한국어 워드넷의 구축[1)]

임성신[1] 이은령[2] 권혁철[3]
부산대학교 컴퓨터공학과[1,3], 부산대학교 한국어정보처리연구실[2]
{sslim, eunryounglee, hckwon}@pusan.ac.kr

Construction of Korean WordNet

Sung-Shin Lim[1] Eun-Ryoung Lee[2] Hyuk-Chul Kwon[3]
Dept. of Computer Science and Engineering, Pusan National University[1,3]
Korean Language Processing Laboratory, Pusan National University[2]

요 약

사람의 언어를 이해하는 자연언어처리 시스템을 개발하기 위해서는 의미처리를 위한 지식 베이스(knowledge base)가 필요하다. 지금까지 사람이 가진 지식 베이스를 컴퓨터에 도입하려는 많은 노력을 기울이고 있고 그 결과물로 온톨로지(ontology)와 시소러스(thesaurus)가 만들어지고 있다. 외국에서는 지식 베이스의 중요성을 알고 많은 연구를 수행하고 있으며 그 대표적인 사례들에는 Roget's Thesaurus, WordNet, EDR 개념사전, CYC, EuroWordNet 등이 있다.

이 중에서 가장 대표적이며 많은 활용을 보이는 것이 Princeton 대학의 WordNet이다. WordNet은 인간의 어휘지식에 대한 심리 언어학적인 연구의 결과물로써 심리학자와 언어학자들에 의해 10여 년 동안 구축되고 있는 영어에 대한 어휘데이터베이스이다.

본 논문에서는 WordNet을 기반으로 명사에 대해서 영한사전과 국어사전을 이용하여 구축한 한국어 워드넷을 소개하고, 구축시 고려한 기본지침을 소개하도록 하겠다.

1. 서 론

현재 컴퓨터 분야가 많은 발전을 이루고 있지만 아직까지 사람의 언어를 이해하는 자연언어처리 시스템은 개발되지 못하고 있다. 의미해석을 기반으로 하는 자연언어처리시스템의 응용 분야는 의미기반 정보검색, 자연어 질의응답, 지식 자동습득, 담화처리 등이 있다. 사람이 대화를 이해하기 위해서는 지식이나 공통의 개념들이 필요하며 앞에서 언급한 응용분야에 활용할 수 있는 자연언어처리 시스템을 구축하기 위해서는 의미처리를 위한 지식 베이스(knowledge base)가 필요하다. 지금까지 사람이 가진 지식 베이스를 컴퓨터에 도입하려는 많은 노력을 기울이고 있고 그 결과물로 온톨로지(ontology)와 시소러스(thesaurus)가 만들어지고 있다.

외국에서는 지식 베이스의 중요성을 알고 많은 연구를 수행하고 있으며 그 대표적인 것에는 Roget's Thesaurus, WordNet, EDR 개념사전, CYC, EuroWordNet 등이 있다.

그리고 우리나라에서는 울산대학교, KAIST, 포항공과대학교, 서울대학교 등에서 연구를 하고 있다.

이 중에서 가장 대표적이며 많은 활용을 보이는 것이 Princeton 대학의 WordNet이다. WordNet은 인간의 어휘지식에 대한 심리 언어학적인 연구의 결과물로써 심리학자와 언어학자들에 의해 10여 년 동안 구축되고 있는 영어에 대한 어휘데이터베이스이다[1]. WordNet의 구축 결과물은 많은 나라에서 어휘 데이터베이스의 모델로 삼아 연구하고 있을 만큼 대단히 인정을 받고 있다.

본 논문에서는 WordNet을 기반으로 명사에 대해서 영한사전과 국어사전을 이용하여 구축한 한국어 워드넷을 소개하고, 구축시 고려한 기본지침을 소개하도록 하겠다.

2. 관련연구

워드넷을 구축하기 위한 가장 정확하고 신뢰할 수 있는 방법은 수동으로 구축하는 것이다. 하지만, 비용적인 면 때문에 기존의 언어자원들을 활용하여 자동이나 반

1) 이 논문은 과학 기술부(한국과학기술기획평가원)의 국가지정연구실 사업지원으로 이루어진 것임(Contract Number : M1-0412-00-0028-04-J00-00-014-00).

자동으로 대부분 구축하고 있다.

한국어 워드넷을 구축하기 위한 연구는 울산대학교, KAIST, 포항공과대학교, 서울대학교 등에서 수행하고 있다.

울산대학교 한국어 의미망(Korean Semantic-Network)은 국어사전을 중심으로하여 언어처리에 필요한 다양한 정보를 담고 있다. KSN은 자동으로 구축하기 어려운 점이 많기 때문에 수작업과 자동 추출 작업을 병행하여 4만여 단어를 구축하고 있다[2][3].

KAIST의 한국어 어휘의미망은 반자동으로 구축되었다. 기계가독사전과 NTT시소러스를 기반으로 한국어 어휘의미망의 틀을 형성하고 이를 다시 수동으로 정리함으로써 한국어 어휘의미망을 구축하였다. 한국어 어휘의미망은 2,954개의 계층적 개념과 이에 속하는 명사, 동사, 형용사의 의미로 구성되어 있다[4].

포항공과대학교의 한국어 시소러스는 WordNet을 기반으로 한영사전과 국어사전을 이용하여 한국어 명사 개념체계를 자동으로 구축하였다[5].

서울대학교 한국어 명사 WordNet은 초등학교 1-6학년 교과서에서 쓰이고 국어사전에 수록된 명사 중 5,000개를 선정하고 선정된 명사를 기준으로 국어사전을 참고하여 명사의 정의를 자동으로 분석함으로써 명사의 상위개념의 한 level을 개발하였다. 그리고 영어의 WordNet과 국어학자의 단어개념을 참조하여 수정 확장하는 작업 수행하였다. 수정된 기본 틀을 기반으로 하여 분야가 명시된 20,000개 정도의 명사에 확대 적용하여 한국어 명사의 WordNet을 구축하였다[6].

3. 한국어 워드넷 구축

3.1 WordNet

WordNet은 인간의 어휘 지식에 대한 심리 언어학적인 연구의 결과물로써 1985년 Princeton 대학의 심리학자와 언어학자들에 의해 구축된 영어에 대한 어휘 데이터 베이스이다. WordNet에서는 사전을 단순히 자모순보다는 개념적으로 찾는데 이용하기 위해서 단어의 의미에 의해서 어휘정보를 조직화하려고 시도하였다[7].

WordNet의 기본 요소는 비슷한 의미의 단어들의 집합인 synset(set of synonym)이다.

WordNet은 인간의 어휘지식을 모방한 만큼 다의성과 동의관계를 이용하여 의미를 최대한 정확히 표현하고 있다. WordNet에서 단어와 의미 사이의 관계는 다음과 같은 단어형-의미 행렬을 전제로 한다.

단어형 / 의 미	F_1	F_2	F_3	...	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
:				...	
M_m					$E_{m,n}$

[그림 1] 단어형-의미 행렬

같은 의미를 나타내는 단어 F_1과 F_2는 동의어이며 여러 의미를 가지고 있는 단어 F_2는 다의어이다. 이와 같이 단어와 의미사이의 관계는 다대다(多對多) 대응관계이다.

WordNet에서는 한 의미를 표현할 때 이와 같은 동의어 집합을 {F_1, F_2, ...}와 같은 형태를 이용한다. 예를 들면 {board, plank}와 {board, committee}는 각각 하나의 의미를 표현하는 동의어 집합이다. 이 두 동의어 집합을 통해서 "board"라는 다의어의 여러 의미가 실체화될 수 있다. 이와 같은 동의어집합을 synset이라고 한다. 이러한 synset은 표현하는 의미가 무엇인지 명시하지는 않으며 다만 그와 같은 의미가 존재하는 것만을 나타낸다[8].

{board, plank}
{board, committee}
{board, table}
:

[그림 2] 간단한 synset의 예

단어형 / 의 미	board	plank	committee	table
{board, plank}	○	○		
{board, committee}	○		○	
{board, table}	○			○

[그림 3] 'board'와 관련된 단어-의미 행렬의 예

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

Note: N = Nouns Aj = Adjectives V = Verbs Av = Adverbs

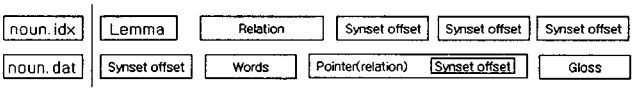
[그림 4] WordNet에서 정의된 관계 유형

WordNet에서는 앞에서 살펴본 바와 같이 synset을 정확히 표현하기 위해서 개념간의 관계를 이용한다. 따라서 의미표현 장치와 관계표현 장치가 분리되어 있지 않다. WordNet에서 사용하는 관계유형은 그림 4와 같다.

- 동의(同意, synonymy) : synset을 이루는 WordNet의 기본적인 관계
- 반의(反意, antonymy) : 단어연상 실험에서 형용사가 제시되었을 때 많은 사람들이 반의어를 연상하는 것이 관찰
- 하의(下意, hyponymy) : 상의(hypernymy)와 함께 관계쌍을 이루어 명사 synset 사이의 계층관계를 표현
- 분의(分意, meronymy) : holonym과 함께 부분-전체 관계를 표현
- 양식(樣式, Troponomy) : 함의(含意)의 일종, 동사 V1이 특정한 한 양식으로서 동사 V2를 행할 때 동사 V1은 동사 V2와 양식관계가 됨
- 함의(含意, entailment) : 동사 사이의 감춰진 내포관계(buy-pay) 또는 인과관계(give-have)를 표현하는데 쓰임

3.2 DB 구축을 위한 WordNet DB 분석

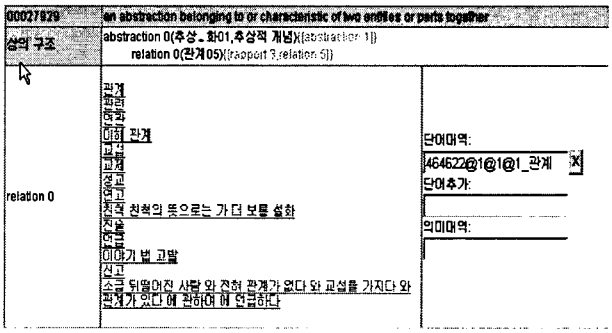
한국어 워드넷을 구축하기 위해 WordNet DB의 파일 구조를 분석하였다. WordNet 브라우저에서 사용하는 DB 파일은 idx파일과 dat파일이다. 이 파일을 분석해 본 결과 한국어로 번역을 하기 위해 RDB로 변환을 수행했다. 그리고 한글화를 수행할때는 해설(gloss)를 보고 단어들(words)를 번역하게 하였다. 이렇게 하면 synset들의 relation들은 워드넷의 구조를 이용함으로서 전체적인 기본 구조의 유지가 가능할 것으로 판단했다.



[그림 5] WordNet DB 파일 구조

3.3 한국어 워드넷 구축 도구

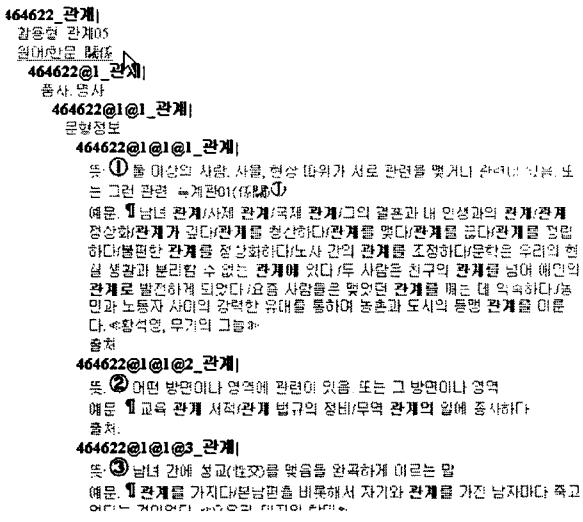
한국어 워드넷을 구축하기 프라임 영한사전과 표준국어대사전을 통합적으로 이용할 수 있는 도구를 개발하였다. 작업자는 구축한 도구를 이용하여 synset단위로 된 표제어를 수작업으로 번역을 하였다.



[그림 6] 영한 사전을 이용한 대역어 추천

이때 그림 6처럼 프라임 영한사전을 이용하여 영어 표제어에 대한 한글 대역어의 일관성을 유지하였다. 또한, WordNet에 상당한 수의 전문용어가 존재하므로 기 구축한 전문용어사전(과학기술분야 약 28만 표제어)을 이용하여 한글 대역어를 자동 추천하였다.

쉽게 표준국어대사전에서 의미를 검색하게 하기 위해 해당 표제어를 클릭하면 자동으로 표준국어 대사전을 검색하여 그 결과를 보여줄 수 있도록 하였다. 그런데 현재의 표준국어대사전은 한 단어가 여러 가지 의미를 가짐으로써 많은 구문적 · 의미적 · 개념적 부담이 있다. 이를 해결하기 위해 다의어일 경우 의미적으로 세분화를 수행하였다. 구체적으로 설명을 하면 표제어와 뜻풀이의 관계를 1 : 1로 하여 한 단어가 가질 수 있는 구문적 · 의미적 처리의 부담감을 줄이고자 하였다.



[그림 7] 표준국어대사전의 의미 세분화

3.4 한국어 워드넷 구축의 원칙

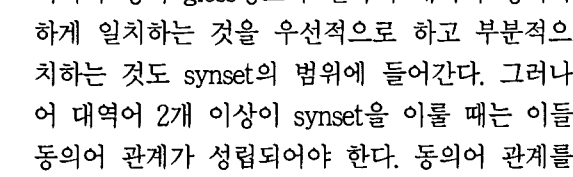
본 논문에서 제안하는 한국어 워드넷 구축의 원칙은 KIPONTO2004에서 제공한 1차 번역결과물에 대한 추가 정제작업에 대한 것이다[9].

- 어떤 synset을 기준으로 상하위 synset을 모두 검색했을때 상위에서 이미 존재하는 대역어가 노드를 하나 건너서 재귀적으로 하위에서 나오는 일이 없도록 해야 한다. KIPONTO2004에서 제공한 1차 번역의 결과 이와 같은 경우가 있어서 모두 수정하였다.
- 원칙적으로 영어의 gloss정보와 한국어 대역어 사전의 정의와의 일치를 기준으로 한국어 대역어를 선택한다. 여기서 영어 gloss정보가 한국어 대역어 정의와 정확하게 일치하는 것을 우선적으로 하고 부분적으로 일치하는 것도 synset의 범위에 들어간다. 그러나 한국어 대역어 2개 이상이 synset을 이룰 때는 이들 간의 동의어 관계가 성립되어야 한다. 동의어 관계를 테스트하는 기준은 "____이 명사1이면 ____은하는 적당한 표제어를 찾을 수 있는 경우에는 표준국어대사전에서 찾아 의미대역을 삭제하고 검색된 표제어를 등록해 준다.
- 대역어를 정제할 때 한국어 사전의 정의에서 "-인 무엇"에서 "무엇"에 해당하는 명사의 사전적 의미를 참조한다. 이것은 보통 그 단어의 상위어인 경우가 대부분이다. 예를 들어, synset "animal, animate_being, beast, brute, creature, fauna"의 대역어로 선정되어 있는 "짐승"의 사전적인 의미는 "몸에 털이나고 네발을 가진 동물"에서 "동물"은 "짐승"의 상위어로 상정되고 동물이 갖는 모든 의미자질이 "짐승"에도 전제된다. 따라서 한국어 대역어의 사전적 정보를 통해 상위어를 결정할 수 있으며 상하위의 관계로 설정된 것들은 동의어 관계로 중복되어 선택될 수 없다.
- 영어 명사의 synset에 대한 한국어 대역어는 명사를 원칙으로 한다. 현재 동사형으로 번역되어 있는 것은 명사형으로 수정한다.
- 현재 표기 불가나 의미대역으로 분류되어 있는 올바른 번역으로 나온 대역어가 명사구일 경우(정제시 작업자가 새로운 복합어를 만드는 경우도 마찬가지)에는 corpus에서 출현 여부에 따라 등록 여부를 결정한다.
- 한국어 대역어가 부재하여 의미대역을 하였을 경우 번역이 바르게 되었는지 재검토한다.
- 워드넷 gloss번역으로 되어 있는 경우라도 작업자가 이에 해당하는 적당한 표제어를 찾을 수 있는 경우에는 표준국어대사전에서 찾아 의미대역을 삭제하고 검색된 표제어를 등록해 준다.
- 하나의 영어 synset의 개념이 한국어에서 두 개 이상으로 세분화될 수 있는지를 항상 염두에 두고 검토한다.

3.5 한국어 워드넷 구축의 문제점 및 해결방안

WordNet의 전체데이터를 기반으로 한국어 워드넷을 구축하면서 발생한 문제점을 소개하고 해결방법을 소개하도록 하겠다.

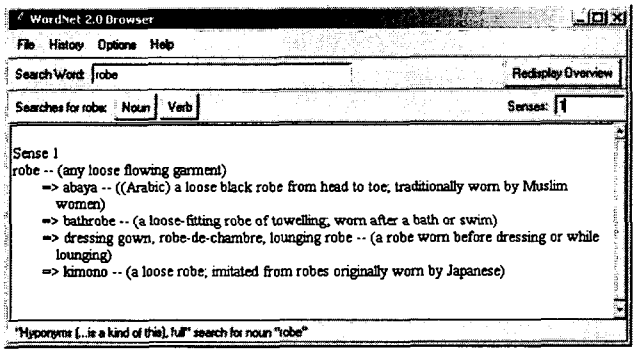
- 영어와 한국어의 의미세분화(depth)가 같지 않은 것이 존재 : 영어의 의미세분화가 한국어에서는 불필요한 경우가 존재하였다. 이와 같은 경우에는 하나의 노드로 묶어줄 수 있다.



[그림 8] 한국어의 경우 의미세분화가 되지 않는 경우

- 번역자의 역량 문제 : 기 구축되어있는 지식베이스를 번역하여 자체적인 지식베이스를 구축할 때 품질을 보증하기 위해서는 작업자의 능력이 매우 중요한 요소이다. 그리고 여러 사람이 작업을 할 경우에는 일관성을 유지하기가 힘들다. 이를 해결하기 위해 객관적인 지식베이스 구축을 위한 도구의 개발이 필요하고 체계적인 작업지침의 수립과 교육이 필요하다. 특히 작업도구에는 개인의 언어적 지식차이를 극복하기 위한 영어사전, 전문용어사전, 동의어/유의어 사전과 같은 기구축한 사전의 활용이 중요하다.
- 한국어에는 개념이 있으나 영어에는 개념이 없는 것

이 존재 : 현재 WordNet 2.0의 경우에 robe(any loose flowing garment)의 하위어로 "기모노(kimono)"는 포함되어 있지만 "한복"은 그렇지 못하다. 이처럼 우리나라의 고유한 전통이나 문화적 개념들은 정리하여 추가 구축하여야 한다.

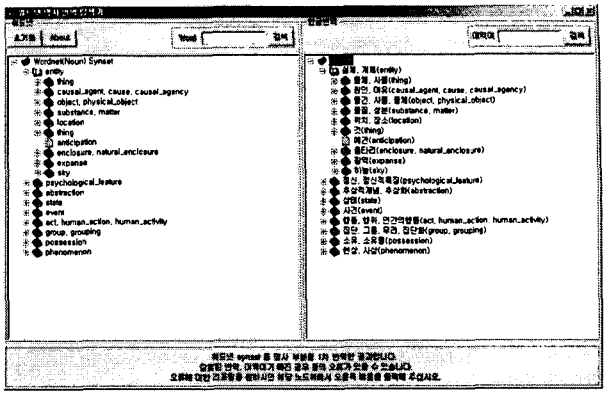


[그림 9] 한국어에는 있으나 영어에 개념이 없는 사례

- WordNet 자체의 문제 : formal과 informal 한 용어가 혼재되어 synset으로도 또는 하위어로도 나타나는 현상이 있다. 그리고 반의어의 경우 WordNet에서는 단어대 단어의 대립이 아니고 synset이 나타내는 개념의 대립이다. 따라서 이는 일반적인 반의어 사전에서 볼수 있는 의미대립의 단어쌍과는 다를수 있다. 한국어 대역어는 워드넷의 반의어 관계를 반영함으로 실제로 한국어에서의 의미대립은 WordNet과는 매우 다를 수 있다. 따라서 한국어에서 반의어 관계 설정시 WordNet의 반의어 관계수정이 요구된다. 그 실제적인 이유는 특정 단어와 단어의 대립관계가 synset의 반의어 쌍이 포함되기도 하고 그렇지 않기도 한다. 그리고 개념 대립의 유형은 여러 가지로 나타난다. 특히 단어 X의 반의어는 "not-X"라고 할 수 있지만 항상 그런 것은 아니다. EuroWordNet의 경우에는 near synonym관계로 synset 관계에서의 반의어가 관계가 아니라 단어 간에 반의어 관계를 설정해 놓고 있다. 그리고 WordNet에서는 대립되는 개념이 실제로 반의어관계뿐만 아니라 hyponym관계로도 설정되어 있다.

3.6 한국어 워드넷 브라우저

구축한 한국어 워드넷을 효과적으로 탐색하기 위해서 브라우저를 개발하였다. 브라우저는 영어와 한글을 동시에 볼 수 있도록 두 개의 화면으로 구성되어 있으며 클릭을 하면 동시에 내용을 볼 수 있도록 서로 동기화가 되어있다.



[그림 10] 한국어 워드넷 브라우저

3.7 구축한 명사 한국어 워드넷의 정보

구축한 명사 한국어 워드넷은 모두 109,586개의 개념으로 구성되어 있다.

표 1. 구축한 명사 한국어 워드넷의 의미 정보

구 분	개 수
표준국어대사전에 이미 있던 의미의 개수	69,134
새로 등록한 의미 개수	23,157
의미대역의 개수	17,295
총 계	109,586

전체 구축한 한국어 워드넷은 synset을 기준으로 영어워드넷의 73.6%인 58,656개이다.

4. 결론 및 향후 연구과제

본 논문에서는 의미기반 자연언어처리 시스템 개발을 위한 핵심 기술인 지식 베이스(knowledge base) 구축에 대해 알아보았다. 외국에서는 지식 베이스의 중요성을 간파하고 오래전부터 많은 연구를 수행하고 있으며 그 대표적인 것에는 Roget's Thesaurus, WordNet, EDR 개념사전, CYC, EuroWordNet 등이 있다.

본 논문에서는 가장 대표적이며 많은 활용을 보이는 Princeton 대학의 WordNet을 기반으로 명사에 대해서 영한사전과 국어사전을 이용하여 구축한 한국어 워드넷을 소개하고, 구축시 고려한 기본지침을 소개하였다. KIPONTO2004에서 1차 결과물을 공개하였으나 1차 작업의 결과물이다 보니 상당한 오류가 있었다. 그래서 기본적인 지침을 새롭게 마련하고 언어학 박사들을 주축으로 재정비를 수행하여 상당한 수준의 개선을 보였다. 또한 WordNet의 동사도 번역을 수행하여 기본적인 동사 어휘망을 구축하였다.

그리고 향후 WordNet의 형용사와 부사를 번역하여

전체적인 한국어 워드넷을 구축할 것이며 EuroWordNet을 기본구조로 다국어에 사용할 수 있는 워드넷을 구축을 검토하고 있다. 또한, 한국의 문화와 전통을 나타내는 개념들을 이미 구축한 워드넷에 추가하여 좀 더 완성도 높은 한국어 워드넷을 구축할 것이다.

향후 구축한 한국어 워드넷을 번역시스템과 의미태깅 시스템에 적용하여 성능을 검증할 것이다.

5. 참고 문헌

[1] Christiane Fellbaum. "WORDNET : AN ELECTRONIC LEXICAL DATABASE", MIT Press, 1998

[2] 조평옥. "한국어 명사의 의미 계층 구조 구축", 울산대학교 석사학위논문, 1996

[3] 최호섭. 옥철영. "한국어 의미망 구축과 활용 - 명사를 중심으로", 한국어학 제17집, 2002

[4] 코텀. "카이스트 한국어 어휘의미망",제15회 한글 및 한국어 정보처리 학술대회, 2003

[5] 이창기. 이근배. "WordNet을 이용한 한국어 시소러스 자동 구축", 제11회 한글 및 한국어 정보처리 학술대회, 1999

[6] 문유진. "한국어 명사 WordNet의 구축 방안 및 구현에 관한 연구", 한국과학재단연구보고서, 1995

[7] Gorge A. Miller, etc. "Introduction to WordNet : An On-line Lexical Database", International Journal of Lexicography, 1990

[8] 이재윤. 김태수. "WordNet과 시소러스", 언어정보개발연구 창간호, 1998

[9] 이은령. 임성신. "WordNet 2.0의 한국어 번역 작업과 결과물", 제2회 지식정보처리와 온톨로지 워크숍 발표자료집, 2004

[10] Piek Vossen편, 한정한 외 6명 공역. "EuroWordNet, 유로워드넷", 한국문화사, 2004