영한 기계번역에서 전치사구를 해석하는 시스템

강 원 석†

요 약

영한 기계번역에서 전치사구의 해석은 부착의 문제(Attachment Problem)와 의미 해석의 문제, 그리고 해석에 필요한 정보 획득의 문제가 있다. 이 세 가지 문제를 해결하기 위하여 본 논문은 전치사구 해석 시스템을 제시한다. 이 시스템은 규칙 제어기와 신경망의 하이브리드 구문해석 시스템, 격의미 해석 시스템, 그리고 신경망의 입력 정보를 자동으로 생성하는 의미속성 생성기로 구성된다. 의미속성 생성기는 시스템의 입력이 되는 의미속성을 자동으로 생성하는 방법으로 인위적인 방법의 단점을 보완하여 객관성 있는 전치사구 해석을 하게 한다. 격의미 해석 시스템은 영한 기계번역에 맞는 격의미를 찾아내어 자연스런 한국어 생성을 하게 하고 구문해석 시스템은 규칙 방법의 장점과 신경망 방법의 장점을 취한 하이브리드 방식의 시스템으로 전치사구 부착의 문제를 해결한다.

An Analysis System for Prepositional Phrases in English-to-Korean Machine Translation

Won Seok Kang†

ABSTRACT

The analysis of prepositional phrases in English-to-Korean machine translation has problem on the PP-attachment resolution, semantic analysis, and acquisition of information. This paper presents an analysis system for prepositional phrases, which solves the problem. The analysis system consists of the PP-attachment resolution hybrid system, semantic analysis system, and semantic feature generator that automatically generates input information. It provides objectiveness in analyzing prepositional phrases with the automatic generation of semantic features. The semantic analysis system enables to generate natural Korean expressions through selecting semantic roles of prepositional phrases. The PP-attachment resolution hybrid system has the merit of the rule-based and neural network-based method.

1. 서 론

영한 기계번역에서 전치사구의 해석은 다음과 같은 세 가지의 문제를 가지고 있다. 그것은 전치사구 부착의 문제, 의미 해석의 문제, 그리고 번역시에 필요한 정보의 정의와 획득의 문제이다. 전치사구 부착의 문제는 전치사구가 어떠한 구와 관계를 가지는지에 대한 구조적 애매성의 문제이다. 지금까지 이 전치사구 부착의 문제를 해결하기 위하여 많은 연구[2, 5, 6, 16, 17]가 진행되어 왔다. [5]의 연구는 단지 구문정보만을 이용하여 전치사구 문제를 해결하려 하였

† 정 회 원:안동대학교 컴퓨터공학교육과
논문접수:1996년 3월 15일, 심사완료:1996년 8월 12일

다. 그렇지만 구문 정보만을 이용하여 전치사구 부착의 문제를 해결하는 것은 어렵다. 이 문제를 해결하기 위하여 선호규칙 방법[6, 14, 16, 17]이 적용되었다. 선호규칙 방법은 전치사구 부착의 두 경쟁 후보 가운데 하나를 선호하는 것을 규칙화한 것이다. 그렇지만 이 방법은 경쟁하는 후보들에 대한 모든 조합을 고려해야 된다.

신경망 방법은 신경망 체계[11, 13]를 이용하여 애매성의 문제를 해결한다. 신경망 체계는 신경망이 자동으로 구축되고 계속 수정 보완하기 쉬운 이점이 있는 반면, 정해진 크기의 입력 단위소자를 가져야 한다. 일반적으로 주어진 문장의 길이는 변하므로 가변의 입력을 고정된 크기의 구조로 처리해야 하는 문제가 있다. recurrent 신경망[4]이 이 문제를 해결하지만 학습 시간이 너무 길다. 본 논문에서는 전치사구 부착의 문제를 해결하기 위하여 규칙 제어기와 신경망의 하이브리드 시스템을 구축하고, 이 시스템으로 학습 시간의 문제와 가변의 입력 문제를 해결한다.

영한 기계번역에서는 전치사구 부착의 문제 뿐 아니라 목적언어인 한국어를 생성하기 위한 전치사구의 격의미 해석의 문제를 해결해야 한다. 격의미 해석은 격의미 체계를 이용하여 전치사구가 어떠한 격의미를 가지는지를 밝히는 것이다. 격의미 체계는 전치사구의 모든 가능한 격의미를 표현할 수 있어야 하고, 그리고 자연스런 한국어를 생성할 수 있어야 한다. 이미 격의미 체계에 대한 연구[4, 11]가 있었지만 이 격의미 체계는 한국어의 표현에 맞지 않다. 본 논문은 한국어 표현에 맞는 격의미 체계를 설계하고 2단계의 신경망 시스템을 구축하여 전치사구 의미 해석에 적용한다.

전치사구 해석 시스템은 해석에 필요한 정보를 시스템이 필요로 하는 표현으로 획득해야 한다. 본 논문의 시스템은 자동으로 시스템이 구축되는 신경망을 이용한다. 신경망은 객관성과 확장성에 이점이 있으나 이 이점을 뒷받침하기 위해서는 시스템에 필요한 의미속성의 객관적인 획득 방법이 필요하다. 본 논문은 영한 기계번역의 전치사구 해석 시스템에 적합한 의미 속성 집합을 정의하고 그 의미 속성을 자동으로 획득하는 의미 속성 생성기를 제안한다.

본 논문의 2장은 전치사구 해석에 필요한 격의미 체계를 기술하고 3장은 이미 속성 집합의 정의와 획득 방법을 설명한다. 4장은 전치사구 해석 시스템의 전체적인 구조와 동작을 설명하고 5장은 시스템의 실험과 그 결과를 기술한다. 마지막으로 6장은 본 논문의 전치사구 해석 시스템에 대한 결론을 맺는다.

2. 격의미 체계

영한 기계번역의 목적언어인 한국어는 구성 성분의 격역할을 표현하는 방법이 영어와 많이 다르다. 한국어는 교착어이기 때문에 구성 성분에 붙는 첨가어-조사나 어미가 그 구성 성분의 역할을 나타낸다. 첨가어는 그 구성 성분이 위치가 변하더라도 그 성분의 역할을 그대로 전달한다. 영어는 전치사가 그 역할을 하는데 전치사는 위치가 바뀌면 역할이 달라진다. 이 사실은 한국어의 첨가어가 전치사보다 역할의 표현을 구체적으로 한다는 것을 의미한다. 따라서 영한 기계번역에서 좋은 번역을 하기 위해서는 첨가어의 구체적인 격역할을 반영하는 격의미 체계가 필요하다.

다음과 같은 문장이 격의미 체계의 필요성을 보여준다.

(1) He put the fertilizer *in the hole*.
(2) The ear thquake occurred *in Japan*.
(3) The pencil is *in the drawer*.

전치사 in은 같은 장소격이지만 한국어 조사 표현은 (1)의 경우는 에가, (2)의 경우는 에서가, (3)의 경우는 다시 에가 적합하다. 이것은 전치사구와 부착되는 동사의 유형이 그 격의미 표현의 조사를 결정짓는다는 것을 알려 준다. 동사가 상태 동사인 경우 그 전치사는 물체가 어떠한 상태에 놓여 있는가를 나타내게 되고 그 전치사에 대한 한국어 조사는 에가 사용된다. 동사가 사건을 나타내는 경우 그 때 사용되는 전치사는 사건의 발생 장소를 표현하며 한국어 조사 에서가 사용된다. 그리고 동사가 목적지를 갖는 행위 동사이면 전치사는 그 목적지를 나타내며, 그 전치사에 해당하는 한국어 조사는 에가 된다. 이 에는 (1)의 에와 표현은 같지만 그 의미가 다르다.

전치사 from의 경우를 보자. 이 전치사는 한국어 격조사 에서, 에게서, 부터, 에서 등의 용례로 표현된다.

(4) The baby gets nutriments *from the milk.*
(5) I buy the desk *from him.*
(6) The man finds *from the front* to rear.
(7) It is few feet *from the window.*

먼저 전치사구의 목적구가 무생물인 경우와 유생물인 경우를 구분해 본다. 각 경우에 대응될 한국어 격조사가 다르다. 무생물인 경우 에서라는 한국어 조사가 선택되어 사용되고 유생물인 경우 존칭의 의미가 깃든 에게서가 선택되어 사용된다. 그리고 전치사가 시간의 시점으로 사용되는 경우는 한국어 조사 부터를 선택하여 표현되어야 하고, 전치사가 어떠한 상태의 표현을 위해 사용되는 경우는 다른 의미의 한국어 조사 에서가 선택되어야 한다.

예문은 영어의 전치사가 여러 개의 한국어 조사로 번역될 수 있음을 보인다. 영한 번역에서는 전치사에 대해 어떠한 한국어 조사가 대역되어야 하는지에 대한 구체적인 격의미 해석이 필요하다. 본 논문은 구체적인 격의미 해석에 필요한 한국어 격조사 표현의 격의미 체계를 구성하였다[9]. 이 격의미 체계는 코퍼스의 예문을 분석하여 전치사가 가질 수 있는 격의미를 모두 포함시켰다. 이 격의미 체계는 자연스런 표현의 격조사를 의미하는 63개의 격의미로 구성된다.

3. 의미 속성

3.1 의미 속성 집합

기계 번역 시스템은 번역에 필요한 정보를 시스템에 이용하기 편리한 형태로 표현하고 이를 이용한다. 의미 속성 집합도 그 정보 표현 방법 가운데 하나로서 신경망으로 구성된 시스템에서 많이 사용된다. 본 논문의 시스템도 신경망을 이용하므로 정보를 표현하기 위해 의미 속성 집합을 사용한다.

의미 속성 집합은 전치사구를 해석하여 어떠한 한국어 표현으로 번역해야 적당한지를 선택할 수 있는 정보를 내포하고 있어야 한다. 즉, 의미 속성 집합이 전치사구의 의미를 구분할 수 있는 변별성(distinction)의 특징을 가지고 있어야 한다. 변별성이 부족하게 되면 전치사구의 의미를 파악할 수가 없게 되어 전치사구의 뜻에 맞는 한국어 문장을 생성할 수가 없다. 의미 속성 집합은 변별성과 함께 보편성(generality)의 특징도 가지고 있어야 한다[1, 2, 11]. 기계번역에서 선택의 문제를 해결하기 위하여 사용하는 의미 속성 집합은 변별성이 높을수록 문제를 간단히 해결할 수가 있다. 그러나 변별성을 높이려고 하면 의미속성 집합의 요소들의 수가 많아진다. 그리고 많아진 의미 속성 집합의 요소들은 그 사용빈도가 아주 작을 수 있다. 잘 사용하지 않는 의미 속성들은 의미 속성 집합의 엔트리 수를 늘여 시스템의 수행 효율을 저하한다. 따라서 의미 속성 집합은 변별성과 보편성을 함께 갖춘 요소로 구성해야 한다. 그러나 변별성과 보편성을 갖춘 의미 속성 집합의 설계는 쉽지 않다. 본 논문에서는 변별성을 갖추고자 의미 분류에 대한 연구[1, 7, 9]를 토대로 한국어의 특성을 반영하였고 보

의미속성	예
1 thing	entity
3 animate-thing	life form
6 human	person, man
5 animal	beast, horse
5 plant	tree
3 inanimate-thing	inanimate object
4 material	substance, material
5 edible-thing	food
4 man-made-thing	artificial thing
....	.....
3 intellectual-thing	knowledge, method
.....	.....

(그림 1) 명사류의 의미 속성의 예
(Fig. 1) Example of Semantic Features of Noun

편성을 갖추고자 일반의 범용 사전과 시소러스[3, 7, 12]의 의미 분류를 토대로 의미 속성 집합을 설계하였다.

WordNet[12]는 프린스턴 대학에서 언어 심리학자들과 어휘론 학자들에 의해 개발된 시소러스이다. 이 시소러스에는 최상위에 40개의 개념을 가지고 있고 각 단어에 대하여 해당하는 개념들을 추출할 수 있다. 40개의 개념을 의미 속성으로 하여 해석 시스템의 신경망에 이용하였다. 그러나 40개의 개념은 너무 추상적이어서 신경망 시스템이 전치사구 구문/의미 해석할 때 좋은 결과를 가져오지 못하였다. 이 시소러스는 전부 49000여 개의 개념을 가지고 있다. 이 개념들을 모두 의미 속성으로 하면 변별성은 얻을 수 있으나 보편성은 얻지 못한다. 본 논문은 이와 같은 점을 고려하여 [1, 3, 7, 9, 12]의 의미 상하위 관계를 기초로 전치사구 부착의 문제와 격의미 해석의 문제를 효율적으로 해결할 수 있고 보편성을 잃지 않는 104개의 의미 속성 집합을 정의했다.

의미 속성 집합은 동사류와 명사류, 형용사류의 의미 속성으로 구성된다. 명사류에 대한 의미 속성 집합의 구성요소는 그림 1과 같다. 그림 1에서 의미 속성의 앞에 나오는 숫자는 상하위 관계에서의 레벨을 표시한다.

동사류의 의미 속성 분류는 그림 2와 같다. 동사류의 의미 속성은 mesh 형태의 상하위 관계를 가진다. 즉, 각 의미 속성은 그 의미 속성의 부모 의미 속성이 둘 이상이 될 수 있기 때문에 의미 속성들의 관계가 복잡하다. 이와 같은 사실은 규칙 중심의 기계번역 시스템에 규칙 설계를 어렵게 하지만, 신경망 중심의 기계번역 시스템에는 문제가 되지 않는다. 신경망 시스템에서는 의미 속성들의 관계를 인위적인 조사와 연구를 통하여 얻지 않고 자동의 알고리즘을 통하여 얻기 때문이다[11]. 형용사류의 의미속성의 예는 그림 3과 같다.

의미속성	예
1 action	action, behave
1 event	event
2 accomplishment	succeed
2 nonaccomplishment	fail
2 composition	construction, destruction
2 movement	movement, put
2 intellectual-behavior	rational action
2 control-behavior	manage
2 ownership	buy, pay
2 change	change, soften
.....	.....

(그림 2) 동사류의 의미 속성의 예
(Fig. 2) Example of Semantic Features of Verb

의미속성	예
1 location	location
2 location-point	floor
2 territory	space
1 time	time
2 time-point	10 o'clock
2 duration	3 days
1 direction	in-direction
1 measure	distance
1 state	stable
1 feature	beautiful
.....	.....

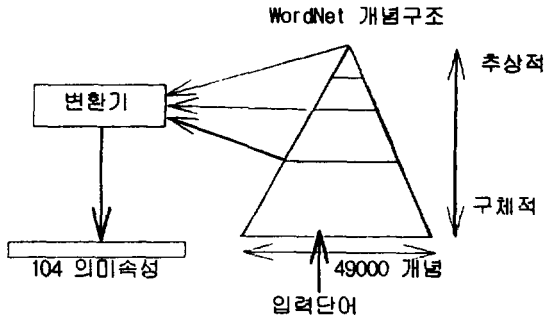
(그림 3) 형용사류의 의미 속성의 예
(Fig. 3) Example of Semantic Features of Adjective

3.2 의미 속성 생성기

의미 속성 집합의 설계가 잘 되었다고 하더라도 이를 획득하는 방법이 시스템에 의존한다면 보편성을 잃게 된다. 대부분의 기계번역 시스템은 시스템에 필요한 정보를 시스템의 전자 사전으로부터 얻는다. 이 방법은 시스템을 확장하거나 다른 영역에 응용할 때 전자 사전을 다시 구축해야 하는 문제가 있다. 또한 전자 사전의 재 구축은 근본적으로 시스템의 구조를 변경시켜야 하는 사태도 일어날 수가 있다. 본 논문에서는 의미 속성 획득에 보편성을 얻기 위하여 시스템에 의존하지 않는 대규모의 범용 전자 시소러스인 WordNet로 부터 필요한 의미 속성 정보를 추출한다.

WordNet[12]는 3.1에서 이미 언급한 바와 같이 범용의 목적으로 만든 대규모의 시소러스로 총 49000여 개의 개념을 가지고 있고 54000여 개의 단어에 대한 개념을 추출할 수 있다. 그림 4는 의미 속성 생성기의 동작을 묘사하는 그림이다. 그림의 WordNet 개념 구조는 피라미드 형태의 상하위 관계를 형성한다. 상위에 있는 개념들은 추상적인 것이고 하위로 내려갈수록 구체적인 개념들이 위치하고 있다. 이 WordNet는 주어진 단어에 대하여 하위의 개념부터 상위의 개념까지의 관련된 개념들을 추출할 수 있다. 의미 속성 생성기는 WordNet의 이와 같은 기능을 이용하여 주어진 단어의 개념들을 획득하고 이를 정의된 104개의 의미 속성 집합으로 변환한다.

의미 속성 생성기가 변환 과정을 거치지 않는다면



(그림 4) 자동 의미 속성 생성기의 동작

(Fig. 4) Overview of Automatic Semantic Feature Generator

의미 속성들은 WordNet의 개념 가운데서 추출한 것이어야 한다. WordNet의 개념 분류는 한국어의 특성들에 대해 배려하지 않았으므로 이를 직접 영한 기계번역에 사용하기에는 변별성이 부족하다. 변환기는 결국 WordNet의 개념 분류를 변별성이 있는 의미 속성 집합으로 재분류하는 역할을 하여 변별성의 특징을 향상시키는 효과를 가져온다.

변환기는 변환표를 이용하여 추출된 WordNet 개념들을 정의한 의미 속성으로 변환한다. 그 변환표의 일부 예는 그림 5와 같다. 그림 5의 WordNet 개념란에는 여러 개의 단어와 콤마로 하나의 개념을 표시한다. 해당 의미속성란에 나타나는 단어는 정의된 의미 속성의 이름을 나타내고 콤마가 들어 있는 것은 해당

WordNet 개념	해당의미속성
(1) entity	thing
(1.1) life form,..	animate-thing
(1.1.1) person,...	human
(1.1.2) animal,...	animal
(1.1.3) plant,...	plant
(1.1.4) microorganism	microorganism
...	...
(1.2) object,...	inanimate-thing
(1.2.1) natural object	natural-thing
(1.2.2) substance, matter	material
...	...
(5.1) change, alter	action, change
(5.1.1) affect, ...	effect
...	...

(그림 5) 변환표의 일부

(Fig. 5) Part of Mapping Table

하는 의미속성이 두 개 이상이라는 것을 의미한다. 즉, WordNet 개념란의 콤마는 구분자가 아닌 개념 이름을 기술하기 위한 표기이고 의미속성란의 콤마는 구분자이다. WordNet 개념란의 앞에 나오는 숫자는 개념의 상하위 정도를 표시한다. 이 변환표는 명사류의 변환표, 동사류의 변환표, 형용사류의 변환표로 구성된다. 명사류의 변환표는 990개의 엔트리로 구성되고, 동사류는 907개의 엔트리로 구성되며, 형용사는 405개의 엔트리로 구성된다.

그림 6은 단어 rock에 대한 의미 속성을 생성하는 과정을 보여 준다. 단어 rock은 그림 6에서 보는 바와 같이 5개의 서로 다른 뜻을 가지고 있다. 각 뜻에 대하여 그림은 부분적인 상하위 관계를 보여 준다. 그림의 좌측란은 WordNet의 기능을 통하여 출력된 단어 rock의 WordNet 개념들이다. 각 뜻마다 WordNet는 하위 개념부터 점차 상위 개념으로 해당하는 개념들을 차례로 출력한다. 우측란에 콜론(:) 기호 다음에 나오는 것은 변환기를 통하여 생성된 의미 속성들이다. 단어 rock은 최종적으로 material, inanimate-thing, thing, natural-thing, edible-thing, domain, social, action, moving의 의미 속성들이 추출된다.

4. 전치사구 해석 시스템

본 논문에서 정의한 격의미 체계와 의미 속성 생성기를 이용하는 전치사구 해석 시스템의 구조는 그림 7과 같다. 이 해석 시스템은 입력을 정규화하는 전처리기, 자동으로 의미 속성을 생성하는 의미속성 생성기, 입력 문장의 구문적인 형태를 가려내는 구문해석기, 전치사구의 격의미를 밝혀내는 격의미 해석기로 구성된다. 각 구성요소 가운데 의미 속성 생성기에 대한 내용은 3장에 기술되어 있다. 전처리기는 입력을 받아들여 품사가 붙여진 단위구의 리스트를 만들어 구문해석기로 결과를 넘겨준다. 단위구는 하나의

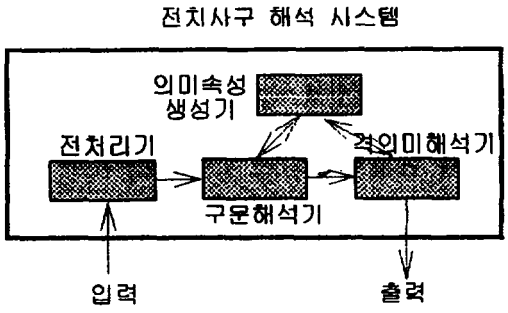
```
5 senses of rock                                          semantic features

Sense 1
rock
 => material, stuff
  => substance, matter                                    : material
   => object, inanimate object, physical object, thing    : inanimate-thing
    => entity                                             : thing
Sense 2
rock, stone                                               : material
 => natural object                                        : natural-thing
  => object, inanimate object, physical object, thing     : inanimate-thing
   => entity                                              : thing
Sense 3
rock candy, rock
 => candy
  => sweet, sweetmeat, confection, confectionery
   => dainty, delicacy, goody, kickshaw, treat
    => aliment, nourishment, nutriment, sustenance, victuals
     => food, nutrient                                    : edible-thing
      => substance, matter                                : material
       => object, inanimate object, physical object, thing inanimate-thing
        => entity                                         : thing

.........
```

(그림 6) 자동 의미 속성 생성기를 통한 의미 속성 생성
(Fig. 6) Generation by Automatic Semantic Feature Generator

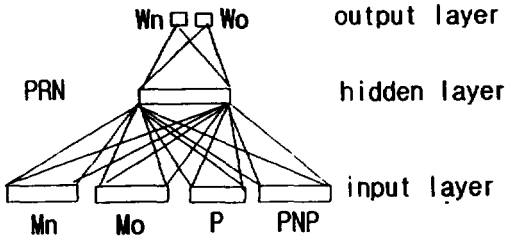
동사 또는 시제가 첨부된 동사, 양상이 첨부된 동사, 단순 명사구, 하나의 전치사구 등이 될 수 있다. 여기에서 단순 명사구라고 하는 것은 관사, 형용사들, 명사들로 구성된 복문이 아닌 명사구를 의미한다.



(그림 7) 전치사구 해석 시스템
(Fig. 7) PP Analysis System

4.1 전치사구 구문해석기

전치사구의 구문적 부착의 문제는 구문 해석기에 의하여 해결된다. 구문 해석기는 하나의 신경망과 제어기로 구성된 하이브리드 시스템이다. 제어기는 신경망의 동작을 관할하고, 신경망의 입력을 검사하여 no-crossing 제한 조건을 충족하는 입력을 제공한다. 신경망은 부착 가능성이 있는 두 개의 후보 가운데서 하나의 후보를 선택한다. 그 구조는 그림 8과 같다.



(그림 8) 구문해석 신경망의 구조
(Fig. 8) Structure of PP Attachment Analysis Neural Network

신경망은 입력층, 은닉층, 출력층의 3개의 층으로 구성되었다. 입력층은 4개의 부분으로 구분되는데 Mn과 Mo는 전치사구의 부착 후보에 해당하는 부분이고 P는 전치사구의 전치사 부분, PNP는 전치사의 목적어에 해당하는 부분이다. 각 부분은 입력 문장의 해당하는 단위구의 의미 속성으로 이루어진다. Mn, Mo, PNP 부분은 의미 속성의 수인 104개의 단위 소자와 품사를 나타내는 2개의 소자로 구성된다. 이 때 Mn과 Mo는 거리 정보를 나타내는 2개의 소자가 더 추가된다. 거리 정보는 PNP에는 포함되지 않고 Mn과 Mo에만 포함된다. 각 단위 소자는 1 또는 0의 정보가 주어진다. P는 전치사를 나타내므로 전치사의 종류만큼 소자의 수가 형성된다. 본 시스템에서는 35개의 단위 소자를 사용하였다. 은닉층은 입력층과 출력층의 단위 소자의 수를 고려하여 120개의 단위 소자로 구성하였다. 출력층은 2개의 소자로 구성되는데 각 소자는 전치사구에 어디에 부착되어야 할 것인가를 표시한다. Mn이 Wo의 결과보다 큰 값이 나온다면 Mn이 전치사구의 피수식어가 되고 그렇지 않다면 Mo가 전치사구의 피수식어가 된다. 신경망의 학습은 [15]의 역전파 알고리듬에 의하여 이루어진다.

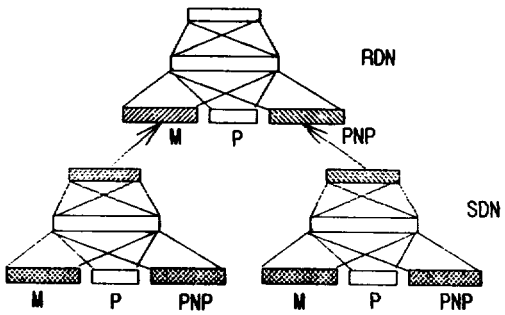
시스템의 신경망은 단지 2개의 후보 가운데서 하나만 선택할 수 있다. 실제로 입력 문장은 전치사구가 부착될 가능성이 있는 후보의 수가 가변적이다. 시스템의 구문 해석기는 제어기를 두어 가변적인 후보의 수에 관계없이 구문해석을 할 수 있도록 한다. 즉 제어기는 여러 후보 가운데서 하나를 선택할 수 있도록 신경망을 감독하여 반복적인 비교 작업을 행한다.

구문해석기의 구문 해석 과정은 다음과 같다.

1. 제어기가 no-crossing 제한 조건을 충족하면서 전치사구에 가까운 후보 2개를 선택하고 의미속성 생성기를 통하여 후보 2개의 의미속성들과 전치사구의 의미 속성들을 추출하고 이를 신경망의 입력으로 전달하고 신경망을 동작한다.
2. 신경망은 주어진 의미속성들을 입력으로 2개의 후보 가운데 보다 나은 후보 하나를 선택한다.
3. 제어기는 선택된 후보 하나와 전치사구에 가까우면서 경쟁에 고려되지 않은 후보 하나를 골라서 신경망의 입력으로 다시 전달한다. 이때 no-crossing 제한 조건을 만족하는 후보임을 검사한다.
4. 경쟁에 고려되지 않은 후보가 없을 때까지 2, 3 과정을 반복하여 최종의 하나의 선택된 후보를 격의미 해석 시스템에 넘겨준다.

4.2 전치사구 격의미 해석기

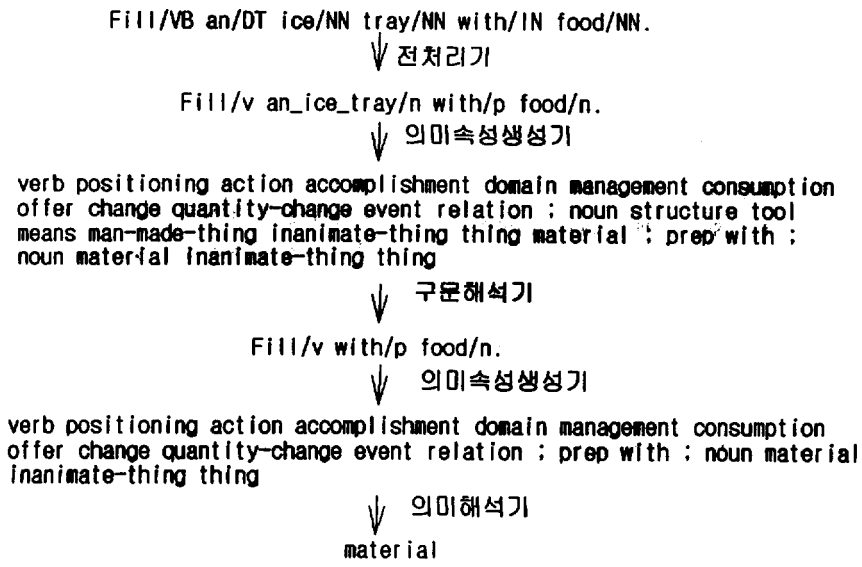
격의미 해석기는 세 개의 신경망으로 구성된다. 두 개의 신경망은 주어진 단어의 의미 애매성 해소를 목적으로 하고 나머지 하나의 신경망은 전치사구의 격의미 해석을 목적으로 한다.



(그림 9) 격의미 해석기
(Fig. 9) Semantic Analyzer

즉 SDN(Sense Disambiguation Network)의 좌측 신경망은 전치사구의 부착 후보에 M에 대한 의미 애매성 해소를 시도하고 우측 신경망은 전치사구의 명사구 PNP의 의미 애매성 해소를 시도한다. 그리고 RDN (Role Disambiguation Network)은 애매성이 해소된 단위구의 의미 속성들을 입력으로 하여 전치사구가 부착된 후보에 대한 격의미를 파악한다. 각 신경망은 입력층에 전치사구의 피수식어에 해당하는 M 부분, 전치사에 해당하는 P 부분, 전치사의 피수식어에 해당하는 PNP 부분이 있다. M 부분과 PNP 부분은 104개의 단위 소자와 품사를 표현하는 2개의 소자로 구성되고, P 부분은 35개의 단위 소자로 구성된다. 그리고 각 신경망의 은닉층은 120개의 단위 소자로 구성된다. 출력층은 SDN과 RDN이 서로 다르다. SDN은 출력으로 나오는 것이 애매성이 해소된 의미 속성이므로 104개의 단위 소자로 구성되고 RDN은 전치사구의 격의미가 출력으로 나오므로 격의미의 숫자만큼의 단위 소자를 지닌다. RDN과 SDN은 입력에 있어서도 단위 소자 수는 같지만 그 의미가 다르다. RDN의 입력 단위 소자들은 애매성이 해소된 것임에 비해 SDN의 입력 단위 소자들은 애매성이 해소되지 않은 것이다. 각 신경망의 학습은 역시 [15]의 역전파 알고리듬을 따른다.

그림 10은 입력 문장에 대한 전치사구 해석 시스템의 처리 예를 보여 준다. 입력 문의 단어 뒤의 기호는 Pentree Bank[10]의 품사 구분 코드이다. 전처리기는 입력 문을 단위구의 리스트로 만든다. 구문 해석기는



(그림 10) 전치사구 해석 시스템의 처리 예
(Fig. 10) Procedure of PP Analysis System

의미속성 생성의 결과를 받아 no-crossing 조건을 검사하고 거리 정보를 추가한다. 의미 해석기는 결과로 전치사구의 격의미 material을 출력한다.

5. 실험과 결과

본 시스템의 실험은 Penntree Bank[10]의 코퍼스 데이터에서 학교 교재의 예문을 선택하여 이루어졌다. 시스템의 각 신경망의 학습의 인수들은 다음과 같다.

〈표 1〉 각 신경망의 인수
〈Table 1〉 Parameters of Neural Networks

	PRN	SDN	RDN
learning rate η	0.2	0.2	0.2
momentum term α	0.2	0.2	0.2
# of epoch	1000	1500	1500
# of input units	357	247	247
# of hidden units	120	120	120
# of output units	2	106	63

PRN은 그림 8의 전치사구 구문해석기의 신경망에 대한 인수이고 SDN은 그림 9의 격의미 해석기의 신경망 중에 의미 애매성 해소를 하는 SDN의 인수를 나타내고 RDN은 격의미 해석기의 격의미 해석을 하는 RDN의 인수를 나타낸다. 구문 해석 시스템의 실험 결과는 표 2와 같다.

〈표 2〉 전치사구 구문 해석기 실험
〈Table 2〉 Experiment of PP-attachment Analyzer

	데이타크기		실험(학습)		실험(비학습)	
실험	PRN	System	PRN	System	PRN(1020)	System(275)
1	1234	413	.995	.985	.877	.756
2	2443	782	.989	.967	.932	.804
3	3648	1076	.993	.944	.944	.844

표 2에서 각 열에 항목이 두개씩 있다. 하나는 PRN에 대한 것이고 하나는 시스템에 대한 것이다. PRN 항목은 구문 해석에 이용한 신경망의 실험에 대한 것이고 시스템 항목은 제어기와 신경망의 하이브리드 시스템의 실험에 대한 것이다. 구문 해석 시스템의 실험의 결과를 보면 학습에 사용한 데이타의 양이 많아짐에 따라 학습에 사용하지 않은 데이타의 실험의 성공율이 점차 높아짐을 알 수 있다. 첫번째 실험의 경우 75%의 결과가 나왔고 두번째의 경우 80%, 세번째의 경우 84%의 결과가 나왔다. 학습 데이타의 양이 많아짐에 따라 성공율이 높아지므로 더 많은 데이타의 학습은 비학습 데이타에 대한 더 높은 성공율을 얻을 수 있다.

실험에 대해 성공하지 못한 사례를 분석한 결과 제어기의 오류, 문맥 정보의 부족, 거리 정보의 부족, 품사 분류의 일반성으로 오류가 일어난다는 것을 알 수 있었다. 시스템의 성공율을 높이기 위해서는 이와 같은 문제를 개선해야 한다.

〈표 3〉 격의미 해석기 실험
〈Table 3〉 Experiment of Semantic Analyzer

실험	학습데이타	실험(학습)	실험(비학습)
1	497	.994	.490(.631)
2	936	.996	.694(.796)
3	1471	.994	.804(.902)
4	2065	.994	.902(.945)

표 3은 격의미 해석의 실험 결과이다. 첫번째 실험의 경우 비학습 데이타에 대해 49%의 성공율을 얻었고, 두번째 실험의 경우 69.4%, 세번째의 경우 80.4%, 네번째의 경우 90.2%의 성공율을 얻었다. 학습에 사용한 데이타의 양이 많아짐에 따라 비학습 데이타에 대한 성공율이 높아지므로 더 많은 데이타의 학습은 더 높은 성공율을 기대할 수 있다. 본 논문의 실험에서 두 번째로 선택된 격의미가 성공한 경우까지 포함한 성공율은 94.5%까지 얻었다. 실패한 경우에 대하여 분석해 본 결과 대부분의 경우가 SDN의 잘못으로 발생하였다. SDN은 문맥 정보의 부족으로 인해 단어의 중의성을 해결하지 못하였다. 나머지 경우는 의미 속성에 기인한 것이었다. 이 문제의 해결을 위해 보다 나은 의미 속성의 정의와 획득이 필요하다.

6. 결 론

본 논문은 전치사구 해석에서 일어나는 전치사구 부착의 문제, 의미 해석의 문제, 의미속성 획득의 문제를 해결하는 전치사구 해석 시스템을 제시한다. 전치사구 해석 시스템은 전치사구 부착의 문제를 해결하는 구문 해석기, 격의미를 파악하는 격의미 해석기,

각 시스템에 필요로 하는 의미속성을 제공하는 의미 속성 생성기로 구성된다. 구문 해석기는 신경망과 제어기로 구성된 하이브리드 시스템으로 입력의 크기가 변하는 문제를 해결하고 학습 시간의 문제를 해결한다. 시스템의 실험 결과를 통해 규칙 중심 방법의 장점과 신경망 방법의 장점을 살릴 수 있는 하이브리드 시스템의 가능성을 보았고 세분된 거리 정보의 추가, 품사 정보의 추가, 제어기의 개선으로 시스템을 개선할 수 있음을 보았다.

본 논문에서 정의한 격의미 체계는 격의미 해석기에 좋은 결과를 가져왔다. 이것은 기계번역 시스템과 자연어 처리의 다양한 영역에 기여 가치가 높다. 특히 의미 속성 생성기는 의미 속성 획득에 객관성을 부여하여 시스템의 신경망의 장점을 더욱 높인다. 이 의미속성 생성기는 시스템의 확장과 타 영역의 응용에 큰 기여를 할 수 있다.

참 고 문 헌

[1] 천기석, *Research on the system of Active Verb and Stative Verb in Korean*, 형설출판사, 1984.

[2] D. K. Dahlgren and J. McDowell, "Using Commonsense Knowledge to Disambiguate Prepositional Phrase Modifiers," *Proceedings of International Conference on Artificial Intelligence*, pp. 589-593, 1986.

[3] EDR Technical Report, Concept Dictionary, Japan Electronic Dictionary Research Institute, 1988.

[4] F. J. Eisenhart, "Instantiating Thematic Roles with a Recurrent Network," Artificial Intelligence Programs, Research Report AI-1993-01, Univ. of Georgia, 1993.

[5] L. Frazier and J. D. Fodor, "The Sausage Machine: A New Two-Stage Parsing Model," *Cognition*, Vol. 6, 1979.

[6] J. R. Hobbs and J. Bear, "Two Principles of Parse Preference," *Proceedings of International Conference on Computational Linguistics*, pp. 162-167, 1990.

[7] Japan National Japanese Institute, *Technical Research on Meanings and Usage of Verb*, Su-Young Publishing Co., 1972.

[8] W. S. Kang, J. Y. Seo, and G. C. Kim, "A Hybrid Method for the Resolution of Prepositional Phrase Attachment Problems," *Proceedings of Natural Language Pacific Rim Symposium '93*, 1993.

[9] W. S. Kang, J. Y. Seo, G. C. Kim, and K. S. Choi, "A Neural Network Method for the Semantic Analysis of Prepositional Phrases in English-to-Korean Maching Translation," *Computer Processing of Chinese and Oriental Languages*, 1994.

[10] M.P.Marcus, B.Santorini, and M.A.Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penntree Bank," *Computational Linguistics*, Vol. 19, No. 2, June 1993.

[11] J. L. McClelland and A. H. Kawamoto, "Mechanisms of Sentence Processing: Assigning Roles to Constituents of Sentences," *Parallel Distributed Processing*, Volume 2, pp. 272-325, 1986.

[12] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Introduction to WordNet: An On-line Lexical Database," Report of WordNet, Princeton University, 1990.

[13] Y.H.Pao, *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley Publishing Co., 1989.

[14] D. Petitpierre, S. Krauwer, D. Arnold, and G. B. Varile, "A Model for Preference," *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp. 134-139, April 1987.

[15] D.E.Rumelhart, G.E.Hinton, and R.J.Williams, "Learning Internal Representations by Error Propagation," *Parallel Distributed Processing*, Vol. 1, 1986.

[16] D.S.Touretzky, "Connectionism and PP Attachment," *Proceedings of the 1988 Connectionist Models Summer School*, 1988.

[17] Y. Wilks, "Right Attachment and Preference Semantics," *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp. 89-92, March 1985.

강 원 석

1985년 경북대학교 전자공학과 졸업(학사)
1988년 한국과학기술원 전산학과(공학석사)
1995년 한국과학기술원 전산학과(공학박사)
1994년~1995년 한국과학기술원 인공지능연구센터 위촉 연구원
1995년~현재 안동대학교 컴퓨터공학교육과 전임강사
관심분야 : 자연어처리, 기계번역, 한국어정보처리