대규모 말뭉치와 전산 언어 사전을 이용한 의미역 결정 규칙의 구축

강 신 재† · 박 정 혜††

요 약

본 논문은 한국어정보처리 과정에서 구문 관계를 의미역으로 사상시키기 위한 규칙을 효과적으로 구축하는 방법을 제시하고 있다. 의미역의 결정은 의미 분석의 핵심 작업 중 하나이며 자연어처리에서 해결해야 하는 매우 중요한 문제 중 하나이다. 일반적인 언어학 지식과 경험만 가지고 의미역 결정 규칙을 기술하는 것은 작업자의 주관에 따라 결과가 많이 달라질 수 있으며, 또 모든 경우를 다룰 수 있는 규칙의 구축은 불가능하다. 하지만 본 논문에서 제시하는 방법은 대량의 원시 말뭉치를 분석하여 실제 언어의 다양한 사용례를 반영하며, 또 수십 명의 한국어 학자들이 심도 있게 구축하고 있는 세종전자사전의 격틀 정보도 함께 고려하기 때문에 보다 객관적이고 효율적인 방법이라 할 수 있다. 의미역을 보다 정확하게 결정하기 위해 구문관계, 의미부류, 형태소 정보, 이중주어의 위치정보 등의 자질 정보를 사용하였으며, 특히 의미부류의 사용으로 인해 규칙의 적용률이 향상되는 효과를 가져올 수 있었다.

Rule Construction for Determination of Thematic Roles by Using Large Corpora and Computational Dictionaries

Sin-Jae Kang† · Jung-Hye Park††

ABSTRACT

This paper presents an efficient construction method of determination rules of thematic roles from syntactic relations in Korean language processing. This process is one of the main core of semantic analysis and an important issue to be solved in natural language processing. It is problematic to describe rules for determining thematic roles by only using general linguistic knowledge and experience, since the final result may be different according to the subjective views of researchers, and it is impossible to construct rules to cover all cases. However, our method is objective and efficient by considering large corpora, which contain practical usages of Korean language, and case frames in the Sejong Electronic Lexicon of Korean, which is being developed by dozens of Korean linguistic researchers. To determine thematic roles more correctly, our system uses syntactic relations, semantic classes, morpheme information, position of double subject. Especially by using semantic classes, we can increase the applicability of the rules.

키워드 : 의미역(Thematic Roles), 의미 분석(Semantic Analysis), 말뭉치 분석(Corpus Analysis), 세종 전자사전(Sejong Electronic Lexicon of Korean), 기계번역(Machine Translation)

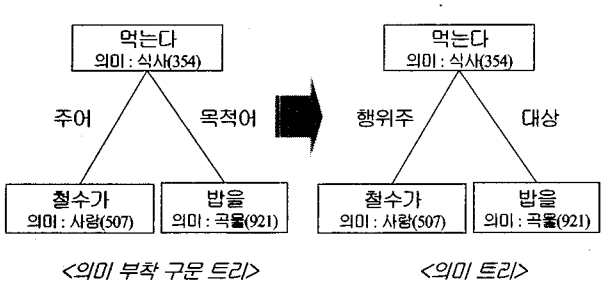
1. 서 론

최근 널리 보급된 인터넷과 통신망에서 언어적 장벽을 극복하고, 또 대량의 정보 중에서 필요한 정보를 정확하고 빠르게 습득하기 위해서는 자연언어처리 기반 기술의 확보가 필수적이라 할 수 있다. 일반적으로 언어를 분석할 때는 형태소 분석과 구문 분석의 과정을 거쳐 의미 분석을 하게 되는데, 의미 분석에서는 단어의 의미 중의성 해소(word sense disambiguation)와 단어간 의미역(thematic role)의 결정이 주요한 작업이다. 이러한 여러 과정 가운데 본 논문에서는 의미 분석에서의 의미역 결정에 대해 논하고자 한다.

일반적으로 의미역의 결정은 격틀(case frame)과 같은 언어 지식(linguistic knowledge)을 필요로 하지만, 지식 구축의 어려움 때문에 그다지 연구가 활발하지는 못한 실정이다. 그래서 본 연구에서는 사용 가능한 격틀 정보가 없거나 부족한 경우에, 단어 의미(word sense)가 태깅된 구문 트리(syntac-

※ 이 논문은 2002년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2002-003-D00330).
† 정 회 원 : 대구대학교 정보통신공학부 교수
†† 준 회 원 : SemanticQuest Inc. 연구원
논문접수 : 2002년 8월 20일, 심사완료 : 2003년 1월 29일

tic tree)를 입력으로 받아, 주어/목적어와 같은 구문관계를 행위주/대상과 같은 의미역으로 사상하여 의미 트리(semantic tree)를 생성하는 시스템을 구축하고자 한다(그림 1)[1].



(그림 1) 구문관계에 따른 의미역 결정의 예

구문관계에서 의미관계로 사상할 때 어떠한 경우에 트리가 변형되는지에 관해, 현재까지의 연구 결과로는 완전히 정리가 되지 않기 때문에, 본 논문에서는 구문 트리가 변형되지 않는다는 가정 하에서 연구를 진행하였다. 의미역 사상 후의 의미 트리를 표현하기 위해서는 개념 그래프(conceptual graph)[10, 15]를 사용하고 있는데, 이는 개념 노드(conceptual node)와 그 개념을 연결해 주는 개념 관계 노드(conceptual relation node)로 개념 그래프가 이루어진다는 점에서 본 연구의 결과와 매우 유사한 특성을 가지고 있기 때문이다.

구문관계에 따른 의미역 결정에 관한 본 연구는 개념지식 베이스인 온톨러지(ontology)의 구축시 개념간 관계(semantic relation)의 추출이나 기계 번역(machine translation), 질의 응답 시스템 등과 같은 응용분야에서 활용될 수 있다.

2. 기존 연구

언어학에서는 언어현상의 본질적인 규명을 위해 의미역의 분류에 관한 연구가 주로 이루어진 반면, 전산언어학에서는 언어의 전산처리를 쉽게 효율적으로 하기 위해 의미역의 분류 뿐 아니라 결정에 관한 연구도 이루어져 왔다.

2.1 언어학에서의 의미역

의미역을 결정하기 위해서는 어떤 구문관계들을 어떠한 의미역으로 사상할 것인지를 미리 정의해야 한다. 세종계획(전자사전 개발)[1]의 용언사전에서는 술어가 요구하는 통사적인 논항 뿐만 아니라 의미적인 논항에 대해서도 의미역을 정의하여 격틀정보를 구축했다. 대상, 행위주, 경험주, 동반주, 처소, 출발점, 도착점, 방향, 도구, 이유, 수령주, 자격, 기준치, 정도 등 총 14개의 의미역을 정의했다. 구별 가능한 의미역을 최대한 구분하여 기술하고 추후에 필요가 없다고 판단되면 구분했던 의미역을 다시 하나로 통합한다는 방침을 세우고 있다. 이렇듯 의미역의 구분을 고정시켜 놓지 않고 융통성 있게 하는 이유는 누구나 공감할 만한 의미역 분류를 하기가 매우 힘들기 때문이다.

박정운[5]은 부사어 '로'의 다의성에 관한 연구로 부사어 '로'가 갖는 의미들과 그 의미들의 연관성을 논하면서 부사어 '로'가 갖는 의미를 밝히고 있다. 의미역은 경로, 방향, 지향점, 시간의 경로, 상태변화, 자격, 도구나 수단, 재료, 원인, 양태 10개로 정의하고 있다.

조일영[12]에서는 'NP'로에 관한 의미역을 논하는데 15개의 의미역을 정의하고 있다. 동사에 의한 의미역과 명사에 의한 의미역으로 나누어 두 단계에 걸쳐 의미역을 결정한다. 동사에 의해서 방법, 결과, 처소, 원인 중 하나가 결정되며 동사에 의한 의미역의 하위 집합에 속하는 명사에 의해 의미역이 결정된다. 명사에 의해 결정되는 의미역은, 방법의 하위 집합인 도구, 수단, 재료와 처소의 하위 집합인 경로, 방향, 지향점 등이다. 그러나 동사에 의한 의미역은 결정 능력이 충분치 않다. "버스로 학교에 가다"의 '버스로'는 동사에 의해 '방법'이, 명사에 의해 '수단'이 할당되므로 그럴 듯 하다. 그러나, "산길로 학교에 가다"의 '산길로'는 '경로'의 의미역인데, 경로의 의미역을 가지려면 동사에 의해 처소의 의미역을 부여받아야 한다. 따라서, 동사 '가다'는 두 가지 추상적인 의미역을 갖게 되므로, 동사에 의해 추상적인 의미역을 결정한다는 논리에는 모순이 있다.

남기심[3]에서는 부사격 조사 '-에'와 '-로'의 쓰임에 관한 연구를 하였는데, 조사의 쓰임을 밝히기 위해서 조사와 결합하는 체언과 그것을 논항으로 취하는 술어와의 관계를 통해 의미를 제시하고 있다. 이를 통해 논항과 부가항이 구분되므로 논항이 가질 수 있는 의미와 부가항이 가질 수 있는 의미를 구분하여 기술하고 있다. 부사격 조사 '-에'의 논항은 장소, 대상, 기준점, 원인, 이유, 도구, 행위자, 수혜주로 8개의 의미를, 부가항은 장소, 시간, 부가, 원인, 도구, 인용, 대응, 기준으로 8개의 의미를 가지며, 부사격 조사 '-로'의 논항에 대해서는 지향점, 방향, 경로, 속성, 변성, 재료, 원인으로 7개의 의미를, 부가항에 대해서는 양태, 순서, 시간, 진술, 정도, 빈도, 원인, 수량으로 8개의 의미를 제안했다. 다른 기존 연구와 비교해 볼 때, 의미를 매우 자세하게 분류하였다.

이희자[11]는 말뭉치 분석을 통해서 국어 조사에 대한 특성을 밝힌 연구이다. 사전식으로 기술되어 있는 이 연구를 통해 조사에 따른 의미를 파악할 수 있다. 주격조사, 목적격조사, 보격조사와 인용격조사를 제외한 부사격조사에 대해 30개 가량의 조사를 언급하였으며, 조사가 가질 수 있는

1) 단어의 의미를 구분하기 위해 사용된 의미 코드는 가도카와 시소러스[19]의 의미 부류인데, 4장에서 자세히 설명된다.

의미역은 기존 연구[1, 3, 5, 12]를 기준으로 재해석하여 제시하였다.

2.2 전산 언어학에서의 의미역

의미역은 언어 전산 처리의 성능을 높이는 데에 큰 역할을 한다. 그러나 의미역의 분류 자체가 어려운 문제이므로 전산 언어학에서는 언어학에서의 의미역 분류를 그대로 이용하거나 전산처리에 용이하게 의미역을 분류하여 사용하는 추세이다.

조정미[13]에서는 한국어의 의미격을 30가지로 구분한 후, 23가지의 대표 조사를 그 의미격에 따라 분류하였으며, 명사와 동사의 의미부류와 조사만을 이용해 의미역을 결정하는 신경망 기반 방법을 제안했다. 실험 결과는 보이지 않았으나, 세 개의 자질(feature)[2]만으로는 의미역을 결정하기에 부족하다는 사실을 짐작할 수 있다.

양단희[7]에서는 격 원형성(case prototypicality)이라는 개념을 도입하였는데, 이는 모든 격에 대해 명사와 동사가 갖고 있는 의미의 정도를 말한다. 각 용언과 명사에 대해 격 원형성을 말뭉치로부터 미리 계산해 둔 후, 논항의 격을 이로부터 결정하는 방법을 제시하였다. 이 방법은 말뭉치로부터 기계 학습을 통해 지식을 구축했기 때문에 은유나 환유 현상을 다룰 수 있는 장점이 있으나, 대량의 학습데이터가 필요하며, 격조사가 표현할 수 있는 격 종류를 3가지로 제한한 점이 문제점으로 나타난다.

박성배[18]는 한영 기계번역에서 관계절의 의미역을 결정하기 위해서, 한영 기계번역을 위해 구축한 동사 패턴[2]에 수작업으로 의미 정보를 추가하여 규칙을 만들고, 수작업으로 구축한 규칙을 이용해 통계 정보를 추출하여 고빈도 의미역을 할당하는 규칙을 자동으로 생성하였다. 그러나 수작업에 따른 비용 증가와 어휘를 이용한 학습으로 인한 규칙의 적용률 저하가 문제점으로 지적되었다. 박성배[4]에서는 한영 기계번역에서 결정트리(decision tree)를 사용하여 부사격 조사의 의미 중의성을 해소하기 위한 연구를 하였는데, 사용한 자질은 200개의 클래스로 클러스터링된 명사와 동사의 의미부류, 보조사 유무, 명사와 동사가 떨어진 거리(D), 전체 문장에 대한 거리(D)의 상대 거리인데, 실험 결과를 통해 거리와 상대거리는 의미역 결정 능력이 약하다고 밝히고 있다. 이는 한국어가 어순이 비교적 자유롭기 때문으로 풀이해 볼 수 있겠다. 이전 연구에 비해 단어의 클래스를 이용해 학습 데이터 부족 문제를 완화시키기는 하였지만, 단어의 클래스를 200개로 고정한 점과 미경험 단어의 출현시 적용이 불가능하다는 단점을 여전히 가지고 있다.

Gildea[14]는 자질의 적절한 조합을 이용한 확률 모델을 제안하였는데, 자료 부족(data sparseness) 문제를 해결하기 위해서 선형 보간 방법(linear interpolation method)와 back-off 방법을 함께 사용했다. 선형 보간법은 구체적인(specific) 자질을 통한 확률과 일반적인(general) 자질을 통한 확률 모두를 항상 고려해서 원하는 값을 얻는 것인 반면, 선형 보간법에 backoff를 결합한 방법은 구체적인 자질을 통한 확률이 있을 경우에는 그 확률로 원하는 값을 얻지만, 자료 부족 문제로 인해 구체적인 자질을 이용한 확률이 없을 경우에는 좀 더 일반적인 자질의 확률 값을 보간(interpolation)하여 원하는 값을 추정하는 효과적인 방법이다. 이 연구에서는 일종의 의미부류인 프레임(frame)과 의미역이 태깅된 말뭉치를 포함하고 있는 FrameNet이라는 지식베이스를 사용하고 있다. 특정 프레임은 그 의미에 속한 단어들과 그 단어들이 가질 수 있는 의미역에 대한 정보를 갖고 있다. 프레임은 단어가 가질 수 있는 의미역 만을 나열한 것으로 논항 정보와 선택 제약 정보가 없다는 점에서 단순화된 격틀이라 할 수 있다. 이 연구는 괜찮은 성능을 보이고 있기는 하지만, 한국어에 대해서는 FrameNet과 같이 활용될 만한 지식베이스가 아직 여의치 않기 때문에 본 접근 방법을 그대로 한국어에 적용해 보기는 어렵다.

3. 구문관계와 의미역의 분류

구문관계에 대응하는 의미역을 결정하기 위해서는 어떤 구문관계를 어떤 의미역으로 사상(mapping)할 것인지를 먼저 정의해야 한다. 의미역은 논항들이 문장 내에서 수행하고 있는 역할[8]을 의미하므로 필수 논항인 구성요소에만 할당하는 것이 원칙이다. 하지만 남기심[3]에서 밝히고 있듯이 논항과 부가항의 구분이 어렵고, 또 궁극적인 의미분석을 위해서는 부가항에 대해서도 의미역을 결정해야 하므로 본 연구에서는 논항 뿐만 아니라 부가항에 대해서도 의미역을 결정하는 것을 목표로 한다.

본 연구에서 의미역 결정을 위해 대상으로 삼는 구문관계는 주어, 목적어, 보어, 부사어이다. 주어는 체언과 그에 상당하는 주격조사와 결합한 문장성분을 이르며, 목적어는 서술어의 동작 대상이 되는 문장 성분을 이른다. 보어는 현행 학교 문법에 따라 '되다/아니다' 앞에 오는 성분만을 인정한다. 주어와 목적어가 아닌 논항을 보어(complement)로 정의하기도 하지만[9], 논항과 부가항을 구별하지 않는 본 논문에서는 그런 정의는 무의미하다. 부사어는 용언, 관형사, 부사, 동사구, 관형사구, 부사구와 절이나 문장 전체를 수식하는 부사뿐만 아니라 그와 같은 수식 기능을 보이는

2) 자질은 의미역 결정에 이용되는 정보를 이른다.

여러 형태의 어구들을 망라하여 이르지만[6], 본 논문에서는 '체언+부사격조사'의 형태를 가지는 부사어만을 고려한다. 왜냐하면 일반 부사어에 대해 의미역을 결정하기는 어렵기 때문이다.

그리고, 구문관계가 사상될 의미역으로는 세종전자사전[1][3]에서 기술된 14개의 의미역에 재료, 경로, 시간을 더해 17개를 정의하였다. 세종사전은 논항만을 대상으로 하여 격틀정보를 구축했기 때문에 도구를 재료와, 도착점을 경로와 구분하기 어려울 수도 있다. 이는 의미역 정의가 논항과 부가항의 구별과 밀접히 관련되어 있기 때문이다. 하지만 본 논문에서는 논항 뿐만 아니라 부가항도 의미역을 결정하는 대상으로 고려하고 있으므로 도구와 재료, 도착점과 경로와의 구분이 논항만을 고려할 때보다 명확하다. 시간은 부가항에만 나타나는 의미역이므로 새로이 추가되었다. 또 다른 이유로는 본 연구결과를 중간언어 방식의 기계번역에 적용하고자 할 때, 재료, 경로, 시간의 의미역 정도는 구분하여야 대상 언어의 생성이 용이할 것으로 판단하여 추가하게 되었다.

대부분의 의미역 정의는 세종 전자사전을 따르지만, 필요에 따라 일부 내용을 수정하였다. 구체적인 정의는 아래와 같다.

3.1 행위주(Agent)

동사의 논항 가운데 행위를 야기시키거나 행위의 주체(subject)가 되는 논항에 주어지는 의미역이다. 행위주는 문장의 주어 자리에 나타나지만 그 역은 성립하지 않는다.

예) ① **철수가** 도둑을 잡았다.
② **정부에서** 실시한 조사 결과가 발표되었다.
③ **영희에게** 선물을 받았다.

3.2 대상(Theme)

문장에서 동작(action)이나 과정(process)의 영향을 입는 요소에 할당되는 의미역이다. 많은 경우 목적어 자리에 위치하는 논항이 대상의 의미역을 할당받는다.

예) ① 철수가 **책을** 옮겼다.
② **창문이** 흔들거린다.
③ 영희는 **선생님이** 아니다.
④ **논문에** 대해서 이야기 하자.

3.3 경험주(Experiencer)

어떤 사건에 대한 느낌이나 감정을 느끼는 심리적 주체나 사태를 경험하는 자를 가리키는 논항에 주어지는 의미역이다. 주로 심리형용사(좋다, 싫다, 밉다 부류)나 지각동사(느끼다 부류)의 유정물 논항이 경험주로 해석된다.

예) ① **철수는** 축구를 좋아한다.
② **영희에게** 노래를 들려주었다.

3.4 동반주(Companion)

행위주 이외에 그 행위주와 동등한 지위에 서는 다른 구성요소가 있을 경우 이 구성요소에 할당되는 의미역이다. 주로 문법표지 '-와/과'와 함께 주어 자리가 아닌 위치에서 실현된다.

예) 철수는 **영수와** 싸웠다.

3.5 장소(Location)

장소와 관련된 의미역이다. 사건(event)이나 사태(state-of-affair)가 일어나는 공간적 배경을 가리키는 구성요소(constituent)에 처소의 의미역이 배당된다.

예) ① **철수는 밭에** 씨를 뿌렸다.
② 영희는 **도서관에서** 공부했다.

3.6 출발점(Source)

동작의 시작이 이루어지는 시점이나 지점, 어떤 행위의 유래점을 가리키는 의미역이다. '-부터'가 첨가될 수 있는 경우 출발점으로 처리한다.

예) ① **서울에서** 대구까지 4시간이 걸린다.
② 고마운 **마음에서** 드리는 말씀입니다.

3.7 도착점(Goal)

객체(object)가 미치는 도달 지점을 나타내는 구성요소에 배당되는 의미역으로 출발점에 대조되는 개념이다.

예) ① 서울에서 **대구까지** 4시간이 걸린다.
② 영희는 **선생님이** 되었다.
③ 철수는 **학교에** 갔다.
④ 인삼을 **당의정으로** 가공한다.
⑤ 아버지는 **딸에게** 외출을 금지하셨다.

3.8 도구(Instrument)

동사가 나타내고 있는 사건, 상태를 변화시키거나, 행위를 작동시키는 데 도구로써 관여되는 구성요소가 갖는 의미역을 가리킨다.

예) 영희는 **칼로** 떡을 썰었다.

3.9 이유(Reason)

사건의 이유나 원인을 나타내는 구성요소에 주어진다. 도

3) 세종전자사전은 다양한 현대 한국어 정보처리를 지원할 수 있는 범용적/대규모 기반 전자사전의 구축과 언어학적 타당성, 전산적 효율성을 조화시킨 전산 어휘자료체의 형태를 띠는 것을 목표로 수십 명의 언어학자들에 의해 구축되고 있는 전자사전이므로, 본 연구에서 활용하기에 가장 적합한 언어 자원이라 판단된다.

구와 다소 구별되면서 이유나 원인의 의미가 두드러지게 나타나는 구성요소의 의미역이다. 주로 격표지 '-으로, -에'로 실현되며 도구의 의미역과 달리 주로 자동사에서 많이 나타난다.

예) ① 철수는 **감기로** 고생했다.
② 철수는 영희의 **합격에** 무척 기뻐했다.

3.10 수령주(Recipient)

소유의 이동이 일어나는 경우 소유를 넘겨받는 참여자에 수령주의 의미역을 부여하기로 한다.

예) ① 철수는 **영희에게** 꽃을 주었다.
② **철수는** 영희에게서 선물을 받았다.

3.11 자격(Appraisee)

평가 동사류, 즉 '~을 ~으로 V'에서 '보다, 판단하다, 생각하다, 간주하다, 평가하다, 여기다, 삼다' 등의 V로 나타나는 '-으로' 논항을 도착점이 아닌 자격으로 설정한다.

예) 철수는 영희를 **반장으로** 밀었다.

3.12 기준치(Criterion)

술어가 기술하는 대상의 특정 속성에 대한 도량적 평가의 기준이 되는 정도를 나타내는 구성요소를 출발점이 아닌 기준치로 설정한다.

예) ① 사람들은 철수를 **영희와** 비교한다.
② **이에서** 더 기쁘지는 않을 것이다.
③ 이것은 **분에** 넘치는 것이다.

3.13 정도(Degree)

구체적인 수량, 가격 따위의 차이를 보여 주는 구성요소이다. 전형적으로 조사 '-만큼'에 의해 표시될 수 있다.

예) 철수는 **아버지만큼** 크다.

3.14 방향(Direction)

행동이 진행되는 방향을 나타내는 의미역이다. 반드시 도달 지점을 전제하지 않는다는 점에서 도착점과 대조되며, 따라서 별도의 의미역으로 설정될 수 있다. 이동이 나타나지만 도착 지점이 구체적으로 나타나지 않고 방향만 나타나는 경우이다.

예) 철수는 **앞으로** 달려갔다.

3.15 시간(Time)

시간이나 횟수 등의 단위를 나타내는 명사에 준하는 구성요소에 할당되는 의미역이다. 시간의 의미역을 갖는 구성요소는 대부분 부가항에 해당된다.

예) ① 영희는 **아침에** 일찍 일어났다.
② 약속을 **1시로** 정하자.

3.16 경로(Path)

경로의 의미역은 도착점(goal)이나 방향(direction)처럼 이동의 개념을 갖고 있긴 하지만, 그들과 달리 단순히 지나가는 경유지인 경우에 할당된다.

예) 학생들은 **정문으로** 다닌다.

3.17 재료(Material)

사건이나 상태를 변화시킨다거나, 행위를 작동시키기 위해 이용되는 구성요소에 할당되는 의미역인 도구와 달리, 재료의 의미역은 결과물의 요소를 이룰 경우 할당된다.

예) 철수는 **나무로** 책상을 만들었다.

지금까지 본 연구에서 고려할 구문관계와 의미역에 대해 살펴보았는데, 이를 언어학 논저에서 제시하고 있는 일반적 원칙과 말뭉치 분석 결과, 그리고 세종전자사전의 정보 등을 종합하여 구문관계에 따른 의미역을 정리해 본 것이 <표 1>이다.

〈표 1〉 구문관계에 따른 의미역

구 분 관 계		의 미 역
주 어		행위주, 대상, 경험주, 수령주
목적어		대상
보 어		대상, 도착점
부사어	에	장소, 도착점, 기준치, 대상, 이유, 도구, 시간
	로	도구, 재료, 경로, 방향, 도착점, 자격, 이유, 시간
	에서	장소, 출발점, 행위주, 기준치
	에게	경험주, 행위주, 수령주, 도착점
	기타	기준치, 정도, 동반주, 자격, 도구, 출발점

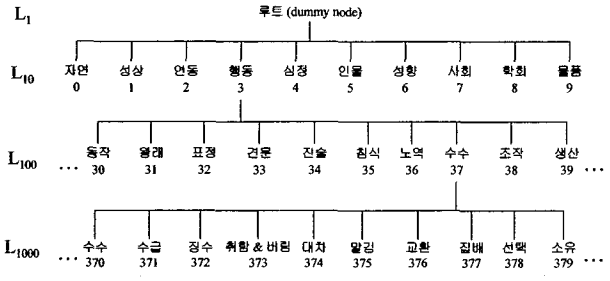
4. 의미역 결정 규칙

4.1 규칙 기술에 사용되는 자질

의미역을 결정하기 위한 규칙의 기술을 위해서는 구문관계, 의미부류, 형태소 정보와 같은 자질을 사용한다. 지배소와 의존소간의 구문관계에 따라 가능한 의미역의 후보가 달라지므로 구문관계는 모든 규칙에서 사용될 수 있는 중요한 자질이며, 목적어의 유무와 같은 정보도 의미역 결정에서 유용하게 사용된다. 사동사의 주어는 행위주의 의미역을 가지고, 피동사의 주어는 대상의 의미역을 갖기 때문에 동사가 사동사인지에 대한 정보는 의미역을 결정하는데 상당한 기여를 한다. 사동(causativization)은 주어 자리의 동

작자가 다른 동작자로 하여금 어떤 동작을 일으키게 만드는 것을 의미하기 때문에, 사동문에는 항상 목적어가 나타나며 목적어 유무 정보를 이용해서 사동사 주어의 의미역을 결정할 수 있다.

또 본 시스템은 단어의 의미 중의성이 해소된 결과를 입력으로 받기 때문에, 지배소의 의미부류와 의존소의 의미부류를 얻을 수 있는데, 가도카와 시소러스[19]를 그 의미부류로 사용하고 있다. 가도카와 시소러스는 총 1,110개의 개념과 4단계의 계층구조를 가지고 있으며, L_1, L_{10}, L_{100} 레벨에 속해 있는 개념들은 각각 10개의 하위 개념들로 나뉜다(그림 2). 명사와 동사의 분류는 하나의 계층구조에 공존하며, 동사의 의미 부류는 주로 L_{1000} 레벨의 의미 코드 2xx, 3xx, 4xx에서 나타난다.



(그림 2) 가도카와 시스러스의 개념 계층 구조

일반적으로 한국어는 어순이 자유롭기 때문에 위치 정보가 중요하지 않다고 알려져 있다. 이는 박성배[4]에서 지배소와 의존소 간의 거리(D), 문장 전체 길이에 대한 D의 상대거리와 같은 자질이 의미역 결정에 유용하지 않다고 증명된 사실과도 그 맥락을 같이 한다. 그러나 아래 예와 같이 두 주어가 모두 격조사를 가지고 있다거나 모두 보조사를 가지고 있을 경우에는 위치 정보로 의미역을 결정할 수 있다.

① 철수는(경험주) 영희는(대상) 싫다.
② 철수**가**(경험주) 영희**가**(대상) 싫다.

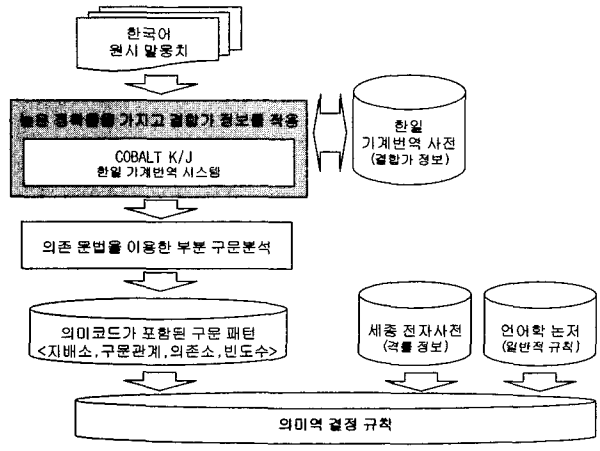
또 지배소의 어휘 또는 품사 정보, 의존소의 명사형 전성어미 포함 유무와 같은 형태소 정보도 의미역 결정에 사용될 수 있다. 지배소의 어휘를 고려하는 경우는 '느끼다'와 같은 동사가 목적어를 가지면서 주어의 의미역으로 행위주가 아닌 경험주를 취하는 경우로만 제한한다.

③ 나는(경험주) 슬픔을 **느끼다**.
④ 영희가(경험주) **예쁘다**(형용사).
⑤ 키가(대상) **크다**(형용사).
⑥ 얼마나 잤**기**에(이유) 눈이 부었니?

지배소의 품사는 주어가 가질 수 있는 의미역 후보를 줄여주는 역할을 한다. ④의 '예쁘다', ⑤의 '크다'와 같은 형용사들은 주어가 의미역으로 경험주와 대상만을 가지게 제한한다. 그리고 의존소에 명사형 전성어미를 포함하고 있으면 ⑥에서처럼 부사어 '에'는 이유의 의미역을 가진다.

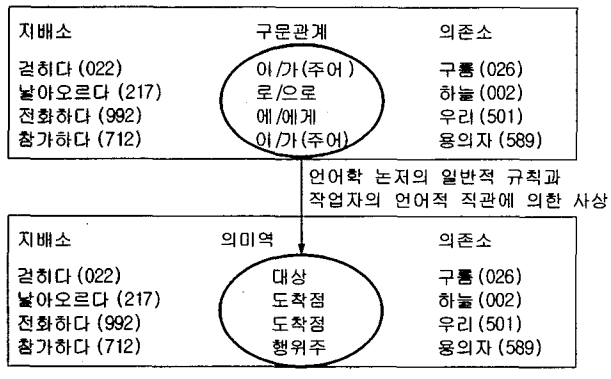
4.2 규칙의 구축

지금까지 살펴본 자질들이 이용되어 규칙의 조건부를 형성하게 되는데, (그림 3)에서 제시된 절차를 거쳐 규칙을 구축하게 된다.



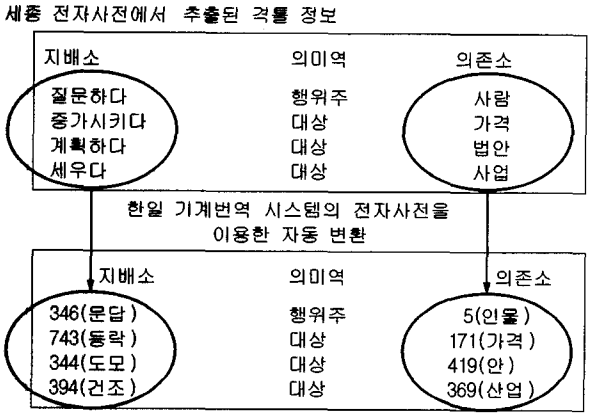
(그림 3) 규칙 구축 방법

의미코드가 포함된 구문패턴은 포항공대 지식 및 언어공학 연구실에서 개발한 한일 기계번역 시스템(COBALT-KJ)[16]을 사용해서 추출했다. 이 기계번역 시스템은 내부적으로로 단어 의미 중의성 해소를 위해 가도카와 시소러스의 의미코드로 표현된 결합가 정보를 사용하고 있는데, 단어 의미 중의성 해소가 끝난 단계에서 의미코드를 한국어 어휘와 동시에 출력하게 수정하였다. 그러면, 번역되어 출력된 문장의 각 어휘는 의미코드가 부가되어 있게 되는데, 이렇게 번역된 문장을 의존 문법을 이용하여 부분 구문 분석[17]을 하면 (그림 4)와 같은 구문 패턴을 얻을 수가 있게 된다. 7,000만



(그림 4) COBALT-KJ를 이용해서 추출한 구문패턴의 의미역 패턴으로의 사상

어절의 KIBS(Korean Information Base System, 1994~1997) 한국어 원시 말뭉치를 분석하여 총 208,088개의 의미 태깅된 구문 패턴을 생성하였다. (그림 4)는 생성된 구문 패턴을 언어학 논저의 일반적 규칙과 작업자의 언어적 직관에 의하여 의미역 패턴으로 변환하는 예를 보여 주고 있으며, (그림 5)는 세종 전자사전의 격틀 정보로부터 의미역 패턴 정보를 얻어내는 과정을 예시하고 있다. 세종 전자사전(용언사전)의 격틀 정보 (그림 6)에는 격틀의 예문도 함께 기술되어 있으므로 이를 추출하여 수정된 한일 기계번역 시스템에서 번역하면 역시 대상 어휘의 의미 코드를 알 수 있게 된다.



(그림 5) 세종 전자사전 격틀 정보의 의미역 패턴으로의 사상

```
<?xml version = "1.0" encoding = "euc-kr"?>
- <용언>
 - <논리구조>
  - <표제어정보그룹>
     <표제어> 태우 </표제어>
   </표제어정보그룹>
  - <형태정보그룹>
     <품사> 동사 </품사>
     <활용> 규칙형 </활용>
     <성분구성> 타,이,우 </성분구성>
   </형태정보그룹>
  - <구문정보그룹>
   - <구문정보그룹 type = "FTR">
    - <격틀정보>
       <격틀> N0 N1-을 V </격틀>
     - <구문의미>
         <선택제약> N0 = 인물 N1 = 음식(밥 | 생선 | 고기 | 국) | 사물(냄비 | 후라이팬) </선택제약>
         <의미역할> N0 = AGT N1 = THM </의미역할>
         <예제>영희는 생선을 새카맣게 태웠다.</예제>
         <예제> 불을 너무 가열해서 냄비를 태워버렸다.</예제>
       </구문의미>
     - <구문의미>
         <선택제약> N0 = 인물 N1 = (담배) </선택제약>
         <의미역할> N0 = AGT N1 = THM </의미역할>
         <예제> 그는 담배를 연신 태웠다. </예제>
       </구문의미>
     - <구문의미>
         <선택제약> N0 = 인물 N1 = (에 | 속 | 가슴) </선택제약>
         <의미역할> N0 = 인물 N1 = THM </의미역할>
         <예제> 셋째 아이가 내 속을 무척이나 태웠어요. <예제>
         <예제> 애를 태우며 그녀를 기다렸어요. </예제>
       </구문의미>
    </격틀정보>
```

(그림 6) 세종전자사전 중 용언사전의 격틀정보

이러한 과정에서 얻어진 구문 패턴과 의미역 패턴의 사상 관계를 분류해서 분석해 보면 특정 어휘 혹은 의미 부류 사이에 존재하는 의미역 결정 규칙을 보다 쉽게 발견할 수 있다. 이 분석 결과와 언어학 논저에 기술된 일반적인 규칙들을 종합적으로 정리하여 <표 2>와 같이 총 55개의 규칙을 구축하였다.

〈표 2〉 구축된 규칙의 수

구 문 관 계		규칙의 수
주 어		12
목적어		1
보 어		2
부사어	에	11
	로	10
	에서	4
	에게	4
	기타	11
합 계		55

의미역 결정 규칙의 예는 (그림 7)과 (그림 8)에 제시되어 있으며, 예시된 규칙에서 사용된 기호의 의미는 다음과 같다.

lex : 지배소의 어휘
lex_in : 지배소 어휘에 포함된 형태소
pos : 지배소의 품사
pos_in : lex_in의 품사
gov_con : 지배소의 개념코드(가도카와 의미코드)
dep_con : 의존소의 개념코드
role : 결정된 의미역

```
IF (lex == '하다' AND pos == 일반동사)
        THEN role = 행위주
ELSE IF (lex_in == '되' AND pos_in = 동사파행접사
        THEN role = 대상
ELSE IF ((lex_in == '시키' OR lex_in = '하') AND
          pos_in == 동사파생접사)
        THEN role = 행위주
ELSE IF ((lex == '듣다' AND gov_con ==332) AND
          (lex == '받다' OR lex == '물려받다') AND
          (gov_con == 295 OR gov_con ==370)))
        THEN role = 수령주
.....
위 규칙에서 사용된 개념코드의 의미
    332 : 청취, 295 : 영향, 370 : 수수
```

(그림 7) 주어의 의미역 결정 규칙 예

```
IF (gov_con == 210 OR gov_con ==217 OR gov_con ==230)
        THEN role = 출발점
ELSE IF ((710 <= dep_con <= 729) AND (pos == '일반동사'))
        THEN role = 행위주
.....
위 규칙에서 사용된 개념코드의 의미
   210 : 이동, 217 : 승강, 230 : 출입, 710~719 : 집단, 720~729 : 시설
```

(그림 8) 부사어 '에서'의 의미역 결정 규칙 예

규칙의 적용은 지배소의 어휘와 같은 구체적인 자질을 이용하는 규칙을 먼저 적용하게 되며, 의미부류와 같은 정보를 이용하는 규칙은 나중에 적용된다[14]. 만약 적용되는

규칙이 없는 경우에는 구문관계와 의미역간에 고빈도로 나타나는 의미역을 기본(default) 의미역으로 할당하게 된다.

5. 실험 및 평가

실험은 크게 두 가지로 나뉘어져서 이루어졌다. 하나는 형태소 분석, 구문 분석 및 단어 의미 중의성 해소 등 전단계에서 포함하고 있는 오류를 모두 수정하고 한 실험이고 다른 하나는 오류를 수정하지 않고 한 실험이다.

또 의미역 결정 문제의 기본 성능을 알아보기 위해 특정 구문관계에 대해 주로 나타나는 의미역을 기본적으로 할당하는 기본(baseline) 모델로도 실험을 따로 하였다. 주어, 목적어, 보어는 대상으로, 부사어 '에, 에서'는 장소로, 부사어 '에게, 로'는 도착점으로, 부사어 '와'는 동반주로 기본 의미역을 설정하였다.

실험말뭉치로는 한국전자통신연구원에서 주관한 한국어 형태소 분석기 및 품사태거 평가 워크숍(MATEC'99)에서 제공받은 말뭉치에서 임의로 추출한 340문장을 사용하였다. 먼저 형태소 분석, 구문 분석 및 단어 의미 중의성 해소 등 전단계의 오류를 모두 수정한 후의 적용 결과가 <표 3>에 제시되어 있다.

〈표 3〉 전단계 오류를 포함하지 않은 실험 결과(%)

	주 어	목적어	보 어	부 사 어					평 균
				에	로	에서	에게	기타	
기 본	55	90	59	42	36	73	64	96	64
규 칙	84	90	100	74	68	96	93	100	88

기본 모델에서 발생한 오류 중 규칙 모델에서 정확히 의미역을 결정한 예로는 "재옥은(행위주) 전에 없이 정현의 옷차림을 탓하였다.", "어머니의 예상대로 영아는 시험 전날 집에(도착점) 왔다.", "그제서야 어두운 상념에서(출발점) 벗어나 …" 등이 있었다.

본 연구를 통해 구축된 규칙 모델은 기본 모델에 비해 37%의 성능 향상을 보이고 있다. 그런데 실험 결과에서 목적어가 가질 수 있는 의미역이 대상, 하나임에도 불구하고 정확률이 90%인 것은 의미역을 할당해서는 안 되는 구성요소가 있기 때문이다. 의미역이 할당되지 않는 구성요소는 '서술성 명사 + 하다'형의 용언에서 '서술성 명사'가 분리되어 나타나는 경우로 주로 주어나 목적어로 실현이 된다. '생각하다'에서 서술성 명사 '생각'이 '생각을 하다', '생각이 나다', '생각이 되다', '생각이 들다'와 같은 행태를 보이는데, 주어로 실현될 때 결합하는 동사는 '나다, 되다, 들다' 뿐만 아니라 서술성 명사의 특성에 따라 다양하며 이 경우 동사는 큰 의미를 갖지 않는다. 이런 예는 실험 말뭉치에서 주어의 경우는 1%에도 못 미치지만 목적어의 경우는 대략 10%를 차지하기 때문에 앞서 언급된 것처럼 목적어의 경우는 정확률에 상당한 영향을 끼치게 된다. 물론 구문 분석 단계에서 이와 같은 유형의 용언을 하나의 단위로 묶어서 처리한다면 이 부분은 문제가 되지 않을 수 있다. 또 다른 해결방안으로 구문 트리를 변형하는 전처리를 생각할 수 있다. '생각을 하다'가 '생각하다'와 사실상 같은 의미를 갖고 있기 때문에 의미론적으로는 같은 트리를 가져야 한다. 따라서 이런 경우에 트리의 변형이 필요하다는 사실을 알 수 있는데, 본 연구에서는 구문 트리에서 의미 트리로 사상시 트리 변형이 없다고 가정을 했기 때문에 추후 좀더 고려해 보아야 할 부분이라 하겠다.

부사어의 경우 [3, 5]과 같이 '-에'와 '-로'의 문제가'-에서'와 '-에게'보다 상대적으로 어렵다는 것을 실험 결과를 통해 알 수 있다.

전단계 오류의 수정 없이 본 시스템을 적용했을 때의 결과는 <표 4>와 같다.

〈표 4〉 전단계 오류를 포함한 실험 결과(%)

	주 어	목적어	보 어	부 사 어					평 균
				에	로	에서	에게	기타	
기 본	38	77	59	29	37	75	60	69	55
규 칙	56	77	100	68	61	78	87	73	68

이 실험의 오류 원인을 단계별로 분석한 결과는 <표 5>와 같다. POS는 품사 태거(tagger)를, Parser는 구문분석기[17]를, WSD는 단어 의미 중의성을 해소하는 시스템[20]을, Roles는 본 연구에서 구축된 시스템을 말하며, 각 단계별 오류는 전단계의 오류를 포함한 수치이다.

〈표 5〉 단계별 오류 분석(%)

	주 어	목적어	보 어	부 사 어					평 균
				에	로	에서	에게	기타	
POS	2	0	0	1	2	0	7	14	3
Parser	32	15	0	11	15	15	13	27	16
WSD	33	15	0	21	29	17	13	27	19
Roles	44	23	0	32	39	22	13	27	32

6. 결론 및 향후 계획

본 논문에서는 의미 분석의 한 부분인 의미역 결정을 위한 규칙을 대규모 말뭉치와 기계번역 시스템, 세종 전자사전을 활용하여 구축하는 방법을 제시하였다. 사실 격틀과 같은 언어 지식이 의미역 결정에서 대단히 중요한 역할을 하지만, 이것이 가용치 않은 경우가 대부분이므로 기존 언어 자원을 최대한 활용하여, 규칙의 구축이 보다 객관적이고 효율적으로 이루어지게 하였다. 일반적인 언어학 지식과 경험만 가지고 의미역 결정 규칙을 기술하는 것은 작업자의 주관에 따라 결과가 많이 달라질 수 있으며, 또 미처 생

각지 못한 부분들도 있을 수 있기 때문에 규칙의 적용률이 많이 떨어질 수 있다. 하지만 본 논문에서 제시하는 방법은 대량의 원시 말뭉치를 기계번역 시스템으로 분석하여 의미정보가 태깅된 구문패턴을 추출함으로써 실제 언어의 다양한 사용례를 반영하였으며, 또 다수의 언어학자들이 심도있게 구축하고 있는 세종전자사전(용언사전)의 격틀 정보도 함께 고려하였기 때문에 본 방법에 의해 구축된 규칙들은 보다 객관적이고 효율적이라 할 수 있다.

또 의미역을 보다 정확하게 결정하기 위해 사용될 수 있는 자질 정보(구문관계, 의미부류, 형태소 정보, 이중주어의 위치정보 등)를 가능한 모두 포함시켰다. 특히 의미부류가 규칙의 기술에 이용되었기 때문에 규칙의 적용률이 향상되는 효과를 가져올 수 있다. 규칙 모델의 본질적인 장점이 적용되기만 하면 정확한 결과를 얻을 수 있다는 것이지만, 모든 경우를 규칙으로 해결할 수 없다는 단점도 존재한다. 그래서 규칙으로 해결되지 않는, 즉 규칙이 적용되는 않는 부분에 대한 처리를 현재는 기본 의미역(빈도수로 결정)으로 결정하나, 향후에는 확률모델을 도입하여 성능을 개선하는 연구를 할 예정이다. 규칙 모델의 또다른 단점이라면 규칙의 수가 많아질수록 기존 규칙과의 충돌이 발생할 가능성이 높아진다는 점을 들 수 있는데, 본 연구에서는 어휘정보를 사용한 규칙을 먼저 적용하고 의미코드 정보를 사용한 규칙을 나중에 적용하는 등 규칙 간 적용순서를 정해서 이 문제를 해결하고 있다.

본 연구의 결과는 온톨러지(ontology)의 구축시 개념간 개념관계의 추출이나 기계 번역(machine translation), 질의응답 시스템 등과 같은 응용분야에서 활용될 수 있다.

기존 의미역 결정 연구와는 연구 범위와 대상, 방법들이 달라서 성능의 직접 비교에는 무리가 있지만, 겉으로 드러난 정확률 만을 비교한다면 본 연구에서 구축한 시스템이 70~82%에 이르는 기존 연구에 비해 88%로 다소 좋은 성능을 보이고 있다.

향후에는 본 시스템을 이용하여 의미역이 태깅된 말뭉치를 반자동으로 구축하는 방법과 의미역이 할당되지 않는 구성요소를 고려하기 위해 트리가 변형되는 부분을 고려할 예정이며, 정확률이 상대적으로 낮은 부사어 '-에'와 '-로'의 성능 향상을 위한 새로운 방법을 연구할 예정이다.

참 고 문 헌

[1] 21세기 세종계획 전자사전 개발 연구보고서, 문화관광부, 2000.

[2] 김나리, 김영택, "한국어 동사 패턴에 기반한 한국어 문장 분석과 한영 변환의 모호성 해결", 한국정보과학회논문지, 제23권 제7호, pp.766-775, 1996.

[3] 남기심, "국어 조사의 용법 '-에' 와 '-로'를 중심으로", 서광학술자료사, 1993.

[4] 박성배, 김영택, "한영 기계번역에서 결정 트리 학습에 의한 한국어 부사격 조사의 의미 중의성 해소", 한국정보과학회논문지, 제27권 제6호, pp.668-677, 2000.

[5] 박정운, "한국어 도구격 조사의 다의어 체계 언어", 제24권 제3호, pp.405-426, 1999.

[6] 서정수, "국어 문법", 뿌리 깊은 나무, 1994.

[7] 양단희, 송만석, "기계학습에 의한 단어의 격 원형성 자동 획득", 한국정보과학회논문지, 제25권 제7호, pp.1116-1127, 1998.

[8] 이익환, "의미론 개론", 한신문화사, 1995.

[9] 이홍식, "국어문장의 주성분 연구", 서울대학교 박사학위논문, 1996.

[10] 이휘봉, "구문의존구조에서 중간언어 방식 기계번역을 위한 개념그래프의 생성", 포항공과대학교 전자계산학과 박사학위논문, 1998.

[11] 이희자, 이종희, "사전식 텍스트분석적 국어 조사의 연구", 한국문화사, 1998.

[12] 조일영, "'NP로'의 의미역", 제16차 한국어학회 전국 학술 대회, pp.56-65, 1998.

[13] 조정미, 김길창, "한국어 의미 해석시 중의성 해소에 대한 연구", 정보과학회지, 제14권 제7호, pp.71-83, 1996.

[14] D. Gildea and D. Jurafsky, "Automatic Labeling of Semantic Roles," In Proceedings of the 38th Annual Meeting of Association of Computational Linguistics, Hong Kong, pp. 512-520, 2000.

[15] J. F. Sowa, "Using a Lexicon of Canonical Graphs in a Semantic Interpreter," in Relational Models of the Lexicon : Representing knowledge in Semantic Networks, Edited by M. W. Evens, Cambridge University Press, pp.113-138, 1988.

[16] K. H. Moon and J. H. Lee, "Representation and Recognition Method for Multi-Word Translation Units in Korean-to-Japanese MT System," In the 18th International Conference on Computational Linguistics (COLING 2000), Germany, pp.544-550, 2000.

[17] M. Y. Kim, S. J. Kang and J. H. Lee, "Resolving Ambiguity in Inter-chunk Dependency Parsing," NLPRS 2001 (6th Natural Language Processing Pacific Rim Symposium), Tokyo, Japan, pp.263-270, Nov., 2001.

[18] S. B. Park and Y. T. Kim, "Semantic Role Determination in Korean Relative Clauses using Idiomatic Patterns," In Proceedings of the 17th International Conference on Computer Processing of Oriental Languages, pp.1-6, 1997.

[19] S. Ohno and M. Hamanishi, "New Synonyms Dictionary," Kadokawa Shoten, Tokyo, 1981.(Written in Japanese).

[20] Y. J. Chung, S. J. Kang, K. H. Moon and J. H. Lee, "Word Sense Disambiguation Using Neural Networks with Concept Co-occurrence Information," NLPRS 2001 (6th Natural Language Processing Pacific Rim Symposium), Tokyo, Japan, pp.715-722, Nov., 2001.

강 신 재

e-mail : sjkang@daegu.ac.kr
1995년 경북대학교 컴퓨터공학과(학사)
1997년 포항공과대학교 컴퓨터공학과 (공학석사)
2002년 포항공과대학교 컴퓨터공학과 (공학박사)
1997년~1998년 SK Telecom 정보기술연구원 주임연구원
2002년~현재 대구대학교 정보통신공학부 전임강사
관심분야 : 기계번역, 정보검색, 자동요약, 기계학습 등

박 정 혜

e-mail : jhpark@semanticquest.com
2000년 충남대학교 언어학과(학사)
2002년 포항공과대학교 정보통신대학원 (공학석사)
2002년~현재 SemanticQuest Inc. 연구원
관심분야 : 자연어처리, 한국어 분석, 기계번역, 정보검색 등