

동사의 의미분석처리를 위한 SENKOV 시스템의 설계와 구현

문 유 진[†]

요 약

컴퓨터의 사용시 편안함과 친근감을 제공하기 위해서 사용자 인터페이스는 사용하기 쉽고 융통성이 있게 설계되어야 한다. 이를 위해 여러 가지 측면이 고려되어야 하는데 그 중 자연어 처리의 역할은 중요하다고 할 수 있다. 이 논문은 자연어 처리 중 동사의 의미분석 처리를 위하여 차별화 이론과 Levin 동사 클래스를 근간으로 하여 한국어 동사의 상위개념을 의미망 시스템(SENKOV)으로 설계 및 구현하였다. 이를 활용하여 부사와 동사 간의 공기 제약관계 설정에 유효한 개념 분류를 수행하였다. 일반적으로 많이 쓰이는 한국어 동사 658개를 대상으로 SENKOV 시스템을 구현한 결과, SENKOV는 44개의 최상위 노드를 가지고 있으며 깊이는 약 2.35이었다. 한국어 동사의 의미망은 명사의 의미망에 비하여 최상위 노드의 개수는 많고 깊이가 훨씬 더 얕았다. 그리고 성상부사의 선택적 제약에 SENKOV 시스템을 적용한 결과 공기계약 관계에 유효한 개념분류를 제공할 수 있음이 증명되었다. 한국어 동사의 의미분석 처리를 위한 실용적인 계층구조 시스템의 구축을 최초로 수행하였다는 점에 이 논문의 의의가 있다.

Design and Implementation of SENKOV System for the Semantic Processing of Korean Verbs

Yoo-Jin Moon[†]

ABSTRACT

Human-computer interface (HCI) should be designed with ease and flexibility for users for providing convenience and friendliness to users. The natural language processing is one of the aspects of HCI. This paper presents the method of design and implementation of SENKOV (Semantic Networks for Korean Verbs) System, which deals with its hierarchies of Korean verbs for the natural language processing. The system's architecture is based on the differential theory and Levin verb classes. This paper selects about 600 Korean verbs which are commonly used in the daily life, and implements the SENKOV System. The experiments show that SENKOV has 44 top nodes and depth of about 2.35. In addition, this paper applies the SENKOV System to co-occurrence constraint relationships among adverbs and verbs, and proves the validity of the system. This paper is important in that it has made the first trial to classify Korean verb concept for HCI.

1. 서 론

컴퓨터의 활용은 지식 정보화 사회를 살아가는 우리

에게 필수적인 요소이다. 컴퓨터의 활용시 편안함과 친근감을 제공하기 위해서 사용자 인터페이스(user interface)는 사용하기 쉽고 융통성이 있게 설계되어야 한다. 이를 위해 여러 가지 측면이 고려되어야 하는데 그 중 자연어 처리의 역할은 중요하다고 할 수 있다. 자연어 처리는 어휘분석단계, 구문분석단계, 그리고 의

※ 이 논문은 1999년도 한남대학교 학술연구 조성비 지원에 의하여 연구되었음

† 성 회 원 · 한남대학교 경영정보학과 교수

논문접수 : 1999년 4월 15일, 심사완료 : 2000년 3월 16일

미분석단계로 구성되어 있다. 이 중에서 어휘분석단계와 구문분석단계에 관한 시스템 개발은 상당한 진전이 이루어졌으나 의미분석단계에 관한 시스템 개발은 미진한 상태이다[1, 2].

한국어 문장의 의미분석단계에 관한 시스템을 개발하기 위해서는 한국어 동사의 의미분석이 핵심을 이룬다고 할 수 있다[3] 또한 한국어 동사의 의미분석은 추론의 핵심을 이루고, 한국어 담화, 텍스트 이해 및 정보소통을 위한 어휘정보 기술에 필요하며, 어휘부의 이질적 의미정보를 통합하는 데에도 필요하다[4] 한국어 동사의 의미분석을 수행하기 위해서는 문장에 쓰인 동사들을 문자 자체로만 인식해서는 안되고 동사의 개념에 의한 상관관계를 네트워크 형식으로 구축해주어야 한다[5-7].

이 논문에서는 동사의 개념인식을 위하여 동사의 상위개념(上位概念)을 의미망 시스템(SENKOV, Semantic Networks for Korean Verbs)으로 구축하고자 한다.

일반적으로 하위개념(下位概念)은 이행적(transitive)이고 비대칭적(asymmetrical)이며 계층적 의미구조를 생성한다. 이러한 계층적 의미구조를 상속 시스템이라 하며[8], 이 시스템은 하위개념이 상위개념들의 특성을 물려받고 상위개념에 없는 고유의 특성만 첨가하는 것이다 이 연구의 목적은 동사의 상위개념 정보를 계층적 상속 시스템인 Isa 계층구조로 구축하기 위하여 SENKOV 시스템을 설계하고 구현하는 것이다. 이를 위해 차별화 이론을 기반으로 하여 국어 사전, WordNet 동사 클래스 그리고 Levin 동사 클래스를 이용한다. 또한 부사와 동사 간의 공기 제약관계 설정에 유효하도록 개념 분류를 수행하는 데 SENKOV를 활용한다.

2. 관련 연구

현재까지 진행된 영어 동사 계층구조 시스템에는 WordNet 동사 클래스와 Levin 동사 클래스가 있다 그리고 유럽의 Euro WordNet 등이 있다.

2.1 영어의 WordNet

WordNet[8]은 약 120,000개의 영어 단어(명사, 동사, 형용사 그리고 부사)를 어휘개념으로 표현하여 의미망(semantic network)을 구성한 일종의 온라인 영어 데이터베이스이다 WordNet의 기본적 골격구조는 어휘개

념과 계층구조이다 어휘개념은 동의어 집합(synonym set, synset)으로 표현되고 있으며, 계층구조는 이 동의어 집합들 간의 상위위개념 관계로 표현되고 있다. WordNet은 이 동의어 집합을 사용하여 시소러스의 역할도 수행할 수 있도록 한다 인간의 어휘 지식을 모방하려는 시도에서 WordNet은 어휘형과 어휘개념 사이의 관계에 대한 정보를 차별화 이론을 기반으로 하여 표현하고 있다

WordNet 동사 클래스는 동사의 최상위 노드 분류가 체계적으로 되어 있지 않고 너무 많으며, 유사한 의미를 세분하여 다른 노드로 만든 경우도 있고, 자동사와 타동사를 구분하지 않았다는 단점이 있으니 어휘개념은 차별화 이론에 직합하여 이 연구에서 참조하였다

2.2 유럽의 WordNet

독일, 스페인, 프랑스 등 유럽에서도 영어의 WordNet을 기반으로 하여 각 나라별 의미망과 Euro WordNet에 관한 작업을 계속하고 있다. 명사의 의미망에 관한 작업은 활발히 진행 중에 있으나, 동사의 의미망에 관한 작업은 아직 미미한 상태이다.

2.3 Beth Levin의 동사 클래스

Levin 동사 클래스[9,10]는 의미 중심의 분류 구조로서, 통사적 관계를 고려하여 선택적 제약(selection restriction)의 유무 등을 감안한 것으로 3,000여 개의 동사를 49개의 클래스로 분류한 것이다 동사가 취할 수 있는 통사구조는 클래스 멤버십(class membership)을 결정한다. 기본적인 가정은 동사의 통사구조가 내재한 의미를 직접 반영한다는 것이다

Levin 동사 클래스는 영어의 동사를 중심으로 하였으므로 한국어 동사의 분류에는 맞지 않는 경우가 많이 있으나 전체적인 계층구조의 분류는 이 연구에서 참조하였다

3. SENKOV 시스템의 설계

3.1 SENKOV 시스템의 어휘개념과 계층구조

SENKOV 시스템의 설계에 있어서 중요한 관건은 기본 골격인 어휘개념을 표현하는 형식과 계층구조 관계 설정이다.

1) <http://www.ilc.pcnr.it/EAGLE96/rep2/>.

의미론에서 어휘화된 개념을 정의에 의하여 어떻게 표현할 수 있는가에 대한 이론은 개념을 '구성화(constructive)'하리는 것인지 혹은 '차별화(differential)'하리는 것인지에 달려 있다[8]. 구성화 이론에서는 어휘 표현이 어휘개념을 정확히 구성할 수 있도록 충분한 정보를 포함해야 한다는 전제조건이 있다. 그러나 이러한 전제조건은 충족되기 어렵고 대부분의 사전에 있는 정의(definition)에서도 충족되지 않는다. 한편, 차별화 이론에서는 어휘개념을 개념별로 구별될 수 있는 상징(symbol)에 의하여 표현할 수 있다고 본다.

SENKOV 시스템의 기본 골격인 어휘개념은 차별화 이론에 의하여 두 가지 방법으로 표현될 수 있다. 첫째, 한국어 동사에서 동의어 집합을 구성하여 한국어 동사 어휘개념을 표현하는 것이다. 둘째, 영어 동사 WordNet의 어휘개념을 이용하여 한국어 동사 어휘개념을 표현하는 것이다. 첫째 방법은 국어학자, 언어학자 및 심리학자들과 공동으로 막대한 작업을 하여야 하는 것이므로 이 연구에서는 채택하지 못하였다. 이 연구에서는 둘째 방법을 기본으로 하여 한국어 동사 계층구조에서의 어휘개념을 표현하고 영어 동사에 없는 어휘개념은 국어 사전, 시소러스 등을 참조하여 한국어 동사에 알맞은 어휘개념을 만들어 표현한다.

3.2 SENKOV 시스템의 추진전략

SENKOV 시스템의 한국어 동사 분류에는 국어 사전을 근간으로 하고, 계층구조 구축시 Levin 동사 클래스를 한국어 동사에 맞게 수정하여 참조하였으며, 어휘개념 구축시 국어사전과 WordNet 동사 클래스를 수정하여 참조하였다[2]. Levin 동사 클래스는 영어를 중심으로 하였으므로 한국어에는 맞지 않는 분류가 있으나 동사 클래스의 계층구조는 구성이 논리적이어서 참조하였다. 그리고 WordNet 동사 클래스는 동사의 최상위 노드 분류가 체계적으로 되어 있지 않고 너무 많으며, 유사한 의미를 세분하여 다른 노드로 만든 경우도 있고, 자동사와 타동사를 구분하지 않았다는 단점이 있으나 어휘개념은 차별화 이론에 적합하여 참조하였다.

SENKOV 시스템의 설계를 위한 추진전략은 다음과 같다

- ① 우리말 용언 중 동사의 원형을 중심으로 개념망

구축을 하였다. 후에 피생어, 피동사, 그리고 사동사 등으로 확장할 수 있는 기본틀을 구축하기 위함이다.

- ② 동사의 구문론적인 면과 의미론적인 면의 상관관계를 동시에 고려하였고, 구문론적인 면은 동사의 논항구조와 하위 범주화 유형을 중심으로 연구하였다. 왜냐하면 동사의 구문론적인 면을 배제하고 의미론적인 면만 고려하여 같은 동사 클래스에 넣을 경우 문장의 정확한 의미분석이 어렵기 때문이다.

- ③ 국어 사전의 정의를 근간으로 하여 한국어 동사의 의미와 상위개념을 추출하였다. 한국어 동사는 영어 동사와는 다른 의미와 특징이 많으므로 국어 사전을 근간으로 하였다.

- ④ 최상위 노드의 기준은 Levin 동사 클래스로 하였고, 국어 사전을 참고하여 한국어 동사 특성에 맞게 삽입, 삭제하여 보완하였다. WordNet 동사 클래스는 동사의 최상위 노드 분류가 체계적으로 되어 있지 않고 너무 많기 때문에 참조하지 않았다.

- ⑤ 중간 및 터미널노드의 계층구조는 국어 사전의 정의와 Levin 동사 클래스를 참조하였다. Levin 동사 클래스는 영어를 중심으로 하였으므로 한국어에는 맞지 않는 분류가 있으나 동사 클래스의 계층구조는 구성이 논리적이어서 참조하였다.

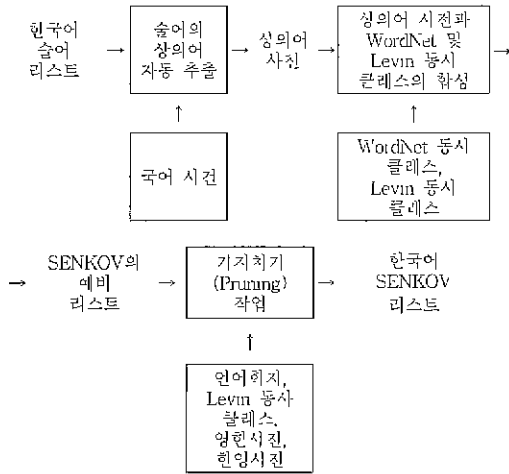
- ⑥ WordNet 동사 클래스는 자동사와 타동사를 구분하지 않았다는 단점이 있으므로, 한국어 동사 클래스 분류에는 자동사와 타동사를 구분하였다.

- ⑦ 차별화 이론에 근거하여 어휘개념을 동의어 집합으로 표현하였다. 즉, 국어 사전, 시소러스, 그리고 WordNet 동사 클래스의 어휘개념을 참조하여 한국어 동사 클래스의 어휘개념을 동의어 집합으로 표현하였다. WordNet 동사 클래스는 유사한 의미를 세분하여 다른 노드로 만든 경우도 있다는 단점이 있으나 어휘개념은 차별화 이론에 적합하여 참조하였다.

3.3 SENKOV 시스템의 설계 및 가지치기 작업

SENKOV 시스템은 생성시스템과 DB구축시스템으로 구성되어 있다.

생성시스템은 3.2에서 기술한 추진전략에 의하여 (그림 1)과 같이 설계되었다. (그림 1)에서 도시하듯이 먼저 한국어 술어를 선정하여 한국어 술어 리스트를 작성한다. 이 선정된 술어를 중심으로 국어 사전을 참고하여 술어의 정의를 자동으로 분석하므로써[6, 11] 술



(그림 1) SENKOV의 생성시스템

어의 상의어(hypernym, 上意名) 자동 추출 작업을 수행하고 상의어 사전을 출력한다. 상의어는 계층구조의 상위개념을 표현하기 위한 밑바탕이 된다. 이를 기반으로 하여 SENKOV의 터미널노드와 그 위 레벨을 개발한다. 이 상의어 사전과 WordNet 동사 클래스 및 Levin 동사 클래스를 합성하여 SENKOV의 개발된 레벨을 수정 확장하는 작업을 수행하고, SENKOV의 예비 리스트를 출력한다. 언어학자와 전산학자는 한영사전, 영한사전 그리고 상식을 바탕으로 리스트의 가지치기(pruning) 작업을 수행하여 SENKOV 리스트를 최종적으로 생성한다.

생성시스템에서의 가지치기 작업은 한국어 동사 ISA 계층구조에 근거하여 수행된다. 이를테면, Levin 동사 클래스와 WordNet 동사 클래스의 분류가 한국어의 ISA 계층구조에 적합하지 않은 경우에 삽입, 삭제, 이동 등의 작업이 수행된다. 이 생성시스템에서 수행한 가지치기 작업은 35 종류인데, 그 중 일부를 예시하면 다음과 같다.

① 영어 WordNet 동사 클래스는 자/타동사를 구별하지 않고 한 동사 내에 sense만 달리하여 기술하였으나 한국어는 한 동사를 자/타동사로 집하여 쓰는 경우가 드물기 때문에 각 어휘 엔트리(lexical entry)를 달리하였다. (계층구조 삽입)

② Levin 동사 클래스 263에 속한 preparing 동사는 실제 한국어 슬어와 상응하는 의미가 있으므로 가지치기

작업을 수행하였다. (계층구조 삭제)

③ Levin 동사 클래스에서 9.1과 9.2는 한국어의 경우에 의미상 거의 동일하여 따로 분류하지 않고 같은 클래스에 넣었다

④ Spray 동사의 경우 Levin 동사 클래스에서는 putting 동사 클래스에 속해있으나 WordNet 동사 클래스에서는 putting 동사 분류에 속하지 않았다. Levin 동사 클래스의 분류가 옳다고 판단하여 최상위 개념에 putting을 넣어주었다 (계층구조 이동)

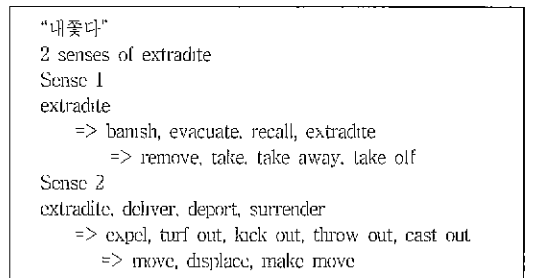
⑤ 선택한 동사에 관한 Levin 동사 클래스의 계층구조와 WordNet 동사 클래스 계층구조를 비교하여, Levin 동사 클래스의 계층구조를 그 동사의 Sense 1로 기술하고 나머지 WordNet sense는 Sense 2, ... 등으로 SENKOV 계층구조를 구성하였다.

⑥ 한국어 슬어의 의미 중 회귀하게 쓰이는 의미는 그 슬어 자체의 독창적 의미를 약화시키므로 제거하였다

DB구축시스템은 SENKOV의 생성시스템에 의해 출력된 리스트를 정렬하여 데이터베이스로 구축한다. DB구축시스템은 먼저 SENKOV 리스트의 파일을 모아서 정렬을 하고 그 다음 데이터베이스인 트리 파일과 인덱스 파일을 생성하는 것이다[2].

4. SENKOV 시스템의 구축 및 결과

3.3절에서 전술한 설계방법에 의하여 썬 워크스테이션을 사용하여 C와 C++언어로 SENKOV 시스템을 (그림 2)와 같이 구축하였다. (그림 2)는 “내쫓다”의 계층구조를 나타낸 것으로 Sense 1과 Sense 2의 두 가지 의미를 표현하고 있다



(그림 2) SENKOV 시스템의 구축 예

일반적으로 많이 쓰이는 한국어 동사 658개를 대상

으로 한 결과, SENKOV는 44개의 최상위 노드를 가지고 있으며 계층구조의 평균 깊이는 약 2.35이었다. 한국어 동사의 의미망은 한국어 명사의 의미망에 비하여 최상위 노드의 개수는 많고 깊이가 훨씬 더 얕았다. 구축된 SENKOV 시스템과 [2]에서 구축한 한국어 명사 계층구조 시스템에 근거하여 <표 1>은 한국어 명사와 동사의 최상위 노드 수와 계층구조의 평균 깊이를 비교하였다. 한국어 동사는 명사에 비하여 계층구조를 형성하기 보다는 유사한 의미를 갖는 동사그룹을 더 많이 형성한다는 것을 보여준다.

<표 1> 품사별 계층구조의 평균 깊이

한국어 품사	최상위 노드의 수	계층구조의 평균 깊이
명사	11	5.40
동사	44	2.35

구축된 SENKOV 시스템은 기계 번역 시스템(word-sense ambiguity, phrase attachment problem), 기계 이해 시스템, 정보 검색 시스템, 문장 검사/교정 시스템, 자연어 인터페이스 시스템 그리고 질의어 처리 등에 활용될 수 있다.

5. 부사의 공기제약 관계에 유효한 개념 분류

한국어 부사와 동사간의 공기제약(co-occurrence restriction) 관계 설정에 유효한 개념 분류를 하는 데 SENKOV를 활용하였다.

한국어 부사는 크게 특정한 성분을 수식하는 성분 부사와 문장 전체를 수식하는 문장부사로 나뉘어진다 [3] 성분부사는 동사, 형용사 및 명사를 수식하는 성상 부사와 의성어, 의태어를 표현하는 상징부사 그리고 지시부사를 포함한다 문장부사는 양태부사와 접속부사를 포함하는데, 이 부사는 문장과 관계가 있고 동사의 공기제약과는 관계가 없으므로 이 연구에서는 논외로 한다. 그리고 성분부사 중 지시부사도 동사의 공기제약과는 관계가 없으므로 이 연구에서는 논외로 한다.

성상부사에는 “꽤”, “세게”, “잘게”, “잘”, “높이”, “빨리” 등이 속하며, 이 부사와 동사간의 공기제약에 유효한 개념분류를 하는데 SENKOV를 다음 예와 같이 활용할 수 있다. 다음에서 나오는 정수는 SENKOV의 동사 클래스를 표시하며, 소수는 동사 하위클래스를 표

시한다. 부사 “꽤”은 동사 클래스 12, 9.5, 9.7 등과 공기제약 관계를 가질 수 있다는 것을 보여주고 있다.

- 예) 꽤 + (12, 9.5, 9.7, 10.1, 10.2, 10.7, 11.2, 17.1, 18.1, 18.4, 21, 26.6, 43.4, 45, 51.1, 51.3, 51.4, 57 중 “빨다”)
- 세게 + (18.1, 9.3 ~ 9.7, 10.4, 10.7, 11.2, 12, 15.1, 17, 18, 21, 22, 38, 40.1 ~ 40.3, 45, 51.3, 57)
- 잘게 + (21, 10.7 중 “넋기다”)

상징부사에는 “땡땡”, “도란도란”, “까옥까옥”, “테굴테굴” 등이 속하며, 이 부사는 대부분 한 두 개의 동사와 공기제약 관계를 형성하므로 SENKOV의 활용이 큰 도움을 주지는 않는다. 다음 예는 상징부사와 동사간의 공기제약 관계를 나타낸 것으로 SENKOV의 동사 클래스가 큰 도움을 주지 않는다는 것을 보여준다

- 예) 땡땡 + (18.1 중 “치다”)
- 도란도란 + (37.1 중 “이야기하다”, 37.2, 37.3 중 “속삭이다”)

위와 같이 SENKOV 시스템을 부사와의 공기제약 관계에 활용하여 본 결과, 성상부사와 동사간의 공기제약 관계에 유효한 개념분류를 제공할 수 있음을 보여주었다.

6. 결 론

컴퓨터의 사용자 편안함과 친근감을 제공하기 위해서 사용자 인터페이스(user interface)는 사용하기 쉽고 용통성이 있게 설계되어야 한다. 이를 위해 여러 가지 측면이 고려되어야 하는데 그 중 자연어 처리의 역할은 중요하다고 할 수 있다. 이 논문은 자연어 처리 중 동사의 의미분석 처리를 위하여 한국어 동사의 상위개념을 의미망 시스템(SENKOV)으로 설계 및 구현하였으며, 이를 활용하여 부사와 동사간의 공기제약관계 설정에 유효한 개념 분류를 수행하였다. 이 연구는 한국어 동사의 의미분석 처리를 위한 실용적 계층구조 시스템의 구축을 최초로 수행하였다는 점에서 의의가 있다.

SENKOV 시스템의 한국어 동사 분류에는 국어 사전을 근간으로 하고, Levin 동사 클래스를 한국어 동사에 맞게 수정하여 계층구조 구축시 참조하였으며, 계층구조의 어휘개념 구축시 WordNet 동사 클래스를

참조하였다 SUN 워크스테이션을 사용하여 C와 C++언어로 일반적으로 많이 쓰이는 한국어 동사 658개를 대상으로 하여 SENKOV 시스템을 구축하였다. SENKOV 시스템은 44개의 최상위 노드를 가지고 있으며 깊이는 약 2.35이었다. 한국어 동사의 의미망은 한국어 명사의 의미망에 비하여 최상위 노드의 개수는 많고 깊이가 훨씬 더 얕았다.

향후 연구계획은 다음과 같다. 첫째, 3,000여 개의 한국어 동사에 대한 SENKOV를 구축하여 완성하는 것이다. 둘째, 주어와 동사, 목적어와 동사간의 선택적 제약에 유효한 SENKOV를 구축하는 것이다. 셋째는 위에서 제시한 활용분야에 효율적으로 활용하는 것저자이다.

참 고 문 헌

[1] 김영택, 자연언어처리, 교학사, 1994.

[2] 문유진, 의미론적 어휘개념에 기반한 한국어 명사 WordNet의 설계와 구축, 서울대학교 대학원 컴퓨터공학과 박사학위 논문, 1996.

[3] 남기십, 고영근, 표준 국어 문법론, pp.169-174, 탑출판사, 1985.

[4] 이정민, 강범보, 남승호, "한국어 술어의 의미구조 연구". 제2회 소프트과학워크숍 학술회의, 1997

[5] C. Fellbaum, "English Verbs as a Semantic Net," *Report of WordNet*, 1993

[6] P. Hernert, "KASSYS: A Definition Acquisition System in Natural Language," *Proc. of COLING-94*, pp.263-267, Aug 1994

[7] 문유진, 김영택, "한영기계번역에서 개념기반의

동사 번역". 한국정보과학회 논문지, 제22권 제8호, pp.1166-1173, 1995.

[8] G. A. Miller, R. Bockwith, C. Fellbaum, D. Gross and K. Miller, "Introduction to WordNet: An Online Lexical Database," in *Five Papers on WordNet*, CSL report, Cognitive Science Laboratory, Princeton University, 1993

[9] B. Levin, 'English Verb Classes and Alterations: A Preliminary Investigation,' The MIT Press, 1997.

[10] B. Levin and M. Hovav, 'Unaccusativity: At the Syntax-Lexical Semantics Interface,' The MIT Press, 1996

[11] S. Montemagni and L. Vanderwende, "Structural Patterns vs. String Patterns for Extracting Semantic Information from Dictionaries," *Proc of COLING-92*, pp.546-552, Aug. 1992.



문 유 진

e-mail : yjmoon@eve.hannam.ac.kr

1979년 한국외국어대학교 졸업

1986년 펜실베니아 주립대학교

전산학과 졸업(이학석사)

1996년 서울대학교 컴퓨터공학과

졸업(공학박사)

1996년~1997년 펜실베니아대학교 인지과학연구소

Post-doc.

1990년~1998년 호남대학교 컴퓨터공학과 부교수

1999년~현재 한남대학교 경영정보학과 부교수

관심분야: 자연언어처리, 인공지능, 전자상거래