백과사전 질의응답 시스템을 위한 어휘개념망 구축

최 미 란　　오 효 정　　장 명 길
한국전자통신연구원 음성/언어정보연구부
{miranc,ohj,mgjang}@etri.re.kr

Constructing Korean Lexical Concept Network for Encyclopedia Question-Answering System

Miran Choi, Hyo-Jung Oh, Myung-Gil Jang
Speech/Language Technology Research Center, ETRI

요 약

백과사전 질의응답 시스템은 사용자의 자연어 질문과 검색 대상 문서인 백과사전 내용의 의미를 파악하기 위한 고정밀 자연어 처리 기술이 요구된다. 이러한 고정밀 자연어 처리 기술을 위한 중요한 언어자원을 제공하기 위하여 한국어 명사와 동사로 구성되는 대규모 어휘개념망을 구축하였다. 한국어 어휘개념망은 명사와 동사의 상하위 관계를 주요 계층구조로 하여 다양한 한국어 어휘 기초 자료를 바탕으로 구축되었다. 구축된 규모는 일반 명사 약 6만 어휘와 동사 약 2만 어휘를 포함한다. 이 논문에서는 어휘개념망을 구축하기 위한 방법과 과정을 소개하고 지금까지 구축된 어휘개념망의 특성에 대해 기술하며, 백과사전 질의응답 시스템에서 어떻게 활용되는지 시스템 구성요소의 예를 들어서 설명한다. 또한 현재 구축된 어휘개념망의 성능 평가를 위해 일반 코퍼스에 대한 커버리지 측정 결과를 기술한다.

1. 서 론

백과사전 질의응답 시스템은 사용자의 자연어 질문에 대하여 그 의도를 파악한 후에 백과사전 도메인으로부터 질문에 대한 답을 찾는 정보 검색 시스템이다. 이러한 질의응답 시스템이 정확한 답을 제공하기 위해서는 고정밀의 자연어처리 기술이 필요하다. 정교한 자연어 처리 시스템의 기반이 되는 언어 정보를 제공하기 위하여 대용량 어휘사전과 같은 잘 구성된 대규모의 언어자원으로 이루어진 지식 베이스가 필수적이다. 그러한 언어자원에는 다양한 사전류, 시소러스 워드넷, 의미망등의 예가 있다.

백과사전 질의응답 시스템의 핵심 기초 언어자원으로 활용하기 위하여 온톨로지나 시소러스와 유사한 한국어 명사와 동사를 위한 어휘개념망을 구축하고 있다. 어휘개념망은 한국어 어휘들의 개념을 고려하여 각 어휘들을 의미관계로 연결시켜 놓은 어휘 데이터베이스로, 2001년 명사 개념망 구축을 시작으로 현재 동사 개념망도 구축 중이다.

한국어 명사 개념망의 경우, 상하위관계를 주요 계층구조로 하고 있는데, 현재 약 6만여 단어로서 31개의 최상위 레벨의 어휘와 깊이 12레벨로 구성되어 있으며, 약 25만 여의 고유명사들이 InstanceOf 관계로 연결되어 있다. 구축된 명사 개념망은 의미기반 정보검색 시스템의 기초 지식베이스로서 활용되었으며, 현재 연구를 수행중인 백과사전 질의응답 시스템을 위한 핵심 지식베이스로 활용되고 있다[1].

동사 개념망은 일반 코퍼스 추출 고빈도 동사 1만개와 백과사전 고빈도 동사 1만개를 대상으로 구축 중이다. 동사개념망은 어휘개념망 워크벤치를 이용하여 대상 동사를 기초 동사와 파생동사로 분류하여 동의어/유의어/관련어 관계를 중심으로 반자동으로 구축하고 있다. 향후 동사개념망은 명사 개념망과 연결하여 백과사전 질의응답시스템의 언어처리와 기타 일반 언어처리를 위한 핵심 언어 자원으로 활용할 예정이다. 현재 개발되고 있는 백과사전 질의 응답 시스템에서 어휘개념망은 언어분석 모듈의 여러 부분에서 활용되는데 어휘의 미태깅을 위한 어휘 중의성 해소 기능, 사용자 질의 확장, 정답색인, 주제 추출 및 확장 등의 다양한 기능을

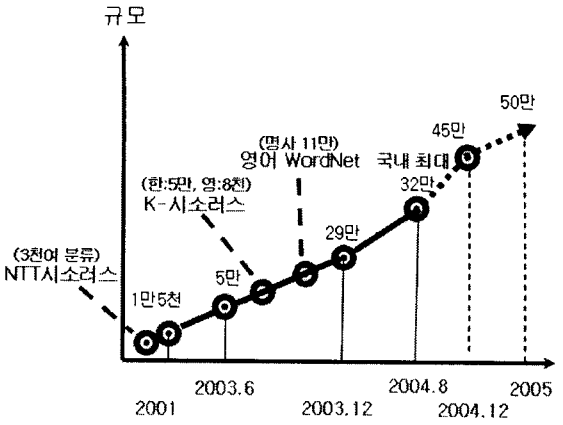
위하여 사용된다.

2. 명사 개념망 구축

응용 서비스 시스템에 언어 정보를 제공해주는 언어 자원에는 다양한 종류가 존재한다. 이 중에는 각종 사전이나 WordNet, 시소러스와 의미망등이 있는데, 우리가 구축한 한국어 어휘개념망도 이러한 언어 자원의 일종이다 [1][2]. 한국어 어휘개념망을 정의하면 한국어 명사 어휘들 중, 개념어에 해당하는 어휘들과 이들 어휘들을 의미관계로 연결시켜 놓은 어휘 데이터베이스이다. 어휘개념망은 분별 이론에 기반하여 한 단어의 개념을 각각의 개념에 분별된 심볼로 표현한다. 또한 그림 2에서와 같이 어휘개념망은 어휘간의 의미관계에 기반하여 한국어 명사의 개념을 구성하는 계층적인 구조를 형성하고 있다.

WordNet에서는 단어들이 유의어 세트를 통하여 연결되어 있는 반면에, 한국어 어휘개념망에서는 개별적인 어휘가 개념망 상에서 관계를 가지고 연결되어 있다. 그러므로, 한국어 어휘개념망은 세밀화된 구조를 가지고 있으므로 자연어 처리 응용 시스템에서 보다 유리한 언어 자원의 역할을 할 수 있다. 어휘개념망에는 동의어와 유의어들이 계층적인 의미 구조와는 별개로 독립적으로 구축되어있다. 어휘개념망은 어휘들간에 구체적인 의미관계와 계층적인 구조가 형성된다는 점은 WordNet이나 시소러스와 비슷하지만 일반적인 개념 구조에 기반한다는 점이 다르다고 할 수 있다.

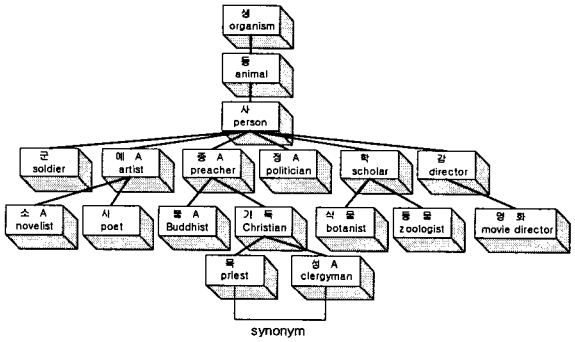
한국어 명사 개념망의 주요 목표는 여러 도메인을 포괄하는 일반개념망을 개발하는 것이었다. 하지만 처음 작업부터 일반화된 개념망을 목표로 작업을 시작하면 작업이 너무 방대해져서 정작 활용하기 어려운 개념망을 구축할 수 있게 되기 때문에 특정 분야의 개념망을 먼저 구축한 다음, 단계적으로 일반 개념망을 구축하였다. 이에 따라 한국어 명사 개념망은 경제 도메인의 개념망 구축으로부터 시작하여 범용의 일반 개념망 구축을 목표로 개발되어왔고, 현재는 일반개념망에 백과사전의 개념 어휘들을 보완함으로써 개념망을 확장, 구축하는 형태로 작업이 진행되고 있다. 그림 1에서와 같이 초기에는 경제개념망 1만5천개의 어휘에서 시작하여 현재 6만개의 일반명사를 포함하여 32만개의 노드로 구성되어 있다.



[그림 1] 명사 개념망의 규모

2.1. 명사 개념망의 구조

어휘개념망의 핵심관계는 "IS-A" 관계인데 그것은 어떤 단어와 단어를 상위어 관계(hyponymy)를 통하여 의미적으로 연결하는 것이다. 어휘개념망에서는 하나의 어휘는 단 하나의 상위어만을 가질 수가 있다. 즉, 그림 2에서와 같이 한 어휘가 다중의 상위어에 연결될 수 없다. 이러한 1대1 매핑 전략은 한 단어의 개념은 단 하나이며 다중 개념은 단일 개념으로부터 파생된다는 논리에서 출발하였다. 현재 명사 개념망은 일반 고빈도 명사와 백과사전 고빈도 명사를 포함한 일반 명사 6만 노드와 고유명사 25만 노드, 백과사전 표제어 1만 노드로 구성되어 있으며 최대 레벨은 12 레벨로 구성되어 있다.



[그림 2] 명사 개념망의 예

2.2. 명사 개념망 구축 프로세스

한국어 명사개념망을 구축하기 위하여 다음과 같은 순차적인 단계가 필요하다.

2.2.1 명사 리스트 작성

명사 리스트를 작성하고 문법과 의미를 분석하기 위하여 기본적인 언어 데이터가 필요하다. 이러한 기본

언어 데이터를 위하여 한국어 사전이 우선적으로 선택되었는데 그 이유는 사전에는 문법적인 정보나 의미 정보는 충분하지 않지만 대량의 어휘 정보가 제공되기 때문이다. 그러므로 명사 개념망은 한국어 사전에 등록된 어휘들과 어휘 정의에 기초하여 구축되기 시작했다. 다음 언어 자원 목록은 명사 개념망을 구축하는데 사용한 기본적인 사전 목록이다.

- 코난시소러스(ETRI) : 금성출판사 국어사전을 이용한 시소러스
- 표준국어대사전(두산동아)
- 연세한국어사전(연세대)
- 국어사전(금성출판사) : 야후 코리아(Yahoo Korea) 금성출판사 국어사전
- 두산세계대백과 사전 : 네이버(NAVER), 야후 코리아(Yahoo Korea)

또한, 다른 연구 결과에서 도출된 문제점이나 쟁점들도 고려되었으며 한국어 언어학자들과의 토론을 통하여 최상의 어휘개념망 구축을 위해 사용되었다.

2.2.2 최상위 개념 결정

명사 개념망의 어휘간 중심 관계는 어휘를 상위어 또는 상위 개념에 IS_A 관계로 연결하는 상하위어 관계이다. 그러므로 개념어의 균형된 분포를 얻기 위해서는 최상위어를 잘 선택해야하는 작업이 필수적이다. 상위어와의 연결이나 균형 조정 작업 전에 먼저 최상위어를 결정하여 개념망의 기본 프레임을 형성해야한다. 최상위어를 결정하기 위하여 ETRI 코난시소러스와 WordNet, 세종계획의 전자 사전 등이 이용되었다. 최상위어 선정을 위한 기초 작업을 통해 개념망의 최상위어 선정 기준을 마련한 것을 정리하면 다음과 같다.[3]

첫째, 최상위어는 사전을 기반으로 하는 만큼 사전에 등재되어 있는 어휘를 사용한다.

둘째, 최상위어는 의미적으로 명확하게 인지되는 어휘를 사용하고, 형태적으로 사람들이 자주 사용하고 인식하는 어휘가 되어야 한다.

셋째, 최상위어는 다른 최상위어와 개념적 중복성이 적어야 한다.

넷째, 최상위어는 하위어의 구성을 고려하여 선택한다.

위와 같은 최상위어 선정 원칙에 따라서 명사 31개의 최상위어가 선택되었다.

2.2.3 상위어와의 연결

어휘간의 상하위어 관계 수립을 위하여 다음과 같은 원칙에 의하여 상위어와의 연결이 이루어졌다.

1. 한자에서 유래된 단일어 어휘는 한자의 핵심 의미에 기초하여 관계 설정한다.
2. 한자 접미사가 붙을 경우에는 그 접미사의 의미를 기준으로 관계설정한다.
3. 어휘의 정의에 상의어의 정보가 있을 때는 해당 상의어에 연결한다.
4. 어휘의 정의에 복수의 상의어 정보가 있을 때는 하나만 선택하여 연결한다.
5. 복합명사의 경우에는 우측 명사의 의미에 의해 연결한다.

2.2.4 형제 노드의 균형 조정

모든 대상 명사에 대한 상하위어 관계가 설정된 후에는 전체 개념망에 대하여 같은 레벨에 존재하는 어휘들이 개념 정도가 너무 일반적이거나 너무 세부적인지를 점검하여 레벨의 균형을 맞춘다. 개념의 균형이 맞지 않을 경우에는 어휘 노드의 위치를 변경하거나 중간 노드를 삽입하는 방법으로 해결한다.

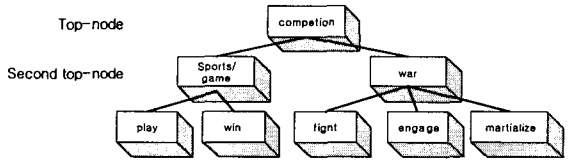
3. 동사 개념망 구축

그동안 한국어 동사에 대한 시소러스나 의미망을 구축하기위한 시도가 있었으나 구축 결과의 규모나 내용면에서 일반 언어 처리 응용에 사용하기에는 만족스럽지 않았다. 한국어 버전의 WrodNet을 위한 영어 WordNet의 번역 프로젝트도 최근에 이루어지고 있는데 한국어 동사로부터 시작한 것이 아니기 때문에 실제 응용에서의 역할에 대해 아직까지 검증된 상태가 아니다.

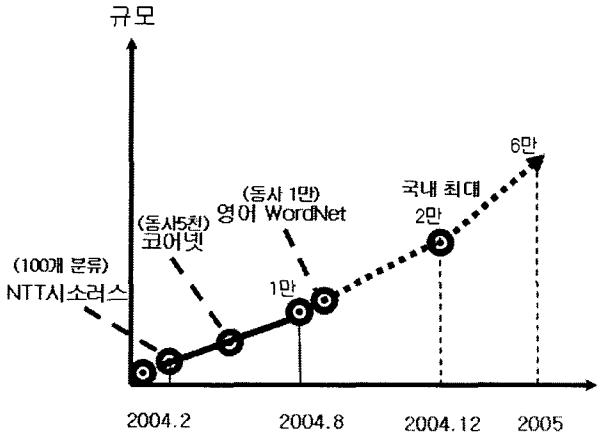
3.1. 동사 개념망 구조

한국어 어휘개념망 중 동사 개념망은 최상위 레벨의 개념과 차상위 레벨의 개념 그리고 하위 어휘들로 이루어진 계층 구조이다. 즉, 그림 3에서와 같이 상위 2개의 개념 레벨을 제외하면 하위 레벨들은 동사 어휘 자체로 이루어져있으며 동사들 간의 상하위 관계에 의해서 연결되어 있다. 동사 개념망은 동사의 정의와 파생동사의 근원, 동사들간의 상하위관계, 동의어, 반의어, 용례 문장과 동사의 격틀과 같은 언어 정보를 제공한다. 그림 4에서와 같이 동사 개념망은 다의어 포함하여 2만여개의 동사 노드로 구성되며 내년까지 6만 어휘까지 확장될 예정이다. 동사 개념망의 평균 노드 깊이는 4.5 레벨이다. [6]

한국어 동사는 기본동사와 파생동사로 분류할 수 있다. 파생 동사는 명사에 "하다" 나 "되다"와 같은 어미를 추가하여 동사로 파생된다. 표1에서와 같이 사전이나 코퍼스에 나타나는 파생 동사의 비율은 60% 이상으로 그 중요도가 크다고 볼 수 있다. 표의 자료는 표준국어대사전과, 세종 계획 일반 코퍼스의 고빈도 동사 그리고 두산 동아 대백과 사전의 고빈도 동사를 대상으로 하였다. 효율적인 동사 개념망의 구축을 위하여 기본 동사와 파생동사는 분리하여 개념망에 연결하였는데 구축과정은 다음 절에서 상세히 설명한다.



[그림 3] 동사 개념망의 예



[그림 4] 동사 개념망의 규모

[표 1] 기본동사와 파생동사의 비율

동사종류	사전	코퍼스	백과사전
기본동사	15000	4739	2861
파생동사	53000	5991	7139
비율	77.94%	55.83%	71.39%

3.2. 동사 개념망 구축 프로세스

한국어 동사개념망을 구축하기 위하여 명사개념망의 구축과 같이 다음과 같은 순차적인 단계를 거쳐야 한다.

3.2.1 동사 리스트 작성

동사리스트는 다음과 같은 언어 자원으로부터 생성하였다. 전체 2만개의 동사가 선택되었으면 다의어를 포함하기 위해 확장되었다.

- 다양한 코퍼스로부터 추출한 국립국어연구원의 고빈도 동사 1000 개.
- 현재 시스템의 이전 버전인 질의응답 시스템 1.0 에서 사용된 유의어 세트에 포함된 동사 1500개.
- 질의응답 시스템의 적용 도메인인 두산동아 대백과사전에서 추출한 고빈도 동사 10000개.

3.2.2 최상위 개념 결정

동사 개념망의 최상위어로 WordNet의 15개 의미 도메인인 state, motion, perception, contact, communication, competition, change, cognition, consumption, creation, emotion, possession, bodily care and functions, social behavior and interaction을 사용하였다. "weather" 도메인은 한국어 동사에 적절하지 않으므로 제외되었다. 최상위 개념과 하위 동사들 사이에 차상위 개념을 두기로 선택한 이유는 질의응답 시스템과 같은 응용 프로그램에서 동사의 그루핑이 활용이 될 수 있기 때문이다. 예를 들어서 문장에서의 주제 추출시에 중간 레벨의 동사 개념을 활용하면 효율적인 주제 추출이 가능하다. 차상위 개념 중에는 "학습", "평가", "예술활동", "경험", "종교" 등이 있다. 차상위 개념을 결정하기 위하여 1400여 동사에 대한 파일롯 스터디를 수행하여 57개의 개념이 선정되었고 개념망 동사의 연결 작업을 하면서 50개의 개념이 추가되었다.

3.2.3 상위어와의 연결

앞절에서 언급한 바와 같이 파생 동사의 비중이 크기 때문에 기본동사와 파생동사의 연결 방법은 다른 접근 방법을 사용하였다. 먼저 기본동사를 개념망에 연결할 때는 동사의 의미에 기준하여 차상위 개념이나 상하위 관계가 성립되는 다른 동사에 연결을 하였다. 반면에 파생동사 리스트에 대해서는 반 자동 연결 방식을 사용하였다.

파생동사는 파생된 명사와 밀접하게 의미적으로 관련되어 있기 때문에 이미 계층 관계가 수립되어 있는 명사 개념망을 이용하여 연결하였다. 즉, 파생동사의 상하위 관계는 명사 개념망으로부터 자동적으로 추출하였다. 명사 개념망의 레벨은 동사에 바로 적용할 경우에 레벨 수가 너무 크기 때문에 중간 패스는 축소하여 최소 5개 레벨로 제한하였다. 동의어/유의어와 같은 다른 어휘관계도 명사로부터 추출하여 직접 동사에 적용하였다. 이와 같은 자동 추출 방법은 대량의 동사 개념망 구축을 가능하게 한다.

3.2.4 기본 동사와 파생 동사의 통합

이전 단계에서 독립적으로 연결된 기본 동사와 파생 동사의 상하위 관계는 통합 단계를 거쳐야한다. 파생

동사의 개념망 연결을 반자동 과정에 의해 수행해야하는 이유는 명사 개념망으로부터 추출한 상하위 관계 패스는 차상위 개념이나 단순동사의 패스와 비교하여 통합해야하기 때문이다. 파생동사의 패스에서 선두에 있는 동사를 차상위 개념과 연결할 수 있는 경우에는 바로 연결과정을 거친다. 이 때 이미 연결된 기본 동사가 있을 경우에는 기본 동사와의 통합 과정이 필요하다. 만일 적합한 차상위 개념이 없을 경우에는 새로운 차상위 개념을 선정해야한다.

3.2.5 동사 정보 저장

어휘간의 상하위 관계가 수립된 후에는 동사의 정의를 사전으로부터 추출하여 데이터베이스에 저장하는데 이 내용은 어휘 중의성 해소나 주제 추적 응용과 같은 질의응답 시스템의 자연어처리 모듈에서 사용된다. 다음 단계로 동의어, 유의어, 사동, 피동 등과 같은 파생동사들의 다른 어휘 관계에 관한 정보들을 명사 개념망으로부터 추출하여 데이터베이스에 저장한다. 기본동사의 관계 정보도 사전으로부터 추출하여 이미 데이터베이스에 포함되어 있는지 확인한 후에 저장한다.

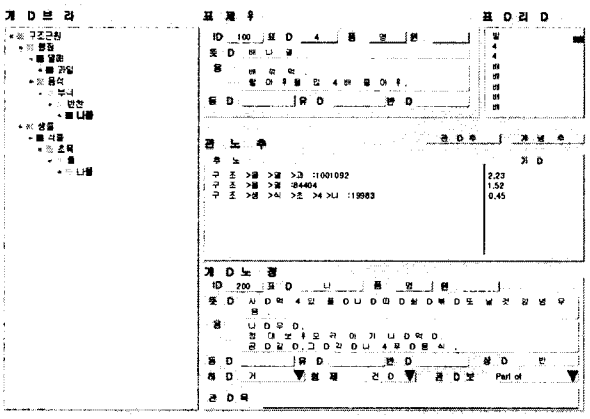
동사 개념망에 추가될 수 있는 기타 다른 정보에는 용례 문장, 격틀 정보, 파생동사의 명사등이 있다. 파생동사의 근원 명사는 명사 개념망과 동사 개념망을 연결해주는 역할을 할 수 있다. 격틀의 논항 정보 역시 명사 개념망의 해당 노드와 연결하여 질의응답 시스템의 구문분석 모듈에서 사용될 수 있다. 이러한 연결 정보의 구축은 향후 연구시에 수행될 예정이다.

4. 어휘개념망 워크벤치

어휘개념망 워크벤치는 한국어 정보처리에서 단어 개념간의 의미의 모호성을 해소하여, 보다 정확한 분석 결과를 사용자에게 제시하기 위하여 한국어 명사들 사이의 관계를 망으로 구축, 관리하는 기능을 제공하는 프로그램이다. 한국어 어휘개념망 구축을 위하여 어휘간의 관계를 네트워크로 구성하기위해 다양한 어휘 관계를 설정하고 연결하기 위한 기능과 데이터 베이스 관리 기능을 지원하는 워크벤치를 개발하여 구축하였다.

개념망 워크벤치는 다음과 같은 세 가지 주요 기능을 가지고 있다. 첫번째 기능은 어휘 리스트를 사전의 형태로 관리하는 기능이다. 두번째 기능은 사전에 있는 어휘를 개념망에 트리의 형태로 연결해주는 네트워크 관리 기능과 개념망 안에 있는 어휘를 찾아주는 검색 기능이다. 세번째 기능으로는 개념망 구축작업의 히스토리를 로그 파일의 형태로 유지하여 수정 구축시에 참고할 수 있도록 하는 관리 기능이 있다. 위와 같은 개념망 워크벤치에서 제공하는 기능들을 이용하여 개념망 작업자는 직접 단어를 등록하고, 단어들 간의 관계를 정의할 수 있으며, 여러 작업자가 병렬로 작업할 때 발생하는 문제를 해결할 수 있다.

현재 새로 개발 중인 개념망 워크벤치는 그림 5에서와 같이 한 어휘의 상위 개념을 추천해주는 보다 지능적인 기능이 포함되어 있어서 향후의 개념망 구축의 정확도와 시간을 단축해 줄 수 있을 것이다.



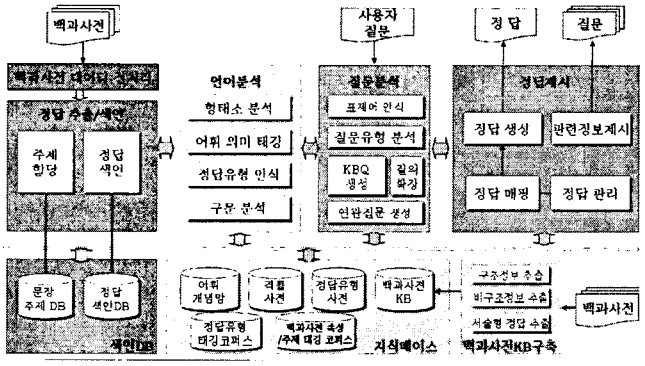
[그림 5] 어휘개념망 워크벤치

5. 어휘개념망의 활용

어휘개념망은 그림 6에서와 같이 백과사전 질의응답 시스템의 각 모듈에서 핵심 언어자원으로 활용되고 있다. 본 절에서는 구축된 어휘개념망을 활용하는 예로써 단락 주제 추출 모듈에서 동사의 동의어/유의어를 이용하여 단락의 주제를 추출하는 방법을 설명한다.

백과사전 질의응답 시스템에서 정답을 추출하기 위한 첫번째 관문은 정답이 들어있는 문서 또는 단락을 검색하는 것이다. 사용자의 질문에 적합한 단락을 제공하기 위해서는 백과사전의 도메인별 특성을 반영하는 주제로 의미적인 단락을 생성하고, 이를 검색하는 방법이 효과적이다[3]. 문장주제를 통한 의미적 단락 생성이란 학습기를 통해 구축된 문장주제 할당기를 통해 해당 주제 범주를 각 문장에 할당한 후, 이를 주제별로 문장을 재구성하여 단락을 생성하는 과정인데, 여기서 가장 중요한 역할을 하는 것이 바로 '문장패턴'이다. '문장패턴'이란 주어진 문장의 주제를 파악할 수 있는 가장 중요한 자질로 동사와 이를 중심으로 주변 명사(NN) 및 개체명 태그(혹은 정답유형 태그)를 셋으로 가진 형태로 정의된다. 예를 들어, 예문 "〈LOC : 하버드대학/nn〉+에서 공부/nc+하/xsv+였으며"에서 추출되는 문장 패턴은 〈LOC, 공부, 하〉이다.

문장패턴에서 주제를 부여하기 위한 가장 중요한 단서는 동사이다. 그러나 일반적으로 학습에 사용되는 문서의 양이 전체 문서집합에 비해 매우 적기 때문에, 학습문서에서 구축된 패턴에 출현한 동사의 양이 매우 적다는데 문제점이 있다. 이와 같이 학습문서에서 나타나지 않은 동사에 대한 적용 범위를 확장시키기 위해 동사 개념망을 활용한다. 동사 개념망은 동사가 하나의 개념노드로 상하위 관계로 연결되어 있으며, 유의어, 동의어, 사동/피동의 관계가 설정되어 있다. 예를 들면 '출생하다'의 경우 유의어로 '태어나다', '탄생하다' 등이 있고, '사망하다'의 유의어로는 '죽다', '급사하다' 등이 있다. 또한 '사망하다'의 피동형인 '암살당하다', '처형당하다' 등을 문장 패턴에 추가함으로써 학습문서에서 추출된 패턴의 적용범위를 넓힐 수 있게 된다. 실험결과, 학습문서에서 구축된 22,916개의 패턴을 동사개념망을 활용한 패턴 확장을 통해 38,214개로 확장시킬 수 있었다.



[그림 6] 질의응답 시스템 구조

6. 어휘개념망 평가

어휘개념망의 평가를 위하여 일반 코퍼스를 이용하여 개념망 사전의 커버리지를 측정하였다. 명사 개념망의 경우에는 사전 어휘 중 30.7%만이 개념망에 연결되어 있으므로 개념망 연결 어휘와 사전 어휘를 분리하여 커버리지를 측정하였다. 커버리지 측정을 위해 사용된 일반 코퍼스는 세종 계획의 의미 분석 말뭉치 400만 어절로 의미 태그가 부착되어 있기 때문에 동형이의어를 구분하여 측정할 수 있다. 세종말뭉치는 1998년부터 2003년까지 년도 별로 구축한 코퍼스로 인문·뉴스·신문잡지·정치분야 현대국어 말뭉치를 종류별로 원시 말뭉치, 형태소 말뭉치, 구문분석 말뭉치, 의미분석 말뭉치 등으로 분류하여 구축하였다. 이중 의미분석 말뭉치는 2001년부터 2003년까지 구축한 것으로 표준국어대사전의 의미 구분자를 사용하였다. (예 : 정원1, 정원2) 전체 300개 파일로 구성되어 있으며 어절수는 약 560만 어절이다.

명사 개념망의 경우, 커버리지 조사 방법은 세종 코퍼스로부터 NNG (명사) 태그가 부착된 어휘를 추출하고 각 어휘의 빈도수를 추출한다. 빈도수 고려한 coverage 프로그램으로 측정하는데 어휘가 출현한 횟수를 고려하여 계산하므로 자주 나오는 어휘는 중요도가 높게 평가된다.

조사 결과 세종 코퍼스에 출현하는 명사 중에서 빈도수 고려하여 96.19%가 개념망 사전에 존재하며 83.43%는 개념망 연결 어휘에 존재한다는 결과가 나왔다. 빈도수를 고려하지 않으면 사전에 없는 명사는 세종 코퍼스 추출 명사인 총 68328개 중에서 36.97%이며 연결 안된 명사는 59.44%이다.

[표 2] 명사 개념망 커버리지 측정 결과

	개념망 사전 어휘	개념망 연결 어휘
Exact match	96.09%	77.92%
의미태그 불문	0.10 %	5.5%
Exact+의미불문	96.19%	83.43%

동사 개념망의 경우, 커버리지 조사 방법은 세종 코퍼스로부터 VV (단순동사)와 NNG+XSV(파생동사) 태그가 부착된 어휘를 추출하고 각 어휘의 빈도수를 추출하여 계산한다.

전체동사를 단순동사와 파생동사로 구분하여 각각의 커퍼리지를 계산하였다. 그 결과로 빈도수를 고려할 때 세종 코퍼스 출현 동사의 98.43 %가 동사 개념망에 존재한다는 것을 알 수 있었다. 단순 동사와 파생동사의 차이는 1.22%로 큰 의미는 없는 것으로 생각된다.

[표 3] 동사 개념망 커버리지 측정 결과

	전체동사	단순동사	파생동사
Exact match	96.55%	96.45%	97.35%
의미태그 불문	1.88%	2.11%	0%
Exact+의미불문	98.43%	98.57%	97.35%

어휘개념망의 커버리지가 나타내는 의미는 어휘개념망에 있는 어휘들이 대상 문서의 어휘에 대하여 어느 정도로 처리할 수 있는지의 정도를 보여줄 수 있다. 그러나 어휘개념망의 성능을 평가하기 위해서는 어휘개념망에 포함되어 있는 언어 정보가 응용 시스템에 어떤 영향을 끼치는지에 대하여 측정해야한다. 즉, 상하관계나 동의어 관계와 같은 어휘간의 관계가 얼마나 정확한지, 이러한 관계 정보가 시스템의 각 모듈의 성능에 얼마나 기여하는지 평가해보아야한다. 커버리지가 양적인 평가의 일부를 담당한다면 질적인 평가를 수행하는 다

른 척도가 필요하며 어휘개념망의 정확한 평가를 위해서 이것은 향후에 필수적으로 연구되어야 할 것이다.

7. 결론 및 향후 연구

본 논문에서는 한국어 명사 개념망과 동사 개념망의 구축과정을 소개하고 어휘개념망의 활용 내용을 주제 추출의 예를 들어 기술하였다. 일반 어휘 데이터베이스에 도메인 관련된 어휘들을 포함함으로써, 지금까지 구축된 어휘개념망은 백과사전 기반의 질의 응답 시스템을 위한 필수적인 언어 자원의 역할을 할 수 있다. 구축된 어휘개념망이 우리의 백과사전 질의 응답 시스템에 어휘 지식 베이스를 제공하는 역할을 주요 목표로 삼고 있지만 정교한 자연어 처리를 요구하는 다른 일반 분야 응용 서비스 시스템에서도 충분히 활용될 수 있을 것이다. 어휘개념망의 활용도를 높이기 위해서 향후에는 명사 개념망과 동사 개념망을 연결하여 격틀에 관련된 정보를 제공하는 등의 통합과정에 관한 연구가 예정되어 있다. 또한 형용사나 부사와 같은 다른 내용어와의 연결과 더불어 동사 개념망의 동사구를 포함하는 확장 분야도 미래의 연구로 고려하고 있다.

참고 문헌

[1] Lenat, Doug, George Miller, and Toshio Yokoi. 1995. "CYC, WordNet, and EDR : Critiques and Responses," *Communications of the ACM* 38(11) : 45-48.

[2] Yokoi, Toshio. 1995 "The EDR Electronic Dictionary," *Communications of the ACM* 38(11) : 42-44.

[3] Wang, JiHyun and Jang, Myung-Gil. 2003 ``Study on Construction of Korean Noun Concept Network for Information Retrieval", *First International Forum of Korean Thesaurus Proc.*,.

[4] 황이규, 김현진, 장명길, "질의응답 기술 개발", 한국정보처리학회지, VOL. 11 NO. 02, 2004,03

[5] Chung-Hee Lee, Hyo-Jung Oh, Hyeon-Jin Kim, Myung-Gil Jang, "Semantic Passage Retrieval for Encyclopedia QA System," to appear in AIRS 2004 proc.

[6] 최승권, 김현숙, 최미란 "연구센터 탐방 : ETRI 음성/언어정보연구부", 한국정보처리학회지, VOL. 11 NO. 02, 2004,03

[7] Resnik, P. 1995b. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy." In Proceedings of IJCAI-95, Montreal, Canada, pp.448-453.

[8] Fellbaum, Christiane. edited 1999, *WordNet, an Electronic Lexical Database.* The MIT Press.

[9] Lee, Wang-Woo, et al. 2001. "An Improved Homonym Disambiguation Model based on Bayes Theory", *13th Hangeul and Korean Information Processing Proc.*, pp.465-471....