최대 엔트로피 부스팅 모델을 이용한 품사 모호성 해소

박성배°, 장병탁
서울대학교 컴퓨터공학부

Resolving Part-of-Speech Tagging Ambiguities by a Maximum Entropy Boosting Model

Seong-Bae Park° and Yung Taek Kim
School of Computer Science and Engineering, Seoul National University
°{sbpark,btzhang}@bi.snu.ac.kr

요 약

품사 결정 문제는 자연언어처리의 가장 기본적인 문제들 중 하나이며, 기계학습의 관점에서 보면 분류 문제(classification problem)로 쉽게 표현된다. 본 논문에서는 품사 결정의 모호성을 해소하기 위해서 최대 엔트로피 부스팅 모델(maximum entropy boosting model)을 이 문제에 적용하였다. 그리고, 품사 결정에서 중요한 요소 중의 하나인 미지어 처리를 위해서 특별히 설계된 일차 자질을 고려하였다. 최대 엔트로피 부스팅 모델의 장점은 쉬운 모델링인데, 실제로 품사 결정을 위한 일차 자질만 작성하는 노력만 들이고도 96.78%의 정확도를 보여 지금까지 알려진 최고의 성능과 거의 비슷한 결과를 보였다.

1. 서 론

품사 결정 문제는 자연언어처리의 가장 기본적인 문제들 중의 하나이기 때문에 많은 기계학습이나 통계 기반의 방법들이 이 문제에 적용되어 왔다. 이런 방법에 기반한 연구는 일반적으로 96% 이상의 적확도를 보고하고 있다[1]. 품사 결정 문제에서 반드시 고려해야 할 이슈 중의 하나는 미지어를 어떻게 다루느냐 하는 점이다. 규칙 기반의 방법에서는 이런 미지어를 다루기 위한 방법을 자연스럽게 둘 수 있지만, 통계 기반의 방법에서는 이들을 다루기 위한 특별한 방법을 준비하여야 한다. Weischedel et al.은 어휘 정보를 사용하는 것이 이 이슈를 해결하기 위한 좋은 방법이 될 수 있음을 보였다[1]. 이 경우, 품사 태거는 미지어의 접두사나 접미사가 특별한 태그와 함께 나타날 확률을 계산하게 된다.

이런 품사 결정 문제는 기계학습(machine learning)의 입장에서 보면 일종의 분류 문제(classification problem)이다. 그리고, 조건부 확률은 분류 문제를 위한 분류기를 구현하는 가장 좋은 방법 중의 하나이다. 조건부 확률을 계산하는 여러 확률 모델 중에서 최대 엔트로피 모델은 자연언어처리의 여러 문제에 성공적으로 적용된 모델이다. 또한, Park 과 Zhang 이 제시한 최대 엔트로피 부스팅 모델(maximum entropy boosting model)은 최대 엔트로피 모델의 장점을 살리면서도 이 모델의 주요한 문제를 해결한 모델이다[2]. 이 모델은 문서 단위화(text chunking)에 성공적으로 적용되었다. 본 논문에서는 최대 엔트로피 부스팅 모델을 품사 결정 문제에 적용한다. 실험 결과, 모델링 비용이 거의 들지 않았음에도 지금까지 알려진 최고의 정확도와 거의 비슷한 성능을 보였다.

2. 최대 엔트로피 부스팅 모델

최대 엔트로피 모델의 성능은 이 모델의 자질의 좋고 나쁨에 의해 크게 영향을 받는다. 하지만, 최대 엔트로피 모델의 자질을 구성하는 일은 쉬운 일이 아니다. 많은 경우에 최대 엔트로피 모델의 자질은 주어진 데이터의 특성을 잘 파악할 수 있는 모델러(modeler)에 의해 만들어 진다. 따라서, 만약 이 모델러가 문제 영역에 대한 충분한 지식이 없다면 자질을 구성 하는 일은 매우 어렵다. 최대 엔트로피 모델을 학습하는 데 있어 또 다른 문제는 GIS 알고리즘의 각 반복마다 모든 자질의 기대값을 추정해야 한다. 기대값을 추정할 때 모든 학습 예제에 대한 합을 계산하여야 하므로, 이 값을 추정하는 것은 자연언어와 같이 학습 예제의 수가 아주 많은 문제인 경우에 계산이

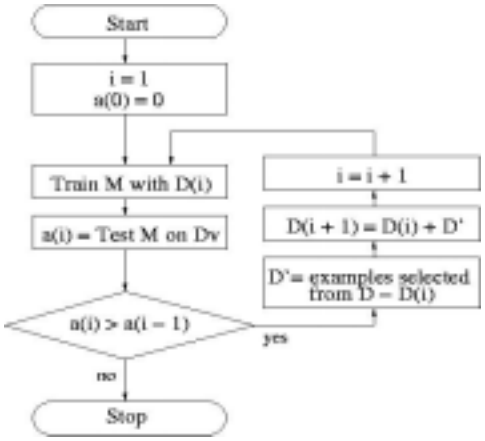

그림 1. 능동 학습 과정. 이 그림에서 M 은 학습 모델, D_v 는 검증집합, D 는 전체 학습 데이터 집합이다.

불가능할 수 있다.

Park 과 Zhang[1]은 이런 문제를 해결하기 위해서 최대 엔트로피 부스팅 모델을 제시하였다. 이 모델에서는 간단한 일차자질로부터 복잡하고 적합한 자질을 자동으로 구성하기 위해서 결정트리(decision tree)를 사용하였다. 학습된 결정트리는 쉽게 if-then 규칙의 집합으로 쉽게 변경할 수 있기 때문에 , 자질은 결정트리를 if-then 규칙으로 변환함으로써 자동으로 구성할 수 있다. 또한, 결정트리는 *n*-gram 과 같은 간단한 언어 모델로 학습될 수 있다.

그리고, 이 모델에서는 자질의 기대값을 계산할 때의 복잡도 문제를 해결하기 위해서, 능동 학습(active learning)을 기법을 사용한다. 능동 학습에서는 필요한 학습 데이터의 크기를 줄이기 위해, 전체 학습 데이터를 다 사용하기 보다는 정보량이 많은 예제를 먼저 학습한다. 그림 1 은 능동 학습 과정을 보여준다. 마지막으로, 이 모델은 자연언어 자체에 내재하는 학습 데이터의 불균형 분포 문제를 해결하기 위하여, 최대 엔트로피 모델 위에 AdaBoost 를 적용한다.

3. 최대 엔트로피 모델에 의한 품사 결정

w_1, …, w_N 을 문장 내의 단어열이라고 하자. POS 결정의 목적은 $p(t_1, \ldots, t_N \mid w_1, \ldots, w_N)$ 을 최대화하는 품사열 t_1, …, t_N 을 찾는 것이다. 이 열에 대한 완벽한 확률추정을 위해서는 너무 많은 데이터가 필요하게 되므로, 일반적으로 적당한 근사법을 이용하게 된다. 본 논문에서는 독립 가정을 이용하여, 확률을 좀 더 간단히 계산한다. 즉,

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

여기서, h_i 는 w_i 의 문맥 정보이다. 확률 $p(t_I \mid h_i)$ 는 최대 엔트로피 부스팅 모델을 이용해서 아래와 같이 계산된다.

표 1. 품사 결정을 위한 일차 자질.

미지어일 때		미지어가 아닐 때	
w_{i-2}	i-2 위치의 단어	w_{i-2}	i-2 위치의 단어
w_{i-1}	i-1 위치의 단어	w_{i-1}	i-1 위치의 단어
w_{i+1}	i+1 위치의 단어	w_i	i 위치의 단어
w_{i+2}	i+2 위치의 단어	w_{i+1}	i+1 위치의 단어
t_{i-2}	t-2 위치의 품사	w_{i+2}	i+2 위치의 단어
t_{i-1}	i-1 위치의 단어	t_{i-2}	t-2 위치의 품사
prefix(w_i, j)	길이 j 의 접두사	t_{i-1}	i-1 위치의 단어
suffix(w_i, j)	길이 j 의 접미사	w_i	i 위치의 단어
hasnumber	숫자를 가졌는가?		
hasupper	대문자가 있는가?		
hashyper	하이펀이 있는가?		

$$p(t_i \mid h_i) = \frac{1}{Z}\exp\left(\sum_i \lambda_i f_i(h_i, t_i)\right)$$

여기서, $f_i(h_i, t_i)$는 일차 자질이고 λ_i 는 f_i 의 가중치이다.

본 논문에서는 일차 자질로 왼쪽과 오른쪽의 두 단어를 문맥으로 사용한다. 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 문맥의 품사 정보는 사용하지 않았다. 또한, 미지어를 위해서 접두사와 접미사, 그리고 세 종류의 일차 자질을 추가로 사용하였다 (표 1). 어떤 단어가 미지어인지를 결정하는 근거는 이 단어가 학습 집합에서 3 번 이하로 나타났는지를 보는 것이다. 미지어를 위한 접두사나 접미사의 길이로는 3 을 사용하였다.

N 이 문장 길이이고 T 가 태그 집합일 때, 가능한 품사열의 수는 $N^{|T|}$. 가장 좋은 품사열을 찾기 위한 검색 공간을 줄이기 위해, 폭 검색(beam search)를 사용하였다. M 을 폭 크기, t_{ij} 를 단어 w_i 의 j 번째로 확률이 높은 품사 태그라고 하자. 새 단어를 만날 때마다 M 개의 가장 좋은 열을 유지한다. s_{ij} 를 w_i 까지 j 번째로 좋은 품사열이라고 하자. 각 열 $s_{(i-1)j}$ 마다 w_i 의 M 태그가 부착되기 때문에, M^2 개의 열이 생성된다. M^2 개의 태그열 중에서 가장 좋은 M 개만 남기고 나머지는 버린다. 따라서, 결국 폭 검색에서는 $M \cdot |T|$ 개만 고려하게 된다. 아래의 모든 실험에서는 M=3 으로 하였다.

4. 실 험

4.1 데이터집합

품사 결정 실험을 위해서 Penn Treebank II 의 Wall Street Journal 말뭉치를 사용하였다. 이 데이터집합은 1,173,765 단어, 49.206 개의 어휘, 45 개의 품사 태그를 가지고 있다. 이 데이터집합을 세 부분으로 나누었다. 60%인 704,251 단어를 학습 집합으로, 20%인 234,819 단어를 검증 집합으로, 나머지 20%인

표 2. 품사 결정 문제의 성능.

방법	정확도
AdaBoost.MI	96.72%
최대 엔트로피 모델	96.89%
결정트리를 통한 최대 엔트로피 모델	96.36%
최대 엔트로피 부스팅 모델	96.78%

표 3. 미지어를 위한 자질의 유용성.

	정확도
미지어를 위한 자질을 고려했을 때	96.78%
미지어를 위한 자질을 고려하지 않았을 때	92.19%

234,695 단어를 테스트 집합으로 사용하였다. 학습 집합에 있는 미지어의 수는 23,237 개이고, 접두사의 수는 3,713, 접미사의 수는 3,199 개이다. 테스트 집합에는 있지만 학습 집합에는 나타나지 않은 단어의 수는 4,578 개이다.

4.2 실험 결과

표 2 는 다양한 방법과 제시된 방법의 성능을 비교한 것이다. 이 표의 AdaBoost.MI 는 [3]에서 제시된 방법으로 AdaBoost 의 메모리 한계를 뛰어넘기 위한 방법이다. 본 논문에서 제시된 방법과의 차이는 기저 분류기로 “attribute=value”의 술어를 사용하였다는 점이다. 최대 엔트로피 모델은 품사 결정 문제에 있어서 가장 높은 성능을 보이는 학습 모델이다[4]. 이 모델에서는 자질이 학습 집합에서의 빈도에 의해 선택된다.

본 논문에서 제시된 방법은 부스팅을 하지 않고도 96.36%의 정확도를 보였다. 이는 ‘AdaBoost.MI’나 ‘최대 엔트로피 모델’보다 조금 낮은 수준이지만, 부스팅을 한 후에는 96.78%로 정확도가 높아졌다. 같은 학습 방법인 ‘AdaBoost.MI’보다 성능이 높아진 이유는 우리가 사용한 기조 분류기가 술어보다 훨씬 더 강력하기 때문인 것으로 여겨진다. 하지만, 불행하게도 본 논문에서 제시된 방법은 최대 엔트로피 모델보다 조금 낮은 정확도를 보인다. 그렇지만, 이 정확도 차이는 통계적으로 의미가 있는 수준은 아니다.

표 3 은 미지어에 대한 자질의 유용성을 보이고 있다. 이들을 위한 자질을 고려하지 않았을 때는 정확도가 92.19%에 머물러 다른 기계학습 방법보다 훨씬 낮은 성능을 보인다. 따라서, 미지어를 위해서 특별한 자질을 구성하는 것이 중요하 다고 할 수 있다

마지막으로, 그림 2 는 품사 결정에서의 능동학습의 유용성을 보인다. 능동학습을 하지 않아도 전체 학습 예제의 약 40%만 쓰고도 학습은 거의 끝난다. 하지만, 능동학습을 하였을 때에는 약 25%만 쓰고도 학습이 끝나므로 능동학습의 효과는 매우 확실하다.

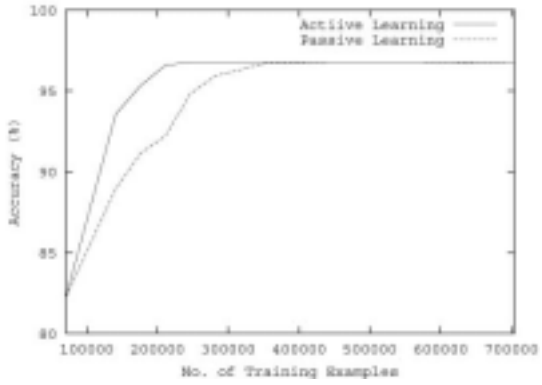

그림 2. 품사 결정 문제에 있어서 능동학습의 효과.

5. 결 론

본 논문에서는 최대 엔트로피 부스팅 모델을 품사 결정 문제에 적용하였다. Wall Street Journal 말뭉치에 대한 실험 결과, 이 모델은 96.78%의 정확도을 보여, 지금까지 알려진 가장 좋은 알고리즘의 정확도와 거의 차이가 없는 성능을 보였다. 하지만, 이 모델을 구성하기 위한 사전 지식이 거의 들지 않았고, 최대 엔트로피 부스팅 모델의 과도한 계산량도 능동 학습을 통해 해소되었다.

감사의 글

이 논문은 과기부 BrainTech 프로그램과 교육부 BK 21 사업에 의하여 지원되었음.

참고문헌

[1] R. Weischedel, M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmucci, “Coping with ambiguity and unknown words through probabilistic models,” *Computational Linguistics*, 19(2), pp. 359-382, 1994.

[3] S.-B. Park and B.-T. Zhang, “A boosted maximum entropy model for learning text chunking,” In *Proceedings of the 19th International Conference on Machine Learning*, pp. 482-489, 2002.

[3] 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.

[4] 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.