려고 하였다[3]. E. Fox 연구팀은 정보검색에 이용할 수 있는 큰 시소러스(thesaurus)를 구축하였다[2].

반면에 한국어에서는 쉽게 이용할 수 있는 상용 MRD는 매우 적다. 이러한 환경에서도 MRD를 사용하여 시소러스를 구축하여 형태소 분석이나 구문 분석에 이용하는 연구가 활발히 진행되고 있다[8]. 본 논문에서는 연구를 위해서 제작된 MRD를 이용하여 명사의 상위 개념을 WordNet으로 구축하는 작업을 수행하고자 한다.

3. 국어 MRD의 구성

본 논문에서 사용하는 국어 MRD는 '엣센스 국어사전'이다. MRD로부터 자동으로 어떤 정보를 추출하는 방법은 사전의 구조에 의존하기 때문에 여기에서는 먼저 이 MRD의 구조를 소개한다.

2.1 3.1 MRD의 구성 형식

본 논문에 사용된 '엣센스 국어사전'의 구성은 다음과 같다[13].

"단어"

<<품사>>

{1. 단어에 대한 설명. 2. …… @ 예제. [단어의 특징 설명].}$

쌍따옴표 안의 단어가 있고 그에 대한 설명은 괄호('{', '}') 안에 쓰여져 있으며 설명의 끝은 $로서 표시하였다. '<<'와 '>>' 사이에는 단어에 대한 품사가 표시된다. 이것으로 이 단어가 명사인지 아닌지를 구분할 수 있다. '@' 다음에 나오는 내용은 그 단어에 대한 예제로서, 예를 들어 '지느러미'에 대한 설명중 '등~'과 같이 '등지느러미'로 쓰이는 경우가 있음을 나타낸다. 괄호('[', ']') 안의 내용은 단어에 대한 특징을 설명하는 부분으로 상위어 추출에 관계가 없는 내용이다. 여기에서 단어에 대한 정의가 '1. …… 2. ……'처럼 여러개 되어있는 것은 같은 어원에서 파생된 뜻을 나타낸다. 어원이 다른 경우는 위와 같은 구조가 같은 단어에 다른 내용으로 계속에서 표현하고 있다. 이러한 단어를 다의어(equivocal word)라고 한다.

본 논문에서 이용한 MRD는 '<<'와 '>>'사이에 품사를 결정하는 단어가 있다. 검색하는 사전의 크기를 줄이기 위해 MRD에서 명사만을 뽑아내었다. 각 품사별로 구분함으로써 상위어에 대한 품사를 일정하게 유지할 수 있다. 즉 어떤 단어가 명사이면, 그 상위어도 명사이어야 한다. 또한 명사는 명사끼리, 동사는 동사끼리, 형용사는 형용사끼리 사전을 따로 관리할 수 있다.

여기에서는 명사에 대한 상위어를 추출하는 방법을 제시한다. 각 품사마다 단어에 대한 정의 형식이 다르다. 다만, 일반적인 공통점은 중심적인 어구는 문장 뒷부분에 나온다는 것이다. 이러한 특징을 살리어 명사에 대한 상위어 사전뿐만 아니라 동사나 형용사에 대한 상위어 사전을 구축할 수 있다.

3.2 명사 상위어 사전의 구성

상위어 추출 방법에 따라 알고리즘을 수행하면 다음과 같은 형식으로 나타난다.

"단어"

#1n 상위어1 상위어2…… 상위어n

#2m 상위어1′ 상위어2′ … 상위어m′

쌍따옴표 안의 단어에 대한 상위어는 아래에 '#' 표시로서 시작되고 다음에 나오는 숫자는 어원이 다른 단어가 있는 경우 '#'이 여러개 나올 수 있다. 그리고 이렇게 어원이 다른 경우를 다의어(equivocal word)라 한다. 숫자 n은 상위어 갯수에 해당된다. 명사 상위어 사전은 이렇게 명사에 대한 상위어를 나열하고 있으며, 또한 상위어로 쓰인 단어도 자신의 상위어를 가질 수 있다. 따라서 이렇게 구성된 상위어 사전으로 단어들 간의 상관관계를 망으로 구성할 수 있다.

4. 상위어 추출 방법

4.1 국어 MRD의 분석

여기에서 사용한 한국어 명서 상위어 추출 방법은 국어사전에 있는 단어의 정의를 기반으로 하여 구성한 MRD을 이용하는 방법이다. 따라서 MRD의 특징을 분석하고 자동으로 상위어를 추출할 수 있는 방법을 찾아야 한다.

한국어의 특징은 중심어 후행성과 어순의 자유성을 들 수 있다[9]. 동사나 형용사 등의 서술어로 사용되는 용언 구는 그 절의 맨 뒤에 위치한다. 그러나 서술어에 대한 보조어는 순서에 상관없이 자유로이 나타날 수 있다. 명사의 경우는 주어나 목적어로 쓰이기 때문에 문장 중간에 나타난다. 그렇지만, 사전에서는 명사를 설명하는 것이므로 중심이 되는 명사는 대부분 마지막에 나타난다. 이것이 영어와의 차이점이다. 즉, 영어는 어떤 명사가 나오고 그 설명은 'that'나 'which' 등의 관계사나 'of'나 'for' 등의 전치사를 연결하여 뒤에서 설명한다. 이러한 한국어적 특징에 따라 명사를 설명하는 구별어와 설명 받는 중심어를 구분할 수 있다.

한국어 사전에서 명사에 대한 정의는 일반적으로 구별어와 중심어의 결합된 형태를 이룬다. 그리고 중심어는 어떠한 기능을 갖는 어구의 수

식을 받거나 그 어구의 목적어가 되기도 한다. 이러한 어구는 미리 사전을 분석하여 알 수 있다.

예를 들어,

"지기"

{1.종이로 만든 그릇의 총칭. ……}$

위에서처럼 '종이로 만든'은 구별어이고 '~의 총칭'은 포괄적인 의미를 가진 어구로서 중심어를 부연 설명하는 것으로 볼 수 있다. 따라서 '지기'의 상위어 '그릇'을 추출하기 위해서는 '~의 총칭'과 같은 패턴이 알려져 있어야 한다.

4.2 추출 방법

상위어 명사를 추출하는 방법은 다음 10가지 규칙에 따라 수행한다.

▣ 규칙 1

사전에서 명사에 대한 정의는 명사로 끝나는 것이 일반적이다. 또한, 한국어의 특성상 중심적인 어구는 보통 문장 뒷부분에 나타난다. 따라서, 문장 마지막에 나타나는 명사를 추출하는 것이 바람직하다. 다시 말해서, 문장 구성이 구별어와 중심어로 되어있는 경우에서 중심어를 추출하는 방법이다[10]. 예를 들면,

"미량"

{1. 아주 적은 양. @~의 극약.}$

을 들 수 있다. 이 문장 중 중심어는 '양'으로서 문장의 마지막에 나타난 단어에 해당됨을 알 수 있다. '아주 적은'은 중심어를 설명하여주는 구별어에 해당된다. 이것을 수식으로 표현하면 다음과 같다.

$$M_1(S_1) = F_{\text{구별어}}(S_1) \cdot F_{\text{중심어}}(S_1).$$

▣ 규칙 2

어떤 단어를 지칭하는 의미를 가지고 있는 동사의 목적어를 상위어로 추출하는 방법이다. 이러한 의미를 갖는 동사는 '~을 부르다', '~을 뜻하다', '~을 이르다', '~을 나타내다', '을 일컫다', '을 가리키다', '을 비유하다', '을 말하다' 등이다. 예를 들면,

"도령"

{총각을 대접하여 일컫는 말. @이~. (높)도련님}$

여기에서 문장 뒷부분에 '일컫다'라는 동사가 나타나고 그 동사의 목적어로 '총각'이 쓰였다. 따라서 '총각'이 문장의 중심어가 된다.

이것을 수식으로 표현하면 다음과 같다.

$$M_4(S_1) = F_{\text{구별어}}(S_1) \cdot F_{\text{중심어}}(S_1) \cdot (\text{'을' or '를'}) \cdot (\cdots) \cdot F_{\text{이용패턴}}(S_1)$$

▣ 규칙 3

소유격 조사 다음에 어떤 부분을 표시하는 단어가 나타날 때, 조사 앞의 명사를 상위어로 추출하는 방법이다. 이러한 부분을 표시하는 단어는 '~종류', '~총칭', '~비유', '~한가지', '~덩이', '일종' 등이다. 예를 들면,

"돈견"

{1. 돼지와 개. 2. 못난 사람의 비유. 또는, 자기 아들의 겸칭. 돈아.}$

여기에는 '의 비유'와 '의 겸칭'이 사용되었다. 이것은 소유 받는 명사의 다른 표현이나 포괄하는 것을 뜻하는 의미를 갖는다. 따라서 위와 같은 패턴이 사용되었을 경우에는 소유격 조사 앞의 단어(여기서는 '사람'과 '아들')를 중심어로 선정하는 게 타당하다.

소유격 조사를 이용한 방법에 대한 수식은 다음과 같다.

$$M_3(S_1) = F_{\text{구별어}}(S_1) \cdot F_{\text{중심어}}(S_1) \cdot (\text{'의'}) \cdot (\cdots) \cdot F_{\text{이용패턴}}(S_1).$$

▣ 규칙 4

접속조사나 접속부사의 앞·뒤 단어를 상위어로 추출하는 방법이다. 등위 접속조사와 부사로는 '와', '과', '이나', '또는', '및' 등이 있다. 예를 들면,

"덕량"

{어질고 너그러운 마음씨와 생각.}$

문장 뒷부분에 접속조사나 접속부사가 사용될 경우에는 접속사의 양쪽 단어가 설명하는 명사와 많은 상관관계를 갖는다. 따라서 이러한 접속사 양쪽의 단어를 중심어로 하는 것이 타당하다.

접속조사나 접속부사가 이용된 방법으로 수식

표현은 아래와 같다.

$M_4(S_i) = F_{구별어}(S_i) \cdot F_{중심어}(S_i)$
$\cdot F_{(접속조사\ 또는\ 접속부사)}(S_i)$
$\cdot F_{중심어}(S_i)$.

▣ 규칙 5

보어로 사용되는 명사를 상위어를 추출하는 방법이다. 이러한 경우는 대게 뒤에 사용되는 동사가 보어를 사용하는 의미를 갖거나 어떤 상태로 변한다는 의미를 갖는 동사들이 해당된다. 예를 들면, '~쓰다', '~이용되다', '~되다', '~되는 부분' 등이다. 예를 들면,

"데님"

{능직의 두꺼운 면직물. 가구의 커버나 작업복으로 이용됨.}$

이러한 경우는 설명 명사의 용도를 나타내는 정의에 해당되고 설명 명사와 상관관계를 갖는다. 즉, '데님'은 의류이면서 주로 작업복으로 이용되기 때문에 '작업복'으로 불러도 큰 문제가 되지 않는다.

이것을 수식으로 표현하면 다음과 같다.

$M_3(S_i) = F_{구별어}(S_i) \cdot F_{중심어}(S_i)$
('로', '으로', '에' 등)$\cdot F_{이용패턴}(S_i)$

▣ 규칙 6

명사에 동작의 의미를 가미하기 위해 동사형 어미('하다', '되다', '시키다' 등)를 첨가한 단어로써 다시 명사형 어미 '링'를 붙인 것을 추출하는 방법이다. 또한, 동사에 명사형 어미 '링'을 붙여서 명사화한 것을 추출하는 방법이다. 예를 들면,

"명오"

{((가))사물에 대하여 밝게 인식함. 또는, 그러한 힘.}$

위의 예에서 쓰인 것처럼 '인식함'은 '인식하다'에서 명사형으로 변형된 것이다. 그리고 이러한 경우는 명사 '인식'에 동작의 의미를 나타낸 것이다. 따라서, 동사형 어미를 뺀 '인식'이란 명사만으로도 충분한 의미를 전달할 수 있다. 따라서 이러한 경우도 동사형 어미 앞의 명사를 중심어로 추출할 수 있다.

이렇게 동사형 어미가 접합된 표현의 수식은 아래와 같다.

$M_5(S_i) = F_{구별어}(S_i) \cdot F_{중심어}(S_i)$
$\cdot F_{(동사형\ 어미)}(S_i)$

▣ 규칙 7

규칙 6에서처럼 명사에 동작의 의미를 첨가한 단어로서 그 동작을 행하고 난 후의 상태나 또는 그러한 행위를 나타내는 명사('~ 일', '~ 것' 등)가 문장의 마지막에 나타날 때 동사형 어미 앞의 단어를 상위어로 추출하는 방법이다. 예를 들면,

"더블 스틸"

{((체))야구에서, 두 사람의 주자가 동시에 도루하는 것.}$

'하는 것'은 행위의 상태를 나타내고 '하는 일'은 행위의 동작을 나타낸다. 그리고 이러한 행위의 목적이 되는 단어는 주로 위 어구 바로 앞에 나타난다. 따라서 그 앞 단어(여기서는 '도루')를 중심어로 추출하는 것이 바람직하다.

이렇게 '일'이나 '것'으로 끝나는 문장 표현을 이용한 방법의 수식은 아래와 같다.

$M_7(S_i) = F_{구별어}(S_i) \cdot F_{중심어}(S_i) \cdot$
$F_{(동사형\ 어미)}(S_i) \cdot$('일', '것').

▣ 규칙 8

두 문장을 연결시켜주는 접속의 의미로 사용되는 어구('~하거나', '~되거나', '~시키거나', '이거나' 등)가 사용될 경우 그 앞의 단어를 상위어로 추출하고 뒤에 연결되는 절은 위에 제시한 방법으로 다시 검색하는 방법이다. 예를 들면,

"독공"

{혼자서 공부하거나 혼자서 일하는 것.}$

여기에서 사용되는 어구 '하거나'와 같은 단어가 사용될 경우는 대게 문장 뒷부분에 규칙 5, 규칙 6, 규칙 7과 같은 형식이 나타난다. 따라서 문장 뒷부분에서 추출된 단어와 대등한 의미를 갖는 단어가 '하거나'와 같은 어구 앞에 사용될 것으로 추측할 수 있다. 따라서 '하거나'와 같은 어구 앞에 쓰인 단어를 상위어로 추출하는 것도 문제가 되지 않는다.

이것에 대한 수식표현은 아래와 같다.

$M_8(S_i) = F_{\text{구별어}}(S_i) \cdot F_{\text{중심어}}(S_i)$
$\cdot$('하거나', '되거나' 등)$\cdot$(…).

▣ 규칙 9

규칙 1이나 규칙 6의 경우는 원하지 않는 단어가 상위어로 추출될 수가 있다. 예를 들면, '~씀', '~음'으로 끝나는 경우는 대게 '~하였음', '~있음' 등처럼 상태나 존재를 나타내는 동사의 명사형 단어이다. '~만함', '~같음' 등처럼 동급 비교를 나타내는 동사의 명사형 단어도 해당된다. 이러한 경우 불용어(stopword)로 처리하여 상위어 리스트에서 삭제하는 방법이다. 예를 들면,

"도역 유도"

{도둑에게도 도둑 나름의 도리가 있음을 이르는 말.}$

위의 예처럼 '을 이르는 말'의 경우는 규칙 2의 추출 방법이 적용된다. 따라서 '있음'이라는 단어가 상위어로 추출되게 되는데, 이 단어는 '도역 유도'와 관계가 적다는 것은 통례로 알 수 있다. 따라서 이 단어는 상위어로는 부적당하다.

불용어를 처리하는 문장 표현을 수식으로 표현하면 다음과 같다.

$M_9(S_i) = F_{\text{구별어}}(S_i) \cdot$
('있음', '없음', '하였음' 등).

▣ 규칙 10

예제로서 사용되는 구절이나 문장 등은 상위어 추출 방법에 적용하지 않는 방법이다. 예제가 사용된 예를 들어보면 다음과 같다.

"더블 펀치"

{((체))권투에서, 두 번 연달아 치는 펀치. @~를 날리다.}$

위에서 '@'표시 다음에는 사용 예제를 보여주고 있다. 이러한 구성은 사전마다 다르겠지만 본 논문에서 이용한 사전에는 '@'다음에는 항상 예제가 나오고 '['와 ']'로 묶어서 설명이나 속담 등의 예가 표시되었다. 이러한 내용들은 상위어를 추출하는 것과 상관관계가 적으므로 상위어에서 제외하도록 한다.

예제나 반대어 속담 등의 표현은 일반적으로 시작 문자를 같은 문자로 표기한다.

$M_{10}(S_i) =$ ('@', '<', '[' 등)
$\cdot F_{\text{나머지문장}}(S_i)$

위에서 제시한 10가지 방법을 종합적으로 표현하면 아래와 같다.

$$H = (\overline{M_{10} \wedge M_9}) \vee (M_8 \wedge M_7 \wedge M_6 \wedge M_5 \wedge M_4 \wedge M_3 \wedge M_2) \vee (M_1)$$

규칙 9와 규칙 10은 상위어로 사용되지 않는 부분이니 위에 바를 그렸다. 그리고 규칙 1의 경우는 모든 패턴 검사가 끝난 후에 처리되는 부분이므로 마지막에 따로 두었다.

5. 실험 결과

본 논문에서 채택한 국어사전은 명사가 124,850 단어로 이루어져 있다. 12만여 개의 명사 단어 중 2,500개의 연속적인 표본을 추출하여 검사해 보았다. 본 논문에서 제시한 방법으로 명사에 대한 상위어를 추출한 결과 88.4%의 명사가 정확한 의미의 상위어를 추출하였고, 5.6% 정도의 명사가 부분적으로 부정확한 상위어가 추출되었다. 그리고 전혀 엉뚱한 결과를 낳은 경우가 2%이며, 상위어가 한 단어도 추출되지 않는 경우가 4% 정도 되었다. 따라서, 94% 정도의 명사는 상위어를 찾을 수 있다고 할 수 있다.

실험결과를 각 규칙별로 올바른 추출 횟수와 잘못된 추출 횟수를 비교해 보았다(표 1). 실험결과를 보면 규칙 1에 의해서 추출한 상위어 갯수가 많음을 알 수 있다. 또한, 전체적인 정확도에도 비중이 큼을 알 수 있다. 규칙 2, 규칙 3, 규칙 4 그리고 규칙 7은 상대적으로 정확히 상위어를 추출하였다. 규칙 5의 경우는 부사가 추출되는 경우가 상당히 있었고, 규칙 6의 경우는 절단에 의하여 동사나 형용사의 의미를 파악하기 힘든 경우가 있었다. 규칙 5와 규칙 6은 다른 방법에 비하여 정확도 면에서 뒤떨어졌다. 이것은 구문 구성을 더 연구함으로써 극복될 수 있을 것이라고 본다. 불용어 처리 방법인 규칙 9에 의해 걸러진 상위어 갯수는 123개였다.

실험결과에서 상위어 갯수는 각 명사 당 평균 1.5개 정도가 추출되었고 92.43%가 정확한 상위어로서의 의미를 갖고 있었다. 나머지 부정확한

상위어가 추출되는 경우는 국어사전의 오류와 한국어 특성 때문이다. 따라서 상위어 추출 방법은 다음과 같은 문제점을 갖고 있다.

첫째, 국어사전 자체가 상위어 추출 목적으로 쓰여진 것이 아니기 때문에 정확한 상위어 추출의 문제점이 나타났다. 단어에 대한 정의가 서술적인 설명으로 되어 있어서 추출된 단어가 정확한 상위어로써의 의미를 갖고 있지 않는 경우가 있다. 예를 들면,

"선시선종"

<<(명)>>

{처음부터 끝까지 한결같이 잘한다는 뜻.}$

처럼 고사성어의 경우 그 뜻 풀이가 나타나 어느 부분을 상위어로 추출하기가 곤란한 경우이다.

둘째, 불용어로 처리하여야 하는 경우가 규칙 9와 규칙 10 외에도 더 존재한다는 것이다. 예를 들어,

"보급"

<<(명)>>

{1. 널리 펴서 골고루 미치게 함. @~률/새 기술을 ~하다.}$

위에서처럼 '함'을 불용어로 처리하는 것은 신랑이 신부에게 보내는 '함'의 경우가 있어서 곤란하다. 그리고 '~등'처럼 여러 가지를 나열할 때 사용하는 것은 상위어로서 불필요한 것이지만 불을 켜는 '등'이란 명사가 있어서 불용어로 처리하는 것이 곤란하다.

셋째, 제시한 방법으로 추출한 명사가 설명하는 명사와 상관관계가 적은 경우가 있다. 예를 들면,

"볏귀"

<<(명)>>

{((농)) 1. 쟁기 뒷바닥의 삼각형으로 된 부분.}$

처럼 설명하는 명사 '볏귀'와 추출된 명사 '삼각형'과의 의미적 상관관계가 적다.

넷째, 한국어는 동사나 형용사 형태소에 'ㅁ'등을 붙여서 명사형으로 변환시켜서 사용하는 경우가 많다. 결국 이러한 동사나 형용사의 명사형이 상위어로 추출되는 경우, 추출된 단어가 사전에 정의되어 있지 않기 때문에 다음 상위어를 추출할 수가 없게 된다.

다섯째, 명사의 정의에 있어서 그 특징을 서술하는 부분이 간혹 있다. 이러한 부분은 예외 상황으로 추출 방법에서 벗어나게 된다. 예를 들면,

"가다랭이"

<<(명)>>

{ 1. ((동))고등어과의 바닷물고기. 몸길이 1미터 가량. 방추형으로 살지고,몸빛은 등이 검은 자주색,배는 은백색임.}$

"가돌리늄"

<<(명)>>

{((화)) 1. 희토류 원소의 하나. 기호 Gd, 원자 번호 64, 원자량 157\.25\. 흰빛의 금속으로 단단함. 수은과 합금을 만들어 충치의 구멍을 메우는 데 씀.}$

여기에서는 '가다랭이'라는 물고기의 특징과 '가돌리늄'이라는 화학 원소의 특징을 설명하는 부분이 나와 있어서 제시한 방법으로 상위어를 추출할 경우 '가량', '은백색', '단단' 등이 추출된다. 이것은 설명하는 명사와 직접적인 관계가 적어서 상위어로 부적합하다.

〈표 1〉 각 추출 규칙의 결과
〈Table 1〉 Result for extracting by each Rules

	올바른 추출갯수	잘못된 추출갯수	정확도 (%)
규칙 1	2687	209	92.8
규칙 2	51	0	100
규칙 3	142	0	100
규칙 4	95	1	99.0
규칙 5	38	9	80.9
규칙 6	276	58	82.6
규칙 7	97	0	100
규칙 8	22	2	91.7
합 계	3408	279	92.43

6. 결론 및 개선 방향

본 논문에서는 한국어 MRD(Machine Readable Dictionary)의 명사의 정의를 이용하여 자동으로 한국어 명사 WordNet을 구축하는 방법을 제안하였다. 한국어 문장에서는 중심적인 말이 보통 뒤에 나타난다는 구조적인 특성과 명사의

정의문의 특수한 구조적인 특징을 분석하여 상위어를 추출하는 규칙을 제안하였다. 결과로 본 논문에서는 명사의 상위 개념을 추출할 수 있도록 명사 정의 부분을 분석하여 10가지의 규칙을 제시하였다.

본 논문에서 제시한 방법으로 실험을 한 결과 약 92여 퍼센트가 명사에 대한 상위 개념을 정확히 추출할 수 있었다. 부정확한 결과는 한국어의 특성에 따른 것이기 때문에, 한국어 문장의 구조적 특징을 분석하여 제시한 방법을 보충 또는 개선함으로써 더 나은 결과를 얻을 수 있을 것이다. 5장에서 제시한 실험결과는 연구에서 사용한 MRD에 종속적 면이 있지만 어느 정도 구문의 특성을 이용한 방법인 만큼 기타 다른 국어 사전에서도 이용될 수 있다고 본다.

본 논문에서 제시한 프로토타입으로 컴퓨터가 자연언어 처리를 조금 더 효율적으로 수행할 수 있으리라 본다. 따라서, 이 연구 결과는 모든 명사에 확대 적용된 WordNet을 구축하는데 많은 기여를 할 것이다. 또한, 명사 뿐만 아니라 형용사, 동사, 부사 등의 단어 연관관계를 트리구조로 구축하는 데에도 동일한 방법의 적용이 가능하므로 시간과 비용의 절감효과가 크다고 하겠다.

따라서, 앞으로 형태소 분석을 통하여 동사나 형용사를 명사에 연계시키는 작업, 명사 정의의 특성의 파악에 대한 연구가 필요하며, 마지막으로 명사 정의에 있어서 설명부분의 처리가 연구과제로 남아 있다.

참 고 문 헌

[1] R.Byrd and G.Heidorn, "Extracting Semantic Hierarchies from a Large On-line Dictionary", Proc. of COLING-85, pp.299-304, July, 1985.

[2] E.A.Fox and J.T.Nutter, "Building a Large Thesaurus for Information Retrieval", Proc. of ACL Conference, pp.101-108, 1988.

[3] R.Krovetz and W.B.Croft, "Word Sense Disambiguation Using Machine-Readable Dictionaries", Proc. of ACM SIGIR Conference, pp.127-136, 1989.

[4] Y.Moon and Y.Kim, "Critiques for Semantic Errors in the Korea Text", Proc of the Int'l Conference on AI in Engineering, pp. 155-161, 1990.

[5] Y.Tomiura and T.Nakamura, "Logical Form of Hierarchical Relation on Verbs and Extracting it from Definition Sentences in a Japanese Dictionary", Proc. of COLING-92, pp.574-580, Aug., 1992.

[6] P.Hernert, "KASSYS : a Definition Acquisition System in Natural Language", Proc. of COLING-94, pp.263-267, Aug., 1994.

[7] M.Chung and D.I.Moldovan, "Parallel Natural Language Processing on a Semantic Network Array Processor", IEEE Trans. on Knowledge and Data Engineering, Vol. 7, No. 3, pp.391-405, Jun., 1995.

[8] 이상조, "기계번역 시스템을 위한 사전 구성", 정보과학회지, Vol. 7, No. 6, pp.25-30, 1989.

[9] 나동렬, "한국어 파싱에 대한 고찰", 정보과학회지, Vol. 12, No. 8, pp.33-46, 1994.

[10] 문유진, 김영택, "한국어 명사의 Hypernym 자동 추출 방법", 한국정보과학회 학술발표논문집, Vol. 21, No. 2, pp.613-616, 1994.

[11] 신기철, 신용철, "새우리말 국어사전", 삼성출판사, 1984.

[12] 이숭녕, "표준 국어 대사전", 한영출판사, 1990.

[13] "엣센스 국어사전", 민중서관, 1993.

김 민 수

1993년 전남대학교 전산통계학과 졸업(학사)

1995년 전남대학교 대학원 전산통계학과(이학석사)

1995년~현재 전남대학교 시간강사

관심분야 : 데이타 통신, 신경망, 자연언어 처리, 통신 보안 등.

김 태 연

1986년 전남대학교 계산통계학과 졸업(학사)
1988년 전남대학교 대학원 계산통계학과(이학석사)
1993년~현재 광주예술전문대학 컴퓨터그래픽 디자인학과 전임강사
관심분야 : 통신망 관리, 분산처리시스템, 통신 보안, 컴퓨터 그래픽스 등.

노 봉 남

1978년 전남대학교 수학교육과 졸업(학사)
1982년 한국과학기술원 전산학과(공학석사)
1994년 전북대학교 대학원 전산통계학과(이학박사)
1983년~현재 전남대학교 전산학과 교수
관심분야 : 객체지향 시스템, 통신망관리, 정보 보안, 컴퓨터와 정보사회 등.