□ 기술해설 □

자연어처리 연구동향 : 통계 기반의 자연어 처리

고려대학교 임해창* · 임희석** · 윤보현**

목 차

1. 서 론

인간이 사용하는 자연어(natural language)를 컴퓨터를 이용하여 처리하고자 하는 자연어처리에 대한 연구는 크게 규칙 기반의 접근법(rule-based approach)과 통계 기반의 접근법(statistics-based approach)으로 나눌 수 있다.

규칙 기반의 접근법은 자연어가 사용될 때 적용되는 일반적인 규칙을 찾아내고, 그 규칙을 이용하여 자연어를 처리하고자 하는 방법을 말하며, AI 기반의 접근법(AI-based approach) 또는 지식 기반의 접근법(knowledge-based approach)이라고도 한다. 규칙 기반의 접근법은 제한된 영역에 대해서는 높은 정확도를 보이나, 복잡한 언어 현상들로부터 규칙을 추출하기가 매우 어렵고, 다른 영역으로의 확장성이 좋지 않다는 단점을 가지고 있다.

통계 기반의 접근법은 사람들이 실제로 사용하는 많은 데이타로부터 확률 정보 및 통계 정보를 추출하고, 이를 이용하여 여러 가지 언어 현상을 규명하고자 하는 접근방법을 말하며, 코퍼스 기반의 접근법(corpus-based approach)이라고도 한다[17,41,42]. 통계 기반의 접근법에서는 언어가 실제 사용된 용례들을 모아둔 대량의 코퍼스로부터 추출한 통계 정보와 확률 정보를 자연어 처리에 사용하므로 확장성이 좋으며, 어떤 영역의 데이타에 대해서도 견고하다는 장점이 있다.

통계 기반의 접근법은 1950년대에 처음 제시되었으나, Chomsky의 N-그램에 대한 비평 이후 그 열기가 점차 줄어들게 되었다. 또한 통계 기반의 접근법은 그 당시 사용 가능한 대량의 데이타가 없었다는 점과 대량의 데이타를 빠른 시간내에 처리할 수 있는 컴퓨터 시스템의 부재로 많은 연구자들로부터 매력을 잃게 되었다.

그러나 최근에 들어서는 ACL/DCI(Association for Computational Linguistics' Data Collection Initiative), ECI(European Corpus Initiative), ICAME, BNC(British National Corpus), LDC(Linguistic Data Consortium), CLR(Consortium for Lexical Reasearch) 등과 같은 기관의 꾸준한 노력으로 대량의 기계 가독형 코퍼스(machine readable corpus)를 얻을 수 있게 되었다[17,29,42]. 또한 컴퓨터 하드웨어의 눈부신 발전에 힘입어 최근에는 통계 기반의 접근법을 이용한 자연어처리에 대한 연구가 다시 부흥하게 되었다. 이러한 부흥에는 "All but Only"로 표현되는 규칙 기반 접근법에 대한 회의적인 반응도 크게 작용하였다.

통계 기반 접근법은 어휘 지식 획득(lexical

*종신회원
**준회원

knowledge acquisition), 품사 태깅(part of speech tagging), 문법 개발(grammar construction), 구문 분석(syntactic analysis), 기계 번역(machine translation) 등 많은 자연어처리 분야의 발전 가능성을 제시하였다[41,42,53].

이에 본 논문은 코퍼스 구축, 품사 태깅, 구문 분석 그리고 기계 번역 분야를 중심으로 통계 기반의 자연어처리 연구 동향에 대하여 살펴보고, 자연어처리 연구의 향후 발전 방향에 대해서 논의하고자 한다.

2. 통계 기반의 자연어 처리

2.1 코퍼스 구축 및 코퍼스 분석 도구의 개발

코퍼스는 인간에 의해서 사용된 자연어의 실제 용례와 그에 따른 여러 가지 언어 정보(linguistic knowledge)를 묶어놓은 기계 가독 형태(machine readable form)의 자료를 말한다[6,21,29,41,50]. 코퍼스는 일반적으로 용례의 종류에 따라 문서 코퍼스(written corpus), 음성 코퍼스(spoken corpus) 그리고 통합형 코퍼스로 분류할 수 있다. 문서 코퍼스는 신문이나 잡지, 소설 등과 같은 문어체에서 추출된 용례를 가지고 있는 코퍼스를 말하며, 음성 코퍼스는 대화, 연설, 뉴스, 연극 대본과 같이 인간의 음성을 통해서 발현된 음성 용례의 묶음을 의미한다. 통합형 코퍼스는 문어체 용례 및 음성 용례를 통합하여 구축한 코퍼스를 의미한다.

문서 코퍼스의 예로는 Brown 코퍼스, LOB 코퍼스 그리고 EDR 코퍼스가 있으며[21,29,58], 음성 코퍼스의 대표적인 예로는 LLC (London-Lund Corpus of British Spoken English)와 SEC (IBMM/Lancaster Spoken English Corpus) 등을 들 수 있다. 통합형 코퍼스의 대표적인 예로는 Birmingham 코퍼스가 있다[50].

Brown 코퍼스는 미국 Brown 대학에서 만들어진 것으로서 미국식 영어 문체에서 추출된 약 100만개의 단어로 구성되어 있다. LOB 코퍼스는 Lancaster 대학에서 개발된 것으로서 영국식 영어라는 점을 제외하고는 Brown 코퍼스와 구성이 같다. Brown 코퍼스와 LOB 코퍼스는 그 자체로 이용되기 보다는 새로운 코퍼스 제작에 많은 도움을 주는 역할을 하고 있다. EDR 코퍼스는 EDR 전자 사전 개발을 위하여 구축된 것으로서 현재 일본어와 영어에 대해 각각 2,000만개의 문장으로 구성되어 있는 방대한 크기의 코퍼스이다[21]. LLC는 약 50만 단어의 영국식 영어 음성 용례와 강세 억양 등의 운율 주석이 아주 자세히 설명되어 있는 코퍼스이다. SEC는 영어 음성 용례로 구성된 코퍼스로서 기본적인 태깅 정보뿐 아니라 음성 신호, 운율 주석 등 여러 정보를 가지고 있어서 상당히 유용하다. Birmingham 코퍼스는 처음에 COBUILD 사전 개발을 위하여 Birmingham 대학과 Collins publishers에 의해 1960년대부터 구축되기 시작하였고, 현재에는 약 2억개 정도의 단어로 구성되어 있는 통합형 코퍼스이다.

코퍼스는 통계 기반의 접근법에서 이용하는 확률 정보와 통계 정보 추출을 위해 반드시 필요한 자료로서, 최근에는 위에서 설명한 코퍼스 이외에도 효율적인 언어 처리를 위한 코퍼스 구축이 활발히 진행되고 있다. 이러한 연구의 대표적인 예로는 Penn Treebank와 EDR 코퍼스를 들 수 있다.

Penn Treebank는 품사가 태깅된 4백 5십만 단어로 구성되어 있고, 이 중 2/3가 braket을 이용하여 구문 주석이 달려있는 코퍼스이다[40]. Penn Treebank처럼 구문 구조가 기술되어 있는 코퍼스는 정확한 파싱과 정보 검색 분야의 구단위 색인(phrase indexing) 등 많은 연구 분야에서 효율적으로 사용될 수 있다. 최근 Penn 단체는 코퍼스의 세부적인 주석화에 많은 흥미를 가지고, predicate-argument 구조로 이루어진 3백만 단어 treebank를 구축하려고 하고 있다[40,42].

또한 최근에는 문장의 구문 구조와 아울러 단어간의 개념 관계(concept relation)를 기술하는 의미 정보 태깅에 대한 연구도 매우 활발한데, 그 대표적인 예로는 EDR 코퍼스를 들 수 있다[21]. EDR 코퍼스는 전체 4,000만개의 문장 중에서 50만개의 문장을 선정하여 이들 문장에 대해서는 기본적인 언어 정보 뿐만 아니라 구문 구조 및 의미 정보를 부여하였다.

최근에는 한 가지 언어에 대한 단일어 코퍼스

(mono-lingual corpus) 뿐만 아니라, 두 가지 또는 여러 가지 언어들의 용례와 언어 정보를 병렬적으로 저장하여 이를 이용하기 위한 양국어 코퍼스(bi-lingual corpus) 및 다국어 코퍼스(multi-lingual corpus)에 대한 연구에도 많은 관심이 집중되고 있다[11,21,22,53]. 양국어 코퍼스와 다국어 코퍼스의 개발은 기계 번역, 어휘 지식 획득(lexical knowledge acquisition)과 사전 편찬과 같은 자연어처리 분야에서 뿐만 아니라 외국어 교육 분야를 위한 도구로도 유용하게 사용될 수 있다[21,22,52,53]. 기계 번역에서는 이미 번역되어 있는 많은 양의 예를 분석하여 다음 번역 작업에 유용하게 사용될 수 있는 통계 정보 및 확률 정보를 추출한다. 어휘 지식 획득 및 사전 편찬을 위해서는 양국어 또는 다국어 코퍼스를 병렬적으로 구성하고, 여기서 원하는 정보를 자동적으로 추출한다.

대량의 코퍼스로부터 유용한 정보를 추출하여, 이용하는 통계 기반의 접근법에서는 코퍼스 분석 도구가 반드시 필요하다. 현재까지 개발되어 있는 코퍼스 분석 도구의 예로는 Church와 Hanks의 시스템[16], Xtract[51], INTEXT[49]와 MULTEXT[28] 등을 들 수 있다. Chruch와 Hanks는 코퍼스로부터 단어간의 상호 정보(mutual information)를 이용하여 연관된 단어들의 쌍을 찾고자 하였다. XTRACT는 연어 정보(collocation information)를 찾기 위하여 단일 구문 요소에 있는 단어 쌍들의 관련성을 조사하였고, 적절하지 못한 단어 쌍을 제거하기 위하여 구문 정보를 사용한다는 특징을 가지고 있다. 1000만 단어의 코퍼스를 이용하여 실험한 결과 XTRACT는 80%의 정확도를 보였다. INTEXT는 대량의 데이타로부터 특정한 단어 유형(lexical pattern), 구문 유형(syntactic pattern) 그리고 용례(concordance) 등 다양한 언어 정보 추출을 위한 코퍼스 분석기로서 사용자 자신의 사전과 문법을 추가할 수 있는 장점을 가지고 있다. MULTEXT 프로젝트는 다국어 문서 코퍼스를 구축하고 이를 분석하는 데 사용될 수 있는 도구 개발을 위해 수행되고 있으며, 크게 품사 태거, 후처리기 등 문서 주석을 위한 도구와 용례 색인/추출기, 통계 처리기 등 텍스트 분석 도구 개발로 나누어 수행되고 있다. MULTEXT 연구 결과 얻게 될 모든 데이타와 도구들은 무료로 공개될 예정이다.

2.2 품사 태깅(Part-of-Speech tagging)

품사 태깅은 문장내의 각 단어에 대한 품사 정보를 결정하는 작업을 말한다. 대부분의 품사 태깅 시스템의 처리 단계는 품사 중의성 탐색 단계(part-of-speech ambiguity detection)와 품사 중의성 해소(part-of-speech amibuity resolution) 단계로 세분 할 수 있다.

품사 중의성 탐색 단계는 형태소 해석에 의해서 주어진 문장 내에 나타나는 모든 단어의 가능한 품사를 찾는 단계이고 품사 중의성 해소 단계는 중의성 해결을 위한 규칙 또는 어휘 확률, 문맥 확률 등 확률 정보를 이용하여 정확한 태그를 결정하는 단계이다.

일반적으로 품사 태깅 시스템은 품사 모호성 해소 단계에서 어떠한 정보를 이용하느냐에 따라 규칙 기반 품사 태깅[2,9,23,24,27,33,41]과 통계 기반 품사 태깅으로 구분할 수 있다[17,18,20,34,37].

규칙 기반 품사 태깅에 관한 연구의 예로는 Klein과 Simmons의 시스템[33], TAGGIT[23] 그리고 Brill의 연구[7,8,9]를 들 수 있다. Klein과 Simmons의 시스템은 품사가 발생하는 조건을 문맥 틀 규칙(context frame rule)으로 기술하고, 이를 이용하여 품사를 태깅하고자 한 시스템으로 표본 데이타에 대해서 90%의 정확도를 보였다. TAGGIT은 86가지의 태그 집합과 3300개의 규칙을 사용한 시스템으로서 정확도는 76% 정도이나, 표본 데이타가 아닌 대량의 데이타를 가지고 실험한 첫번째 시도였다는데 의의가 있다. Brill의 시스템은 규칙 추출이 어렵고 다른 영역으로의 적용성(portability)이 좋지 않다는 규칙 기반 방법의 한계를 극복한 시스템으로서 대량의 코퍼스로부터 중의성 해결을 위한 규칙을 자동으로 학습할 수 있도록 한 시스템이다. 이 시스템의 정확도는 중의성 해결을 위해 사용되는 patch의 수에 따라 다르게 나타나는데, 71개의 patch를 사용하여 Brown 코퍼스에 대해 실험한

결과 94.9%의 정확도를 보였다.

일반적으로 규칙 기반의 방법은 표본 데이타에 대해서 높은 정확도를 보이나 규칙 추출이 어렵고, 확장성이 없다는 문제점을 가지고 있다. 또한 실제 적용되는 규칙 중 단순한 규칙이 대부분을 차지하는 비합리성을 가지고 있다. 이러한 문제점을 해결하기 위해 인접 태그 간의 관계와 통계를 고려하는 통계 기반 접근법이 대두되었다.

통계 기반 품사 태깅은 근본적으로 입력으로 주어진 문장 S에 대하여 확률 정보를 이용하여 다음 식을 만족하는 태그열 T를 찾는 것이라 할 수 있다.

$$\phi(S)=\arg\max_t P(T|S)=\arg\max_t \frac{P(T,S)}{P(S)}$$
$$=\arg\max_t P(T,S)$$

통계 기반 품사 태깅의 장점은 코퍼스로부터 모호성 해소에 필요한 정보를 손쉽게 얻을 수 있다는 점과 처리 영역에 제한 되지 않는 견고성이 있다는 점이다. 그러나 확률 정보를 저장하기 위한 기억 공간이 많이 필요하고 태거를 구현, 수정이 어렵다는 단점이 있다.

최근 품사 태깅을 위해서 많이 이용되고 있는 통계 기반 접근법은 품사 간의 단어 확률을 이용하는 방법, 은닉 마코프 모델(Hidden Markov Model)을 이용하는 방법 그리고 신경망(neural network)을 이용하는 방법으로 나눌 수 있다.

품사간의 단어 확률을 이용하는 방법은 통계적 기법을 이용한 시스템으로 CLAWS 시스템[6, 41], VOLSUNGA 시스템[20] 등을 들 수 있다. CLAWS 시스템은 태깅된 Brown 코퍼스로부터 특정 품사 태그들 간의 동시 발생 확률 정보(probability information)를 추출하여, 이를 품사 태깅에 사용하였으며, 96~97%의 정확도를 보였다. 최근에는 언어 정보를 사용하여 태깅의 일관성과 정확도를 향상하고 적응성을 향상시킨 CLAWS4가 개발되기도 하였다[37]. VOLSUNGA 시스템은 CLAWS 시스템을 동적 프로그래밍 기법을 이용하여 시간 복잡도(time complexity)와 공간 복잡도(space complexity)를 선형으로 줄이고자 한 시스템이다.

은닉 마코프 모델을 이용하는 품사 태깅 방법은 베이지언 정리(Bayes' theorem)를 이론적 배경으로 하고, 품사 결정을 위해서 국지 정보(local information)만을 이용한다는 마코프 가정(Markov assumption)을 근간으로 하는 방법이다[18,34,42,47]. 은닉 마코프 모델을 이용하는 많은 모델들은 보통 2개 이상의 문맥 정보를 사용하는데, 품사 태깅을 위한 식은 다음과 같다.

$$\phi(S)=\arg\max \prod_{i=1}^{n} P(t_i|t_{i-k}\ t_{i-1})\ P(w_i|t_i)$$

위 식에서 $P(t_i|t_{i-k}\ t_{i-1})$은 장거리 정보(long distance information) 대신 k 거리만큼의 국지 정보를 이용한다는 마코프 가정을 고려한 품사 전이 확률(transition probability)을 나타내며, $P(w_i|t_i)$는 어휘 관찰 확률을 의미한다.

은닉 마코프 모델에 기초한 최근 연구로는 Kupiec과 Schütz와 Singer 등의 연구를 들 수 있다. Kupiec의 연구는 단어 동일 부류(word equivalence class)를 이용하여 은닉 마코프 모델의 매개 변수를 감소시키고, 미등록어 추정을 위해 국지 문맥과 접미 정보(suffix information)을 이용하고자 한 것이다[18,34]. 이 방법은 어휘 사전과 원시 학습 자료 만을 이용하여 견고하고 정확하게 태깅할 수 있는 방법으로, 96%의 정확도를 보였다[18]. Schütz와 Singer는 오랜 학습 시간을 요구하고, 매개변수의 초기 값에 매우 의존적인 은닉 마코프 모델의 단점을 해결하기 위하여 VMM(Variable Memory Markov) 모델을 이용한 품사 태깅 방법을 제시 하였다[47]. VMM 모델은 학습 데이타에 기초한 history length를 동적으로 적응시키며, 보다 적은 매개변수를 사용할 수 있는 모델이며, Brown 코퍼스로 검증한 결과, VMM 기반 방법은 정확도가 95.81%였다.

신경망을 이용한 태깅 시스템의 예로는 Benello, Mackle, Anderson의 시스템[2]과 Schmid의 Net-Tagger를 들 수 있다[46]. Benello 등에 의해서 개발된 시스템은 236개의 입력 노드와 1개의 은닉층으로 구성된 시스템으로 학습 알고리즘으로는 역전파(backpropagation) 학습 알고리즘을 사용한 시스템이다[2,43]. Schmid의 Net-Tagger는 MLP-network를 이용한 시스템으로

역전파 학습 알고리즘을 이용하여 Penn-Treebank의 2백만 단어에 대해서 학습을 시켰으며, 10만 단어에 대해서 실험한 결과 96.22%의 정확도를 보였다.

2.3 통계 기반 파싱(stochastic parsing)

자연어를 파싱하는 목적은 입력으로 주어진 문장의 구문 구조 혹은 문법 구조를 파악하는 것이다. 여기서 말하는 구문 구조는 문장의 각 요소들 사이의 관계 뿐만 아니라, 그것들이 구문적으로 어떠한 역할을 하는 지를 결정하는 것이다. 이와 같은 작업을 수행하기 위해서 문법, 파싱 알고리즘, 그리고 구문적 중의성 해소 방법이 필요하다. 문법은 어떤 언어에서 사용되는 구조를 조직적으로 기술하기 위한 것이고, 파싱 알고리즘은 문법을 이용하여 문법 구조를 결정하는 분석을 위한 것이다. 구문적 중의성 해소 방법은 구문적 중의성을 갖는 문장이 현재 문맥에서 어떤 분석으로 의도된 것인지 결정할 수 있도록 하는 것이다.

모든 자연어는 중의성을 갖는다. 다시 말하면, 일반적으로 자연어에서는 분석하고자 하는 문장이 하나 이상의 구문 구조를 갖는 경우가 많아서 하나 이상의 의미로 분석되는 경우가 많다는 것이다. 이것은 문법 규칙들이 과생성적이어서 문법을 만든 사람이 의도하지 않았던 구조가 나오거나 또는 문장 구조 자체가 중의적인 경우가 있기 때문이다.

지금까지 구문분석 단계에서 중의성을 해소하고자 하는 연구는 여러 형태의 규칙에 조건을 부가하여 처리하는 규칙 기반 구문분석과 통계 및 확률 정보를 이용해서 구문 구조에 순서 매김하는 통계 기반 구문 분석으로 세분 될 수 있다. 규칙 기반 구문분석은 규칙의 획득과 확장이 어렵고, 구문적 중의성을 갖는 문장을 적절하게 처리할 수 없는 문제점이 있다. 이런 문제점을 극복하기 위해서 최근에는 통계 기반 접근법을 이용한 파싱(stochastic parsing)에 대한 연구가 이루어지고 있다.

통계 기반 구문분석 기법은 1979년 IBM T. J. Watson 연구소의 Baker가 처음으로 PCFG (Probabilistic Context Free Grammar)를 사용함으로써 시작되었다. Baker는 확률적 파서(probabilistic parser)를 자동으로 생성할 수 있는 Inside-Outside 알고리즘을 개발하였다. 이 Inside-Outside 알고리즘은 PCFG의 매개 변수 추정을 위해서 은닉 마코프 모델의 매개 변수를 추정하기 위한 Baum-Welsh 알고리즘을 일반화한 것이다[17,44]. 즉, Baum-Welsh 알고리즘이 반복 작업으로 은닉 마코프 모델의 매개 변수를 개선하는 것처럼, PCFG 매개 변수를 반복적으로 개선시킨다.

Inside-Outside 알고리즘을 단순히 적용했을 때 발생하는 문제점은 효율적인 파서를 만들 수 없다는 것인데, 그 이유는 다음과 같이 두 가지 요인 때문이다[41,42]. 첫째, PCFG을 위해서 추정될 매개변수가 너무 많아서 신뢰성 있는 매개변수를 구하기 어렵다. 둘째, Inside-Outside 알고리즘은 그 특성상 단어의 품사 태깅을 위한 방법이므로 파서의 학습을 위한 최선의 방법이 되지 못한다. 즉, Inside-Outside 알고리즘은 파서가 정확한 문법 구조(grammatical structure)를 선택할 수 있도록 학습시키는 데 적절하지 못하다는 것이다.

위와 같은 문제점을 해결하고, 효율적인 통계 기반 구문분석을 수행하기 위한 연구가 활발하게 진행되고 있는데, 새로운 문법인 어휘화된 문법(lexicalized grammar)을 이용하는 연구와 Inside-Outside 알고리즘을 여러 방법으로 제약하는 연구가 있다.

어휘화된 문법(lexicalized grammar)은 모든 단어가 최소한 한가지의 구문 구조와 관련이 되어 있으며, 이런 구조를 조합할 수 있는 조합 규칙으로 구성되는 문법을 의미한다[30,35,45]. 이런 문법 구성의 목적은 Inside-Outside 알고리즘이 제대로 적용되어 정확한 문법 구조를 찾도록 하자는 것이다.

어휘화된 문법 형식의 대표적인 예는 Combinatory Categorial Grammars(CCGs), Lexicalized Tree Adjoining Grammars(LTAGs)[45], 그리고 Link Grammar[35]을 들 수 있다. 이들 문법의 특징은 모든 단어들이 구문적, 의미적 의존 관계를 나타내는 supertag 또는 tree frag-

ment라 불리는 구문 구조와 연관이 된다는 것이다[30,35,45].

어휘화된 문법을 이용한 파싱 과정은 입력 문장의 각 단어들이 어떤 supertag 또는 tree fragment에 해당하는가를 결정하고, 그것들을 적절하게 결합하는 문제로 볼 수 있다. 그러므로 단어의 품사 태깅을 위해 사용되는 Inside-Outside 알고리즘을 적절히 변형 또는 확장한다면, 어휘화된 문법을 이용하여 효율적이고 정확한 파싱이 가능하게 된다.

Inside-Outside 알고리즘을 제약하기 위한 방법으로 Pereira와 Schabes는 구문 구조가 기술되어 있는 Penn Treebank 코퍼스에 Inside-Outside 알고리즘을 수정하여 적용함으로써 파싱의 정확도를 향상시키고자 하였다[44]. Pereira와 Schabes의 연구는 언어 지식(linguistic knowledge)를 이용하여 Inside-Outside 알고리즘을 제약하기 위한 것으로 코퍼스 내의 구문 구조에 위배되지 않는 PCFG 규칙만을 고려하였다. 언어 지식을 이용하여 Inside-Outside 알고리즘을 제약한 경우 파싱의 정확도는 90% 정도로 단순히 Inside-Outside 알고리즘을 이용한 경우에 비하여 3배나 높았다. 또 다른 제약으로 PCFG을 언어 지식으로 제약하는 방법도 있다. 이 방법은 PCFG에 그 규칙을 적용할 수 있는 부가적인 언어 지식을 사용한 것으로 이 방법을 적용한 시스템의 예로는 Magerman의 구문 분석기를 들 수 있다[38,39].

효율적인 파싱과 정확도 향상을 위한 연구는 위에서 설명한 것 이외에도 통계적인 접근법과 언어 지식을 접목하려는 연구를 들 수 있다[4,5,15]. Brown 대학[15], Pennsylvania 대학, 음성 인식의 결과를 토대로 좋은 결과를 내고 있는 IBM T. J. Waston 연구소[4,5], 그리고 영국에서는 Leeds 대학[48], Cambridge 대학[10], Lancaster 대학 등이 이에 대한 연구를 진행 중에 있다.

2.4 기계 번역(machine translation)

자연어처리 연구의 한 분야인 기계 번역은 컴퓨터를 이용하여 원시 언어를 번역 전문가와 유사하게 목적 언어로 번역하는 작업을 의미한다. 기계 번역을 수행하는 방법은 여러 가지가 있는데, 크게 언어학 기반 기계 번역(LBMT: Linguistics Based Machine Translation)과 통계 기반 기계 번역(SBMT:Statistics Based Machine Translation) 두 가지로 분류할 수 있다.

기계 번역 연구의 초기에는 번역 작업을 위하여 언어학적인 지식(linguistic knowledge)을 이용하는 언어학 기반 기계 번역에 대한 연구가 활발하였다. 그러나, 언어의 무한한 표현 방법을 방대한 지식 베이스(knowledge base)로 구축한다는 일은 거의 불가능하며, 입력 문장과 번역될 문장 사이의 복잡한 의미 관계를 언어학적인 지식으로 표현한다는 것 또한 매우 어려운 일이었다.

최근에 이와 같은 문제점을 극복하기 위한 방법으로 인간의 번역 과정에 대한 경험적인 연구의 필요성이 제기되었고, 인간이 실제 사용한 언어 용례로부터 추출한 확률 및 통계 정보를 이용하고자 하는 통계 기반 기계 번역에 대한 연구가 시도되고 있다[17,21,25,26,52].

통계 기반의 기계 번역에서 원시 언어(s)는 목적 언어(t)의 어떤 문장으로도 번역될 수 있다고 생각된다. 따라서 모든 (s, t)의 쌍에 대해서 입력으로 s가 주어졌을 때 그 번역문이 t가 될 확률값을 나타내는 P(t|s)를 할당할 수 있다. 원시 언어 s를 t로 번역한다는 것은 t를 자국어로 사용하는 사람(native speaker)이 t를 말할 때, 번역어로서 생각하는 s를 찾는 작업으로 생각할 수 있다. 즉, s를 t로 번역한다는 것은 P(s|t)가 최대가 되도록 하는 s'을 찾는 것이라 할 수 있고, 이를 베이지언 이론(Bayes' theorem)을 이용하여 나타내면 다음과 같다.

$$P(s|t)=\frac{P(s)\,P(t|s)}{P(t)}$$

위 식에서 분모 P(t)는 s에 독립적이므로 s'을 찾는 일은 P(s)P(t|s)가 최대값을 갖는 s를 찾는 작업이 된다. 따라서 통계 기반 기계 번역에서 가장 적절한 번역어의 선택을 위해 사용될 확률 계산 식은 다음과 같다[12,13,42].

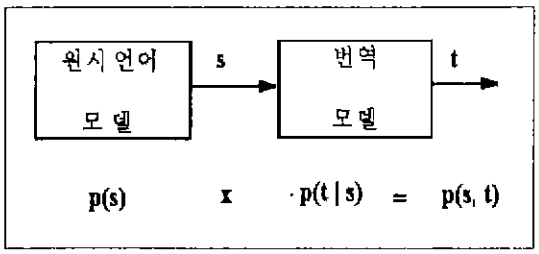


그림 1 통계 기반 기계 번역 시스템

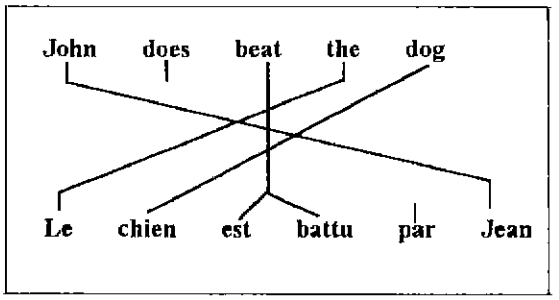


그림 2 복잡한 정렬의 예

$s' = \arg\max P(s)P(t|s)$

아래의 그림 1은 위 식에 근거하여 통계 기반 기계 번역 시스템을 나타낸 것이다.

통계 기반 기계 번역 시스템은 그림 1에서 나타낸 바와 같이 원시 언어 모델(source language model)과 번역 모델(translation model)로 구성된다. 따라서, 통계 기반 기계 번역을 위해서는 원시 언어 모델의 확률 값을 계산하기 위한 방법, 번역 모델의 확률 값을 계산하기 위한 방법 그리고 P(s)P(t|s)을 최대화시키는 s를 찾기 위한 탐색 방법이 필요하다.

원시 언어 모델의 확률 값 계산을 위해서는 주로 N-gram 모델을 사용하며, 번역 모델의 확률 값 P(t|s)을 계산하기 위해서는 원시 언어 s와 목적 언어 t 사이의 정렬(alignment) 정보를 이용한다. s와 t의 정렬 작업은 s의 어떤 단어가 t의 어떤 단어로 번역되었는가를 나타내기 위한 작업이다. 정렬되어 있는 코퍼스는 기계 번역 뿐 아니라 다른 응용 분야를 위한 언어 지식 추출에 많이 사용될 수 있으며, 최근에는 정렬된 코퍼스 구축에 많은 노력을 하고 있다[11,22,32].

"John loves mary"는 프랑스어로 "Jean aime Marie"로 번역되는데, 이것을 정렬하여 나타내면 (Jean aime Marie|John(1) loves(2) Mary(3))와 같다[12]. 여기서 "|"는 두 문장을 분리하는 표시이고, "John(1)"와 같은 "e_t(n)"은 e_t가 n번째의 프랑스 단어로 번역된다는 것을 의미한다. 위의 정렬은 간단한 예를 나타낸 것이며, 다음의 예처럼 복잡한 경우의 정렬도 있을 수 있다.

(Le chien est battu par Jean|John(6) does beat(3,4) the(1) dog(2))

위의 예는 "does"는 앞 프랑스 문장의 어떤 단어와도 대응되지 않으며, "beat"는 "est", "battu"와 같은 두개의 단어와 대응된다는 것을 의미한다. 정렬에서 한 단어가 번역문의 몇개의 단어와 대응되는 가는 나타내는 척도를 산출력(fertility)이라 하는데[12,13], 그림 2에서 "does"의 산출력은 0, "John", "the", "dog"의 산출력은 1이고 "beat"의 산출력은 2가 된다. 이와 같은 대응 관계를 그림으로 나타내면 그림 2와 같다.

번역 모델에서 확률 값 P(t|s)를 계산하기 위해서는 한 단어가 몇개의 단어와 대응되는 가를 나타내는 산출력 확률, P(n|e)와 s가 t로 번역 확률, P(t|s) 그리고 변조 확률, P(i|j, l) 값이 필요하다. P(i|j, l)은 원시 언어의 j번째 단어가 목적 언어의 i번째 단어와 대응될 확률을 의미한다. 위와 같은 확률 값을 이용하여 위에서 예를 든 (Le chien est battu par Jean | John(6) does beat(3,4) the(1) dog(2))에 대한 P(t|s) 값을 계산하면 아래의 식과 같다[12].

$$
\begin{aligned}
P(t|s) = {} & P(\text{산출력}=1|\text{John}) \times P(\text{Jean}|\text{John}) \times \\
& P(\text{산출력}=0|\text{does}) \times \\
& P(\text{산출력}=2|\text{beat}) \times P(\text{est}|\text{beat})P(\text{battu}|\text{beat}) \times \\
& P(\text{산출력}=1|\text{the}) \times P(\text{Le}|\text{the}) \times \\
& P(\text{산출력}=1|\text{dog}) \times P(\text{chien}|\text{dog}) \times \\
& P(\text{산출력}=1|\langle\text{null}\rangle) \times P(\text{par}|\langle\text{null}\rangle)
\end{aligned}
$$

위에서 통계 기반 번역을 위한 확률 계산에 대하여 설명하였는데, 이러한 확률 정보를 이용하는 기계 번역 시스템의 예로는 Brown 등의 시스템을 들 수 있다[12]. 이 시스템은 번역 결

과를 exact(정확하게 번역되었음), alternate(정확하지는 않지만 의미는 같음), different(잘못 번역됨)과 같이 세가지 기준으로 나누어 조사하였고, exact인 경우 48%, exact와 alternate 두 가지를 고려한 경우 60%의 정확도를 얻게 되었다.

기계 번역은 자연어처리 연구의 초창기부터 계속해서 연구되어 왔는데, 현재까지의 연구 결과 많은 연구자들은 다음과 같은 교훈을 얻게 되었다[54]. 첫째, 규칙 기반 기계 번역과 통계 기반 기계 번역 중 그 어느 것 하나 만으로는 만족할 만한 결과를 얻을 수 없다. 둘째, 모든 처리 영역에 만족할 수 있는 기계 번역 시스템을 구현한다는 것은 매우 어려운 일이다. 첫째 교훈은 두 가지 번역 방법의 통합을 암시하고 있으며, 둘째 교훈은 앞으로의 기계 번역 시스템이 갖추어야할 적응성(adaptability)을 강조하는 것이라 할 수 있다.

기계 번역 분야의 이와 같은 교훈을 바탕으로 최근 통계 기반 시스템과 규칙 기반 시스템의 통합형에 대한 연구와 기계 번역 시스템의 적응성 향상에 대한 연구가 활발히 진행되고 있고, 그 예로는 CMU의 Despatcher, AT&T의 통합 시스템, ET-6 그리고 ATR, IBM-Japan의 번역 시스템을 들 수 있다[32,54].

3. 결 론

통계 기반의 접근법은 처리 영역에 제한없이 대량의 데이타를 처리할 수 있다는 장점으로 현재 많은 과학 분야에서 이를 이용한 연구가 진행중이며, 본 논문은 통계 기반 접근법을 중심으로 최근 자연어처리 연구 분야의 동향에 대해서 살펴보았다.

통계 기반 접근법을 이용한 품사 태깅은 태깅의 목적을 입력 문장의 어휘 확률(lexical probability) 및 품사 전이 확률(transition probability)의 곱이 최대가 되는 품사열을 찾기 위한 확률 모델을 이용한다. 이 확률 모델에 사용될 어휘 확률, 품사 전이 확률 그리고 연어 정보(collocation information)등 태깅에 필요한 정보는 코퍼스를 분석하여 자동으로 추출하게 된다.

파싱 분야에서는 초기에 대두된 규칙에 의한 문제점을 해결하고자 규칙에 확률 정보를 첨가한 통계 기반의 PCFG에 대한 많은 연구가 수행되고 있다. 또한 효율적이고 정확한 파싱을 위하여 어휘화된 문법(lexicalized grammar)의 개발과 Inside-Outside 알고리즘의 제한에 관한 연구도 활발히 진행되고 있다.

기계 번역 분야에서는 최근 통계 기반의 번역 방법과 규칙 기반의 번역 방법을 통합하고자 하는 노력이 많아지고 있으며, 기계 번역 시스템의 적응성(adaptability)을 높이기 위한 연구도 아울러 수행되고 있다.

통계 기반의 접근법을 이용한 자연어처리 연구 방법들은 처리 영역에 제한없이 어떤 입력 데이타도 처리할 수 있으며, 확장성이 좋다는 장점을 가지고 있다. 하지만 통계 기반의 접근법은 시스템의 수정이 어렵고, 제한된 영역에 대하여 높은 정확도를 보이는 규칙 기반 접근법과 비교하여 정확도가 떨어진다는 단점을 가지고 있다. 따라서 통계 기반 접근법과 규칙 기반 접근법의 선택은 응용 분야에 따라 결정되어야 한다.

예를 들면, 질의어 응답 시스템 등과 같이 처리 영역이 매우 제한적이나 정확성을 요구하는 분야를 위해서는 규칙 기반 접근법을 이용하는 것이 효율적일 것이지만, 음성 인식, 문자 인식 그리고 음성 합성 분야 등 약간의 오류율을 감수하더라도 모든 입력 데이타를 처리하기 위해서는 통계 기반의 접근법이 더욱 효과적일 것이다.

위에서 설명한 바와 같이 각 응용 분야에 따라 적합한 자연어처리의 접근법이 있을 수 있지만, 처리 영역에 제한되지 않고 대량의 데이타를 정확하게 분석하기를 원하는 궁극적인 자연어처리 연구 목표를 달성하기 위해서는 상호 보완적인 규칙 기반 접근법과 통계 기반 접근법의 통합에 대한 연구가 필요하다. 또한 이를 위해서 규칙 기반 접근법에서는 처리 영역의 확장에 대한 연구가 필요하며, 통계 기반 접근법에서는 정확도 향상을 위한 연구가 요구된다.

Leech는 자연어처리 연구의 목표를 터널에 비유하여 "현재 규칙 기반의 접근법과 통계 기반의 접근법은 터널을 뚫기 위해 서로 다른 끝에서 땅을 파고 있는 것이라 할 수 있으며, 언

제가는 터널의 중간에서 만날 것이다"라고 하였는데[23], 이와 같은 비유도 규칙 기반 접근법과 통계 기반 접근법의 통합을 암시하는 것이라고 할 수 있다.

최근 자연어 처리 시스템의 효율을 증진시키고 정확도를 향상시키고자 통계 기반의 접근법과 규칙 기반 접근법의 통합이 시도되고 있으며, 앞으로 두 가지 접근법의 장점을 최대화 시키고 단점은 최소화 시킬 수 있는 효과적인 통합 방법에 대한 연구가 요구되고 있다.

참고문헌

[1] Baker, J. K., "Trainable grammars for speech recognition," Proc. of Spring Conf. of the Acoustical Society of America, 1979.

[2] Benello, J., Mackie, A. W., Anderson, J. A., "Syntactic Category Disambiguation with Neural Networks," Computer Speech and Language, Vol. 3, pp .203~217, 1989.

[3] Black, E., Jelinek, F., Lafferty, J., Mercer, R., Roukos, S. "Decision tree models applied to the labeling of text with parts-of-speech," Proc. of the 1992 DAPRA Speech and Natural Language Workshop, pp. 117~121, 1992.

[4] Black, E., Lafferty, J., Magerman, D., Mercer, R., Roukos, S., "Towards History-based Grammars: using Richer Models for Probabilistic parsing," Proc. of the 31th Annual Meeting of the ACL, pp. 31~37, 1993.

[5] Black, E., Lafferty, J., Roukos, S., "Development and Evaluation of a Board-Coverage Probabilistic Grammar of English-language Computer Manuals," Proc. of the 30th Annual Meeting of the ACL, pp. 185~192, 1992.

[6] Booth, B. M., "Revising CLAWS," ICAME News, Vol.9, pp. 29~3E, 1985.

[7] Brill, E. Magerman, D., Marcus, M., Santorini, B., "Deducing linguistic structure from the statistics of large corpora," Proc. of the DARPA Speech and Natural Language Workshop, pp. 275~285, 1990.

[8] Brill, E., "A Simple Rule-Based Part of Speech Tagger," Proc. of the 3rd Conf. on Applied NLP, Trento, Italy, pp. 153~155, 1992.

[9] Brill, E., "Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach," Proc. of the 31th Annual Meeting of the ACL, pp. 259~265, 1993.

[10] Briscoe, T., Waeger, N.. "Robust Stochastic Parsing Using the Inside-Outside Algorithm," AAAI Workshop Note on Statistically-based Natural Language Programming Techniques, pp. 33-47, 1992.

[11] Brown, P. F., Lai, J. C., Mercer, R. L., "Aligning sentence in Parallel Corpora." Proc. of the 29th Annual Meeting of the ACL, pp. 169~176, 1991.

[12] Brown, P., F., Cocke, J., Della Pietra, S., A., Della Pietra, V., J., Jelinek, F., Lafferty. J., D., Mercer, R., L., Roossin P., S., "A Statistical Approach to Machine Translation," Computational Linguistics, Vol. 16, No. 2, pp. 79-85, 1990.

[13] Brown, P., F., S., A., Della Pietra, V., J., Jelinek, Mercer, R., L., "The Mathematics of Statistical Machine Translation:Parameter Estimation," Computational Linguistics, Vol. 19, No. 2, pp. 263~312, 1993.

[14] Butler, C. S., Computers and Written Texts, Oxford UK: Cambridge USA : Basic Blackwell, Inc, 1992.

[15] Carroll, G., Charniak, E., "Two Experiments on Learning Probabilistic Dpendency Grammars from Copora," AAAI Workshop Note on Statistically-based Natural Language Programming Techniques. pp. 1~7, 1992.

[16] Church, K. W., Hanks, P., "Word association norms, mutual information, and lexicography," Proc. of 27th Meeting of the ACL, pp. 76~83, 1989.

[17] Church, K. W., Mercer, R. L., "Intorduction to the Special Issue on Computational Linguistics Using Large Corpora," Computaional Linguistics, Vol. 19, No. 1, pp. 1~24, 1993.

[18] Cutting, D., Kupiec J., Pedersen, J., S.ibun, P., "A practical Part-Of-Speech Tagger," Proc. of the 3rd Conf. on Applied NLP, Trento, Italy, pp. 134~140, 1992.

[19] Dagan, I., "Automatic Processing of Large Corpora for the resolution of Anaphora Referen-

ces, Project Note," Proc. of the COLING-90, 1990.

[20] DeRose, S. J., "Grammatical Category Disambiguation by Statistical Optimization," Computational Linguistics, Vol. 14, No. 1, pp. 31-39, 1988.

[21] EDR, EDR Electronic Dictionary Technical guide, pp. 69-74, August, 1993.

[22] Gale, W. A., "A Program for Aligning Sentences in Bilingual Corpora," Computational Linguistics, Vol. 19, No. 1, pp. 75-102, 1993.

[23] Garside, R., Leech, G., Sampson, G., The Computational Analysis of English, Longman Inc., New York, 1987.

[24] Greene, B. B., Rubin, G. M. "Automatic Grammatical Tagging of English," Technical Report, Dep. of Linguistics, Brown University, Providence, Rhode Island, 1971.

[25] Grishman, R., Macleod, A. Meyers, A., "Complex Syntax: Building a Computational Lexicon,", Proc. of the COLING-94, pp. 268~272, 1994.

[26] Grishman, R., "Iterative Alignment of Syntactic Structures for a Bilingual Corpus," Second Annual Workshop on Very Large Corpora, Kyoto, Japan, August, pp. 57~68, 1994.

[27] Hindle, D. "Acquiring disambiguation rules from text," Proc. of the 27th Annual Meeting of the ACL, pp. 118~125, 1989.

[28] Ide, N., Veronis, J., "MULTEXT:Multilingual Text Tools and corpora," Proc. of the COLING-94, pp. 588~592., 1994.

[29] Johansson, Stig, The Tagged LOB Corpus : Users' Manual, Bergen, Norwegian Computing Center for the Humanities, 1986.

[30] Joshi, A. K., Srinivas, B., "Disambiguation of Super Parts of Speech(or Supertags): Almost Parsing", Proc. of the COLING-94, Japan, pp. 154~160, 1994.

[31] Kaji, H., Kida, Y., and Morimoto, Y., "Learning translation templates from bilingual text," Proc. of the COLING-92, pp. 672~678, 1992.

[32] Kay, M., Roscheisen, M., "Text-Translation Alignment," Computational Linguistics, Vol. 19., No.1, pp. 121~142, 1993.

[33] Klein, S., Simmons, R. F., "A Computational Approach to Grammatical Coding of English Words," JACM, Vol. 10. pp. 334~47, 1963.

[34] Kupiec, J., "Robust Part-Of-Speech Tagging Using a Hidden Markov Model." Computer Speech and Language, Vol. 6, pp. 225~242, 1992.

[35] Lafferty, J., Sleator, D., Temperley, D., "Grammatical Trigrams: A Probabilistic Model of Link Grammar," Proc. of the AAAI Fall Symposium Series on Probabilistic Approach to NLP, pp. 89~97, 1992.

[36] Lancashire, Ian, The Humanities Computing Yearbook 1989-90: A Comprehensive Guide to Software and other Resources, Oxford Clarendon Press, 1991.

[37] Leech, G., Garside, R., Bryant, M., "CLAWS4: The Tagging of The British National Corpus," Proc. of the COLING-94, pp. 622~628, 1994.

[38] Magerman, D. M., Marcus, M. P., "Pearl: A Probabilistic Chart Parser," Proc. of the 4th conf. of the EACL-91, Berlin, Germany, 1991.

[39] Magerman, D. M., Weir, C., "Probabilistic Prediction and Picky Chart Parser," Proc. of the 1992 DAPRA Speech and Natural Language Workshop, 1992.

[40] Marcus, M. P., Marcinkiewicz M. A., Santorini, B., "Building a Large Annotated Corpus of English : The Penn Treebank," Computational Linguistics, Vol. 19, No. 2, pp. 313-330, 1993.

[41] Marcus, M., "Corpus based Natural Language Processing," Tutorial Program of COLLING-94, Tokyo, Japan, pp. 119-147, 1994.

[42] Marcus, M., "Statistical Natural Language Processing: Current Trends and Future Directions", Proc. of ATR Int'l. Workshop on Speech Translaton, 1993.

[43] McClelland, J. L., Rumelhart, D. E., The PDP Research Group, Parallel Distributed Processing, The MIT Press, 1986.

[44] Pereira. Fernando, schabes, Yves, "Inside-Outside reestimation from partially bracketed corpora." Proc. of the 30th Annual Meeting of the ACL, pp. 128~135, 1992.

[45] Schabes, Y., "Stochastic lexicalized tree-adjoining grammar," Proc. of the COLING-92, pp. 426~432, 1992.

[46] Schmid, H., "Part-of-Speech Tagging with Neural Networks," Proc. of the COLING-94, pp. 172~176, 1994.

[47] Sch tze H., Singer, Y., "Part-Of-Speech Tagging Using a Variable Memory Markov Model," Proc. of the 26th Annual Meeting of the ACL, pp. 181~187, 1994.

[48] Scouter, G., Atwell, E., "A Richly Annotated Corpus for Probabilistic Parsing," AAAI Workshop Note on Statistically-based Natural Language Programming Techniques, pp. 22~32, 1992.

[49] Silberztein, M. D., "INTEXT:A Corpus Processing System," Proc. of the COLING-94, pp. 579~582, 1994.

[50] Singclair, J. M., Looking Up, Collins ELT, 1987.

[51] Smadja, F., "Retrieving Collocations from Text:Xtract," Computational Linguistics, Vol. 19, No. 1, pp. 141~177., 1993.

[52] Sumita, E., Iida, H., "Experiments and prospects of example-based machine translation," Proc. of the 29th Annual. Meeting of the ACL, pp.185~192, 1991.

[53] Teller, V., Kosaka, M., and Grishman, R., "A Comparative study of Japanese and English sublanguage patterns," Proc. of the Second Int' l Conf. on Theoretical and Methodological Issues in Machine Translation, 1988.

[54] Tsujii, J., "Future Directions of MT-Language, Meanings and Translation," Tutorial Program of the COLING-94, Tokyo, Japan, pp. 63~116, 1994.

[55] Uramoto, N., "Extracting a Disambiguated Thesaurus from Parallel Dictionary, " Second Annual Workshop on Very Large Corpora, Kyoto, Japan, pp. 33~41, 1994.

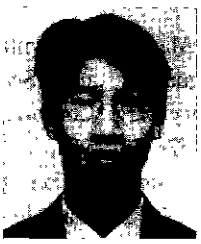
[56] Utsuro, T., Matsumoto, Y., and Nagao, M., "Lexical Knowledge acquisition from bilingual corpora,", Proc. of the COLING-92, pp. 581~587, 1992.

[57] Zernik, U., Lexical Acquisition:Exploiting On-Line Resources to Build a Lexicon, Lawlence Erlbaum Associates, 1991.

[58] de Marcken, C. G., "Parsing the LOB Corpus," Proc. of the 28th Annual Meeting of the ACL, pp. 243~251, 1990.

임 해 창

1979 고려대학교 독어독문학과 학사
1983 Missori 주립대학 전산학 석사
1990 Texas 주립대학 전산학 박사
1994 ~현재 고려대학교 전산과학과 부교수
관심 분야 : 자연어 처리, 정보 검색, 인공지능



임 희 석

1992 고려대학교 전산과학과 학사
1994 고려대학교 전산과학과 석사
1994 ~현재 고려대학교 전산과학과 박사과정
관심 분야 : 자연어처리, 정보검색, 인공지능, 데이타베이스

윤 보 현

1992 목포대학교 전산통계학과 학사
1994 ~현재 고려대학교 전산과학과 석사과정
관심 분야 : 자연어 처리, 정보 검색, 데이타베이스