한국어 어휘 지식 베이스 구축 시스템*

이해중, 조정미, †문준혁, 서정연
서강대학교 컴퓨터 학과

Korean Lexical Knowledge Base Construction System

Haejoong Lee, Jeong Mi Cho, Jungyun Seo
Department of Computer Science, Sogang University

요약

어휘 지식은 자연어 처리에서 매우 중요한 요소이다. 그러나 대규모의 어휘 지식 베이스를 구축하는 것은 많은 시간과 비용을 필요로하는 일이다. 본 논문에서는 온라인 국어 사전을 이용하여 범용의 대규모 한국어 어휘 지식 베이스를 자동으로 구축하는 방법을 제안하고 실제로 시스템을 구현한다. 제안하는 방법론은 비교적 적은 비용으로 단시일 내에 대규모의 어휘 지식 베이스를 구축하는 것을 가능하게 한다. 또한 지식 구축 과정이 자동화 되어 만들어진 지식 베이스의 유지, 보수 및 확장이 용이하다. 구현된 시스템으로 구축한 어휘 지식 베이스는 기계번역에서의 대역어 선정이나 한국어 조사의 의미 분별 등 자연어 처리 과정에서 발생하는 각종 어휘 의미 모호성 해소에 응용될 수 있다.

1. 서론

어휘 지식 베이스는 자연어 이해 및 처리에 필요한 다양한 어휘 정보를 제공하는 일종의 지식 베이스이다. 이러한 어휘 지식 베이스는 견고하고 실용적인 자연어 처리 시스템 개발에 필수적인 요소이다. 21 세기 정보화 시대를 맞이하여 자연어 처리 시스템은 보다 다양하고 방대한 문서 처리가 요구되고 있다. 이에 따라 보다 다양한 어휘에 대처할 수 있는 어휘 지식 베이스에 대한 요구도 점점 커지고 있다. 본 논문은 대규모의 한국어 어휘 지식 베이스를 구축할 수 있는 토대를 마련하고자 한 것이다.

본 논문에서는 온라인 국어 사전을 이용하여 자동으로 한국어 어휘 지식 베이스를 구축하는 시스템을 구현한다. 어휘 지식을 구축하는 방법은 구축 과정의 자동화 정도에 따라 크게 세 가지로 나누어 생각할 수 있다. 먼저, 완전히 사람에 의존하는 방법으로서 전문가가 수작업으로 직접 어휘 지식을 구축하는 것이다. 이 방법의 장점은 어휘 지식 베이스에 포함된 정보가 매우 정확하다는 것이다. 그러나 여기에는 많은 전문 인력과 시간이 필요하고, 따라서 비용도 많이 들게 된다. 또한 수정이나 확장이 매우 어렵다. 두번째 방법은 시스템에 의해 지식 구축 과정이 진행되는 동안 시스템이 판단할 수 없는 모호한 사항들만 사람이 결정해 주는 반자동 방식이다. 이 방법은 첫번째 방법에 비해 시간과 비용이 적게 들고 사람이 관여하므로 포함된 정보의 질도 그리 나쁘지 않다는 장점을 가지고 있다. 그러나 대규모의 어휘 지식 베이스를 만드는 데는 여전히 많은 시간과 비용이 소요된다. 세번째 방법은 컴퓨터 프로그램을 이용하여 어휘 지식 구축 과정을 자동화 하는 방법이다. 이렇게 함으로써 지식 구축에 드는 비용과 시간을 획기적으로 줄일 수 있다. 또한, 지식 구축 과정에서 일관성을 유지할 수 있으며 지식 베이스의 수정 및 확장이 용이하게 된다. 구축된 지식의 질은 다소 떨어지지만 범용의 대규모 어휘 지식 베이스를 구축하기 위해서는 자동화된 방법론을 사용하는 것이 가장 효과적이

* 본 연구는 부분적으로 과학재단 특정기초연구사업 중 "통계적 담화분석"에 대한 지원으로 이루어진 것임.

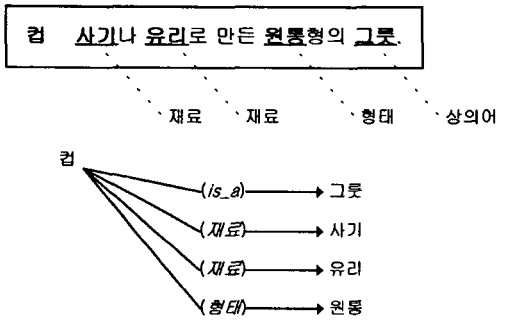
라 할 수 있다. 따라서 본 논문에서 제시하는 한국어 어휘 지식 베이스 구축 시스템은 자동화를 지향한다.

사전은 단어의 뜻, 품사, 용법, 불규칙 활용 등 그 단어에 관한 여러가지 정보를 포함하고 있다. 사전은 또, 일반적인 문서에서 나타날 수 있는 거의 모든 어휘를 망라한다. 따라서 사전은 대규모의 어휘 지식 베이스를 위한 훌륭한 지식원(knowledge source)이 된다. 본 논문에서는 「우리말큰사전」을 지식원으로 삼아서 품사, 불규칙 활용 정보, 유의어, 반의어 등과 단어에 대한 의미 정보를 포함하는 어휘 지식 베이스를 구축한다. 단어에 대한 의미 정보는 단어의 뜻 풀이말로부터 추출하며, 이것은 자연어 처리의 각 분야에서 각종 모호성을 해소하는 데 유용하게 사용할 수 있다.

2. 단어의 뜻 풀이말과 의미 정보

사전의 가장 중요한 기능은 사용자에게 단어의 뜻을 전달하는 것이며 단어의 뜻은 뜻 풀이말이라는 구 또는 문장으로 사전에 나타난다. 따라서 단어의 의미와 관련된 중요하고도 유용한 정보는 대부분 뜻 풀이말에 포함된다. 본 논문에서 추출하고자 하는 단어의 의미 정보는 표제어가 뜻 풀이말에 나타나는 단어 중 어떤 단어와 관련이 있는지, 그들 간의 의미 관계는 무엇인지에 관한 것이다. 또는 뜻 풀이말 속에 있는 단어들은 어떤 단어와 관련이 있으며, 그들끼리의 의미 관계는 무엇인지에 관한 것이다.

[그림 1]은 사전에 나타난 **컵**의 뜻 풀이말과 거기에서 추출할 수 있는 의미 정보를 나타낸 것이다.



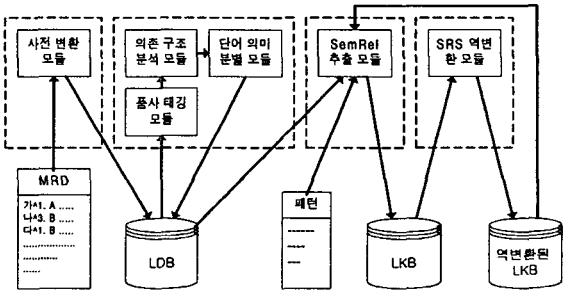
[그림 1] 컵의 뜻 풀이말에서 추출한 의미 정보

[그림 1]의 아래에 있는 트리 모양의 구조가 구하고자하는 컵의 의미 정보이다. **컵**은 **그릇**, **사기**, **유리**, **원통** 등과 같은 단어들과 관련을 맺고 있으며 각각의 의미 관계는 *is_a*, *재료*, *재료*, *형태*와 같다. 여기에서 전체 트리를 이루고 있는 **단어**—(*의미관계*)→**단어**와 같은 단위를 SemRel(Semantic Relation)이라고 하고, 전체 트리를 SRS(SemRel Structure)라고 한다[Richardson, 1997].

뜻 풀이말로부터 SemRel을 추출하는 방법은 특정 패턴을 이용하는 것이다. 예를 들어 **사전의 명사 뜻 풀이말의 끝 단어는 대개 표제어의 상의어(hypernym)가 된다.** 이것이 하나의 패턴이 되는데, [그림 1]의 뜻 풀이말에 이 패턴을 적용하면 **컵**—(*is_a*)→**그릇**과 같은 SemRel을 추출할 수 있다.

3. 한국어 어휘 지식 베이스 구축 시스템

[그림 2]는 본 논문에서 구현하고자 하는 한국어 어휘 지식 베이스 구축 시스템의 구조도이다. MRD(Machine Readable Dictionary)는 기계 가독 형태의 사전을 말하며, 본 논문에서는 텍스트 화일로 되어있는 「우리말큰사전」을 사용하였다. LDB(Lexical Data Base)는 MRD를 데이터베이스로 변환한 것이다. 본 논문에서는 「우리말큰사전」을 PostgreSQL RDBMS의 테이블로 변환하였다. 여기에 뜻 풀이말로부터 얻은 단어의 의미 정보를 추가하여 LKB(Lexical Knowledge Base), 즉 어휘 지식 베이스를 완성한다. 뜻 풀이말로부터 의미 정보를 추출할 때에는 패턴을 사용한다. [그림 2]의 패턴은 2 절의 끝에서 예로 들었던 것과 같이 말로 표현된 패턴을 좀 더 정형화한 것으로서 SemRel 추출 모듈의 입력이 된다. LKB를 보다 효율적으로 사용할 수 있도록 LKB를 역변환(inversion)하며, SemRel 추출 과정에 다시 사용하여 보다 정확한 SemRel을 추출할 수 있도록 한다.

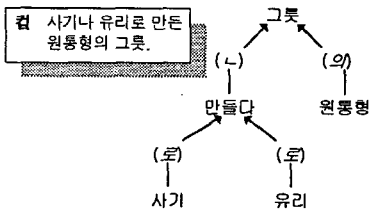


[그림 2] 시스템 구조도

3.1. LDB 구축

「우리말큰사전」은 표제어 수만 40 만 개에 달하는 방대한 사전이다. 이 사전의 엔트리들은 28 가지의 정보를 각종 기호를 사용하여 나타내고 있다. 각각의 표제어는 하나 이상의 뜻을 가지며 각각의 뜻을 의미 레코드라는 LDB의 레코드로 변환한다. 즉, 하나의 의미 레코드는 하나의 단어 의미를 갖는다. 의미 레코드에는 「우리말큰사전」에 나타나는 총 28 개의 정보 중에서 일련번호, 표제어, 품사, 동음이의어 번호, 의미 번호, 불규칙 정보, 분야 정보, 뜻 풀이말, 예문 등 24 가지의 정보를 포함시킨다. 또한 LDB에는 명사, 동사, 형용사, 부사, 관형사 등의 내용어만 포함시키고, 방언이나 사람의 성, 뜻 풀이말이 없으면서 잘못 사용된 단어로 표시된 것들은 제외한다. Perl로 구현한 사전 변환 프로그램을 이용하여 「우리말큰사전」의 엔트리들은 총 398,930 개의 의미 레코드(322,746 개의 뜻 풀이말)로 변환되었다.

의미 레코드의 뜻 풀이말은 SemRel 추출을 위하여 품사 태깅하고 의존 구조 분석해야 한다. 뜻 풀이말은 KTS[이상호, 1992]를 이용하여 품사 태깅한다. 품사 태깅된 뜻 풀이말은 간단한 의존 규칙과, 수식할 수 있는 가장 가까운 어절을 수식하도록 하는 휴리스틱을 이용하여 의존 구조 분석한다. 구현된 의존 구조 분석기는 또한 간단하게나마 등위 접속된 어절들을 고려하여 처리한다. [그림 3]은 **컵**의 뜻 풀이말을 의존 구조 분석한 결과이다.



[그림 3] 컵의 뜻 풀이말을 의존 구조 분석한 결과

본 논문에서는 [그림 3]과 같은 의존 구조 분석 결과에 나타나는 **단어**—(*조사/어미*)→**단어**와 같은 것도 넓은 의미에서 SemRel인 것으로 간주한다. 이는, 기능어가 발달한 한국어의 특성을 고려해 볼 때, 이러한 SemRel도 단어 간의 의미 관계를 설명하는데 중요한 역할을 할 수 있기 때문이다.

의존 구조 분석까지 마친 뜻 풀이말은 단어 의미 분별 과정을 거친다. 뜻 풀이말을 의미 분별하는 이유는 뒤에서 밝히기로 한다. 본 연구에서는 [Karov, 1998]를 기반으로 단어 의미 분별 프로그램을 구현하였으며, 뜻 풀이말의 명사, 동사, 형용사를 「우리말큰사전」의 의미로 의미 분별하였다.

위의 모든 과정을 거친 의미 레코드의 뜻 풀이말은 다시 LDB에 저장되어 SemRel을 추출할 때 사용된다.

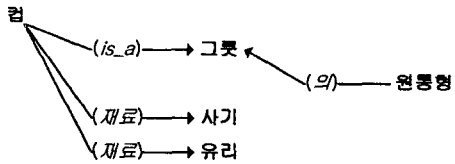
3.2. SemRel 추출

패턴은 SemRel 추출 프로그램으로 하여금 특정 의미 관계를 찾아서 SemRel을 추출하게 하는 지시서이다. [그림 4]는 각각 *is_a*, *재료* 의미 관계를 찾는 패턴의 예 이다. 패턴은 조건부와 실행부로 이루어진다. 조건부는 하나 이상의 SemRel들이 연결된 형태로 표현되며, 실행부에는 하나 이상의 SemRel이 나열된다. 조건부에 나타난 패턴이 뜻 풀이말의 의존 구조 분석 결과와 매치되면, 실행부에 표시된 SemRel이 만들어진다. 그림에서 **[ROOT]**는 의존 구조 분석 결과의 루트 노드를 의미 하고, **[A]**는 패턴의 조건 부분이 의존 구조 분석 결과와 매치될 때 그 자리에 매치되는 단어를 의미한다.

조건부: 무조건
실행부: **[표제어]**—(*is_a*)→**[ROOT]**

조건부: **[A]**—(로)→**만들다**—(ㄴ)→**[ROOT]**
실행부: **[표제어]**—(*재료*)→**[A]**

[그림 4] 패턴의 예



[그림 5] 패턴 매치된 SRS

[그림 5]는 [그림 3]의 의존 구조 분석 결과에 [그림 4]의 패턴을 매치시킨 결과이다. 원래의 의존 구조 분석 결과에서 패턴과 일치하는 '로 만든' 부분이 삭제되고 재료 의미 관계에 의해 연결된 **사기**—(*재료*)→**컵** 등의 SemRel이 추가되어 새로운 SRS가 만들어진 것을

볼 수 있다.

여러가지 의미 관계 중에서 *is_a* 의미 관계는 가장 기본적이다. 또한 그 수도 많아서 사전으로부터 추출한 단어의 의미 정보의 질을 크게 좌우한다. 따라서 *is_a* 의미 관계를 추출할 때는 주의를 요한다. 대부분의 명사 뜻 풀이말에서는 뜻 풀이말의 끝 단어가 표제어의 상의어가 된다. 그러므로 [그림 4]의 *is_a* 패턴은 대부분의 경우 잘 들어 맞는다. 그러나 [그림 6]과 같은 뜻 풀이말에 이 패턴을 적용하면 **용틀임**—(*is_a*)→**하나**, **분생자**—(*is_a*)→**일종**과 같은 부적절한 SemRel이 만들어진다.

아연판 인쇄판의 한 가지.
용틀임 양주별산대놀이 춤사위의 하나.
분생자 균류의 균사 끝에 생기는 포자의 일종.
품류 물건의 갖가지 종류.

[그림 6] 끝 단어가 상의어가 아닌 뜻 풀이말

이러한 경우 무작정 뜻 풀이말의 끝 단어를 상의어로 결정하는 것이 아니라, [그림 7]과 같은 패턴을 적용하여 올바른 상의어를 선택한다. 이렇게 함으로써 **용틀임**—(*is_a*)→**춤사위**, **분생자**—(*is_a*)→**포자**와 같은 올바른 SemRel을 추출한다. 패턴을 적용할 때는 [그림 7]과 같은 패턴을 먼저 적용하고, [그림 4]의 *is_a* 패턴을 나중에 적용하도록 한다.

조건부: [A]—(*의*)→**하나**
[ROOT] = **하나**
실행부: [**표제어**]—(*is_a*)→[A]

조건부: [A]—(*의*)→**일종**
[ROOT] = **일종**
실행부: [**표제어**]—(*is_a*)→[A]

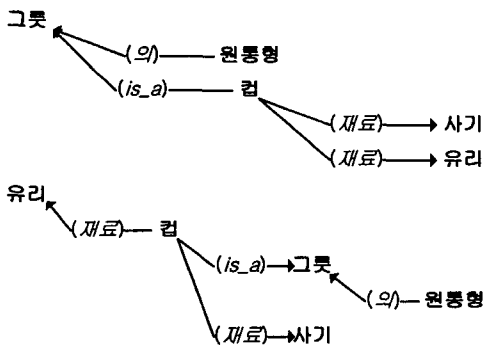
[그림 7] 올바른 상의어를 찾는 패턴

이와 같은 예외 처리를 하더라도, 뜻 풀이말에서 상의어를 찾을 수 없는 경우는 많다. 특히 4 자 성어나 한자로 된 단어의 경우는 뜻 풀이말이 '용언 어간 + ㅁ/exm'과 같은 형태로 끝나는 것이 많으며, 이러한 경우 표제어의 상의어를 찾기란 매우 어렵다. 이와 같은 문제를 포함하여, 패턴을 만드는 일은 어휘 지식 베이스를 구축하는 과정에서 가장 중요하고도 어려운 작업이다.

3.3. SRS 역변환

3.2.절의 과정을 거치면 각각의 뜻 풀이말에 대하여 단어 간의 의미 관계가 추가된 SRS가 하나씩 만들어지게 된다. 이로써 한국어 어휘 지식 베이스는 사실상 완성된다. 만들어진 어휘 지식 베이스는 한 단어에 대한 일반적인 정보를 알아보는 데 사용될 수도 있고, 한 단어가 다른 단어와 어떤 의미 관계에 의해 관련되어 있는지를 알아보는 데도 쓰인다. 후자의 경우와 같이 어떤 단어 w 가 어떤 단어들과 관련을 맺고 있는지 알아보려면 LKB의 각각의 SRS가 w 를 포함하고 있는지 일일이 확인해야 한다. 이것은 부하가 큰 작업이다. 따라서 SRS를 역변환함으로써 이러한 과부하를 줄이고자 한다.

SRS를 역변환하는 것은 SRS에 있는 루트 노드 이외의 노드들을 루트 노드로 하는 새로운 SRS를 만드는 것이다[Richardson, 1997]. 이 때 SRS의 위상은 변함이 없으며 단지 루트 노드만 달라지게 된다. [그림 8]은 [그림 5]의 SRS를 역변환하여 만들어진 SRS 중의 일부를 보인 것이다.



[그림 8] 역변환된 SRS

역변환된 LKB에서 어떤 단어 w 와 관련된 단어들을 검색하는 것은 매우 간단하다. 단순히 SQL의 SELECT 문을 이용하여 루트 노드를 w 로 하는 SRS들을 검색하여, 이들 SRS의 루트 노드와 연결된 단어들만 찾으면 되는 것이다. 이 때 단어 의미 분별의 필요성이 대두된다. 많은 단어들이 의미 중의성을 내포하고 있으며 따라서 올바른 SRS를 검색하기 위해서는 검색하는 단

어의 의미까지 명시해줘야 한다. 예를 들어 car 라는 의미의 차를 검색하는데 tea 라는 의미의 차까지 검색되어 나온다면 검색된 정보의 신뢰도는 현격히 떨어지게 된다. 현재 단어 의미 분별의 정확성은 매우 저조한 편이며 앞으로 여기에 대한 대책이 반드시 필요하다.

역변환 과정까지 마치면 LKB 구축의 첫번째 싸이클은 완전히 끝난다. 첫번째 싸이클에서 얻어진 의미 정보는 SemRel 추출 과정에서 다시 사용됨으로써 다음 싸이클에서 추가적인 그리고 보다 정확한 SemRel을 추출하는데 도움을 준다.

3.4. 토의

본 논문에서 제시하는 시스템으로 구축되는 LKB는 커다란 네트웍과도 같다. 먼저 각각의 뜻 풀이말로부터 만들어지는 SRS는 각기 다른 단어들이 연결된 작은 네트웍을 이룬다. 이 작은 네트웍을 이루는 단어들은 다시, LKB의 다른 곳에서 다른 단어들과 연결되어 있다. 따라서 본 논문에서 구축하는 어휘 지식 베이스는 하나의 커다란 네트웍이라고 할 수 있다. [Dolan, 1993]은 사전은 단어들이 다양한 의미 관계에 의해 긴밀하게 연결된 단어들의 네트웍으로 보았으며, 어휘 지식 베이스를 구축하는 과정은 사전에 숨겨져 있는 이러한 네트웍을 밖으로 드러내는 과정으로 보았다.

이러한 네트웍을 어떻게 이용할 것인가 하는 것은 매우 중요한 문제이다. 본 논문에서 이에 대한 명확한 방안을 제시하지는 못했지만 영어권에서는 이에 대한 활발한 연구가 있었다. [Richardson, 1997]은 SemRel에 가중치를 부여하는 등 LKB를 좀 더 가공하여 여러 개의 SemRel로 연결된 두 단어의 의미 관계를 추론하는 방법을 고안하고, 이를 바탕으로 두 단어의 유사성을 판단하는 시스템을 개발하였다. [Vanderwende, 1994]는 단어 간의 의미 관계를 이용하여 두 단어로 이루어진 복합 명사의 의미를 해석하는 방법을 제시하였다. 또한 Microsoft 사의 자연어 처리 연구 팀은 본 논문에서 제시하는 것과 같은 LKB인 MindNet을 개발하여, 자사의 자연어 처리 시스템의 기본 어휘 지식 베이스로 활용하고 있다[Richardson, 1998].

4. 실험

본 논문에서 구현한 한국어 어휘 지식 베이스 구축 시스템을 이용하여 실험적인 LKB를 구축하고, 구축한 의미 지식의 정확성을 측정하는 실험을 실시하였다.

실험에 사용한 사전은 「우리말큰사전」 LDB로서 자세한 사항은 [표 1]에 정리하였다. 7 가지의 의미 관계를 대상으로 하였으며 [표 2]에 이에 대한 사항을 정리하였다. SemRel 추출을 위하여 7 가지의 의미 관계에 대한 90 개의 패턴을 작성하였다.

[표 1] 실험에 사용한 사전

	명사	동사	계
의미 레코드 수	278,863	53,225	332,088
뜻 풀이말 수	220,118	20,789	240,907

「우리말큰사전」 LDB에서 표제어의 품사가 명사나 동사인 의미 레코드를 실험 대상으로 함.

[표 2] 실험에 사용한 의미 관계들

의미 관계	의미	SemRel 예
is_a	상의어	**코끼리—(*is_a*)→포유동물**
part_of	사물의 일부분	**꽃잎—(*part_of*)→꽃**
이유		**바쁘다—(*이유*)→일**
재료		**위스키—(*재료*)→보리, 밀, 수수**
용도		**칼—(*용도*)→베다, 썰다**
syn	유사어	**나르다—(*syn*)→운반하다**
equ	같은 말	**각설이—(*equ*)→장타령꾼**

추출한 의미 관계들 중 *syn* 의미 관계는 사전에 명시되어 있는 유사어 정보를 그대로 이용하였으므로, 90 개의 패턴 중 *syn* 의미 관계를 추출하기 위한 패턴은 없다. *equ* 의미 관계도 사전에 이미 포함되어 있는 정보를 사용하였으며 일부는 패턴을 이용하여 추출하였다.

[표 3] SemRel의 정확성 측정 실험 결과

의미 관계	전체 개수	측정에 사용된 샘플 수	O / X	정확성
is_a	234,673	298	237 / 61	79.5 %
equ	75,614	92	91 / 1	98.9 %
syn	74,061	100	100 / 0	100.0 %
part_of	978	2	1 / 1	
용도	3,626	4	2 / 2	
재료	2,823	3	3 / 0	
이유	427	1	1 / 0	
전체	392,202	500	435 / 65	83.7 %

[표 4] *part_of, 용도, 재료, 이유* SemRel의 정확성

의미 관계	측정에 사용된 샘플 수	O/X	정확성
part_of	50	33/17	66.0 %
용도	50	43/7	86.0 %
재료	50	42/8	84.0 %
이유	50	39/11	78.0 %

구축된 LKB에 포함된 SemRel은 총 392,202 개이며, 이들 중 500 개의 샘플을 무작위로 추출하여 정확성을 측정하였다. 정확성 측정 결과 LKB에 포함된 SemRel의 정확성은 83.7±3% (신뢰도 95%)인 것으로 나타났다. [표 3]은 이 실험 결과를 정리한 것이다. 이 실험에서는 정확성 측정에 사용한 샘플 중 part_of, 용도, 재료, 이유 등의 의미 관계에 대한 SemRel의 수가 너무 적어서 이들의 정확성을 가늠하기 어렵다. 따라서 이들 의미 관계에 대한 SemRel을 각각 50 개씩 추출하여 정확성을 측정하였다. [표 4]에 이 실험의 결과를 정리하였다.

[표 5]는 첫번째 실험의 *is_a* SemRel 추출 과정에서 발생하는 오류를 유형별로 정리한 것이다. is_a SemRel 추출 과정에서 발생하는 오류의 대부분은 아래의 예처럼 한자어로 되고 행위나 상태를 나타내는 단어에서 발생하였다.

> **무선** 전선을 가설함이 없음.
> **박두** 기일이나 시기가 가까이 닥침.
> **반박** 남의 의견이나 주장에 반대하여 공격함.

이러한 단어의 뜻 풀이말의 특징은 뜻 풀이말이 명사형 전성어미 ㅁ으로 끝난다는 것인데, 이들 뜻 풀이말에서는 표제어에 대한 적절한 상의어나 하의어를 찾을 수가 없다.

[표 5] *is_a* SemRel에서의 오류

오류 유형		발생 빈도
상·하의어를 찾을 수 없음	한자어	38 (62.3 %)
	기타	8 (13.1 %)
패턴이 없음		11 (18.0 %)
품사 태깅 오류		1 (1.6 %)
의존 구조 분석 오류		3 (4.9 %)
계		61 (100 %)

[표 6] *part_of, 용도, 재료, 이유* SemRel에서의 오류

오류 유형	발생 빈도
부적절한 패턴	40
품사 태깅 오류	1
의존 구조 분석 오류	2

[표 6]은 두번째 실험의 *part_of, 용도, 재료, 이유* SemRel 추출 과정에서 발생한 오류들을 정리한 것이다. 이들 오류들은 대부분 패턴이 부적절하여 발생한다. 예를 들어 'A의 부분'과 같은 패턴은 *part_of* SemRel을 추출하는 데 사용되어서 **[표제어]**—(*part_of*)→**[A]**와 같은 SemRel을 추출하게 된다. 이것은 아래와 같은 뜻 풀이말에 적용되어 **늑경골**—(*part_of*)→**갈비뼈**라는 올바른 SemRel을 추출한다.

> **늑경골** 단단한 뼈로 된 갈비뼈의 한 부분.

그러나, 이 패턴은 아래와 같은 뜻 풀이말에서는 **무종아리**—(*part_of*)→**사이**와 같은 잘못된 SemRel을 추출한다.

> **무종아리** 발뒤꿈치와 장딴지 사이의 부분.

이러한 문제는 패턴을 좀 더 정교하게 만듦으로써 해결할 수 밖에 없다.

5. 결론 및 향후 과제

본 논문에서는 온라인 국어 사전을 이용하여 대규모의 한국어 어휘 지식 베이스를 자동으로 구축하는 방법을 제안하고, 그 방법론에 따라 시스템을 구현하였다. 구현한 어휘 지식 베이스 구축 시스템을 이용하여 7 가지 의미 관계에 대한 실험적인 어휘 지식 베이스를 구축하였으며, 실험 결과 구축된 어휘 지식 베이스에 포함된 단어 의미 지식의 정확성은 83.7±3% (신뢰도 95%)인 것으로 나타났다.

본 논문에서 제안한 한국어 LKB 구축 방법론의 장점은 다음과 같다. 첫째, 품사 태거와 파서를 이용하여 적은 비용과 시간을 들여서 대규모의 LKB를 구축할 수 있다. 둘째, 구축 과정이 모두 자동화 되어 있어서 LKB를 수정·확장하기가 쉽고 유지·관리하기가 쉽다. 셋째, MRD의 특성에 거의 얽매이지 않으므로 여러가지의 사전을 통합하여 하나의 LKB를 구축할 수 있다. 넷째, 구축된 지식을 재활용하여 점진적인 LKB 구축이

가능하다.

실제 시스템에 이용할 수 있는 LKB를 위해서는 첫째, 어휘 지식의 질을 향상시킬 수 있도록 노력해야 한다. 이를 위해서는 보다 견고하고 정확한 품사 태거 및 의존 구조 분석기가 필요하고, 양질의 패턴이 만들어져야 한다. 둘째, 어휘 지식의 양이 훨씬 많아져야 한다. 이를 위해서는 보다 다양한 의미 관계에 대한 보다 많은 패턴이 있어야 한다. 셋째, 구축된 LKB를 활용하는 추론 기법이 개발되어야 한다. LKB의 단어 간 의미 관계 정보는 그 자체로도 응용될 수 있겠지만, 보다 고급의 응용을 위해서는 LKB를 활용하는 추론 기법 개발이 필수적이다. 넷째, 보다 정확한 단어 의미 분별 시스템이 필요하다. 단어 의미 분별은 앞에서도 설명했듯이 LKB 활용에 있어서 필수적인 요소이다.

참고 문헌

[이상호,1992] 이상호, 미등록어를 고려한 한국어 품사 태깅 시스템 구현, 석사학위 논문, 한국과학기술원, 1992.

[이해중,1999] 이해중, 온라인 국어 사전을 이용한 한국어 Lexical Knowlwdge Base 구축, 석사학위 논문, 서강대학교, 1999.

[Dolan,1993] W.B. Dolan, L. Vanderwende, S.D. Richardson, "Automatically deriving structured knowledge base from on-line dictionaries," In *Proceedings of the Pacific Association for Computational Linguistics*, pp.5-14, 1993.

[Karov,1998] Y. Karov, S. Edelman, *Similarity-based word sense disambiguation*, Computational Linguistics, 1998.

[Richardson,1998] S.D. Richardson, W.B. Dolan, L. Vanderwende, "MindNet: acquiring and structuring semantic information from text," In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, 1998.

[Richardson,1997] S.D. Richardson, *Determining similarity and inferring relations in a lexical knowledge base*, Ph.D. dissertation, City University of New York, 1997.

[Vanderwende,1994] L. Vanderwende, "Algorithm for interpretation of noun sequence," In *Proceedings of the International Conference on Computational Linguistics (COLING-94)*, 1994.