# 한국어 전산처리에서 규칙과 확률을 이용한 구문관계에 따른 의미역 결정†

# (Determination of Thematic Roles according to Syntactic Relations Using Rules and Statistical Models in Korean Language Processing)

강신재*, 박정혜**
(Sin-Jae Kang, Jung-Hye Park)

***요 약*** 본 논문은 한국어정보처리 과정에서 규칙과 확률을 이용하여 구문 관계를 의미역으로 사상시키는 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 혼합 방법은 대량의 원시 말뭉치를 분석하여 실제 언어의 다양한 사용례를 반영하며, 또 수십 명의 한국어학자들이 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 적용률이 향상되는 효과를 가져올 수 있었다.

***Abstract*** This paper presents an efficient determination method of thematic roles from syntactic relations using rules and statistical model 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 hybrid 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 our system.

## 1. 서 론

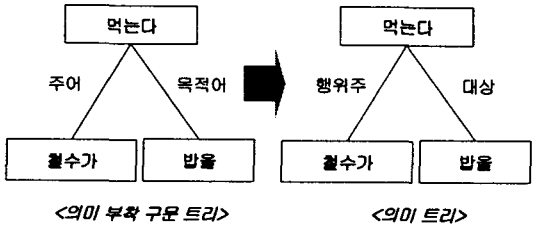
최근 널리 보급된 인터넷과 통신망에서 언어적 장벽을 극복하고, 또 대량의 정보 중에서 필요한 정보를 정확하고 빠르게 습득하기 위해서는 자연언어처리 기반 기술의 확보가 필수적이라 할 수 있다.

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 결

† 이 논문은 2002학년도 대구대학교 학술연구비 지원에 의한 논문임.
* 대구대학교 정보통신공학부 교수
** KOREA WISEnut Inc. 연구원

정이 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(syntactic tree)를 입력으로 받아, 주어/목적어와 같은 구문관계를 행위주/대상과 같은 의미역으로 사상하여 의미 트리(semantic tree)를 생성하는 시스템을 구축하고자 한다<그림 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)[1]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 개념지식베이스인 온톨러지(ontology)의 구축 시 개념간 관계(semantic relation)의 추출이나 기계 번역(machine translation), 질의 응답 시스템 등과 같은 응용분야에서 활용될 수 있다.

## 2. 관련 연구

조정미[2]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)[1]만으로는 의미역을 결정하기에 부족하다는 사실을 짐작할 수 있다.

양단희[3]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[4]는 한영 기계번역에서 관계절의 의미역을 결정하기 위해서, 한영 기계번역을 위해 구축한 동사 패턴[5]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[6]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 출현 시 적용이 불가능하다는 단점을 여전히 가지고 있다.

Gildea[7]는 자질의 적절한 조합을 이용한 확률 모델을 제안하였는데, 자료 부족(data sparseness) 문제를 해결하기 위해서 선형 보간 방법(linear interpolation method)와 backoff 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

1) 자질은 의미역 결정에 이용되는 정보를 이른다.

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.

지금까지 의미역 결정을 위한 방법에 관한 연구를 살펴보았다. 그러나 의미역을 결정하기 위해서는 어떤 구문관계들을 어떠한 의미역으로 사상할 것인지를 미리 정의해야 한다. 세종계획(전자사전 개발)[8]의 용언 사전에서는 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 고정시켜놓지 않고 융통성 있게 하는 이유는 누구나 공감할 만한 의미역 분류를 하기가 매우 힘들기 때문이다.

## 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[9]을 의미하므로 필수 논항인 구성요소에만 할당하는 것이 원칙이다. 하지만 남기심[10]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[11], 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 여러 형태의 어구들을 망라하여 이르지만[12], 본 논문에서는 '체언 + 부사격조사'의 형태를 가지는 부사어만을 고려한다. 왜냐하면 일반 부사어에 대해 의미역을 결정하기는 어렵기 때문이다.

그리고, 구문관계가 사상될 의미역으로는 세종전자사전[8][2]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. 대부분의 의미역 정의는 세종전자사전을 따르지만, 필요에 따라 일부 내용을 수정하였다. 구체적인 정의는 아래와 같다.

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.

② 대상 (Theme)

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

---

2) 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.

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.

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.

⑤ 장소 (Location)

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

⑥ 출발점(Source)

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

⑦ 도착점 (Goal)

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

⑧ 도구(Instrument)

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

⑨ 이유 (Reason)

사건의 이유나 원인을 나타내는 구성요소에 주어진다. 도구와 다소 구별되면서 이유나 원인의 의미가 두드러지게 나타나는 구성요소의 의미역이다. 주로 격표지 '-으로, -에'로 실현되며 도구의 의미역과 달리 주로 자동사에서 많이 나타난다.

⑩ 수령주(Recipient)

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

⑪ 자격(Appraisee)

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

⑫ 기준치(Criterion)

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

⑬ 정도 (Degree)

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

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.

⑮ 시간 (Time)

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

⑯ 경로 (Path)

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

⑰ 재료 (Material)

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

지금까지 본 연구에서 고려할 구문관계와 의미역에 대해 살펴보았는데, 이를 언어학 논저에서 제시하고 있는 일반적 원칙과 말뭉치 분석 결과, 그리고 세종전

자사전의 정보 등을 종합하여 구문관계에 따른 의미역을 정리해 본 것이 <표 1>이다.

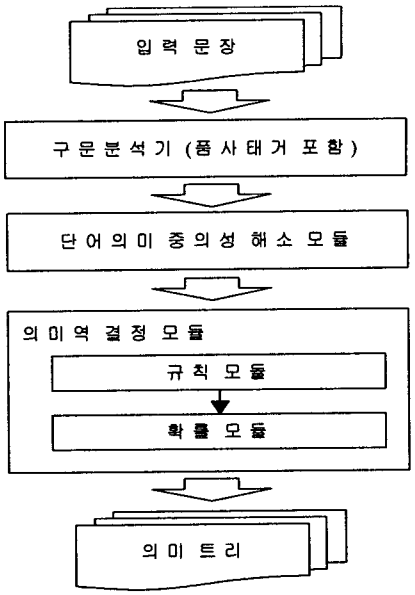
<표 1> 구문관계에 따른 의미역

| 구문관계 | | 의미역 |
|---|---|---|
| 주어 | | 행위주, 대상, 경험주, 수령주 |
| 목적어 | | 대상 |
| 보어 | | 대상, 도착점 |
| 부사어 | 에 | 장소, 도착점, 기준치, 대상, 이유, 도구, 시간 |
| | 로 | 도구, 재료, 경로, 방향, 도착점, 자격, 이유, 시간 |
| | 에서 | 장소, 출발점, 행위주 |
| | 에게 | 경험주, 행위주, 수령주, 도착점 |
| | 기타 | 기준치, 정도, 동반주, 자격, 도구, 출발점 |

## 4. 제안하는 의미역 결정 방법론

의미역이 태깅된 말뭉치와 같은 구체적인 사례를 통해 구축된 규칙은 정확하지만 적용범위가 좁다. 이러한 규칙의 단점을 극복하기 위해서 확률 기반 방법을 적용하게 되면 규칙에 적용되지 않는 부분을 처리할 수 있어서 시스템에 견고성(robustness)을 부여할 수 있게 된다. 하지만 확률 모듈은 의미역이 태깅된 말뭉치에 전적으로 의존하기 때문에 의미역이 태깅된 말뭉치를 반드시 구축해야 한다. 의미역이 태깅된 말뭉치를 좀 더 쉽게 구축하기 위해서 본 논문에서는 규칙 모듈을 활용하였다. 규칙 모듈의 제한된 적용범위를 확률 모델로 보완하여, 즉 규칙 모듈과 확률 모듈의 장점만을 취하여 혼합(hybrid) 기반 시스템을 구축하는 것이다.

<그림 2>는 규칙과 확률을 결합한 혼합 기반 시스템의 전체 구성도이다. 입력 문장은 품사 태거를 포함하고 있는 구문분석기와 단어의미 중의성 해소 모듈을 거쳐서 본 연구에서 구축한 의미역 결정 시스템에 적용된다. 의미역 결정 시스템에서는 규칙으로 정확하게 결정할 수 있는 부분을 먼저 처리하고, 나머지 부분을 확률 모듈에 적용하게 되며 이러한 과정을 통해 최종 결과인 의미 트리를 얻게 된다.



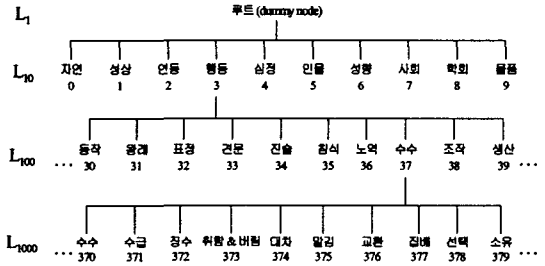

<그림 2> 시스템 전체 구성도

### 4.1 규칙 모듈

의미역을 결정하기 위한 규칙의 기술을 위해서는 구문관계, 의미부류, 형태소 정보와 같은 자질을 사용한다. 지배소와 의존소 간의 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 동작자가 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[13]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 3>. 명사와 동사의 분류는 하나의 계층구조에 공존하며, 동사의 의미 부류는 주로 $L_{1000}$ 레벨의 의미 코드 2xx, 3xx,

4xx에서 나타난다.




<그림 3> 가도카와 시소러스의 개념 계층 구조

일반적으로 한국어는 어순이 자유롭기 때문에 위치 정보가 중요하지 않다고 알려져 있다. 이는 박성배[6]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.

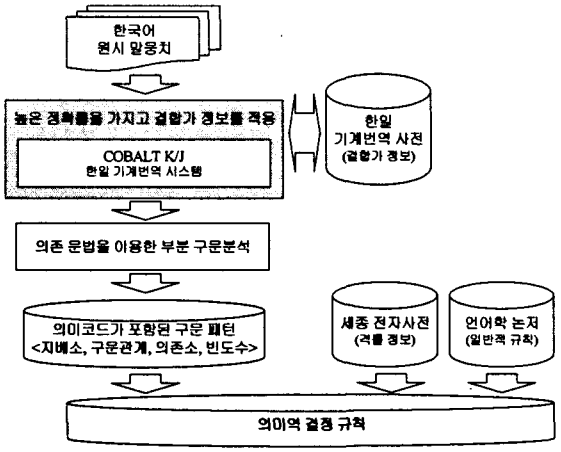
(1) 철수는(경험주) 영희는(대상) 싫다.
(2) 철수가(경험주) 영희가(대상) 싫다.

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 잤기에(이유) 눈이 부었니?

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

지금까지 살펴본 자질들이 이용되어 규칙의 조건부를 형성하게 되는데, <그림 4>에서 제시된 절차를 거쳐 규칙을 구축하게 된다.




<그림 4> 규칙 구축 방법

의미코드가 포함된 구문패턴은 포항공대 지식 및 언어공학 연구실에서 개발한 한일 기계번역 시스템(COBALT-KJ)[14]을 사용해서 추출했다. 이 기계번역 시스템은 내부적으로 단어 의미 중의성 해소를 위해 가도카와 시소러스의 의미 코드로 표현된 결합가 정보를 사용하고 있는데, 단어 의미 중의성 해소가 끝난 단계에서 구문정보와 의미코드를 동시에 출력하게 수정하였다. 7,000만 어절의 KIBS(Korean Information Base System, '94-'97) 한국어 원시 말뭉치를 분석하여 총 208,088개의 의미 태깅된 구문 패턴을 생성하였다<그림 5>.

| 지배소 | 구문관계 | 의존소 |
|---|---|---|
| 걸히다(022) | 이/가(주어) | 구름(026) |
| 날아오르다(217) | 로/으로 | 하늘(002) |
| 전화하다(992) | 에|에게 | 우리(501) |
| 참가하다(712) | 이/가(주어) | 용의자(589) |

<그림 5> COBALT-KJ를 이용해서 추출한 구문패턴의 예

이렇게 추출된 구문패턴을 분류해서 분석해 보면 특정 의미 부류와 구문 관계 간에 존재하는 규칙을 쉽게 발견할 수 있다. 이 분석 결과와 세종 전자사전 중 용언사전의 격틀정보, 그리고 언어학 논저에 기술된 일반적인 규칙들을 종합적으로 정리하여 <표 2>와 같이 총 55개의 규칙을 구축하였다.

<표 2> 구축된 규칙의 수

| 구문관계 | | 의미역 |
|---|---|---|
| 주어 | | 12 |
| 목적어 | | 1 |
| 보어 | | 2 |
| 부사어 | 에 | 11 |
| | 로 | 10 |
| | 에서 | 4 |
| | 에게 | 4 |
| | 기타 | 11 |
| 합계 | | 55 |

규칙의 적용은 지배소의 어휘와 같은 구체적인 자질을 이용하는 규칙을 먼저 적용하게 되며, 의미부류와 같은 정보를 이용하는 규칙은 나중에 적용된다[7].

### 4.2 확률 모듈

규칙 모듈에 적용되지 않은 구문관계의 의미역을 결정하기 위해서 확률 모듈을 적용한다. 규칙에서 중요한 역할을 했던 자질은 지배소의 의미부류, 의존소의 의미부류, 지배소의 어휘와 지배소의 품사이다. 지배소의 품사는 주어의 의미역 후보를 줄여주는 역할을 하기 때문에 주어의 확률 모델에만 적용한다. 자질의 조합을 이용한 확률 모델은 다음과 같다.

$$r = \arg\max_{r_i} p(r_i \mid dc, gc, gmor, gpos) \quad (1)$$

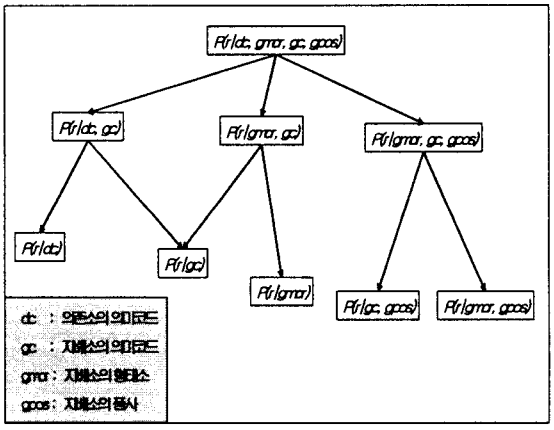
$$r = \arg\max_{r_i} p(r_i \mid dc, gc, gmor) \quad (2)$$

식(1)은 주어의 의미역을 결정하기 위한 것이고, 식(2)는 부사어의 의미역을 결정하기 위한 것이다. $r$은 의미역을 의미하고, $dc$는 의존소의 의미부류, $gc$는 지배소의 의미부류, $gmor$는 지배소의 형태소, $gpos$는 지배소의 품사를 말한다. 식(1), (2)에서 사용되는 확률은 식(3)과 같은 방법으로 구한다. 즉, 자질조합이 출현한 빈도로 의미역과 자질조합이 함께 나타난 빈도로 나눈다.

$$p(r_i \mid dc, gc, gmor, gpos) = \frac{freq(r_i, dc, gc, gmor, gpos)}{freq(dc, gc, gmor, gpos)} \quad (3)$$

확률 모델은 주어, 부사어 '-에, -로, -에서, -에게'에 한해 적용된다. 목적어는 하나의 의미역을 갖기 때문에 확률 모델에 적용하지 않았으며 보어와 '-에, -로, -에서, -에게'를 제외한 나머지 부사어들은 규칙만으로도 정확하게 의미역을 결정할 수 있기 때문에 확률 모델에 적용할 필요가 없다.

그리고, 확률 모델에 적용할 경우 나타나게 되는 자료 부족 문제(data sparseness problem)를 해결하기 위해서는 [7]에서 사용한 선형보간법(linear interpolation method)과 백오프 방법(backoff method)을 결합한 확률 모델을 사용한다. 선형보간법은 원하는 결과를 얻기 위해서 구체적인 자질을 이용한 확률뿐만 아니라 일반적인 자질을 이용한 확률도 사용한다. 즉, <그림 6>[3]에 있는 9개의 확률을 보간(interpolate)해서 원하는 결과를 얻는 것이다.




<그림 6> 주어의 의미역 결정을 위한 확률 격자(lattice)

반면 선형보간법과 백오프 방법을 혼합한 방법은 구체적인 자질을 이용한 확률을 구할 수 있으면 그것으로 원하는 결과를 얻고, 구체적인 자질을 이용한 확률을 구하지 못할 경우에만 좀 더 일반적인 자질들을 사용하는 확률을 이용해서 원하는 결과를 얻는다. 다시 말하자면, <그림 6>에서 자질을 제일 많이 포함하고 있는 최상위 노드의 확률을 구할 수 있으면 그것으로 의미역을 결정하고, 자료부족 문제로 인해서 그 확률을 구하지 못할 경우에는 화살표를 따라 내려가

3) 이 그림은 주어의 의미역 결정을 위한 확률식 간에 어떠한 경로(path)를 갖는지를 보여주고 있다. 부사어의 의미역 결정을 위한 격자(lattice)는 최상위 노드에서 gpos를 삭제하고, gpos자질을 포함한 세 가지 확률을 제거하면 된다.

서 좀 더 일반적인 자질을 이용한 확률을 사용해서 의미역을 결정하게 된다. 이러한 경우에는 확률 값이 3개가 되기 때문에 아래와 같은 식이 필요하게 된다.

$$p(r) = \lambda_1 p(r \mid dc, gc) + \lambda_2 p(r \mid gc, gmor) + \lambda_3 p(r \mid gmor, gpos) \quad (4)$$

[7]에서 식(4)의 λ가 성능에 별다른 영향을 미치지 못함을 증명하였으므로 본 연구에서는 λ는 서로 동일한 값을 가지며 그 합은 1이 되도록 하였다. 만약 식(4)의 세 확률 중 하나의 확률을 구하지 못한 경우에는 <그림 6>의 화살표를 따라 내려가서 더 일반적인 자질을 이용한 확률로 식(4)를 대치시킨다. 만약 *p(r|dc, gc)*의 확률을 구하지 못했으면 *p(r|dc, gc)*를 대신 *p(r|dc)*와 *p(r|gc)*를 보간한 값을 이용한다. 이 경우 λ의 수가 증가되는데, 이 경우 역시 그 합이 1이 되는 동일한 값을 가지도록 λ를 조정한다.

## 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) 모델, 기본 모델에 의미역 결정 규칙 모듈만 결합한 규칙(rule) 모델, 규칙 모델에 확률 모듈을 추가한 혼합(hybrid) 모델로도 실험을 따로 하였다. 먼저 전단계의 오류를 모두 수정한 후의 적용 결과가 <표 3>에 제시되어 있다.

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 갖지않는다. 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. 또다른 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 사상 시 트리 변형이 없다고 가정을 했기 때문에 추후 좀더 고려해 보아야 할 부분이라 하겠다.

<표 3> 전단계 오류를 포함하지 않은 실험 결과(%)

| 구문관계 | | 기본 | 규칙 | 혼합 |
|---|---|---|---|---|
| 주어 | | 55 | 84 | 96 |
| 목적어 | | 90 | 90 | 90 |
| 보어 | | 59 | 100 | 100 |
| 부사어 | 에 | 42 | 74 | 75 |
| | 로 | 36 | 68 | 68 |
| | 에서 | 73 | 96 | 96 |
| | 에게 | 64 | 93 | 93 |
| | 기타 | 96 | 100 | 100 |
| 평균 | | 64 | 88 | 90 |

부사어의 경우 [10, 15]와 같이 '-에'와 '-로'의 문제가'-에서'와 '-에게'보다 상대적으로 어렵다는 것을 실험 결과를 통해 알 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 90%로 다소 좋은 성능을 보이고 있다.

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

이 실험의 오류 원인을 단계별로 분석한 결과는 <표 5>와 같다. POS는 품사 태거(tagger)를, Parser는

구문분석기[16]를, WSD는 단어 의미 중의성을 해소하는 시스템[17]을, Roles는 본 연구에서 구축된 시스템을 말하며, 각 단계별 오류는 전단계의 오류를 포함한 수치이다.

<표 4> 전단계 오류를 포함한 실험 결과(%)

| 구문관계 | | 기본 | 규칙 | 혼합 |
|---|---|---|---|---|
| 주어 | | 38 | 56 | 62 |
| 목적어 | | 77 | 77 | 77 |
| 보어 | | 59 | 100 | 100 |
| 부사어 | 에 | 29 | 68 | 72 |
| | 로 | 37 | 61 | 66 |
| | 에서 | 75 | 78 | 78 |
| | 에게 | 60 | 87 | 87 |
| | 기타 | 69 | 73 | 73 |
| 평균 | | 55 | 68 | 77 |

<표 5> 단계별 오류 분석(%)

| 구문관계 | | POS | Parser | WSD | Roles |
|---|---|---|---|---|---|
| 주어 | | 2 | 32 | 33 | 38 |
| 목적어 | | 0 | 15 | 15 | 23 |
| 보어 | | 0 | 0 | 0 | 0 |
| 부사어 | 에 | 1 | 11 | 21 | 28 |
| | 로 | 2 | 15 | 29 | 34 |
| | 에서 | 0 | 15 | 17 | 22 |
| | 에게 | 7 | 13 | 13 | 13 |
| | 기타 | 14 | 27 | 27 | 27 |
| 평균 | | 3 | 16 | 19 | 23 |

## 6. 결론 및 향후 계획

본 논문에서는 의미 분석의 한 부분인 의미역 결정을 위해서 정확한 규칙 모듈과 견고한 확률 모듈의 장점을 취하여 두 시스템을 결합한 혼합 기반 시스템을 제안했다. 의미역이 태깅된 말뭉치와 같은 구체적인 사례를 통해 구축된 규칙은 정확하지만 적용범위가 제한되는 경우가 많다. 이러한 규칙의 단점을 극복하기 위해서 확률 기반 방법을 적용하게 되면 규칙에 적용되지 않는 부분을 처리할 수 있어서 시스템에 견고성(robustness)을 부여할 수 있게 되었다.

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 의미역 결정에 이용되었기 때문에 적용률이 향상되는 효과를 가져올 수 있었다.

본 연구의 결과는 온톨러지(ontology)의 구축 시 개념간 개념관계의 추출이나 기계 번역(machine translation), 질의 응답 시스템 등과 같은 응용분야에서 활용될 수 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] 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.

[2] 조정미, 김길창, "한국어 의미 해석시 중의성 해소에 대한 연구", *정보과학회지*, 제 14권, 제 7호, pp.71-83, 1996.

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

[4] 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.
[5] 김나리, 김영택, "한국어 동사 패턴에 기반한 한국어 문장 분석과 한영 변환의 모호성 해결", *한국정보과학회논문지*, 제23권, 제7호, pp.766-775, 1996.
[6] 박성배, 김영택, "한영 기계번역에서 결정 트리 학습에 의한 한국어 부사격 조사의 의미 중의성 해소", *한국정보과학회논문지*, 제27권, 제6호, pp.668-677, 2000.
[7] 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.
[8] *21세기 세종계획 전자사전 개발 연구보고서*, 문화관광부, 1999.
[9] 이익환, '*의미론 개론*', 한신문화사, 1995.
[10] 남기심, '*국어 조사의 용법 '-에'와 '-로'를 중심으로*', 서광학술자료사, 1993.
[11] 이홍식, '*국어문장의 주성분 연구*', 서울대학교 박사학위논문, 1996.
[12] 서정수, '*국어 문법*', 뿌리 깊은 나무, 1994.
[13] S. Ohno and M. Hamanishi, '*New Synonyms Dictionary*,' Kadokawa Shoten, Tokyo, 1981. (Written in Japanese)
[14] 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.
[15] 박정운, "한국어 도구격 조사의 다의어 체계", *언어*, 제24권, 제3호, pp.405-426, 1999.
[16] 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.
[17] 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** (Sin-Jae, Kang)

1995년 경북대학교 컴퓨터 공학과 졸업(학사)
1997년 포항공과대학교 컴퓨터 공학과 졸업(공학석사)
2002년 포항공과대학교 컴퓨터 공학과 졸업(공학박사)
1997년 - 1998년 SK Telecom 정보기술연구원 주임연구원
2002년 - 현재 대구대학교 정보통신공학부 전임강사
관심분야 : 기계번역,정보검색,자동요약,기계학습등

**박 정 혜** (Jung-Hye, Park)

2000년 충남대학교 언어학과 졸업(학사)
2002년 포항공과대학교 정보통신대학원 졸업(공학석사)
2002년-2003년 SemanticQuest Inc. 연구원
2003년 - 현재 KOREA WISEnut Inc. 연구원
관심분야 : 자연어처리, 한국어 분석, 기계번역, 정보검색 등