명사 워드넷과 단일어 사전을 이용한 한국어 동사 워드넷 구축

이주호° 배희숙† 김은혜 김혜경 최기선
한국과학기술원 전산학과 전문용어언어공학연구센터
{mywork, eunhye, hgkim, kschoi}@world.kaist.ac.kr
한국표준연구원†
elle@kriss.re.kr†

Construction of Korean Verb Wordnet Using Preexisting Noun Wordnet and Monolingual Dictionary

Juho Lee° Heesuk Bae† Eunhye Kim Hyekyong Kim Key-Sun Choi
KOTERM, Dept. of CS, Korea Advanced Institute of Science and Technology
Korea Research Institute of Standard and Science†

요 약

의미기반 정보 검색, 자연어 질의 응답, 지식 자동 습득, 담화 처리 등 높은 수준의 자연언어처리 시스템에서 의미처리를 위한 대용량의 지식 베이스가 필요하다. 이러한 지식 베이스 중에서 가장 기본적인 것이 워드넷이다. 이러한 워드넷을 이용함으로써 여러 의미 사이의 의미 유사도를 구할 수 있고, 속성을 물려받을 수 있기 때문에 비슷한 속성을 가진 의미들을 한꺼번에 다루는 데 유용하다. 본 논문에서는 기본 어휘를 바탕으로 기존의 명사 워드넷과 단일어 사전을 이용하여 한국어 동사 워드넷을 구축하는 방법을 제시한다. 본 논문에서 1차 작업을 통하여 구축한 동사 워드넷에는 동사 1,757개에 대한 4,717개의 의미(중복을 포함하면 모두 5,235개의 의미)를 포함하고 있으며 특별히 의미가 많이 편중된 14개의 개념에 속한 571개의 의미를 53개의 세부 개념으로 재분류하여 최종적으로 모두 767개의 계층적 개념으로 구성된 동사 워드넷이 만들어졌다.

1. 서론

의미기반 정보 검색, 자연어 질의 응답, 지식 자동 습득, 담화 처리 등 높은 수준의 자연언어처리 시스템에서 의미처리를 위한 대용량의 지식 베이스가 필요하다. 이러한 지식 베이스 중에서 가장 기본적인 것이 워드넷이다. 본 논문에서는 워드넷이라는 용어를 의미에 따라 체계적으로 분류된 다단계 어휘목록이라 정의하고 이용하도록 한다. 이러한 워드넷을 이용함으로써 여러 의미 사이의 의미 유사도를 구할 수 있고, 속성을 물려받을 수 있기 때문에 비슷한 속성을 가진 의미들을 한꺼번에 다루는 데 유용하다.

이제까지 한국어 명사의 의미분류에 대한 몇몇 연구가 있었지만 동사에 대한 연구는 거의 찾아볼 수 없었다[2][3][4]. 본 논문에서는 이미 구축된 명사 워드넷을 기반으로 하여 국어 사전을 이용해 동사 워드넷을 구축하는 과정과 이 때 생긴 문제점, 해결 방법에 대하여 기술하고자 한다.

먼저 관련 연구에서는 영어 워드넷의 동사에 대해서 간략하게 알아보고, 현재 우리가 이용하려고 하는 한국어 명사 워드넷에 대해서 기술한다. 3절에서는 본 논문에서 접근하는 방법에 대해서 대략적으로 알아 보고, 4절에서 실질적으로 한국어 동사 워드넷을 구축하는 과정에 대해서 자세히 기술한다. 이 과정은 자동으로 개념번호를 부여하는 부분과 수동으로 하는 후처리 과정이 포함된다. 5절에서는 1차 구축된 동사 워드넷에 대하여 일부 과도하게 편중된 개념에 속한 의미를 재분류 하는 과정에 대해서 설명하고, 마지막으로 6절에서 결론과 앞으로의 계획을 제시하며 마무리한다.

2. 관련연구

2.1 영어 워드넷(WordNet)의 동사

영어 워드넷은 미국 프린스턴 대학에서 언어 심리학적인 방법으로 구축된 어휘 데이터 베이스이다. 워드넷에서는 의미의 기본 단위를 synset이라 정의하여 유의어들을 모두 하나의 synset에 모았으며 이러한 synset을 기본으로 하여 각 의미간의 관계를 표현한다[10]. 이러한 워드넷은 현재 1.7 버전까지 구축되었으며 연구용으로 무료로 공개되어 영어권에서 언어자원으로 널리 이용되고 있다.

영어 워드넷의 동사 부분은 현재 약 21,000개의

의미(유일한 어휘수는 약 13,000개)에 대한 약 8,400개의 synset으로 이루어져 있다. 동사의 synset은 크게 사건이나 행동을 나타내는 동사와 상태를 나타내는 동사로 나누어 진다. 사건이나 행동을 나타내는 동사는 다시 14개의 세부 항목으로 나누어 지며 그것들의 최상위 synset은 verbs of bodily care and functions, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social interaction, weather verb이다.

동사 synset간의 관계는 기본적으로 함의(entailment)를 기반으로 하여 기술되었으며 다른 품사의 synset간의 관계는 아직까지 명확하게 기술되어 있지 않다[8].

2.2 사용한 한국어 명사 워드넷

본 논문에서는 기계가독사전과 기존의 의미체계를 이용하여 구축한 한국어 명사 워드넷을 기반으로 동사 워드넷을 구축했다. 이 명사 워드넷은 코퍼스의 고빈도어를 중심으로 하여 사전의 정의문을 이용해 사전의 각 명사 의미에 기존의 의미체계의 개념번호를 부여함으로써 구축하였다.

이 한국어 명사 워드넷은 23,823개의 명사와 이에 대한 56,523개의 의미를 포함하는 2,710개의 계층적 개념노드로 이루어져 있다[3].

3. 접근 방법

어느 언어이건 명사 의미체계에 대한 연구가 가장 활발하고 그 성과 또한 상대적으로 크다. 그러나 동사나 형용사와 같은 서술어 의미체계에 대해서는 연구가 아직 미진한 상태이다. 본 논문에서는 이미 구축해 놓은 한국어 명사 워드넷을 기반으로 동사 부분을 추가함으로써 동사 워드넷을 구축하는 방법론을 제시하고자 한다. 동사의 의미에 개념번호를 부여하기 위해서 기본적으로 각 동사의 의미에 대한 사전 정의문을 이용하였다.

3.1 명사 워드넷 기반 동사 워드넷 구축

인구어의 경우 형용사는 명사 앞뒤에서 명사를 한정하면서 명사와 함께 명사구를 이루지만, 한국어의 경우 서술어로 기능하며 동사와 같이 동사구의 핵심 요소로 기능한다. 동사가 동작성 서술어라면 형용사는 상태성 서술어인 것이다. 동사와 형용사는 문장에서의 역할이 명사와는 다르며, 동사의 의미는 동반하는 논항의 의미적, 통사적 특성에 따라 분화되므로 동사의 의미는 근본적으로 논항과 함께 고려되어야 한다. 본 논문에서는 동사 의미체계를 구축함에 있어서 기존의 명사 의미체계에 동사 어휘들을 매칭시켰다. 이는 개별 동사 어휘들이 단어로서의 의미속성을 범주화, 계층화한 것으로 동사 의미체계를 명사 의미체계와 완전히 분리하여 구축하지 않고, 논항에 따라 달라지는 동사의 세분화된 의미 보다는 동사의 보편적 의미에 중점을 두면서 구축하였으며, 의미를 고려한 다양한 언어처리에 유용하게 이용될 수 있다.

3.2 구축의 문제점

본 논문에서 이용한 한국어 명사 워드넷은 크게 구체와 추상으로 모든 현상과 사물, 행위 등을 구분하였다. 각 동사의 의미는 구체적으로 존재하는 사물을 지시하는 말인 구체명사의 하위 개념노드에 해당하지 않고 추상명사의 하위 개념노드에 포함될 것이다. 또한 기존의 워드넷이 명사를 위한 것이기 때문에 일부 개념노드에 대해서는 동사 의미가 편중될 수 있다. 이러한 문제점은 뒷부분에 다시 논의될 것이다.

4. 동사 워드넷 구축

전체적인 작업 순서는 아래의 그림과 같다.

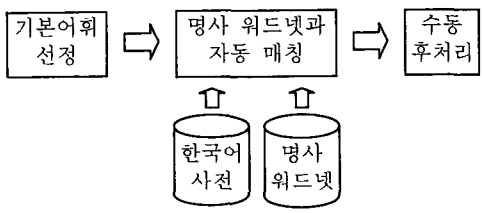

그림 1. 전체과정

먼저 동사 워드넷에 포함될 기본어휘에 대한 목록을 선정한다. 목록이 선정되면 기본동사의 의미에 대하여 사전의 정의문을 추출하고, 그 정의문을 이용하여, 이미 구축되어 있는 명사 워드넷을 기반으로 각 개념노드별로 만들어진 정의문 클러스터와 비교하여 가장 유사한 개념번호 후보들을 제시한다. 그런 다음 수동 후처리를 통하여 동사 워드넷을 구축한다. 그러면 각 과정에 대하여 보다 자세히 알아보겠다.

4.1 기본어휘 선정

1) 기본어휘 선정의 필요성

한국어 사전에서 실제 사용되는 어휘들을 조사하면 대략 10~15%에 불과하다[1]. 『표준국어대사전』에 대한 연구[6]에 따르면, 『표준국어대사전』에는 508,771개의 어휘가 수록되어 있는데 실제로 사용되는 어휘는 이중 20%에 불과하다고 한다. 언어사전의 역할을 고려한다 해도 매우 많은 수의 항목을 다루고 있음을 알 수 있다. 사실 한국어 언어사전에는 백과사전적 항목들이 다수 들어가 있으며, 복합어에 대한 엄밀한 사전학적 기준을 세우지 않고 항목에서 수많은 복합어를 다루는 것도 사전 항목수의 배가에 한 몫을 하고 있음을 부인할 수 없다. 어쨌든 자연언어처리를 목적으로 하는 전자사전이나 어휘목록에서 실제로 잘 사용되지도 않는 단어들을 매번 함께 다루는 것이 낭비 요소가

될 수 있으므로 실제 한국어 사용에서 필요하다고 인정되는 기본어휘목록을 선정하여 이용하는 것이 효율적이다.

2) 기본어휘 선정

기본어휘로 선정되어야 할 어휘와 제외되어야 할 어휘에 대한 명확한 기준을 정하는 것은 쉽지 않다. 이는 코퍼스를 기반으로 하여 고빈도어를 추출해 내는 것만으로는 해결될 수 없다. 완벽하게 균형잡힌 코퍼스를 구성한다는 것이 어렵고, 고빈도로 나타나지만 기본어휘로 보기 어렵거나 저빈도로 사용되더라도 한국어를 사용하는 데 있어서 기본적인 어휘로 보아야 할 것들도 있기 때문이다. 또한 어휘는 개방 집합이며 끊임없이 변화한다. 특히 인터넷을 통해 여러 나라의 정보가 활발히 교류될 수 있는 최근에는 어휘의 생기고 사라짐이 매우 급격하게 이루어지는 만큼 기본어휘에 포함될 단어들이 빨리 변하기 때문에 문제는 더욱 심화된다.

일반적으로 고빈도어는 고정성이 높은 어휘에 대한 정보를 제공하지만 기본어휘는 고정성에 더하여 저빈도이면서 기초적인 어휘들을 보완하여 선정하여야 한다. 이러한 작업은 전적으로 기계에만 의존하여 이루어질 수 없으므로 추출 방법의 객관성 문제가 대두되고 또한 의미 부착 코퍼스를 이용할 수 없는 만큼 동형어 처리의 문제를 안고 있다. 품사 부착된 코퍼스에 근거한 본 작업에서 품사가 다른 경우 자동 분리 되지만 품사까지 같은 동형어의 경우는 일일이 수동으로 조정될 수밖에 없다.

여기에서는 대용량 품사 부착 카이스트 코퍼스를 이용하여 모든 형태소에 대한 출현빈도를 구했다. 이를 내림차순으로 정렬하고 서로 다른 단어 형태에 대해 기존의 자료와 비교하면서 전문가들이 수동으로 난이도에 따라 어휘를 5단계로 등급화하였다. 이때 동형어, 복합어, 파생어 등은 가장 최근에 만들어진 대규모 국어사전인 『표준국어대사전』에 입각하여 처리하였다.

이러한 과정을 거쳐 만들어진 기본어휘목록에 대하여 1등급에서 3등급에 속하는 동사목록을 추출하여 동사 워드넷 구축에 이용하였다. 이는 동사로 쓰인 동작성 명사를 제외하고 모두 1,757개이다.

4.2 명사 워드넷에 자동으로 적용

각 동사 의미의 사전 정의문에 대하여 기존의 명사 워드넷의 각 개념노드에 속한 명사 의미의 사전 정의문 클러스터와 유사도를 계산하여, 가장 유사한 개념노드에 그 동사의 의미가 속하도록 한다. 이 방법은 기본적으로 한국어 명사 워드넷을 구축했을 때 사용했던 방법과 같다[3]. 여기에서는 유사도를 계산할 때 간단하게 <tf> (term frequency), <idf> (inverted document frequency)를 사용했다. 이렇게 하여 각 동사 의미에 대하여 열 개씩 개념번호 후보를 제시하였다.

4.3 수동 후처리

1) 후처리 방법

앞단계의 방법을 통하여 개념번호 후보가 제시된 동사 의미에 대하여 사람이 직접 보고 판별하여 정확한 개념번호만 선택하도록 한다. 개념번호 선택의 기준은 사전의 정의문에 기반하도록 하고, 하나 이상의 개념노드에 속하는 경우 모든 개념번호를 부여하도록 한다.

2) 작업상의 문제점

수동 후처리 과정에서 나타난 여러 가지 문제점들을 아래와 같이 정리하였다.

① 적합한 개념 선택의 어려움

동사 의미에 해당하는 의미속성을 부여하는데 있어서 명사 의미체계에 있는 개념으로 정확히 표현하기 어려운 경우들이 있다. 예를 들어 '가꾸다'의 의미가 한글학회 사전에서는 "식물이 잘 자라도록 매만지거나 보살피다"와 "몸을 잘 매만지거나 거두다"로 구분된다. 전자의 경우 '재배'에 해당하는 1,960번을, 후자의 경우 '치장'에 해당하는 1,583번을 부여하였지만, 후자의 경우 '화장'의 1,609번에도 해당하므로 두 가지 개념을 모두 선택하였다. 또한 '베다'는 "날이 서 있는 물건으로 어떤 물건을 끊거나 자르거나 가르다"의 의미로 2,249번 '절단'에 해당되나 "누워서 어떤 물건을 머리 밑에 괴다"의 의미에 대해서는 명사 워드넷 체계에서 적합한 개념을 찾기 어려웠다.

② 피동형과 사동형 구분

'감기다(몸이나 머리를 물에 씻게 하다)'는 '목욕'의 1,611번을 부여하였다. 그러나 "논밭을 갈게 하다"의 '갈리다'와 같은 사동형은 현 명사 워드넷 체계로는 이를 구분하여 주는 것이 어려웠다.

③ 기본어휘 항목이지만 기본의미가 아닌 경우

기본어휘 항목이지만 그 의미가 기본어휘로 간주되지 않을 경우는 개념번호를 부여하지 않았다. 예를 들어 '갈다'의 경우, "산란기의 민물고기가 잔잔하고 얕은 물로 몰려 나와 암수가 어울려 몸을 비비며 뒤틀다"라는 의미의 '갈다'는 자주 사용되지 않는 의미라 간주하고 처리하지 않았다.

④ 상위개념 참조

필요한 경우 상위개념을 참조하여 개념번호를 결정하였다. '쓰다'는 '집필'의 1,488번을, "글을 짓다"의 '짓다'는 '창작'의 1,555번을 부여하였는데, 이는 '집필'의 상위개념이 1,486의 '쓰기'이고, '창작'의 상위개념은 1,553번 '창조'이기 때문이다.

⑤ 두 개 이상의 개념

각 단어에 가능한 개념번호는 꼭 하나가 아닌 경우가 있을 수 있다. 관점에 따라 여러 개의 개념을 가질 수 있기 때문이다. 따라서 해당 동사가 여러 경로에 해당되는 경우 가능한 모든 개념번호를 부여하였다. 예를 들어 동사 '물려받다'는 '전승(1,146)', '상속(1,619)', '전달(1,550)' 모두의 개념번호를 가질 것이고, 동사 '꾸미다'는 '치장(1,583)', '장식(1,948)'의 개념번호를 부여 받는다.

⑥ 의미체계 하위노드 확장

수동 후처리를 통해 동사들의 의미가 거의 대부분 인간활동이나 사실과 현상 기술과 관련된 노드에 몰려 있음을 확인하였다. 그러나 명사 의미체계가 하위 노드로 분할되지 않은 끝 노드만으로는 한 곳에 집약되어 있는 동사의 의미 특성을 제대로 살려주기 힘들었다. 예를 들어 명사 의미체계에서는 '동작(1,561)'이란 개념이 다시 '전신동작(1,562)', '손동작(1,579)' 등으로 분화되는데, '머리동작'은 분화가 되어 있지 않아서 이에 해당하는 '끄덕이다' 같은 동사에 대해서는 개념번호 부여에 문제가 있었다. 또한 노드에 따라서 하위개념 분화가 잘 되어 있는 경우와 그렇지 않은 경우들이 균형을 잃어, 이러한 문제들을 고려하면서 하위노드로 확장을 꾀하였다. 이러한 재분류에 대해서는 5절에서 자세히 기술하였다.

4.4 작업결과

각 동사 의미당 10개의 의미분류 후보를 제시한 후 수동 후처리를 통해서 만든 1차 결과에 대한 통계값은 아래 표와 같다.

항목	값
동사 의미수	4,717
동사 어휘수	1,757
의미 분류수	728

표 1. 처리된 항목에 대한 통계

각 동사 어휘당 평균 약 2.68개의 의미를 가지며 이는 전체 2,710개 개념노드 중에서 728개의 개념노드에 각각 매칭되었다. 이 728개의 개념노드에 대한 전체적인 분포는 그림 2와 같다. 여기에서 X축은 속해있는 동사 의미수로 정렬된 개념번호를 나타내며 Y축은 그 개념노드에 속한 동사 의미수를 나타낸다

각 개념노드 당 평균 약 6.48개의 동사 의미가 속해 있지만 그래프을 보면 알 수 있듯이 일부 개념노드에 집중적으로 속해있음을 알 수 있다. 가장 많은 동사 의미를 가지는 개념노드는 '발생(2,068)'으로 모두 59개의 동사 의미가 몰려 있었다. 반면에 126개의 개념노드에는 오직 하나의 동사 의미만 속해 있었다. 속해있는 동사 의미수가 30개가 넘는 개념노드는 모두 14개인데 그것들에 속한 동사 의미수는 모두 571개로 전체 동사 의미수의 약 10.9%를 차지했다. 이는 현재 의미체계가 동사의 세분화된 의미를 반영하는데 한계가 있으며 더 세분화 할 필요가 있음을 나타낸다.

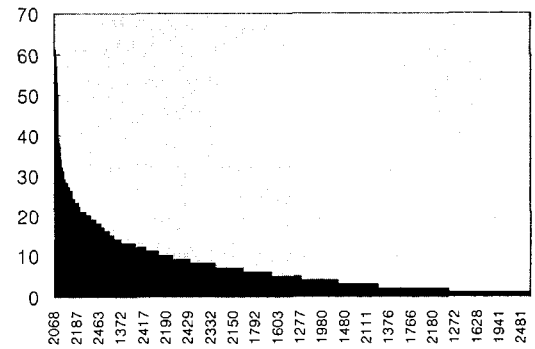

그림 2. 개념노드에 대한 의미 분포

5. 재분류

앞서 말했지만 전체 개념노드에서 기본 동사의 의미 속성에 해당하는 개념번호를 부여하면, 동사가 사물의 상태나 동작을 기술하는 데에 쓰이는 만큼 인간활동이나 사물의 현상 등의 개념노드에 편중되게 된다. 명사 워드넷에서 1번부터 1,000번까지는 '인간', '동물', '식물', '도구' 등과 같은 구체적 개념이고 1,001번부터는 추상적 개념들로 이루어져 있다. 동사 의미는 명사 워드넷에서 1,000번이상의 추상적 개념들에 대한 개념번호가 부여되었고, 1,000번 아래쪽의 구체적 개념번호는 전혀 부여되지 아니하였다. 1,000번 위쪽의 추상적 개념은 다시 '추상물(1,001)', '일(1,235)', '추상적 관계(2,422)'로 나뉘고, 동사들은 대부분 '일(1,235)' 하위의 개념노드에 들어간다.

'일(1,235)'은 '인간활동(1,236)', '사실과 현상(2,054)', '자연현상(2,304)'으로 이루어지며 한국어 기본동사는 거의 대부분 이쪽의 범주에 들어간다. 전체 4,717개 동사 의미는 중복을 포함하여 5,235개 개념번호를 부여 받았는데, 이중 2,636개가 '인간활동(1,236)' 또는 그 하위개념에 속해 있다. '사실과 현상(2,054)'의 하위개념에는 1,895개, '자연현상(2,304)'의 하위개념에는 434개의 동사 의미가 각각 속해 있다. 따라서 단 280개 동사 의미만 앞의 세 개념 및 그것의 하위 개념이 아닌 나머지 개념노드에 해당한다. 이런 것들은 대개 '인간활동(1,236)'에도 속하면서 다른 개념노드에도 동시에 속하는 경우이며 그 소속은 다양하다.

이러한 개념의 분산을 볼 때, 동사의 워드넷은 전체 명사 워드넷 중에서 대부분 '일(1,235)'이라는 개념노드 및 그 하위 개념에 모두 속하므로, 이쪽 부분을 세분화하면 동사의 의미를 더 잘 반영한 동사 워드넷

을 구축할 수 있다. 동사 워드넷을 명사와 독립적으로 구축하더라도 결국은 '인간활동'이나 '사실과 현상', '자연현상' 등에 관한 워드넷으로 구축될 것이므로 전체 워드넷에서 이 부분을 세분화하고 동사의 의미를 잘 살려줄 수 있도록 확장 한다면 좋을 것이다.

표 2는 '도착(2,133)', '분리(2,244)' 개념노드에 대한 재분류의 예이다. 여기에서 동사의 의미는 '표제어/품사/표제어번호/의미번호'로 표기하였다.

분류전	분류후	의미
도착	장소	다다르다/자동사/0/1, 닿다/자동사/1/2, 밟다/타동사/0/3, 이르다/자동사/1/1, 접어들다/자동사/0/2
	지경	다다르다/자동사/0/3, 닥치다/자동사/1/0, 당하다/자동사/0/1, 당하다/타동사/1/1, 들이닥치다/자동사/0/0, 마주치다/자동사/0/4, 만나다/타동사/0/2, 맞다/타동사/1/1, 접하다/타동사/1/2, 처하다/자동사/0/0, 맞히다/타동사/0/7, 만나다/타동사/0/4, 맞닥뜨리다/타동사/1/0
	때	가다/자동사/0/9, 오다/자동사/2/9, 이르다/자동사/1/2, 이르다/자동사/1/3, 돌아오다/자동사/0/3, 만나다/타동사/0/3, 오다/자동사/2/8, 찾아오다/타동사/0/2
	정도	닿다/자동사/1/3, 오다/자동사/2/4, 올라오다/자동사/0/2, 닿다/자동사/1/4, 자라다/자동사/2/0, 먹다/타동사/0/7,
	위치	서다/자동사/1/4, 앉다/자동사/0/4
분리	분리	가르다/타동사/2/2, 갈라지다/자동사/1/2, 갈리다/자동사/1/0, 나누다/타동사/0/1, 나뉘다/자동사/0/0, 벌리다/타동사/2/1, 벌어지다/자동사/0/1, 벌어지다/자동사/0/2, 비집다/타동사/0/1, 째다/타동사/5/1, 째다/타동사/5/3, 쪼개다/타동사/0/0, 찢기다/자동사/0/0, 찢다/타동사/0/1, 찢다/타동사/0/2, 찢어지다/자동사/0/0, 타다/타동사/4/1, 태우다/타동사/4/1
	떼어냄	거르다/타동사/1/0, 내리다/타동사/1/6, 덜다/타동사/3/2, 떨다/타동사/2/1, 떨리다/자동사/2/1, 떨어뜨리다/타동사/0/2, 떨어지다/자동사/0/2, 떼다/타동사/0/1, 떼다/타동사/0/9, 뜨다/타동사/7/1, 뜨이다/자동사/1/1, 뜯다/타동사/0/1, 떼놓다/타동사/0/0, 빼다/타동사/0/3
	벌여놓음	떼다/타동사/0/6 떼다/타동사/0/7 띄다/타동사/2/0 띄우다/타동사/1/0 바르다/타동사/3/2
	기타	떠나다/타동사/1/4 풀다/타동사/0/2

표 2. 재분류의 예

명사 의미체계에서 2,133번 '도착'은 최하위 개념노드이다. 그런데 많은 기본동사들이 '도착'의 개념노드에 연결된다. 그러나 '도착' 개념에 연결된 동사들을 다시 부분집합으로 분류되는 것을 볼 수 있다. 예를 들어 '가다/자동사/0/9', '오다/자동사/2/9', '이르다/자동사/1/2', '이르다/자동사/1/3', '돌아오다/자동사/0/3', '만나다/타동사/0/3', '찾아오다/타동사/0/2' 등은 "시간적 개념의 다다름"이고, '다다르다/자동사/0/1', '닿다/자동사/1/2', '밟다/타동사/0/3', '이르다/자동사/1/1', '접어들다/자동사/0/2'는 "장소의 개념의 다다름"이다. 이렇게 의미의 속성을 재분류하여 하나의 개념노드에 엮인 동사들을 재분류하고 노드를 연장시켰다.

하위개념으로 연장된 개념노드는 분류된 동사들을 대표할 수 있는 개념을 내세워 묶어 주었다. 이때 재분류가 개념의 분화에 대한 철학적 심리적 기준 보다는 실제 기본 동사들의 분포를 동의어 여부에 따라 그룹으로 나누는 것으로부터 시작하였다.

이 과정을 통하여 한 개념노드에 속한 동사 의미수가 30개 이상인 총 14개의 개념노드에 대하여 재분류 작업을 해서 모두 53개의 세부 개념노드로 나누었다.

6. 결론 및 향후연구

본 논문에서는 기본 어휘를 바탕으로 하여 기존의 명사 워드넷과 단일어 사전을 이용하여 한국어 동사 워드넷을 구축했다. 1차 작업을 통하여 구축한 동사 워드넷에는 동사 1,757개에 대한 4,717개의 의미(중복을 포함하면 모두 5,235개의 의미)를 포함하고 있다.

그 후 특별히 의미가 많이 몰린 14개의 개념에 속한 571개의 의미를 53개의 세부 개념으로 재분류하여 최종적으로 모두 767개의 개념으로 구성되어 있다.

향후 연구로 현재 수동으로 되어있는 재분류 작업을 반자동화 하는 것을 계획중이다. 각 동사의 의미에 대한 사전 뜻풀이와 이미 구축되어 있는 명사 워드넷을 기반으로 클러스터링 기법을 이용하면 어느 정도 재분류 작업을 반자동화 할 수 있을 것이다.

또한 구축된 동사 워드넷을 평가하는 과정이 남아 있다. 다양한 방법으로 구축되어 있는 동사 워드넷에 대한 여러 가지 현상을 파악하고, 이를 보완하는 과정이 남아 있다.

궁극적으로는 현재 각각 구축되어 있는 언어자원을 통합하는 일이 남아 있다. 이미 구축되어 있는 동사 격틀 정보를 매개로 하여 동사 워드넷과 명사 워드넷을 통합하면 명사와 동사를 한꺼번에 포함하는 의미체계를 얻을 수 있다.

감사의 글

본 연구는 전문용어언어공학연구센터에서 수행한 문화부의 21세기 세종계획 "전문용어 정비", 과학재단의 국제공동연구사업 "자연언어처리 기반 이동통신 시스템에 관한 기초 연구", 과학기술부의 뇌신경정보학 사업 "인간의 지식처리 모델링을 위한 전문분야 지식베이스 원형 구축 및 활용" 과제의 일환으로 수행되었습니다.

7. 참고 문헌

[1] 김광해. 1999. 형용사 유의어 뜻풀이 정교화 방안에 관한 연구, -'아름답다, 추하다'군을 중심으로-, 선청어문, 제27집.

[2] 문유진. 1996. 한국어 명사를 위한 WordNet의 설계와 구현. 정보과학회논문지(c) 제2권 제4호, 437-445.

[3] 이주호, 은광희, 최기선. 2001. 기계가독사전을 이용한 한국어 시소러스 구축. 제13회 한글 및 한국어 정보처리 학술대회 논문집, 273-278.

[4] 이창기, 이근배. 1999. WordNet을 이용한 한국어 시소러스 자동 구축. 제11회 한글 및 한국어 정보처리 학술대회 논문집, 156-163.

[5] 이정민, 배영남. 1982. 『언어학 사전』. 서울. 한신문화사.

[6] 정호성. 1999. 『표준국어대사전』 수록 정보의 통계적 분석, 새국어생활, 제10권 1호.

[7] 한글학회. 1997. 『우리말 큰사전』. 어문각.

[8] Fellbaum, Chriatiane, The English Verb Lexicon as a Semantic Net. 1990. *International Journal of Lexicography*, 3(4), 278-301.

[9] Ikehara S. *et al.* .1997. *The Semantic System, volume 1 of Goi-Taikei -- A Japanese Lexicon.* Iwanami Shoten.

[10] Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4), 235-244.

[11] Poesio, Massimo, Sabine Schulte im Walde and Chris Brew. 1998. Lexical Clustering and Definite Description Interpretation. In *Proceedings of AAAI Spring Symposium on Learning for Discourse.*