세종 의미 부류와 KorLex 명사 어휘 의미망 자동 맵핑

소길자○ 윤애선 권혁철
영산대학교○ 부산대학교 불어불문학과 / 정보컴퓨터공학부
kjso@ysu.ac.kr {asyoon, hckwon}@pusan.ac.kr

Automatic Mapping of Korean Wordnet 『KorLex』 to Semantic Classes of Sejong Dictionary

Gilja So○ Aesun Yoon Hyuk-Chul Kwon
Youngsan University○ Pusan University

인간이 가진 개념을 지식베이스화하려는 시도 중 하나로 의미망이 구축되고 있다. 한국어를 대상으로 한 어휘 의미망 중 프린스턴 대학의 WordNet을 대역한 KorLex는 1,2단계에서 한국어 어휘의미의 특성을 반영하여 개념 및 의미구조를 재구조화하고 있다. 그러나 현재 KorLex의 동의어 집합을 구성하는 어휘 의미에는 논항정보를 따로 구성할 수 없었다. 본 연구는 세종 전자 사전 격틀정보내의 선택제약조건(selectional restriction)으로 사용되고 있는 의미 부류와 KorLex의 명사 어휘 의미망을 자동 맵핑하는 방안을 제안함으로써 KorLex에서 세종 전자 사전 격틀정보를 활용할 수 있는 가능성을 제공한다.

주제어: 세종의미부류, 코렉스, 자동맵핑, 격틀정보

1. 서 론

인간이 가진 개념을 지식베이스화하려는 시도는 시소러스, 온톨로지, 의미망 등 다양한 형태로 나타나고 있다[7]. 온톨로지는 개념 분류에 중점을 두고 있고, 시소러스는 어휘를 기준으로한 “유의어 집합”으로 볼 수 있다. 시소러스는 유사 개념 연결이 방대하여 정확한 동의어 선택이 어렵고, 온톨로지는 세분화된 의미 분류에 한계가 있다. 이런 이유로 전산언어학자들 사이에 이를 보완할 어휘의망의 필요성이 요구되었다[4].

한국어를 대상으로 한 어휘의미망을 구축한 것은 90년대 중반부터다. 이중 PWN(프린스턴 대학의 WordNet)을 참조한 것은 ‘한국어 시소러스’와 'KorLex'다. 2004년부터 개발되기 시작한 부산대학교의 KorLex(이하 'KL')는 1단계(KorLex 1.0)로 PWN의 동의어집합을 대역하여 어휘의미망의 언어독립적인(language independent) 자질을 계승하였고, 2단계(KorLex 2.0)에서 1.0에 결여된 한국어 어휘의미의 특성을 반영하여 개념 및 의미구조를 재구조화하고 있다. 이때 한국어 어휘의 세분화된 의미 구분은 표준국어대사전의 다의어 분류를 준거로 삼았다. 3단계에서 구축해야 할 부분은 KL의 언어의존적인(language dependent) 정보이다. PWN의 동사와 형용사의 경우 간략하지만 격틀 정보(subcategorization)가 포함되어 있고, 이밖에도 품사 간 파생 정보가 들어있다. 이러한 언어의존적인 정보는 자연언어처리 분야에서 어휘의미망의 효용성을 증가시킬 수 있다.

그 예로 용언의 격틀 정보는 용언에 필요한 격들과 그 격에 사용할 수 있는 명사와의 관계를 설정하므로 구문분석 단계의 의미 중의성이나 기계 번역에서의 한국어 생성, 논항의 의미역 결정 등에 활용될 수 있다. 이런 이유로 격틀을 구축하려는 시도가 다양하게 있었다[2]. 특히 세종 용언 전자 사전에는 형용사 4,398개와 동사 15,181개에 대해서 격틀정보가 구축되어 있어 다양한 활용이 기대된다.

PWN을 대역한 KL도 자연언어 처리에서 유용하게 활용되려면 용언의 격틀정보가 추가되어 문장 내의 환경에 대한 형태적 통사적 분석이 병행되어야 한다. 그러나 현재 KL의 동의어 집합을 구성하는 어휘 의미에는 논항정보를 따로 구성할 수 없었다[3]. KL이 어휘 의미 구분의 준거로 삼는 표준국어대사전의 경우, 자연언어처리에 활용할만한 정도로 유용한 격틀 정보를 포함하지 않으므로, 이러한 정보가 정교하게 기술된 세종 전자 사전을 이용하여 KL을 확장하고자 한다.

세종 전자 사전의 격틀 정보의 선택제약(selectional restriction)은 세종 전자 사전의 명사 의미 부류와 체언 사전을 기준으로 기술되어 있다. 그래서 세종 전자 사전의 격틀정보를 활용하려면 격틀정보내의 선택제약 조건으로 사용하고 있는 의미 부류와 KL의 명사 어휘망을 연결하는 작업이 선행되어야 한다.

본 논문에서는 세종 전자 사전의 의미 부류와 현재 구축된 어휘 의미망과의 자동 맵핑 작업을 통해 KL에서 세종 용언의 격틀 정보를 활용할 수 있는 기반을 제공한다. 본 논문의 전개는 다음과 같다. 2장에서는 세종 국어 전자 사전의 의미 부류와 KL의 의미 계층 구조의 차이점을 살펴본다. 3장에서는 상이한 두 온톨로지를 자동 맵핑하기 위한 방법을 제시한다. 4장의 결론에서는 연구의 한계를 밝히고 연구 의의를 제시한다.

2. 세종의 의미 부류 체계와 KL의 의미 체계의 비교

2.1 세종 전자 사전 의미 부류

세종의 전자 사전은 기존 의미 부류 체계가 통사적 특성을 고려하지 않고 순수하게 개념적 기준으로 만들었다는 점을 고려하여 개념적 기준과 통사적 기준으로 명사 의미 부류 체계를 구축하였다. 의미 부류 체계를 구축하는 방법은 Lattice와 Tree구조가 있는데 세종 의미 부류는 자료의 편의성과 개념의 명확성을 위해 Tree구조를 따른다[1].

최상위 부류로는 <구체물>, <집단>,<장소>,<추상적 대상>,<사태> 등 5개의 부류가 설정되었는데 이중 처음 4개는 논항명사의 의미부류고 나머지 <사태> 부류는 술어명사의 의미 부류이다.[6] 현재 세종 의미 부류는 최상위 노드를 기점으로 최소 2층위에서 최대 7층위까지의 깊이를 갖는다. 세종 체언 상세 사전에 기술된 명사 35,874개에 대해서 의미 부류가 명시되어 있다. 2007년 공개된 세종 의미 부류 수는 아래와 같다.

[표1] 세종 전자 사전 의미 부류

최상위 부류명	최상위 부류별 하위부류	최상위 부류를 포함한 총 의미 부류수
구체물	197	198
집단	28	29
장소	53	54
추상적 대상	150	151
사태	190	191
합계		623

2.2 Korlex 명사 어휘 의미망

KL은 개념을 표상하는 최소 단위로 “동일한 어휘 의미를 가지는 동의어 집합(Synset)”으로 규정한다. 예로 다의어 “배”는 “복부, 선박, 배수”등의 의미를 가지는 데 이를 KL에서는 {복부, 배}, {선박, 배}, {배수, 배}등으로 표현하여 중의성이 없이 하나의 개념을 나타낸다. 예에서와 같이 KL에 표현된 어의는 어의를 나타내는 어휘형태로 표현된다. KL은 명사, 동사, 형용사, 부사에 대해서 구축하였고 그 중 가장 먼저 명사에 대해서 의미망이 구축되었다. 본 연구는 세종 의미 부류와 KL에 구축된 명사 의미망의 Synset 매칭이 주요 관심분야이므로 주로 KL에 구축된 명사 어휘의미망(KorLexNoun, 이하 KLN)에 대해서 다룬다.

KLN은 신셋간, 어의 간 의미 관계가 매우 다양하게 표현되어 있다. KLN의 의미 관계 중 신셋 간의 상위(hypernymy)와 하의(hyponymy) 계층은 IS-A 방식으로 나타낸다. KLN은 9개의 최상위 계층을 갖으며, 25개의 의미범주가 분류되어 있다[3].

2.3 세종 전자 사전 의미 부류와 KLN 의미 체계 비교

의미 범주 체계는 의미 범주 설정 목적에 따라 달라진다. 세계 내 존재 대상이 동일한 것이어도 경험적 지식 기반이 다르면 어휘의 의미 범주가 달라질 수 있다.

2.3절에서는 세종 전자 사전 의미 부류와 KLN의 의미 범주 체계가 달라서 발생할 수 있는 점을 본 연구의 주제와 관련하여 세 가지 기준에서 살펴 본다. 2.3.1에서는 세종 의미 부류와 KLN 신셋간의 맵핑 함수 관계를 살펴보고, 2.3.2에서는 의미 분류 체계 때문에 발생할 수 있는 계층 구조의 차이를 살펴본다. 2.3.3에서는 세종 의미 부류와 KL의 의미 범주 크기로 인해 맵핑할 때 고려할 점을 살펴본다.

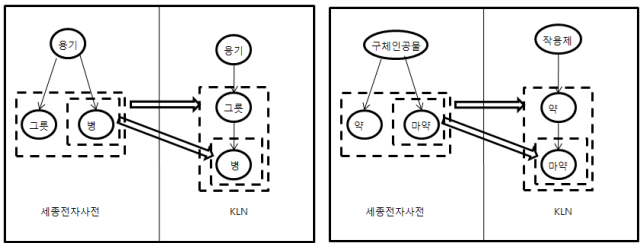
2.3.1 맵핑 함수 관계

세종과 KLN의 의미 분류 체계의 차이는 세종의 의미 부류에 속하는 어휘들이 KLN에서는 서로 다른 의미 범주에 속하도록 한다. 예로 “빗방울”이 KLN에서는 <양>의 하위 노드로 분류된 반면 세종에서는 <기상관련물>로 분류가 된다. 즉, 세종의 의미 부류에 속하는 어휘와 맵핑되는 KL의 신셋들은 서로 다른 상위 개념을 갖게 된다. 일부 신셋들은 최소공통상위노드(Least Upper Bound Node, 이하 LUB)를 통해 세종의 의미 부류와 하나로 맵핑될 수 있다. 그러나 분류 목적과 체계가 다르기 때문에 LUB가 항상 세종의 의미 부류와 맵핑 가능하다고는 볼 수 없다. 만약 맵핑 가능한 LUB를 구할 수 없으면 세종의 의미 부류에 해당하는 어휘 수만큼의 신셋으로 맵핑될 수도 있다. 본 연구에서는 이런 이유로 세종의미 부류: KLN 신셋=1:n의 관계로 본다.

2.3.2 의미 범주 계층 구조

KLN에서 구축된 의미 계층 구조는 사물의 관계를

IS-A방식으로 계층화 하였고 세종의 의미 부류는 개념 및 통사적 사용이 모두 고려되어 구축되었다. 이 같은 이유로 KLN의 상하위 계층으로 구성된 어휘 들이 세종에서는 서로 다른 의미 부류에 소속될 수 있다. [그림1]은 세종의 KLN과 세종의 서로 다른 계층 구조의 예이다. KLN에서는 <약>이 <마약>의 상위 계층으로 개념 계층화가 되어 있지만 세종에서는 <약>과 <마약>이 형제 관계로 설정되어 있다. 이는 맵핑된 결과를 세종의 격틀정보의 선택제약조건으로 활용할 때 적정서술어와의 관계에서 문제가 될 수 있다.



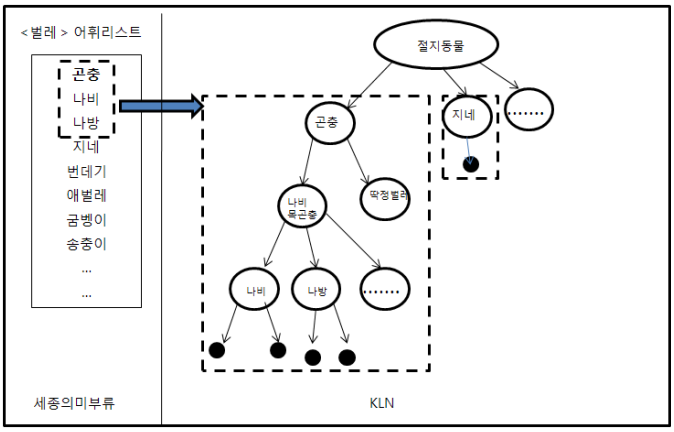
[그림 1] 세종의미 부류와 KLN의 계층 구조 차이

2.3.3 의미 범주 크기 비교

세종 전자 사전의 의미 범주는 약650개로 자연어에서 발생하는 의미범주를 모두 표현하기는 부족하다. 그러나 세종 전자 사전의 의미 범주는 “적정술어”와의 관계를 고려하여 분류 작업이 수행되었다는 점에서 의의를 찾을 수 있다. 세종 전자 사전에 비해서 KL은 의미 세분화가 세밀하기 때문에 세종 전자 사전의 의미 범주가 KL의 의미 범주보다 크다. 예로 세종 전자 사전의 <벌레> 의미 부류에는 “곤충”과 “나비”가 같은 부류에 속해 있고, KL에서는 “나비”가 “곤충”의 하위 노드로 설정되어 있다. 이는 세종 전자 사전의 의미 범주가 KL의 의미 체계보다 포괄적이기 때문이다. 이 같이 세종의 의미 범주는 KLN의 의미 범주보다 크기 때문에 세종의 같은 의미 범주에 속한 어휘들이 KLN에서는 상/하위어로 맵핑될 수 있다. 이런 예에서는 맵핑된 신셋 중 다른 신셋의 최소 공통노드에 해당하는 신셋으로 맵핑되는 것이 바람직하다.

3. 세종 의미 부류와 KLN 신셋의 자동 맵핑

2.3에서 살펴보았듯이 세종 전자 사전의 의미 부류는 다수의 신셋과 맵핑될 수 있다. 그러나 신셋들의 LUB가 하나의 개념으로 의미 부류와 맵핑될 수 있다면 좀 더 일반화된 맵핑 결과를 얻을 수 있을 것이다[2009윤애선]. 신셋들의 LUB를 어느 단계까지 고려할 것인지는 언어학자가 수작업으로 LUB와의 의미관계, 적정 술어와의 통



[그림 2] 세종의미 부류와 KLN의 의미 범주 크기 비교

사적 관계를 구별하여 맵핑하는 것이 가장 좋은 방법이지만 본 연구에서는 문자열 비교를 기반으로 한 형식적 방법과 LUB의 맵핑 가능성을 확률에 의해 결정한다.

세종 의미 부류 이름은 <기상관련물>, <운동경기역할인간>등의 메타언어가 사용되므로 의미 부류 이름이 직접 KLN의 신셋과 맵핑되지 않을 때가 있다. 본 연구에서는 세종 의미 부류 이름과 체언 전자 사전에서 의미 부류별로 추출한 어휘를 이용하여 맵핑을 시도한다. 세종 체언 전자 사전과 KLN은 『표준』기반으로 의미 분화가 되어 있으므로 세종에서 추출된 어휘가 다의어일 때도 같은 의미를 가진 KLN의 신셋 검색이 가능하다.

3.1 세종의미 부류와 KLN의 맵핑 알고리즘

2.3.2절에서 살펴보았듯이 세종 의미 부류에 속하는 어휘와 맵핑된 중에는 다른 신셋들의 LUB가 되는 신셋이 존재한다. 이런 신셋은 하위 신셋들을 대표하는 개념으로 맵핑될 수 있다. 그 외의 신셋들은 상위 노드가 같은 신셋들을 하나의 그룹으로 보고 그룹의 LUB에 해당하는 어휘의 맵핑 가능성을 확률적으로 계산한다.

$$P(SC_i,LUB_i) = C(SC_i,LUB_i) \;/C(LUB_i) \;- \Theta \;(0 <= \Theta<1)$$

SC: 세종 전자 사전 의미 부류,

LUB : KLN의 일부 신셋들의 LUB)

$C(SC_i,LUB_i)$ 는 세종의미부류 SC_i와 KLN에서의 맵핑 가능성을 계산할 LUB 신셋에 공통으로 나타난 어휘의 빈도수이다. $C(LUB_i)$ 는 LUB신셋에 속하는 하위 노드 개수를 나타낸다. 확률값 P가 임계치(Θ)이상이면 LUB를 세종 의미 부류와 맵핑 가능한 신셋으로 본다.

본 연구의 맵핑 과정은 다음과 같은 단계로 수행된다. 우선 세종 상세 체언 사전에 등록된 35,000개의 어휘를 의미 부류별로 추출한다. 여기서 추출된 의미 부류별 어휘는 KLN의 신셋을 검색할 때 사용된다. 맵핑은 의미

세분화가 잘 된 하위 의미 부류부터 상향식(bottom-up)으로 진행된다. 각 의미 부류에 속하는 어휘 어형과 KLN 신셋의 문자열을 비교하여 어형이 같으면 모두 맵핑 후보 신셋으로 설정한다. 이렇게 형식적 방법으로 만들어진 맵핑 후보 신셋리스트는 신셋들의 LUB와의 맵핑 가능성을 계산하여 복수의 LUB로 일반화된다. 맵핑 과정은 맵핑 후보 신셋 리스트가 더 이상 변하지 않을 때까지 (3)~(5)과정을 반복적으로 수행한다.

(1) 세종의 현재 의미 부류에 속하는 어휘들을 KLN에 맵핑해서 SynSetList를 구한다.
(2) SynSetList의 Element들 중 하위 의미 부류와 맵핑된 SynSet은 SynSetList에서 제거한다.
(3) 맵핑된 모든 신셋들을 상위 노드를 기준으로 그룹핑한다.

```
for each s(i) in mapped Synset List of KLN
    Append(s(i),SynsetGroup(k))
     for each s(j) in mapped Synset List of KLN
         if Hypernmy(s(i)) == Hypernmy((s(j))
                                  && !s(j).groupped then
            Append(s(j),SynsetGroup(k))
            s(j).groupped = true
         end if
      next
   next
```

(4) 그룹 내의 신셋들이 LUB로 병합가능 한지 계산한다.

```
    for each g(i) in SynsetGroup
         if p(sc,kc) >0 then
            lub=LUB(g(i))
            delete all element of g(i) from SynsetList
            append(SynsetList,lub)
         end if
    next
```

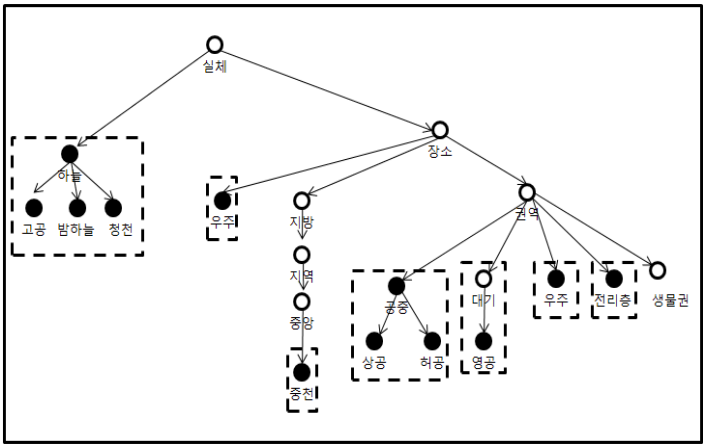
(5) 그룹간의 상/하위 관계를 검사해서 다른 그룹의 Hyponymy에 해당하는 그룹은 삭제한다.

```
  for each g(i) in SynsetGroup
     for each g(j) in SynsetGroup
        if  IsParent(g(i), LUB(g(j)) then
            delete all element of g(j) from SynsetList
        end if
     next
  next
```

세종 전자 사전의 <공중장소>는 총 23개의 어휘가 있다. 이 어휘를 이용해 KLN의 신셋을 검사한 결과 12개의 맵핑 후보 신셋이 검색되었다. 어휘형태 뒤의 숫자는 다의어를 구분하는 표시다.

맵핑된 어휘 리스트 :
고공1 공중1 밤하늘1 상공1 영공1 외계1 우주1 전리층1
조각하늘1 중천1 청천1 *코스모스3* 하늘1 허공1
맵핑되지 않은 어휘 리스트 :
반공중1 서천1 수련천1 우주공간1 저공1 창공1 천공1 한천1

검사된 신셋 후보 리스트를 이용해 위 알고리즘을 적용한 결과 아래와 같이 7개의 신셋이 <공중장소>로 맵핑된다.



[그림 3] 세종 의미 범주 <공중장소>와 맵핑된 KLN신셋

4. 결 론

본 연구에서는 세종 전자 사전의 격틀정보를 KLN에서 활용할 수 있도록 격틀정보의 선택 제약 조건으로 사용되고 있는 의미 부류와 KLN을 자동 맵핑하는 방안에 대하여 연구하였다. 세종 의미 부류별로 해당 어휘들을 KLN의 신셋과 맵핑하여 맵핑 가능한 후보 신셋들을 구하고 LUB에 의해 통합되는 신셋들은 LUB로 일반화하는 작업을 수행하였다. 그러나 자동으로 의미 체계가 다른 두 온톨로지를 완전하게 맵핑하기는 사실상 불가능하다. 자동 맵핑된 결과를 바탕으로 언어 전문가에 의해 좀 더 일반화 시키는 작업이 수행되어야 한다. 본 연구는 세종 의미 부류와 KLN의 맵핑을 통해 KLN의 언어 자원으로서의 유용성을 증대할 수 있는 가능성을 모색하였다는 데 의의가 있다.

참고 문헌

[1] 강범모, 박동호 외, “한국어 명사 의미 부류 체계의 구축과 활용”, 한글 및 한국어 정보처리 학술대회, pp. 247-251, 2001
[2] 최용석 외, “격틀 자동구축과 격틀 평가 방법에 관한 연구”, 한글 및 한국어 정보 처리 학술대회, pp. 272-279, 1999

[3] 윤애선, 황순희 외, “한국어 어휘 의미망 Korlex 1.5의 구축”, 정보과학회논문지:소프트웨어 및 응용 제36권 제 1 호(2009.1), pp. 92-108, 2009
[4] 이은령, 황순희 외 “다국어 어휘의미망 구축의 현황과 문제점”, 프랑스문화예술연구 제12집, pp. 369-401, 2004
[5] 이동혁, “의미 범주 체계의 구축과 사전에서의 활용”, 한국어 의미학 24, pp. 51-82, 2007
[6] 이성헌, “전자사전 구축과 의미 부류-세종 명사 의미 부류 체계의 예”, 한국 사전학, pp. 32-42, 2005
[7] 최호섭, “한국어 의미망 구축과 활용-명사를 중심으로”, 한국어학 제 17집, pp. 301-329, 2002