한국어학 제17집(2002.12.30). 한국어학회(pp.301~329)

한국어 의미망 구축과 활용* **

- 명사를 중심으로 -

최호섭 · 옥철영

Abstract

It is important to construct Knowledge Base(like Thesaurus, Ontology, Semantic Network, etc) which can be applied to the whole field of natural language processing. For example, WordNet, Kadokawa Thesaurus, and Lexical FreeNet represent the most typical Knowledge Base in natural language processing. Many Knowledge Bases constructed in many fields does not come up to our expectations in Korean language processing.

In order to construct an effective Knowledge Base, various language resources such as corpus, dictionary, synonym dictionary and WordNet have to be integrated one another, and the knowledge base has to consist of chain of morpheme-word-phrase-collocation-idiom-corpus.

This paper presents a construction method and application of Korean Semantic Network (KSN). The KSN is based on Korean dictionary and Sejong corpus, and is applied to text processing, word sense disambiguation (WSD), semantic analysis, query pattern analysis in information retrieval, and so on. This paper deals with the following contents: (1) We point out problems of thesaurus and semantic network that look like a hierarchical structure of words, and compare KSN with them. The KSN has 1:1 relationship between word and sense, not 1:n relationship that an existing thesaurus and semantic network has. (2) We present KSN component parts and a construction method. The KSN has noun semantic hierarchy structure linked to predicates, semantic class, proper noun, semantic information, and so on. The links are resulted from consideration of a paradigmatic relation and a syntagmatic relation within sentence. For reference, the KSN consists of

---

* 이 논문은 2002년 한국어학회 국제학술대회에서 발표한 것을 일부 수정한 원고이다.
** 이 논문은 2002년 정보통신부에서 지원하는 대학기초연구지원사업으로 수행되었다.

dictionary, morpheme information, parts of speech information, construction information, proper noun information (name entity), noun semantic hierarchical structure, predicates classification structure, semantic class relation, idiom, semantic information, and so on. (3) We present that WSD using the KSN is more effective than one using an existing thesaurus and semantic network.

주요어 : 지식베이스(knowledge base), 의미망(semantic network), 시소러스(thesaurus), 온톨로지(ontology), 어휘 데이터베이스(lexical database)

1. 서론

컴퓨터의 발달은 자연스럽게 인간의 언어를 어떻게 하면 컴퓨터가 인식하게끔 하는가에 관심을 가지게 하였다. 더불어 컴퓨터를 이용하여 인간의 언어를 어떻게 효과적으로 연구할 수 있는가에도 관심의 방향을 전환할 수 있었다. 이러한 인간과 컴퓨터의 상호작용(Human and Computer Interface; HCI)을 연구하는 것 중의 하나인, 언어와 컴퓨터의 상호작용에 대한 연구는 학문의 중심을 어디에 두는가에 따라 그 명칭을 달리하고 있다. 즉 전산학을 중심으로 한 자연언어처리(Natural Language Processing), 언어학을 중심으로 한 전산언어학(Computational Linguistics)이 바로 그것이다[1]. 국내에서도 1980년대 이후 빠른 속도로 이들 학문에 대한 연구가 진행되었으며, 현재 기초적인 연구 토대를 어느 정도 체계적으로 잡아 가고 있는 실정이다.

자연언어처리와 전산언어학의 하위 분야와 응용 분야는 다양한다. 형태론, 통

1) '자연언어처리'와 '전산언어학'을 구분하는 것은 위에서 언급했듯이, 학문의 중심을 어디에 두는가에 달려 있다. 서상규 · 한영균(1999)에서도 두 학문 분야의 경계를 짓는 것은 무의미하다고 지적하면서, 굳이 경계를 두자면 연구의 목적과 주된 결과가 소프트웨어의 개발인가 아니면 언어 내적 구조에 관한 구명인가에 따라서 두 학문을 구분할 수 있다고 기술하고 있다. 그러나 최근에는 이러한 경계조차도 의미가 없을 정도로 이론적 · 실용적 측면을 모두 고려한 연구가 많아 지고 있는 실정이어서 특별히 구분하지 않는 듯하다.

사론, 의미론 등을 전산 처리(Computational Processing)에 효율적일 수 있도록 하는 연구와 컴퓨터를 이용하여 텍스트를 언어학적으로 연구하는 코퍼스언어학, 통계(계량)언어학 등이 있고, 품사 태깅, 형태소 분석, 구문 분석, 의미 분석 등과 같은 실질적인 전산 처리 연구도 있다. 나아가 정보검색, 기계번역, 문서분류, 문서관리, 지식경영 등은 자연언어처리와 전산언어학의 응용 분야라 할 수 있다. 이러한 많은 분야 들 중 일부는 국외의 연구 못지 않게 상당한 수준의 연구 단계까지 발전하였다. 또한 '언어(국어)정보산업'이라는 명칭을 사용하여 새롭게 급부상하는 사업 중의 하나로 평가하여 국가나 기업 차원의 지원도 계속 이어지고 있다.

이러한 발전 속에서도 아직까지 미흡한 분야가 의미와 관련된 전산적 처리 분야라 할 수 있다. 특히 영어 어휘 데이터베이스인 WordNet과 같이, 어휘들을 긴밀하게 연결시켜 자연언어의 개념적 양상을 체계적으로 정리하려는 연구가 많이 진행되고 있다. 이것은 자연언어처리에서 해결하고자 하는 부눈 중 단어 중의성 해소(Word Sense Disambiguation)와 밀접한 관계가 있을 뿐 아니라, 정보검색·기계번역 등과 같은 응용 분야에서도 필요로 하는 연구이다. 즉 단순한 어휘 나열이 아니라, 어휘들간의 긴밀한 연결 상태를 하나의 망(network)처럼 구성하고, 그것을 데이터베이스화하는 것이다. 대표적인 예들이 소위 지식베이스(Knowledge Base)라 할 수 있는 시소러스(Thesaurus), 의미망(Semantic Network), 온톨로지(Ontology) 등이 있다[2]. 이것은 언어자원(Language Resource)의 효율적인 관리, 자연언어의 전산적 처리 향상, 사용자 인터페이스를 통한 어휘의 개념 파생 양상의 시각화 등의 기대 효과를 가질 수 있다. 학문적인

2) '지식베이스'라는 말은 일반적으로 "전문가 시스템(Expert System)의 구성 요소 중 하나로서, 특정 분야의 전문가가 지적 활동과 경험을 통해 축적한 전문 지식이나 문제 해결에 필요한 사실과 규칙 등이 격납되어 있는 데이터베이스"를 말한다. 즉 문제 해결의 방법이 전문가에 따라 다르므로 지식베이스의 역할과 구성도 달라질 수 있는 것이다. 자연언어처리나 전산언어학적 입장에서 본다면 '지식베이스'는 지식의 표현이라 할 수 있는 자연언어의 전산적 처리 방안을 모색하기 위한 체계적이면서 전문적인 어휘 데이터베이스로 생각할 수 있다. 하지만 이러한 '지식베이스'에서 중요한 지식 획득(knowledge acquisition) 방법이나 전문가의 지식 반영 방법에 대한 구체적인 방법론은 제시되지 못하고 있는 실정이다.

관점에서 본다면, WordNet과 같은 한국어를 대상으로 한 wordnet 구축 방법 연구, 한국어 전체의 개념적 연결 가능성과 타당성 연구 등을 기대할 수 있을 것이다.

본 논문에서는 지식베이스를 구축하기 위한 방법인 지식표현(knowledge representation)[3] 중 의미망(Semantic Network)의 원리를 바탕으로 함과 동시에, 기존에 구축된 시소러스, 의미망, 온톨로지 등의 구축 방법을 비판·수용하여, 새로운 한국어 의미망 구축의 한 방향을 논의할 것이다. 그리고 자연언어처리의 의미 분석(Semantic Analysis) 분야에 적절하게 이용될 수 있고, 나아가 자연언어처리 전반에 도움을 줄 수 있는 의미망의 구성과 실제를 제시하고자 한다.

2. 기존의 지식베이스 검토

의미망를 비롯한 시소러스, 의미망 등과 같은 어휘들간의 문법적·의미적 상호 관계를 중심으로 한 지식베이스는 자연언어를 효과적으로 처리하기 위하여 많은 분야에서 구축·활용되고 있다. 하지만 이러한 지식베이스는 활용 분야에 따라 그 구축 방법론이 다르게 적용되어, 언어처리에 효과적이면서 통용될 수 있는 지식베이스는 실질적으로 구축되지 못한 실정이다. 그래서 일반적이고 효율적인 지식베이스를 구축하기 위해서 사전, 동의어 사전, 말뭉치, 워드넷을 유기적으로 통합하여 하나의 연속적인 "어휘(lexicon)-구(phrase)-연어(collocation)-관

3) '지식표현(knowledge representation)'은 전산학적으로 "지식베이스로의 저장을 위해 임의의 지식을 나타내는 사실과 관계성 등을 코드화하는 방법"이라 할 수 있고, 언어학적으로 "어떤 언어로 기록하는 것, 의사소통에 쓰이는 다른 개체로 기록하는 것, 우리 주위의 세계(World) 또는 그 세계의 상태를 기술하는 것, 또는 그림을 그리는 것과 같은 것"(서상욱, 1994),이라 할 수 있다. 결국 인간의 지식을 어떻게 하면 정확하고 정밀하게 표현할 수 있는가에 중점을 둔다는 점과 지식베이스 구축을 위해서는 반드시 '지식표현'과 같은 표현 방법의 정립화가 필요하는 점에서도 별다른 차이가 없다고 할 수 있다. 참고로 인공지능에서는 지식표현 방법으로 의미망 이외에도 생성규칙(prodution rule), 틀(frame), 술어논리(predicateive logic), 개념적 종속성(conceptual dependency), 스크립트(scripts) 등을 사용하고 있다.

용구(idiomatic expression)-말뭉치(corpus)"를 마련하거나, 세부적으로 품사 정보, 형태소 정보, 의미 관계, 구문 정보, 선택 제약 정보, 의미 정보 등을 긴밀하게 연결시킨 지식베이스 구축이 필요하다. 즉 이러한 지식베이스 구축은 특정 분야에 맞는 개별적 구축도 중요하지만 언어처리 전반에 통용될 수 있는 구축 방향도 모색되어야만 한다. 이 장에서는 시소러스, 의미망, 온톨로지에 대한 기본적인 개념을 간략히 살펴보고, 그들이 가지는 문제점을 지적하여 우리가 구축하고자 하는 한국어 의미망의 원칙을 마련해 볼 것이다.

2.1 시소러스, 의미망, 온톨로지

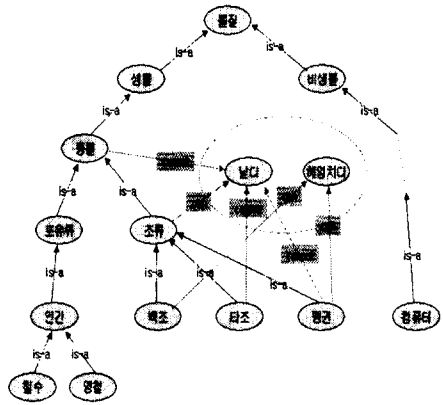
2.1.1 시소러스(Thesaurus)

시소러스는 그리스어에서 유래된 것으로 '지식의 보고(寶庫)'라는 뜻으로 사용되었는데, Peter Roget가 영어의 어휘를 내용상으로 분류하여 관련어(關聯語)를 표시한 사전을 만들어 '시소러스'라는 이름을 붙인 이래, 그러한 사전을 시소러스라고 일컫게 되었다. 단순한 형태의 시소러스는 (1) 주어진 지식 영역에서 미리 수집하여 구성한 중요한 단어들의 목록과 (2) 이 목록의 각 단어에 대해 관련성이 있는 단어의 집합으로 구성된다. 관련어는 여러 가지가 있지만 대부분은 주로 동의(또는 유의) 관계에서 도출된다(김명철 외, 2001:192~196). 일반적인 시소러스의 구성은 상하관계(BT/NT)를 중심으로 하여 동등관계(USE/UF), 부분·전체관계(BTP/NTP), 사례관계(Instance), 연관관계(RT) 등으로 이루어진다. 대표적인 예로 NASA 시소러스, Kadokawa 시소러스, Goi-Taikei[NTT], 신문기사종합 시소러스[한국언론연구원], 경제신문 시소러스[한국경제신문사], 국방과학기술 한글 시소러스[국방과학연구소], 과학기술용어 시소러스[시스템공학연구소], 중앙일보 시소러스[중앙일보사] 등이 있다[4].

4) 덧붙여 <21세기 세종 계획>의 전자사전 개발에서도 전산적 처리와 접목시킨 국어학적 어휘

2.1.2 의미망(Semantic Network)

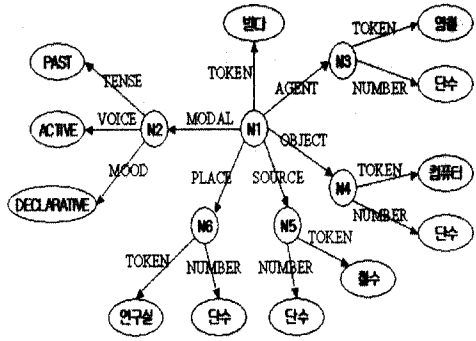
의미망은 형식적인 면에서는 시소러스와 비슷한 구조를 가지지만, 의미망을 구성하는 방법적인 면과 내용적인 면에서는 차이점을 가진다. 의미망은 한 어휘가 가지고 있는 다른 어휘들간의 관계를 망(network)이라는 형태로 나타낸 것으로 정의할 수 있다. 즉 단어의 의미, 개념간의 연상관계, 지식 등을 표현하는 네트워크를 의미한다. 이러한 의미망은 기본적으로 세 가지 특성을 지닌다[5]. 첫째, 개념들이 특정한 관계에 의해 연결되어 있는 지식에 대한 사고 방식을 보여 준다. 둘째, 노드(또는 박스)와 링크(또는 화살표), 관계 표시(또는 라벨) 등의 결합으로 이루어진 도식적 표현(diagrammatic representation)을 가진다. 셋째, 지식 표현을 관리할 수 있는 알고리듬을 이용한 다양한 추론 기술과 데이터베이스와 같은 역할을 고려하는 전산적 표현(computational representation)을 가진다. [그림 1]은 의미망에 의한 문장 표현과 개념 체계 표현의 예를 든 것인데[6], 가장 일반적인 의미망의 구성이라 할 수 있다. 대표적인 예로 WordNet, Euro-WordNet, Lexical FreeNet, 한국어 명사 개념망[ETRI], 명사 의미 계층 구조[울산대, KAIST] 등이 있다.



분류 방식을 채택한 의미 부류 체계를 구축 중인데, 이 의미 부류 체계는 시소러스에 속한다고 할 수 있다.

5) Hartley, R. T. & Barnden, J. A.(1997)에서는 의미망에 대한 개괄적인 설명을 확인할 수 있다.

6) 황도삼 · 최기선 · 김태석 공역(1999:146~147) 참고.



[그림 1] 의미망에 의한 문장 표현과 개념 체계 표현의 예

2.1.3 온톨로지(Ontology)

온톨로지는 본래 철학적인 개념으로 "존재의 본질을 연구하는 형이상학의 한 갈래(the branch of metaphysics that studies the nature of existence)"이나, 자연언어처리에서는 "실세계(혹은 특정 도메인)에 존재하는 모든 개념들(concepts)과 그 개념들의 속성(properties), 그리고 개념들이 상호간 의미적으로 어떻게 연결되어 있는가(semantic relation)에 대한 정보를 가지고 있는 지식베이스(knowledge base)"로 정의할 수 있다. 그리고 온톨로지는 언어 독립적인 정보만 저장하고 있어서 지식 공유와 재사용을 중요시한다는 점과, 개념간 의미관계가 계층관계(taxonomic relation), 격관계, 동의관계 외의 "has-member, material-of, represent"와 같은 다양한 의미관계도 포함하고 있다는 점에서 시소러스와 구별될 수 있다. 그래서 온톨로지는 의미망과 비슷한 원리와 구성을 가진다고 할 수 있다[7]. 이러한 온톨로지는 자연언어처리에서 지식 기반 기계번역 시

7) 실제로 '의미망'과 '온톨로지'에 대한 구별은 그리 중요하지 않은 듯하다. 다만 '온톨로지'라는 말이 철학에서 사용되었다는 말을 통해 어휘들의 개념적 양상이 의미망보다는 더 세부적이다. 단적인 예로, Enterprise Ontology, Mikrokosmos Ontology 등 현재 국내외에서 구축되고 있는 온톨로지의 경우 특정 분야(특히 기계번역)마다 개념관계를 달리 설정하는 경우가 많다. 이러한 특정 도메인별 온톨로지의 구축 원리와 내용들은 Uschold, M. 외(1997), 강신재(2002), Mahesh, K. & Nirenburgm, S(1996) 등을 참고하면 된다.

스템(Knowledge-based Machine Translation System)에 활용되거나 기업의 지식경영(Knowledge Management) 차원에서 지식을 효율적으로 정리하기 위해서 사용되는데, 이러한 온톨로지 구축의 활용적인 면을 살펴보면 아직까지 미흡한 부분이 많다. 온톨로지의 대표적인 예로 Mikrokosmos 온톨로지, LIP 온톨로지, EDR, SENSUS, HowNet, CYC 등이 있다.

2.2 지식베이스 구축상의 문제점

시소러스, 의미망, 온톨로지 등은 각각 지식베이스로서의 역할을 담당할 수 있다. 2.1절에서 간략하게 살펴보았던 지식베이스들은 연구실 수준의 실험적 데이터베이스 구축에서부터 대규모 데이터베이스 구축에 이르기까지 다양하게 연구·개발되고 있지만, 자연언어처리를 비롯한 정보검색, 기계번역 등과 같은 응용 분야에서 큰 효과를 얻지 못하는 실정이다. 이러한 현상은 시스템상의 문제보다는 지식베이스를 구축하는 방법론상의 문제에서 비롯되었다고 할 수 있다. 국내에서 구축된 지식베이스를 대상으로 구축상의 문제점과 어려운 점을 간략하게 기술하면 다음과 같다.

첫째, 특정 분야의 특성을 고려한 지식베이스가 많다는 점이다. 시소러스, 의미망, 온톨로지 등을 연구하고 구축하는 것이 대부분 정보검색이나 기계번역과 같은 특정 분야에 국한되어 진행되어 왔다는 점을 통해 알 수 있다. 기계번역에서는 대상언어와 목적언어의 대역성(translation)에 치중하여, WordNet이나 가도카와 시소러스 등과 같은 국외에서 구축된 지식베이스를 번역하거나 응용하는 경향이 많은 편이다. 그리고 정보검색에서는 사용자의 언어 사용과 학문 분야의 특수성을 고려한 시소러스를 구축하는 것이 일반적이다. 이러한 경향은 통합 지식베이스 구축이라는 측면에서 많은 문제를 야기시킨다. 즉 서로 다른 구성 원리와 방법을 이용해 구축된 지식베이스를 통합하는 것은 쉬운 작업이 아니며, 공통적으로 사용되는 단어(용어, 개념)들이나 전문용어들의 여러 의미(개념) 관계들

을 어떻게 통합하느냐 하는 문제는 계속 논의되어야 할 연구 과제이다.

둘째, 지식베이스들의 구성적인 면을 볼 때, 계층적인(hierarchical) 구조와 분류(부류)적인 구조가 혼합되어 사용되는 경우가 많아 일관된 구조체를 형성하지 못하고 있다. 시소러스에서 가장 기본적인 틀을 이루는 상하관계(BT/NT)의 경우 IS_A 관계의 모호성으로 인해 KIND_OF, PART_OF, USE_OF(FOR) 등이 혼용되는 경우가 많다. 이것은 엄밀하게 상하관계에 의한 계층적 구조라기보다는 사용자나 연구자들의 언어 사용의 습관적 지식에서 비롯된 의미 부류적인 구조로 보는 것이 타당할 것이다. 의미 계층 구조를 담고 있는 온톨로지나 의미망에서도 비슷한 경향을 보이고 있다[8]. 이러한 구조들은 언어의 사회적 성격을 모두 수용할 수 있는 지식베이스 구축이라는 큰 부담감을 가지는 요소로 작용할 수 있다. 즉 사람들의 언어 사용을 중심으로 어휘들의 개념적 양상을 네트워크화하는 것은 지식베이스 내부 구조가 일정한 기준 없이 구축될 가능성 높아진다.

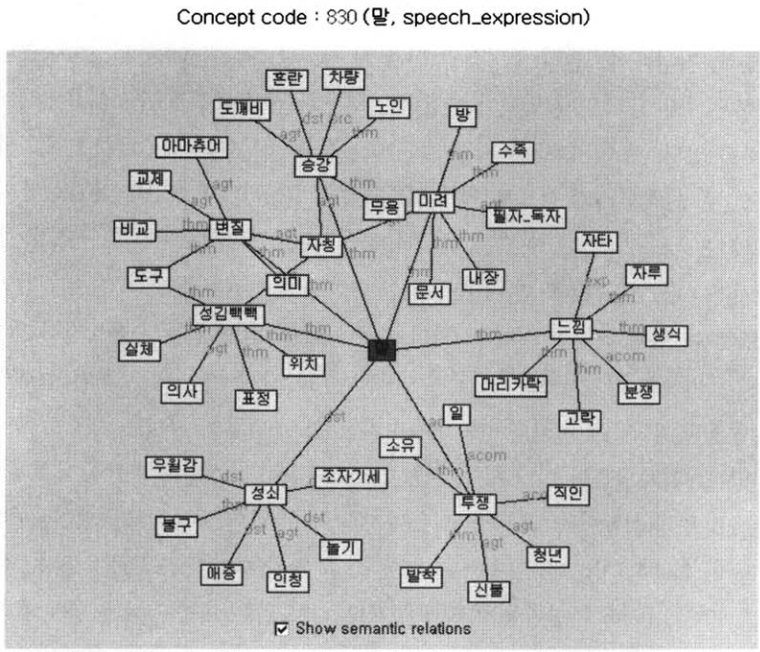
셋째, 기존의 지식베이스는 다의어에 대한 처리 방안을 고려하지 않아, 한 단어가 많은 의미적 부담감을 가짐과 동시에 너무 많은 개념 관계를 가지게 되어, 단어 중의성 해소(Word Sense Disambiguation)과 같은 자연언어의 의미 분석에서 그리 큰 효과를 보지 못하고 있다. 예를 들어 [표준국어대사전]을 기준으로 "말"은 12개의 동형이의어(homograph)로 나눠 있고, "말_1"은 다의어로서 11개의 의미를 가진다[9]. [그림 2]는 "말_1"을 격 관계(case relation)와 의미 관계(semantic relation)에 의해 지식베이스 형태로 표현한 것으로, 여러 의미를 가지는 한 단어가 너무 넓은 개념적 파생성을 가지고 있다는 것을 알 수 있다. 즉 지식베이스 구축상의 다의어 처리 문제는 중요한 논쟁거리가 될 수 있다[10].

---

8) 물론 의미망과 온톨로지는 단어들의 계층적인 구조를 중심으로 구축되지는 않는다. 즉 의미망과 온톨로지는 계층적 구조뿐만 아니라 특정한 의미 또는 개념 관계를 이용하여 단어(개념)들간의 연결 구조체를 만드는 것이 목적이라 할 수 있다.

9) 국어사전에서 동형이의어(동형어)는 "말[1]"과 같이 표제어의 우측 상단에 어깨번호가 붙어 있다. 여기서는 편의상 "말_1"로 표기하기로 한다.

10) <21세기 세종 계획>의 전자사전 개발 분과에서도 의미부류와 관련된 동형어와 다의어 처리 문제가 논의되고 있다.



[그림 2] "말(speech_expression)"의 지식베이스에서의 표현 : LIP 온톨로지[11]

넷째, 각종 사전을 기반으로 구축된 지식베이스의 경우, 사전이 가지고 있는 모든 정보를 이용하고 있지 않다. 국어사전에서는 표제어, 뜻풀이, 동의·유의관계, 형태소 정보, 구문 정보, 관련 정보 등을 추출할 수 있으며, 전문용어사전과 백과사전에서는 전문성을 가진 표제어와 뜻풀이, 관련 정보 등을 추출할 수 있다. 이러한 사전들은 언어처리에 적절한 데이터베이스로 제공되지 않는다는 단점이 있으나, 많은 정보를 추출할 수 있다는 점에서 중요한 지식베이스의 구축 자료로 활용될 수 있다. 하지만 대부분의 지식베이스 구축에서는 이러한 다양한

11) LIP 온톨로지는 가도카와(Kadokawa) 시소러스의 개념 부류와 계층 구조를 그대로 도입하고, 그 계층 구조에 추가적인 의미관계를 삽입하여 구조를 확장한 것이다. 추가적인 의미관계는 격관계와 기타 의미관계로 나눌 수 있다. 자세한 내용은 강신재(2002)를 참고하면 되고, http://bertha.postech.ac.kr/~sejong/ontology. htm에서는 LIP 온톨로지를 시각적으로 확인할 수 있다.

정보를 적극적으로 활용하지 못하고 있는 실정인데, 예를 들어, 지식베이스 구성 면에서 잘 논의되지 않았던 표제어의 경우, 계층적 구조를 유지하거나, 하위 개념을 묶을 수 있는 상위 개념을 마련하기 위하여 사전에 등재되어 있지 않은 '구체물, 추상물, 인공물, 추상적 관계' 등과 같은 새로운 단어를 만들어 이용하는 경우가 많다. 이것은 사전에 등재되어 있지 않은 단어들에 대한 명확한 개념 설정, 신어(新語) 사용 여부의 기준 설정 등의 문제가 야기될 수 있다[12].

이외에도 어휘들간의 관계(relation) 또는 노드(node)간의 연결(link) 문제, 지식베이스 구축 범위, 말뭉치에서의 정보 추출 방법 등의 문제점이 있다. 그리고 지식베이스 구축의 외부적인 문제로 수작업에 대한 부담감, 많은 시간과 연구자의 필요에 따른 구축상의 어려움 등을 들 수 있다.

3. 한국어 의미망(Korean Semantic-Network)의 구성과 구축 방법

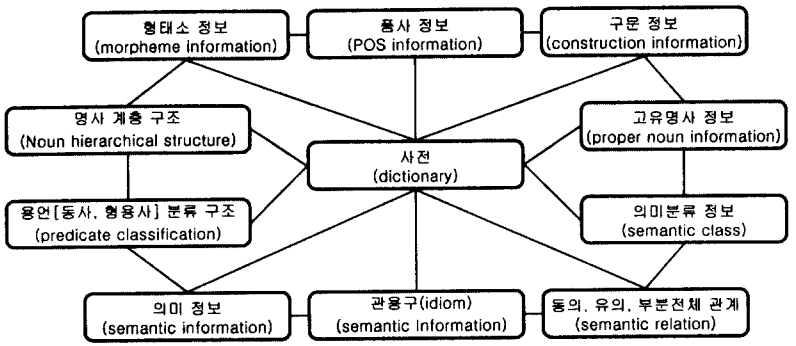
2장에서 언급되었던 지식베이스에 대한 간략한 설명과 문제점 지적을 기초로 하여, 3장에서는 텍스트 언어처리에서 적절하게 이용될 수 있고 나아가 언어처리 전반에 이용될 수 있는 의미망 구축의 한 방향을 제시한다. 이를 위해 우리가 구축하고 있는 한국어 의미망(이하 KSN)의 구성과 구축 방법을 세부적으로 설명할 것이다.

3.1 KSN의 전체 구성

KSN은 [그림 3]을 통해 알 수 있듯이 국어사전을 중심으로 하여 언어처리에

12) 계층적으로 보았을 때, 사전에 등재되지 않은 상위어를 사용하였을 경우에는 하위어(하위 개념)을 포괄적으로 수용할 수 있는 명확한 뜻풀이가 명시되어야 할 것이다. 이것이 정확하게 기술되지 않는다면, 의미의 상속(inheritance), 개념화(추상화) 양상 등의 의미론적 문제가 발생할 수 있을 것이다.

필요한 다양한 정보를 담는다. 여기에서 중요한 것은 모든 사전이 아래와 같은 정보를 모두 가지고 있는 것이 아니기 때문에, 규모가 큰 국어사전을 중심으로 하고, 중소형 국어사전을 부수적으로 활용한다. 그리고 명사 중심으로 기술되어 있는 전문용어사전이나 백과사전은 전문성이 필요한 의미분류 정보나 의미 정보를 보완할 때 사용한다.



[그림 3] KSN의 전체 구성도

이러한 구성은 자연언어처리에서 의미 분석에 필요한 요소들이다. 물론 의미 분석은 형태소 분석이나 구문 분석과 밀접한 관계를 가지고 있으므로, 자연언어처리에서 의미 분석 단계를 형태소 정보와 구문 정보와의 연계성을 고려하지 않으면 안 된다. 그래서 KSN은 의미 분석을 위한 최대한의 언어 정보들을 포함시키고자 하였다. KSN 구성에 대한 기본적인 사항을 정리하면 다음과 같다.

먼저 KSN에서 이용되는 언어 자료인 말뭉치는 국어사전의 뜻풀이, 용례 등을 활용하고, 세종 말뭉치를 부가적으로 활용한다. 이 말뭉치는 원시 말뭉치(raw corpus)와 품사태그 부착 말뭉치(POS tagged corpus), 의미태그 부착 말뭉치(Sense tagged corpus)로 구성되어 있다. 각각의 말뭉치는 각종 정보를 수집하는 데 사용될 뿐 아니라, KSN의 실험·평가에도 이용되고 있다.

다음으로 KSN의 기본축인 사전은 국어사전에 등재되어 있는 모든 단어를 대

상으로 한다[13]. 여기에서 중요한 점은 KSN의 사전은 국어사전의 표제어 개수보다 더 많은 단어를 가진다는 점이다. 표제어와 뜻풀이의 관계를 다의어와 같은 1:n(n은 뜻풀이 개수)의 관계가 아니라 단의어와 같이 1:1의 관계로 설정하기 때문이다. 이것은 2장에서 잠시 언급되었듯이, 한 단어가 여러 가지 의미를 가짐으로써 많은 구문적·의미적·개념적 부담을 가진다는 것은, 그만큼 지식베이스 구축상의 문제점으로 지적될 수 있다. 예를 들어, [그림 4]는 표준국어대사전에서의 "말_1"의 뜻풀이와 용례를 나타낸 것으로, "말_1"과 같은 다의어일 경우 11가지의 의미를 어떻게 하나의 구문적·의미적 관계로 연결시킬 것인가 하는 것은 지식베이스 구축의 문제가 된다.

말[1] [말:] 명
①사람의 생각이나 느낌 따위를 표현하고 전달하는 데 쓰는 음성 기호. 곧 사람의 생각이나 느낌 따위를 목구멍을 통하여 조직적으로 나타내는 소리를 가리킨다. ↔어사[11](語辭)①.
¶말과 글/말을 못하는 벙어리/말을 가르치다/말을 배우다/혀가 짧아져 있어서 말이 제대로 안 들린다.
②음성 기호로 생각이나 느낌을 표현하고 전달하는 행위. 또는 그런 결과물. ↔소리[1]②.
¶고운 말과 바른 말/말이 거칠다/말이 느리다/말이 빠르다/그들은 두 살 터울이 지는데도 말을 놓고 지내는 친구 같은 사이이다.
③일정한 주제나 줄거리를 가진 이야기.
¶말을 건네다/말을 꺼내다.
④단어, 구, 문장 따위를 통틀어 이르는 말.
¶적절한 말을 찾다/내 사전에 불가능이란 말은 없다./이번엔 '미안하다'는 말로는 용서가 안 된다.
⑤소문이나 풍문 따위를 이르는 말.
¶말이 퍼지다/널 두고 말이 많으니 조심해라./항간에 머잖아 물가가 폭등할 것이라는 말이 있다.
⑥ (''-으라는/-다는 말이다' 구성으로 쓰여) 다시 강조하거나 확인하는 뜻을 나타내는 말.
¶나보고 이런 것을 먹으란 말이냐?/가겠다는 말인지 안 가겠다는 말인지 알 수가 없다./제가 어제 과장님께 확실하게 보고를 드렸단 말입니다.
⑦ (''-으니/-기에 말이지' 구성으로 쓰여) '망정이지'의 뜻을 나타내는 말.
¶집에서 조금 일찍 나왔으니 말이지 하마터면 차를 놓칠 뻔했다./그가 있었기에 말이지 없었으면 큰 낭패를 보았을 거야.
⑧ (''-을 말이면', '-을 말로는', '-을 말로야' 구성으로 쓰여) '-을 것 같으면'의 뜻을 나타내는 말.
¶자네가 장가들 말이면 내게 미리 귀띔을 했어야지.
⑨ (''-어(아)야 말이지' 구성으로 쓰여) 어떤 행위가 잘 이루어지지 않음을 탄식하는 말.
¶차를 사고 싶은데 돈이 있어야 말이지./모를 내야 되는데 도대체가 비가 와야 말이지.
⑩ (주로 '말이나', '말이야' 꼴로 명사 뒤에 쓰여) 앞에서 언급한 사실을 강조하여 말하는 뜻을 나타내는 말.
¶돈이라니, 며칠 전에 네가 내게 준 돈 말이냐?/책 좀 빌려 줘. 네가 읽던 책 말이야./그 사람 아십니까? 정거장에서 인사한 사람 말입니다./추수를 해야겠다. 서리가 내리기 전에 말이야.
⑪ (주로 '말이야', '말이죠', '말이지', '말인데' 꼴로 쓰여) 어감을 고르게 할 때 쓰는 군말. 상대편의 주의를 끌거나 말을 다짐하는 뜻을 나타낸다.
¶그런데 말이야./하지만 말이죠./내가 말이지 어제 낚시를 갔는데 말이지./우리끼리라서 말인데.
{말 〈용가〉 1}

[그림 4] 표준국어대사전에서의 "말1"의 뜻풀이

13) 현재 방언, 북한어 등과 같은 특수 용어는 KSN에서 우선 제외시켰다. 이러한 용어들에 대한 처리 방안은 좀더 논의가 필요할 듯하다.

이러한 문제를 해결하기 위해서 KSN에서는 표제어와 뜻풀이의 관계를 1:1로 설정하여 한 단어가 가질 수 있는 구문적·의미적 처리의 부담감을 줄이고자 하였다[14].

3.2 KSN의 세부 구조

3.2.1 어휘 데이터베이스(사전)

KSN의 어휘 데이터베이스는 사전의 기술 방식에 따라 구성된다. 그리고 뜻풀이와 용례 등에는 품사 태그를 부착하여 다른 테이블과 필드를 구성하여 언어 분석이 용이하도록 구성한다. 아래의 [그림 5]와 [그림 6]은 KSN의 어휘 데이터베이스의 일부분이다.

ID	Entry	SemTag	Explain
421	차	차_1	병장기의 하나인 '작살'을 십팔기(十八技)에서 지목하여 부르는 이름.
422	차	차_2	(어근) '차하다'의 어근
423	차	차_3	(의존명사) (大) ① 『주로 한자어 수 뒤에 쓰여』 '번', '차례'의 뜻을 나타내는 말.
424	차	차_3	(의존명사) (大) ② 『'-던 차에', '-던 차이다' 구성으로 쓰여』 어떠한 일을 하던 기회나 순간.
425	차	차_3	(의존명사) (大) ③ [수학] 방정식 따위의 차수를 이르는 말.
426	차	차_4	(大) [음악] '체_7'의 잘못. '체'를 한자를 빌려서 쓴 말.
427	차	차_5	(대명사) (此) '이_5-1'를 분어적으로 이르는 말.
428	차	차_6	(中) ① 바퀴가 굴러서 나아가게 되어 있는, 사람이나 짐을 실어 옮기는 기계.
429	차	차_6	(中) ② 『수량을 나타내는 말 뒤에 쓰여』 화물을 '①'에 실어 그 분량을 세는 단위.
430	차	차_6	(中) ③ [운동/오락] '바'자를 새긴 장기짝.
431	차	차_7	(中) 우리나라 성(姓)의 하나.
432	차	차_8	(조) ① 둘 이상의 사물을 견주었을 때에, 서로 다르게 나타나는 수준이나 정도.
433	차	차_8	(조) ② [수학] 어떤 수나 식에서 다른 수나 식을 뺀 나머지.
434	차	차_9	(초) ① 자나무의 어린잎을 달이거나 우린 물.
435	차	차_9	(초) ② 식물의 잎이나 뿌리, 과실 따위를 달이거나 우리거나 하여 만든 마실 것을 통틀어 이르는 말.
436	차	차_9	(초) ③ [식물] =자나무.
437	차	차_10	(寂) '수 갈래 진 비녀의 하나.
438	차	차_11	(漸) [역사] =차차04(漸次).
439	차	차_12	(접사) 『'것'으로 시작되는 몇몇 명사 앞에 붙어』 '끈기가 있어 자진'의 뜻을 더하는 접두사.
440	차	차_13	(접사) (大) 『일부 명사 뒤에 붙어』 '목적'의 뜻을 더하는 접미사.

[그림 5] '차'에 대한 KSN의 어휘 데이터베이스 구성

14) 물론 사전마다 다의어일 경우, 뜻풀이의 개수에 차이가 난다. 예를 들어 '마음'은 뉴에이스 국어사전에서는 8개, 연세한국어사전에서는 3개, 표준국어대사전에서는 7개로 나누고 있다. 뜻풀이를 세부적으로 검토했을 때 비슷한 의미를 가지고 있다고 할 수 있으나, 통일된 뜻풀이는 존재하지 않는다. 그래서 이 문제는 사전 이용자의 입장에서든 편찬자의 입장에서든 정제된 뜻풀이 작업이 필요할 것으로 보인다. 여기서는 [표준국어대사전]을 기준으로 하여 KSN을 구성하고, 나머지 사전상의 뜻풀이 차이는 각 사전마다 다른 뜻풀이를 서로 뜻풀이별로 묶어줄 수 있는 방안을 하나의 연구 과제로 남기고자 한다.

EntID	ExpOrd	StcOrd	Exam	PosExam	SemPosExam
65283	1	1	제목까지 상기하	제목/NNG+까지/J	제목/NNG+까지/JKB 상기_12/NNG+하/XSV+다/EF
65286	1	1	상기도 감염	상기도/NNG 감	상기도/NNG 감염/NNG
65289	1	1	상갑의 미역	상갑/NNG+의/J	상갑/NNG+의/JKG 미역/NNG
65289	1	2	상갈로 치다	상갈/NNG+로/J	상갈/NNG+로/JKB 치/VV+다/EF
65292	1	1	소녀가 상끗 웃는	소녀/NNG+가/J	소녀_1/NNG+가/JKS 상끗/MAG 웃/VV+는다/EF
65299	1	1	상납전	상납전/NNG	상납전/NNG
65299	2	1	뇌물조로 많은 돈	뇌물/NNG+조/N	뇌물/NNG+조/NNB+로/JKB 많/VA+은/ETM 돈/NNG+을/JKO 상납_2/NNG+하
65300	2	1	상납금을 바치다	상납금/NNG+을/	상납금/NNG+을/JKO 바치/VV+다/EF
65303	1	1	상냥스러운 소녀	상냥스럽/VA+ㄴ/	상냥스럽/VA+ㄴ/ETM 소녀_1/NNG
65304	1	1	상냥한 목소리	상냥하/VA+ㄴ/E	상냥하/VA+ㄴ/ETM 목소리/NNG
65304	1	2	그녀는 누구에게	그녀/NP+는/JX	그녀/NP+는/JX 누구/NP+에게/JKB+이나/JKB 상냥하/VA+다/EF
65309	1	1	같은 상념에 잠기	같/VA+은/ETM	같/VA+은/ETM 상념/NNG+에/JKB 잠기/VV+다/EF
65316	1	1	마을의 상놈은이	마을/NNG+의/J	마을/NNG+의/JKG 상_15/NNG+놈은이/NNG
65317	1	1	상다리가 휘어지	상/NNG+다리/N	상_2/NNG+다리_1/NNG+가/JKS 휘어지/VV+게/EC 음식/NNG+을/JKO 차리/V
65318	1	1	45페이지 상단 들	45/SN+페이지/N	45/SN+페이지/NNB 상단_1/NNG 들춰/NR 줄/NNG
65318	2	1	침대차의 상단	침대차/NNG+의/	침대차/NNG+의/JKG 상단_1/NNG
65325	1	1	하의 상달	하의/NNG 상달/	하의/NNG 상달_2/NNG
65327	1	1	전화 상담	전화/NNG 상담/	전화/NNG 상담_1/NNG
65327	1	2	상담에 응하다	상담/NNG+에/J	상담_1/NNG+에/JKB 응하/VV+다/EF
65327	1	3	변호사와 이혼 문	변호사/NNG+와/	변호사/NNG+와/JKB 이혼/NNG 문제/NNG+를/JKO 상담_1/NNG+하/XSV+다/I
65329	1	1	외국 바이어와 상	외국/NNG 바이어	외국_2/NNG 바이어/NNG+와/JKB 상담_3/NNG+을/JKO 나누/VV+다/EF
65333	1	1	법률 상담소	법률/NNG 상담/	법률/NNG 상담소/NNG
65338	1	1	지위에 상당하는	지위/NNG+에/J	지위/NNG+에/JKB 상당/NNG+하/XSV+는/ETM 수입/NNG
65338	2	1	10만 원 상당의	10/SN+만/JKB	10/SN+만/JKB 원/NNB 상당/NNG+의/JKG 시계/NNG
65338	3	1	상당한 수준	상당/NNG+하/X	상당/NNG+하/XSV+ㄴ/ETM 수준/NNG
65338	3	2	상당한 실력	상당/NNG+하/X	상당/NNG+하/XSV+ㄴ/ETM 실력/NNG
65339	1	1	학생의 상당수가	학생/NNG+의/J	학생/NNG+의/JKG 상당수/NNG+가/JKS 안경/NNG+를/JKO 쓰/VV+었/EP+다/
65340	1	1	잃어 버린 돈이 상	잃/VV+어/EC 버	잃/VV+어/EC 버리/VX+ㄴ/ETM 돈/NNG+이/JKS 상당액/NNG+에/JKB 달하/V
65340	2	1	월급 상당액를 돌	월급/NNG 상당	월급/NNG 상당액를/NNG 돌통/NNG+으로/JKB 지급/NNG+하/XSV+다/EF
65344	1	1	상대로부터 전해	상대/NNG+로/J	상대_1/NNG+로/JKB+부터/JKB 전하/VV+어/EC 오/VX+는/ETM 가보/NNG
65345	1	1	말 상대	말/NNG 상대/N	말/NNG 상대_2/NNG
65345	1	2	그런 사람은 상대	그러하/VA+ㄴ/E	그러하/VA+ㄴ/ETM 사람/NNG+은/JX 상대_2/NNG+하/XSV+지/EC 말/VX+아/
65345	2	1	그는 만만찮은 상	그/NP+는/JX 만	그/NP+는/JX 만만찮/VA+은/ETM 상대_2/NNG+이/VCP+다/EF

[그림 6] 사전 용례에 품사·의미태그를 부착한 데이터베이스

[그림 5]와 같이 표제어를 단의어 형식으로 구성함으로써 한 단어가 가지고 있는 구문적·의미적 부담을 최소화할 수 있을 뿐 아니라, 지식베이스 구축에서는 보다 폭넓은 의미적 상관성을 고려할 수 있다. 또한 [그림 6]과 같은 사전 데이터베이스 구성은 사전적 기능을 담당함과 동시에 원시 말뭉치, 품사태그 부착 말뭉치, 의미태그 부착 말뭉치의 역할까지 담당할 수 있다. 사전의 뜻풀이나 용례 등을 하나의 말뭉치로 활용함으로써 얻을 수 있는 큰 장점은, 다른 말뭉치에 비해 문학적 표현(비유적 표현, 시적 표현 등), 생략형 문장 등과 같은 정제되지 않은 표현들이 적어서 언어학적으로 정제된 문장을 대상으로 언어를 분석할 수 있다는 점이다. 물론 다양한 표현이 없다는 것도 문제가 될 수 있으나, 자연언어처러에서 분석의 오류로 발생하는 대부분이 정제되지 않은 표현들임을 감안한다면, 일반 말뭉치보다는 더 효과적인 분석 대상이 될 수 있을 것이다.

3.2.2 형태소·품사 정보와 구문 정보

KSN에서는 '명사, 대명사, 수사, 동사, 형용사, 부사, 관형사, 감탄사, 조사' 등의 품사 정보와 '어근, 어미, 접사' 등의 형태소 정보 등을 포함시키고 있다. 이러한 정보들은 사전에서 기본적으로 제공하는 품사 표시를 그대로 이용하고, '조사, 어근, 어미, 접사' 등은 뜻풀이에서 밝히고 있는 세부적인 분류를 참고로 이용한다. 또한 품사 태깅 시스템이나 형태소 분석 시스템과의 연계성을 고려하여 품사 중의성 문제에서도 KSN이 어느 정도의 역할을 담당할 수 있도록 구성한다.

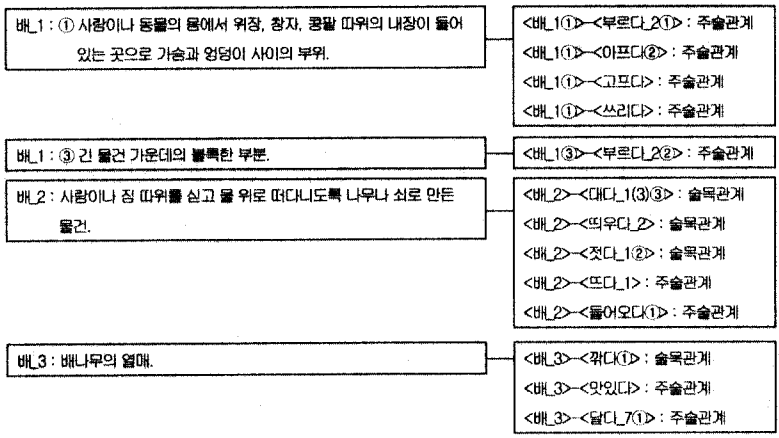
구문 정보는 기존의 격틀(case frame) 구조와 비슷하지만 한국어의 문장성분(sentence component)에 의해 주어-서술어 관계, 목적어-서술어 관계 등과 같이 문장성분간의 관계를 중심으로 연결된 정보를 핵심으로 한다는 점에서 격틀 구조와는 다르다고 할 수 있다. 또한 <21세기 세종 계획> 전자사전 분과에서 개념적 기준과 통사적 기준을 모두 고려하여 구축 중인 명사 의미 부류 체계에서의 논항 구조와 적정 술어를 이용한 구문 정보와도 다르다고 할 수 있다[15]. 즉 KSN에서 주로 다루고 있는 구문 정보는 특정 명사와 특정 용언간의 직접적인 구문 관계를 설정하는 것이다. 한국어의 문장성분 중 필수 성분 즉, 주어, 목적어, 보어, 서술어, 필수적 부사어 등은 문장에서 구문적으로나 의미적으로 핵심 요소에 속한다. 이런 점을 적극 활용하여, 서술어 중심으로 한 주어, 목적어, 필수적 부사어를 각각 연결시켜 구문적·의미적 결정성(decision)을 확보하는 것이다.

서술어(용언)를 중심으로 한 통사적인 제약인 '하위범주화 제약'과 의미적인 제약인 '선택 제약'을 적절히 고려하여, 실질적으로 [그림 7]과 같이 명사와 용언을 직접적으로 연결시킴으로써, 언어 분석상 구문적·의미적 결정성을 확보할 수 있다는 것이다. 이러한 관계 설정은 용언이 명사를 수식하는 관형어의 역할을 담당하더라도 그 관계는 변하지 않는다는 점도 고려한 것이다[16].

---

15) 강범모 외(2001), 송근영 외(2001), 홍재성 외(2001) 등 참고.

16) 이왕우·최호섭·옥철영(2002)에서는 이러한 구문 패턴을 바탕으로 동형이의어 분별에 필요



[그림 7] '배'의 구문 정보 일부

3.2.3 명사 계층 구조와 의미 분류 정보

명사 계층 구조와 의미 분류 정보는 지식베이스의 중점적인 구축 사항이다. 2장에서도 언급했듯이, 계층적(hierarchical) 구조를 이루는 지식베이스는 일반적으로 IS_A 관계를 이용하는데, 특히 시소러스 구축에서는 상위어(Broader Term; BT)와 하위어(Narrower Term; NT)를 구성할 때, IS_A 관계를 사용하는 것을 하나의 기준으로 사용하고 있다. 하지만 IS_A 관계는 의미적으로 상당한 모호성을 가지고 있다. IS_A 관계는 표현상으로 "(무엇)은 (무엇)이다."의 형태를 취하는 것으로, 의미적 유사성에 근거한 문학에서의 은유적 표현을 비롯하여, 함축적인 표현, 상하관계(hyponymy)를 나타내거나 한 단어의 문화적, 사회적 쓰임을 나타내기도 한다.

(1) 자동차는 교통(운송)수단이다.
(2) 황인종{흑인종, 백인종}은 사람이다.

한 의미정보를 획득하고자 하였다.

(3) 칼은 무기다.
(4) 시계는 선물이다.

위와 같은 문장은 모두 우리가 평상시에 자주 쓰는 표현이다. 하지만, '교통수단, 사람, 무기, 선물'이 엄격한 의미에서 상위어가 되는 것은 아니다. 즉 이러한 것은 쓰임새, 형태, 속성 등과 같은 특정한 기준에 의해 나누어진 분류적인 체계에 속한다고 할 수 있다. 예를 들어 (2)의 '황인종, 흑인종, 백인종'을 '사람'의 하위어로 생각할 수 있다. 그렇다면 '인종(人種)'도 '사람'의 하위어인가를 살펴야 한다. 우리는 '황인종, 흑인종, 백인종'의 쓰임을 통해 이 단어들을 '사람'처럼 인식한다. 그러나 '인종'의 한자 '人種'과 뜻풀이 "인류를 지역과 신체적 특성에 따라 구분한 종류" 등을 통해 '황인종, 흑인종, 백인종'이 '사람'의 하위어가 아닌 '인종'의 하위어인 것을 알 수 잇다. 또한 '인종'은 '사람'의 하위어가 아니라 '종류'의 하위어라는 것을 알 수 있다. 이와 같은 맥락에서 '자동차', '칼', '시계' 등의 상위어들도 달리 설정할 수 있을 것이다.

그렇다면 한 단어의 근본적인 의미를 바탕으로 한 상하관계에 의한 명사 계층 구조가 형성되어야 할 것이다[17]. KSN에서의 명사 계층 구조는 명사 자체가 가지고 있는 본래의 의미를 중심으로 상하관계를 형성하며 동의관계, 유의관계, 부분·전체관계와 같은 의미 관계(semantic relation)가 설정되어 있다. 의미 분류 정보는 문맥적 상황에서의 명사의 쓰임새를 중심으로 명사 계층 구조와 연결된다[18]. 앞에서 밝혔듯이 KSN은 텍스트 언어처리에 효과적인 의미망을 구축하는

17) 최호섭(2001)에서는 한국어 의미망 구축과 비슷한 맥락에서 이해할 수 있는, 경제용어를 중심으로 한 '한국어 명사 개념망'에 대한 구축 방법론을 제시하고 있다. 특히 상하관계에 대해 표제어의 형태적 특성, 뜻풀이 등을 이용하여 나름대로의 기준을 설정하고 있다.

18) 사전편찬학적 입장에서는 일반 사람들의 언어 사용을 고려한 의미 부류(분류) 체계가 필요하다. 하지만 이러한 의미 부류 체계를 자연언어처리나 정보검색, 기계번역과 같은 응용 분야에서 이용하기 위해서는 어느 정도의 수정이 필요할 듯하다. 정보검색에 이용할 경우에는 시소러스가 가지고 있어야 할 다양한 의미 관계를 가지고 있지 않다는 점을 보완해야 할 것이다. 기계번역의 경우에는 대상언어와 목표언어 사이의 대역성을 의미 부류 체계가 어떻게 담당해야 하는지에 대한 구체적인 논의가 있어야 할 것이다.

만큼, 어떠한 문장의 표면 구조(surface structure)를 중심으로 단어의 계열 관계(paradigmatic relation)와 통합 관계(syntagmatic relation)를 파악하는 것은 중요한 작업 중의 하나이다.

(5) [장치 / 기계 / 차 / 자동차] 을/를 타다.

(6) [기관? / 교통기관?] 을/를 타다.

(7) [수단* / 교통수단*] 을/를 타다.

(5)는 말뭉치에서 쉽게 추출되는 문장이고, (6)은 사전 뜻풀이·용례 말뭉치에 추출되지는 않으나 드물게 사용되는 문장이며, (7)은 사용되지 않는 문장이다. 이러한 단어의 계열·통합 관계를 이용하는 것은 명사 계층 구조와 의미 분류 정보를 구축하는 데 적절하게 이용될 수 있다. 즉 사전의 뜻풀이를 이용한 상하관계 설정[19], 어휘의미론에서의 상하관계 설정뿐만 아니라 단어의 계열·통합관계에 의한 상하관계 설정까지 고려할 수 있는 것이다. 이러한 관계를 좀더 이용하면 "자동차는 (교통수단, 상품)으로 이용된다."와 같은 '자동차'의 문맥적 상황을 통해 '자동차'가 '교통수단'의 분류에 속하고, '상품'의 분류에도 속할 수 있다는 의미 분류 정보를 얻을 수 있다[20].

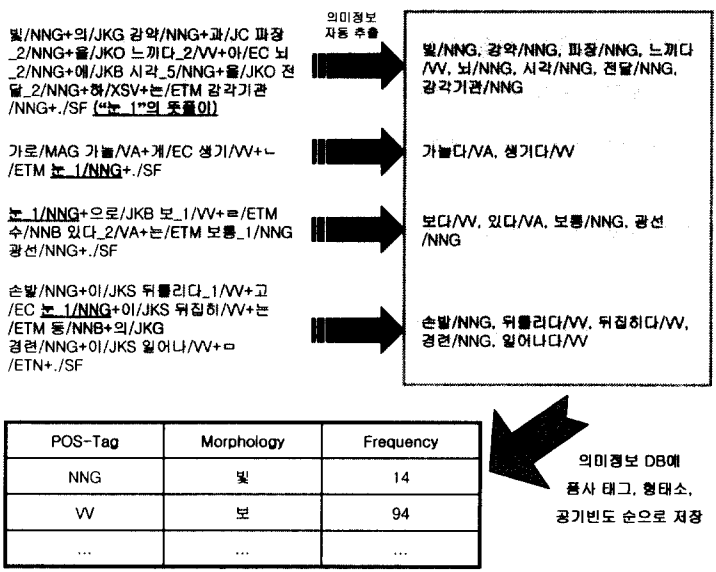
3.2.4 의미 정보

의미 정보는 ① 문장 내에서 한 단어의 의미를 이해하는 데 중요한 역할을 하

19) 한영균(1994)에서도 사전의 뜻풀이를 이용한 의미망 구축 방법을 제시함과 동시에, 사전 뜻풀이의 문제점을 지적하고 있다. 사전의 뜻풀이 문제는 계속적인 논의가 필요할 것으로 보인다.

20) 네이버, 엠파스, 야후 등과 같은 검색엔진에서 제공하고 있는 카테고리(category) 정보를 통해서도 실생활에서 언어들의 쓰임새를 범주별로 정리하여 검토할 수 있을 것이다. 이러한 정보는 의미 부류 체계를 고려할 때 유용한 정보로 활용될 수 있다.

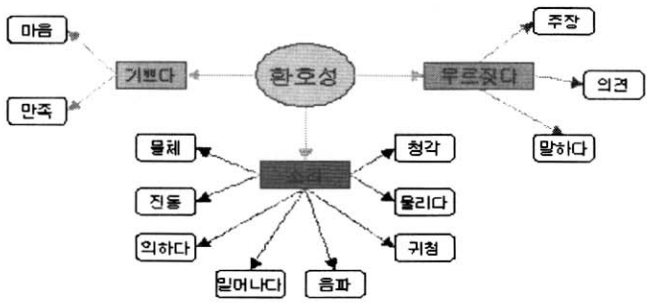
는 단어, ② 국어 사전에 등재되어 있는 표제어를 이해할 수 있도록 하는 뜻풀이 내의 단어, ③ 문서나 문장에서 특정 단어와 함께 출현하는 빈도가 높은 단어(실마리 단어; clue-word) 등으로 구성된다. 앞에서 제시되었던 구문 정보를 의미 정보로 활용할 수도 있다. 이것은 공기(co-occurrence) 정보를 이용하는 의미 정보보다도 언어학적으로나 전산학적으로 도움을 줄 수 있다. 더나아가 이왕우 외(2002)에서 제시되었던 순환 뜻풀이망(recursive definition network)을 이용하여 의미 정보를 확장하여 자동으로 추출할 수도 있다. 이러한 의미 정보는 단어 중의성 해소(WSD)뿐만 아니라 정보검색, 기계번역 등에 널리 이용될 수 있는 것으로, 언어 자원(language resource)의 효율적인 관리 차원에서도 언어학적으로나 전산학적으로 의미가 있을 것이다[21].



[그림 8] 자동으로 추출한 의미 정보의 예

21) 울산대에서 구축 중인 WSD와 정보검색에 이용될 '의미 정보 DB'는 그 대표적인 예라고 할 수 있다.

환호성	기뻐서 부르짖는 소리.
기쁘다	(마음이) 만족스럽다.
부르짖다	어떤 주장이나 의견을 열렬히 말하다.
소리	물체의 진동에 의하며 일어나는 음파가 귀청을 울리어 일어나는 청각.



[그림 9] 순환 뜻풀이망(RND)의 일부분

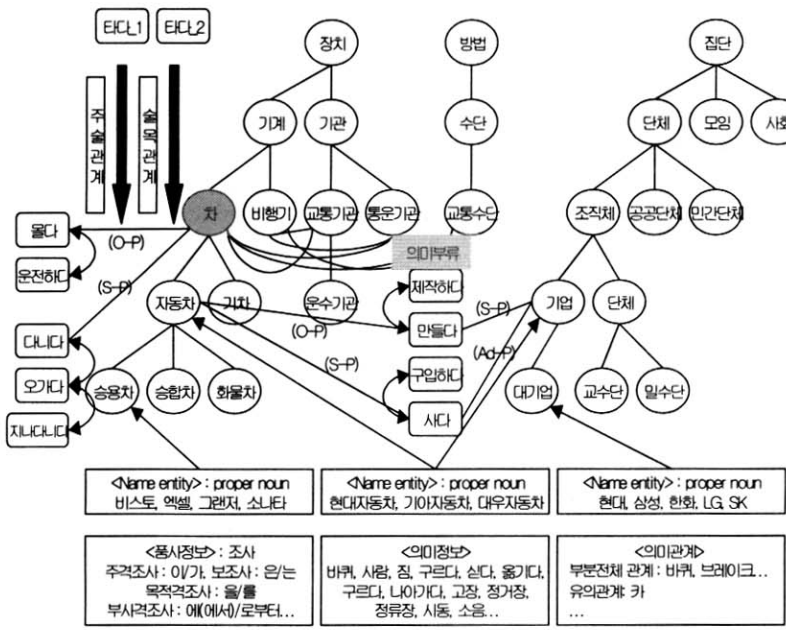
3.2.5 기타

위에서 언급되었던 정보 이외에도 고유명사 정보, 관용구 정보 등이 포함된다. 고유명사 정보는 인명, 회사명, 상품명 등의 고유명사를 일반명사와 분리시켜 개별적으로 관리함과 동시에 일반명사와의 관계를 설정하도록 한다. 물론 고유명사들을 단순한 명사의 나열로 데이터베이스를 구축할 경우에는 아무런 의미가 없을 것이다. 또한 불필요한 단어가 추가됨으로써 발생되는 데이터베이스의 증가로 인한 시스템의 속도 저하 문제를 최대한 줄일 필요가 있다. 그래서 KSN에서는 말뭉치에서 추출되는 고유명사를 우선적 처리 대상으로 삼는다. 즉 말뭉치에 존재하지 않는 고유명사일 경우에는 우선 KSN의 고유명사 정보에 포함시키지 않는다.

관용구(연어) 정보는 사전을 적극 활용하여 KSN에 포함시킨다. 관용구에 대해서는 학계에서 계속 논의되고 있는 실정이며, 자연언어처리에서도 이러한 관용구 처리뿐만 아니라 '수요의 법칙, 공급의 법칙' 등과 같은 전문용어의 처리 방안을 마련하고자 연구가 진행되고 있다. KSN에는 현재 사전에 등재되어 있는

관용구를 해당 표제어에 링크(link)시켜 놓고 있다.

이상과 같이 KSNdml 구성과 구축 방법을 간략하게 살펴보았다. KSN은 자동으로 구축하기 어려운 점이 많기 때문에, 수작업과 자동 추출 작업을 병행하여 구축되고 있다. KSN을 구성하고 있는 의미 정보, 구문 정보 등의 각종 정보들의 대부분은 자동으로 추출되고 있으며, 명사 계층 구조와 의미 분류 정보는 의미적인 분석과 패턴 분석을 위해 수작업이 병행되고 있다. [그림 10]은 '차'와 관련된 KSN의 일부 구조를 나타낸 것이다.



[그림 10] '차_6①'과 관련된 KSN의 일부 구조

4. KSN의 정보검색시스템과 단어 중의성 해소에서의 활용

4.1 정보검색시스템에서의 활용

최근의 정보검색 연구 동향은, 단순한 명사 중심의 색인 및 검색을 하였던 '키워드(keyword) 기반 정보검색'에서 자연언어처리 기술의 발달로 자연언어처리 기술을 바탕으로 한 '대화형·질의/응답형 정보검색'으로 발전하고 있다[22]. 현재 대부분의 상용 정보검색 시스템들에서 사용되는 기술은 형태소 분석이나 n-gram 방식으로 단순히 명사 키워드에 기반하여 색하고 검색하기 때문에 문장의 내용적인 의미를 제대로 색인에 반영하지 못하여 검색이 부정확하게 되는 결과를 낳고 있다(장명길 외, 2001:7). 이러한 점을 해결하기 위하여 '자동차를 구입할 수 있는 곳', '훈민정음은 언제 만들어졌는가'와 같은 일상적인 자연언어 질의를 통해 사용자에게 정확한 문서를 제공하는 기술을 연구하는 것이 최근의 동향이라 할 수 있다.

정보검색과 관련된 지식베이스 측면에서도, 색인어 중심의 시소러스 구축에서 자연언어 질의를 분석하기 위한 폭넓은 지식베이스 구축으로 변화하고 있다. 물론 이 지식베이스는 시소러스의 역할을 담당할 수 있도록 구축하는 통합형 시소러스 형태로 생각할 수도 있다. 즉 사용자의 질의 패턴을 분석하고 웹 문서들이 가지고 있는 속성(attribute)[23]을 부여하기 위해서는 기존의 시소러스를 대폭 수

22) 질의/응답형 정보검색과 관련된 연구는 ETRI의 "의미기반 정보검색"이 대표적인 예가 될 수 있다. 자세한 내용은 장명길 외(2001) 참고.

23) 여기서 말하는 속성이란, (1) 한 단어(개념어)가 가지고 있는 실생활에서의 활용 특성과 (2) 특정한 단어가 포함된 웹 문서의 다른 단어와의 관계 등으로 정의할 수 있다. 예를 들어, '불공정거래'의 속성은 '정의, 종류, 현황, 대책' 등이 될 수 있는데, 이것은 사용자가 원하는 정보를 하나의 속성으로 정의한 것이라 할 수 있다. 참고로 이러한 속성은 공통 속성과 개별 속성으로 나뉘는데, 공통 속성의 경우 명사 계층 구조에서 상하위어가 모두 공통적으로 가지고 있는 속성을 가리킨다. 이에 대한 자세한 논의는 장명길 외(2001), 장문수 외(2000), 장문수 외(2001) 등을 참고.

정할 필요가 있는 것이다.

'질의/응답형 정보검색'에서의 KSN 활용성을 예를 들어 설명하면 다음과 같다. '자동차'의 경우, 용언이나 다른 명사와 함께 쓰여 교통수단(운송수단), 교통기관, 상품(제품) 등과 같은 다양한 쓰임을 가지고 있기 때문에, 기존의 시소러스에서는 이러한 점을 감안하여 상하관계를 설정하고 있는 경우가 많다. 이러한 시소러스의 계층 구조는 바로 위 층위의 상위어만 이용 가능한 경우가 많고, 공통된 속성을 형성하기 어려울 뿐만 아니라 더 나아가 자연언어처리 기술을 이용한 정보검색에서도 큰 효율성이 없다. 하지만 KSN을 이용한다면, '자동차'의 본래 의미를 이용한 계층 구조, 즉 '자동차-차-기계-장치'라는 상하관계를 통해 "자동차(차;기계;장치) 구입"과 같은 사용자의 요구 사항 인식과 '정의', '종류', '뉴스', '상품' 등과 같은 공통 속성의 상하위어 적용 여부를 파악할 수 있다. 또한 KSN이 담고 있는 의미 분류 정보나 의미 정보 등의 다양한 정보를 이용하여 시소러스의 역할뿐만 아니라 검색 정확률 향상에도 기여할 수 있다. 즉 KSN은 사용자가 원하는 정답문서를 보여주는 정보검색 기술 처리 과정의 일부분을 담당할 수 있는 것이다.

4.2 KSN을 이용한 단어 중의성 해소(WSD)

WSD 방법 중 단어의 계층적 구조를 이용한 방법이 있는데, 이것은 시소러스, 의미망 등의 상하관계, 동의·유의관계로 설정된 단어나 노드간의 유사도, 거리 측정을 통해 단어 중의성을 해소하는 방법이다. Xiaobin(1995)의 WordNet을 이용한 WSD, 강신재(2002)의 LIP 온톨로지를 이용한 WSD 등이 대표적이라 할 수 있다. 하지만 국내에서는 동형이의어 분별과 의미에 의한 분별을 명확히 구분하지 않아 기술적 어려움을 겪고 있는 실정이다. 이러한 어려움은 시소러스나 의미망, 온톨로지 구축상의 문제이기도 하다. 즉 단의어와 다의어 둘 다 하나의 표제어를 가지기 때문에, 다의어인 표제어는 많은 의미적 부담감을 가지게 된다.

이러한 문제를 해결하기 위해서 최근 다의어 처리에 대한 논의가 진행되고 있으나 아직까지 실험적인 수준에 그치고 있다.

3장에서 잠시 언급했듯이 KSN은 표제어와 뜻풀이를 1:n이 아니라 1:1의 관계로 다루어, 한 다의어 내의 중의성 해소뿐만 아니라 동형이의어의 중의성 해소 방안을 고려하고 구축되고 있다. 또한 KSN은 기존의 시소러스, 의미망 등과 다르게, 단어 본래의 의미를 중심으로 하여 한 단어의 문장 속에서의 계열·통합 관계를 적절히 파악할 수 있도록 구축되고 있는 만큼 WSD 분석에 적절하게 이용될 수 있다. 즉 KSN은 단어의 계층적 구조를 이용한 WSD에서 단어간의 거리나 유사도, 노드간의 거리나 유사도 측정에 있어 일관성을 부여할 수 있는 구조를 가진 것이다. 또한 의미 정보, 구문 정보, 품사 정보 등과 같은 다양한 정보를 가지고 있으므로, 명사일 경우 명사 계층 구조를 통해 해결하지 못한 동형이의어 분별을 다른 추가적인 정보를 통해서도 해결할 수 있는 것이다[24].

5. 결론

본 논문에서는 기존에 구축되었던 시소러스, 의미망, 온톨로지 등과 같은 지식베이스 구축과 관련된 문제점을 지적하고, 우리가 구축 중인 KSN을 통해 한국어를 대상으로 한 의미망 구축의 한 방향을 제시하고자 하였다. 언어 자원의 체계적인 정리뿐만 아니라 자연언어처리와 관련된 분야에서의 적절한 활용을 고려한 지식베이스 구축은 전산언어학이나 자연언어처리 관련 연구자들이 심각하게 고민해야 하는 연구 과제 중 하나이다. 이러한 의미에서 KSN 구축은 장시간이 요구되는 작업이지만, 언어처리에 필요한 언어 자원이자 지식베이스 확보라는 측면에서는 반드시 필요한 작업이기도 하다.

---

24) KSN의 전체적인 구성과 구축 방법을 설명하는 만큼, KSN을 이용한 WSD에 대한 세부적인 내용은 본 논문에 담지 않는다.

현재 KSN 구축 작업은 4만 여 단어가 구축되었으며, 수작업의 부담감을 줄이기 위해 1997년 자동으로 구축되었던 울산대의 명사 의미 계층 자동 구축 기술을 수정·보완하고 있다. 또한 KSN의 각종 정보는 사전 말뭉치, 세종 말뭉치를 통해 자동 수집·분류하고 있다. 앞으로 형태소·구문 분석과의 상호 연계성 문제, 복잡한 문장 구성에서의 KSN 활용 문제, 명사 계층 구조의 자동 구축 방법 모색 등은 앞으로 해결해야 할 전산학적 과제이며, 명사 계층 구조를 형성하기 위한 서술성 명사(predicative noun)와 같은 명사류의 구조 분석, 어휘의미론적 의미 관계에 대한 세부적인 기준 마련 등은 언어학적 과제로 남는다.

국내에서도 잘 알려진 프린스턴 대학의 WordNet이 10여 년 동안의 긴 연구와 실험으로 어휘 데이터베이스를 구축한 사례를 통해 알 수 있듯이, 지식베이스는 단시간 내에 구축될 수 있는 언어 자원이 아니다. 국내에서도 이러한 연구 사례를 바탕으로 언어 자원 구축에 대한 연구가 활성화시킬 필요가 있을 것이다.

참고문헌

강범모 외. 2001. "한국어 명사 의미 부류 체계의 구축과 활용." 「제13회 한글 및 한국어 정보처리 학술대회 논문집」(한국정보과학회).

강신재. 2002. "실용적인 온톨로지의 반자동 구축 및 어휘 의미 중의성 해소를 위한 응용." 포항공대 박사학위논문.

강정미. 1999. "전문용어사전 표제어 기술 형식에 대한 연구." 「제6회 한국정보관리학회 학술대회 논문집」(한국정보관리학회).

고창수. 1999. 「한국어와 인공지능」 태학사.

김광해. 1990. "어휘소간의 의미 관계에 대한 재검토." 「국어학」(국어학회) 20.

김광해. 2000. "국어 정보화를 위한 어휘 연구의 방향." 「제1회 국어연구소 전국학술대회 자료집」(고려대 민족문화연구원 국어연구소).

김명철 외. 2001. 「최신정보검색론」 홍릉과학출판사.

김영택. 2001. 「자연언어처리」 생능출판사.

문유진. 1996. "의미론적 어휘개념에 기반한 한국어 명사 WordNet의 설계와 구축." 서울대 박사학위논문.

배해수 · 최덕수 · 김광해. 1998. "한국학 정보 처리를 위한 학술용 시소러스 연구." 「한국어전산학」(한국어전산학회) 2.

서상규 · 한영균. 1999. 「국어정보학 입문」 태학사.

송근영 · 홍재성. 2001. "한국어 '크기' 명사 부류에 대하여." 「제13회 한글 및 한국어 정보처리 학술대회 논문집」(한국정보과학회).

오길록 · 최기선 · 박세영. 1994. 「한글공학」 대영사.

옥철영. 1998. "우리말 개념망 명사 데이터 구축." ETRI 최종연구보고서.

옥철영. 2000. "우리말 개념망 구축." ETRI 지식정보검색연구팀 기술세미나 자료.

이왕우 · 최호섭 · 옥철영. 2002. "구문 패턴과 순환 뜻풀이망을 이용한 동형이의어 분별." 「제29회 한국정보과학회 춘계학술발표회 논문집(B)」(한국정보과학회).

이정민 외. 2000. 「의미구조의 표상과 실현」(한림과학원 총서 74) 도서출판 소화.

장명길 외. 2001. "의미기반 정보검색." 「정보과학회지」(한국정보과학회) 19-10.

장문수 외. 2000. "인터넷 질의/응답을 위한 지식베이스 구축." 「제12회 한글 및 한

국어 정보처리 학술대회 논문집」(한국정보과학회).
장문수 · 오효정 · 장명길. 2001. "자동분류를 이용한 정답문서집합 구축." 「제13회 한글 및 한국어 정보처리 학술대회 논문집」(한국정보과학회).
장석진. 2001. "자연언어처리를 위한 중간언어 표상." 「학술원 논문집(인문 · 사회과학 편)」(학술원) 20.
정시호. 1994. 「어휘장 이론 연구」 경북대 출판부.
조평옥. 1996. "한국어 명사의 의미 계층 구조." 울산대 석사학위논문.
최경봉. 1996. "명사의 의미 분류에 대하여." 「한국어학」(한국어학회) 4.
최경봉. 1999. "단어 의미의 구성과 의미 확장 원리." 「한국어학」(한국어학회) 9.
최경봉. 2001. "지식기반 구축을 위한 어휘의 의미 분류." 「담화와 인지」(담화인지언어학회) 8-2.
최호섭. 2001. "한국어 명사 개념망 구축-경제용어를 중심으로." ETRI 지식정보검색연구팀 경제개념망 구축결과보고서.
최호섭 외. 2002. "사전을 기반으로 한 한국어 의미망 구축과 활용." 「제29회 한국정보과학회 춘계학술발표회 논문집(B)」(한국정보과학회).
한영균. 1994. "명사류 의미망 구축을 위한 사전 뜻풀이의 어휘구조분석." 「제6회 한글 및 한국어정보처리 학술대회 논문집」(한국정보과학회).
홍재성 외. 2001. "21세기 세종 계획 〈전자사전 개발분과〉 연구보고서." 문화관광부.
황도삼 · 최기선 · 김태석. 1998. 「자연언어처리」 홍릉과학출판사.
황도삼 · 최기선 · 김태석. 1999. 「자연언어이해」 홍릉과학출판사.

Beeferman, D. 1998. "Lexical discovery with an enriched semantic network." In Proceeding for the Workshop on Applications of WordNet in Natural Language Processing Systems, ACL/COLING 1998.
Davis, R. & Shrobe. H & Szolovits. P. 1993. "What is a Knowledge Representation." *AI magazine* 14(1).
Fellbaum, C. 1998. *WordNet*, The MIT press.
Hartley, R. T. & Barnden J. A. 1997. "Semantic networks: visualizations of knowledge." Trends in Cognitive Sciences Vol 1, No. 5, Elservier Science Ltd.

Lee, J. H. 2002. "A Korean Noun Semantic Hierarchy (WordNet) Constructi -on." Proceeding of The 16th Pacific Asia Conference.

Mahesh, K. 1996. "Ontology Development for Machine Translation: Ideol -ogy and Methodology." Memoranda in Computer Science and Cognitive Science, MCCS-96-292, CRL, New Mexico State University.

Sowa, J. F. 1984. *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley Publishing Company.

Sowa, J. F. 2000. *Knowledge Representation:Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA, http://www.jfsowa.com/krbook/index.htm.

Sowa, J. F. 2001. Ontology, http://www.jfsowa.com/ontology/index.htm.

Uschold, M. & Gruninger, M. 1996. "Ontologies: Principles, Methods and Applications." *Knowledge Engineering Review*. 11-2, The University of Edinburgh.

Xiaobin, Li. 1995. "A WordNet-based Algorithm for Word Sense Disambiguation." IJCAI.

최호섭(Choe, Hoseop)
울산대학교 컴퓨터정보통신공학과 자연언어처리연구실
680-749 울산광역시 남구 무거2동 산29 울산대학교 7호관 324-1호
전화 : (052)259-2625
팩스 : (052)959-5312
전자우편 : hoseop@mail.ulsan.ac.kr

옥철영(Ock, Cheolyoung)
울산대학교 컴퓨터정보통신공학부
680-749 울산광역시 남구 무거2동 산29 울산대학교 7호관 324호
전화 : (052)259-2222
팩스 : (052)959-5312
전자우편 : okcy@mail.ulsan.ac.kr

접수일 : 2002. 9. 23
심사의뢰일 : 2002. 10. 30
게재결정일 : 2002. 12. 9