정보 추출

국민대학교 강승식* · 우종우**
한국전자통신연구원 윤보현 · 박상규

1. 서 론

정보 추출(Information Extraction)은 사용자가 원하는 정보를 검색하거나 추출하는 연구 분야로서 대용량의 정보자료들로부터 미리 정의된 주제나 관심분야의 정보만을 인식하여 요약된 형태로 가공을 한다. 즉, 정보 추출 시스템은 자연언어 처리(Natural Language Processing) 기술을 바탕으로 문서의 내용을 분석하여 주어진 주제영역에서 유용한 정보들을 데이터베이스와 같이 구조화된 형식으로 저장한다. 자연언어 이해(Natural Language Understanding)와는 달리 정보 추출 시스템은 텍스트에서 적합한 부분을 효과적으로 찾아서 추가적인 처리를 한다. 예를 들면, 테러에 관한 영역에서는 범인 및 희생자 이름, 사용된 무기, 사건발생 날짜 및 장소 등의 정보를 추출할 수 있으며, 비즈니스 영역에서는 회사이름, 제품명, 시설, 그리고 재정상태 등의 정보를 추출할 수 있다[1,2,3].

이러한 정보 추출에 관한 연구는 미국정부 주도의 Message Understanding Conferences(MUC) 학술회의로 대표된다[4,5,6,7]. 1990년 후반, 미국정부는 정보 추출 연구의 평가 및 활성화를 위해 MUC을 주관하였는데, MUC 회의 참가자들은 미리 정해진 주제에 대하여 정보 추출 시스템을 구축하고, MUC 본부에서 개발한 공식 평가 프로그램으로 평가를 한다[8]. MUC은 자연언어 처리시스템 평가를 위한 최초의 대규모 노력으로서, 기존의 자연언어 처리 시스템 개발이 연구자들의 관심 분야에서만 독립적으로 개발된 반면, MUC에서는 같은 조건 하에서 서로 다른 NLP 시스템들을 비교 분석할 수 있는 기반을 제공하였다. 더욱이 뉴스기사와 같은 실제 텍스트들을 자료로 사용함으로써 NLP 연구자들에게 실제 텍스트의 활용에 관한 자극을 주었다는데 의미가 있다.

이러한 연구 노력으로 정보 추출은 실세계의 텍스트기반 응용 분야를 위한 필수적 기술로 인정되고 있다. 몇 가지 개발되고 있는 시스템들의 예를 보면, 생명보험 분야를 분석하는 시스템[9], 의료환자의 기록 중에서 진단 및 증세 등을 요약하는 시스템[10], 신문기사, 라디오, TV 등에서 폭력에 관한 기사의 발견 및 요약 시스템[11] 등이 있다. 또한, 인터넷 기반의 응용 시스템들도 정보 추출 기술들을 사용하는데, 예를 들면, 웹 페이지에서 직접 지식베이스를 구축하는 자연언어 시스템[12], 신문기사나 웹 페이지나 또는 광고물 등에서 구직 리스트를 생성하는 시스템, 뉴스 그룹 질의 시스템[13], 그리고 웹 페이지로부터 날씨 예측 데이터베이스 구축 시스템[14] 등이 있다.

MUC에서는 어떤 점에서 이러한 정보 추출 시스템들을 평가할 수 있었지만, 전반적인 성능평가를 내리기는 어렵다[15]. 최소한의 성능평가도 추출 작업의 복잡도, 자연언어처리 시스템의 지식베이스의 질적 문제, 도큐먼트의 문법적 복잡도 등 여러 가지 요인에 의해 객관적인 평가가 쉽지 않기 때문이다. Will(1993)에 의하면, 최고의 자동화된 정보 추출 시스템도 숙달된 사람보다 두 배 이상의 오류를 나타낸다고 한다[16]. 또한 정보 추출 시스템은 최근 눈부신 발전을 하였음에도 불구하고 여전히 정확성과 견고성이 훨씬 더 개선되어야 하며, 추출 작업이 특정 영역에 의존적인 특징 때문에 숙달된 전문가들이 많은 노력을 해도 시간이 많이 소요된다는 문제점이 있다.

* 종신회원
** 정 회 원

2. 정보 추출 시스템의 구조

자연언어 처리 기반 정보 추출 시스템 구조는 몇 년간의 다양한 실험과 연구를 거쳐 현재는 어느 정도 정보 추출을 위한 기본적인 시스템의 구조가 확립되어 있다. 정보 추출의 목적이나 방법에 따라 부분적인 차이는 있지만 대체적으로 연구자들이 동의하는 정보 추출 시스템의 개략적인 구조는 그림 1과 같다[17].

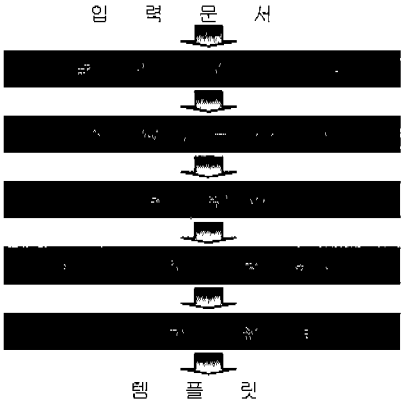

그림 1 정보 추출시스템의 구조

(1) 문장의 토큰화 및 태깅

이 단계에서는 분석할 문장을 단어로 분리하고 각각의 단어에 대한 기본적인 품사 할당과 중의성이 있는 품사들에 대해 중의성 해소를 위한 태깅을 수행한다. 필요한 경우 단어에 간단한 의미 분류도 행할 수 있다. 서구어의 경우, 공백이나 구둣점 등을 이용하여 문장에서 쉽게 단어를 분리할 수 있지만 중국어나 일본어와 같이 단어 경계가 없는 언어의 경우에는 좀 더 복잡한 단계를 거치게 된다[1].

(2) 문장 분석

문장 분석 단계에서는 명사구나 동사구, 부사구와 같은 간단한 구조를 찾아낸다. 몇몇 시스템에서는 문장의 표층 주어나 직접 목적어 같은 간단한 구조나 접속사, 동격의 구, 복잡한 구구조를 인식하기도 한다. 또한 파싱 전후나 도중에 추출 분야와 관련이 있는 엔티티(entity)들을 찾아내고, 적절한 의미를 부여하기도 한다.

문장분석과정에서 살펴보아야 할 중요한 점 중 하나는 문장 분석이 정보 추출을 위한 것이라는 점이다. 정보 추출을 위한 문장 분석은 주어진 문장에 대한 완벽한 구문 분석 트리를 생성할 필요는 없다. 대신에 부분 구문 분석을 통해 문장에 대한 부분적인 파스 트리만으로도 정보 추출에서 필요로 하는 정보를 얻을 수 있기 때문이다[18,19,20].

(3) 정보 추출

추출 단계는 도메인과 관련이 있는 엔티티를 인식하는 것으로 전체 시스템의 구조 중 첫번째로 도메인에 종속적인 부분이다. 예를 들어, 자연 재해와 관련된 도메인의 경우, 이 단계에서 재앙의 종류, 재앙이 일어난 장소, 실제로 발생한 피해의 정도 등이 인식될 수 있다. 우리는 정보 추출을 위한 도메인을 알고 있기 때문에 각 문장별로 어떠한 정보를 추출해야 되는지 알고 있다.

(4) 대용어 참조 해소 및 병합

대용어 참조 해소(coreference resolution)는 정보 추출에서 가장 어려운 분야 중에 하나이다. 이 단계에서는 동일한 엔티티를 가리키는 지시어를 해결하거나 반복해서 사용되는 단어를 나타내는 표현들에 대한 참조를 해결한다. 예를 들어 대명사나 보통 명사가 어떤 특정한 고유명사와 동일한 엔티티를 가리킬 때 이들을 하나의 참조로 보고 각 문장에서 이러한 단어들을 해결한다. 문장에서 반복을 피하기 위해 많이 사용되는 이러한 표현들은 정보의 추출을 위해 반드시 하나의 엔티티 참조로 해결되어야 한다. 예를 들면 다음 두개의 문장에서,

"종로서적에서 책을 샀다."

"그 곳에서 3000원을 지불했다"

"그 곳"과 "종로서적"이 동일한 장소를 나타내는 엔티티를 참조하고 있음을 알 수 있어야 한다. 또한 이름과 별명에 대해서도 참조 문제를 해결해야 하는데, "Mr. Gates", "William H. Gates", "Bill Gates" 등이 동일한 이름임을 파악해야 한다.

(5) 템플릿 생성

템플릿 생성 단계에서는 주어진 문서에서 얻고자 하는 몇 개의 사건에 대해서, 문서에서 추출된 정보를 이용하여 템플릿을 생성한다. 이 단계에서 도메인 종속적인 제약 지식(constraints)을 활용할 수 있으며, 템플릿 생성을 위하여 추론(inference)이 사용될 수 있다. 예를 들어, MUC 테러 관련 문서에서 민간인이 부상당하거나 민간인의 재산이 손해를 입지 않으면 군 목표물에 대한 테러도 정보 추출 대상에서 제외시킨다. 또한, 템플릿의 각 슬롯

도 경우에 따라서는 문장에 나타난 단순한 문자열의 나열이 아닐 수도 있으며, 날짜나 시간, 장소와 같은 슬롯에는 적절한 정규화도 필요하다.

3. 자연언어 분석 기법

정보 추출 시스템의 기본적인 구조는 접근 방법에 따라 다양하게 구성될 수 있지만 대부분의 정보 추출 시스템은 태깅, 부분 파싱, 의미 분석, 격틀(case frame) 활성화와 담화 분석으로 구성된다[21].

3.1 부분 파싱

처음 정보 추출 시스템에 대한 연구가 시작되었을 때, 문장 분석 분야에서 많은 시스템들이 다양한 접근 방법을 채택했었다. 어떤 시스템은 언어학적 이론에 기반하여 주어진 문장에 대한 완벽한 구문 분석 트리를 생성하기도 하였고[22,23,24], 극단적으로 구문 분석 과정이 없는 정보 추출 시스템도 있었다[25]. 정보 추출 시스템에 대한 연구가 점차 진행됨에 따라 구문 분석 분야에서는 어느 정도 표준화된 접근 방법이 대두되기 시작하였다. 현재 개발되거나 연구되고 있는 정보 추출 시스템들은 대부분 부분 파싱 기법을 활용하고 있다. 언어학에 기반한 정보 추출 연구자들조차도 이러한 조류를 받아들이고 있는데, 이는 정보 추출 분야에서만큼은 완전 구문 분석에 드는 추가 비용과 완전구문 분석을 위한 시스템의 견고성 부족에서 원인을 찾을 수 있다. 실제로 신문에서 나타나는 많은 문장들이 복잡하고 길며, 비문법적인 표현이나 중의성을 많이 내포하고 있다.

실제로 정보 추출 분야에서 추출될 정보들은 짧은 절이나 구안에 포함되어 있는 경우가 많다. 문장에서 이러한 구나 절을 제외한 단어들은 정보 추출에서 필요한 유용한 정보를 포함하고 있지 않기 때문에 무시될 수 있다. MUC-3에서 사용되었던 아래의 문장을 살펴 보자.

In an action that is unprecedented in Colombia's history of violence, unidentified persons kidnapped 31 people in the strife-torn banana-growing region of Uraba, the Antioquia governor's office reported today.

실제로 이 문장에서 관심이 있는 정보는 "우라바 지역에서 미확인된 사람들이 31명을 납치하였다"이다. 이러한 정보를 구성하는 문장을 뺀 나머지는 문장을 분석할 필요가 없다. 이 문장에서 우리는 단지 "kidnapped"라는 단어와 "<X> kidnapped <Y> in <Z>"라는 패턴에 의해 범인, 피해자와 장소를 쉽게 찾아낼 수 있다. 이러한 접근 방법은 우리가 문장에서 도메인에 관련된 특별한 종류의 정보만을 인식할 것이고, 이러한 인식은 간단한 구문 분석을 통해 가능하다는 점에 기반하여 출발한 것이다. 문장이 간단한 구조로 구성되어 있을 경우, 이러한 접근 방법을 통해 높은 정보 추출 성능을 보일 수 있다. 그러나 어떠한 정보가 우리가 인식하고자 하는 도메인에 연관된 특별한 종류에 포함되었는지 판단하기 어려운 경우가 종종 발생한다. 이러한 경우에는 단순한 패턴 매칭에 의한 방법으로는 유용한 정보를 추출하기 어렵다. 아래의 간단한 예문을 살펴 보자[21].

1. The mayor was killed by FMLN guerrillas.
2. The mayor was killed by armed men.
3. The mayor was killed by armed men during a hold-up in a convenience store.
4. The mayor was killed by armed men in retaliation for the arrested of a prominent FMLN leader.
5. The mayor was killed and the FMLN claimed responsibility for the murder.

1번부터 4번 문장까지는 "killed by <X>"라는 간단한 패턴에 의해 가해자에 대한 정보를 쉽게 추출할 수 있다. 하지만 무장한 사람이 단순한 강도인지, 목적을 가진 테러리스트인지를 구별하는 것은 어렵다. 5번 문장에서는 이러한 패턴을 이용할 경우, 가해자에 대한 정보를 얻을 수 없다. 그러나 두번째 절에 대한 분석과 추론을 통해 가해자가 FMLN이라는 테러 조직임을 알 수 있다. 이러한 예들은 정보 추출과 관련이 있는 정보인지 아닌지를 단어 수준에서 판단하기가 어려움을 보여주고 있다.

정보 추출 시스템이 이러한 문제점을 포함하고 있기 때문에 최근의 연구에서는 좀 더 자세한 구문 분석이 필요하다는 연구자들도 있다. 그러나 정보 추출 시스템에서 구문 분석은 상대적으로 부분적 역할로 충분한 것으로 여겨지고 있으며, 문장에 나

타난 주요 단어를 이용한 개념 활성화와 개념 실제화가 주요한 역할을 차지하고 있다. 실제의 경우에서, 구문 분석은 단지 문장에 나타나는 패턴들의 구문적 역할을 격틀의 개념적 역할로 사상하는 기능을 하며 이는 간단한 구문 분석(shallow parsing)을 하고 있다. 이러한 파싱 기법으로 인해 손실되는 정보를 보충하기 위해 담화 분석이 정보 추출 시스템의 중요한 부분으로 부각되고 있다. 비록 담화 분석을 위해 완전 구문 분석이 좀 더 유용한 정보를 전달할 수 있지만, 개념에 기반한 문장 분석과 도메인 지식 등이 담화분석의 주요 연구 주제로 부각되고 있다.

3.2 담화 분석

MUC에서 연구자들이 얻은 교훈 중 하나는 문어체 문장에 대한 담화 분석은 매우 많은 연구 과제를 안고 있다는 점이었다. 대부분의 담화 분석에 관한 이론들은 많은 현실 세계 지식을 필요로 하며, 대량의 문서에 대한 실험이 실시된 적이 없었다. 실제로 정보 추출을 위한 문장들은 담화 이론의 관점에서 생각하는 것보다 많은 불규칙성과 중의성을 내포하고 있다[26,27,28,29].

일반적으로 담화 처리는 사건이나 객체를 추적하고 이들에 대한 관계를 이해하는 것이다. 대명사나 고유명사, 한정 명사구들에 대한 대용어 문제 해결은 담화 처리의 한 예로 볼 수 있다. 서로 다른 문장에서 서술하는 하나의 사건에 대한 다중 참조 정보를 효율적으로 추적하여 적절한 하나의 사건으로 병합하거나 하나의 사건인 것처럼 보이지만 실제로는 서로 다른 사건의 나열인 정보를 효율적으로 분해하는 방법들이 필요하다.

언뜻 보기에, 정보 추출을 위한 담화 분석은 추출하고자 하는 목적과 관련이 있는 객체와 사건만을 추적하기 때문에 일반적인 이야기 이해를 위한 시스템보다 쉬운 것처럼 보인다. 그러나 담화 분석이 완벽한 지식이 없이도 이러한 문제들을 해결해야 하기 때문에 연구해야 할 많은 과제를 가지고 있다. 대부분의 정보 추출 시스템은 전체 문서에서 주제와 관련이 있는 문서의 일부분만을 인식하고 저장한다. 결과적으로 사건을 구분짓는 중요한 정보를 포함하지 않을 수 있다. 예를 들어, 정보 추출 시스템은 아래와 같은 단락을 (a). (b), (c)와 같은 정보의 조각으로 단락을 분해할 것이다.

Members of the 8th front of the self-styled revolutionary armed forces of Colombia [FARC] have carried out terrorist attacks in southern Cauca department to protest patriotic union [UP] presidential candidate Bernardo Jaramillo Ossa's murder.

(a) Members of the 8th front of the self-styled revolutionary armed forces of Colombia [FARC] have carried out terrorist attacks
(b) Terrorist attacks in southern Cauca department
(c) Patriotic Union [UP] presidential candidate Bernardo Jaramillo Ossa's murder

이 단락의 조각들은 Jaramillo Ossa의 죽음이 테러에 의한 것이라고 추론할 수도 있다. 이는 중요한 구절인 "to protest"가 테러리즘과 깊은 연관이 있는 구절이라고 판단하지 않아 이 구절을 배제하기 때문이다. 실제 테러리즘과 관련된 지식에 이러한 표현을 인식하기 위한 격틀 정보를 가지고 있는 경우는 드물 것이다. 그러나 기사에서 "to protest" 구절은 살인이 테러 공격보다 먼저 발생했음을 이해하는데 중요한 단서를 제공하고 있다.

이러한 예는 정보 추출과 자연언어 이해 시스템 간의 차이점을 설명할 수 있게 해준다. 좀 더 자세한 이해를 위해 깊은 수준의 문서 이해를 할 것인지, 간단한 문서 이해를 기반으로 정보 추출을 할 것인지는 문서 분석 시스템을 설계할 때 매우 중요한 문제이다.

4. 정보 추출 패턴의 학습

4.1 말뭉치 기반 학습 기법

자연언어처리에서 사용되는 말뭉치기반 학습 방법론은 일반적으로 정보 추출 시스템을 구성하는 핵심 요소 기술을 개발하거나 성능을 향상시키는데 사용된다. 통계적 방법론이 주로 사용되는 요소 기술에는 품사 태깅(part-of-speech tagging), 의미 클래스 태깅(semantic class tagging), 어의 중의성 해소(word-sense disambiguation), 명칭 인식(named entity identification)[30], 부분 구문 분

석(partial parsing), 추출 패턴 학습(extraction pattern learning), 대용어 참조 해소(coreference resolution), 템플릿 생성(template generation) 등이 있다.

말뭉치 기반 학습은 충분히 많은 양의 학습 데이터를 필요로 하며, 학습 알고리즘은 많은 예제들로부터 특정 기능을 수행하는 언어처리 규칙을 습득하여 예제에서 발견되지 않은 경우에 대해 적용될 수 있도록 규칙을 일반화시킨다. 따라서 학습 알고리즘은 학습에 필요한 정보들이 담겨 있는 태그부착 말뭉치(tagged corpus)를 필요로 한다. 문서 유형에 종속되지 않는 범용 정보 추출 시스템은 구문 정보가 표시된 균형 말뭉치(balanced corpus)를 기반으로 구축되는데 영어 정보 추출 시스템은 주로 Penn Tree Bank를 균형 말뭉치로 사용하고 있다[31]. Penn Tree Bank의 Wall Street Journal 말뭉치는 품사 태깅 및 구문 태깅이 되어 있다.

통계적 학습 기법으로는 HMM(Hidden Markov Model)과 Charniak(1993)의 통계적 파싱 기법[32], Brill(1995)의 변형 기반 학습[33], Ramshaw(1995)의 문장 분할(sentence bracketing) 기법[34], Magerman(1995)의 결정 트리 모델(decision tree model)[35], Daelemans(1996)와 Cardie(1993)의 case-based 학습 기법[36,37], Zelle(1994)의 구문학습을 위한 연역 논리 프로그래밍[38] 등이 있다. 이와 같이 통계적 기법을 이용한 품사 태거와 문장 분할 기법은 학습 말뭉치와 유사한 문서들에 대한 정보 추출 시스템을 구축할 때 유용하다. 어의 중의성 해소의 경우에 현재의 학습 말뭉치는 소수의 단어들에 대해서만 의미 클래스가 정의되어 있는데 의미 클래스를 표준화하기는 쉽지 않다. 이러한 제약으로 인하여 어의 중의성 해소가 정보 추출 시스템의 성능을 향상시키는데 어느 정도 기여할 것인지는 불분명하다.

통계적 학습 기법을 추출 패턴의 학습이나 대용어 참조 해소(coreference resolution), 템플릿 생성 등 자연언어 분석 이외의 정보 추출 기능에 적용하기는 쉽지 않다. 그 이유는 의미태그 및 문서 유형에 적합한 학습 정보가 표시된 말뭉치가 구축되어 있지 않기 때문이다. 즉, 정보 추출 시스템을 구축하려면 출력 템플릿과 해답(answer key)이 부착된 말뭉치가 필요하며, 정보 추출 태스크가 달라지면 새로운 말뭉치를 구축해야 하는 문제점이 있다.

또한, 정보 추출용 말뭉치는 태깅, 대용어 참조 해소, 템플릿 생성 등 정보 추출 시스템의 각 모듈을 개발하는데 필요한 정보들이 충분하지가 않다. 특히, 문서에서 문자열이 중복 출현하는 경우에 어떤 문자열을 추출하고 어떤 표시를 해야 하는지 등 모호한 경우가 발생한다. 뿐만 아니라, 정보 추출용 학습 말뭉치는 정보의 유형이라든지 문서내의 불필요한 요소들을 학습하는 방법을 제시하지 않는다. 이러한 문제를 해결하기 위하여 학습 말뭉치를 각자 구축하기도 하는데, 이 말뭉치는 그 크기가 작아서 통계적 기법을 적용하여 학습하는데 충분치 않다.

정보 추출 시스템은 특정 분야의 문서집합에 대한 의미 분석 등 언어 이해 기술이 요구되는데 품사 태깅 및 구문분석기는 정보 추출에 부적합한 결과를 생성하기도 한다. 따라서 정보 추출 시스템의 성능을 향상시키기 위해서는 품사 태깅이나 구문분석 단계에서 발생하는 오류를 효율적으로 처리하는 방안 및 정보 추출에 적합한 결과를 생성하도록 성능을 개선해야 할 필요가 있다. 이러한 문제점에도 불구하고 말뭉치에 기반한 통계적 방법론이 사용되는 이유는 정확도가 높고 이식성이 좋기 때문이다. 특정 분야에 대한 정보 추출 시스템이 구축되면 타 분야에 대한 정보 추출 시스템은 그 분야에 대한 학습 말뭉치를 구축하여 재학습이 필요한 모듈들을 다시 학습시켜 새로운 주제에 대한 정보 추출 시스템으로 재구축하게 된다.

4.2 패턴 학습 알고리즘

문장분석 결과를 이용하여 정보를 추출하는 단계에서는 패턴일치 기법(pattern matching technique)이 사용된다. 특정 주제에 관한 정보를 추출하기 위한 추출 패턴(extraction pattern)들은 말뭉치로부터 자동으로 학습되는데 추출 패턴들은 정확한 정보들을 추출할 수 있도록 하기 위해 범용성(generality)과 부적합한 정보의 추출을 방지할 수 있는 구체성(specialty)을 겸비해야 한다. 이러한 목적으로 말뭉치 기반 통계적 기법을 사용하고 있으나, 구체적인 학습 방법론은 패턴의 유형, 학습 말뭉치, 사람의 개입 정도, 전처리 과정, 지식베이스의 활용 정도에 따라 다르게 구현된다[39,40].

학습 말뭉치로부터 패턴을 추출하는 기법으로 Riloff(1993)는 AUTOSLOG를 제안하였는데[41],

AUTOSLOG는 Cardie(1991)의 CIRCUS 파서를 이용하여 개념노드(concept-node)라고 정의되는 추출 패턴들을 학습하는 시스템이다[41,42,43]. 이 개념노드는 격 프레임(case frame)에 최대 1개의 슬롯(slot)을 갖는 domain-specific semantic case frame 형태이다. 예를 들어, "Witnesses confirm that the twister occurred without warning at approximately 7:15 p.m. and destroyed two mobile homes."로부터 "two mobile homes"를 추출하는데 사용되는 개념노드의 예는 다음과 같다.

<개념노드의 예>

Concept = Damage
Trigger = "destroyed"
Position = direct-object
Constraints = physical-object
Enabling Conditions = active-voice

이 개념노드는 "destroyed"가 능동태(active-voice)로 출현하는 문장에서 직접 목적어(direct-object)를 damage라는 개념으로 추출한다. 이 개념노드에서 concept 속성은 생성 템플릿에 채워야 할 개념의 유형이고, trigger는 개념노드가 적용될지를 판단하는 역할을 하는 단어이다. position은 주어, 목적어, 혹은 전치사의 목적어 등과 같이 문장 내에서 추출되는 정보의 위치이다. constraints는 개념을 선택할 때 적용되는 의미 선택 제약(selectional restriction)으로 2가지 형태의 제약조건이 있다. hard constraints는 개념이 포함된 구/절을 추출할 때 반드시 만족되어야 하는 조건으로 이를 위반하면 정보 추출 대상에서 제외된다. 그러나 soft constraints는 개념이 포함된 구/절을 추천하는 제약조건으로 이를 위반하더라도 정보 추출의 대상이 될 수 있다. enabling conditions는 패턴이 적용되기 전에 trigger word가 만족해야 할 조건이다.

개념노드를 이용하여 정보를 추출하는 과정은 다음과 같다. 구문분석기가 입력문장을 파싱하여 enabling 조건을 만족하는 trigger word가 발견되면, 해당 구문으로부터 구문 요소(syntactic constituent)를 추출한다. 이 때, 구문 요소가 제약조건을 만족하는 경우에만 개념노드의 concept에 대한 정보가 추출된다.

AUTOSLOG는 학습 알고리즘에 의해 개념노드들을 학습하는데, 학습에 사용되는 말뭉치는 answer key가 부착된 문서집합이다. 또한, AUTOSLOG는 개념노드를 학습하기 위하여 부분 구문분석기(partial parser)와 의미 클래스 사전, 그리고 범용 언어 패턴들을 이용하고 있다. AUTOSLOG의 개념노드 학습 방법은 다음과 같다.

(1) 학습 말뭉치에서 명사구(answer key)를 포함하는 문장을 검색한다.
(2) 검색된 문장에 대해 부분 구문분석을 한다.
(3) 구문분석 결과에 범용 언어 패턴을 적용하여 명사구의 위치 및 역할을 인식한다.
(4) 개념노드(추출 패턴)를 생성한다.

AUTOSLOG를 테러 분야에 적용한 실험 결과에 의하면, 수동으로 패턴을 추출했을 때 약 1500시간이 소요되던 작업을 5시간으로 줄일 수 있었다. 그리고 추출된 패턴의 정확도는 수작업으로 추출된 패턴과 비교했을 때 98%이다.

대부분의 정보 추출 시스템에서 추출 패턴을 학습하는 방법은 AUTOSLOG에서 사용되는 학습 알고리즘과 매우 유사하다. Rillof(1996)는 태그가 부착되지 않는 학습 말뭉치(unannotated corpus)로부터 패턴을 학습하는 AUTOSLOG_TS를 제안하였다[45,46,47]. AUTOSLOG_TS에서는 answer key가 표시되지 않은 대신에 각 문서들이 특정 분야에 적합한지, 그렇지 않은지를 구분하였으며, 학습된 개념노드의 유형을 수동으로 할당하였다.

Kim(1995)은 PALKA 시스템에서 AUTOSLOG와 유사한 형태의 개념노드를 추출하는 방법을 제안하였는데[48], 범용 언어 패턴 대신에 각 패턴을 적용할 때 사용되는 trigger word들을 개념구조(concept hierarchy) 형태로 구축하였다. 또한, Soderland(1995)는 CRYSTAL 시스템에서 의미격 구조(semantic case frame) 형태의 패턴을 학습하는 방법을 제안하였으며[49], CRYSTAL의 패턴은 triggering 제약조건을 좀 더 구체적으로 기술하는 방식을 취하고 있다.

이외에도 Huffman(1996)의 LIEP 시스템은 출력 템플릿의 두 slot filler 명사구 사이의 의미관계를 인식하여 패턴을 학습하였고[50], Cardie(1993)는 기호 학습 알고리즘(symbolic machine-learning algorithm)을 이용하여 trigger word를

인식하였다[37]. 또한, Califf(1997)는 관계 학습(relational learning) 기법을 이용하여 뉴스 그룹의 구인 문서로부터 패턴을 추출하는 방법을 제안하였다[51]. 이러한 패턴 학습 알고리즘들은 동일한 문서집합에 대해 동일한 조건하에서 실험하기가 어려우므로 그 성능을 객관적으로 비교-평가하기는 쉽지 않다.

5. 대용어 참조 해소

대용어 참조 해소(coreference resolution)는 "빌 게이츠"가 "William H. Gates", "Mr. Gates", "William Gates", "Bill Gates", "Mr. Bill H. Gates"와 같이 다양한 형태로 표현되었을 때 동일 개념에 대한 용어들의 관계를 인식하는 것이다[52,53,54,55]. 예를 들어, "Bill Gates is the richest man in the world.... Many people in the software industry fear and respect the guy."에서 "the richest man in the world"와 "the guy"는 모두 "Bill Gates"를 가리킨다. 그런데 "the guy"는 앞에 출현한 대용어에 의존적인 반면에 "the richest man in the world"는 이와 무관하다는 점에 차이가 있다.

대용어 참조 해소는 MUC-6에서 대용어 참조 태스크가 추가되었으나 그 전에도 정보 추출 시스템이 정보를 추출하기 위해 그 기능이 구현되어 왔다. 대용어 참조 해소의 대상이 되는 용어로는 고유명사(names), 별칭(aliases), 동격명사(appositives), 특정 명사를 가리키는 명사구(definite noun phrases), 대명사(pronouns), 서술 명사(predicate nominals) 등이 있다. 대용어 해소의 대상이 되는 용어들의 예는 아래 문장에서 대괄호로 둘러 싼 용어이다.

[Motor Vehicles International Corp.] announced a major management shakeup.... [MVI] said the chief executive officer has resigned.... [The Big 10 auto maker] is attempting to regain market share.... [It] will announce significant losses for the fourth quarter.... A [company] spokesman to Mexico in a cost-saving effort.... [MVI, [the first company to announce such a move since the passage of the new international trade agreement],] is facing increasing demands from unionized workers.... [Motor Vehicles International] is [the biggest American auto exporter to Latin America].

대용어 참조 해소 기법으로는 경험규칙에 의한 방법과 통계적 방법이 사용된다. 경험규칙에 의한 방법은 referent가 앞 부분에 출현한다는 전제조건에 의해 앞 부분에 출현하는 대용어 후보들을 수집하여 성(gender), 수(number) 등의 제약조건을 만족하는 후보들 중에서 경험적으로 가능성이 높은 후보를 선택하는 규칙을 적용하여 referent를 선택한다. 그런데 경험규칙에 의한 방법은 여러 유형의 대용어 관계를 규칙화하는 것이 쉽지 않을 뿐만 아니라, 품사 태깅이나 구문분석, 담화분석 등 문서 분석 과정에서 발생하는 오류로 인해 정확도를 향상시키는데 제약이 있다.

이에 비해, 통계적 기법은 학습 말뭉치로부터 대용어 관계를 결정하는 규칙을 학습하는 접근 방법이다. 그 예로서, Aone(1995)의 MLR(machine-learning-based resolver)과 McCarthy(1995)의 RESOLVE는 통계적 학습 알고리즘에 의해 학습 말뭉치에 표시된 대용어 관계 정보로부터 임의의 두 용어들에 대한 대용어 관계 여부를 좌우 문맥정보와 함께 추출하여 결정 트리(decision tree)로 구현하는 결정 트리 추론 시스템이다[56,57]. 학습 말뭉치의 대용어 참조 정보는 학습 알고리즘에 따라 다르지만, 시스템이 대용어 관계를 구별할 수 있는 <속성, 값>쌍 형태의 특성 정보와 대용어가 출현한 문맥으로 구성된다. MLR의 경우에 학습 예제들은 66개의 특성 정보들로 기술되어 있다.

두 시스템을 각각 50개 문서집합과 250개 문서집합에 대해 실험한 결과에 따르면, RESOLVE는 재현율-정확율이 80%-85%, 87%-92%로 나타났으며 MLR의 경우는 67%-70%, 83%-88%였다. 두 가지 실험결과는 결정 트리의 절단 여부에 따른 실험결과이다. 그런데 MUC-6의 대용어 태스크에서 25개 문서집합에 대해 학습한 실험 결과에 의하면, RESOLVE는 재현율이 41%에서 44%, 정확률이 51%에서 59%이다. 이 결과는 재현율이 51%에서 63%, 정확률이 62%에서 72%인 상위 5개 시스템보다 낮으며, 이 시스템들은 모두 대용어 해결 알고리즘을 수동으로 작성한 것이다.

6. 템플릿 생성

정보 추출 시스템의 궁극적인 목표는 문서 내용으로부터 가치가 있는 특정 정보들을 추출하는 것으로 어떤 정보를 추출할 것인지는 사용자가 결정한다. 지금까지 개발된 정보 추출 시스템들은 '사실정보(factual information)'를 추출하는데 중점을 두고 있으며, 문서 작성자의 의도라든지 문서 내의 모든 정보들을 추출하는데는 적합하지 않다. 따라서 정보 추출 시스템의 개발자와 사용자는 어떤 정보를 어떤 형태로 추출할 것인지에 대해 결정하고 이에 적합한 형태의 정보 추출 시스템을 설계하여야 한다. 추출된 정보를 저장할 템플릿은 텍스트 문서로부터 추출된다는 특성을 고려하여 정보 표현 방식에 적합한 구조로 설계되며, 또한 템플릿이 정보를 가공하는 시스템의 입력으로 사용될 수 있음을 고려하여 설계되어야 한다.

MUC-6의 '경영권 이동' 태스크의 템플릿은 기관명, 직위, 이유, 'in_and_out' 표시 등으로 구성되는데, 정보 추출 시스템은 경영권 이동과 관련된 사건을 중심으로 정보를 추출한다. 그런데 "홍길동은 A기업을 사직하고 B기업의 사장으로 취임하였다."와 같이 한 문장에 2가지 사건이 발생하는 경우와 두 개 이상의 문장이 하나의 사건을 구성하는 경우가 발생한다. 또한, "홍길동은 최근 5년 동안 C기업의 사장으로 재직하고 있다."의 예와 같이 특정 시점에 누구에게 경영권이 있었는지에 관한 정보를 저장할 필요성도 발생한다. 즉, 템플릿은 '경영권 이동'뿐만 아니라 특정 시점의 경영권 상황에 관한 정보를 표현할 수 있어야 한다.

정보를 추출하는 패턴을 구축하는 방법에는 molecular approach와 atomic approach가 있다. 구체적인 템플릿 구조가 결정되면 신뢰도가 높은 정보를 추출할 수 있는 패턴 규칙들을 구축한다. 패턴 규칙을 구축할 때는 우선 적용 범위가 넓은 보편적인 규칙을 구축한 후에 점차적으로 적용 범위가 좁은 규칙들을 추가하는 접근 방식을 취한다. 이를 molecular approach라고 하며 정확도를 중요시하는 시스템에서 사용된다.

이에 비해, atomic approach는 재현율을 중요시하는 시스템에서 사용하는 방법으로 주제와 관련된 명사구와 술어 관계로부터 부분적인 정보들을 추출하여 이로부터 '통합 연산(unification operation)'에 의해 부분적인 정보를 통합하여 전체적인 정보를 생성하고 불필요한 정보들은 제거한다. 부분적인 정보들을 통합하여 전체적인 정보를 생성하는 정보통합에 관한 연구는 많지 않으나, 수동으로 데이터를 분석하여 통합규칙을 발견하는 방법과 자동으로 통합규칙을 학습하는 방법이 있다.

현재까지 정보 추출 시스템은 특정 분야의 문서들에 대해 특정 정보를 추출하는 기능으로 구현되어 왔다. 따라서 사용자가 다른 분야의 문서에 적용하여 원하는 항목의 정보를 추출하려면 그 용도에 적합하게 새로운 시스템으로 재구성하여야 한다. 따라서 정보 추출 시스템들은 새로운 분야에 대한 적응성(adaptability) 및 이식성(portability)을 고려하여 구축되고 있다. 즉, 정보 추출에 사용되는 규칙들을 특정 분야에 종속적인 규칙(domain-dependent rules)과 문서 유형과 무관하게 적용되는 범용 규칙(domain-independent rules)으로 구분하여 구축한다. 범용 규칙은 다양한 구문 구조를 포괄하는 형태의 규칙이고, 분야 종속적 규칙은 해당 정보를 추출하는데 필요한 어휘로서 술어(동사, 형용사)를 중심으로 기술되는 규칙이다.

7. 결 론

정보 추출의 연구방향은 경험적 방법과 학습 알고리즘에 의한 방법으로 구분되는데, 정보 추출에 학습 알고리즘을 적용하는 것은 더욱 새로운 접근 방식이다. 정보 추출 시스템의 성능을 향상시키기 위해서는 다음과 같은 새로운 접근 방법에 대한 연구가 필요하다. 첫째, 통계적 언어 학습의 경향처럼 비 지도 학습 알고리즘을 탐구하는 것이다. 왜냐하면, 일반적으로 학습 알고리즘의 다양함에 비해 효율적으로 학습하는데 필요한 학습 데이터의 크기가 작기 때문이다.

둘째, 조금 다른 방향의 접근 방법은 최종 사용자들이 NLP 시스템 개발자들의 도움없이 사용자와 정보 추출 시스템간의 상호작용을 통하여 정보 추출 시스템을 훈련시키는 기술의 개발이다. 그리고 이러한 작업이 성공하기 위해서는 새로운 학습 기법들이 요구된다. 셋째, 현재 학습 알고리즘의 견고성 및 범용성과 관련하여 사건들 사이의 시간적, 인과관계 등 복잡한 관계를 포함할 수 있도록 정보 추출의 개념을 확장시키는 방안이 탐구되어야 한다.

정보 추출 시스템은 텍스트가 온라인화 됨에 따라 산업체, 정부, 교육 등 사회 각 분야에서 그 필요성이 가속화되고 있다. 자연언어 분석 기술을 발전시키고 정보 추출 시스템의 정확도를 개선하는 학습 알고리즘을 개발함으로써 이러한 요구를 만족시킬 수 있을 것이다. 궁극적으로 정보 추출 시스템은 풍부한 지식베이스와 지식구조, 추론 시스템 등이 요구될 것이다.

미래의 정보 추출 시스템은 보다 개념적인 지식이 요구될 수 있고, 이러한 지식은 최소의 노력으로 얻을 수 있어야 한다. 이런 점에서 가장 보편적인 접근 방법은 현재의 정보 추출 기술 위에서 점진적으로 개발되어야 할 것이다. 정보 추출 기술의 실질적인 면을 지속적으로 탐구하는 것과 자연언어 분석 기술 및 이론을 기반으로, 우리는 지능적인 자연언어 처리 시스템을 구축할 수 있음을 확신할 수 있다.

참고문헌

[1] Appelt, D. E. and David J. Israel, "Introduction to Information Extraction Technology", A Tutorial Prepared for IJCAI-99, 1999.

[2] Yangarber, R., R. Grishman, P. Tapanainen, and S. Huttunen, "Automatic Acquisition of Domain Knowledge for Information Extraction", the 18th International Conference on Computational Linguistics(COLING'2000), pp.940-946, 2000.

[3] Riloff, E., and W. Lehnert, "Information Extraction as a Basic for High-Precision Text Classification", ACM Transactions on Information Systems, vol. 12, no. 3, pp.296-333, 1994.

[4] MUC-3, Proceedings of the Third Message-Understanding Conference (MUC-3), Morgan Kaufmann, 1991.

[5] MUC-4, Proceedings of the Fourth Message-Understanding Conference (MUC-4), Morgan Kaufmann, 1992.

[6] MUC-5, Proceedings of the Fifth Message-Understanding Conference(MUC-5), Morgan Kaufmann, 1994.

[7] MUC-6, Proceedings of the Sixth Message-Understanding Conference (MUC-6), Morgan Kaufmann, 1995.

[8] Chinchor, N., L. Hirschman, and D. Lewis, "Evaluating Message-Understanding Systems", An Analysis of the Third Message-Understanding Conference (MUC-3), Computational Linguistics, vol. 19, no. 3, pp.409-449, 1993.

[9] Glasgow, B., A. Mandell, D. Binney, L. Ghemri, and D. Fisher, "MITA: An Information-Extraction Approach to Analysis of Free-Form Text in Life Insurance Applications", Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence, American Association for Artificial Intelligence, pp.992-999, 1997.

[10] Soderland, S., D. Aronow, D. Fisher, J. Aseltine, and W. Lehnert, "Machine Learning of Text-Analysis Rules for Clinical Records", Technical Report, TE39, Department of Computer Science, University of Massachusetts, 1995.

[11] Lehnert, W. G., "Plot Units and Narrative Summarization", Cognitive Science, vol. 5, no. 4, pp.293-331, 1981.

[12] Craven, M., D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and C. Y. Quek, "Learning to Extract Symbolic Knowledge from the World Wide Web", Internal report, School of Computer Science, Carnegie Mellon University, 1997.

[13] Thompson, C. A., R. J. Mooney, and L. R. Tang, "Learning to Parse Natural Language database Queries into Logical Form", Proceedings of the ML97 Workshop on Automata Induction, Grammatical Inference and Language Acquisition, Association for Computational Linguistics, 1997.

[14] Soderland, S., "Learning to Extract Text-Based Information from the World Wide Web", Proceedings of the Third International Conference on Knowledge Discovery and Data Mining(KDD-97), AAAI Press, pp.251-254, 1997.

[15] Lehnert, W., and B. Sundheim, "A Performance Evaluation of Text Analysis Technologies", AI Magazine vol. 12, no. 3, pp.81-94, 1991.

[16] Will, C. A., "Comparing Human and Machine Performance for Natural Language Information Extraction: Results from the TIPSTER Text Evaluation", Proceedings, TIPSTER Text Program (Phase I), Morgan Kaufmann, pp.179-194, 1993.

[17] Cardie, C., "Empirical Methods in Information Extraction", AAAI-97, pp.65-79, 1997.

[18] Appelt, D. E., Jerry R. Hobbs, John Bear, David Israel, and Mabry Tyson, "FASTUS: A finite-state processor for information extraction from real-world text", In Proceedings of the 13th International Joint Conference on AI, 1993.

[19] Grishman, R., "The NYU system for MUC-6 or where's syntax", MUC-6, 1995.

[20] Aone, C., Lauren Halverson, Tom Hampton, and Mila Ramos-Santacruz, "SRA: description of the IE system used for MUC-7", MUC-7, 1998.

[21] Riloff, E., "Information Extraction as a Stepping Stone toward Story Understanding", Understanding Language Understanding: Computational Models of Reading, MIT Press, 1999.

[22] Dahlgren, K., Carol Lord, Hajime Wada, Joyce McDowell, and Edward P. Stabler Jr. "ITP: Description of the Interpretext System as Used for MUC-3", Proceedings of the Third Message Understanding Conference (MUC-3), pp.163-170, Morgan Kaufmann, 1991.

[23] Grishman, R., John Sterling, and Catherine Macleod, "New York University: Description of the Proteus System as Used for MUC-3", Proceedings of the Third Message Understanding Conference (MUC-3), Morgan Kaufmann, pp.183-190, 1991.

[24] Montgomery, C. A., Bonnie Glover Stalls, Robert S. Belvin, and Robert E. Stumerger, "Language Systems, Inc.: Description of the DBG System as Used for MUC-3", Proceedings of the Third Message Understanding Conference (MUC-3), Morgan Kaufmann, pp.171-177, 1991.

[25] Lin, D., "University of Manitoba: Description of the NUBA System as Used for MUC-5", MUC-5, 1993.

[26] Soderland, S. and W. Lehnert, "Wrap-Up: A trainable discourse module for information extraction", Journal of Artificial Intelligence Research(JAIR), vol. 2, pp.131-158, 1994.

[27] Soderland, S., and W. Lehnert, "Corpus-Driven Knowledge Acquisition for Discourse Analysis", Proceedings of the Twelfth National Conference on Artificial Intelligence, American Association for Artificial intelligence, pp.827-832, 1994.

[28] Lenat, D. B., M. Prakash, and M. Shepherd, "CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge-Acquisition Bottlenecks", AI Magazine, vol. 6, pp.65-85, 1986.

[29] Lehnert, W., "Symbolic-Subsymbolic Sentence Analysis: Exploiting the Best of Two Worlds", In Advances in Connectionist and Neural Computation Theory, pp.135-164, 1990.

[30] Sekine, S., and Y. Eriguchi, "Japanese Named Entity Extraction Evaluation - Analysis of Results", the 18th International Conference on Computational Linguistics (COLING'2000), pp.1106-1110, 2000.

[31] Marcus, M., M. Marchinkiewicz, and B. Santorini, "Building a Large Annotated Corpus of English: the Penn Tree Bank", Computational Linguistics vol. 19, no. 2, pp.313-330, 1993.

[32] Charniak, E., "Statistical Language Learning", MIT Press, 1993.

[33] Brill, E., "Transformation-Based Error-Driven Learning ad Natural Language Processing: A Case Study in Part-of-Speech Tagging", Computational Linguistics vol. 21, no. 4, pp.543-565, 1995.

[34] Ramshaw, L. A., and M. P. Marcus, "Text Chunking Using Transformation-Based Learning", Proceedings of the Thirty-Third Annual Meeting of the ACL, Association for Computational Linguistics, pp.82-94, 1995.

[35] Magerman, D. M., "Statistical Decision-Tree Models for Parsing", Proceedings of the Thirty Third Annual Meeting of the ACL, Association for Computational Linguistics, pp.276-283, 1995.

[36] Daelemans, W., J. Zavrel, P. Berck, and S. Gillis, "A Memory-Based Part-of-Speech tagger Generator", Proceedings of the Fourth Workshop on Very Large Corpora, ACL SIGDAT, pp.14-27, 1996.

[37] Cardie, C., "A Case-Based Approach to Knowledge Acquisition for Domain-Specific Sentence Analysis", Processings of the Eleventh National Conference on Artificial Intelligence, American Association for Artificial Intelligence, pp.798-803, 1993.

[38] Zelle, J., and R. Mooney, "Inducing Deterministic Prolog Parsers from Tree Banks: A Machine-Learning Approach", Proceedings of the Twelfth National Conference on Artifical Intelligence, American Association for Artificial Intelligence, pp.748-753, 1994.

[39] Quinlan, J. R., "c4.5: Programs for Machine Learning", Morgan Kaufmann, 1992.

[40] Weischedel, R., D. Ayuso, S. Boisen, H. Fox, T. Matsukawa, C. Papageorgiou, D. MacLaughlin, T. Sakai, H. J. H. Abe, Y. Miyamoto, and S. Miller, "BBN's PLUM Probabilistic Language-Understanding System", Proceedings, TIPSTER Text Program(Phase I), Morgan Kaufmann, pp.195-208, 1993.

[41] Riloff, E., "Automatically Constructing a Dictionary for Information-Extraction Tasks", Proceedings of the Eleventh National Conference on Artificial Intelligence, American Association for Artificial Intelligence, pp.811-816, 1993.

[42] Cardie, C., and W. Lehnert, "A Cognitively Plausible Approach to Understanding Complicated Syntax", Proceedings of the Ninth National Conference on Artificial Intelligence, American Association for Artificial Intelligence, pp.117-124, 1991.

[43] Lehnert, W., C. Cardie, D. Fisher, E. Riloff, and R. Williams, "Description of the CIRCUS System as Used in MUC-3", Proceedings of the Third Message-Understanding Conference (MUC-3), Morgan Kaufmann, pp.223-233, 1991.

[44] Lehnert, W., C. Cardie, D. Fisher, J. McCarthy, E. Riloff, and S. Soderland, "Description of the CIRCUS System as Used in MUC-4", Proceedings of the Fourth Message-Understanding Conference(MUC-4), Morgan Kaufmann, pp.282-288, 1992.

[45] Riloff, E., and J. Shoen, "Automatically Acquiring Conceptual Patterns Without an Annotated Corpus", Proceedings of the Third Workshop on Very Large Corpora, pp.148-161, 1995.

[46] Riloff. E., "Automatically Generating Extraction Patterns from Untagged Text", Proceedings of the thirteenth National Conference on Artificial Intelligence, American Association for Artificial Intelligence, pp.1044-1049, 1996.

[47] Riloff, E., "An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains", Artificial Intelligence, vol. 85, pp.101-134, 1996.

[48] Kim, J. T. and D. I. Moldovan, "Acquisition of Lingistic Patterns for Knowledge-Based Information Extraction", IEEE Transactions on Knowledge and Data Engineering, vol. 7, no. 5, pp.713-724, 1995.

[49] Soderland, S., D. Fisher, J. Aseltine, and W. Lehnert, "CRYSTAL: Inducing a Conceptual Dictionary", Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, pp.1314-1319, 1995.

[50] Huffman, S., "Learning Information-Extraction Patterns from Examples", In Symbolic, Connectionist, and Statistical Approaches to Learning for Natural Language Processing, Lecture Notes in Artificial Intelligence Series, pp.246-260, 1996.

[51] Califf, M. E., and R. J. Mooney, "Relational Learning of Pattern-Match Rules for Information Extraction", Proceedings of the ACL Workshop on Natural Language Learning, Association for Computational Linguistics, pp.9-15, 1997.

[52] Cardie, C., "Corpus-Based Acquisition of Relative Pronoun Disambiguation Heuristics", Proceedings of the Thirtieth Annual Meeting of the ACL, Association for Computational Linguistics, pp.216-233, 1992.

[53] Cardie, C., "Learning to Disambiguate Relative Pronouns", Proceedings of the Tenth National Conference on Artificial Intelligence, American Association for Artificial Intelligence, pp.38-43, 1992.

[54] Kehler, A., "Probabilistic Coreference in Information Extraction", Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp.163-173, 1997.

[55] Weischedel, R., M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmucci, "Coping with Ambiguity and unknown Words through Probabilistic Models", Computational Linguistics vol. 19, no. 2, pp.359-382, 1993.

[56] Aone, C., and W. Bennett, "Evaluation Automated and Manual Acquisition of Anaphora Resolution Strategies", Proceedings of the Thirty-Third Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp.122-129, 1995.

[57] McCarthy, J. F., and W. G. Lehnert, "Using Decision trees for Coreference Resolution", Proceedings of the Fourteenth International Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, pp.1050-1055, 1995.

[58] Appelt, D. E., Jerry R. Hobbs, John Bear, David Israel, M. Kameyama, A. Kehler, D. Martin, K. Myers, and Mabry Tyson, "SRI International FASTUS System: MUC-6 Test Results and Analysis",

Proceedings of the Sixth Message-Understanding Conference(MUC-6), Morgan Kaufmann, pp.237-248, 1995.

[59] Dolan, C., S. Goldman, T. Cuda, and A. Nakamura, "Hughes Trainable Text Skimmer: Description of the TTS System as Used for MUC-3", Proceedings of the third Message-Understanding Conference (MUC-3), Morgan Kaufmann, pp.155-162, 1991.

[60] Holowczak, R. D., and N. R. Adam, "Information Extraction-Based Multiple-Category Document Classification for the Global Legal Information Network", Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence, American Association for Artificial Intelligence, pp.1013-1018, 1997.

[61] Hastings, P., and S. Lytinen, "The Ups and Downs of Lexical Acquisition", Proceedings of the Twelfth National Conference on Artificial Intelligence, AAAI Press/The MIT Press, pp.754-759, 1994.

강 승 식

1986 서울대학교 컴퓨터공학과(학사)
1988 서울대학교 컴퓨터공학과(석사)
1993 서울대학교 컴퓨터공학과(박사)
1994~2001 한성대학교 정보전산학부
2001~현재 국민대학교 컴퓨터학부
관심분야:형태소분석, 구문분석, 정보검색, 정보추출, 기계번역
E-mail:sskang@kookmin.ac.kr

우 종 우

1978 서울대학교 농생물학과(학사)
1983 미국 Minnesotta State University at Mankato 전산학과(석사)
1991 미국 Illinois Institute of Technology 전산학과(박사)
1992~1993 국방정보체계연구소 선임연구원
1994~현재 국민대학교 컴퓨터학부
관심분야:지능형교육시스템(Intelligent Tutoring System), 에이전트, 정보추출, 전문가시스템
E-mail:cwwoo@kookmin.ac.kr

윤 보 현

1992 목포대학교 전산통계학과(학사)
1995 고려대학교 컴퓨터학과(석사)
1999 고려대학교 컴퓨터공학과(박사)
2001~현재 한국전자통신연구원 선임연구원 언어이해연구팀 팀장
관심분야:정보검색, 자연언어 처리, XML/SGML, 지식정보처리
E-mail:ybh@etri.re.kr

박 상 규

1982 서울대학교 컴퓨터공학과(학사)
1984 KAIST 전산학과(석사)
1997 KAIST 전산학과(박사)
1984~1987 대림산업(주) 전산실
1987~현재 ETRI 책임연구원, 언어공학연구부 부장
관심분야:정보검색, 정보추출, 기계번역, 지능형 에이전트, 바이오인포메틱스
E-mail:skpark@computer.etri.re.kr