사전을 기반으로 한 한국어 의미망 구축과 활용

최호섭[0]* 옥철영* 장문수** 장명길**
*울산대학교 컴퓨터정보통신공학부 **한국전자통신연구원 휴먼정보검색연구팀
(hoseop, okcy)@mail.ulsan.ac.kr / (cosmos, mgjang)@etri.re.kr

Construction and application of Korean Semantic-Network based on Korean Dictionary

Ho-Seop Choe[0]*, Cheol-Young Ock*, Moon-Soo Chang**, Myung-Gil Jang**
Dept. of Computer Engineering and Information Technology, University of Ulsan
Human Information Retrieval Reaserch Team, ETRI

요 약

시소러스, 의미망, 온톨로지 등과 같은 지식베이스는 자연언어처리와 관련된 여러 분야에서 중요한 언어자원의 역할을 담당하고 있다. 하지만 정보검색, 기계번역과 같은 특정 분야마다 다르게 구축되어 이러한 지식베이스는 실질적인 한국어 처리에는 크게 효과를 보지 못하고 있는 실정이다. 본 논문은 한국어를 대상으로 한 시소러스, 의미망 등의 구축 방법론적 문제를 지적하고, 말뭉치를 중심으로 한 텍스트 언어처리에 필요한 의미망의 구축 방법과 포괄적인 활용 방안을 모색한다. 의미망 구축의 기반이 되는 지식은 각종 사전(dictionary)를 이용했으며, 구축하고 있는 의미망의 활용 가능성을 평가하기 위하여 ETRI의 '의미기반 정보검색'과 언어처리의 큰 문제 중 하나인 단어 중의성 해소(WSD)에서 어떻게 활용되는지를 살핀다. 그리하여 언어자원의 처리 방안 중의 하나인 의미망을 구축함으로써 언어를 효과적으로 처리하기 위한 기본적이면서 중요한 어휘 데이터베이스 마련과 동시에 언어자원 구축의 한 방향을 제시하고자 한다.

1. 서 론

시소러스를 비롯한 온톨로지, 의미망 등과 같은 어휘들간의 문법적·의미적 상호 관계를 중심으로 한 지식베이스(knowledge base)는 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 구성하기 위해서 사전, 동의어 사전, 말뭉치, WordNet을 유기적으로 통합하여 하나의 연속적인 "어휘(lexicon)-구(phrase)-연어(collocation)-관용어(idiomatic expression)-말뭉치"를 마련하거나[1], 세부적으로는 품사정보, 형태소 정보, 의미 관계, 구문 정보, 선택 제약 정보(selectional preference), 의미 정보(semantic information) 등을 긴밀하게 연결시킨 지식베이스 구축이 필요하다[2]. 즉 이러한 지식베이스 구축은 특정 분야에 맞는 개별적 구축도 중요하지만 언어처리 전반에 통용될 수 있는 구축 방향도 모색되어야만 한다.

본 논문에서는 말뭉치와 같은 텍스트 언어처리에서 적절하게 이용될 수 있고 나아가 언어처리 전반에 적절하게 사용될 수 있는 의미망 구축의 한 방향을 제시하고자 한다. 먼저 국내에서 구축되었던 시소러스, 의미망 등의 구성상 문제점을 간략하게 설명하고, 이러한 문제점을 보완하면서 각종 사전(dictionary)을 기반으로 새롭게 구축하고 있는 한국어 의미망(Korean Semantic Network; 이하 KSN)의 구축 방법을 밝힌다. 다음으로 ETRI에서 개발 중인 "의미기반 정보검색"에서 이용되고 있는 '한국어 명사 개념망'은 본 논문의 의미망 구축 방법으로 구성되고 있으므로, KSN의 정보검색에의 이용을 살펴본다. 마지막으로 단어 중의성 해소(이하 WSD)를 통해 KSN이 어떻게 이용될 수 있는지를 살핌으로써 KSN의 포괄적인 적용 양상을 검토하고자 한다.

2. 기존 지식베이스 구성상의 문제점

프린스턴 대학의 워드넷(WordNet), 가도카와(Kadokawa) 시소러스, 마이크로코스모스 온톨로지(Mikrokosmos ontology), EDR 개념사전 등은 시소러스, 의미망, 온톨로지로 명명하면서 구축된 국외의 지식베이스들이다. 국내에서도 각 분야마다 상당수의 시소러스가 구축되었으며, 더 나아가 ETRI의 한국어 명사 개념망[3], 포항공대의 LIP(Language Independent and Practical) 온톨로지[2], 울산대와 한국과학기술원의 한국어 명사 의미 계층 구조[4][5], 한국어 명사 워드넷[6] 등의 다각적인 지식베이스 구축 방법과 실제를 보여주고 있다. 덧붙여 "21세기 세종 계획"의 전자사전 개발에서도 전산적 처리와 접목시킨 국어학적 어휘 분류 방식을 채택하여 이용 가능성이 논의되고 있다.

하지만 다양한 방법과 구성 원리로 구축된 지식베이스들이 실질적인 언어처리에서 큰 효과를 거두지 못하는 실정이다. 국내에서 구축된 지식베이스를 대상으로 하여 문제점을 간략하게 기술하면 다음과 같다.

첫째, 기계번역, 정보검색 등과 같은 특정 분야의 특성을 많이 고려하였기 때문에, 한국어를 대상으로 한 언어처리에는 그리 큰 실효성을 거두지 못하고 있다. 기계번역에서는 대상언어와 목적언어의 대역성(translation)에 치중하여, WordNet이나 가도카와 시소러스 등과 같은 국외에서 구축된 지식베이스를 번역하거나 응용하는 경향이 많다. 그리고 정보검색에서도 사용자의 언어 사용과 학문 분야의 특수성을 고려한 시소러스를 구축하다 보니, 통합적인 시소러스 구현에 어려움이 많다는 지적도 있다.

둘째, 시소러스나 의미망의 구성적인 면에서, 계층적인(hierarchical) 구조와 분류(부류)적인 구조가 혼합되어 사용되는 경우가 많아 일관된 구조체를 형성하지 못하고 있다. 시소러스에서 가장 기본적인 틀을 이루는 상하관계(BT/NT)에 경우 is_a 관계의 모호성으로 인해 kind_of, part_of, use_of 등이 혼용되는 경우가 많다. 이것은 정확한 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. 이러한 경향은 언어의 사회적 성격을 모두 수용할 수 있는 지식베이스 구축이라는 큰 부담감을 가지는 요소로 작용할 수 있다.

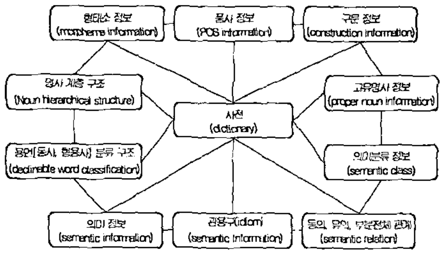
셋째, 각종 사전을 기반으로 구축된 지식베이스가 사전이 가

지고 있는 모든 정보를 이용하고 있지 않다. 국어사전에서는 표제어, 뜻풀이, 동의·유의관계, 형태소 정보, 구문 정보, 관련(related) 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 활용할 수 있다.

이외에도 어휘들간의 관계(relation) 또는 노드(node)간의 연결(link) 문제, 지식베이스 구축 범위 등의 문제점이 있지만, 대부분의 문제가 수작업에 대한 부담감, 많은 시간과 연구자의 필요에 따른 어려움에서 비롯한 것이라 할 수 있다.

3. KSN의 구성과 구축 방법

KSN 구축은 언어처리에서 효과적으로 사용할 수 있는 지식베이스를 마련하기 위한 작업이다. KSN은 [그림1]를 통해 알 수 있듯이, 사전을 기반으로 하여 언어처리에 필요한 다양한 정보를 담는다.



[그림 1] KSN의 기본 구성도

여기에서 중요한 것은 모든 사전이 이러한 정보를 다 가지고 있는 것이 아니기 때문에, 규모가 큰 국어사전을 중심으로 하고, 명사 중심으로 기술되어 있는 전문용어사전이나 백과사전은 전문성이 필요한 의미분류 정보나 의미 정보를 보완할 때 사용한다.

KSN의 구성적인 면을 간략하게 설명하면, 우선 KSN에 기본축인 사전은 국어사전에 등재되어 있는 모든 단어를 대상으로 한다. 여기에서 중요한 점은 KSN의 사전은 사전의 표제어보다 많은 단어를 가진다는 점이다. 표제어와 뜻풀이의 관계를 다의어와 같이 1:n(n은 뜻풀이 개수)의 관계가 아니라 단의어와 같은 1:1의 관계로 보느냐에 따라 달라진다. [표1]과 같이 KSN에서는 다의어를 분리하여 단의어로 둠으로써, 한 단어가 가질 수 있는 많은 구문적·의미적 처리의 부담감을 줄이고자 하였다.

[표 1] '차'에 대한 KSN의 사전 DB 구성

ID	Entry	SemTag	Explain
421	차	차_1	명장기의 하나인 '착살'을 십팔기(十八技)에서 지목하여 부르는 이름
422	차	차_2	(어근) '차하다'의 어근
423	차	차_3	(의존명사)(次) 1 「주로 한자어 수 뒤에 쓰여」 '번', '차례'의 뜻을 나타내는 말.
424	차	차_3	(의존명사)(次) 2 「-던 차에, -던 차이다 구성으로 쓰여」 어떠한 일을 하던 기회나 순간.
425	차	차_3	(의존명사)(次) 3 [수학] 방정식 따위의 차수를 이르는 말
426	차	차_4	(次) [음악] '제 7'의 잘못. '제'를 한자를 빌려서 쓴 말
427	차	차_5	(대명사)(此) '이, 5 1'를 문어적으로 이르는 말.
428	차	차_6	(車) 1 바퀴가 굴러서 나아가게 되어 있는, 사람이나 짐을 실어 옮기는 기관
429	차	차_6	(車) 2 「수량을 나타내는 말 뒤에 쓰여」 트럭을 1 에 실어 그 분량을 세는 단위.
430	차	차_6	(車) 3 [운동/오락] 車 자를 새긴 장기짝.
431	차	차_7	(車) 우리나라 성(姓)의 하나
432	차	차_8	(差) 1 둘 이상의 사물을 견주었을 때에, 서로 다르게 나타나는 수준이나 정도
433	차	차_8	(差) 2 [수학] 어떤 수나 식에서 다른 수나 식을 뺀 나머지
434	차	차_9	(茶) 1 차나무의 어린잎을 달이거나 우린 물.
435	차	차_9	(茶) 2 식물의 잎이나 뿌리, 과실 따위를 달이거나 우리거나 하여 만든 마실 것을 통틀어 이르는 말
436	차	차_9	(茶) 3 [식물] =차나무.
437	차	차_10	(釵) 두 갈래 진 비녀의 하나
438	차	차_11	(箚) [역사] =차자04(箚子)
439	차	차_12	(접사) 「ㅊ으로 시작되는 몇몇 명사 앞에 붙어」 끈기가 있어 차진 의 뜻을 더하는 접두사
440	차	차_13	(접사)(次) 「일부 명사 뒤에 붙어」 '목적'의 뜻을 더하는 접미사

형태소 정보에는 어근, 어미, 접사 등 품사와 겹치지 않는 정보를 가지며, 품사 정보는 명사, 대명사, 수사, 동사, 형용사, 부사, 관형사, 감탄사, 조사 등의 정보를 가진다. 형태소 정보와 품사 정보는 형태소 분석과 연관성이 많으므로 두 정보를 하나로 묶을 수도 있을 것이다.

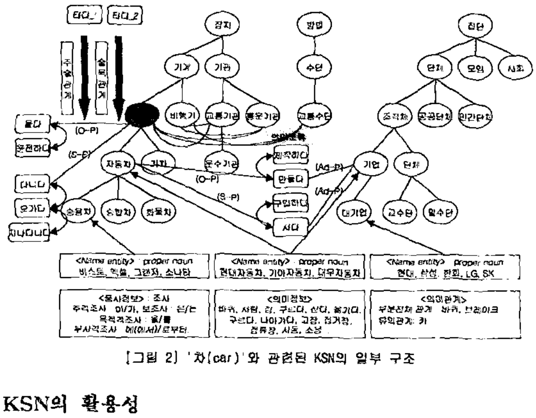
구문 정보는 기존의 격틀(*case frame*) 구조과 비슷하지만 한국어의 문장성분(sentence component)에 의해 주어-서술어 관계, 목적어-서술어 관계, 보어-서술어 관계 등과 같이 문장성분간의 관계를 중심으로 연결된 정보를 핵심으로 한다는 점에서 격틀 구조와는 다르다고 할 수 있다. 구문 정보의 주를 이루는 것은 명사와 용언의 문장성분을 이용한 관계이다. 한국어의 문장성분은 필수 성분과 부속 성분으로 구성되는데, 전자에는 주어, 목적어, 보어, 서술어, 필수적 부사어 등이 속하고, 후자에는 부사어, 관형어, 독립어 등이 속한다. KSN 구축에서 주목하고자 하는 것은 서술어를 중심으로 하여 주어, 목적어, 보어, 필수적 부사어를 각각 연결시켜 구문적·의미적 결정성(*decision*)을 확보하는 것이다. 예를 들어 '배'와 '타다'라는 단어의 구문적 연결은 주어-서술어 관계, 목적어-서술어 관계, 부사어-서술어 관계를 형성하는데, 이러한 관계 설정은 서술어가 명사를 수식하는 관형어의 역할을 담당하더라도 변하지 않는다는 점을 고려한 것이다.

고유명사 정보는 인명, 회사명, 상품명 등의 고유명사를 일반명사에서 분리시켜 개별적으로 관리함과 동시에 일반명사와의 관계를 설정하도록 한다. 또한 관용구(연어 포함) 정보, 동의·유의관계 정보는 사전을 적극 활용한다.

의미 정보는 ① 문장 내에서 한 단어를 이해하는 데 중요한 역할을 하는 단어, ② 국어사전에 등재되어 있는 표제어를 이해할 수 있도록 하는 뜻풀이 내의 단어, ③ 문서나 문장에서 특정 단어와 함께 출현하는 빈도가 높은 단어(실마리단어; clue-word) 등으로 구성된다.

KSN 구축의 중점 사항 중 하나는 명사 계층 구조와 용언 분류 구조, 의미분류 정보의 구축으로, 이들은 구문·의미 관계를 이용해 서로 연결된다. 또한 앞에서 기술한 구문 정보와 밀접한 관계를 가지며, 동의·유의·부분전체 관계 등의 의미 관계(semantic relation)가 명사 계층 구조에 포함된다. 명사 계층 구조는 명사 자체가 가지고 본래의 의미를 중심으로 상하관계를 형성하며, 문맥적 상황을 많이 고려하는 명사는 그 사용성에 중점을 둔 의미부류 관계로 설정한다. 이러한 명사 의미 계층 구조와 다른 정보와의 관계 설정은 (1) 시소러스, 온톨로지 등에서 많이 사용되는 *최상위 부류나 최상위어 설정의 부담을* 줄이며, (2) 어떠한 문장의 표면 구조(surface structure)를 중심으로 단어의 계열 관계(paradigmatic relation)와 통합 관계(syntagmatic relation)를 파악하여 계층 구조의 확충과 동일 의미부류 집합 구축에 용이하다.

의미망은 자동적으로 구축하기 어려운 점이 많기 때문에, KSN도 수작업과 자동 추출 작업을 병행하여 구축되고 있다. 하지만 KSN 구축에서는 명사 계층 구조, 의미정보, 구문정보, 품사정보 등 많은 정보를 자동으로 추출하여 관계를 설정하고 있다. [그림 2]는 KSN의 일부 구조를 나타낸 것이다.



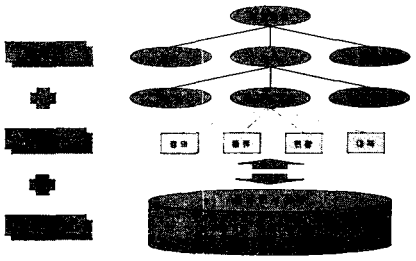
[그림 2] '차(car)'와 관련된 KSN의 일부 구조

4. KSN의 활용성

KSN의 활용 가능성을 모색하기 위하여, ETRI에서 구축 중인 '의미기반 정보검색'의 지식베이스의 일부인 명사 개념망(Noun Conceptual-Network)의 사용과 WSD에서의 KSN의 적용 방법을 간략하게 살펴보기로 한다.

4.1 ETRI 한국어 명사 개념망

ETRI의 의미기반 정보검색은 텍스트 문장의 의미를 정확하게 하는 자연어처리 기술을 적용한 의미 색인 기술과 지식베이스로 자동 구축된 정답문서 기반 검색으로서, 인터넷이 제공하는 방대하고 다양한 미디어 정보를 사용자가 편리하고 정확하게 찾을 수 있도록 의미에 기반하여 분석, 색인하여 검색할 수 있는 핵심기술 개발을 최종 목표로 삼고 있다. 현재 문장 의미기반 정보검색 기술 개발 단계에까지 이르고 있다. 이 기술에 사용된 지식베이스의 구성은 [그림 3]과 같다.



[그림 3] ETRI 지식베이스의 구성

ETRI 명사 개념망이란 한국어 명사 어휘로 표현되는 개념을 정확하게 파악하기 위하여 개념들간의 다양한 관계를 연결시켜 놓은 어휘 데이터베이스를 말한다. 이 개념망은 국어학적인 의미관계를 이용하여 상하관계를 기본축으로 하고 있으며, 상하관계의 보완적 측면에서 동의·유의관계, 부분-전체관계, 반의관계 등을 추가로 정의하고 있다.[3][7][8]

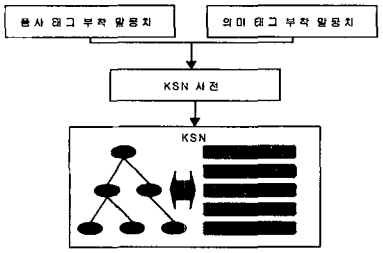
이 개념망의 명사 계층 구조는 정답문서집합(answer set)과 개념망을 연결시켜 주는 속성(attribute)과 연관성이 많다. is_a 관계의 모호성으로 인해 상하관계의 명확한 기준이 없었던 기존의 시소러스는 상하관계를 통해 공통된 질의 패턴과 속성을 추출하기 어렵다. 예를 들어 '자동차'의 경우, 용언이나 다른 명사와 함께 쓰여 교통수단(운송수단), 교통기관, 상품(제품) 등과 같은 다양한 쓰임을 가지고 있기 때문에, 기존의 시소러스에서는 이러한 점에 감안하여 상하관계를 설정하고 있는 경우가 많다. 이러한 시소러스의 계층 구조는 바로 위 층위(level)의 상위어만 이용 가능한 경우가 많고, 공통된 속성을 형성하기 어려울 뿐만 아니라 더 나아가 자연언어처리 기술을 이용한 정보검색에서도 큰 효율성이 없다. 하지만 사전에서 제공하는 '자동차' 본래의 의미를 이용한 계층 구조를 구축하여 '자동차-차-기계-장치'라는 상하관계를 설정함으로써, 대치가능성을 고려한 질의 분석과 공통 속성의 연계성을 획득할 수 있다. 즉 "자동차(차;기계;장치) 구입"과 같은 사용자의 요구 사항 인식과 '정의', '종류', '뉴스', '상품' 등과 같은 공통 속성의 상하위 개념어의 적용 여부 파악을 통하여, 사용자가 원하는 정답문서를 보여주는 처리 과정의 일부분을 담당할 수 있는 것이다.

KSN의 명사 계층 구조는 ETRI 명사 개념망과 비슷한 구조를 가지고 있으므로, 의미기반 정보검색 시스템을 통해 정보검색에서의 KSN 활용성을 실험하고 평가할 수 있을 것이다.

4.2 KSN을 이용한 WSD

WSD 방법 중 단어의 계층적 구조를 이용하는 방법이 있다. 이것은 시소러스, 의미망 등의 상하관계, 동의·유의관계로 설정된 단어나 노드간의 유사도, 거리 측정을 통해 단어 중의성을 해결하는 방법이다. WordNet을 이용한 WSD[9], 온톨로지를 이용한 WSD[2] 등이 대표적이라 할 수 있다. 하지만 국내에서는 동형어 분별과 의미에 의한 단어 분별을 명확히 구분하지 않아 많은 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 부담을 가지게 된다. 이러한 문제를 해결하기 위해 최근 다의어 처리에 대한 논의가 진행되고 있다.

3장에서 잠시 언급했듯이 KSN은 표제어와 뜻풀이를 1:n이 아니라 1:1 관계로 다루어, 한 다의어 내의 중의성 해소뿐만 아니라 동형어의 중의성 해소 방안을 고려하여 구축되고 있다. 또한 KSN은 기존의 시소러스, 의미망 등과 다르게, 단어 본래의 의미를 중심으로 하여 한 단어의 문장 속에서의 계열관계와 통합관계를 적절히 파악할 수 있도록 구축되고 있다. 즉 KSN은 단어의 계층적 구조를 이용한 WSD에 단어간의 거리나 유사도, 노드간의 거리나 유사도 측정에 일관성을 부여할 수 있는 구조를 가지는 것이다. 현재 말뭉치에서 빈도가 높은 동형어를 대상으로 KSN을 이용한 WSD를 실험 중에 있다.



[그림 4] KSN을 이용한 WSD 1차 실험 단계

5. 결론

본 논문에서는 기존의 시소러스, 의미망 등과 같은 언어자원의 효과적인 구조체인 지식베이스 구축과 관련된 문제점을 지적하고, 현재 구축 중인 KSN의 구성과 활용 가능성을 제시하고자 하였다.

KSN 구축은 장시간이 요구되는 작업이지만, 언어처리에 필요한 언어자원이자 지식베이스 확보라는 측면에서는 꼭 필요한 작업이기도 하다. 수작업의 부담감을 줄이기 위해, 기존의 울산대 명사 의미 계층 자동 구축 기술을 수정·보완하고 있으며, KSN의 각종 정보는 사전, 말뭉치를 통해 자동 수집·분류하고 있다.

형태소·구문 분석과의 상호 연계성 문제, KSN을 통합 관리할 수 있는 도구 개발 문제, 복잡한 문장 구성에서의 KSN 활용 문제 등은 본 연구가 앞으로 해결해야 할 과제이다.

6. 참고문헌

[1] 김영택 외, 자연언어처리, 생능출판사, 2001
[2] 강신재, 실용적인 온톨러지의 반자동 구축 및 어휘 의미 중의성 해소를 위한 응용, 포항공대 박사학위논문, 2002
[3] 장명길 외, 인터넷 질의/응답을 위한 지식베이스 구축, 제12회 한글 및 한국어 정보처리 학술대회, p198~202, 2000
[4] 조평옥, 한국어 명사의 의미 계층 구조, 울산대 석사학위논문, 1996
[5] J.H. Lee 외, A Korean Noun Semantic Hierarchy (Wordnet) Construction, Proceedings of The 16th Pacific Asia Conference, p290~295, 2002
[6] 문유진, 한국어 명사를 위한 WordNet의 설계와 구현, 정보과학회 논문지, 제2권, 4호, p437~445, 1996
[7] 장문수 외, 경제개념망 구축 결과보고서, TDP, ETRI, 2001
[8] 장문수 외, 의미관계연구회 결과보고서, TDP, ETRI, 2001
[9] Xiaobin Li 외, A WordNet-based Algorithm for Word Sense Disambiguation, IJCAI, 1995