IT 인물 관련 텍스트 정보의 효율적인 검색을 위한 Sub-language의 속성 연구

고승희*, 김소연*, 천승미*, 남지순*, 김권양**, 박세영***, Ivan Berlocher****
*DICORA, 한국외국어대학교, **경일대학교, ***경북대학교, ****(주)솔트룩스
kohsh@hufs.ac.kr, claire831@hanmail.net, smcheon@hufs.ac.kr, namjs@hufs.ac.kr,
seyoung@knu.ac.kr, kykim@kiu.ac.kr, ivan_berlocher@hotmail.com

Studies on the linguistic properties of the IT-People documents for an efficient Information Retrieval

Seung-Hui Koh*, So-Yeon Kim*, Seung-Mi Cheon*, Jee-Sun Nam*
Kweon-Yang Kim, Se-Young Park***, Ivan Berlocher****
*Hankuk University of Foreign Studies, **Kyungil University, ***Kyungpook University, ****Saltlux Inc.

요 약

본 연구는 IT 인물 관련 텍스트 정보의 효율적인 검색을 위하여 문서 내에서 인물과 관련된 정보를 담고 있는 문장들이 어떠한 특징을 가지고 실현되는가를 살펴보고 언어적 속성을 어떻게 구조화하고 형식화할 것인가를 논의하는 것을 목적으로 한다. 언어적 속성 분석을 위해서 전자신문 내에서 인물 관련 코퍼스를 수집하고 이들의 분석을 통해 다음과 같이 문제가 되는 특징들을 확인하였다. 즉 외래어 음차 표기 문제, 복합 명사 및 명사구 그리고 서술 명사적 표현의 문제 등으로 요약된다. IT라는 특정 영역에 대해 텍스트 내에서의 어휘-통사적 패턴을 분석하고 언어적 특징에 대한 효율적 기술을 위해서는 LGG 부분 문법 그래프 모델을 활용하도록 한다. 본 연구는 특정 영역인 IT 관련 문서에서 자연언어 텍스트를 대상으로 정보 검색할 때 문제가 되는 다양한 언어학적 현상들을 다루며, 향후보다 확장된 영역에서의 효율적 언어 처리에 대한 방법론적 대안을 제시할 수 있을 것으로 기대된다.

1. 서 론

본 연구는 IT 인물[1] 관련 텍스트 정보의 효율적인 검색을 위하여 IT 문서에서 인물과 관련된 정보를 담고 있는 문장들이 어떠한 특징을 가지고 실현되는가를 밝혀내고 이러한 언어적 속성을 어떻게 구조화하고 형식화할 것인가를 논의하는 것을 목적으로 한다.

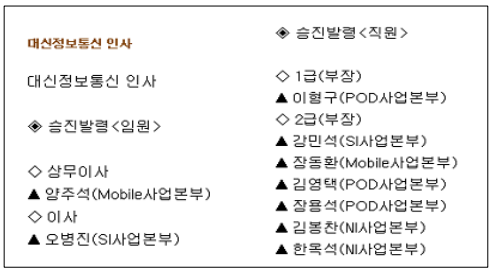
사용자가 요구하는 정보의 키워드(key-word)는 실제 자연언어 텍스트(full-text) 문서에서는 단순 명사의 형태뿐 아니라 복합 명사, 또는 통사적 명사구의 형태로 실현될 수 있다. 또는 일정 서술어를 구성하는 서술명사의 형태로 실현될 수도 있으며, 표기상으로도 여러 철자법으로 실현될 수도 있다. 예를 들어 외래어가 차용된 경우 동일한 대상을 지시함에도 불구하고 한국어로 전사 표기되면서 발생될 수 있는 음절 표기 방식의 다양성으로 인해 그 검색 결과는 매우 상이하게 나타날 수 있다. 복합 명사의 경우도 띄어쓰기 문제나 복합명사 정규화문제(normalization problem of compound noun)[2] 등으로 인해 규칙 기반의 기존 처리 방식으로는 만족스러운 결과를 기대하기 어려운 경우가 많다.

현대와 같이 기하급수적으로 증가하고 있는 디지털 문서의 양을 고려할 때 사용자의 질의나 검색 요청에 보다 정확한 결과를 제시해 주기 위해서는 검색 결과가 되는 문서에 나타난 구문들을 올바르게 분석하는 기술이 절실하게 요구된다. 본 연구에서는 일반적인 성격의 텍스트의 범위를 좁혀서 IT분야의 인물과 관련된 코퍼스에 나타나는 문장의 핵심적인 언어적 속성들을 분석하여 이에 대한 정보를 검색하고자 할 때 어떠한 언어적 특징을 고려해야 할 것인지 그리고 어떠한 체계적인 접근이 이루어져야 하는지에 대해 논의할 것이다. 이러한 언어적 특징에 대한 효율적인 기술 방법론으로 프랑스 파리 7대학과 마른느-라-발레 대학에서 제안한 LGG(Local-Grammar Graph)[3]

1) 여기서 인물은 <사람>과 함께 주체가 될 수 있는 <회사>나 <조직>을 포함한다.

2) 여기서 논의하는 복합 명사 정규화 문제는 가령 아래 ㄱ)에서 복합 명사 ㄴ)을 색인어로 추출하는 문제를 의미한다(강병주, 최기선, 윤준태, 1998).
ㄱ) "정보를 검색하다"
ㄴ) = '정보 검색'

부분문법 그래프 문법 모델을 활용하도록 한다. 본 연구를 위해 수집된 코퍼스는 전자신문 2006년 11월부터 2007년 3월까지의 기사 중 자연어 문장의 형태로 기술된 인사 및 개인 경력 관련 기사들이다. 문서를 수집함에 있어 아래 <그림 1>과 같이 <직위-이름-부서>의 일정 패턴이 반복되어 정보를 제공하고 있는 정형성을 지닌 문서는 연구 대상에서 제외하고 <그림 2>와 같이 자연어로 기술된 문장만을 연구 대상으로 한다.


대신정보통신 인사

대신정보통신 인사

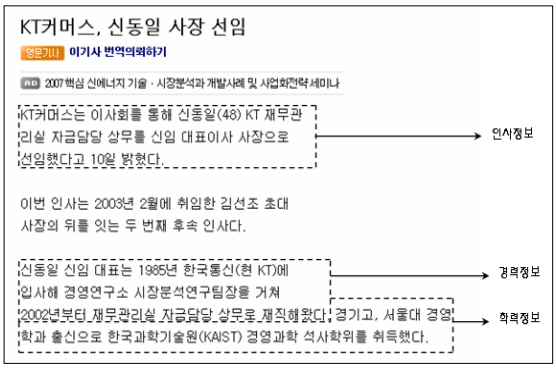
◈ 승진발령<임원>

◇ 상무이사
▲ 양주석(Mobile사업본부)
◇ 이사
▲ 오병진(SI사업본부)

◈ 승진발령<직원>

◇ 1급(부장)
▲ 이형구(POD사업본부)
◇ 2급(부장)
▲ 강민석(SI사업본부)
▲ 장동환(Mobile사업본부)
▲ 김영택(POD사업본부)
▲ 장용석(POD사업본부)
▲ 김봉찬(NI사업본부)
▲ 한목석(NI사업본부)

<그림1> '대신정보통신 인사' 기사
(전자신문 2007.03.23)


KT커머스, 신동일 사장 선임

영문기사 이기사 번역의뢰하기

RTD 2007 핵심 신에너지 기술 · 시장분석과 개발사례 및 사업화전략 세미나

KT커머스는 이사회를 통해 신동일(48) KT 재무관리실 자금담당 상무를 신임 대표이사 사장으로 선임했다고 10일 밝혔다. → 인사정보

이번 인사는 2003년 2월에 취임한 김선조 초대 사장의 뒤를 잇는 두 번째 후속 인사다.

신동일 신임 대표는 1985년 한국통신(현 KT)에 입사해 경영연구소 시장분석연구팀장을 거쳐 2002년부터 재무관리실 자금담당 상무로 재직해왔다. → 경력정보

경기고, 서울대 경영학과 출신으로 한국과학기술원(KAIST) 경영과학 석사학위를 취득했다. → 학력정보

<그림2> 'KT커머스인사관련' 기사
(전자신문 2006.12.11)

2. IT 인물 관련 텍스트의 언어적 특징

2.1. 외래어 전사 표기와 관련된 특징

오늘날 우리는 자국어, 외래어, 외국어, 그리고 국적을 알 수 없는 언어까지, 언어 사용에 있어서 적지 않은 혼란을 느끼며 살고 있다. 이들 외래어 또는 외국어 중에는 국가 간의 학문·문화적 교류나 정치적 접촉 등의 여러 가지 요인으로 최근 언어 사용에 있어 필수불가결인 요소로 자리잡은 형태들이 빈번하며 특히 IT 관련 문서들의 경우 그 출현 빈도는 생각보다 훨씬 중요한 비중을 차지하고 있는 것을 발견할 수 있다.

외국어가 자국에 들어와 외래어로 사용이 되는 경우나 전문용어로써 사용되는 경우, 또는 국내외 기업명, 인물명, 제품명과 같은 고유명사 부류, 그 외 다른 이유에 의하여 외국어가 자국어와 섞이어 사용되는 경우에 이를 자국어에 있는 단어로 번역할 수 없는 경우는 그 발음을 따라 한국어로 음차 표기(transliteration)를 하게 된다. 이때 실제로 발화되는 영어 발음 자체가 일관성이 없을 뿐더러, 한국식 발음에 의한 한국식 영어에 의존하여 음차를 하는 경우도 있고, 영어에 익숙한 사람은 그렇지 않은 사람보다 원어의 발음에 가깝게 음차 표기를 하려는 경향을 보이기도 한다. 이러한 문제가 발생하는 가장 중요한 이유는 외국어와 한국어 사이의 음운 구조 상의 차이로 외국어의 발음에 대응하는 정확한 한국어 글자가 존재하지 않기 때문이다. 또 다른 원인으로 일반인의 표준 외래어표기법에 대한 이해의 부족을 들 수 있다. 이러한 혼란을 없애기 위해서 국립국어연구원에서 표준 외래어 표기법을 만들어 각 언어마다의 표기 세칙을 둔 상황이지만, 표기법 자체가 충분히 자세하지 않아서 모든 외래어를 적는 데는 여전히 어려움이 있고, 보통 사람이 이 표기법을 정확히 이해하고 실천하다는 것이 그리 용이하지도 않다. 또한 표기법 자체에도 예외 사항이 많아 다양한 외래어 표기를 만들어 내는 원인이 되고 있다. 따라서 이러한 다양한 외래어 표기법의 존재는 한국어 정보처리에 큰 걸림돌이 될 수밖에 없는 상황이다.

최근 웹페이지와 같은 온라인 문서에서, 정보통신 기술의 급속한 발전으로 외국과의 교류가 증대됨에 따라 영어 및 외래어의 사용이 급격히 증가하고 있고 특히 과학 전문분야 문서의 경우에는 이러한 정도가 훨씬 심한 것을 발견할 수 있다. 전문분야 문서의 경우 대부분의 전문용어들이 영어에서 도입된 것들이라 적절한 한국어 번역어를 발견할 수 없는 경우 해당 영어 표현을 한국어로 음차 표기해서 사용하거나 또는 정확한 의미 전달을 위해서 원어 그대로를 사용하기도 한다. 또한 국외의 기업명, 인물명, 제품명과 같은 영어 고유명사의 경우도 한국어로 음차 표기되어 사용되기도 하고 또는 원어대로 인용되기도 한다. 따

3) LGG 문법은 "부분적인" 언어 현상을 다루기 때문에 복합명사의 분석 및 처리, 부분적으로 굳어진 관용어구를 포함하는 문장이나 연어 관계를 갖는 구, 의미적으로 동의 관계(synonymy)를 갖는 일정 어구들, 또는 특정 분야에 관련된 언어 표현들에 대한 기술에 유용하게 쓰일 수 있다. 부분 문법으로 구현된 언어 정보들은 텍스트에 대한 자동 처리 시 분석이나 생성을 위한 문법적 정보로 사용될 수 있으며, 분석시 발생하는 중의성(ambiguity)을 제어하는 데에도 효율적으로 기능할 수 있다. 부분 문법은 비순환 그래프(Directed Acyclic Graph, DAG)의 형태로 실현될 수 있으며, 이는 텍스트 처리시 유한 트랜스듀서(Finite-State Transducer)나 유한 오토마타(FSA)의 형태로 바로 전환되어 시스템에서 작동할 수 있다(남지순, 2002: 432~433).

라서 한국어 문서에는 같은 개념을 지칭하기 위하여 영어와 한국어 음차표기(외래어), 그리고 경우에 따라 한국어 번역 용어 등이 혼재되어 사용되고 있어 이러한 비일관성을 가중시키고 있다. 결국 같은 문서 내에서 또는 같은 데이터베이스 내에서 사용되는 영어 및 다양한 외래어 표기의 혼용은 정보검색에서 심각한 단어불일치문제(word mismatch problem)를 야기하며 정보 검색의 성능을 크게 저하시키는 역할을 하게 된다.

실제 한국어 문서에서의 외래어 사용실태를 조사해 본 결과, 상품명과 사람명과 같은 명사에서 음차표기한 외래어의 사용이 현저하게 나타나며, 특히 IT 계열의 기사는 전문용어와 국내외 기업명, 인물명 및 제품명 등의 사용 도수가 잦아 외래어의 사용 빈도수가 일반 기사에 비해 현저히 높은 비율을 나타내고 있는 것을 발견할 수 있다. 본 논문에서 연구의 대상이 된 코퍼스의 일부를 추출하여 외래어의 출현 비율을 살펴본 결과, 예를 들어 1,220개의 단어에 대하여 313개의 외래어가 사용되어 전체의 약 25.6% 이상의 비중을 차지하고 있는 것을 관찰할 수 있었다. 이와 대응되는 크기의 일반 기사 코퍼스의 경우, 가령 1,430개의 단어로 구성된 텍스트에 나타난 외래어는 221개로, 외래어 사용 비중이 15% 정도로 IT 분야 기사보다 그 출현 빈도가 상대적으로 낮은 것을 확인할 수 있다.

본 연구의 대상이 된 IT 분야 인물 관련 정보를 담고 있는 텍스트에서 사용된 외래어를 살펴보면, 이 중 94% 이상이 전문용어와 고유명사로 구성되어 있는 것을 발견할 수 있다.

다음 <표 1>은 고유명사와 전문용어의 몇 가지 예를 보인다.

고유명사	전문용어
모토로라	모바일
비즈니스위크	네트웍
에드 젠더	홈네트워크
케시 레스잭	디티비로
봅 웨이먼	IPTV
칼리 피오리나	셋톱박스

<표 1> IT 인물 관련 기사에서 실현된 외래어의 예

여기서 외래어와 관련하여 심각한 문제는, 동일한 문서에서 하나의 외래어가 여러 유형의 음차 표기 형태로 혼용되어 사용되는 경우가 빈번하다는 점이다. 아래 <표 2>는 동일 원어에 대한 여러 유형의 음차 표기가 발견되는 예를 보인다.

원어	음차표기	%
digital	디지털	0.1%
	디지탈	48%
	디지틀	50%
	디디틀	0.1%
data	데이타	97%
	데이터	3%

<표 2> 외래어의 다양한 음차 표기의 예

이때 국내에 들어와 자리를 잡은 국외기업명이나 제품명, 인물명의 경우는 한 가지 음차 표기로 굳어져 사용되는 경향을 보이는 반면, 그 외 전문용어나 국내에 처음 소개된 기업이나 인물, 제품명과 같은 고유명사의 경우는 한 단어에 대해 몇 가지의 음차 표기가 동시에 사용되면서 통일성을 갖추지 못한 경우가 많다. 다음 <표 3>은 외래어 고유명사 부류 중에서 거의 표기상의 변이형이 나타나지 않는 유형들의 예를 보인다.

국외기업명	제품명	인물명
애플컴퓨터	소니바이오	빌 게이츠
노키아	모토로라	워렌버핏
아이비엠	마이크로소프트	세르게이 브린
델컴퓨터	윈도우	래리 페이지

<표 3> 표기 변이형이 발견되지 않는 외래어 고유명사의 예

반면 아래 <표 4>에서 보이는 외래어 고유명사들의 경우는 표기상의 변이형이 상대적으로 다양하게 분포되는 예를 보인다.

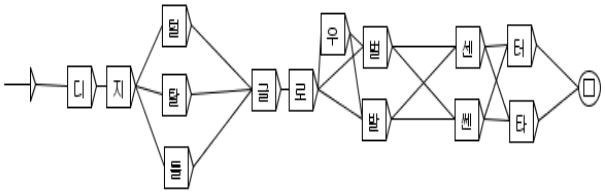
분류	원어	음차표기
국외기업명	lindenlab	린든랩
		린덴랩
인물명	Ed Zander	에드잰더
		에드젠더
		애드잰더
		애드젠더
제품명	Razr	레이저폰
		레이져폰
	Windows Vista	윈도우비스타
		윈도비스타
		윈도우즈비스타
		윈도비스타

<표 4> 표기 변이형이 발견되는 외래어 고유명사의 예

위의 <표 4>에서 나타난 것처럼 사용되고 있는 외래어가 고유명사 부류이거나 전문용어 신조어인 경우에는 거의 사전에 등재되어 있지 않으며, 또한 사전에 등재되어 있는 외래어의 경우라 할지라도 음차 표기가 다양한 경우는 이러한 형태들을 모두 검색하는 것이 불가능하므로 실제로 검색 엔진의 효율성을 크게

저하시키는 원인이 된다.

이런 관점에서 볼 때 이러한 어휘 부류를 키워드로 하는 문서들에 대한 정보 검색 시스템의 성능을 향상시키기 위해서는, 외래어 및 외국어의 고유명사와 전문용어 부류에 대해서 원어 정보와 더불어 여러 가지 가능한 음차 표기 정보를 함께 제시한 데이터베이스의 제공이 요구된다[4]. 이를 위하여 개별 외래어의 음차 표기의 다양성을 어휘적으로 기술하는 LGG 부분문법 그래프를 활용하는 것이 가능하다. 다음 <그림 3>은 "Digital Global Center"라는 원어에 대한 다양한 음차 표기가 가능한 경로를 보이는 유한 오토마타 LGG 그래프의 예를 보인다.



<그림 3> LGG 그래프를 이용한 외래어 음차 표기 기술 문법

위의 LGG 그래프는 하나의 외래어가 실제 문서에서 실현될 수 있는 최대 48가지의 음차 변이형을 기술하고 있다.

이와 같은 방식으로 LGG 문법을 구현하여 이를 바탕으로 IT 인물 관련 텍스트에서 출현 빈도가 높은 외래어 유형들에 대한 정보 검색을 수행한다면 실제로 IT 분야에 대한 검색의 질을 높일 뿐 아니라, 일반적인 유형의 텍스트에서도 사용될 수 있는 좀 더 확장되고 정교한 데이터베이스의 구축이 가능할 것이며, 이를 통한 보다 유의미한 정보 추출 작업이 가능해질 것이다.

2.2. 명사 연결 구성의 복합 명사구 특징

IT 인물 관련 텍스트에서 실현되는 정보의 키워드를 살펴보면, 이들 중의 많은 형태는 위에서 논의한 바와 같은 외래어 계열의 음차 표기 어휘들이면서 동시에 두 개 이상의 어휘가 결합되거나 일반 명사와 고유어 명사가 결합하여 복합 명사구의 형태를 구성하는 어휘들이 빈번하게 나타남을 관찰할 수 있다.

예를 들어 IT 인물과 관련된 정보는 대부분 고유명사로 실현된 <인명>을 나타내는 어휘와 함께 <소속>이나 <직위>, <학력>이나 <경력>을 나타내는 어휘들의 복합적인 나열형으로 나타난다. 문법 범주적 관점에서 이들은 대부분 명사들의 나열 형태로 이루어지는데 이러한 나열은 실제로 띄어쓰기의 문제나 동일 지시 대상에 대한 생략, 이동 등의 특징, 그리고 특정 문법적 표지의 삽입 등의 현상에 의해서 동일 정보가 다양한 형태로 실현되는 것을 볼 수 있다.

[1] 명사 나열형

명사 나열형은 여러 개의 명사(N)가 언어학적으로 특정 표지의 삽입 없이 선조적으로 나열되는 형태를 일컫는다. IT 인물 관련 명사구 키워드에서 특징적으로 발견되는 형태로 이들은 경우에 따라 띄어쓰기가 일정치 않아 구성 명사(N)들의 인식과 처리에 어려움이 많이 발생한다. 예를 들면 다음과 같다.

KT커머스는 이사회를 통해 신동일(48) KT 재무관리실 자금담당 상무를 신임 대표이사 사장으로 선임했다.

[2] 생략, 이동 등에 의한 변형 명사구

생략, 이동 등에 의한 변형 명사구는 앞서 살핀 명사 나열형이 다양한 방식으로 변형이 이루어지는 형태를 말한다. 실제로 동일 텍스트 내에서는 위와 같은 명사구가 반복되어 나타날 때에는 동일한 형태로 반복되는 경우를 찾기가 힘들다. 아래와 같은 방식으로 축약이나 생략이 일어난 간결한 형태로 변형되어 텍스트 내에서 실현되는 것이 일반적인 현상이다. 이처럼 형태적으로 전혀 다른 대상이 동일한 의미를 지니고 동일한 대상을 지시함을 기계는 인식하기가 어렵다. 그러므로 이들에 대한 처리가 가능하도록 구조화해 주어야 할 필요성이 있다.

김중태 신임 대표이사
= 김 신임 대표
= 김 대표이사
= 김 대표 등등..

[3] 특정 문법적 표지에 의해 연결된 명사구 형태

IT분야 인물 관련 정보는 아래와 같은 특정 문법적 표지에 의해 연결된 명사구 형태로 실현되는 경우들이 빈번하게 발견된다. 주로 특정 인물의 직위나 직책을 나타낼 때 이러한 언어적 표현이 사용됨을 확인할 수 있다. 이들 문법적 표지들이란 "겸"이나 "및"과 같은 의존명사 유형으로서, 예를 들면 다음과 같다.

회장 겸 최고경영자
사장 겸 최고운영책임자
국내 및 해외 영업총괄 부서 담당

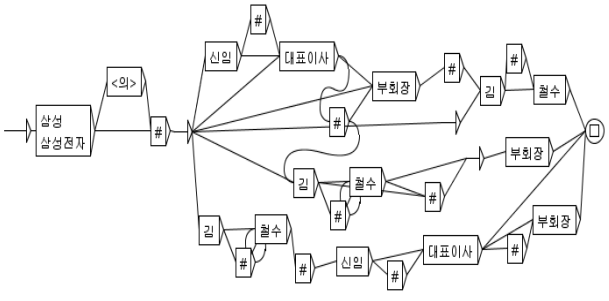
4) 이러한 문제점에 대한 다른 관점에서의 해결 방안으로 이재성(1999)에서는 다양한 방식의 외래어 음차표기 형태에 대하여 원어정보를 자동으로 매핑(mapping)시키는 방식이 제안된 바 있다.

위와 같은 인명 관련 명사구 표현들은 통사적 변형 및 위치 이동이 상대적으로 자유롭게 일어나기 때문에 동일 정보가 실제로 아래와 같이 다양한 형태로 실현 가능한 것을 수 있다(남지순, 2002: 431~433).

예를 들어 <삼성전자의 신임 대표이사 부회장 김철수>는 가령 다음과 같은 여러 유형의 명사구 패턴으로 실현 가능하다.

삼성전자<의>#신임#대표이사#김#철수부회장
삼성전자<의>#신임#대표이사#김#철수#부회장
삼성전자<의>#신임#대표이사#김철수부회장
삼성전자<의>#신임#대표이사#김철수#부회장
삼성전자<의>#신임#대표이사부회장#김#철수
삼성전자<의>#신임#대표이사부회장#김철수
삼성전자#김#부회장
삼성전자#김#철수
삼성전자#김#철수부회장
삼성전자#김#철수#부회장
삼성전자#김철수
삼성전자#김철수부회장
삼성전자#김철수#부회장
삼성전자#대표이사#부회장#김#철수
삼성전자#대표이사#부회장#김철수
삼성전자#대표이사#김#부회장
삼성전자#대표이사#김#철수부회장
삼성전자#대표이사#김#철수#부회장
삼성전자#대표이사#김철수부회장
삼성전자#대표이사#김철수#부회장

위와 같이 “삼성전자의 신임 대표이사 부회장 김철수”와 동일한 의미의 시퀀스의 가능한 형태들을 기술하기 위해서는 이들 시퀀스를 구성하는 어휘 성분들 사이의 모든 어휘적 조합의 가능성을 고려하여 이에 대한 부분적인 문법을 구현해야 한다. 아래 <그림 4>와 같은 LGG 그래프는 모두 192가지의 명사구 패턴을 보이는 문법으로 이는 기존의 고유명사 사전이나 복합명사 사전만으로 처리할 수 없는 다양한 표현들에 대한 처리를 가능하게 한다.



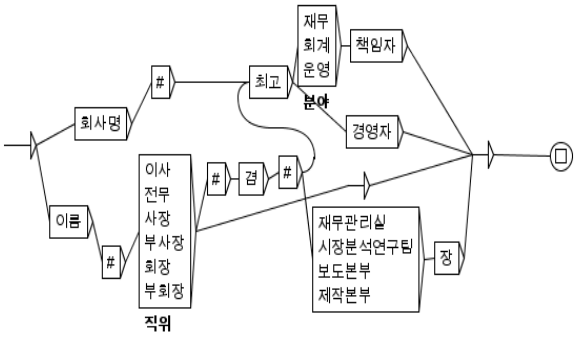
<그림 4> 명사 연결형 시퀀스에 대한 LGG 문법 예

일정 의존 명사와 같은 특정 문법적 표지에 의해 연결된 형태를 살펴보면 특히 의존 명사 중에서 “겸”과 “및”이 주로 사용되는 것을 볼 수 있다. 이때 직책이나 직위를 나타내는 어휘 중 두 직위를 한 사람이 맡고 있을 경우 “겸”이라는 의존 명사가 주로 사용되고 부서명에서는 “국내 및 해외 영업”처럼 “및”이라는 의존 명사가 주로 사용된다. 그러나 “겸”과 “및”은 결합 가능한 어휘들의 의미적 분포에서 차이를 보인다. “겸”의 앞과 뒤에는 직책이나 직위를 나타내는 어휘만이 나타나지만 “및”의 경우는 다양한 의미 분야의 어휘들이 사용되고 있다.

<겸>
“교수 **겸** 학회장”
“회장 **겸** 최고경영자”
“이사 **겸** 본부장”

<및>
“국내 **및** 해외 영업총괄”
“전 동원증권 **및** 한국증권”
“부사장광주 광산업육성 **및** 집적화 사업계획”
“전자여권 **및** 공공사업 분야”
“IT 분야 학회 활동 **및** 각 기관, 대학교를 두루 다니며 특강을 통해 ‘지식혁신의 전도사’ 역할을 하고 있다.”

아래 <그림 5>는 의존 명사 “겸”에 의해 연결된 명사구 패턴에 대한 LGG 문법의 예를 보인다.



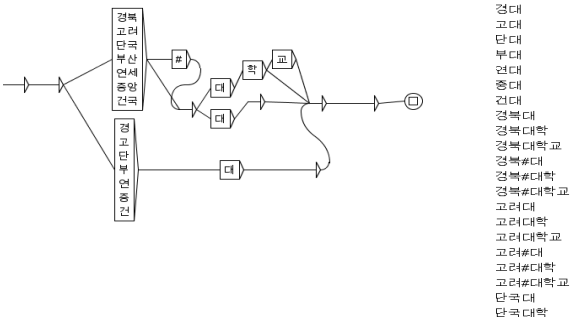
이름#회장#겸#제작본부장/직위
이름#회장#겸#보도본부장/직위
이름#회장#겸#시장분석연구팀장/직위
이름#회장#겸#재무관리실장/직위
이름#회장#겸#최고운영책임자/직위분야
이름#회장#겸#최고회계책임자/직위분야
이름#회장#겸#최고재무책임자/직위분야
이름#회장#겸#최고경영자/직위
이름#부사장/직위
이름#부사장#겸#제작본부장/직위
이름#부사장#겸#보도본부장/직위
이름#부사장#겸#시장분석연구팀장/직위
이름#부사장#겸#재무관리실장/직위
이름#부사장#겸#최고운영책임자/직위분야
이름#부사장#겸#최고회계책임자/직위분야
이름#부사장#겸#최고재무책임자/직위분야
이름#부사장#겸#최고경영자/직위
이름#사장/직위
이름#사장#겸#제작본부장/직위
이름#사장#겸#보도본부장/직위
이름#사장#겸#시장분석연구팀장/직위

<그림 5> 의존 명사 “겸”이 포함된 명사 시퀀스의 예

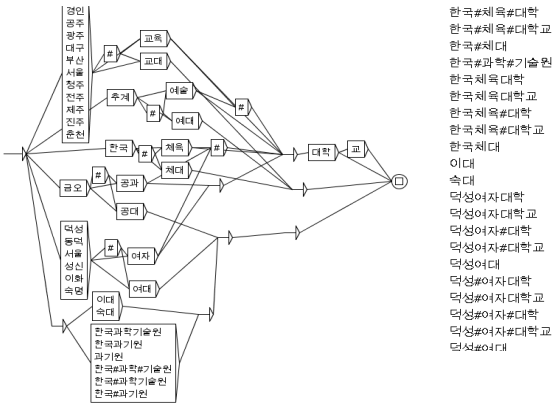
앞서 살펴본 패턴들 이외에도 IT 인물 관련 정보 유형에는 학력과 관련된 정보가 있다. 이들 정보에 대한 표현도 실제 텍스트에서는 매우 다양하게 나타나는데, 예를 들어 “A는 어느 대학을 졸업하고....” 등과 같은 표현에서 “대학”을 나타내는 개체명은 아래와 같이 여러 다양한 양상을 보이는 것을 확인할 수 있다.

1. “연세 대학교/연세대학교”
   “연세 대학/연세대학”
   “연세대/연대”
2. “서울 대학교/서울대학교”
   “서울 대학/서울대학”
   “서울대”
3. “육군사관학교”
   “육사”
4. “한국예술종합학교”

“대학명”은 위와 같이 4가지 유형으로 분류될 수 있는데, 첫째는 “연대”와 같은 형태로 축약이 되는 대학명 유형이며, 둘째는 “서울대”와 같이 축약은 되지만 “연대”와 같은 형태는 취할 수 없는 형태가 여기 해당한다. 셋째는 “육사”, “해사”, “공사”와 같은 형태로 축약될 수 있는 형태들이며, 마지막으로 일반인들에게는(즉 해당 학교 관련자나 구성원이 아닌 경우) 어떤 형태로든 축약하는 것이 부자연스러운 대학명을 한 유형으로 범주화할 수 있다. 이들 4개 유형은 IT 인물 관련 정보를 표현하는 문장의 시퀀스를 나타내는 주 그래프(main-graph)에 4가지 형태로 세분된 <대학>이라는 하위 그래프(sub-graph) 형태로 삽입된다. 다음 <그림 6>은 예를 들어 99개 대학에 대한 487개의 서로 다른 다양한 언어 패턴을 표상하는 하위 그래프의 일부 예를 보인다.



<그림 6> 대학명 하위 그래프의 일부 예(1)



<그림 7> 대학명 하위 그래프의 일부 예(2)

IT 인물 관련 텍스트에서 관찰되는 이와 같은 직책, 학력 등의 표현은 일정 기관명 고유명사의 다양한 축약 패턴과 함께 실현되는 경우가 많아 이들에 대한 체계적이고 빠짐없는 기술이 이루어지지 않으면 기존의 사전만으로는 이들 인물 관련 키워드에 대한 정확한 검색을 기대하기 어렵다는 점을 확인할 수 있다.

2.3. 서술어 구성 명사 키워드의 특징

IT 인물 관련 정보를 검색하는 데에 있어 이들 키워드가 실현될 수 있는 문장의 문형 구조에 대한 기술은 중요한 언어 자원이 된다. 문형 구조를 기술하기 위해서는 그 핵이 되는 <술어(predicate)> 성분에 대한 연구가 수반되어야 하는데, 예를 들어 A라는 사람의 현재 소속과 경력을 알고자 한다면 “A가 … 근무하다/재직하다/입사하다” 등의 관련 서술어가 포함된 문장에서 사용자가 요구하는 정보를 추출할 수 있는 가능성이 높아진다.

IT 인물 관련 텍스트에서 발견되는 서술어의 특성을 살펴보기 위하여, 본 연구에서는 현재 사용한 코퍼스의 일부에서 서술어의 종결형/연결형 활용 변화형들을 분석해 보았다. 이렇게 추출된 표본 코퍼스에서 총 446개의 서술어가 나타났고, 10회 이상의 빈도를 보이는 서술어의 목록을 데이터베이스화하여 제시하면 아래 <표 5>와 같다.

	서술어	빈도수	백분율(%)	예
1	선임하다/되다	37	8.2	선임됐다, 선임하고, 선임될
2	밝히다	22	4.9	밝혔다, 밝히고
3	거치다	21	4.7	거쳐, 거치며
4	졸업하다	14	3.2	졸업하고, 졸업한, 졸업해
5	역임하다	10	2.2	역임한, 역임했다, 역임했으며
6	입사하다	10	2.2	입사한, 입사해, 입사했다

<표 5> 10회 이상 빈도의 서술어의 예

위의 서술어의 빈도 분석 결과를 보면, IT 인물의 인사 발령이나 경력과 관련된 내용이 문서의 주요 정보를 구성함을 알 수 있다. 또한 “선임하다”, “졸업하다”와 같이 “한자어+하다/되다” 유형의 서술어가 전체적으로 높은 빈도를 나타내는데, 코퍼스 전체에서 “한자어+하다/되다” 유형의 서술어와 “일하다”, “거치다”와 같은 고유어 비중을 비교해 보면 아래 <표 6>에서 보이는 바와 같다.

서술어 유형	빈도	백분률(%)	예
한자어+하다/되다	228	51.1	취임하다, 겸임하다
고유어	186	41.7	밝히다, 얻다, 일하다
기타	32	7.2	때문이다, 것이다

<표 6> “N+하다/되다”와 “V” 유형의 서술어 비교

표본 코퍼스에서 <종결형> 서술어를 중심으로 살펴보면 “선임했다고 밝혔다”, “이끌게 됐다” 등의 복합 서술어 구성이 많이 나타나는 것을 볼 수 있다. 코퍼스에서 종결형 서술어의 유형을 분석한 결과를 제시하면 아래 <표 7>에서 보이는 바와 같다.

연결 유형	예	빈도	백분율
지속	겸임하고 있다, 일해 왔다	23	42%
인용	선임했다고 밝혔다	15	27%
피동	취임하게 된다	13	24%
기타	취득하기도 했다	4	7%
		55	100%

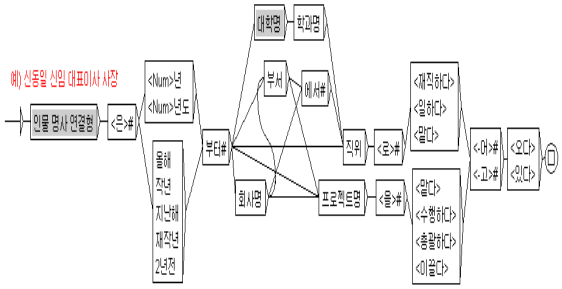
<표 7> 종결형 서술어 패턴의 예

위에서 “...고 밝혔다/말했다” 등의 직접 또는 간접 인용 형태는 전체 종결형에서 27%의 비중을 차지하는 것으로 나타났다. 이는 코퍼스가 인터넷 신문 기사를 중심으로 구성되어 있기 때문에 나타나는 현상으로, 이것은 실제로 IT 인물 관련 텍스트의 특징에 속한다 기보다는 인터뷰 등의 간접 인용문이 주로 나타나는 신문 기사의 담화적 특징에서 기인하는 현상으로 보인다. 그러나 “(겸하)고 있다, (담당)해 왔다” 등과 같은 ‘지속’을 나타내는 보조 동사가 높은 빈도를 보이는 현상은 코퍼스의 핵심 정보에 <인물의 경력/직책> 등의 특징이 포함되어 있기 때문인 것으로 보인다. 실제 표본 코퍼스에서 이러한 보조 동사가 나타난 문장을 살펴보면 아래 <표 8>과 같다.

신동일 신임 대표는 … 2002년부터 재무관리실 자금담당 상무로 재직해왔다.
전 부회장은… 2003년 6월부터 제2대 상근부회장으로 재직해 왔다.
신임 이 CTO는 … 2004년 3월부터 한북대 컴퓨터공학과 교수로 재직해 왔다.
그는 … 지난 99년 온더아이티에 입사해 올해부터 해외사업을 총괄하고 있다.
심 신임 대표는 … 1996년부터 … 상품운용본부장, 중부지역본부장을 맡아왔다.

<표 8> 인물의 경력/직책 관련 복합 술어 표현의 예

위의 문장들에서 관찰할 수 있듯이 이들 술어 구조는 “2002년부터”, “올해부터”, “지난 99년” 등과 같은 시간(time) 표현과 함께 실현되는 것을 볼 수 있으며, 이들 정보는 그 시간 표현(시작상 또는 지속상 등의 표현)의 범위 내에서 현재 유효하다는 사실을 보이고 있다. 이처럼 ‘지속’의 보조 동사는 경력에 관련된 정보를 구성하는 데에 매우 중요한 술어 성분으로 기능하고 있기 때문에 일반 텍스트에서 나타나는 술어 성분들과는 별도의 특별한 기술이 수반되어야 한다. 따라서 이들 특정 보조 동사가 포함된 문장들에 대한 체계적인 기술이 이루어진다면 IT 인물 관련 정보의 효율적인 추출이 가능할 것이다. 아래 <그림 8>은 이러한 목적으로 구현된 LGG 그래프 문법의 예를 보인다[5].

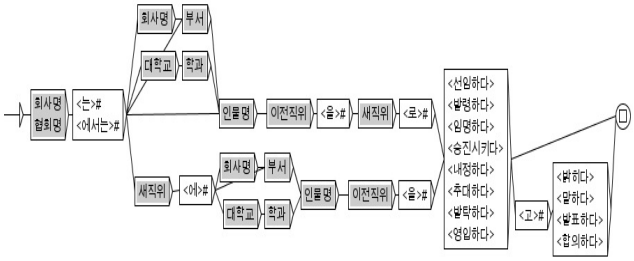


<그림 8> LGG 그래프 문법으로 표상된 술어 구조 예

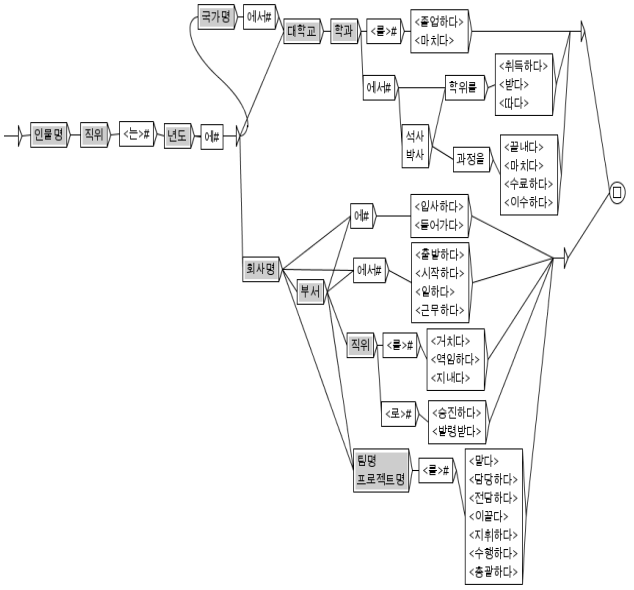
이상과 같이 표본 코퍼스에서 술어 성분을 중심으로 그 핵심 정보의 유형을 분석한 결과, IT 인물 관련 텍스트에서 주요 정보를 구성하는 것은 <인물의 인사 발령>과 <경력 및 직책>에 관한 내용임을 알 수 있다. 따라서 이들 정보를 담고 있는 적절한 서술어의 유형을 찾아내어 텍스트에서 이들 서술어가 포함된 문장들을 추출하고 이로부터 사용자가 요구하는 정확한 정보를 제공해야 할 것이다. 이를 위해서 표본 코

5) 위의 그래프에서 회색으로 표시된 box는 2.2.에서 구축된 개체명들에 대한 그래프가 <하위 그래프(Sub-graph)>로 삽입되는 것을 의미하며, 이들 하위 그래프는 여러 LGG 그래프에 재사용 가능하다.

퍼스에서 사용된 술어 성분에 대한 분석을 바탕으로, <인사 발령>과 <경력 및 직책>의 두 부류에 해당하는 서술어 유형을 수집하고, 그 서술어들에서 가능한 문장 구조를 LGG 문법으로 구축하였다. 아래 <그림 9>와 <그림 10>은 각각 <인사 발령>과 <경력 및 직책> 관련 문형 구조에 대한 LGG 그래프 문법의 예를 보인다.



<그림 9> 인사 발령 LGG 문법의 예



<그림 10> 경력 및 직책 LGG 문법의 예

위의 <그림 9>의 LGG 그래프는 전체 800 여개의 <인사 발령> 관련 문장 구조를 표상하고 있으며, <그림 10>의 LGG 그래프는 <경력 및 직책>과 관련된 문장 구조를 120개 이상 표상하고 있다. 이상에서 보인 바와 같이 LGG 그래프 방식으로 구현된 술어 문형 패턴 문법은 IT 인물과 관련된 정보를 유한한(finite) 방식으로 빠짐없이 기술하고 있어 이와 관련된 정보를 보다 정확하고 효율적으로 검색할 수 있는 시스템의 개발을 가능하게 할 것이다.

3. 결론

본 연구에서는 IT 인물 관련 텍스트 정보의 효율적인 검색을 위하여 이러한 정보를 담고 있는 특정 영역 텍스트의 언어적 속성을 분석하고 이를 기반으로 LGG 그래프 문법을 구현하는 방법에 대하여 논의하였다. IT 관련 문서의 예로서 <전자 신문> 텍스트를 대상 코퍼스로 하였고 이로부터 표본 코퍼스를 구성한 후 IT 인물 관련 핵심 정보를 담고 있는 언어 형식들을 분석하였다. 그 결과 본 연구의 대상인 IT영역 내에서의 텍스트는 다음과 같은 3가지 언어적 특징으로 요약될 수 있는 것을 관찰할 수 있었다.

- 외래어 음차 표기 명사들의 출현
- 명사 연결 구성의 복합 명사구의 출현
- 일정 유형의 서술어를 구성하는 명사들의 출현

외래어 음차 표기 명사들은 특히 IT 분야라는 분야적 속성에 의해 핵심 키워드의 중요한 비중을 차지하고 있었으며 이들은 더욱이 한국어 텍스트 내에서 다양한 음차 표기 형태로 실현될 수 있어 이들에 대한 체계적인 검색 및 분석에 어려움을 유발하는 요인이 되고 있음을 살펴볼 수 있었다. 따라서 이들에 대한 LGG(Local-Grammar Graph) 문법을 구현하여 향후 효율적인 검색이 가능한 시스템의 개발에 사용될 수 있는 중요한 데이터베이스가 되도록 하였다.

복합 명사구의 경우는 IT 인물 관련 명사들이 나열된 연결 구성이 주를 이루었는데, 이들은 크게 다음의 3가지 유형으로 나누어질 수 있음을 살펴보았다. 첫째는 고유 명사인 사람의 이름과 함께 소속이나 직위, 학력이나 경력을 나타내는 어휘들이 연거푸 나열되는 방식이고, 둘째 부류는 이동 및 생략, 그리고 동일 지시 대상에 대한 축약 등의 변형에 의해 실현될 수 있는 다양한 패턴들의 경우이며, 끝으로는 일정 의존 명사에 의해 연결된 형태로 실현되는 명사구의 유형으로 정리된다. 이러한 명사 연결 구성 복합어들은 모두 LGG 그래프 문법으로 구현되어 실제 텍스트에서 실현 가능한 모든 구조들을 인식하고 검색하는 것이 가능하다는 사실을 제시하였다. 또한 대학명과 같은 일정 기관명의 다양한 언어적 패턴에 대해서도 LGG 그래프 방식으로 효율적으로 제어하고 기술할 수 있음을 논의하였다.

술어 성분을 구성하는 일정 서술 명사들의 경우, “한자어+하다/되다” 유형이 중요한 비중을 차지했고 종결형에서는 ‘지속’의 보조 동사 “-고 있다/-어 오다” 유형이 가장 높은 빈도를 나타내고 있음을 볼 수 있었다. 이때 이러한 문형 패턴에 실현되는 서술 명사들의 어휘적 속성은 매우 한정적이어서 ‘임명’ ‘발령’ ‘발탁’ 등의 <인사 발령> 관련 개념어 부류와 ‘졸업’ ‘수행’ 등의 <경력 및 직책> 관련 개념어 부류가 주를

이루었다. 이들 문서에서 일정 유형의 '지속'의 보조동사는 <시간> 표현과 함께 경력에 관한 정보를 제공하기 위해 사용되고 있어 이들에 대한 LGG 그래프 문법의 구성에 대하여 논의하였다. 이러한 연구를 바탕으로 IT 인물 관련 텍스트에서 전형적으로 나타나는 문장 구조들을 LGG 문법으로 표상하였다.

본 연구에서는 IT 분야 인물들에 관련된 정보 유형이 어떠한 것들이 있는지 살펴보고 이러한 정보들이 어떤 언어적 패턴을 통해 실현되는지를 분석하였다. 그리고 도식화된 규칙으로 예측하고 설명할 수 없는 이와 같이 다양한 어휘적 제약 및 문법적 속성들을 유한한 방식으로 표상할 수 있도록 LGG 오토마타 문법을 구현하는 문제에 대하여 논의하였다. 실제로 LGG 그래프의 구축이 수동으로 이루어진다는 점에서 많은 시간과 노력이 요구된다. 하지만 앞서 언급하였듯이 규칙에 의해 처리할 수 없는 인간의 언어적 현상을 표현하고 시스템 내에서 인식, 처리 가능하게 하는 데 많은 도움이 될 것으로 기대된다. 본 연구에서 제시한 연구 방법론은 IT 분야와 같이 특정 도메인에서의 정보 검색의 질을 높이는데 매우 효과적인 방법으로 사용될 수 있을 것이며, 향후 보다 다양한 도메인에서 적용할 수 있기 위해서 지속적이고 확장된 데이터베이스의 구축이 절실하게 요구된다.

참고문헌

강병주, 최기선, 윤준태, 1998, 한국어 정보검색에서 복합명사 색인 실험, 제10회 <한글 및 한국어 정보처리> 학술대회

남기심, 고영근, 2004, <표준국어문법론>, 탑출판사

남지순, 2002, 고유명사 자동 처리를 위한 전자 데이터베이스의 구축, <어학연구> Vol.38, No.1.

남지순, 2005, 프랑스어 언어자원 구축을 위한 부분문법 방법론의 소개, <한국프랑스학논집> 제 49집.

남지순, 2007, 웹문서 의미 지식 추출을 위한 LGG의 구축, <한국프랑스어문교육학> 제 25집.

박선옥, 2005, <국어 보조동사의 통사와 의미 연구>, 도서출판 역락.

윤보현, 1995, 복합명사 구성 패턴과 통계 정보를 이용한 한국어 복합명사 분석, 고려대 석사학위논문

은종진, 박선영, 2000, 고성능 한국어 형태소 분석을 위한 어미 분류, 제 12회 <한글 및 한국어 정보처리 학술대회>.

이강만, 2001, 외래어 표기법의 문제점 연구. 충남대학교

이상권, 1996, 외래어 표기와 사용실태 연구. 계명대 교육대학원

이응백, 1987, 외래어의 표기와 발음 문제. 국어 교육. Vol. 1987 No.59

이재성, 1999,『다국어 정보검색을 위한 영한 음차 표기 및 복원 모델』한국과학기술원 전산학과 박사학위논문

조용준, 1995, 서술성 명사의 통사-의미론적 특성 연구, 건국대학교 대학원 국어국문학과 석사학위논문

최기선, 1993, 한국어에서의 복합 명사구 인식에 대한 연구, 최종 연구 보고서, 한국전자 통신연구소

황화상, 2001, 국어 형태 단위의 의미와 단어 형성, 월인

국립국어연구원, 1995, 기본외래어용례집

Chinchor, N. & Marsh, E. 1998. MUC-7 Information Extraction Task Definition (version 5.1), *Proceedings of MUC-7.*

Gross, M. 1997. The construction of local grammars. In E. Roche and Y. Schabes, editors, *Finite State Language Processing,* 329-354. The MIT Press.

Paumier, S. 2003. De la reconnaissance de formes linguistiques à l'analyse syntaxique, Thèse de Doctorat, Université de Marne-la-Vallée.

Sastre, J. M. 2006. "Computer Tools for the Management of Lexicon-Grammar Databases", poster, Actes de la 13e conférence sur le traitement automatique des langues naturelles (TALN), Leuven, 10~13 april 2006, UCL Presses Universitaires de Louvain, 600~608.

Silberztein, M. 1993. Dictionnaire électronique et analyse automatique de textes - le système INTEX, Paris: Masson.

Soderland. 1997. Learning to extact text-based information from the World Wide Web, *Proceedings of third International Conference on Knowledge Discovery and Data Mining.*

Soderland. 1999. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning,* 34(1-3):233~272.