최대 엔트로피 부스팅 모델을 이용한 영어 전치사구 접속과 품사 결정 모호성 해소

(Resolving Prepositional Phrase Attachment and POS Tagging Ambiguities using a Maximum Entropy Boosting Model)

박 성 배 †

(Sung-Bae Park)

요 약 최대 엔트로피 모델은 자연언어를 모델링하기 위한 좋은 방법이다. 하지만, 최대 엔트로피 모델을 전치사구 접속과 같은 실제 언어 문제에 적용할 때, 자질 선택과 계산 복잡도의 두 가지 문제가 발생한다. 본 논문에서는, 이런 문제와 자연언어 자원에 존재하는 불균형 데이터 문제를 해결하기 위한 최대 엔트로피 부스팅 모델(maximum entropy boosting model)을 제시하고, 이를 영어의 전치사구 접속과 품사 결정 모호성 해소에 적용한다. Wall Street Journal 말뭉치에 대한 실험 결과, 문제의 모델링에 아주 작은 노력을 들였음에도 불구하고, 전치사구 접속 문제에 대해 84.3%의 정확도와 품사 결정 문제에 대해 96.78%의 정확도를 보여 지금까지 알려진 최고의 성능과 비슷한 결과를 보였다.

키워드 : 최대 엔트로피 모델, 전치사구 접속, 품사 결정, 부스팅, 능동 학습

Abstract Maximum entropy models are promising candidates for natural language modeling. However, there are two major hurdles in applying maximum entropy models to real-life language problems, such as prepositional phrase attachment: feature selection and high computational complexity. In this paper, we propose a maximum entropy boosting model to overcome these limitations and the problem of imbalanced data in natural language resources, and apply it to prepositional phrase (PP) attachment and part-of-speech (POS) tagging. According to the experimental results on Wall Street Journal corpus, the model shows 84.3% of accuracy for PP attachment and 96.78% of accuracy for POS tagging that are close to the state-of-the-art performance of these tasks only with small efforts of modeling.

Key words : Maximum Entropy Model, PP Attachment, POS Tagging, Boosting, Active Learning

1. 서 론

전치사구 접속 모호성 해소와 품사 결정 문제는 자연언어처리에서 가장 중요한 문제들 중의 하나이다. 따라서, 기계학습이나 통계에 기반을 둔 여러 방법들이 이 문제들에 적용되어 왔다. 전치사구 접속 문제는 분류 문제로 볼 수 있기 때문에 말뭉치 기반의 해결 방법에서는 일반적으로 감독 학습 방법이 사용되었다[1,2,3]. 특히, [1]에서는 최대 엔트로피 모델(maximum entropy model)을 이 문제에 적용하였는데, 단어와 단어 클래스를 모두 사용하여 81.6%의 정확도를 얻었다. 반면에 Collins와 Brooks는 학습 시 경험하지 못한 예제를 위한 평탄화 모델로 back-off 모델을 사용하여 84.5%의 정확도를 보였다[2]. 또한, 이들은 전치사구 접속 모호성 해소를 위한 가장 중요한 정보가 전치사 자신이라는 것과 빈도수가 낮은 데이타를 학습에 사용하는 것이 성능을 높이는 데 도움이 된다는 것을 발견하였다. 비통계적 방법은 [3]에서 사용되었는데 변형기반(transformation-based) 방법을 사용하여 81.8%의 정확도를 보였다. 이들은 또한 전치사가 이 문제에 있어서 가장 중요한 자질임을 보여 Collins와 Brooks의 주장을 다시 한번 뒷받침하였다.

품사 결정에서는 대부분의 연구가 96% 이상의 정확도를 보였다[4,5,6]. 품사 결정에서 가장 중요한 문제 중 하나는 어떻게 미지어(unknown words)를 다룰 것인가

† 학생회원 : 서울대학교 컴퓨터신기술공동연구소
sbpark@bi.snu.ac.kr
논문접수 : 2002년 10월 5일
심사완료 : 2003년 3월 10일

하는 것이다. 규칙 기반 방법에서는 이런 단어들을 다루기 위한 규칙을 만들면 되지만, 통계적 방법에서는 이들을 처리하기 위한 적당한 방법을 새로이 마련하여야 한다[5]. Weischedel et al.은 어휘 정보가 이 문제에 대한 훌륭한 해결책임을 보였다[6]. 이 경우에 품사 태거는 미지어의 접두사나 접미사가 특정한 태그를 가질 확률을 계산하게 된다.

전치사구 접속 문제와 품사 결정 문제는 기계학습(machine learning)의 입장에서 보면 일종의 분류 문제(classification problem)이다. 그리고, 조건부 확률(conditional probability)은 분류 문제를 구현하는 좋은 방법이다. 조건부 확률을 계산할 수 있는 여러 확률 모델 중 최대 엔트로피 모델(maximum entropy model)은 자연언어처리의 여러 문제에 성공적으로 적용되어 왔지만[7], 이 모델은 두 가지 문제를 가지고 있다. 첫 번째 문제는 최대 엔트로피 모델이 풀고자 하는 대상 문제에 대한 과도한 사전 지식을 요구한다는 점이고, 두 번째 문제는 이 모델을 학습하는데 필요한 계산량이 너무 많다는 점이다. 자연언어 학습 문제에서는 보통 수백만개의 학습 예제가 사용되므로, 계산량을 많은 점은 치명적인 단점이다.

본 논문에서는, 전치사구 접속 문제와 품사 결정 문제를 위해 최대 엔트로피 부스팅 모델(maximum entropy boosting model)을 제안한다. 위에서 언급한 두 문제를 해결하기 위해서, 최대 엔트로피 부스팅 모델은 결정트리(decision tree)와 능동 학습(active learning)을 이용한다. 그리고, 자연언어 학습이 본질적으로 가지는 또 다른 문제는 대부분의 자연언어 자원들이 불균형 데이타 분포를 가진다는 점이다. 이런 데이타 불균형은 기계학습 알고리즘이 낮은 재현도(recall)를 갖게 함으로써 성능을 떨어뜨린다[8]. 이 문제를 해소하기 위해서, 최대 엔트로피 모델에 AdaBoost 알고리즘을 적용한다. AdaBoost처럼 학습 알고리즘을 결합하면 분류 경계 지점에서의 분산(variance)이 줄기 때문에, 높은 성능을 얻을 수 있다.

실험적으로 이 모델을 전치사구 접속 문제와 품사 결정 문제에 적용하여 유용성을 검증하였다. 전치사구 접속 문제의 실험 결과, 사전에 미리 주어진 정보가 어휘 정보 뿐임에도 불구하고 지금까지 알려진 최고의 성능에 버금가는 높은 정확도를 얻을 수 있었다. 품사 결정 문제에서도 보통 단어에 대해 7개의 자질, 미지어에 대해 11개의 자질만 사용하였지만 매우 높은 정확도를 보였다. 따라서, 최대 엔트로피 모델의 문제점인 많은 사전 지식과 과다한 계산량 문제가 최대 엔트로피 부스팅 모델에 의해 효과적으로 해결되었다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 최대 엔트로피 부스팅 모델을 설명한다. 3장에서 전치사구 접속 문제와 품사 결정 문제를 분류 문제로 재정의하여 최대 엔트로피 부스팅 모델이 각 문제에 어떻게 적용되는지를 보인다. 4장에서는 실험 결과를 제시하고, 마지막으로 5장에서 결론을 맺는다.

2. 최대 엔트로피 부스팅 모델

위에서 언급한 바와 같이 많은 자연언어처리 문제들은 분류문제로 생각되어질 수 있다. 즉, 각 문제는 어떤 언어학적 문맥 $x \in X$를 관찰한 후, 이 데이타의 클래스 $y \in Y$를 결정하는 문제로 생각될 수 있다. 조건부 확률 $p(y \mid x)$는 y를 추정하는 분류기를 구현하는 좋은 방법 중의 하나인데, 그 중 특히 최대 엔트로피 모델은 확률 $p(y \mid x)$를 추정하기 위해 다양한 종류의 정보를 통합할 수 있는 장점을 추가적으로 가지고 있다.

최대 엔트로피 모델은 일반적으로 다음과 같은 형태를 갖는다.

$$p(y \mid x) = \frac{\prod_{i=1}^{k} \exp(\mu_i f_i(x, y))}{Z} \quad (1)$$

여기서, μ_i는 최대 엔트로피 모델의 모수(parameter), Z는 정규화 상수, 그리고 $f_i(x, y)$는 한 예제 (x, y)에 대한 계산 가능한 자질(feature)이다. 최대 엔트로피 모델의 성능은 이 자질이 얼마나 좋으냐 하는데 크게 영향을 받는다. 따라서, 일반적으로는 이런 자질들은 데이타의 특성을 잘 반영할 수 있도록 전문가에 의해 정의된다. 하지만, 이 자질을 정의하는 사람이 풀고자 하는 문제에 대해 충분한 지식이 없을 경우에는 좋은 자질을 만들기 어렵고, 따라서 최대 엔트로피 모델은 좋은 성능을 보이기 힘들다.

이 문제를 해결하기 위해서, 최대 엔트로피 부스팅 모델에서는 고차 자질(high-order feature)의 생성자로서 결정트리를 사용한다. 결정트리가 if-then 규칙의 집합 형태로 쉽게 표현될 수 있기 때문에, 자질은 결정트리를 if-then 규칙으로 변경함으로써 자동적으로 만들어 질 수 있다. 또한, 결정트리는 간단한 일차 자질(first-order feature)만 알고 있으면 쉽게 학습할 수 있다.

최대 엔트로피 모델의 또 다른 문제는 이 모델을 학습하는데 필요한 계산량이 너무 과도하다는 점이다. μ_i의 최적값은 주로 GIS 알고리즘[9]에 의해 결정된다. 그러나, 이 알고리즘에서는 각 반복마다 모든 f_i에 대해서 기대값 $E_p[f_i]$을 계산해야 하는데, 이것의 계산 복잡도

가 $O(M|Y|N)$이다[7]. 여기서, M은 주어진 예제에 대해 적용되는 자질의 평균 수, Y는 클래스 레이블의 집합, N은 학습 예제의 수이다. 따라서, 자연언어처리 문제와 같이 학습 예제의 수가 매우 많은 문제에서는 $E_p[f_i]$를 추정하는 것이 불가능할 수 있다.

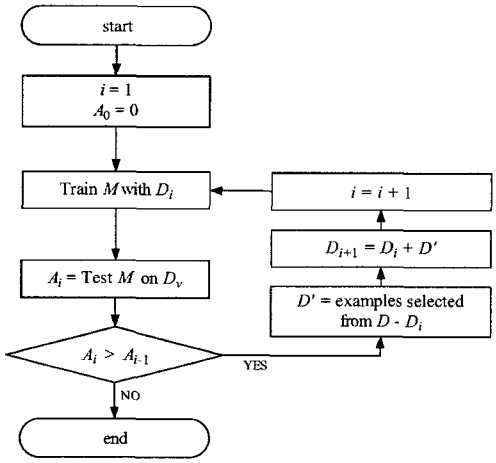

그림 1 최대 엔트로피 부스팅 모델의 능동 학습 과정

결정트리는 예제 공간을 나눔으로써 자질을 생성하기 때문에, 최대 엔트로피 부스팅 모델에서는 M이 고정된다. 따라서, 계산 복잡도를 낮출 수 있는 유일한 방법은 N을 줄이는 것이고, 이는 최대 엔트로피 모델에 능동 학습(active learning) 기법을 적용하여 이루어질 수 있다. 그림 1은 이 과정을 설명하고 있다. N개의 학습 예제로 이루어진 데이타 집합 D가 주어질 때, 최대 엔트로피 부스팅 모델의 학습은 무작위로 선택된 $D_0 \subset D$로 시작한다. 그리고, 독립된 검증 집합 D_v가 있음을 가정한다. 각 i번째 반복마다, 모델은 D_i로 학습되고, 학습된 모델이 D_v에 대해서 이전 반복보다 더 좋은 성능을 보일 때에만 학습 집합이 확장된다. 학습 집합을 확장하기 위해서, $D-D_i$에서 λ개의 예제를 선택한 후 현재의 D_i에 더하여 D_{i+1}을 만든다. 만약 학습 집합을 증가한 후에 D_v에 대한 정확도가 나아지지 않으면 학습은 중단된다. 다음 반복에서 사용되기 위해서 선택되는 학습 예제들은 $D-D_i$ 중에서 가장 정보량이 많은 것들이다. 정보이론에 따르면 정보량이 많은 예제는 불확실한 것들이다[10]. 가장 불확실한 예제 e^*는 다음의 수식에 의해 결정된다.

$$e^* = \underset{e_j \in D-D_i}{\arg\min} KL(U \parallel P(C \mid e_j)) \qquad (2)$$

여기서, U는 균일 분포이고, $P(C|e)$는 예제 e의 클래스 기반 사후분포이다.

마지막으로, 자연언어 데이타의 심한 불균형을 극복하기 위해서 AdaBoost 알고리즘[11]을 사용한다. AdaBoost 알고리즘은 각 반복마다 현재 상태에서 분류하기 힘든 예제에 집중하기 때문에 반복될 때마다 드문 클래스에 우선 순위를 두어 예제를 샘플링한다. AdaBoost나 Support Vector Machines와 같이 마진(margin) 기반 알고리즘은 초평면(hyperplane) 부근의 예제에 집중함으로써, 드문 클래스를 우선하여 샘플링할 수 있다. 특히, AdaBoost는 일종의 위원회 모델로 생각될 수 있기 때문에 높은 재현도(recall) 뿐만 아니라 더 나은 정밀도(precision)를 기대할 수 있다[12].

3. 문제 정의

3.1 전치사구 접속 모호성 해소

전치사구 접속 문제는 문장 내에 나타나는 전치사구의 접속 위치를 결정하는 것이다. 예를 들어, 다음의 두 문장에서 이 문제는 전치사 '*with*'가 앞선 명사구(NP)를 수식하는지, 동사구(VP)를 수식하는지를 결정하는 것이다.

(1) I bought the shirt *with* pockets.

(2) I washed the shirt *with* soap.

첫 번째 문장에서 *with*는 *shirt*가 *pocket*을 가지기 때문에 *shirt*를 중심어로 하는 명사구를 수식한다. 두 번째 문장에서는, *with*는 비누(*soap*)로 셔츠(*shirt*)를 씻기 때문에 *washed*를 중심으로 하는 동사구를 수식한다.

앞에서 밝힌 바와 같이 전치사구 접속 문제는 기계학습(machine learning)의 입장에서 보면 일종의 분류 문제이다. 이 문제의 목적은 (v, n_1, p, n_2)형태로 주어진 4-튜플(tuple)에 대해 정확한 접속 $y \in \{N, V\}$를 결정하는 것이다. 여기서, v는 중심동사, n_1은 v의 목적어인 중심명사, p는 전치사, n_2는 전치사구의 중심명사이다. 즉, 이 문제의 목적은 부사적 접속(VP 접속)인지 형용사적 접속(NP 접속)인지를 결정하는 것이다. 예를 들어, 첫 번째 문장에 해당하는 튜플은 (*bought, shirt, with,* pockets)가 되고, 두 번째 문장에 해당하는 튜플은 (*washed, shirt, with, soap*)가 된다. 그리고, 첫 번째 문장의 올바른 접속은 N이고, 두 번째 문장은 V이다.

앞에서 말한 바와 같이, 전치사구 접속 문제는 분류 문제로 생각될 수 있다. 즉, 이 문제는 다음과 같이 각 접속의 확률을 비교하는 것으로 표현될 수 있다.

$$f(v, n_1, p, n_2) = \underset{y \in \{N, V\}}{\arg\max} P(y \mid v, n_1, p, n_2) \qquad (3)$$

최대 엔트로피 모델이 확률 모델이기 때문에,

$P(y \mid v, n_1, p, n_2)$는 v, n_1, p, n_2를 일차 자질로 하는 최대 엔트로피 모델로 추정될 수 있다. 즉, 식 (3)의 $P(y \mid v, n_1, p, n_2)$은 다음의 식 (4)와 같이 표현될 수 있다.

$$P(y \mid v, n_1, p, n_2) = \frac{1}{Z} \exp(\sum_i \mu_i f_i(v, n_1, p, n_2)) \quad (4)$$

여기서, $f_i(v, n_1, p, n_2)$는 일차 자질들로 구성된 고차 자질이다.

전치사구 접속 문제에서는 주어진 데이타 외에 추가적인 정보를 사용할 수 있을 때에 더 높은 정확도를 기대할 수 있다[13,14]. [13]에서는 WordNet을 사용하여 88.1%의 정확도를 얻었고 이는 지금까지 알려진 최고의 성능이다. 하지만, 본 논문에서는 이런 추가적인 정보를 사용하지 않는다. 이런 정보를 사용하면 더 높은 정확도를 얻을 수 있겠지만, 학습 알고리즘의 성능은 이런 정보가 존재하지 않을 때 공정하게 비교될 수 있다. 특히, 최대 엔트로피 부스팅 모델은 결정트리로 최대 엔트로피 모델을 효율적으로 구현하는 것과 부스팅 기법을 통해 최대 엔트로피 모델을 뛰어넘는 것을 목표로 하기 때문에, 결정트리와 최대 엔트로피 모델과의 성능 비교가 중요하다.

3.2 품사 결정

$w_1, \ldots, w_N$을 문장 내의 단어열이라고 하자. 품사 결정 문제의 목표는 확률 $p(t_1, \ldots, t_N \mid w_1, \ldots, w_N)$을 최대로 만드는 품사열 $t_1, \ldots, t_N$을 찾는 것이다. 일반적으로 이런 문자열에 대한 정확한 확률을 추정하는 것이 현실적으로 불가능하므로 몇 가지 가정을 통해 이 확률을 추정한다. 우리는 이런 가정으로서 독립 가정을 이용하며, 다음과 같이 확률을 추정한다.

$$p(t_1, \ldots, t_N \mid w_1, \ldots, w_N) = \prod_{i=1}^{N} p(t_i \mid h_i)$$

여기서, h_i는 단어 w_i의 문맥정보이다. 언어학적으로는 이런 독립 가정이 맞지 않겠지만, 이렇게 해서 추정된 확률은 실험적으로 매우 정확하다.

본 논문에서는 모든 단어 w_i에 대한 확률 $p(t_i \mid h_i)$를 최대 엔트로피 부스팅 모델로 계산한다. 즉, $f_j(h_i, t_i)$를 자질 함수라고 할 때, 확률 $p(t_i \mid h_i)$는 다음과 같이 추정된다.

$$p(t_i \mid h_i) = \frac{1}{Z} \exp(\sum_j \mu_j f_j(h_i, t_i))$$

본 논문에서는 품사 결정 문제의 일차 자질로 왼쪽과 오른쪽 문맥의 두 단어씩을 이용한다. 즉,

$$h_i = \{w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, t_{i-2}, t_{i-1}\}$$

이다. 품사 결정이 순차적으로 이루어지기 때문에, 오른쪽 문맥에 있는 단어들의 품사는 이용하지 않는다.

미지어를 위해서는 [15]에서처럼 몇 가지 특별한 자질을 마련한다. 표 1은 단어 w_i의 품사를 결정하기 위해 사용되는 일차 자질을 정리하여 보여준다. 학습 데이타에 자주 나타나지 않는 예제도 실험적으로 미지어와 같은 효과를 보이므로, 학습 데이타에서 3 번 이하로 나타나는 단어들도 미지어로 처리한다. 접두사와 접미사는 미지어에 대해서만 추출하고, 그 길이도 3으로 제한한다. 그리고, w_i가 미지어일 때 w_i의 품사를 결정하기 위해 추가로 세 개의 자질을 더 둔다. 자질 hasnumber는 w_i가 숫자를 포함하는지를 나타내고, hasupper는

표 1 단어 w_i의 품사를 결정하기 위한 일차 자질

w_i가 미지어일 때		w_i가 보통 단어일 때	
자질	설명	자질	설명
w_{i-2}	i-2 번째 위치의 단어	w_{i-2}	i-2 번째 위치의 단어
w_{i-1}	i-1 번째 위치의 단어	w_{i-1}	i-1 번째 위치의 단어
w_{i+1}	i+1 번째 위치의 단어	w_{i+1}	i+1 번째 위치의 단어
w_{i+2}	i+2 번째 위치의 단어	w_{i+2}	i+2 번째 위치의 단어
t_{i-2}	w_{i-2}의 품사	t_{i-2}	w_{i-2}의 품사
t_{i-1}	w_{i-1}의 품사	t_{i-1}	w_{i-1}의 품사
prefix(w_i, j)	w_i의 길이 j의 접두사	w_i	i번째 위치의 단어
suffix(w_i, j)	w_i의 길이 j의 접미사		
hasnumber	w_i가 숫자인지		
hasupper	w_i가 대문자를 지녔는지		
hashyphen	w_i가 hyphen을 지녔는지		

w_i가 대문자를 포함하는지, hashyphen은 w_i가 hyphen을 포함하는지를 나타낸다.

품사 결정 문제의 계산 복잡도를 줄이기 위해서 폭검색(beam search)과 태그 사전(tag dictionary)을 사용한다. N을 문장의 길이라 하고 T를 가능한 모든 품사들의 집합이라고 할 때, 품사 결정 시 계산하여야 할 품사열의 종류는 $N^{|T|}$개이다. 최대 엔트로피 부스팅 모델이 마코프 가정(Markov assumption)을 따르지 않으므로, Viterbi 알고리즘이 적용될 수 없다. 따라서, 가장 좋은 열을 구하기 위해 폭검색을 이용한다(그림 2).

1. 단어 w1에 대한 품사 태그를 생성한다.
2. 가장 좋은 M 개의 태그 $t_{11}, \ldots, t_{1M}$를 찾는다.
3. $1 \le j \le M$에 대해 s_{1j}를 설정한다.
4. i = 2, j = 1
5. Do While $i \le N$
   5.1 Do While $j \le M$
      (a) 단어 wi에 대한 품사 태그를 생성한다.
      (b) 가장 좋은 M 개의 태그 $t_{i1}, \ldots, t_{iM}$를 찾는다.
      (c) $s_{(i-1)j}$에 새로운 태그를 붙여 s_{ij}를 설정한다.
      (d) j = j + 1
   5.2 s_{ij}에 따라 M 개의 가장 좋은 열을 찾는다.
   5.3 $i = i+1$

그림 2 품사 결정을 위한 폭검색(beam search) 알고리즘

이 알고리즘에서, M은 폭의 크기이고 t_{ij}는 w_i의 j번째로 좋은 품사이다. 이 그림에서 s_{ij}는 단어 w_1부터 w_i까지의 품사열 중에서 j번째로 좋은 품사열이다. 새로운 단어의 품사를 결정할 때마다 우리는 M개의 가장 좋은 열만 유지한다. 따라서, 새로운 단어의 품사를 결정할 때마다 $s_{(i-1)j}$에 새로운 M개의 가장 좋은 열을 붙인다. 이렇게 하면 M^2개의 열이 생기므로, 이 중에서 가장 좋은 M개만 남기고 나머지는 버린다. 결과적으로, 우리는 폭검색을 통해 최종적으로 $M \cdot |T|$개의 열만 고려하게 된다. 아래의 모든 실험에서는 M을 3으로 고정하였다.

폭검색 외에 계산량을 줄이기 위해, 태그 사전(tag dictionary)도 사용하였다. 이 사전은 학습 데이타에 나타난 각 단어가 가지는 품사들을 기록해 둔 것이다[16]. 이 사전이 있으면, 검색 알고리즘은 각 단어에 대해 사전에 있는 품사의 확률만 계산하므로 계산량이 줄게 된다. 많은 단어들이 실제로는 단지 몇 개의 품사가 가지므로, 이 방법은 계산량 감소에 매우 유익하다.

계산량을 줄일 수 있는 또 하나의 방법은 각 단어마다 독립된 분류기(classifier)를 만드는 것이지만, 이 방법은 매우 심각한 데이타의 분열을 야기할 것이다. 따라서, 우리는 단어 대신에 각 품사 태그마다 독립된 분류기를 학습한다. 최대 엔트로피 부스팅 모델에서는 약 학습자(weak learner)가 최대 엔트로피 모델이므로, 최대 엔트로피 모델의 출력 결과인 확률을 AdaBoost의 신뢰도로 생각한다. 따라서, AdaBoost의 최종 함수 $f(x)$는 다음과 같이 변경된다.

$$f(x) = \arg\max_{j \in T} \sum_{i=1}^{R} (\beta_i h_{ti}(x))$$
$$t = \arg\max_{j \in T} h_{ji}(x)$$

여기서 R은 AdaBoost의 라운드 수이고, h_{ti}는 품사 태그 t를 위해 i번째 라운드에서 학습된 약 학습자이고, β_i는 h_{ti}의 가중치로 $f(x)$를 학습할 때 결정된다. 아래의 실험에서 R의 값은 검증집합을 가지고 결정하였다.

4. 실 험

4.1 실험 데이타

전치사구 접속 모호성 해소를 위한 데이타집합은 Penn Treebank Wall Street Journal 데이타에서 추출된 것이다[3][1]. 이 데이타는 20,801개의 학습 예제와 3,097개의 테스트 예제로 이루어졌다. 이 데이타의 각 예제는 4-튜플과 목표 분류, 즉, (v, n_1, p, n_2, y)의 다섯 항목으로 이루어져 있다. 여기서 y는 N 또는 V의 값을 갖는다. 이 데이타집합이 4,309개의 예제로 이루어진 독립적인 개발 집합(*development set*)을 가지고 있으므로 이를 능동 학습의 검증 집합(*validation set*)으로 사용한다.

이 데이타집합은 많은 오류를 포함하고 있다. 예를 들어, 테스트 집합 중 133개의 예제는 *the*를 n_1이나 n_2에 포함하고 있다. 그리고, (*sing, birthday, to, you, N*)과 같이 잘못된 접속 정보를 가지는 오류도 상당히 존재한다. 하지만, 본 논문의 목적은 여러 학습 알고리즘과 최대 엔트로피 부스팅 모델을 전치사구 접속 문제에서 비교하는 것이기 때문에, 아래의 실험은 이러한 오류들을 정정하지 않고 수행되었다.

품사 결정 문제를 위한 데이타집합도 Penn Treebank Wall Street Journal 데이타를 사용하였다. Treebank II의 구문부착 말뭉치에서 chunklink[2]를 사용하여 단어-품

1) 이 데이타집합은 ftp://ftp.cis.upenn.edu/pub/adwait/PPattachData에서 다운로드받을 수 있다.

2) 이 Perl 스크립트는 http://ilk.kub.nl/~sabine/chunklink에서 다운로드받을 수 있다.

사 쌍을 추출하였다. 이 데이타집합의 통계정보는 표 2에 있다. 전체 단어의 수는 1,173,765개, 어휘수는 49,206개이고 품사 태그의 수는 45개다. 이 중, 약 60%인 704,251개의 단어는 학습 집합으로, 20%인 234,819 개의 단어는 검증 집합으로, 나머지 234,695개의 단어는 테스트 집합으로 사용하였다. 학습 집합에 있는 미지어의 수는 23,237개였고, 접두사의 수는 3,713개, 접미사의 수는 3,199개였다. 그리고, 테스트 집합에 나타난 약 4,578개의 단어가 학습 집합에는 나타나지 않았다.

표 2 Wall Street Journal 품사 결정 말뭉치에 대한 통계

단어의 수	어휘	문장의 수	품사 태그의 수
1,173,765	49,206	49,209	45

4.2 전치사구 접속 문제의 실험 결과

본 실험에서는 최대 엔트로피 부스팅 모델의 성능을 확인하기 위해서 네 가지 종류의 분류기와 비교하였다. 비교대상이 된 분류기는 (i) 기본 모델, (ii) 결정트리, (iii) 최대 엔트로피 모델, (iv) back-off 모델이다. 기본 모델은

$$f_{baseline}(v, n_1, p, n_2) = \begin{cases} N & \text{if } p = of \\ V & \text{otherwise} \end{cases} \quad (5)$$

이다. 이 수식은 전치사 *of*만 명사구를 수식하는 경향을 갖고, 나머지는 동사구를 수식하려는 경향을 갖기 때문에 만들어졌다. 이 논문에서 사용된 결정트리는 C4.5 release 8[17]이다. back-off 모델은 Katz가 제안한 것으로[18], 확률 추정이 더 짧은 문맥 정보를 통해 보정된다. 즉, 다음과 같이 재귀적으로 정의된다.

If $f(w_1, \ldots, w_{n-1}) > c_1$,

$$p(w_n \mid w_1, \ldots, w_{n-1}) = \frac{C(w_1, \ldots, w_n)}{C(w_1, \ldots, w_{n-1})},$$

Else if $f(w_2, \ldots, w_{n-1}) > c_2$,

$$p(w_n \mid w_1, \ldots, w_{n-1}) = \alpha_1 \times \frac{C(w_2, \ldots, w_n)}{C(w_2, \ldots, w_{n-1})},$$

Else backing-off continues in the same way.

여기서, C는 빈도 함수이다. back-off 모델의 아이디어는 빈도 (c_1, c_2, ...)가 정확한 추정을 하기에 충분할 정도로 높지 않을 때 낮은 차원의 n-gram 모델에 기초하여 최대 가능도(maximum likelihood) 추정을 이용하는 것이다. 그리고, α_i는 조건부 확률의 합을 1로 만드는 정규화 상수이다.

전치사구 접속 문제에서의 back-off는 4-튜플, 3-튜플, 2-튜플와 전치사 p를 결합하는 것이며, $c_1 = c_2 = \ldots = 0$이다[2]. 이 c_i 값은 전치사구 접속 문제에서는 빈도수가 낮은 데이타도 중요하다는 것을 뜻한다. 그리고, $\alpha_i = 1$이다.

표 3은 각 분류기의 정확도를 보이고 있다. 이 표에서 최대 엔트로피 모델은 Ratnaparkhi의 실험 결과[1]이고, 최대 엔트로피 부스팅 모델은 본 논문에서 사용된 모델이다. 그리고, back-off 모델은 Collins와 Brooks의 실험 결과[2]이다. 사람의 전치사구 접속 모호성 해소 정확도는 문장 전체가 주어졌을 때 93.2%이고, 실험에서 쓰인 튜플만 주어졌을 때에는 88.2%이다[14]. 따라서, 88.2%를 상한(upper bound)으로 정할 수 있다. 또한, 기본 모델의 정확도가 70.4%이므로 이를 하한(lower bound)으로 생각할 수 있다.

표 3 전치사구 접속 문제의 정확도

알고리즘	정확도
기본 모델	70.4%
결정트리	80.2%
최대 엔트로피 모델	77.7%
back-off 모델	84.5%
부스팅을 하지 않은 최대 엔트로피 부스팅 모델	81.8%
부스팅을 한 최대 엔트로피 부스팅 모델	84.3%

최대 엔트로피 모델의 정확도는 77.7%로 결정트리의 정확도인 80.2%보다 낮다. 이는 약간 놀라운 결과이다. 결정트리가 잘 작동하기 위해서는, 각 일차 자질들이 서로 독립이어야 하는데 직관적으로 볼 때 전치사구 접속의 일차 자질들은 완전히 독립적이지 않다. 특히, 이들 중 전치사구의 명사 중심어 n_2는 전치사 p와 매우 밀접하게 관련되어 있다. 또한 n_1이 v의 목적어이므로, n_1과 v도 서로 밀접하게 관련되어 있다. 그럼에도 결정트리가 최대 엔트로피 모델보다 높은 정확도를 보이는 것은 이 문제의 모든 클래스와 일차 자질들이 이산 값을 취하고 예제 공간이 이 일차 자질들에 의해 잘 나누어지기 때문인 것으로 생각된다.

그러나, 최대 엔트로피 부스팅 모델의 정확도는 결정트리보다 높다. 심지어, 부스팅을 하지 않았을 때에도 81.8%의 정확도를 보이고, 이는 [1]의 실험결과와 결정트리보다 훨씬 좋은 결과이다. [1]에서 Ratnaparkhi는 최대 엔트로피 모델의 정확도로 단어 정보만 사용하였을 때 77.7%, 단어와 단어 클래스를 모두 사용하였을 때 81.6%를 보고하였다. 따라서, 결정트리로부터 추출한 자질로 최대 엔트로피 모델을 학습시키는 최대 엔트로피 부스팅 모델이 [1]의 최대 엔트로피 모델보다 좋은

성능을 보인다.

부스팅 기법을 사용하였을 때에는 정확도가 84.3%까지 올라간다. 이 값은 Collins와 Brook가 보고한 결과(표 3의 back-off 모델)와 거의 비슷한 정확도이다. back-off 모델이 전치사구 접속 문제에서 높은 성능을 보일 수 있었던 이유는 이 문제의 데이타 부족(data sparseness) 현상이 심하지 않기 때문이다. back-off 모델은 데이타 부족이 심할 때 성능이 나빠지므로[19], 품사 결정처럼 데이타의 분포가 다양한 문제에서는 독립적으로 쓰이지 못한다. 이 모델의 또 다른 문제는 이 모델로 계산된 확률이 실제 학습 데이타와 일치하지 않는다는 점이다. 그렇기 때문에, 학습 데이타의 수가 많아 진다고 해서, 그에 비례해서 성능이 좋아진다는 보장을 할 수 없다. 이에 비해, 최대 엔트로피 모델은 보간법(interpolation)의 일종이 아니므로, 학습 데이타를 충실히 반영한다.

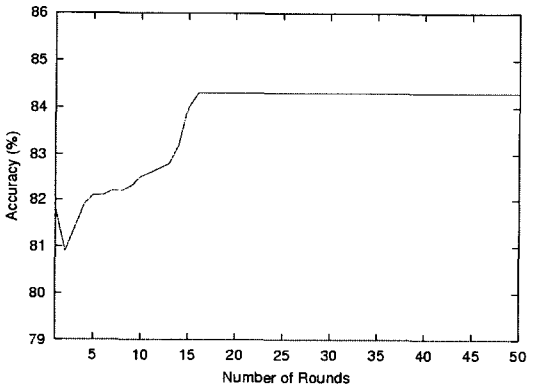

그림 3 전치사구 접속 문제에 대한 AdaBoost의 라운드 수에 따른 정확도의 변화

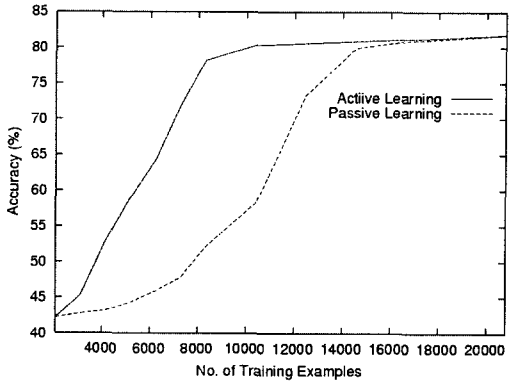

그림 4 전치사구 접속 문제에 있어서 능동 학습의 효과

그림 3은 AdaBoost의 라운드 수에 대한 정확도의 변화를 보이고 있다. 16번째 라운드 이후, 정확도는 84.3%로 올라간 후 평탄화된다. 그림 4는 전치사구 접속 문제에 있어서 능동 학습의 효과를 보인다. 이 그림의 실선은 능동 학습을, 점선은 학습예제를 주어진 순서대로 학습하는 수동 학습을 나타낸다. 따라서, 이 그림은 최대 엔트로피 모델이 학습 예제의 수에 따라 얼마나 빨리 학습되는지를 나타낸다. 즉, 두 선 사이의 차이가 능동 학습의 효과이다. 두 선이 정확도 80% 정도에서 안정되었을 때, 수동 학습이 약 14,500 개의 예제를 사용했는데 비해 능동 학습은 9,000 개의 예제만 사용하였다.

4.3 품사 결정 문제의 실험 결과

표 4는 다양한 학습 방법의 품사 결정 정확도를 보인다.

표 4 품사 결정 문제에 대한 여러 가지 방법의 성능

방 법	정확도
AdaBoost.MI	96.72%
최대 엔트로피 모델	96.89%
부스팅을 하지 않은 최대 엔트로피 부스팅 모델	96.36%
부스팅을 한 최대 엔트로피 부스팅 모델	96.78%

이 표의 'AdaBoost.MI (multiclass, independent discriminators)'는 [20]에서 사용된 것이다. 최대 엔트로피 모델은 품사 결정 문제에 있어서 가장 좋은 성능을 보이는 방법으로 [16]에서 사용된 것이다. 이 모델의 자질은 학습 데이타에 나타난 빈도수에 따라 선택된 것이다. 하지만, 최대 엔트로피 부스팅 모델은 부스팅하지 않았을 때 96.36%의 정확도를 보인다. 이는 AdaBoost.MI와 최대 엔트로피 모델보다 조금 낮은 수준이다. 하지만, 부스팅을 했을 때는 96.78%로 정확도가 올라간다. 이 모델이 AdaBoost.MI보다 높은 정확도를 보이는 이유는 이 모델이 AdaBoost.MI에서 쓰인 술어(predicate)보다 강력한 약 학습자를 사용하기 때문이다. 하지만, 실망스럽게도 이 모델은 최대 엔트로피 모델[16]보다는 낮은 성능을 보인다. 그렇지만, 이 차이는 통계적으로 의미가 없는 수준이며[3], 오히려 본 논문에서 제시한 방법은 [16]에서 고려한 모델링 비용이 들지 않는 장점이 있다.

미지어에 대한 자질의 유용성을 표 5에서 확인할 수 있다. 미지어를 위한 특별 자질을 고려하지 않았을 때에

3) 5번 반복 실험한 후, t-test로 검증하였다. 두 실험의 임계치(critical value)가 1.34이고 5% 유의 수준에서의 임계치가 1.86이므로, 두 실험의 차이는 통계적으로 의미가 없다.

는 제시된 방법의 성능이 92.19%까지 떨어진다. 이는 표 4의 다양한 기계학습 기법들보다 훨씬 낮은 수준이다. 하지만, 미지어를 위한 자질을 사용하였을 때의 성능은 지금까지 알려진 최고의 성능과 거의 비슷한 수준이다. 따라서, 품사 결정에 있어서 미지어를 위한 특별 자질은 중요하다고 할 수 있다.

표 5 미지어를 위한 자질의 유용성

	정확도
미지어를 위한 자질을 고려할 때	96.78%
미지어를 위한 자질을 고려하지 않을 때	92.19%

표 6은 태그 사전의 효과를 보인다. 품사를 결정한 각 단어에 대해 있을 수 있는 모든 품사 태그를 다 고려하면, 정확도는 0.15% 정도 떨어진다. 비록 태그 사전을 사용하여 얻을 수 있는 성능의 향상이 0.15% 뿐이라고 하더라도, 실제로는 품사를 결정하는 시간이 훨씬 짧게 걸리므로 이 방법은 매우 유용하다. 즉, 태그 사전을 사용하면 만들어질 수 있는 품사열의 수가 크게 줄게 되므로, 단어의 품사를 빠르게 결정할 수 있다.

표 6 품사 결정에서 태그 사전의 효과

	정확도
태그 사전을 사용했을 때	96.78%
태그 사전을 사용하지 않았을 때	96.63%

마지막으로, 그림 5는 품사 결정에 있어서 능동 학습의 유용성을 보인다.

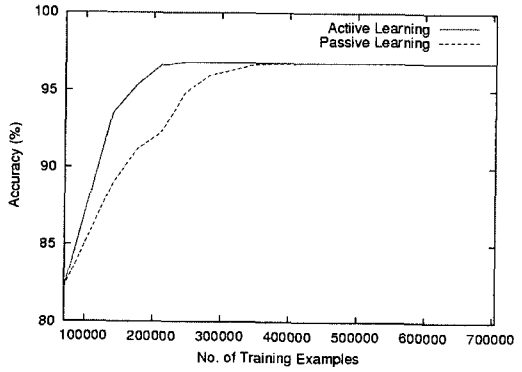

그림 5 최대 엔트로피 부스팅 모델을 품사 결정에 사용했을 때 나타나는 능동 학습의 효과

그림 4와 마찬가지로 이 그림의 실선은 능동 학습의 학습 커브를, 점선은 수동 학습의 학습 커브를 나타낸다. 품사 결정에서는 많은 데이타가 사용되므로, 정확도 커브는 양쪽 방법 모두의 경우 빠르게 올라간다. 수동 학습을 하였을 때에도 학습 데이타의 약 40%만 가지고 95.93%의 정확도를 기록했다. 그렇지만, 능동 학습의 효과도 이 그림에서 매우 확실하게 나타난다. 즉, 능동 학습 커브의 기울기가 수동 학습 커브보다 훨씬 급하다. 다시 말하면, 능동 학습을 할 때에는 학습 데이타의 25%만 사용하고도 거의 비슷한 정확도를 얻을 수 있다. 이 말은 능동 학습에 필요한 학습 데이타의 수는 수동 학습에 사용되는 것의 절반 수준임을 의미한다.

5. 결 론

본 논문에서 우리는 최대 엔트로피 부스팅 모델을 제시하고 이를 전치사구 접속 모호성 해소와 품사 결정 문제에 적용하였다. Wall Street Journal 말뭉치에 대한 실험 결과, 전치사구 접속 모호성 해소 문제에서 부스팅을 하지 않았을 때 81.8%, 부스팅을 했을 때 84.3%의 정확도를 얻었다. 이 결과는 결정트리나 자질이 사람 전문가에 의해 만들어진 최대 엔트로피 모델보다 좋은 성능이다. 특히, 결정트리보다는 부스팅을 하지 않아도 높은 성능을 보인다. 이 문제에 대해서 지금까지 보고된 최고의 성능은 back-off 모델에 의한 84.5%로, 최대 엔트로피 부스팅 모델보다 0.2% 정도 높다. 하지만, 이 차이는 통계적으로 중요하지 않다. 최대 엔트로피 부스팅 모델은 품사 결정 문제에 대해서는 96.78%의 정확도를 보였다. 또한, 우리는 태그 사전과 미지어를 위한 특별 자질의 중요성도 실험적으로 증명하였다.

두 문제 모두에서 사람이 해 주어야 할 일은 일차 자질로 무엇을 쓸 것인가를 결정하는 것에 국한된다. 즉, 좋은 성능을 얻기 위해서 좋은 고차 자질을 디자인할 필요가 없다. 따라서, 풀고자 하는 문제를 모델링하는데 비용이 거의 들지 않았다. 이렇게 비용이 거의 들지 않았음에도 불구하고 최대 엔트로피 부스팅 모델은 각 문제에 대해 지금까지 알려진 최고의 성능과 비슷한 성능을 보였다. 따라서, 특정 문제 도메인에 대한 지식 없이도 매우 뛰어난 시스템을 개발할 수 있음을 보였다.

참 고 문 헌

[1] A. Ratnaparkhi, J. Reynar, and S. Roukos, "A maximum entropy model for prepositional phrase attachment," In *Proceedings of the Human*

Language Technology Workshop, pp. 250-255, 1994.

[2] M. Collins and J. Brooks, "Prepositional phrase attachment through a backed-off model," In *Proceedings of the Third Workshop on Very Large Corpora*, pp. 27-38, 1995.

[3] E. Brill and P. Resnik, "A rule-based approach to prepositional phrase attachment disambiguation," In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 1198-1204, 1994.

[4] E. Brill, "Some advances in transformation-based part of speech tagging," In *Proceedings of the 12th National Conference on Artificial Intelligence*, pp. 722-727, 1994.

[5] H. Schmid, "Part-of-speech tagging with neural networks," In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 172-176, 1994.

[6] R. Weischedel, 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, 1994.

[7] A. Ratnaparkhi, *Maximum Entropy Models for Natural Language Ambiguity Resolution*, Ph.D thesis, University of Pennsylvannia, 1998.

[8] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," In *Proceedings of the 14th International Conference on Machine Learning*, pp. 179-186, 1997.

[9] D. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *The Annals of Mathematical Statistics*, Vol. 43, No. 5, pp. 1470-1480, 1972.

[10] T. Cover and J. Thomas, *Element of information theory*, John Wiley, 1991.

[11] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," In *Proceedings of the 13th International Conference on Machine Learning*, pp. 148-156, 1996.

[12] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," In *Proceedings of the 14th International Conference on Machine Learning*, pp. 179-186, 1997.

[13] J. Steina and M. Nagao, "Corpus based PP attachment ambiguity resolution with a semantic dictionary," In *Proceedings of the Fifth Workshop on Very Large Corpora*, pp. 66-80, 1997.

[14] P. Pantel and D. Lin, "An supervised approach to prepositional phrase attachment using contextually similar words," In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 101-108, 2000.

[15] H. Baayen adn R. Sproat, "Estimating lexical priors for low-frequency morphologically ambiguous forms," *Computational Linguistics*, Vol. 22, No. 2, pp. 155-166, 1996.

[16] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," In *Proceedings of the Empirical Methods in Natural Language Processing*, pp. 133-142, 1996.

[17] R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[18] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 35, No. 3, pp. 400-401, 1987.

[19] S. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 310-318, 1996.

[20] S. Abney, R. Schapire, and Y. Singer, "Boosting applied to tagging and PP-attachment," In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 38-45, 1999.

박 성 배

1994년 한국과학기술원 학사. 1996년 서울대학교 컴퓨터공학과 석사. 2002년 서울대학교 전기컴퓨터공학부 박사. 현재 서울대학교 컴퓨터신기술공동연구소 연구원. 관심분야는 기계학습, 자연언어처리, 정보검색, 바이오인포매틱스.