대규모 말뭉치와 전산 언어 사전을 이용한 의미역 결정 규칙의 구축

강 신재¹·박 정혜²

요 약

본 논문은 한국어정보처리 과정에서 구문 판계를 의미역으로 사상시키기 위한 규칙을 효과적으로 구축하는 방법을 제시하고 있다. 의미역의 결정은 의미 본체의 해석 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 많은 경우를 다룰 수 있는 규칙의 구축은 불가능하다. 하지만 본 논문에서 제시하는 방법은 대량의 문서를 본문으로서 실제로 언어의 다양한 사용례를 반영하며, 또 수십 명의 한국어 학자들이 심도 있게 구축하고 있는 서울전자사전의 적절 정보도 함께 고려하기 때문에 보다 객관적이고 효율적인 방법이라 할 수 있다. 의미역을 보다 정확하게 결정하기 위해 구문체계, 의미부류, 형태소 정보, 이중주의의 위치정보 등의 자세 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 Kang¹·Jung-Hye Park²

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 is 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), 의미본체(Semantic Analysis), 말뭉치 분석(Corpus Analysis), 서울전자사전(Sejong Electronic Lexicon of Korean), 기계번역(Machine Translation)

1. 서론

최근 널리 보급된 인터넷과 통신망에서 언어적 장벽을 극복하고, 또 다양한 정보 중에서 필요한 정보를 정확하고 빠르게 습득하기 위해서는 자연어처리 기반 기술의 확보가 필수적이라 할 수 있다. 일반적으로 언어를 분석할 때는 형태소 분석과 구문 분석의 과정을 거쳐 의미 본체를 하게 되는데, 의미 본체에서는 단어의 의미 중하성 해소(word sense disambiguation)와 단어간 의미역(thermatic role)의 결정이 중요한 작업이다. 이러한 여러 과정 가운데 본 논문에서는 의미 본체에서의 의미역 결정에 대해 논하고자 한다.

일반적으로 의미역의 결정은 적절(case frame)과 같은 언어적(linguistic knowledge)을 필요로 하지만, 지식 구축의 어려움 때문에 그저 연구가 활발하지는 못한 실정이다. 그래서 본 연구에서는 사용 가능한 적절 정보가 없거나 부족한 경우에는 단어 의미(word sense)가 대장된 구문 트리(syntax-
tic tree)를 입력으로 받아, 주어/목적어와 같은 구문관계를 행위주/대상과 같은 의미적으로 상상하여 의미 트리(semantic tree)를 생성하는 시스템을 구축하고자 한다(그림 1)[1].

(그림 1) 구문관계에 따른 의미 결정의 예

구문관계에서 의미관계로 상상할 때 어려운 경우에 트리가 변형되는지에 관해, 현재까지의 연구 결과로는 완전히 정리가 되지 않기 때문에, 본 논문에서는 구문 트리가 변형되지 않는다는 가정 하에 연구를 진행하였다. 의미 사상 후의 의미 트리를 표현하기 위해서는 개념 그래프(conceptual graph)[10,15]을 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 의미 결정에 관한 본 연구는 개념적인 베이스온 온톨로지(ontology)의 구축과 개념간 관계(semantic relation)의 추출이나 기계 번역(machine translation)의 응용 시스템 등과 같은 응용분야에서 활용될 수 있다.

2. 기존 연구

언어학에서는 언어현상의 본질적인 규명을 위해 의미의 본류에 관한 연구가 주로 이루어진 반면, 전산언어학에서는 언어의 전산처리를 쉽게 효율적으로 하기 위해 의미의 본류 뿐 아니라 결정에 관한 연구도 이루어지고 있었다.

2.1 언어학에서의 의미

의미를 결정하기 위해서는 어떤 구문관계들에 어떤 의미의 본류를 사상할 것인지와 다른 정의해 하였다. 세종계획(전산사전 개념)[1]의 유용사전에서는 슬어나 요구하는 동시에 것이 관념없이 아니라 의미에서의 본류에 대해서도 의미를 정의하여 각별한 구축하였다. 대상, 행위주, 경험주, 동반주, 참조, 출발점, 도착점, 발행, 도구, 이유, 수명주, 작격, 기준들, 정도 등 총 14개의 의미를 정의하였다. 구별 가

능한 의미의 최대의 구분하여 기술하고 추후에 필요가 없다고 판단되면 구본론적 의미론을 다시 하나로 통합한다 는 방침을 세우고 있다. 이와듯 의미의 구분을 고정시켜 놓지 않고 융통성 있게 하는 이유는 누구나 공감할 만한 의미 분류를 하기 매우 힘들기 때문이다.

박정운[5]은 부사어 '로'의 다의성에 관한 연구로 부사어 '로'가 갖는 의미들간의 연관성을 통해 부사어 '로'가 갖는 의미를 분류하고 있다. 의미론은, 경로, 방향, 지향점, 시간의 경로, 상대변화, 작격, 도구나 수단, 재료, 원인, 양태 10개로 정하고 있다.

조양기[12]에서는 'NP'로의 구분한 의미론을 논하는데 15개의 의미론을 정하고 있다. 동시에 의미의 분류와 명사에 의한 의미론으로 나누어 두 단계에 걸쳐 의미론을 결정한 다. 동시에 의미의, 방향, 결과, 원인 주어 하나가 결정되며 동시에 구분한 의미론의 하위 집합에 속하는 명사에 의해 의미론이 결정된다. 명사에 의해 결정되는 의미론은, 방향의 하위 집합은 도구, 수단, 재료와 체소의 하위 집합인 경로, 방향, 지향점 등이다. 그러나 동시에 의한 의미론은 결정 능력이 충분하지 않다. "바로 학교에서 가다"의 "바로"는 동시에 하위로, 명사에 의해 '수단'이 활동되는 것에 그만한다. 그러나, "산길로 학교에 가다"의 "산길로"는 '경로'의 의미론이지만, 경로의 의미론과 겹치게 동시에 의미론의 하위 집합을 부여받아야 한다. 따라서, 동시에 가다 는 두 가지 추상적인 의미론을 갖게 되므로, 동시에 의미론의 결정을 결정하는 논리에는 도움이 필요하다.

남기심[3]에서는 부사적 조사 '에'와 '로'의 측면에 관한 연구를 하였는데, 조사의 측면을 바탕으로 조사와 결정하는 체인과 그것을 논향으로 취하는 숫자와의 관계를 통해 의미론을 제시하고 있다. 이를 통해 논향과 부사학이 구 분되므로 논향이 가질 수 있는 의미와 부사학이 가질 수 있는 의미를 구분하여 기술하고 있다. 부사적 조사 '에'의 논향은 장소, 대상, 기준점, 원인, 이유, 도구, 행위자, 수혜 주로 8개의 의미를 부사름은 장소, 시간, 부가, 원인, 도구, 인용, 대용, 기준으로 8개의 의미를 가지며, 부사적 조사 '로'의 논향에 대해서는 대상, 학부, 경로, 속성, 변상, 채 표, 원인으로 7개의 의미를 부사름에 대해서는 양태, 순서, 시간, 진출, 정도, 빈도, 원인, 수혜로 8개의 의미를 제한한다. 다른 기존 연구와 비교해 볼 때, 의미론을 매우 자세하 게 구분하였다.


---

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

22 전산 언어학에서의 의미역

의미역은 언어 전산 처리의 성능을 높이는 데 큰 역할을 한다. 그러나 의미역의 본류 자체가 어려운 문제였으므로 전산 언어학에서의 의미역 본류론이 고요한 것으로 이어져 전산처리에 용이하게 의미역을 본류하여 사용하는 추세이다.


양단희[7]에서는 각 원형(case prototypicality)이라는 개념을 도입하였는데, 이는 모든 격에 대해 명사와 동사가 갖고 있는 의미의 정도를 말한다. 각 용언과 명사에 대해 각 원형성을 말로지로부터 미리 계고해 둔 후, 논의 격을 이로부터 결정하는 방법을 제시하였다. 이 방법은 말로지로부터 기계 학습을 통해 지식을 구축하기 때문에 은유적 혹은 현실성을 다루는 것이 있는데, 대량의 학습 데이터가 필요하며, 적조사가 표현할 수 있는 격 종류를 3가지로 제한한 점이 문제점으로 나타난다.


Gildea[14]는 자질의 적절한 조합을 이용한 확률 모델을 제안하였는데, 자료 부족(data sparseness) 문제를 해결하기 위해서 선형 보간 방법(linear interpolation method)과 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 보간법의 back-off를 결합한 방법은 구체적인 자질을 통한 확률이 있는 경우에는 그 확률로 원하는 값을 얻지만, 자료 부족으로 인해 구체적인 자질을 이용한 확률이 없을 경우에는 좀 더 일반적인 자질의 확률을 보간(interpolation)하여 원하는 값을 추정하는 효과적인 방법이다. 이 연구에서는 일부의 의미부류인 프레임(frame)과 의미가 대신된 말로지를 포함하고 있는 FrameNet이라는 지식베이스를 사용하고 있다. 특정 프레임은 그 의미에 속한 단어들의 그 단어들이 가질 수 있는 의미에 대한 정보들을 갖고 있다. 프레임은 단어가 가질 수 있는 의미만을 나열한 것으로 논의 정보와 선택 계약 정보가 있다는 점에서 단순화된 것이라 할 수 있다. 이 연구는 웹을 성능을 보이고 있기지만, 한국어에 대해서는 FrameNet과 같이 활용될 만한 지식베이스가 아직 여전히 없기 때문에 본 접근 방법을 그대로 한국어에 적용해 보기는 어렵다.

3. 구문관계와 의미역의 분류


본 연구에서 의미역 결정을 위해 대상으로 삼는 구문관계는 주어, 목적어, 보어, 부사격이다. 주어와 목적어의 경우 주어적 조사와 함께 문장 성분을 이루며, 목적어는 서술어의 동작 대상이 되는 문장 성분을 이룬다. 보어는 현행 문법에 따라 '더이아나라'로 알기 오는 성분을 인정한다. 주어와 목적어가 아닌 논향을 보이(complement)로 정의하기도 하지만[9], 논향과 부사격을 구별하지 않는 본 논문에서는 그 단정하는 논리한다. 부어와 부사격은 문장 전체를 수식하는 부사격이 아니라 그와 같은 수식 기능을 보이는
여러 형태의 이구들을 망라하여 이르지만(6), 본 논문에서는 ‘체험+부서가조사’의 형태를 가지는 부자이면을 고려한 다. 그러나 반면 부자이에 대해 의미의 결합하기는 어렵기 때문이다.

그리고, 구문구조가 사상의 의미의 경우에는 세종천사전 [1]에서 기술된 14개의 의미에 제료, 경로, 시간을 더해 17개를 정리하였다. 세종사전은 논항말을 대상으로 하여 각

d음을 고쳐 구축한 것에 대해 도구로 제료와, 도착점 경로와 구분하기 어려울 수도 있다. 이는 의미의 정의가 논항과 부

g항의 구별과 밀접히 관련되어 있기 때문이다. 하지만 본

논문에서는 논항뿐만 아니라 부기항도 의미의 결정하는

대상으로 고려하고 있으므로 도구 및 제료, 도착점 경로와

의 구분이 논항만을 고려할 때보다 명확하하다. 시간은 부가

항에만 나타나는 의미력으로 새로운 추가되었다. 또한

이론으로는 본 연구결과의 중간이어 방식의 기계평역에 적용

하고자 할 때, 제료, 경로, 시간의 의미적 정보는 구분하여

이어 상의 생성이 용이할 것으로 판단하여 추가하게

되었다.

대부분의 의미의 정의는 세종 천사전을 따르지만, 필요에 따라서 일부 내용을 수정하였다. 구체적인 정의는 아래와 같다.

3.1 행위주(Agent)

동사의 논항 가운데 행위를 야기시키거나 행위의 주체

(subject)가 되는 논항에 주하는 의미력이다. 행위주는 동

장의 주어 자리에 나타나지만 그 역은 상호적이다.

예) ① 철수가 도독을 잡았다.

② 정부에서 실시간 조사 결과가 발표되었다.

③ 영화에 신출을 받았다.

3.2 대상(Theme)

문장에서 동작(action)이나 과정(processing)의 영향을 입는

요소에 해당하는 의미역이다. 많은 경우 목적의 자리에 위

치는 논항이 대상을 의미의 함의를 함당한다.

예) ① 철수가 책을 읽었다.

② 꽃문이 훼손되었다.

③ 영화는 신생님이 아니다.

④ 논문에 대해서 이야기 하자.

3.3 경험주(Experimenter)

여러 사전에 대한 느낌이나 감정을 느끼는 심리적 주체

나 사태를 경험하는 자를 가리키는 논항에 주하는 의미

역이다. 주로 체험형용사와(좋다, 싫다, 답답..., 부기나

나 부기)와 자각동

사(느끼는 부기)의 유형을 논항이 경험주로 해석한다.

예) ① 철수는 축구를 좋아한다.

② 영화에 노래를 들려주었다.

3.4 동반주(Companion)

행위주 이외에 그 행위주와 동등한 지위에 서는 다른 구

성요소가 있을 경우 이 구성요소에 할당되는 의미역이다.

주로 문법표지 ‘-/와/와’와 함께 주어 자리가 아닌 위치에서

설현된다.

예) 철수는 영수와 서웠다.

3.5 장소(Location)

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

(constitutional)에 부속의 의미역이 해당된다.

예) ① 서울에서 대구까지 4시간이 걸린다.

② 고마운 마음에서 드리는 말씀입니다.

3.6 출발점(Source)

동작의 시작이 이루어지는 시점이나 지점, 어떤 행위의

유래를 가리키는 의미역이다. ‘-부터’가 참가할 수 있는

경우 출발점으로 처리한다.

예) ① 서울에서 대구까지 4시간이 걸린다.

② 영화는 신생님이 되었다.

③ 철수는 학교에 갔다.

④ 인상을 당의정으로 가깝다.

⑤ 아버지는 밤에게 의술을 급하였다.

3.7 도착점(Goal)

결재(object)가 미치는 도달 지점을 나타내는 구성요소에

배당되는 의미역으로 출발점에 대조되는 개념이다.

예) ① 서울에서 대구까지 4시간이 걸린다.

② 영화는 신생님이 되었다.

③ 철수는 학교에 갔다.

④ 인상을 당의정으로 가깝다.

⑤ 아버지는 밤에게 의술을 급하였다.

3.8 도구(Instrument)

동자가 나타내고 있는 사건, 상태를 변화시키거나, 행위

를 작동시키는 데 도구로서의 관여는 구성요소로 갖는 의

미역을 가리킨다.

예) 영화는 볼레 미역을 벗었다.

3.9 이유(Reason)

사건의 이유나 원인을 나타내는 구성요소에 주어진다. 도
3.16 경로(Path)
경로의 의미적은 도착점(goal)이나 방향(direction)처럼 이 동의 개념을 갖고 있긴 하지만, 그들과 달리 단순히 지나가는 경로지기 경우에 할당된다.
예) 학생들은 정문으로 다닌다.

3.17 재료(Material)
사건이나 상태를 변화시킨다거나, 행위를 작동시키기 위해 이용되는 구성요소에 할당되는 의미인 도구와 달리, 재료의 의미적은 결과물의 요소를 이루 경우 할당된다.
예) 철수는 나무로 하대를 만들었다.

지금까지 본 연구에서 고려할 구문관계와 의미적에 대해 살펴보았는데, 이를 언어학 논지에서 제시하고 있는 일반적 원칙과 말뭉치 분석 결과, 그리고 세종천사사천의 정보 등을 종합하여 구문관계에 따른 의미적을 정리해 본 것이 <표 1>이다.

(표 1) 구문관계에 따른 의미적

<table>
<thead>
<tr>
<th>구문 관계</th>
<th>의미적</th>
</tr>
</thead>
<tbody>
<tr>
<td>주 어</td>
<td>행위주, 대상, 경위주, 수평주</td>
</tr>
<tr>
<td>목적어</td>
<td>대상</td>
</tr>
<tr>
<td>보 어</td>
<td>대상, 도착점</td>
</tr>
<tr>
<td>부사어</td>
<td>예, 장소, 도착점, 기준점, 대상, 이유, 도구, 시간</td>
</tr>
<tr>
<td>로</td>
<td>도구, 재료, 경로, 방향, 도착점, 자격, 이유, 시간</td>
</tr>
<tr>
<td>에서</td>
<td>장소, 출발점, 행위주, 기준점</td>
</tr>
<tr>
<td>에게</td>
<td>경위주, 행위주, 수평주, 도착점</td>
</tr>
<tr>
<td>기타</td>
<td>기준점, 경로, 동반자, 자격, 도구, 출발점</td>
</tr>
</tbody>
</table>

4. 의미적 결정 규칙

4.1 규칙 기술에 사용되는 자질
 의미적을 결정하기 위한 규칙의 기술 위해서는 구문관계, 의미부류, 형태소 정보와 같은 자질을 사용한다. 지배소와 의존소간의 구문관계에 따라 가능한 의미적의 후보가 달라지므로 구문관계는 모든 규칙에서 사용될 수 있는 중요한 자질이며, 목적어의 유무와 같은 정보도 의미적 결정에서 유효하게 사용된다. 사동사의 주어는 행행위의 의미적을 가지고, 파동사의 주어는 대상의 의미적을 갖기 때문에 동사가 사동사인지에 대한 정보는 의미적을 결정하는데 상당한 기여를 한다. 사동(causativization)은 주어 자격의 동
작가가 다른 동작자로 하여금 어떤 동작을 일으키게 만드는 것을 의미하기 때문에, 사물문에는 항상 특이하게 나타나며 목적이 유무 정보를 이용해서 사물 주어의 의미를 결정할 수 있다.

또 본 시스템은 단어의 의미 중심성이 해소된 결과를 입혀므로 받기 때문에, 지배아의 의미부류와 의존소의 의미부류 값을 얻을 수 있는데, 가도카와 시스러시즈[19]의 그 의미부류를 사용하고 있다. 가도카와 시스러시즈는 총 1,110개의 개념과 4단계의 계층구조를 가지고 있으며, L1, L2a, L3b, L3c 레벨에 속해 있는 개념들은 각각 10개의 하위 개념들로 나뉜다 (그림 2). 명사와 동사의 분류는 하나의 계층구조에 존재하며, 동사의 의미 부류는 주로 L100 레벨의 의미 코드 2xx, 3xx, 4xx에서 나타난다.

(그림 2) 가도카와 시스러시즈의 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 상 대거리와 같은 차이가 의미 결정에 유용하지 않다고 증명한 사실과도 그 백작을 같이 한다. 그러나 아래 예와 같이 두 주어가 모두 접속사를 가지고 있으나 모두 보조사를 가지고 있을 경우에는 위치 정보로 의미를 결정할 수 있다.

① 철수(경향주) 영화는(대상) 산다.
② 철수가(경향주) 영화가(대상) 산다.

또 지배소의 이외 또는 품사 정보, 의존소의 명사형 전성 어미 포함 유무와 같은 형태소 정보도 의미 결정에 사용될 수 있다. 지배소의 이외를 고려하는 경우에는 '느라다'와 같은 동사가 목적어가 되면서 주어의 의미로 행위주가 아닌 경험주를 취하는 경우로만 제한한다.

③ 나는(경향주) 술을 느꼈다.
④ 영화가(경향주) 예쁘다(행용사).
⑤ 키가(대상) 크다(행용사).
⑥ 얼마나 갔기에(이유) 눈이 부시나?

지배소의 품사는 주어가 가질 수 있는 의미의 부호를 줄여주는 역할을 한다. ④의 '예쁘다', ⑤의 '크다'와 같은 행용사는 주어가 의미적으로 경험주와 대상임을 가지게 해주며, 그리고 의존소에 명사형 전성어미를 포함하고 있으면 6에서처럼 부사이 '예'는 의외의 의미를 가진다.

4.2 규칙의 구축

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

(그림 3) 규칙 구축 방법

의미코드가 포함된 구문패턴은 포함공덕 기초 및 언어공학 연구실에서 개발한 한일 기계번역 시스템(COBALT-KJ)[16]을 사용해서 추출되었다. 이 기계번역 시스템은 내부적으로 단위 의미 중심성 해소를 위해 가도카와 시스러시즈의 의미 코드로 표현된 접속어 정보를 사용하고 있는데, 단위 의미 중심성 해소가 끝난 단계에서 의미코드를 한국어 어휘와 동시에 출력하게 수정하였다. 그렇다면, 번역이 출력된 문장의 각 이외는 의미코드가 부가되어 있게 되는데, 이렇게 번역된 문장은 의외 문법을 이용하여 부분 구문 분석[17]을 하여 (그림 4)와 같은 구문 패턴을 얻을 수가 있게 된다. 7.000만

(그림 4) COBALT-KJ를 이용해서 추출한 구문패턴의 의미의 패턴으로의 사상
어질의 KIBS(Korean Information Base System, 1994~1997) 한국어 원시 발명된 본질환 구성에 총 208,088개의 의미 패턴을 생성하였다. (그림 4)는 선택된 구문 패턴을 이해할 논지의 일반적 규칙 및 작업자의 언어적 적관에 의하여 의미의 패턴으로 변환하는 예를 보여 주고 있으며, (그림 5)는 세중 전자사전의 적절 정보로부터 의미의 패턴 정보를 얻어내는 과정을 예시하고 있다. 세중 전자사전(용언사전)의 적절 정보 (그림 6)에는 격자의 유형도 함께 기술되어 있으므로 이를 추출하여 수정된 한글 기계번역 시스템에서 번역하면 역시 대상어휘의 의미 코드를 알 수 있게 된다.

(그림 5) 세중 전자사전 격자 정보의 의미 패턴으로의 사상

(그림 6) 세중전자사전 중 용언사전의 격자정보

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

<table>
<thead>
<tr>
<th>주 어</th>
<th>의미의 수</th>
</tr>
</thead>
<tbody>
<tr>
<td>서</td>
<td>12</td>
</tr>
<tr>
<td>목적어</td>
<td>1</td>
</tr>
<tr>
<td>보 어</td>
<td>2</td>
</tr>
<tr>
<td>부사어</td>
<td>11</td>
</tr>
<tr>
<td>에</td>
<td>11</td>
</tr>
<tr>
<td>표</td>
<td>10</td>
</tr>
<tr>
<td>에서</td>
<td>4</td>
</tr>
<tr>
<td>에게</td>
<td>4</td>
</tr>
<tr>
<td>기타</td>
<td>11</td>
</tr>
<tr>
<td>합 계</td>
<td>56</td>
</tr>
</tbody>
</table>

이미지 결정 규칙의 예는 (그림 7)과 (그림 8)에 제시되어 있으며, 예시된 규칙에서 사용된 기호로 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 의미

IF (lex == "하하") AND pos == "일반동사"
THEN role = 행위주
ELSE IF (lex == "와" AND pos_in = "동사과행절")
THEN role = 대상
ELSE IF (lex_in == "시키" OR lex_in == "시") AND
pos_in == "동사과행절"
THEN role = 행위주
ELSE IF (lex_in == "하하" AND gov_con == "333") AND
(lex == "받다" OR lex == "물리받다") AND
(gov_con == "265 OR gov_con == 579")
THEN role = 수행주

위 규칙에서 사용된 개념코드의 의미
332 : 청취, 265 : 영향, 370 : 수수

(그림 7) 주어의 의미결정 규칙 예

IF (gov_con == 210 OR gov_con == 217 OR gov_con == 230)
THEN role = 출발점
ELSE IF (710 <= dep_con <= 725) AND (pos == "일반동사")
THEN role = 행위주

위 규칙에서 사용된 개념코드의 의미

(그림 8) 부사어 예시의 의미결정 규칙 예

규칙의 적용은 지배소의 어휘와 같은 구체적인 자료를 이용하는 규칙을 먼저 적용하게 되며, 의미부류와 같은 정보를 이용하는 규칙은 나중에 적용된다[14]. 만약 적용된다
규칙이 없는 경우에는 구문관계와 의미역간에 고민도로 나타나는 의미역을 기본(default) 의미역으로 할당하게 된다.

5. 실험 및 평가

실험은 크게 두 가지로 나누어져서 이루어졌다. 하나는 형태소 분석, 구문 분석 및 단어 의미 중의 성, 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(lineal) 모델로도 실험을 따로 하였다. 주어, 목적어, 보이는 대상으로, 부사어 ‘에’ 예시는 장소로, 부사어 ‘에게’는 도착점으로, 부사어 ‘와’는 동반주로 기본 의미역을 설정하였다.

실험방법론으로는 한국전자통신연구원에서 주관한 한국어 형태소 분석기 및 품사대화 모형을 위크숍(MATEC99)에서 제공받은 말뭉치에서 임의로 추출한 340문장 사용하였다. 먼저 형태소 분석, 구문 분석 및 단어 의미 중의 성, 해소 등 전단계의 오류를 모두 수정한 후의 적용결과가 <표 3>에 제시되어 있다.

<table>
<thead>
<tr>
<th>주어</th>
<th>목적어</th>
<th>보이</th>
<th>부사어</th>
</tr>
</thead>
<tbody>
<tr>
<td>예제</td>
<td>예제</td>
<td>기타</td>
<td></td>
</tr>
<tr>
<td>기본</td>
<td>55</td>
<td>90</td>
<td>50</td>
</tr>
<tr>
<td>구축</td>
<td>84</td>
<td>90</td>
<td>70</td>
</tr>
</tbody>
</table>


본 연구를 통해 구축된 구체 모델은 기본 모델에 비해 37%의 성능 향상을 보이고 있다. 그런데 실험결과에서 목적어가 가질 수 있는 의미역이 대부분, 하나인에도 불구하고 정확률이 90%인 것은 의미역을 할당해야 하는 다른 구성요소가 있기 때문이다. 의미역이 할당되지 않는 구성요소는 ‘서술성 명사 + 하다’형의 경우에서 ‘서술성 명사가 분리되어 나타나는 경우로 주로 주어나 목적어로 설정된다. ‘생각하게’에서 서술성 명사 성격이 생각을 하다, 생각을 나타나, 생각의 되다, 생각이 등과 같은 형태를 보이며, 주어로 실험할 때 결정하는 동사는 하나, 뒤다, 뒤다 라는 아니라 서술성 명사의 특성에 따라 다양하며 이 경우 동사

우는 정확률에 상당한 영향을 미치게 된다. 물론 구문 분석 단계에서 이와 같은 유형의 음영을 하나의 단위로 묶어서 처리한다면 이 부분은 문제가 되지 않을 수 있다. 또 다른 해결방안으로 구문 트리에 변경하기 전처리를 생각할 수 있다. ‘생각을 하다가 생각하다’가 사실상 같은 의미를 갖고 있기 때문에 의미학적으로는 같은 트리를 가져야 한다. 따라서 이런 경우에 트리의 변경이 필요하다는 사실을 알 수 있는데, 본 연구에서는 구문 트리에서 의미 트리로 사상 시 트리 변경이 없다고 가정했기에 추후 좀 더 고려해 보아야 할 부분이라 하겠다.

부사어의 경우 및 같이 ‘-에’와 ‘-로’의 구문이 ‘에서’와 ‘에’가보다 상대적으로 어렵다는 것을 실험 결과를 통해 알 수 있다.

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

<table>
<thead>
<tr>
<th>주어</th>
<th>목적어</th>
<th>보이</th>
<th>부사어</th>
<th>평균</th>
</tr>
</thead>
<tbody>
<tr>
<td>예제</td>
<td>예제</td>
<td>기타</td>
<td></td>
<td></td>
</tr>
<tr>
<td>기본</td>
<td>38</td>
<td>77</td>
<td>59</td>
<td>29</td>
</tr>
<tr>
<td>구축</td>
<td>56</td>
<td>77</td>
<td>100</td>
<td>68</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>주어</th>
<th>목적어</th>
<th>보이</th>
<th>부사어</th>
<th>평균</th>
</tr>
</thead>
<tbody>
<tr>
<td>예제</td>
<td>예제</td>
<td>기타</td>
<td></td>
<td></td>
</tr>
<tr>
<td>POS</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Parser</td>
<td>32</td>
<td>15</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>WSD</td>
<td>33</td>
<td>15</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Roles</td>
<td>44</td>
<td>23</td>
<td>0</td>
<td>32</td>
</tr>
</tbody>
</table>

6. 결론 및 향후 계획

본 논문에서는 의미 분석의 한 부분인 의미역 결정을 위한 구문을 대규모 말뭉치와 기계학습에서 추출한 쌍자사전 활용하여 구축하는 방법을 제시하였다. 사실 적합한 연구적이 의미역 결정에서 매우 중요한 역할을 하지만, 이것이 가용하는 것은 각 논문이 구체적으로 기존 연구 자원을 최대한 활용하여, 구체적 구축이 보다 간편적이 고 효율적으로 이루어지게 하였다. 일반적인 연구적 의미역 결정 기법이 가지고 의미역 결정을 기술하는 것은 작업자 주관에 따라 결과가 많이 달라질 수 있으며, 또 미처 생
각지 못한 부분들도 있을 수 있기 때문에 규칙의 적용률이 떨어질 수 있다. 하지만 본 논문에서 제시하는 방법은 대량의 원시 발음치를 기계변환 시스템으로 분석하여 의미 정보가 대량된 문장패턴을 추출함으로써 실제 언어의 다양
한 사용례를 반영하였으며, 또 다른 언어학자들에 의한 교육적
적 연구에 있는 세부적인 사항은 본 논문에 의해 구축된 규칙들은
보다 객관적이고 효율적이라 할 수 있다.
또 의미학을 보다 정확하게 결정하기 위해 사용될 수 있
는 자질 정보(구문관계, 의미부류, 형태소 정보, 어중주의
의치정보 등)를 가능한 모든 포함시켰다. 특히 의미부류가
구체의 기술에 이용되었기 때문에 규칙의 적용률이 항상
될 효과를 가질 수 있다. 규칙 모델의 본질적인 장점이
적용되지만 하면 정확한 결과를 얻을 수 있다는 것이지만,
모든 경우를 규칙으로 해결할 수 없다는 단점도 존재한다.
그러나 규칙으로 해결되지 않는 부분에 대한 처리를 현재는 기본 의미(의사소로서 경위)로
정점을 부여하고, 향후에는 허용모델을 도입하여 사항을 개선하
는 연구를 할 예정이다. 규칙 모델의 또 다른 단점이라면 규
칙의 수가 많아질수록 기존 규칙의 충돌이 발생할 가능
성이 높아지다는 점도 들 수 있는데, 본 연구에서는 이의
정보를 사용한 구체적 적용액과 의미코드 정보를 사
용한 구체적 앱용지를 통해
이 문제를 해결하고 있다.
본 연구의 결과는 응용기(ontology)의 구축시 개념간
 개념관계의 추출이나 기계 변환(automated translation), 정보
 응답 시스템 등과 같은 응용분야에서 활용될 수 있다.
기존 의미학 결정 연구는 연구 범위가 많고, 방법이 다
려서 성공적 비교는 두가지만지만, 필요로 도러
난 정확성을 증명하기 위해서 본 연구에서 구축한 시스템의
70~82%에 이르는 기존 연구에 비해 88%로 다소 좋은 성
능을 보이고 있다.
향후에는 본 시스템을 이용하여 의미학적 대상
어를 반자동으로 구축하는 방법과 의미학적 환경에서는
구성요소를 고려하기 위해 트리가 변환되는 부분을 고려할
예정이며, 정확화가 향상으로 낮은 부분을 "-에"와 "-로"의
성도를 위한 새로운 방법을 연구할 예정이다.

참고 문헌
[2] 김나리, 김영택, "한국어 문학 패턴에 기반한 한국어 문학 분
석과 현대 변화의 모호성 해결", 한국정보과학회논문지, 제23
권 제7호, pp.766-775, 1996.
[3] 남기철, "국어 조사의 유용"-에"와 "-로"를 중심으로", 서해학
술자료, 1993.
[4] 박성해, 김영태, "한국 기계변환에서 결정 트리 활용에 의한
한국어 부사작 조사의 의미 정의 방법", 한국정보과학회논
제3호, pp.405-426, 1999.
[7] 양범철, 음자석, "기계변환에 의한 단어의 적 현형성 자동 확
[9] 이종석, "국어문학의 주체론 연구", 서울대학교 박사학위논
문, 1996.
[10] 이호봉, "구문론론구조에서 중심언어 방식 기계변환을 위한
개념그래프의 생성", 포항공과대학교 전자계산학과 박사학
[12] 조영일, "NP로의 의미학", 제16권 한국어학회 전국 학회 대
[13] 조영일, 김강원, "한국어 의의 체계적 종성의 해석에 대한 연구
", 정보과학회, 제14권 제7호, pp.71-83, 1996.
Roles," In Proceedings of the 38th Annual Meeting of
Association of Computational Linguistics, Hong Kong, pp.
tactic Interpreter", in Relational Models of the Lexicon:
Representing knowledge in Semantic Networks, Edited by
M. W. Evans, Cambridge University Press, pp.113-138,
[16] K. H. Moon and J. H. Lee, "Representation and Recogni-
tion Method for Multi-Word Translation Units in Korean-
to-Japanese MT System., in The 18th International Confer-
cence on Computational Linguistics (COLING 2000), Ger-
ity in Inter-chunk Dependency Parsing", NPLRS 2001 (6th
Natural Language Processing Pacific Rim Symposium), Tokyo,
Korean Relative Clauses using Idiomatic Patterns," In
Proceedings of the 17th International Conference on Com-
Sense Disambiguation Using Neural Networks with Con-
cept Co-occurrence Information", NPLRS 2001 (6th Natural
Language Processing Pacific Rim Symposium), Tokyo, Ja-
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, 기계
번역, 정보검색 등