# 자연언어처리를 위한 기계학습

부산대학교 ▌정성원* · 권혁철**

## 1. 서 론

**자연언어처리란 컴퓨터로 인간의 언어를 분석하여, 이를 바탕으로 응용 프로그램을 구현하는 전 과정을 뜻한다.** 인간이 언어를 어떻게 습득하고 처리하며 이해하는지를 알아내고자 하는 시도는 오래전부터 있었다. 이러한 노력의 결과로 사람들은 언어의 구조를 일련의 규칙으로 표현하기 시작했으며, 바른 언어 표현과 바르지 않은 언어 표현을 구분하고자 문법(grammar)을 만들었다. 하지만, '모든 문법은 불완전하다[1]' 라는 유명한 경구에서도 알 수 있듯이 바른 표현과 바르지 않은 표현을 정확하고 완전하게 구분하는 문법을 만들기는 불가능하다. 또한, 문법을 통하여 언어를 해석하면 규칙에 맞는 언어적 현상은 잘 해석되지만, 언어의 다양성과 변화 때문에 계속해서 생기는 예외적인 현상을 능동적으로 반영하지 못한다. 이에 따라 사람들은 현재 언어를 잘 설명하는 문법이나 일반적인 규칙을 찾는 대신에 언어에서 사용되는 일반적인 패턴이 무엇이며, 이를 어떻게 효과적으로 찾을 수 있을지에 관심을 두기 시작하였다. **본 논문은 이와 같은 배경에서 기계학습과 자연언어처리의 관계를 살펴보고, 자연언어처리의 단계별로 적용된 기계학습과 관련한 최신 연구 동향에 대해서 알아보도록 하겠다.**

## 2. 자연언어처리 단계와 기계학습

**자연언어처리는 여러 단계를 거치면서 다양한 문제에 직면하게 되는데, 그중 중의성 문제는 기계학습을 적용하기에 가장 적합하다.** 중의성은 자연언어처리의 여러 단계에서 전반적으로 나타나는 다양한 선택의 문제이며, 언어 이해를 위해 해결해야 할 가장 어려운 문제들을 대부분 포함한다.

**중의성 문제의 대표적인 예로는 문장 분석에서의 품사 중의성, 음성 인식에서의 단어 선택, 다의어의 의미적 중의성, 구문 분석에서의 구문 구조의 중의성, 엑센트 복원, 대명사의 참조 선택, 기계 번역에서의 단어 선택, 문맥 의존적 맞춤법 교정 등이 있다.** 그림 1은 일반적인 자연어처리 단계와 단계별로 해결해야 하는 중요한 중의성 관련 문제들을 정리한 것이다.

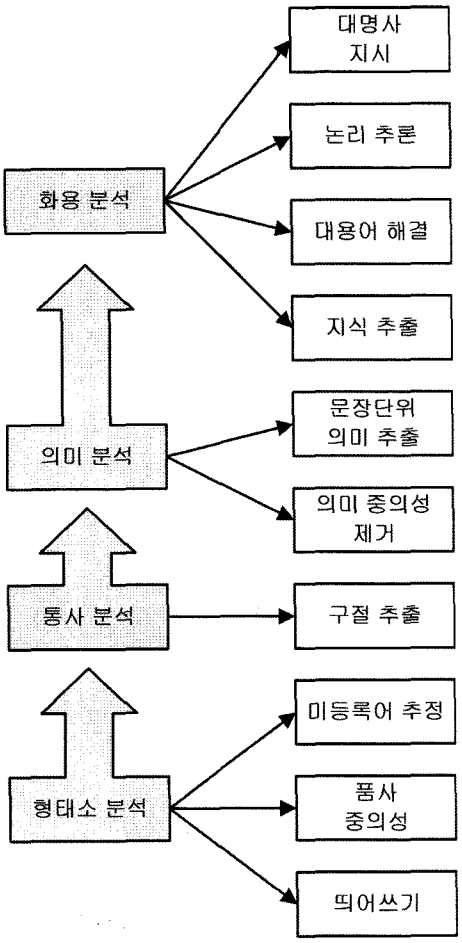


그림 1 자연언어처리의 단계별 중의성 문제

이러한 중의성 문제 해결을 위한 일반적인 처리 과정은 다음과 같다. 먼저, 자연언어처리 시스템은 태깅이나 구문 분석과 같이 자연언어처리의 어떤 문제를 해결하는 과정에 여러 가지로 해석될 수 있는 부분을

* 학생회원
** 종신회원

만나게 된다. 예를 들어, 형태소 분석에서는 어떤 단어가 여러 가지 품사로 해석될 수 있다. 이때, 해당 문맥에 가장 적절한 해석(태깅에서는 품사 선택)을 해야 한다. 이는 인간이 자연언어 문장을 처리할 때 순간적으로 겪게 되는 혼돈과 같다. 예를 들어 다음과 같은 문장을 보자.

'나는 오리 고기를 먹었다.'

이 문장에서 '나는'을 '인칭대명사+조사'로 분석하여 '나+는'으로 해석해야 할지 동사 '날다'의 활용형으로 '날다+는'으로 해석해야 할지는 '오리'까지만 읽었을 때는 판단을 내릴 수 없다. 하지만 '고기를'이라는 어절을 보고 '나는'을 '인칭대명사+조사'로 분석할 수 있다. 사람이 이렇게 판단하는 이유는 선택을 요하는 부분의 주변 문맥을 보고 자연언어의 모호성을 없애는 단서를 인지하기 때문이다. 자연언어처리에 적용되는 기계학습의 기본적인 방법도 이와 다르지 않다. **기계학습에서는 판단을 내리고자 하는 부분의 주변에 존재하면서 선택에 도움이 되는 '특성'이나 '패턴'을 여러 기법으로 추출한 후, 이를 이용하여 형태론적, 통사구조적, 어휘의미적 구별을 하게 된다.**

**자연언어처리에서의 이와 같은 중의성 문제는 기계학습에서 분류의 문제로 해석되며, 이렇게 해석된 문제는 전통적인 여러 기계학습 기법(사례 기반 학습, 결정 트리, 선형 분리, 비감독 클러스터링, 부스팅, 강화 학습 등)을 적용할 수 있다.** 하지만, 기계학습을 자연언어처리에 실제 적용해보면 자연언어의 고유한 특성 때문에 나타나는 다양한 유형의 어려움에 부딪히게 된다. 이를테면, 자연언어처리에서 해결하고자 문제는 거대한 자질 공간을 다루기 때문에 아주 많은 양의 학습 데이터가 필요하며, 매우 드물게 나타나는 자질을 처리해야 하는 일도 빈번하고, 학습에 관련이 없는 자질이나 잉여자질이 데이터 중에 많이 포함되어 있기도 하다. 해결하려는 문제에 따라서는 학습 데이터 자체를 구하기 어려워 비감독학습이나 준 감독학습을 적용해야 할 경우도 있으며, 간단한 분류의 문제로 해석되기 어려운 경우도 있다. 이에 따라, **기계학습 분야에서도 잘 연구되지 않았던 경험적인 학습 방법들, 예들 들어 변형기반 오류에 의한 학습, 최대 엔트로피 등도 자연언어처리를 위해 도입되었으며, 이는 다시 기계학습 분야의 발전에 이바지하고 있다.**

## 3. 자연언어처리에 기계학습이 도입된 배경

기계학습이 자연어처리에 활발하게 응용되기 시작한 것은 그다지 오래되지 않았다. **1990년대 초에 기계학습 기법이 자연언어처리에 적용되기 시작한 이후 현재까지 기계학습기법을 이용한 자연언어처리가 활발히 연구되고 있다.** 이와 같은 연구의 배경에는 다음과 같은 여러 요인이 있다[2].

- 지속적으로 어휘가 추가되고 오류가 많으면서 개체명이나 전문용어가 다양하게 쓰이는 언어현상을 규칙에 의한 접근만으로는 따라잡을 수 없다는 것을 깨달았다.
- 다양한 분야에서 다양한 목적의 표식(tag)을 부착한, 기계가 읽을 수 있는 말뭉치가 여러 언어로 만들어졌다.
- 소프트웨어와 하드웨어의 성능이 향상되면서 과거에는 다룰 수 없었던 대용량의 자료와 많은 시간이 드는 알고리즘들을 처리 가능한 비용과 시간 내에서 적용할 수 있게 되었다.
- 기초적인 수준의 언어적 문제를 통계적인 기법을 사용하여 성공적으로 해결함으로써 기계학습 방법이 자연언어처리에 적용 가능하다는 것을 확인하였다.
- 자연언어처리의 응용 분야가 확대됨에 따라 다양한 응용 목적에 개별화한 시스템이 필요하게 되었다.

이에 따라 **자연언어 습득과 이해 문제(품사 태깅, 어휘 의미중의성 해소, 문법 추론, 강건한 파싱, 정보추출과 검색, 자동 요약, 기계 번역 등)에 기계학습 기법을 적용하여 큰 효과를 거두었다. 또한, 세계 각국은 기계학습에 필요한 대용량 말뭉치를 응용 목적에 따라 다양하게 확보했으며, 그 결과 기계학습 기법의 개선도 같이 이루어졌다.**

## 4. 기계학습과 말뭉치

말뭉치는 사용 목적에 따라 다양한 방법으로 구성된다. 가장 간단한 말뭉치는 단어의 빈도, 단어의 공기 정보 등 언어에서 발견되는 여러 가지 통계를 추출하고자 일반 텍스트를 모아서 구성한다. 이때, 텍스트의 띄어쓰기나 맞춤법 교정 등의 정제 과정을 거치기도 하며, 다양한 분야에서 텍스트를 추출하여 균형적으로 구성하기도 한다. 대표적인 원시 말뭉치로 국외에는 Brown Corpus, Wall Street Journal Corpus 등이 있으며, 국내에서는 문화관광부의 21세기 세종계획에 의한 세종 말뭉치, 연세대학교의 언어정보개발연구원에 의한 연세 한국어 말뭉치, 고려대학교 민족문화연구소의 고려대 한국어 말모둠, 한국과학기술원의 과기원 코퍼스, 국립국어연구원의 국립국어연구

원말뭉치 등이 있다. 또 다른 구성 방법으로 말뭉치를 병렬적으로 구성하기도 한다. 예를 들면 우리말 대 북한말, 방언 대 표준말 등이 있을 수 있으며, 둘 이상의 다국어 텍스트를 정렬하여 구성하기도 한다.

이러한 말뭉치에 어떤 특별한 목적에 의해서 추가의 표식을 부착하기도 하며 이를 특별히 표식이 부착된 말뭉치(Annotated Corpus)라 일컫는다. **표식이란 기계학습을 위하여 어떤 분석 정보를 말뭉치에 덧붙여 놓은 것을 말하며, 해결하려는 문제에 따라서 다양한 표식이 부착된다.** 예를 들어 품사중의성을 해소하고자 할 경우에는 품사 태그를 부착하며, 어휘 의미 중의성을 해소하고자 할 때는 어휘 의미 태그, 구조적 중의성을 해소하고자 할 때는 구문 분석 결과 태그를 부착할 수 있다. 이와 같은 **표식이 부착된 말뭉치 중 가장 대표적인 것으로는 Wall Street Journal Corpus, Switchboard Corpus, ATIS Pilot Corpus, Brown Corpus 등의 말뭉치에 표식을 붙여 놓은 Penn Treebank[1]가 있다. 또 다른 예로 국외에서는 The Lanscater-Leeds Treebank, The Associated Press Treebank 등이 있으며, 국내에서는 한국전자통신연구원의 품사 부착 말뭉치 및 구문구조 부착 말뭉치, 21세기 세종계획에 의한 형태소 분석 말뭉치, 구문 분석 말뭉치, 어휘 의미 분석 말뭉치, 한국과학기술원의 품사 부착 말뭉치 등이 있다.** 이와 같은 말뭉치가 없었던 시기에는 연구자 개인이 연구의 목적에 맞게 말뭉치를 구축하여 학습에 사용하였으나 현재는 다양한 말뭉치가 구축되어 있으므로 연구의 객관성 확보와 다른 연구와의 비교를 위하여 이러한 말뭉치를 적극적으로 활용하는 것이 좋다.

## 5. 기계학습을 이용한 자연언어처리의 최근 연구 동향

앞장까지 자연언어처리와 기계학습의 전반적인 특징을 살펴보았다. 이 장에서는 자연언어처리 관련 연구에서 기계학습이 어떻게 응용되었는지 알아보겠다. 자연언어처리 분야에서는 기계학습을 위하여 일반적으로 앞장에서 설명한 말뭉치를 사용한다. 말뭉치는 크게 표식이 부착되어있는 말뭉치와 표식이 부착되어있지 않은 말뭉치로 나눌 수 있으며, 이에 따라 감독학습과 비감독학습으로 나뉜다.

### 5.1 감독학습

감독학습과 비감독학습은 학습에 사용하는 자료의 특성이 다르다. 감독학습은 목적에 따라 학습 대상이 되는 자연언어의 자질이나 구조에 대한 표식이 부착되어 있는 말뭉치를 사용하며, 표식이 명시되어 있는 말뭉치 중 공인되어 널리 쓰이는 말뭉치는 기계학습에서 학습과 평가를 위한 표준(gold standard)이 된다. **감독학습은 크게 기호적 학습(symbolic learning)과 통계 기반 학습(statistical-based learning)으로 나눌 수 있다. 기호적 학습 방법은 표식이 부착된 말뭉치로부터 규칙을 추출하거나 분류절차를 학습한 후, 표식이 붙지 않은 말뭉치에 적용하며, 대표적인 방법으로는 결정 트리(decision tree), 변형 기반 오류에 의한 학습(transformation-based error driven learning), 선형 분리기(linear separator), 사례 기반 학습(instance-based learning) 등이 있다. 통계 기반 학습 방법은 학습 말뭉치 내에서 확률 변수(속성)를 추출하고 여러 문맥에 걸쳐 관찰되는 확률 변수들 간의 확률 분포를 결정하는 모델을 구축하며, 대표적으로 나이브 베이스 분류기(naive Bayes classifier), 최대 엔트로피(maximum entropy principle), 마코프 모델(Markov model) 등이 있다.**

**감독학습은 앞서 기술한 여러 가지 자연언어의 중의성 문제 중 품사 태깅 문제의 해결에 가장 성공적으로 적용되었다.** 이 분야의 대표적인 연구로써 Church (1988)[3]와 Charniak(1993)[4]의 마코프 모델을 이용한 통계적 품사 태깅, Brill(1992[5], 1994[6])의 변형 기반 오류에 의한 학습을 적용한 변형 기반 태깅을 들 수 있다. 이들 연구 이후 현재의 여러 감독학습 기반 태깅 관련 연구들의 정확도는 Penn Treebank와 같은 태그가 부착된 말뭉치에 대해 평가했을 때 96% 정도이다. 국내에서는 이상호(1992)[7], 이운재(1992)[8], 김진동(1996)[9], 김재훈(1996)[10], 이상조(2000)[11] 등이 마코프 모델을 한국어의 특성에 맞게 변형한 방법을 사용하였으며, 강미영(2007)[12]은 한국어의 어절과 어절 간의 전이 관계를 형태소 uni-gram을 이용하여 추정하는 방법을 제안하였다. 이들 연구 결과에 의하면 한국어를 대상으로 한 태깅 정확도는 95% 전후의 성능을 보인다.

**확률적 구문 분석도 감독학습이 성과를 보인 영역 중의 하나로 Charniak(1997)[13]과 Collins(1998)[14]의 연구가 대표적이다.** 이들은 확률자유문맥문법(PCFG)을 제안하였는데, Penn Tree Bank로부터 확률 값이 명기된 문맥자유문법(CFG) 규칙을 추출한 것이다. Charniak의 PCFG는 어휘적 하위범주화 자질을 포함하고 있으며, Collins의 문법은 논항(argument)-보충어(adjunct)를 구분한다. 이 두 연구는 월스트리트저널과

1) http://www.cis.upenn.edu/~treebank/

Penn Treebank에 대해 90%에 달하는 재현율과 정확도를 얻었다. 국내에서는 김형근(1995)[15]이 처음 확률적 구문 분석을 한국에서 적용하였으며, 이후 최선화(2002)[16]가 KAIST의 코퍼스를 이용하여 한국어에 대한 확률 의존문법 자동 생성 기술을 연구하였다.

**감독학습 기반 구문 분석과 관련한 또 다른 연구로 Clark(2004)[17]이 제안한 최대 엔트로피 통계적 결합범주문법(Maximum Entropy Statistical Combinatory Categorical Grammar)이 있다.** 결합범주문법(CCG) 구문 분석기는 무한의존구문(unbounded dependencies)과 국부적 기능-논항(function-argument) 구조를 표현하고 있으며, CCG 어휘 태그 및 어휘 의존구조가 부착된 말뭉치를 이용하여 학습한다. 의존구조는 자질로서 표현되며 시스템은 표식이 달린 표준 말뭉치로부터 가장 확률이 높은 의존구조들의 집합으로 표현되는 모델을 계산하였다.

이와 유사한 기법은 의미표현에도 사용되었다. Bos(2004)[18]는 통계적 CCG의 구문 분석 구조에 논리적인 형태로 의미표현을 부여하는 방법을 제안하였다. 이러한 논리적인 형태의 의미 표현은 다양한 형태의 문장에 적용하는데 한계가 있고, 높은 차원의 수량 한정사 등을 다루지 못하지만, 실제 문장에 적용될 가능성을 제시했다는데 그 의의를 찾을 수 있다.

**감독학습은 생략된 구문에 대한 문제를 해결하는데도 효율적으로 이용되었다.** Nielsen(2003)[19]은 품사 자질이 명기된 문맥에 기초하여 생략된 '동사구'를 판별해 내기 위해, 변형 기반 오류에 의한 학습과 최대엔트로피 기법을 적용하였다.

**감독학습은 이외에도 귀납적 논리 프로그래밍(Inductive Logic Programming)을 이용하여 텍스트로부터 특정 분야의 특성을 학습하는데도 사용되었으며**(Liakata(2004)[20]), 일상 대화 중에 자주 나타나는 완전하지 않은 문장(non-sentential utterances)을 해석하기 위한 학습(Fernández(2005)[21])에도 사용되었다. 국내에서는 한국어의 특성상 발생하는 띄어쓰기 문제에 감독학습 기법이 널리 쓰였다. 심광섭(1996)[22]과 강승식(2001)[23]은 통계적 기법을 사용하였으며, 이도길(2003)[24]은 한국어의 음절 특성을 마코프 모델에 적용하였고, 강미영(2006)[25]은 범주 패턴을 이용하여 어절을 추정하는 기법을 사용하였다.

지금까지 살펴본 바와 같이 **기계학습 중 감독학습은 자연언어처리의 광범위한 영역에 성공적으로 적용되었다. 표식이 부착된 말뭉치를 이용한 기호적, 통계적인 학습 방법은 빠르고 효율적으로 자연언어의 구조와 의미에 관한 지식을 습득할 수 있다.** 그러나 이러한 학습 방법은 학습해야 할 정보가 학습을 위한 데이터에 미리 명시적으로 표현되어 있어야 함을 전제로 한다.

### 5.2 비감독 학습

**비감독 학습에 쓰이는 학습 말뭉치는 학습하고자 하는 언어적 자질이나 구조에 대한 표식이 부착되어 있지 않다. 비감독 학습은 비슷한 요소들의 집단을 인식함으로써 그 분포와 집단의 패턴을 밝히는 작업이며, 클러스터링이 가장 대표적이다.** 비감독학습 방법을 이용하여 언어적 구조와 내용을 성공적으로 획득함으로써 귀납적 방법이나 투사와 같은 일반적인 인지 기법을 결합하여 언어적 구조나 규칙을 추측할 수 있다는 가정이 힘을 얻게 되었다.

GoldSmith(2001)[26]는 유럽 언어를 대상으로 단어를 형태소로 분해하기 위하여 최소기술길이(Minimal Description Length: MDL)를 사용하였다. 말뭉치 내의 단어를 어간과 접미사(suffix)로 구분하기 위하여 경험규칙을 확률적으로 적용하며, MDL을 이용하여 경험규칙의 적용 여부를 결정한다. 이 연구는 500,000개의 영어 단어로 이루어진 말뭉치에서 추출한 1,000개의 단어를 알파벳 순으로 정렬한 테스트 집합을 대상으로 82.9%의 형태소분석 정확도를 실험 결과로 얻었다.

Schone(2001)[27]은 접미사뿐만 아니라 일반적인 접사의 3가지 형태(suffix, prefix, circumfix)로 형태소 분석 대상을 확대했다. Schone은 제안한 알고리즘을 영어(6,700,000단어), 독일어(2,300,000단어), 네덜란드어(6,700,000단어)를 대상으로 테스트하였으며, 영어 접미사에 대해서 88.1% F-score[2]을 얻어 Goldsmith의 시스템보다 나은 성능을 보였으며, 독일어에서 92.3%(Goldsmith는 84%), 네덜란드어에서 85.8%(GoldSmith는 75.8%)의 성능을 보였다.

**Cutting(1992)[28]은 Baum-Welch 알고리즘을 이용한 비감독학습을 통해 은닉 마코프 모델의 파라메터를 획득하는 품사 태깅 모델을 구성하였다.** Brown corpus의 절반을 이용하여 학습한 후 품사 태깅된 나머지 절반을 이용하여 평가하였으며, 96%의 품사태깅 정확도를 얻었다. 국내에서는 임철수(1994)[29]와 김재훈(1995)[30]이 한국어 품사 태깅에 비감독학습을 사용하여 파라메터를 추정한 은닉 마코프 모델을 적용하였으며, 약 90%의 정확도를 얻었다.

**Clark(2000)[31]는 클러스터링을 통하여 어휘 통사범주를 유도하는 방법을 제안하였다.** 이 방법에서 생

2) 정확률(P)과 재현율(R)의 조화평균, 즉, 2PR/(P+R)

성된 품사 태그 집합은 British National Corpus(BNC)에서 사용하는 CLAWS 품사 태그 집합에 비견될 만했다. Clark의 연구에서는 태깅을 위해 확률적 유한 상태 모델을 만들고 각 품사 태그 집합을 적용해 본 결과 CLAWS의 품사 태그 집합보다 비감독학습에 의해서 얻어진 태그 집합이 낮은 혼잡도(perplexity)를 보였는데, 이는 비감독학습의 결과 태그들이 좀 더 우수한 분포 특성을 보인다는 뜻이다. 따라서, 비감독 품사 태깅은 어느 정도 신뢰할만하며, 비감독 문법 습득을 위한 기초를 마련했다고 볼 수 있다. 하지만, 이 연구를 응용한 통계적 문맥 자유 문법을 학습하는 비감독 시스템(Clark(2001)[32])은 ATIS 말뭉치[3]에서 41%의 F-score을 얻는데 그쳤다.

**Klein(2002)[33]은 품사가 부착된 입력 데이터를 이용하여 트리 내부에 문장 구성요소로서 품사가 부착된 요소의 연속을 두며, 이를 위한 확률 값을 부여함으로써 문장 구성요소의 구조를 비감독학습하는 문법 추론 시스템을 제안하였다.** 이 시스템은 모든 문장을 이진 트리로 분석하며, 가장 적합한 트리 구조를 선택하기 위해 EM알고리즘을 사용하였다. 또한, Klein(2004)[34]은 어휘적인 head 의존 문법을 비감독학습하는 확률 모델을 제안했다. 이 시스템은 문장 내 각 단어의 왼쪽이나 오른쪽에 나타나는 단어 연속들을 논항이나 보충어로 취하여 각 단어가 head가 될 가능성을 추정하여 문장 구성요소의 구조에 확률을 부여한다. 확률들은 head가 나타나는 문맥, 즉 그 head의 양쪽에 인접하여 나타나는 단어나 단어의 집합으로 구성된 문맥을 기반으로 계산된다. Klein의 문장 구성요소 구조와 같이 의존관계 구조 모델 또한 이진 트리로 만들어진다. 이와 같이 비감독학습기반 문법 추론 연구는 획득 가능한 최소의 언어 범주와 규칙 가정의 설정을 기반으로 하였으며, 일반적인 기계학습 방법을 이용하여 언어 지식이 취득될 수 있다는 것을 보여주고 있다.

앞선 연구들은 기계학습을 이용하여 인간이 문법을 학습하는 방식에 대한 잠재적 모형을 유추할 수 있다는 것을 전제로 한다. 이에 대해 비판론자들은 기계학습 시스템이 오직 문법적으로 올바른 문장들만으로 이루어진 말뭉치를 이용해서 학습하기 때문에 올바른 문장만을 인식할 수 있다는 한계점이 있다고 주장한다. 하지만, **기계학습 시스템이 만들어 내는 구문 분석기는 문법적인 문장과 비문법적인 문장을 구분할 수 있으며, 문법적 · 비문법적 구분이 있는 문장의 구조 또한 판별할 수 있다.** 또한, 문장 분석이 실패한 경우 에러 메시지를 제공하게 할 수도 있다. 이와 같이, 비록 한정되어 있기는 하지만, 기계학습을 통해 언어적 구조나 규칙을 추측할 수 있다는 가능성을 확인했다는 것은 중요하다.

3) Air Travel Information System(ATIS) pilot corpus, http://www.ldc.upenn.edu/Catalog/docs/LDC93S4B/corpus.html

## 6. 결 론

**인터넷이 보급되고 자연언어처리의 응용 분야가 확대됨에 따라, 사람들은 다양한 영역에서 응용 목적에 맞는 시스템을 빠른 시간에 개발하고자 하였으며, 그에 따라 자연언어처리에 기계학습을 적극적으로 응용하기 시작하였다.** 이는 기존 언어학적 연구의 바탕 위에 언어학에서 보완하지 못한 언어의 다양성에 능동적으로 대처하기 위한 노력이며, 컴퓨터의 성능 향상, 새로운 알고리즘의 개발, 대용량 말뭉치의 구축과 함께 언제나 얻을 수 있는 대용량의 언어 자원을 가진 인터넷의 발달에 기인하고 있다. 더불어 자연언어처리의 여러 가지 난해한 문제를 기계학습으로 풀어냄으로써 기계학습 분야 자체의 발전도 일구어 내었다. **앞으로도 기존 언어학적인 연구를 바탕으로 기계학습을 적용하려는 노력은 계속될 것이며, 기계학습을 이용하여 인간의 언어에 대한 새로운 사실들을 알아낸다면, 기존 연구자들이 밝혀내지 못했던 새로운 이론으로 기존 이론을 보완할 수 있을 것이다.**

## 참고문헌


[1] Sapir, E. Language: an introduction to the study of speech. New York: Harcourt Brace, 1921.

[2] Màrquez, L., Machine learning and natural language processing. Technical Report LSI-00-45-R, Departament de Llenguatges i Sistemes Informàtics(LSI), Universitat Politècnica de Catalunya(UPC), Barcelona, Spain, 2000.

[3] Church, K., "A stochastic parts program and noun phrase parser for unrestricted text", Second Conference on Applied Natural Language Processing, Austin, TX, pp. 136-143, 1988.

[4] Charniak, E., Hendrickson, C., Jacobson, N., and Perkowitz, M., "Equations for Part-of-Speech Tagging", Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-93), pp. 784-789, 1993.

[5] Brill, E., "A simple rule-based part of speech tagger", Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy, pp. 152- 155, 1992.
[6] Brill, E., "Some advances in transformation-based part of speech tagging", Proceedings of the Twelfth National Conference on Artificial Intelligence, pp. 722- 727, 1994.
[7] 이상호, 미등록어를 고려한 한국어 품사 태깅 시스템 구현, 한국과학기술원 전산학과 석사학위논문, 1992.
[8] 이운재, 한국어 문서 태깅 시스템의 설계 및 구현, 한국과학기술원 전산학과 석사학위논문, 1992.
[9] 김진동, 어절 문맥을 고려하는 형태소 단위의 한국어 품사 태깅 모델, 고려대학교, 컴퓨터학과, 석사학위논문, 1996.
[10] 김재훈, 오류-보정 기법을 이용한 어휘 모호성 해소, 한국과학기술원, 전산학과, 박사학위논문, 1996.
[11] Lee, S.Z., Tsujii, J.I., and Rim, H.C., "Hidden Markov Model-based Korean Part-of-speech Tagging Considering High Agglutinativity, Word-Spacing, and Lexical Correlativity", Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp. 384-391, 2000.
[12] 강미영, 한국어의 형태·통사적 특징을 고려한 범주 기반 가변 n-gram 품사 태깅 모델, 부산대학교, 컴퓨터공학과, 박사학위논문, 2007.
[13] Charniak, E., "Statistical parsing with context-free grammar and word statistics", Proceedings of the Fourteenth National Conference on Artificial Intelligence, pp. 598-603, 1997.
[14] Collins, M., Head-Driven Statistical Models for Natural Language Parsing, PhD thesis, University of Pennsylvania, 1998.
[15] 김형근, 확률적 의존문법과 한국어 구문분석, 한국과학기술원, 전산학과, 석사학위논문, 1995.
[16] 최선화, 박혁로, "형태소 단위의 한국어 확률 의존 문법 학습", 한국정보처리학회 논문지B, 9B(6), pp. 791-798, 2002,
[17] Clark, S. and Curran, J., "Parsing the WSJ using CCG and log-linear models", Proceedings of the Forty- Second Meeting of the Association for Computational Linguistics, Barcelona, Spain, pp. 104-111, 2004.
[18] Bos, J., Clark, S., Curran, J., Hockenmaier, J. and Steedman, M., "Widecoverage semantic representations from a CCG parser", Proceedings of COLING 18, Geneva, Switzerland, 2004.
[19] Nielsen, L., "Using machine learning techniques for VPE detection", Proceedings of Recent Advances in Natural Lanuage Processing, Borovets, Bulgaria, pp. 339-346, 2003.
[20] Liakata, M. and Pulman, S., "Learning theories from text", Proceedings of COLING 18, Geneva, Switzerland, 2004.
[21] Fernández, R., Ginzburg, J. and Lappin, S., "Using machine learning for nonsentential utterance classification", Proceedings of the Sixth SIGdial Workshop on Discourse and Dialogue, Lisbon, pp. 77-86, 2005.
[22] 심광섭, "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기", 정보과학회논문지: 소프트웨어 및 응용, 23권, 9호, pp. 991-1000, 1996.
[23] 강승식, "음절 bigram를 이용한 띄어쓰기 오류의 자동 교정", 음성과학회논문지, 8권, 2호, pp. 83-90, 2001.
[24] 이도길, 이상주, 임희석, 임해창, "한글 문장의 자동 띄어쓰기를 위한 두 가지 통계적 모델", 정보과학회논문지: 소프트웨어 및 응용, 30권, 4호, pp. 358- 370, 2003.
[25] 강미영, 정성원, 권혁철, "어절 내의 형태소 범주 패턴에 기반한 통계적 자동 띄어쓰기 시스템", 정보과학회논문지: 소프트웨어 및 응용, 33권, 11호, pp. 193-206, 2006.
[26] Goldsmith, J., Unsupervised learning of the morphology of a natural language, Computational Linguistics 27, 153-198, 2001.
[27] Schone, P. and Jurafsky, D., "Knowledge-free induction of inflectional morphologies", Proceedings of the North American Chapter of the Association for Computational Linguistics(NAACL-2001), Pittsburgh, PA, 2001.
[28] Cutting, D., Kupiec, J., Pedersen J., and Sibun, P., "A Practical Part-of-Speech Tagger", Proceedings of the third conference on Applied natural language processing, pp.133-140, 1992.
[29] 임철수, HMM을 이용한 한국어 품사 태깅 시스템

구현, 한국과학기술원 전산학과 석사학위논문, 1994.
[30] 김재훈, 임철수, 서정연, "은닉 마르코프 모델을 이용한 효율적인 한국어 품사의 태깅", 정보과학회논문지, 22권, 1호, pp.136-146, 1995.
[31] Clark, A., "Inducing syntactic categories by context distribution clustering", Proceedings of CoNLL 2000, Lisbon, Portugal, 2000.
[32] Clark, A., "Unsupervised induction of stochastic context-free grammars using distributional clustering", Proceedings of CoNLL, Toulouse, France, 2001.
[33] Klein, D. and Manning, C., "A generative constituent-context model for improved grammar induction", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 128-135, 2002.
[34] Klein, D. and Manning, C., "Corpus-based induction of syntactic structure: Models of dependency and constituency, Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics, Barcelnoa, Spain, 2004.

**정 성 원**

2001년 부산대학교 전자계산학과(학사)
2003년 부산대학교 전자계산학과(석사)
2006년 부산대학교 컴퓨터공학과(박사)
2006년 9월~현재 부산대학교 U-Port 정보기술산학공동사업단 post-doc
관심분야 : 자연언어처리, 정보추출, 정보검색, 기계학습
E-mail : swjung@pusan.ac.kr

**권 혁 철**

1982년 서울대학교 공과대학 전산학(학사)
1984년 서울대학교 공과대학 전산학(석사)
1987년 서울대학교 공과대학 전산학(박사)
1988년~현재 부산대학교 전자전기정보컴퓨터공학부 교수
1988년~현재 한국어정보과학회 프로그래밍언어 연구회 운영위원
1990년~현재 한국어정보과학회 한국어정보처리연구회 운영위원
1992년~1993년 미국 Stanford 대학 CSLI연구소 연구원
1992년~1993년 Xerox Palo Alto Research Center 자문위원
2004년~현재 한국정보과학회 이사
2006년~현재 한국인지과학회 이사
관심분야 : 자연언어처리, 정보검색, 프로그래밍언어, 인공지능, 시맨틱웹
E-mail : hckwon@pusan.ac.kr