합성명사 의미해석용 사전 구축을 위한 워크벤치

이경순, 김도완, 최기선
전문용어언어공학연구센터, 첨단정보기술연구센터, 한국과학기술원
{kslee, kschoi}@world.kaist.ac.kr dwkim@csone.kaist.ac.kr

Workbench for Constructing Dictionary for Semantic Analysis of Compound Noun

Kyung-Soon Lee, Do-Wan Kim, Key-Sun Choi
KORTERM, AITRC, KAIST

요약

본 논문에서는 한국어에서 빈번하게 나타나는 합성명사의 의미해석을 하기 위한 워크벤치를 설계하고 구현하였다. 합성명사 의미해석을 위한 사전 구축 지원 워크벤치의 기능은 합성명사를 이루고 있는 명사와 명사가 어떠한 의미관계로 결합하고 있는지를 밝히기 위해서 의미관계 패턴을 정의한다. 정의된 의미관계 패턴을 이용하여 합성명사를 자동적으로 추출한다. 추출된 합성명사 사전을 이용해서 각 명사의 상위개념에 대해서도 의미관계를 반영시켜서 합성명사의 의미관계를 해석할 수 있도록 하는 환경을 제공하고 있다.

1 서론

한국어 문장에서 명사의 연결로 이루어지는 명사 연결구 또는 합성명사의 출현은 보편적인 현상이고, 그 생성 또한 비교적 자유롭다.

합성명사 사전은 합성명사를 이루는 각 명사들이 서로 어떠한 의미 관계로 연결되는지의 정보를 갖는다. 예를 들어, '모퉁이 가게'라는 명사 연결구의 경우 수식 명사인 '모퉁이'는 핵심 명사인 '가게'가 존재하는 위치를 알려준다. 따라서 '모퉁이 가게'는 수식 명사가 핵심 명사의 <위치>를 알려주는 의미 관계로 해석된다.

합성명사를 이루는 명사사이의 의미관계를 추출할 수 있는 방법은 문장에 나타난 의미관계패턴을 이용해서 가능하다. 즉, '모퉁이에 있는 가게'에서 '에 있는'과 같은 패턴을 이용해서 <위치>관계를 추출할 수 있다.

합성명사 해석을 위한 사전 구축을 지원하는 워크벤치는 명사와 명사가 결합할 때 나타날 수 있는 의미관계 패턴을 정의하고, 정의된 의미관계 패턴과 상위관계를 이용하여 합성명사를 자동적으로 추출하고, 의미관계를 해석할 수 있도록 하는 환경을 제공한다.

워크벤치에는 의미관계 패턴을 정의하기 위한 기능과 합성명사 의미관계 추출 기능이 있다.

의미관계패턴 정의기에는 코퍼스에서 명사와 명사 사이에 나타날 수 있는 의미관계패턴을 정의하고, 의미관계를 지정한다.

합성명사 의미관계 추출을 위한 기능에는 정의된 의미관계 패턴을 이용하여 합성명사의 의미관계를 추출하고, 추출된 합성명사 사전을 이용해서 각 명사의 상위개념에 대해서도 의미관계를 반영시켜서 합성명사를 추출한다. 따라서, 시소러스 분류체계 색인기, 어휘에 대한 상위개념 색인기, 합성명사 사전 색인기, 의미관계패턴 색인기능이 포함되어 있다.

2 의미관계 패턴

의미 관계(Semantic Relation) 정보는 합성명사를 이루는 단어와 단어 사이의 의미적 연관성을 나타내는 것

표 1. 합성명사의 의미관계 정보를 제공하는 의미관계 패턴

예제	의미관계 패턴	의미관계	합성명사
경찰이 명령하다	*이*	주체-행위 관계	경찰 명령
자동차를 운전하다	*를*	대상-행위 관계	자동차 운전
사용자를 위한 설명서	*를 위한*	목적 관계	사용자 설명서
구내에 위치한 식당	*에 위치한*	위치 관계	구내 식당
가죽으로 만든 가방	*으로 만든*	재료-물건 관계	가죽 가방
자동차의 부분인 바퀴	*의 부분인*	전체-부분 관계	자동차 바퀴
버스에 의한 사고	*에 의한*	원인 관계	버스 사고

이다. 의미관계를 이용한 합성명사의 해석에 관한 연구는 [1,2,3,4]에 나타나있다.

다음은 합성명사와 두 명사사이에 존재하는 의미관계를 나타낸다.

대통령 연설 <주어-서술어 관계>
자동차 운전 <목적어-서술어 관계>
아버지 지갑 <소유 관계>
모퉁이 가게 <위치 관계>
자동차 바퀴 <전체-부분 관계>
가죽 가방 <재료-물건 관계>.

의미관계에는 상·하위 관계(Hypernym), 목적 관계(Purpose), 위치 관계(Located-at), 시간 관계(Time-of), 원인·결과 관계(Caused-by) 등이 있다.

코퍼스에서 의미관계를 표현해줄 수 있는 패턴이 나타날 때 그것을 이용해서 합성명사의 의미관계를 파악할 수 있다. 예를 들어, "버스에 의한 사고"라는 내용에서 '에 의한'이라는 것을 통해서 '버스'와 '사고'로 연결되는 합성명사 '버스 사고'는 <원인 관계>를 갖는다는 것을 알 수 있을 것이다.

표1은 문장에서 합성명사를 이룰 수 있는 두 명사 사이에 의미관계 정보를 제공하는 패턴이 나타나는 것에 대한 의미관계를 나타낸 것이다[1].

3 합성명사 의미해석용 사전 구축을 위한 워크벤치

합성명사 사전 구축을 위한 워크벤치는 코퍼스를 이용해서 의미관계 정보를 제공할 수 있는 의미관계 패턴을 추출하고, 추출한 의미관계 패턴을 이용해서 합성명사 사전을 구축할 수 있는 환경을 제공한다.

3.1 합성명사 의미관계 패턴 정의기

합성명사를 구성하는 의미관계패턴을 정의하는 기능은 품사부착된 파일에 대해서, 합성명사를 구성하고 있는 명사와 명사의 의미관계를 나타내는 패턴을 추출하고, 의미관계를 정의한다.

가. 입력 파일

품사부착된 파일은 <숫자, 어절, 어휘/품사+어휘/품사 ... > 형태로 나타날 수 있는데, <숫자, 어절>은 필수 요소는 아니다.

나. 새로운 패턴 입력 창

품사부착된 문장에서 합성명사를 이룰 수 있는 부분이 되는 명사로 시작하고, 명사로 끝나면서 중간에 어떠한 패턴을 포함할 수 있도록 하는 패턴을 입력하여, 검색할 패턴 목록에 추가한다.

패턴의 예. 패턴에 나타날 수 있는 기호와 그 의미는 다음과 같다.

N1/n* */j* #4 N2/n*
N1 : 합성명사에서 앞에 나타나는 명사
N2 : 합성명사에서 뒤에 나타나는 명사
* : 아무거나 매칭 가능
/ : 어휘와 품사의 구분자 (예; 바보/ncn)
#숫자: #다음의 숫자 범위내에서 다음의 패턴이 나타나는 것을 처리

N2/n* */j* #4 N1/n*

N1과 N2의 순서가 위와 다른 경우는 수동태와 같은 형태로 이루어진 의미관계패턴으로 나타나는 합성명사의 순서를 고려하기 위한 것이다. 이 패턴으로 찾은 합성명사는 사전에 <N1 N2> 형태로 저장

검색할 패턴 목록에는 2 개이상의 패턴을 추가할 수

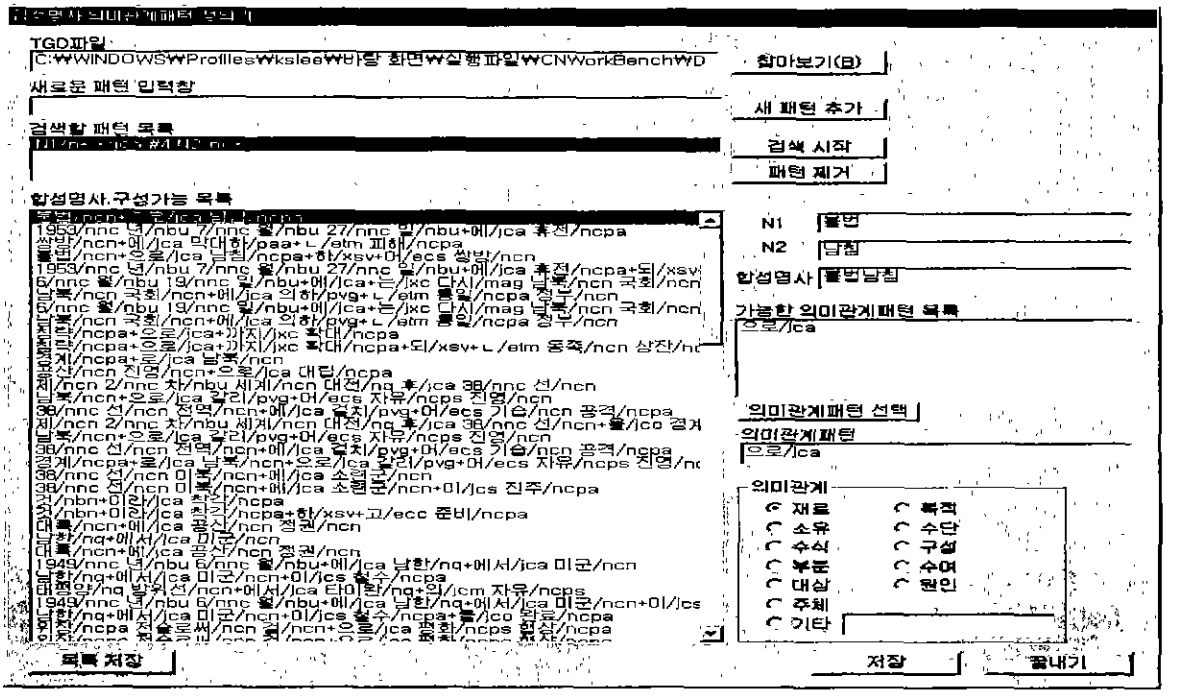

그림 1 : 의미관계 패턴 정의기

있는데, 2개중 하나라도 일치하면 합성명사 구성가능 목록에 추가가 된다.

다. 합성명사 구성가능 목록

품사부착된 코퍼스에서 검색할 패턴 목록에 있는 패턴에 일치하는 부분을 출력한다. 합성명사 구성가능 목록에 들어있는 부분에 대해서 의미관계패턴을 찾고, 의미관계를 정의하게 된다.

합성명사 구성가능 목록을 제시할 때, 명사류를 나타내는 품사인 n*에 대해서는, 연속되어있는 n*을 하나의 n*으로 결합한다. 현재의 위치에서 n*이면, 앞으로 거슬러가면서 n*이 아닌 다른 것이 나올 때 까지 합친다.

합성명사를 이루는 명사 각각이 두 개 이상의 명사가 결합될 수 있는 경우를 고려하기 위함이다. 즉, '북한동포 탈출'에서와 같이 '북한동포'는 '북한'과 '동포'로 이루어진 명사이다.

예: (n*: 북한동포)

북한/nq

동포/ncn+의/jcm

예: (n*: 1953년7월27일)

1953/nnc

년/nbu

7/nnc

월/nbu

27/nnc

일/nbu+에/jca

합성명사 구성가능 목록에 있는 것을 하나 선택하면, 앞의 명사를 N1에, 뒤의 명사를 N2에 출력하고 두 명사의 합성명사를 나타내준다. 그리고, N1과 N2사이에 나올 수 있는 가능한 의미관계 패턴을 나열한다.

예를 들어, 검색할 패턴 목록에 대한 합성명사 구성가능 목록이 다음과 같다고 하자.

N1/n* */j* #4 N2/n*

버스/ncn+에/jca 의하/pvg+ㄴ/etm 사고/ncn

이때, 가능한 의미관계 패턴 목록은 띄어쓰기 되어있는 단위로 가능한 형태를 나열한다.

에/jca

에/jca 의하/pvg+ㄴ/etm

이들 중에서 합성명사의 의미관계를 나타내는 의미관계 패턴인 '에/jca 의하/pvg+ㄴ/etm'을 선택하고, 의미관계에 해당하는 <원인> 관계를 선택한다.

합성명사 가능목록 부분이 다음과 같을 때,

A/ncn+B/jca C/pvg+D/etm E/ncn+F/jco G/ncps

가능한 의미관계 패턴 목록은 다음과 같다.

B/jca
B/jca C/pvg+D/etm
B/jca C/pvg+D/etm E/ncn+F/jco
B/jca E/ncn+F/jco

합성명사 의미관계 패턴 정의기에 나타나는 의미관계는 <재료-물건> 관계, <목적>관계, <원인>관계, <주체>관계, <위치>관계 등이다.

<재료-물건> 관계: 으로 만든, 으로 된, ...
예) 종이 <로 만든> 가방
<목적> 관계 : 을 위한, ..
예) 사용자 <를 위한> 메뉴얼
<원인> 관계 : 에 의한, 으로 인한, ..
예) 버스 <에 의한> 사고
<주체> 관계 : 이, 가, 의, ..
예) 대통령 <이> 연설
<위치> 관계 : 에 있는, 에 위치한,
예) 모퉁이 <에 위치한> 가게

의미관계 패턴 정의기에서 저장하는 정보는 합성명사 사전과 의미관계 패턴 정보이다.

라. 합성명사 사전의 저장 형식

합성명사 의미해석을 위한 사전의 저장형식은 명사1, 명사2, 그리고 의미관계 정보이다.

N1, N2 <의미관계>

다음은 저장되는 합성명사 사전의 예이다.

버스, 사고 <원인>
불법, 남침 <수단>
구내, 식당 <위치>
종이, 가방 <재료>

마. 의미관계패턴목록의 저장 형식

의미관계패턴 [N1/n* N2/n*] <의미관계>

여기서, N1/n* N2/n*는 패턴 추출시 적용된 패턴을 나타낸다. 만약, N1/n* */j* N2/ncps 의 패턴에 의해 추출된 목록일 경우는 저장되는 형태는 의미관계패턴 [N1/n* N2/ncps] <의미관계> 이다.

이/jcs [N1/n* N2/ncps] < 주체>
가/jcs [N1/n* N2/ncps] < 주체>

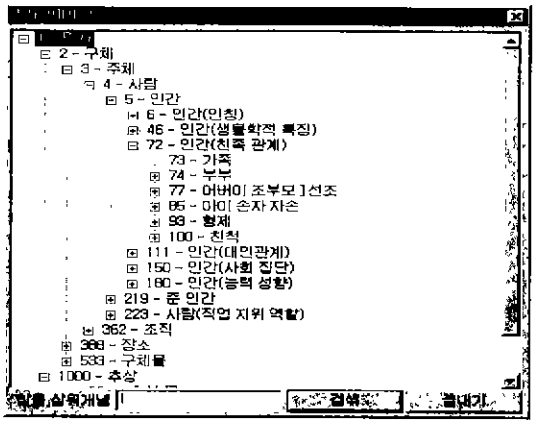

그림 2: 시소러스 분류체계 보기

의/jcm [N1/n* N2/n*] < 소유>
의/jcm [N1/n* N2/n*] < 수식>
에/jca 의하/pvg+ ㄴ/etm [N1/n* N2/n*]< 원인>
에/jca 의하/pvg+ ㄴ/etm [N1/n* N2/n*]< 수단>
으로/jca [N1/n* N2/n*] < 수단>
으로/jca [N1/n* N2/n*] < 위치>
을/jco 위하/pvg+ 어/ecs [N1/n* N2/n*]< 목적>

3.2 시소러스 어휘, 분류체계 색인기

명사에 대한 상위개념을 이용해서 합성명사의 의미관계를 추론하기 위해 각 명사에 대한 의미정보를 색인한다.

가. 시소러스 분류체계 색인기

워크벤치에서 이용한 시소러스는 KAIST시소러스를 번역한 것이다. 전체가 11개의 계층으로 되어있고, 분류체계의 항목은 2502개이다.

시소러스를 이용해서 어휘가 속하는 상위개념을 찾아서, 패턴의 일반화를 가능하도록 한다. 즉, '배추벌레'라는 고유명사에서 '벌레'앞에 나타나는 명사가 가지는 시소러스상의 개념이 '<식물>'인 것을 통해서, <식물>+'벌레'는 동물 분류에 속한다는 것을 얻기 위한 부분이다.

시소러스의 분류체계는 다음과 같이 정의되어 있다.

<개념번호> <개념> [<계층위치> <상위개념번호> < 하위개념 시작번호, 하위개념 끝번호>]

다음 시소러스 분류체계를 나타내는 일부 예에서, 2는 개념번호를 나타내고, '구체'는 개념이름, L은 계층구조에서의 단계, P는 현재 개념이 속하는 상위개념 번호, C는 현재 개념이 가지는 하위개념의 시작번호에서

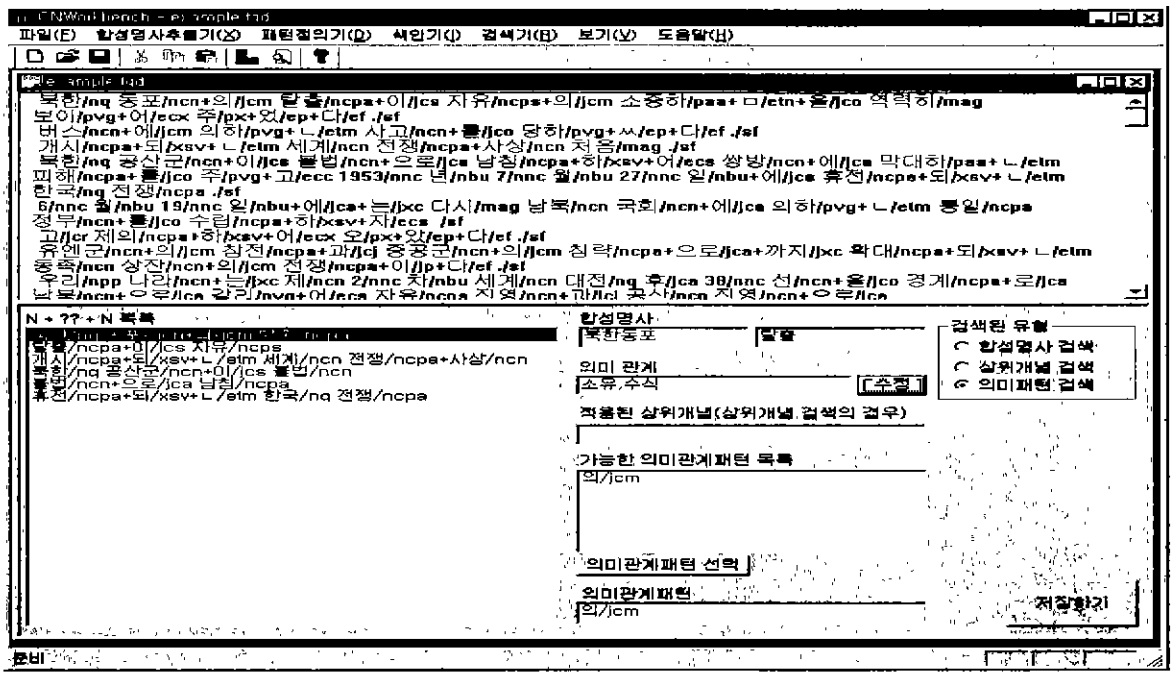

그림 3. 합성명사 추출기

끝 번호를 나타낸다.

1 명사[L 1, P -, C 2-2502]

2 구체[L 2, P 1, C 3-994]

3 주체[L 3, P 2, C 4-387]

4 사람[L 4, P 3, C 5-361]

5 인간[L 5, P 4, C 6-218]

...

1000 추상[L 2, P 1, C 1001-2502]

1001 추상물[L 3, P 1000, C 1001-1234]

1002 추상물(정신)[L 4, P 1001, C 1003-1153]

1003 지적 생산물(사고·학습)[L 5, P 1002, C 1004-1036]

1004 학문·학과[L 6, P 1003, C 1005-1006]

...

나. 시소러스의 어휘 색인

각 어휘에 대해 어휘의 상위개념을 색인한다. 상위개념은 분류체계에서의 번호를 나타내는 것으로, 하나 이상이 나타날 수 있다.

광고 <1542>

광명 <1363><2346>

광물 <714><723>

광물학 <1006>

우주 <462><526>

우주공간 <526>

우주비행사 <292>

3.3 의미관계 패턴의 색인기

의미관계 패턴정의기를 통해서 추출한 의미관계 패턴 목록이 들어있는 파일을 읽어서 색인한다.

입력형식은 앞에서 의미관계패턴목록의 저장형태와 동일하다. 즉, 어휘패턴 [N1/n* N2/n*] <의미관계>

색인기를 통해서 추가로 색인되는 의미관계패턴은 코퍼스에서 합성명사를 추출할 때 새로운 의미관계패턴을 적용시킬 수 있다.

색인의 키는 의미관계패턴이 되고, 데이터부분이 명사의 순서와 의미관계가 된다.

3.4 합성명사 사전 색인기

합성명사를 구성하는 각 어휘에 대해 색인을 한다.

합성명사 사전에 들어있는 각 합성명사에 대해서 색인은 다음과 같이 N1과 N2에 대해서 결합할 수 있는 위치와 의미관계를 색인한다.

합성명사 사전: N1, N2 <의미관계>

N1에 대해서 N2,r,<의미관계>
N2에 대해서 N1,f,<의미관계>

이때, r은 뒷부분에 결합, f은 앞부분에 결합을 나타낸다.

예를 들어서,

버스, 사고 <원인>
key가 버스, data가 사고, r,<원인>
Key가 사고, data가 버스, f,<원인>

색인된 어휘가 가지는 정보는

- 어휘의 상위개념
- 어휘의 앞/뒤에 결합될 수 있는 명사와 그때의 의미관계
- 어휘의 앞/뒤에 결합될 수 있는 명사의 상위개념과 의미관계

3.5 합성명사 추출기

합성명사 추출기는 품사부착된 파일에서, N + <의미관계패턴> + N을 찾아서 합성명사와 의미관계를 추출하고, 사용자의 피드백을 받는다.

품사가 n*으로 시작하는 N1, 그 다음 n*이 나오는 N2에 대한 합성명사 추출을 시도한다. 처음 n*과 다음 n*의 범위를 결정하기 위해 기본적으로는 5로 한다. 이때, 5는 5이내에 포함되는 것도 가능하다.

합성명사를 추출하는 방법은 다음 3가지에 의해서 이다.

가. 합성명사 사전 검색:

합성명사가 사전에 등록되어 있는 것인지 검색한다. N1에 대해 합성명사 사전을 검색해서, 결합가능한 어휘에 N2,r,<의미관계> 정보가 있으면, 성공, 합성명사 인식 결과 제시한다.

나. 합성명사의 상위개념 검색을 통한 추출

기본적으로는 상위개념을 2개 위까지 검색하여 합성명사를 추출한다. 상위개념 검색할 단계는 수정이 가능하다.

N1에 대해 합성명사 사전을 검색해서, 의미패턴에 N2의 상위개념,r,<의미관계> 정보가 있으면, 성공, 합성명사 추출 결과 제시한다. 없으면, N2에 대해 합성명사 사전을 검색해서, 의미패턴에 N1의 상위개념,f,<의미관계> 정보가 있으면, 성공, 합성명사 추출 결과 제시한다.

다. 의미관계 패턴을 이용한 추출

정의된 의미관계 패턴을 이용하여 합성명사를 추출한다. N1/n* N2/n* 사이에 있는 의미관계패턴을 의미관계패턴의 색인정보에서 검색한다.

예를 들어, 다음과 같은 N1/n* +?+ N2/n*목록에서

가죽/ncn+으로/jca
만들/pvg+ㄴ/etm
가방/ncn+을/jco

'으로/jca 만들/pvg+ㄴ/etm'을 key로 해서 검색했을 때, [N1/n* N2/n*] <재료> 가 data로 있으면, 합성명사 추출결과를 제시한다.

사용자 피드백은 합성명사, 의미관계, 적용된 상위개념, 의미관계 패턴 선택에 대해서 확인을 하고 수정을 한다.

추출된 합성명사에 대해서 저장할 합성명사사전을 입력받는다. 합성명사 사전의 파일을 cn-dict.txt로 하면, 의미관계패턴목록은 cn-dict_semrel.txt로 자동지정된다.

저장되는 정보의 형식은 합성명사사전은 입력받는 합성명사사전 형태와 동일하게 저장한다.
의미관계목록 또한 입력받은 의미관계목록과 동일하게 저장한다.

3.6 합성명사의 의미관계 검색기

합성명사 검색기는 입력된 어휘에 대해서 색인된 정보를 이용하여 합성명사를 검색한다.

합성명사 검색기는 합성명사 추출기의 사전과 상위개념을 이용한 검색과 동일하다.

가. 합성명사 사전을 이용한 의미관계 검색

명사가 사전에 등록되어 있는 항목인지 검색하여, 의미관계를 제시한다.

나. 합성명사에서 각 명사의 상위개념을 이용한 의미관계 검색기

명사의 상위개념을 찾아서 합성가능한지 검사하여, 의미관계를 제시한다. 상위개념의 몇 단계까지 적용을 시킬지 결정은 상위개념 단계를 통해서 재 지정할 수 있도록 하였다.

그림4에 나타난 합성명사 검색기의 예는 '버스사고'를 입력으로 했을 때, 합성명사 사전을 검색해서 '버스'와 '사고'라는 명사는 <원인>관계를 갖는다는 것을 나타내

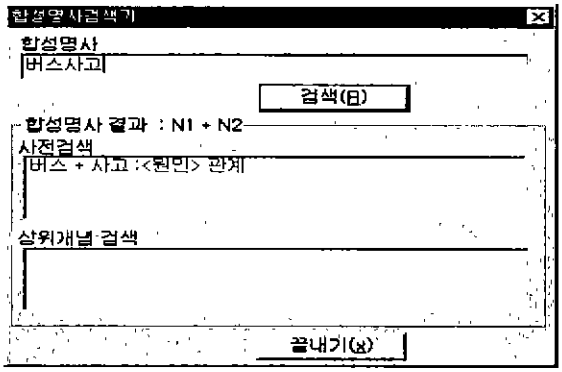

그림 4. 합성명사 의미관계 검색기 - 사전 검색

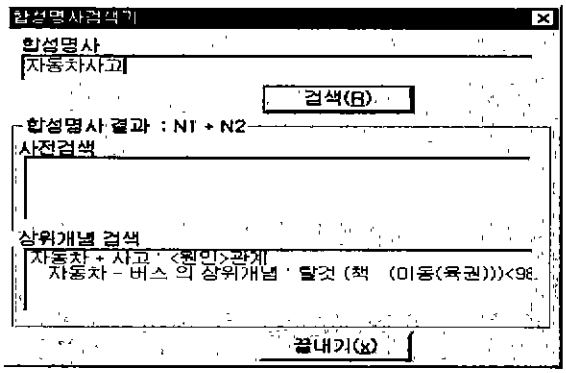

그림 5. 합성명사 의미관계 검색기- 상위개념 검색

고 있다.

그림5에 나타난 합성명사 검색기의 예는 '자동차사고'를 입력으로 했을 때, '사고'와 결합하는 '자동차'의 상위개념이 '사고'와 결합해서 합성명사 사전에 등록된 '버스'의 상위개념과 같다는 정보를 이용해서 '자동차'와 '사고'라는 명사는 <원인>관계를 갖는다는 것을 나타내고 있다.

4. 결론

본 논문에서는 한국어에서 빈번하게 나타나는 합성명사의 의미해석을 하기 위한 지식 구축 작업을 지원하는 워크벤치를 설계하고 구현하였고, 구축된 지식을 이용하여 의미해석을 수행하였다.

합성명사의 의미해석을 위한 사전 구축을 지원하는 워크벤치는 명사와 명사가 결합할 때 나타날 수 있는 의미관계 패턴을 코퍼스를 이용해서 정의한다. 정의된 의미관계 패턴을 이용하여 합성명사를 자동적으로 추출한다. 또한, 합성명사 사전과 합성명사를 이루는 명사의 상위개념을 이용하여 의미관계를 추론하도록 하는 환경을 제공하였다. 이렇게 구축된 합성명사의 의미관계 정보는 자연언어의 생성과정에서 이용될 수 있을 것이다.

감사의 글

본 연구는 전문용어언어공학연구센터에서 수행한 과학기술부와 KISTEP의 핵심소프트웨어사업 중 "대용량 국어정보 심층처리 및 품질관리 기술개발" 과제의 일환으로 수행되었으며, 부분적으로 첨단정보기술연구센터를 통하여 과학재단의 지원을 받았습니다.

참고문헌

[1] 김도완. 2000. MRD와 코퍼스를 이용한 명사 연결구의 의미해석. 석사논문, 한국과학기술원.

[2] 김도완, 이경순, 김길창. 1999. 의미관계와 문형정보를 이용한 복합명사 해석. 제11회 한글 및 한국어정보처리 학술대회.

[3] Richardson, S. 1997. Determining Similarity and Inferring Relations in a Lexical Knowledge Base. Ph.D. thesis, The City University of New York.

[4] Vanderwende, L. 1995. The Analysis of Noun Sequences using Semantic Information Extracted from On-Line Dictionaries. Ph.D. thesis, Georgetown University.