인지과학, 제19권 제2호
Korean Journal of Cognitive Science
2008, Vol. 19, No. 2, 143~161.

코퍼스를 이용한 상하위어 추출 연구*

방 찬 성† 이 해 윤
한국외국어대학교 언어인지과학과

본 논문에서는 코퍼스를 이용하여 어휘들의 상하위 관계 패턴들을 추출하는 방법을 제안한다. 기존 연구들에서는 어순 교체가 자유로운 한국어의 특성으로 인해 주로 사전의 정의문을 이용하여 어휘들의 의미관계 패턴들을 추출하는 방법을 취하고 있으나, 본 논문에서는 코퍼스를 이용하여 보다 다양한 의미관계 패턴들을 추출하여 제시하고자 한다. 이를 위해 먼저 기존의 사전들을 이용해 상하위어 쌍들의 목록을 선정하였다. 다음 이 목록의 어휘 쌍들을 포함하는 문장들을 코퍼스에서 추출한 이후, 이로부터 다시 체계적으로 패턴화 할 수 있는 문장들을 추출하여 21 가지 상하위 관계 패턴들로 일반화하였다. 21가지 패턴들을 정규식으로 표현한 뒤 각각 동일한 패턴들을 가진 문장들을 코퍼스에서 다시 추출한 결과 57%의 정확률이 측정되었다.

주제어 : 상하위 관계, 의미 관계, 상하위어, 패턴 추출, 코퍼스

---

* 2007학년도 한국외국어대학교 교내학술연구비의 지원을 받음.
† 교신저자: 방찬성, 한국외국어대학교 언어인지과학과, 연구분야: 전산언어학
E-mail: pcsalice@hufs.ac.kr

한 언어에서 어휘들은 다양한 의미 관계들을 맺고 있으며, 연구 목적에 따라 그 유형들은 다양하게 제시되어 왔다. 예를 들어 동의 관계, 반의 관계, 상하위 관계, 부분-전체 관계 등은 일반적으로 가정되는 의미 관계들이며, 이들 중에서 상하위 관계와 부분-전체 관계는 일종의 계층 구조를 보여준다는 점에서 계층적 관계로 다시 그룹지어진다. 이러한 계층적 관계는 온톨로지 구축 등에 직접적으로 활용되고, 그 외 기계 번역, 정보 검색, 데이터 마이닝 등의 응용 분야에서 사용될 수 있는 중요한 의미 관계로 파악되어 왔다.

영어를 대상으로 하여 계층적 관계를 자동 추출하는 기존 연구들은 주로 어휘-통사 정보를 이용한 패턴을 사용하여 코퍼스로부터 관계들을 추출해내었다(Berland & Charniak 1999; Cederberg & Widdows 2003; Girju et al. 2003; Hearst 1992). 그러나 한국어의 경우 영어에 비해 어순이 자유롭기 때문에 이와 동일한 방법으로 상하위어를 추출하기에는 어려운 점이 있다. 따라서 한국어에 대한 기존 연구들에서는 코퍼스보다 정형화된 문장 구조를 보이는 사전의 정의문을 이용해 왔다(옥철영 2002; 최선화 2006).

그러나 본 연구에서는 영어권의 기존 연구 방법론을 토대로 하여 코퍼스로부터 상하위 관계를 추출하는 방법을 수정, 제시하고자 한다. 비록 코퍼스가 사전에 비해 단어 사이의 의미 관계를 정의하는 패턴의 출현 빈도가 적을 수도 있겠지만, 사전보다는 보다 다양한 어휘들과 문형들이 출현할 가능성이 높으므로 다양한 유형의 상하위 관계들을 추출할 것으로 보이기 때문이다.

본 논문의 구성은 다음과 같다. 먼저 기존 연구들을 개괄해보도록 한다. 여기서는 영어를 대상으로 한 계층 관계 추출 방법들을 제시한 대표적인 몇 개의 연구들을 살펴보기로 한다. 다음으로 본 논문에서 제안하는 상하위 관계 패턴 추출 과정과 그 결과를 제시한다. 끝으로 실험과정에서 나타난 오류들에 분석과 향후 연구 과제를 제시하기로 한다.

기존 연구

코퍼스에서 의미 관계를 추출하기 위해 어휘통사 패턴을 이용한 관련 연구로

Hearst(1992), Girju et al. (2003) 등이 있고, 상하위어의 공기 관계를 이용하여 패턴으로 추출한 연구로 Verginica(2006) 등이 있다.

Hearst(1992)

텍스트로부터 상하위 의미 관계를 자동적으로 추출하기 위해서는 패턴 인식과 의미 관계를 이용하는 것이 구문 분석된 결과를 이용하는 것보다 정확하고 효과적이라는 가정 하에, Hearst(1992)에서는 몇 개의 어휘-통사적 패턴들을 (lexico-syntactic patterns) 설정한다. 이러한 패턴들을 토대로 하여 상하위 관계들을 추출하였다. 다음은 NP 사이의 상하위 관계를 추출하는 데 이용된 하나의 패턴을 제시하고 있다.

(1) a. NP_0 such as $\{NP_1 , NP_2 ..., (and|or)\}$ NPn

b. such NP as $\{NP,\}^*$ $\{(or|and)\}$ NP

이러한 방법은 무한한 텍스트를 대상으로 할 경우 순전히 구조에 의존하는 추출방법보다도 보다 정확히 상하위어들을 예측할 수 있다고 본다.

Hearst(1992)은 텍스트에서 추출된 어휘 패턴들을 사용하여 자동으로 의미 관계를 추론하는 방법들의 전형으로서 이후 연구들에 기초가 되었다는 점에서 의의가 있다. 그러나 환유 표현이 나타나거나 자질 미명세화(underspecification)된 표현이 나타날 때 혹은 문맥과 관점에 의존하는 표현들이 나타날 경우 이를 처리하기 위해서 수작업이 필요로 하며, 또한 이후의 작업을 위해서는 워드넷과 같은 언어 자원 등에 의존해야 한다는 문제점이 있다.

이러한 Hearst (1992)의 방법론을 사용하여 의미 관계를 추출한 연구로 Berland & Charniak(1999)와 Cederberg & Widdows(2003)를 들 수 있다. Berland & Charniak(1999)에서는 'North American News Corpus(NANC)' 코퍼스로부터 부분-전체 관계를 추출하였고, 그 결과는 55%의 정확률을 보여주었다. Cederberg & Widdows(2003)에서는 Hearst(1992) 방법론에 의미적 분석 방법을 가미하여 상하위 관계의 자동 추출에서 정확률과 재현율을 향상시켰다.

Girju et al.(2003)

Girju et al.(2003)에서는 코퍼스로부터 자동으로 부분-전체 관계를 추출하는 방법을 제시하고 있다. 먼저 부분-전체 관계를 표현하고 있는 어휘-통사 패턴을 찾기 위해 Hearst(1992)의 알고리즘을 적용하였고, 다음으로 워드넷의 의미 관계와 Winston et al.(1987)의 의미 분류를 고려하여 패턴을 추출하였다. 실험 결과로는 20,000 개의 문장 가운데 535 개가 부분-전체 관계였으며, 그 중 구 차원의 패턴이 493 개, 문장 차원의 패턴이 42 개로 나타난다. 또한 중의적인 패턴들이 나타나는 문제를 해결하기 위하여 다음과 같이 모든 패턴들을 세 유형으로 분류하였다.

(2) a. Positive example

<X_hierarchy#sense; Y_hierarchy#sense; Yes>

b. Negative example

<X_hierarchy#sense; Y_hierarchy#sense; No>

c. Ambiguous example

<X_hierarchy#sense; Y_hierarchy#sense; Yes/No>

중의적인 유형의 경우, 워드넷의 명사 IS_A 계층구조를 이용하여 해당 어휘의 하위어로 대체시킨다. 중의적 유형에 해당되지 않을 때까지 이러한 과정을 반복 적용하여 중의적 패턴들을 처리하였다. 이 방법의 실험 결과, TREC-9 LA Times 뉴스 기사들을 대상으로 한 평가에서 83%의 정확률과 98%의 재현율을 보였다.

Verginica(2006)

Verginica(2006)는 코퍼스에서 상위어와 하위어가 공기하는 패턴들을 찾아 상하위 관계의 패턴으로 확정하는 방법을 제안하였다. 먼저 상하위 관계 패턴들을 자동 추출하기 위해 'British National Corpus(BNC)' 코퍼스에서 명사와 동사 위주로 한 하위어가 직접적인 또는 간접적인 상위어와 함께 출현하는 문장을 모두 추출하였다. 그리고 상위어와 하위어 사이의 통합적 거리(syntagmatic distance)에 따라 추출된 문

장을 그룹지운 후, 단어의 출현 순서에 따라 그 구조를 in_paticular, including (the), especially (the), for_example (the) 등과 같은 표현을 중심으로 다시 그룹화 시켰다. 그러나 'waters particularly the reservoirs', 'rats and sometimes other creatures'과 같은 예들에서 보듯이 관사나 부가어 등의 존재 여부에 따라 추출된 구조가 완전히 맞지 않는다는 문제점이 있다.

패턴 추출 실험

본 절에서는 '21세기 세종계획 균형 말뭉치' 가운데 색인된 문어 코퍼스 900만 어절로부터 상하위 관계를 추출하는 방법을 설명한다.[1] 사전들을 이용하여 상하위어 쌍의 목록을 설정하고, 이를 이용하여 코퍼스로부터 상하위 관계 패턴들을 추출하여 일반화한 몇 개의 패턴들을 확정한다. 그리고 이 패턴을 가지고 코퍼스에서 이와 같은 패턴들을 가진 문장들을 모두 추출하여 패턴의 정확률을 살펴본다. 끝으로 전형적인 상하위 관계를 나타내는 패턴들을 분석하고 패턴으로는 포착할 수 없지만 상하위 관계를 나타내는 표현들도 분석한다.

추출 실험은 (그림 1)에서 제시하는 바와 같은 단계를 밟아 이루어진다.
즉 세종전자 사전에서 상하위어 목록을 추출하고, 코어넷을 이용하여 이 목록을 보충한다.[2] 다음으로 추출된 상하위어 목록의 어휘 쌍들을 포함한 문장들을 코퍼스로부터 추출한다. 그리고 추출된 문장들의 구조를 조사하여 유사한 구조를 가진 문장들을 그룹화하고 이를 일반화하여 하나의 패턴을 확정한다. 이를 통해 21 개의 패턴을 얻을 수 있었다. 다음으로 이 패턴들을 정규식으로 표현하고, 이 정규식 표현의 패턴들을 사용하여 코퍼스로부터 문장들을 추출하였다. 단계별로 자세히 살펴보면 다음과 같다.

---

1) 『21세기 세종계획 균형 말뭉치』, CD-ROM (국립국어연구원, 2002).

2) 세종전자사전: http://www.sejong.or.kr

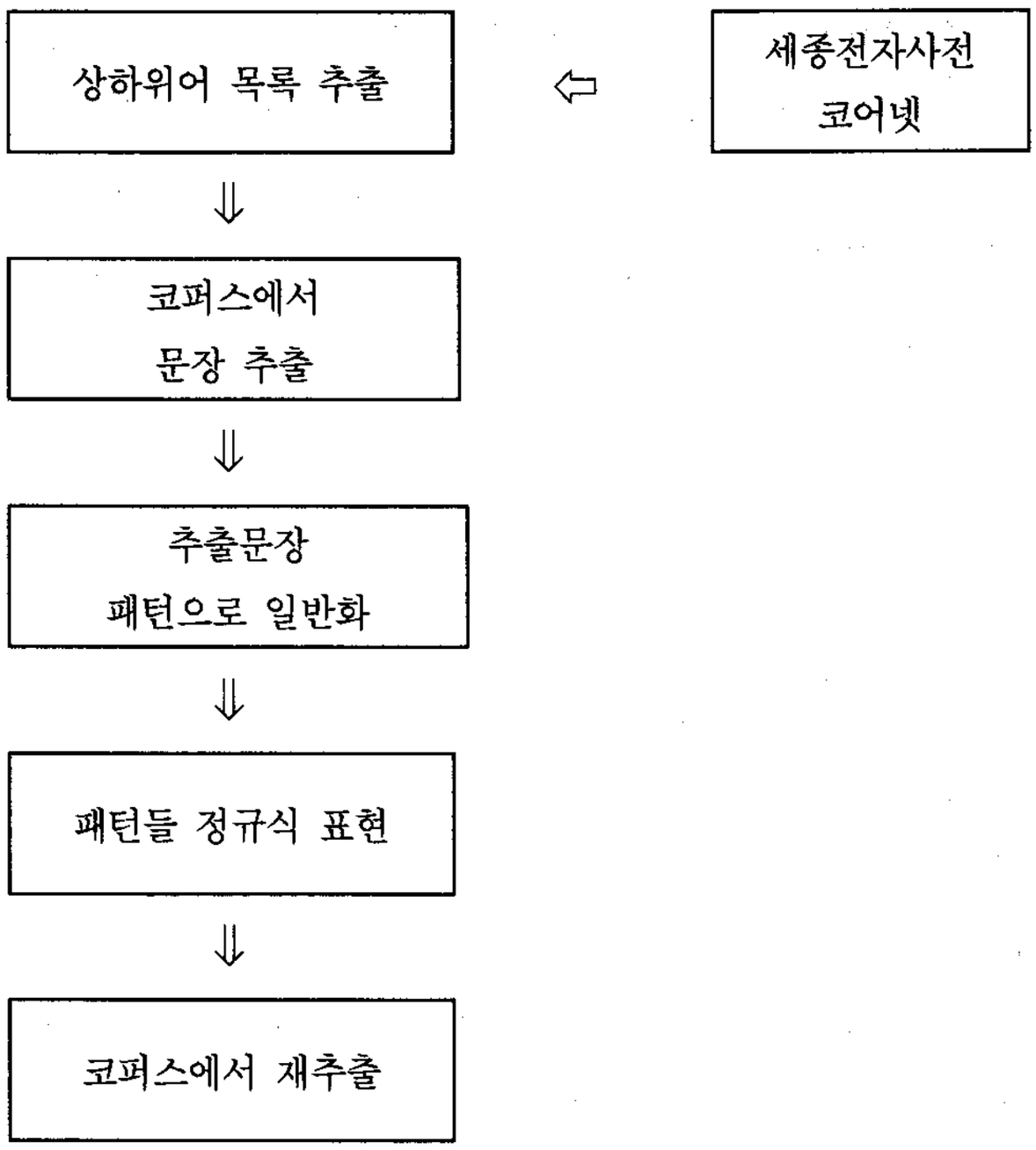

그림 1. 상하위 관계 추출 방법

패턴 추출

먼저 세종전자 사전을 이용하여 "hyper, hypo"로 태깅된 상하위 관계 어휘들의 쌍을 모아 수작업을 통하여 정제한 뒤 80 개의 목록으로 만들었다. <표 1>은 세종전자 사전에서 추출한 명사 상위어-하위어 쌍들 중 일부 예이다.

표 1. 세종전자 사전에서 추출한 상하위어 목록

상위어	하위어
가구	옷장, 침대, 의자, 화장대
감기	독감, 목감기, 코감기
교통기관	기차, 지하철, 버스, 자동차

표 2. 세종전자 사전과 코어넷의 상하위어 비교

<세종전자 사전>		<코어넷>	
상위어	하위어	상위어	하위어
꽃	코스모스, 백합, 장미	꽃	장미
		화초/들풀	코스모스, 백합
등	호롱불, 스탠드	불	호롱불
		등	스탠드

그러나 세종 전자사전에서 추출한 목록은 상하위어 대응의 정확성이 떨어지는 경우들이 있다. 이런 경우 코어넷의 정보를 이용하여 보다 정확한 상하위어 쌍들로 수정, 대치하였다. <표 2>는 두 사전의 상하위어 정보에 있어서 차이가 나는 몇 가지 예들을 보여주고 있다.

다음으로 앞서 선정한 목록을 가지고 세종 색인 코퍼스로부터 관련 문장들을 추출하였다. '글잡이 2' 프로그램을 사용하여 목록에 있는 어휘 쌍을 포함한 문장들을 각각 1,311 개, 692 개로 추출하였다. 그 중 중복되거나 패턴화 할 수 없는 문장들을 제외하면 총 73 개의 문장으로 정리된다. 그리고 추출된 문장들을 토대로 하여 <N1-나/이나 N2 등의 N>, <N1, N2, N3, N4 등의 N> 와 같이 POS를 사용한 패턴으로 일반화하여 최종적으로 21 개의 패턴들을 얻을 수 있었다.[3] 다음은 각 패턴들과 이에 해당하는 추출 문장들을 제시한다.

(2) a. 패턴 1. N1-나/이나 N2 등의 N

보통 소나무나 전나무 등의 상록수에 아름다운 장식물이나 ...

b. 패턴 2. N1, (N2,)* Nn 등의 N

문학은 과장, 수사, 비유 등의 기법을 사용해서 ...

c. 패턴 3. N1, (N2,)* Nn 따위의 N

이 커다란 질문에 대한 가장 간단한 대답은, 소설, 시, 희곡 따위의 글이

3) 보다 구체적으로는 세종전자 사전의 목록을 통하여 15 개 패턴이, 코어넷을 이용하여 6 개 패턴이 추출되었다.

라는 것이다.

d. 패턴 4. N1-나/이나 N2 등 N

이쑤시개는 대체로 밥그릇이나 반찬 그릇 등 식기 속에 방치되기 일쑤다.

e. 패턴 5. N1, (N2,)* Nn 등 N

처음엔 단맛, 짠맛, 신맛 등 맛의 구별이 분명한 것부터 시작해 ...

f. 패턴 6. N1-와/과 (N2,)* Nn 등 N

전체수심은 얕은 편이며 연안에 갈대와 말풀 등 수초가 많다.

g. 패턴 7. N1, (N2,)* Nn 등 온갖 N

몸무게가 늘면 당뇨병, 심장병, 고혈압 등 온갖 병의 위험도 늘어나기 때문이다.

h. 패턴 8. N1, (N2,)* Nn 등 각종 N

8-1. N1· (N2 ·)* Nn 등 각종 N

8-2. N1 (N2)* Nn 등 각종 N

맥주, 소주, 양주 등 각종 술을 종류별로 갖다 놓고 ...

i. 패턴 9. 대표적 N인 N1

환경오염으로 발생되는 대표적 암인 폐암은 미국의 경우 25년 전에 ...

j. 패턴 10. N1-와/과 N2 같은 N

앞으로는 상하수도 요금과 버스요금 같은 공공요금이 인상용 티켓을 사들고 대기하고 있는 형편입니다.

k. 패턴 11. N1-나/이나 N2 같은 N

어떤 애는 농민들에게 옷가지와 크림을 주고 마늘이나 콩 같은 농작물을 받아 집에 특산물로 가져간다.

l. 패턴 12. N1, (N2,)* Nn 같은 N

동글동글한 꽃잎이 뭉친 듯한 모양에 빨강, 파랑 같은 원색을 쓴 오색 구름을 휘황찬란하게 그리는 것이다.

m. 패턴 13. N에는 N1, (N2,)* Nn 등이

떡에는 모시잎송편, 만경떡, 쑥굴레, 칡떡 등이 유명하다.

n. 패턴 14. N인 N1, (N2,)* Nn 등

취악기인 태평소, 나발, 나각 등과 타악기인 징, 북, 장구, 바라 등으로 구성되어 웅장하고 경쾌한 장단을 빚어낸다.

o. 패턴 15. N1 · (N2 ·)* Nn 등의 N

첫째, 물수리 · 독수리 · 매 등의 조류는 식량 자원으로 가치가 없다.

p. 패턴 16. N1-와/과 여타 N

우리뿐만이 아니라 일본, 중국이 쌀과 여타 곡물을 지원하고 ...

q. 패턴 17. N1-며/이며 (N2-며/이며)* Nn 같은 N

도리천에서는 그리고 또 천도복숭아며 도리능금이며 제석자두 같은 실과들이 온갖 육신의 질병과 괴로움을 없이 해줄 것이었습니다.

r. 패턴 18. N1, (N2,)* Nn-와/과 같은 N

수박색, 벽돌색과 같은 색깔이 유행될 것으로 보이고 ...

s. 패턴 19. N1 · (N2·)* Nn 등 N

텐트 · 침낭 · 버너 · 고무보트 등 텐트용품 일체를 묶어 판매하는 데 따른 소비자 피해 사례도 많다.

t. 패턴 20. N1, (N2,)* Nn 등 .* 종류의 N

이 밖에도 사과, 수박, 바나나 등 여러 종류의 과일이 많다.

u. 패턴 21. N1 · (N2·)* Nn-와/과 같은 N

알칼리성 식품에는 채소 · 과일과 같은 식물성 식품과 우유가 포함된다.

이러한 패턴에 대한 정확률은 <표 3>에 제시하고 있다. 이러한 패턴들은 정규식으로 변환되어 코퍼스에 다시 적용됨으로써 각 패턴들과 동일한 형식을 가진 문장들을 추출하였다.

표 3. 각 패턴의 정확률

패턴 번호	해당 문장 수	추출된 문장 수	정확률 (%)
1	58	105	55.2
2	380	718	52.9
3	30	41	73.2
4	101	168	60.1
5	394	626	62.9
6	234	403	58.1
7	7	10	70.0
8	86	107	80.4
9	13	17	76.5
10	34	89	38.2
11	114	254	44.9
12	91	229	39.7
13	16	30	53.3
14	21	26	80.8
15	78	182	42.9
16	3	4	75.0
17	3	3	100.0
18	60	92	65.2
19	470	722	65.1
20	6	7	85.7
21	111	220	50.5
합계	2,310	4,053	57.0

패턴 분석

전형적인 상하위어 패턴 분석

전형적인 상하위 관계 패턴은 다음과 같은 과정을 거쳐 확정되었다.

(3) a. 2310 개 상위어 중 5 회 이상 출현하는 고빈도 상위어 49 개 추출
b. 3 회 이상 서로 다른 패턴에 출현하는 상위어 42 개 선정
c. 42 개 상위어가 출현하는 패턴 분석, 각 패턴 별 상위어가 5 회 이상 출현하는 21 개 패턴 추출
d. 21 개의 패턴들 중 13 개의 패턴들로 압축
e. 13 개의 패턴들의 재분류하여 최종 3 개 패턴으로 확정

각 단계별로 자세히 살펴보면 다음과 같다. 먼저 전체 2,310 개[4] 가운데 5 회 이상 출현하는 고빈도의 어휘들을 목록으로 모은 뒤 21 개 패턴 가운데 가장 전형

표 4. 고빈도 상위어 목록

횟수	상위어
10회 이상	음식, 증상, 도시, 정보, 사람, 공업, 나라, 단체, 분야
9회	공공요금
8회	금융기관, 명절, 지하자원
7회	공공부문, 기관, 매체, 방법, 선진국, 시설, 식품, 은행, 작가, 편의시설
6회	거시경제, 과목, 동물, 사업, 예술, 인물, 재료, 행사, 협의, 회사
5회	가치, 곡물, 과일, 구조, 국가, 대도시, 대학, 물질, 부작용, 생선, 서비스, 성인병, 시설물, 조직, 품목, 활동

4) 상위어 <표 3>에서 제시한 바와 같이 전체 추출된 4,053개의 문장에서 2,310개의 문장들이 21개의 패턴들에 해당되었다. 2,310개의 문장들은 하나의 문장을 기본 단위로 하여 한 문장에 상위어는 하나가 있고 하위어는 여러 개 있다. 따라서 2,310개의 상위어라고 하여도 같은 의미이다.

적으로 나타나는 패턴을 분석하였다. 전체 2,310 개 상위어에서 5 회 이상 출현하는 고빈도의 상위어 49 개로 나타났다(표 4 참조).

다음 단계로 고빈도의 상위어 49 개를 가지고 추출된 21 개의 패턴 가운데 3 회 이상 서로 다른 패턴에 출현하는 빈도를 살펴보았다. 총 49 개 상위어 중 42 개의 상위어가 최소 3 개에서 최대 8 개의 패턴까지 다양하게 나타났다(표 5 참조).

표 5. 42 개 상위어가 갖는 패턴의 수

패턴 수	상위어
8개	분야
7개	성인병, 식품, 음식
6개	동물, 사업, 나라
5개	과일, 부작용, 서비스, 품목, 명절, 도시, 사람, 단체
4개	가치, 국가, 대도시, 생선, 시설물, 조직, 활동, 인물, 재료, 행사, 회사, 공공부문, 기관, 선진국, 시설, 작가, 편의시설, 공공요금, 정보
3개	곡물, 과목, 예술, 혐의, 방법, 은행, 금융기관, 증상

표 6. 패턴별 상위어의 출현 횟수

패턴 번호	42개 상위어의 출현 횟수	11	16
1	7	12	8
2	29	13	0
3	4	14	2
4	10	15	13
5	24	16	1
6	11	17	0
7	0	18	8
8	9	19	26
9	1	20	2
10	5	21	12

그리고 다음 단계에서 3 개 이상의 패턴을 갖는 42 개의 상위어에 대해 가장 많이 나타나는 패턴을 조사하였다. 패턴 2에 29 개의 상위어가 출현하였고, 패턴 19에 26 개의 상위어가 출현하였으며, 패턴 5에 24 개의 상위어가 출현하였다(표 6 참조).

그리고 42 개의 상위어를 가지고 21 개 각각의 패턴 중 어떤 패턴들에 중복되어 나타나는지를 살펴보았다. 한 패턴에 출현하는 횟수가 5 이상의 상위어들을 모아 패턴들을 살펴본 결과, 앞서 제시하였던 21 개의 패턴들이 13 개의 패턴들로 압축되었다(표 7 참조).

끝으로 위 13 개의 패턴들을 3 개의 패턴 유형으로 재분류하였다. 이와 같이 재분류된 패턴들은 전체 패턴들 중 95.7%를 차지하는 것으로 나타났다. 패턴 1, 2, 4, 5, 6, 15, 19번은 '등/등의'을 중심으로 다양한 변이형이 나타나 패턴 1로 일반화하

표 7. 추출된 13개의 패턴들

패턴 번호	패턴 유형
1	N1-나/이나 N2 등의 N
2	N1, (N2,)* Nn 등의 N
4	N1-나/이나 N2 등 N
5	N1, (N2,)* Nn 등 N
6	N1-와/과 (N2,)* Nn 등 N
8	N1, (N2,)* Nn N1 · (N2 ·)* Nn 등 각종 N N1 (N2)* Nn
10	N1-와/과 N2 같은 N
11	N1-나/이나 N2 같은 N
12	N1, (N2,)* Nn 같은 N
15	N1 · (N2 ·)* Nn 등의 N
18	N1, (N2,)* Nn-와/과 같은 N
19	N1 · (N2 ·)* Nn 등 N
21	N1 · (N2 ·)* Nn-와/과 같은 N

표 8. 패턴 1

N1, (N2,)* Nn	
N1 · (N2 ·)* Nn	
N1-나/이나 N2	등/등의 N
N1-와/과 (N2,)* Nn	

표 9. 패턴 2

N1, (N2,)* Nn	
N1 · (N2 ·)* Nn	
N1-와/과 N2	-(과/와) 같은 N
N1-나/이나 N2	
N1-며/이며(N2-며/이며)* Nn	

표 10. 패턴 3

N1, (N2,)* Nn	
N1 · (N2 ·)* Nn	등 각종 N
N1 (N2)* Nn	

였다(표 8 참조). 패턴 1은 2924 개 문장 중 1715 개 문장이 적합하여 정확률은 58.7%이며 전체 적합 문장 수 2310 개 가운데 1715 개가 해당되어 74.2% 가량이 패턴 1로 나타났다.

다음 패턴 10, 11, 12, 18, 21번은 '(과/와) 같은'을 중심으로 변이형이 나타나 패턴 2로 일반화하였다(표 9 참조). 패턴 2는 884 개 문장 중 410 개 문장이 적합하여 정확률은 46.4%이며 전체 적합 문장 수 2310 개 가운데 410 개가 해당되어 17.7% 가량이 패턴 2로 나타났다.

그리고 기존의 패턴 8번은 '등 각종'을 중심으로 변이형이 나타나 패턴 3으로 일반화하였다(표 10 참조). 패턴 3은 10 7개 문장 중 86 개 문장이 적합하여 정확

률은 80.4%이며 전체 적합 문장 수 2310 개 가운데 86 개가 해당되어 3.7% 가량이 패턴 3으로 나타났다. 그러나 패턴 3은 출현하는 횟수는 적으나 가장 높은 정확률을 보여준다.

기타 패턴

그 밖에 상하위어 목록에 나타나지 않아 패턴으로 포착하지 못하였으나 '종류, 즉, 일종, 한 가지' 등 상하위 관계를 명시적으로 나타내는 단어들로 인하여 다음과 같은 패턴들이 추출되었다.

(3) a. [N-의 종류에는/로는]
아동 문학의 종류에는 동화, 소년 소녀 소설, 동요, 동시, 희곡, 전기, 수필 등이 있다.

b. [N, 즉 N1, (N2,)* Nn]
보이저'우주선은 1호와 2호의 쌍둥이 우주선으로 태양계 외곽에 위치한 거대한 행성들, 즉 목성, 토성, 천왕성, 해왕성 탐사를 목적으로 발사되었다.

c. [N의 일종/한 가지]
[나의 살던 고향은]이라는 공해풀이 굿은 알다시피 마당극의 일종이다.

그리고 상하위 패턴들로는 포착할 수 없으나 최상급 표현이나 '나누다/분류하다/구분하다'의 표현을 가진 문장들 속에서 상위어와 하위어들이 많이 출현하였다. 예문 (4)는 이에 대한 예들을 보여주고 있다.

(4) a. 미국 유권자의 투표율은 1960 년 이래로 계속 하락세를 보여 왔는데, 실제로 미국은 세계에서 가장 낮은 투표율을 기록하고 있는 나라들 중 하나이다.

b. 또 유형물을 크게 나누면, 눈에 보이는 무생물 (無生物)과 생명이 있는 생물 (生物)로 구분된다.

오류 분석

패턴 추출 과정에서 정확률을 떨어뜨리는 오류들은 다음과 같이 나타났다.

첫째, 'N, N 등의 N' 패턴에서 상위어 자리에 나오는 명사가 반드시 상위어를 나타내지 않고 예측할 수 없는 어휘들이 나타낼 때가 있다.

(5) a. 윤동주와 같은 분도 계시고, 해방 뒤에는 김수영, 신동엽 등의 시가 있지요.
b. 그 뒤 베이컨은 계속해서 주로 외부적·사회적 문제, 이를테면 우정, 결혼, 논쟁, 여행 등의 문제를 썼다.

둘째, 상위어 자리에 '것, 곳, 놈' 등과 같은 의존 명사가 오는 다음과 같은 경우들이 있다.

(6) a. 플루트나 하프 같은 것은 너무 비싸지 않느냐고?
b. 거리에서 흔히 보는 다방이나 카페 같은 곳이 아니었다.

이외에도 다음과 같은 유형의 오류들을 들 수 있다. 먼저 열거를 나타낼 때 사용하는 다양한 접속 조사나 가운뎃점 등의 문장 부호의 변이로 인하여 하나의 패턴에 대해 여러 변이형들을 보여 고정된 패턴으로 포착하기가 어렵다. 그리고 같은 의미를 지닌 다양한 표현이 존재한다는 자연 언어의 특성상 다양한 어휘 패턴을 포착하는 것은 어렵다. 또한 문맥에 의존적인 어휘들이 나타날 때 어휘 자체만으로 상하위어 판단하는 것은 오류를 불러온다.

결 론

본 논문에서는 코퍼스 내 어휘들 간에 나타난 의미 관계들 중 상하위 관계에 한정하여 문장 패턴을 추출하는 과정 및 그 결과를 제시하였다. 의미 관계별로 추출

된 패턴들은 어휘들 간의 의미 관계에 대한 이론적 고찰에 기여할 수 있으며, 다양한 시스템에 응용 가능하다. 예를 들어, 정보검색에서 사용자가 필요로 하는 적절한 문서를 찾는데 도움을 줄 수 있고, 용례추출 프로그램에서 의미 분석을 하는데 활용될 수 있으며, 또한 온톨로지나 시소러스와 같은 개념망의 확장에 직접 이용될 수 있다.

그러나 오류 분석에서 제시한 바와 같이 본 논문에서 제시한 방법에는 몇 가지 해결해야 할 문제점들이 있다. 그럼에도 불구하고 본 논문에서 제시한 방법은 상하위 관계 이외의 다른 의미 관계들에 대한 자동 추출작업에 응용할 수 있으며, 이는 추후 과제로 남겨놓는다. 그리고 다양한 영역의 코퍼스들에 본 논문의 방식으로 의미 관계의 패턴들을 추출하여 비교하는 작업도 의미 있는 결과를 가져올 것으로 기대되며 이 또한 향후 연구 과제로 남겨놓기로 한다.

참고문헌

김광해 (1990), "어휘소간의 의미 관계에 대한 재검토", **국어학** 20, 28-46.

윤평현 (1995), "국어 명사의 의미 관계에 대한 연구", **한국언어문학** 35, 91-115.

옥철영 (2002), "수식관계 구문에서 공기 제약 어휘간의 정보량 측정", **한글** 255, 129-154.

조평옥, 옥철영 (1999), "의미속성에 기반한 한국어 명사 의미 체계", **정보과학회 논문지** 26:4, 584-594.

최선화 (2006), 사전 정의문의 구문 패턴에 기반한 상위어 판별 규칙 학습, 전남대학교 전자학과 박사학위논문.

한국과학기술원 전문용어언어공학연구센터 (2004), 어휘의미망 구축론, KAIST PRESS.

한정한, 도원영 (2005), "한국어 동사 의미망 구축을 위한 어휘의미 관계 유형", **한국어학** 28, 245-268.

황도삼, 최기선, 김태석 공역 (1999), **자연언어처리**, 홍릉과학출판사.

Berland, M. & Charniak E. (1999), Finding parts in very large corpora, in *Proceeding of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, 57-64,

University of Maryland.

Cederberg, S. & Widdows, D. (2003), Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction, in *Proc. of CoNLL-2003*, 111-118.

Cruse, D. A. (1975), Hyponymy and lexical hierarchies, *Archivum Linguisticum* 6, 26-31.

Cruse, D. A. (1986), *Lexical Semantics*, Cambridge University Press.

Fellbaum, C. (1998), *WordNet: An electronic Lexical Database*, MIT Press.

Girju, R., Badulescu, A., & Moldovan, D. (2003), Learning semantic constraints for the automatic discovery of part-whole relations, in *Proceedings of the 3rd Human Language Technology Conference/ 4th Meeting of the North American Chapter of the Association for Computational Linguistics Conference (HLT-NAACL2003)*, 80-87, Canada.

Hearst, M. A. (1992), Automatic Acquisition of Hyponyms from Large Text Corpora, in *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING-92)*, 539-545.

Hearst, M. A. (1998), Automated Discovery of WordNet Relations, in C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, 131-151, MIT Press, Cambridge, MA.

Jurafsky, D. & Martin, J. H. (2000), *Speech and Language Processing*, Prentice-Hall.

Miller, G. A. et al. (1990), Introduction to WordNet: An On-line Lexical Database, *International Journal of Lexicography* 3: 235-244.

Verginica B. M. (2006), Automatic Extraction of Patterns Displaying Hyponym-Hypernym Co-Occurrence from Corpora, in *Proceedings of First Central European Student Conference in Linguistics,* Budapest, Hungary.

Winston, M. E., Chaffin, R., & Hermann, D. (1987). A Taxonomy of Part-Whole Relations, *Cognitive Science* 11, 417-444.

1 차원고접수 : 2008. 5. 30
최종게재승인 : 2008. 6. 4

(*Abstract*)

A Study of the Automatic Extraction of Hypernyms and Hyponyms from the Corpus

Chanseong Pang Hae-Yun Lee

Dept. of Linguistics & Cognitive Science, Hankuk University of Foreign Studies

The goal of this paper is to extract the hyponymy relation between words in the corpus. Adopting the basic algorithm of Hearst (1992), I propose a method of pattern-based extraction of semantic relations from the corpus. To this end, I set up a list of hypernym -hyponym pairs from Sejong Electronic Dictionary. This list is supplemented with the superordinate-subordinate terms of CoreNet. Then, I extracted all the sentences from the corpus that include hypernym-hyponym pairs of the list. From these extracted sentences, I collected all the sentences that contain meaningful constructions that occur systematically in the corpus. As a result, we could obtain 21 generalized patterns. Using the PERL program, we collected sentences of each of the 21 patterns. 57% of the sentences are turned out to have hyponymy relation. The proposed method in this paper is simpler and more advanced than that in Cederberg and Widdows (2003), in that using a word net or an electronic dictionary is generally considered to be efficient for information retrieval. The patterns extracted by this method are helpful when we look for appropriate documents during information retrieval, and they are used to expand the concept networks like ontologies or thesauruses. However, the word order of Korean is relatively free and it is difficult to capture various expressions of a fixed pattern. In the future, we should investigate more semantic relations than hyponymy, so that we can extract various patterns from the corpus.

Keywords : hyponymy, semantic relation, hypernym-hyponym, pattern extraction, corpus