한국어 동사 의미처리를 위한 SENKOV의 구축과 공기제약 관계에의 활용*

고병수, 정성훈, 문유진
호남대학교 컴퓨터공학과

Implementation of SENKOV and Its Application to the Selectional Restriction for Semantic Analysis of Korean Verbs

Byung-Soo Ko, Seong-Hoon Chung, Yoo-Jin Moon
Computer Engineering Dept. of Honam University
kbs@honam.honam.ac.kr, yjmoon@honam.honam.ac.kr

요 약

본 논문은 의미론적 어휘개념에 기반한 한국어 동사 Isa 계층구조의 시스템을 설계하고 한국어 동사의 Isa 계층구조 시스템을 이용한 Semantic Network을 구축하며, 이를 활용하여 부사와 동사 간의 공기제약관계 설정에 유효한 개념 분류를 수행한다. 일반적으로 많이 쓰이는 한국어 동사 658개를 대상으로 semantic network을 구축한 결과, SENKOV는 44개의 top node를 가지고 있으며 depth는 약 2.35이었다. 한국어 동사의 semantic network은 영어에서와 마찬가지로 명사보다 top node의 개수는 많고 depth가 훨씬 더 얕았다. 그리고 성상부사의 selectional restriction에 유효한 개념분류를 하는데 SENKOV를 활용하였다.

1. 서론

한국어 동사의 의미분석을 수행하기 위해서는 문장에 쓰인 동사들을 문자 자체로만 인식해서는 안되고 동사의 개념에 의한 상관관계를 네트워크 형식으로 구축해주어야 한다[1, 5]. 한국어 문장의 의미분석을 위해서는 한국어 동사의 의미분석이 핵심을 이룬다고 할 수 있다. 따라서 동사의 개념인식을 위하여 동사의 상위개념을 Semantic Network로 구축하는 것이 급선무라고 할 수 있다

한국어 동사의 Semantic Network(Semantic Networks for Korean Verbs SENKOV)이 구축되면 동사의 개념에 의하여 부사의 selectional restriction을 정의하므로써, 한국어 정보처리 및 자연어 처리를 효율적으로 하는데 기여할 수 있다.

본 논문은 의미론적 어휘개념에 기반한 한국어 동사 Isa 계층구조의 시스템을 설계하고 한국어 동사의 Isa 계층구조 시스템을 이용한 Semantic Network을 구축하며, 그리고 부사와 동사 간의 공기제약 관계 설정에 유효한 개념 분류를 수행한다.

2. 관련 연구

현재까지 진행된 영어 동사 분류 방법에는 WordNet의 synonym set과 Levin의 verb class가 있다

WordNet[4]은 약 120,000개의 영어단어(명사, 동사, 형용사 그리고 부사)를 어휘 개념으로 표현하여 semantic network을 구성한 일종의 온라인 영어 데이터베이스이다 WordNet의 기본적 골격구조는 어휘개념인데 동의어 집합(synonym set, synset)으로 표현되고 있으며 이 동의어 집합들 간의 상하위개념 관계는 계층구조를 표현하고, WordNet은 이 동의어 집합을 사용하여 시소러스의 역할도 수행할 수 있도록 한다. 인간의 어휘 지식을 모방하려는 시도에서 WordNet은 어휘형과 어휘개념 사이의 관계에 대한 정보를 표현하고 있다. 그리고 독일, 스페인, 프랑스 등 유럽에서도 영어의 WordNet을 기반으로 하여 semantic network에 관한 작업을 계속하고 있다.

Levin verb class[2,3]는 의미중심의 분류구조로서, 통사적 관계를 고려하여 selectional restriction의 유무 등을 감안하여 3,000여 개의 동사를 49개의 class로 분류한 것이다. 동사가 취할수 있는 통사구조는 class membership을 결정한다. 기본적인 가정은 동사의 통사구조가 내재한 의미를 직접 반영한다는 것이다

3. SENKOV의 설계와 구축

3.1 SENKOV의 어휘개념과 계층구조

한국어 동사 Isa 계층구조를 위한 시스템의 설계에 있어서 중요한 관건은 기본 골격인 어휘개념을 표현하는 형식과 계층구조 관계 설정이다.

의미론에서 어휘화된 개념을 정의에 의하여 어떻게 표현할 수 있는가에 대한 이론은 개념을 '구성화(constructive)'하려는 것인지 혹은 '차별화(differential)'하려는 것인지에 달려 있다([4]). 구성화 이론에서는 어휘 표현이 어휘개념을 정확히 구성할 수 있도록 충분한 정보를 포함해야 한다는 전제조건이 있다. 그러나 이러한 전제조건은 충족되기 어렵고 대부분의 사전에 있는 정의(definition)에서도 충족되지

* 이 연구는 과학기술처의 STEP 2000 지원에 의한 것임

않는다 한편, 차별화 이론에서는 어휘개념을 개념별로 구별될 수 있는 상징(symbol)에 의하여 표현할 수 있다고 본다.

한국어 동사 Isa 계층구조의 기본 골격인 어휘개념은 차별화 이론에 의하여 두 가지 방법으로 표현될 수 있다. 첫째, 한국어 동사에서 동의어 집합을 구성하여 한국어 동사 어휘개념을 표현하는 것이다. 둘째, 영어 동사 WordNet의 어휘개념을 이용하여 한국어 동사 어휘개념을 표현하는 것이다. 첫째 방법은 국어학자, 언어학자 및 심리학자들과 공동으로 막대한 작업을 하여야 하는 것이므로 이 연구에서는 채택하지 못하였다. 이 연구에서는 둘째 방법을 기본으로 하여 한국어 동사 계층구조에서의 어휘개념을 표현하고 영어 동사에 없는 어휘개념은 국어사전, 시소러스 등을 참조하여 한국어 동사에 알맞은 어휘개념을 만들어 표현한다.

WordNet은 동사의 top node 분류가 체계적으로 되어 있지 않고 너무 많으며, 유사한 의미를 세분하여 다른 node로 만든 경우도 있고, 자동사와 타동사를 구분하지 않았다는 단점이 있다. 그리고 Levin verb class는 영어를 중심으로 하였으므로 한국어에는 맞지 않는 분류가 있다 그리하여 SENKOV의 한국어 동사 분류에는 Levin verb class를 한국어 동사에 맞게 수정하여 기준을 삼았으며 계층구조의 어휘개념 구축시 WordNet을 참조하였다[9]

3.2 SENKOV의 설계

SENKOV의 설계를 위한 추진전략은 다음과 같다. 3,000여 개의 동사를 49개의 class로 분류한 Beth Levin의 verb class를 근간으로 하였으며, Princeton 대학의 WordNet을 활용하였다. 우리말 용언 중 동사를 중심으로 추진하였으며, 동사의 구문론적인 면과 의미론적인 면의 상관관계를 동시에 고려하였고, 구문론적인 면은 동사의 논항구조와 하위 범주화 유형을 중심으로 연구하였다. top node의 기준은 Levin의 verb class로 하였으며, 중간 및 terminal node의 기준은 WordNet의 verb class와 Levin의 verb class로 하였다. 또한 술어의 원형을 중심으로 개념망

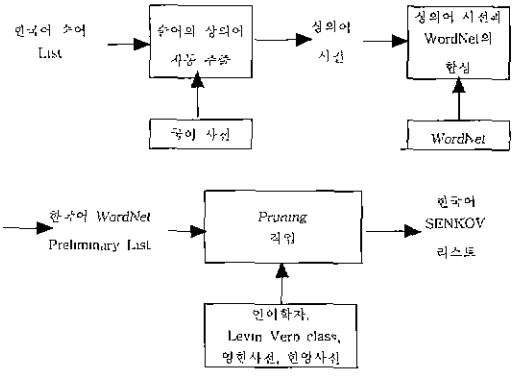

그림 1 SENKOV의 생성시스템

구축을 하였고, WordNet 자체의 문제점을 제거하였으며, 자동사와 타동사를 구분하였다

SENKOV의 전체시스템은 생성시스템과 DB구축시스템으로 구성되어 있다. 생성시스템의 순서도는 그림 1과 같다. DB구축시스템은 SENKOV의 생성시스템에 의해 출력된 리스트를 정렬하여 데이터베이스로 구축한다. DB구축시스템은 먼저 SENKOV 리스트의 파일을 모아서 정렬을 하고 그 다음 데이터베이스인 트리 파일과 인덱스 파일을 생성하는 것이다. SENKOV의 pruning작업 방법은 ISA hierarc-hy에 근거하며 Levin class의 분류가 한국어의 의미분류에 적합하지 않은 경우에 재분류를 다음과 같이 수행하였다. (삽입, 삭제, 이동)

1) 영어 WordNet은 자/타동사를 구별하지 않고 한 동사 내에 sense만 달리하여 기술하였 으나 한국어는 한 동사를 자/타동사로 겸하 여 쓰이는 경우가 드물기 때문에 각 lexical entry를 달리하였다.
2) Levin class에서 9.1과 9.2는 한국어의 경우에 의미상 거의 동일하여 따로 분류하지 않고 같은 class에 넣었다.
3) Spray verb의 경우 Levin class에서는 putting verb class에 속해있으나 WordNet에서는 putting verb class에 속하지 않아서 최상위 개념에 putting을 넣어주었다 (Hierarchy 수정)
4) Levin class 9.7 Spray/Load class 중에서 load를 9.2의 Levin class에 첨가하였다.
5) 한 동사의 Levin class의 hierarchy와 WordNet hierarchy를 비교하여, Levin class의 hierarchy를 그 동사의 Sense 1로 기술하고 나머지 WordNet sense는 Sense 2, 등으로 한국어 WordNet hierarchy를 구성하였다.

.
.
.

35) Levin class의 51.2는 한국어에서는 51과 의미상 관련이 없으므로 삭제하였다.

3.3 SENKOV의 구축

```
"깨끗이하다"
2 senses of clear
Sense 1
clear up, clear, light up, brighten
        => clean, clear, drain, empty
Sense 2
clear, become fair, become sunny
    => change state, turn, fail
        => change, undergo a change, become different
```

그림 2. SENKOV의 구축 예

3.2에서 기술한 설계방법에 의하여 SUN Workstaion (Ultra Sparc 1)을 사용하여 C와 C++언어로 SENKOV를 그림 2와 같이 구축하였다 일반적으로 많이 쓰이는 한국어 동사 658개를 대상으로 한 결과, SENKOV는 44개의 top node를 가지고 있으며 계층구조의 평균 depth는 약 2.35이

었다 한국어 동사의 semantic network은 영어에서와 마찬가지로 명사보다 top node의 개수는 많고 depth가 훨씬 더 얕았다. 표 1은 한국어 명사와 동사의 top node수와 계층구조의 평균 depth를 비교한다.

표 1 품사별 계층구조의 평균 depth

한국어 품사	top node의 수	계층구조의 평균 depth
명 사	11	5.40
동 사	44	2.35

4. 부사의 공기제약 관계에 유효한 개념 분류

한국어 부사와 동사간의 공기제약 관계 설정에 유효한 개념 분류를 하는데 SENKOV를 활용하였다.

한국이 부사는 크게 특정한 성분을 수식하는 성분 부사와 문장 전체를 수식하는 문장 부사로 나뉘어진다[7]. 성분부사는 동시, 형용사 및 명사를 수식하는 성상부사외 의성어, 의태어를 표현하는 상징부사 그리고 지시부사를 포함한다. 문장부사는 양태부사와 접속 부사를 포함하는데, 이 부사는 문장과 관계가 있고 동사의 공기제약과는 관계가 없으므로 이 연구에서는 논외로 한다 그리고 성분부사 중 지시부사도 동사의 공기제약과는 관계가 없으므로 이 연구에서는 논외로 한다

성상부사에 "홱", "세게", "잘게", "잘", "높이", "빨리" 등이 속하는데, 이 부사와 동사간의 공기제약에 유효한 개념 분류를 하는데 SENKOV를 다음 예와 같이 활용할 수 있다 여기에서 나오는 숫자는 SENKOV의 verb class를 표시한다.

예) 홱 + (12, 9.5, 9.7, 10.1, 10.2, 10.7, 11.2, 17.1, 18.1, 18.4, 21, 26.6, 43.4, 45, 51.1, 51.3, 51.4, 57 중 "붙다")
세게 + (18.1, 9.3 ~ 9.7, 10.4, 10.7, 11.2, 12, 15.1, 17, 18, 21, 22, 38, 40.1 ~ 40.3, 45, 51.3, 57)
잘게 + (21, 10.7 중 "빗기다")

상징부사에 "땡땡", "도란도란", "까옥까옥", "데굴데굴" 등이 속하는데, 이 부사는 대부분 한 두 개의 동사와 공기제약 관계를 형성하므로 SENKOV의 활용이 큰 도움을 주지는 않는다

예) 땡땡 + (18.1 중 "치다")
도란도란 + (37.1 중 "이야기하다".
37.2,
37.3 중 "속삭이다")

5. 결론

이 연구는 3.2에서 기술한 설계 방법에 의하여 SUN Workststion(Ultra Sparc 1)을 사용하여 C와 C++언어로 일반적으로 많이 쓰이는 한국어 동사 658개를 대상으로 하여 SENKOV를 구축하였다 SENKOV는 44개의 top node를 가지고 있으며 depth는 약 2.35이었다 한국어 동사의 semantic network은 영어에서와 마찬가지로 명사보다 top node의 개수는 많고 depth가 훨씬 더 얕았다. SENKOV의 활용분야는 다음과 같다. 기계 번역 시스템 (word-sense ambiguity, phrase attachment problem), 기계 이해 시스템, 정보 검색 시스템, 문장 검사/교정 시스템, 자연어 인터페이스 시스템 그리고 질의어 처리 등이다[1,6]

향후 연구계획은 다음과 같다 첫째, 3,000여개의 한국이 동사에 대한 SENKOV를 구축하여 완성하는 것이다 둘째, 주어와 동사, 목적어와 동사간의 selectional restriction에 유효한 SENKOV를 구축하는 것이다 셋째는 위에서 제시한 활용분야에 효율적으로 활용하는 것이다

6. 참고 문헌

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

[2] Levin, B, English Verb Classes and Alterations - A Preliminary Investigation, The MIT Press, 1997

[3] Levin. B. and Hovav, M., Unaccusativity At the Syntax-Lexical Semantics Interface, The MIT Press, 1996

[4] Miller, G A., Beckwith, R, Fellbaum, C. Gross, D. and Miller, K., "Introduction to WordNet An On-line Lexical Database," in Five Papers on WordNet, CSL report, Cognitive Science Laboratory, Princeton University. 1993

[5] Montemagni, S and Vanderwende, L., "Structural Patterns vs String Patterns for Extracting Semantic Information from Dictionaries," Proc. of COLING-92, pp.546-552, Aug. 1992.

[6] Sumita, E., Furuse, O., and Iida, H., "An Example-Based Disambiguation of Preposition Phrase Attachment", Proceedings of TMI, pp.80-91, 1993

[7] 남기심, 고영근, 표준 국이 문법론, 탑출판사, pp 169-174, 1985.

[8] 문유진, 김영택, "한영기계번역에서 개념기반의 동사 번역", 한국정보과학회 논문지, 제22권 제8호, pp.1166-1173, 1995

[9] 문유진, 의미론적 어휘개념에 기반한 한국어 명사 WordNet의 설계와 구축, 서울대학교 대학원 컴퓨터공학과 박사학위 논문, 1996.

[10] 이정민, 강범모, 남승호, "한국이 술어의 의미구조연구", 제2회 소프트과학워크숍 학술회의, 1997