한국어 개념사전의 구축에 관한 연구

A Study on the Construction of a Korean Concept Dictionary

김수정, 김태수, 연세대학교 대학원 문헌정보학과

Kim Soo-Jung, Kim Tae-Soo. Dept. of Library and Information Science, Graduate School of Yonsei University

개념 정보를 제공하는 어휘 데이터베이스로 WordNet, CYC, EDR 등이 출현하였다. 본 연구는 WordNet의 개념 기술 방식에 따라 한국어 개념 사전을 구축하기 위한 것이다. 우선 개념을 분류할 적절한 분류 체계를 설정하고, 연세 말뭉치에서 빈도수가 높은 상위 300개 명사를 추출하여 사전의 뜻풀이에 나타난 명사와 연관관계로 표시된 명사를 함께 제시함으로써 개념을 표현하였다. 이러한 한국어 개념 사전은 의미모호성을 해소하는데 기여할 수 있을 것이다.

1. 서론

시소러스에서 제공되는 어휘 관계와 일반 사전에서 제공하는 개념 정보를 모두 포함하는 어휘 데이터베이스로 WordNet, CYC, EDR 등이 1980년대에 구축되기 시작하였다. 이들은 어휘 단위로 정보를 제공하는 대신 "개념"을 기본 단위로 기술하고 개념간의 의미 관계를 표현하는데 중점을 둔다는 데에 특징이 있다.

그 중에서도 WordNet은 개념 기술과 의미 관계 설정이 간단하면서도 명확하여 존재론(ontology)의 표준으로 인식될 만큼 이 분야의 선구자적 역할을 하고 있다.

따라서 본 연구에서도 WordNet과 유사한 형태로 개념 정보를 제공하는 한국어 개념 사전을 구축하였다.

2. 개념 단위의 어휘 데이터베이스

2.1 WordNet

WordNet은 프린스톤 대학 심리학과 인지과학연구실에서 1985년부터 구축되기 시작한 어휘 데이터베이스로써 동의어 집합(synonym set) 즉 synset을 개념 표현의 기본 단위로 삼는다. synset이란 한 단어가 지닌 여러 의미를 인정하여 개별 의미마다 숫자를 붙여 구별하고 각 의미의 동의어들을 { }로 표시하여 이를 하나의 그룹으로 묶은 것을 말한다. 예를 들어 'case'라는 단어의 synset은 아래와 같다(<그림 1>).

{carton, **case0**, box,@ (a box made of cardboard; opens by flaps on the top)}
{**case1**, bag,@ (a portable bag for carrying samll objects)}
{**case2**, pillowcase, pillowslip, slip2, bed linen,@ (a removable and washable cover for pillow)}
:

<그림 1> 'case'의 synset

WordNet의 명사는 행위, 동물, 인공물, 속성, 신체, 인지, 커뮤니케이션, 사건, 감정, 음식, 집합, 위치, 동기, 자연물, 자연현상, 사람, 식물, 소유, 과정, 수량, 관계, 모양, 상태, 물질, 시간과 같은 범주로 구분되고, 어휘 관계는 동의어, 상위어가 주를 이룬다.

2.2 EDR

1986년부터 1994년에 걸쳐 일본 전자 사전 연구위원회(Japanese Electronic Dictionary Research Institute)에서는 EDR 전자 사전을 구축하여 발표하였다. EDR 전자 사전에서 핵심이 되는 것은 개념 사전이다.

개념 사전에서는 개념을 크게 5범주 즉 인간/행위자, 물질, 사건/발생, 장소, 시간으로 나누고 있고 어휘 관계로는 격관계, 전체/부분 관계, 상하관계, 동등관계, 소유관계 등의 구문적 관계를 주로 제시하고 있다.

2.3 CYC

CYC는 미국의 Cycorp사에서 1984년부터 단어의 의미와 규칙, 단어의 관계에 대한 일반적인 지식을 제공하기 위해 구축되었다.

CYC의 한 항목을 살펴보면 다음과 같다.

#$Skin
A (piece of) skin serves as outer protective and tactile sensory covering for (part of) an animal's body.....
isa#$AnimalBodyPartType
genls#$BiologicalLivingObject#$AnimalBodyPart #SheetOfSomeStuff.....

<그림 2> CYC의 한 항목

CYC의 범주는 43개로 기본적인 범주, 최상위 수준의 범주, 시간과 날짜, 술어의 형태, 공간 관계, 량, 수학, 상황, 집합, 행위, 변형, 상태의 변화, 소유의 변화, 이동, 물체의 부분, 구성요소, 행위자, 기관, 역할, 소유, 감정, 태도, 사회, 생물학, 화학, 심리학, 일반 의학, 물질, 파동, 기계 장치, 생산, 재정, 음식, 의복, 날씨, 지리, 교통, 정보, 지각, 동의, 언어학 용어, 문헌으로 이루어져 있으며 어휘 관계는 설정하기에 따라 수천 개도 될 수 있다.

그러나 CYC는 WordNet이나 EDR 개념 사전과는 달리 어휘 수준 이상의 일반적 상식 즉 추론 지식까지 포함한다는 특징을 지니고 있다.

3. 한국어 개념 사전의 구축

3.1 구축 절차와 방법

한국어 개념 사전의 구축 절차와 방법은 다음과 같다.

1) 분류체계 설정 - 어휘를 단순히 주제별로 나열하는 열거형 분류체계가 아닌 몇 개의 특성으로 분류하는 패싯 분류 체계를 선택하였다.

2) 단어선정 - 연세 말뭉치로부터 품사에 관계없이 빈도수가 높은 상위 1000개의 단어를 추출하고 중에서 상위 300개의 명사를 대상으로 하였다.

3) 개념기술 - 각 단어마다 일반 사전의 뜻풀이에 나타난 명사 그리고 연관관계로 표시된 명사를 함께 제시함으로써 개념을 기술한다. 다의어의 경우 각 의미를 개별적으로 기술한다.

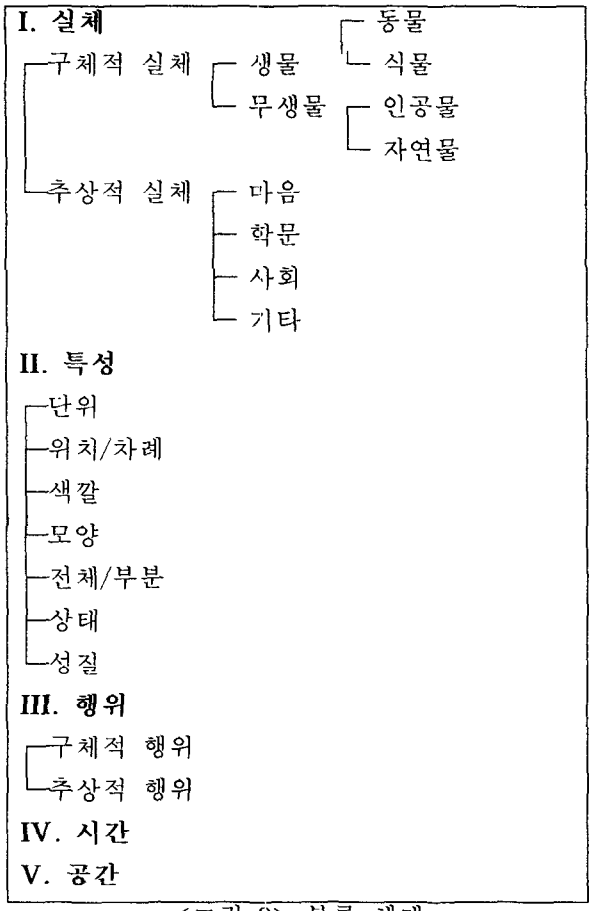
4) 어휘관계설정 - 사전의 뜻풀이에 나타난 상위어와 반의어를 추가한다.

5) 분류체계에 맞춰 기술된 개념들을 분류한다.

3.2 분류체계 설정

본 연구에서는 랑가나단의 기본 범주와 Aitchison의 시소러스 작성법에서 규정한 분류

체계를 참고하여 다음과 같은 분류 체계를 설정하였다(<그림 3>).



<그림 3> 분류 체계

3.3 개념 기술

개념을 기술하는 방법에는 동의어를 제시하는 방법, 뜻풀이를 제시하는 방법, 용례를 제시하는 방법 등 여러 가지가 있으나 아래의 예와 같이 동의어를 제시하는 방법이 표현이 간단하고 명료하여 직관적으로 이해하기 쉽다는 장점이 있다. 따라서 WordNet에서도 우선적으로 동의어를 나열하고 있다.

예) board : group of persons controlling a business or a government department (뜻풀이)
board : the board of governors (용례)
board : committee (동의어)

그러나 본 연구에서는 개념을 더욱 다양하고 직접적으로 표현하기 위하여 일반적인 동의어뿐만 아니라 사전의 뜻풀이에 나타난 모든 단어들을 추출하여 제시하였다. 예를 들어 '가격'은 사전에서 '물건이 지니고 있는 가치를 돈으로 나타낸 것. 값. 고가(估價)'로 뜻풀이되어 있으므로 여기에 출현한 명사를 모두 추출하면 '물건, 가치, 돈, 값, 고가(估價)'가 된다.

그리고 사전에는 표제어로 올라와 있는 단어에 대해 반의어, 참고어, 준말/본말, 큰말/작은말 등 연관관계로 표시된 단어들도 포함시켰다. 이렇게 함으로써 결과적으로 엄격한 의미의 동의어뿐만 아니라 관련된 모든 단어들을 함께 제시하여 보다 적극적으로 개념을 기술할 수 있다.

또한 뜻풀이에 나타난 단어들 중에 상위의 의미를 갖는 단어들에는 @, 반의어에는 ! 표시를 하였다.

이상의 원칙을 바탕으로 '여자'란 표제어의 개념을 아래와 같이 기술하였다(<그림 4>).

여자(女子) ① 자성(雌性)을 가진 사람을 노소에 관계없이 예사롭게 이르는 말. ↔남자
② [역] 신라 때 나인의 하나. 침방(針房)에서 바느질하는 일을 맡아 보았음.
여자(餘資) 쓰고 남은 자금.

↓

여자1 (자성, 사람@, 노소, !남자)
([역] 신라, 나인@, 침방, 일)
여자2 (자금)

<그림 4> '여자'의 개념 기술

이와 같은 방법으로 300개 명사의 개념을 기술하였다.

3.4 장점과 제한점

이와 같이 구축된 개념 사전이 지니는 장점은 다음과 같다.

첫째, 관련어들을 함께 제시하기 때문에 개념을 직관적으로 파악하는 것이 용이하다.

둘째, 다의어가 지닌 모든 의미를 포함할 수 있다.

셋째, 분류 체계를 통해 의미를 그룹화할 수 있다.

그러나 본 연구는 대상 단어의 수가 적고 흔히 쓰이지 않는 단어와 전문어가 다수 포함되어 있어 분류 작업이 완벽하지 못하다는 제한점이 있다.

3.5 활용 가능성

WordNet은 중의성을 가진 단어의 의미를 밝혀 의미 태그를 달아주는 자동 의미모호성 해소 연구, 단어 사이의 유사도를 구하는 연구, 질의와 문헌 사이의 유사도를 구하는 연구, 자동 분류, 정보 추출, 탐색 확장 등 여러 분야에서 응용되고 있다.

한국어 개념 사전도 WordNet과 같이 다양한 연구에서 응용될 수 있지만 특히 분류체계 내에서 개념을 기술하는 것에 중점을 두었기 때문에 아래와 같은 분야에서의 역할이 기대된다.

첫째, 시소러스 구축에 기여할 수 있다. 다양한 동의어와 관련어를 제공함으로써 용어의 선정에 도움을 주고 개별 의미를 기술함으로써 한정어를 선택할 수 있게 한다.

둘째, 분류체계의 수립에 기여할 수 있다. 패싯 분류를 이용하여 특성에 따른 개념분류를 시도함으로써 새로운 분류체계를 수립시 참조 가능하다.

셋째, 의미모호성을 해소하는데 기여할 수 있다. 예를 들어 '말'이라는 단어는 여러 의미로 사용되지만 '말'과 함께 (포유류, 동물, 다리, 목...)이라는 단어를 열거하게 되면, 이것은 "타고 다니는 말"로 사용된 것임을 쉽게 알 수 있고 (사람, 생각, 느낌, 표현, 전달, 음성 기호...)라는 단어들과 함께 열거되면 "사람이 소리내어 하는 말"이라는 뜻으로 쉽게 파악할 수 있다.

Resnick과 Lesk도 이와 유사한 방식을 사용하여 단어가 갖는 의미의 모호성을 해소하고 명확한 의미를 획득하는 실험을 한 바 있다.

4. 결론

본 연구에서는 분류체계를 지닌 단어들의 개념 정보를 제공하는 한국어 개념 사전을 구축하였다. 사전의 뜻풀이에 나타난 명사 그리고 연관관계로 표시된 명사들을 함께 제시함으로써 보다 적극적으로 개념을 표현할 수 있다.

외국에서는 이미 의미모호성 연구에서 개념 정보의 효용성이 입증되었고 앞으로 우리 나라에서도 한국어 개념 사전을 이용한 활발한 실험과 연구가 이루어져야 할 것이다.

이를 위해 대상 단어의 수를 확장하여 좀 더 합리적이고 포괄적인 분류 체계를 확립하고, 현재 상위어, 반의어로 제한되어 있는 어휘 관계 외에 더욱 다양한 어휘 관계를 설정하는 연구가 필요할 것으로 보인다.

참고문헌

CYCorp. Inc., Available from Internet: <http://www.cyc.com/>

EDR Technical Dictionary, Japan Electronic Dictionary Research Institute, LTD., 1998.

Lenat, D., Miller, G., Yokoi, T., "CYC, WordNet, EDR; Critiques and Responses," *Communication of the ACM 38(11)*, 1995, pp. 45-48.

Miller, G. A, "WordNet: An On-Line Lexical Database," *International Journal of Lexicography 3(4)*, Oxford University Press, 1990.