최대 엔트로피 부스팅 모델을 이용한 전치사 접속 모호성 해소

박성배°, 장병탁
서울대학교 컴퓨터공학부

Resolving Prepositional Phrase Attachment Using 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

Park 과 Zhang 은 최대 엔트로피 모델(maximum entropy model)을 실제 자연언어 처리에 적용함에 있어서 나타날 수 있는 여러가지 문제를 해결하기 위한 최대 엔트로피 모델(maximum entropy boosting model)을 제시하여 문서 단위화(text chunking)에 성공적으로 적용하였다. 최대 엔트로피 부스팅 모델은 쉬운 모델링과 높은 성능을 보이는 장점을 가지고 있다. 본 논문에서는 최대 엔트로피 부스팅 모델을 영어 전치사 접속 모호성 해소에 적용한다. Wall Street Journal 말뭉치에 대한 실험 결과, 아주 작은 노력을 들였음에도 84.3%의 성능을 보여 지금까지 알려진 최고의 성능과 비슷한 결과를 보였다.

1. 서 론

전치사 접속 문제는 자연언어처리의 가장 어려운 문제들 중의 하나이다. 이 문제는 문장 내에 나타나는 전치사의 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.

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

이런 전치사구 접속 문제는 기계학습(machine learning)의 입장에서 보면 일종의 분류 문제(classification problem)이다. 이 문제의 목적은 (v, n_1, p, n_2) 형태로 주어진 4-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, 첫번째 문장에 해당하는 튜플은 (bought, shirt, *with*, pockets)가 되고, 두번째 문장에 해당하는 튜플은 (washed, shirt, *with*, soap)가 된다. 그리고 첫번째 문장의 올바른 접속은 N이고, 두번째 문장은 V이다.

2. 최대 엔트로피 부스팅 모델

분류문제의 목적은 언어학적 문맥 정보 $\mathbf{x} \in X$를 관찰한 후, $\mathbf{x}$의 클래스 $y \in Y$ 를 추정하는 것이다. 전치사 접속 문제에 있어서 X 는 (v, n_1, p, n_2)로 이루어진 4 차원 벡터들의 집합이고, Y는 $\{N, V\}$이다. 조건부 확률 $p(y \mid \mathbf{x})$는 y를 추정하는 분류기(classifier)를 구현하는 좋은 방법 중의 하나이다. 그리고, 최대 엔트로피 모델은 조건부 확률을 추정할 수 있으면서도 다양한 문맥 정보를 결합할 수 있는 장점을 가지고 있다.

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 다

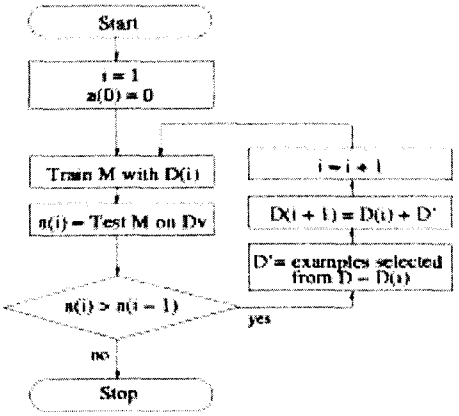

그림 1. 능동 학습 과정. 이 그림에서 M 은 학습 모델, D_v는 검증집합, D는 전체 학습 데이터 집합이다.

른 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 불가능할 수 있다.

Park 과 Zhang[1]은 이런 문제를 해결하기 위해서 *boosted maximum entropy model* 을 제시하고, 이 모델이 문서 단위화에 있어서 매우 효율적임을 보였다. 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[2]를 적용한다. 학습 데이터의 불균형은 대부분의 기계학습 알고리즘의 재현도를 낮춤으로써 성능을 떨어뜨리는 원인이 된다[3]. AdaBoost 가 각 반복마다 분류하기 어려운 예제에 집중하기 때문에, 작은 클래스에 점점 더 집중하게 된다. 즉, AdaBoost 나 support vector machine 과 같이 마진(margin)에 기초한 알고리듬들은 초평면(hyperplane) 부근의 예제에 집중하기 때문에, 작은 클래스를 더 많이 샘플링하는 편중(bias)의 효과를 갖게 된다. 또한, AdaBoost 는 일종의 위원회 모델(committee model)이기 때문에 높은 재현도 뿐만 아니라 더 좋은 정확도도 기대할 수 있다[4].

3. 최대 엔트로피 모델에 의한 전치사 접속 모호성 해소

위에서 언급한 바와 같이, 전치사 접속 문제는 분류문제로 생각될 수 있다. 즉, 이 문제는 다음과 같이 각 접속의 확률을 비교하는 것으로 공식화될 수 있다.

$$f(v,n_1,p,n_2) = \arg\max_{y\in\{N,V\}} p(y \mid v,n_1,p,n_2)$$

최대 엔트로피 모델이 확률 모델이기 때문에 $p(y \mid v, n_1, p, n_2)$은 이 모델에 의해 쉽게 추정될 수 있다. 여기서, v, n_1, p, n_2은 최대 엔트로피 모델을 위한 일차자질이 된다. 즉,

$$p(y \mid v,n_1,p,n_2) = \frac{1}{Z}\exp\left(\sum_i \lambda_i f_i(v,n_1,p,n_2)\right)$$

여기서, Z 는 정규화상수이고, f_i 는 일차자질로 구성된 고차자질들이다.

우리는 다른 어휘 정보나 의미 사전을 사용하지 않았다. 이런 부가적인 정보를 사용하면 정확도가 증가되기는 하나[6], 학습 알고리즘의 성능을 공정하게 비교하기 위해서 부가 정보의 사용을 배제하였다.

4. 실 험

4.1 데이터집합

전치사 접속 모호성 해소를 위한 데이터집합[1]은 [5]에서 사용된 것을 사용하였다. 이 데이터는 Penn Treebank Wall Street Journal 의 구문태그 부착 말뭉치에서 추출된 것으 로, 20801 개의 학습 예제와 3097 개의 테스트 예제로 이루어져 있다. 각 예제는 4-tuple 과 목적 클래스로 이루어져 있다. 즉, v, n_1, p, n_2, y 로 이루어져 있으며, y 는 N 혹은 V 이다. 이 집합에는 4039 개의 예제로 이루어진 독립적인 개발 집합(development set)이 있기 때문에, 우리는 이를 능동 학습 시의 검증 집합으로 사용하였다.

이 데이터 집합은 많은 오류를 포함하고 있다. 예를 들어, 테스트 집합의 133 개의 예제가 the 를 n_1 이나 n_2 에 포함하고 있다. 또한, (*sing, birthday, to, you, N*)과 같이 잘못된 접속 정보를 가지는 예제도 포함하고 있다. 하지만, 본 논문의 목표가 최대 엔트로피 부스팅 모델을 다른 기계학습

[1] 이 데이터집합은 다음 사이트에서 구할 수 있다:
ftp://fp.cis.upenn.edu/pub/adwait/PPattachData

표 1. 전치사 접속 모호성 해소의 정확도.

알고리즘	정확도
Baseline	70.4%
결정트리	80.2%
최대 엔트로피 모델	77.7%
Back-off 모델	84.5%
최대 엔트로피 부스팅 모델	84,3%

알고리즘들과 비교하는 것이므로, 우리는 이런 오류들을 고치지 않았다.

4.2 실험 결과

최대 엔트로피 부스팅 모델의 성능을 알아보기 위해 우리는 다음과 같은 네 종류의 알고리즘과 비교하였다: (i) baseline, (ii) 결정트리, (iii) 최대 엔트로피 모델, (iv) back-off 모델. Baseline 모델은 다음과 같이 정의된다.

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

이 수식은 전치사 of 만 많은 경우에 명사를 수식하고, 나머지 전치사들은 동사를 수식하는 경향을 가졌기 때문에 만들어졌다. 이 실험에 사용된 결정트리는 C4.5 release 8 이다.

표 1 은 각 분류기의 정확도를 보인다. 최대 엔트로피 부스팅 모델은 84.3%로 지금까지 알려진 최고의 성능인 back-off 모델[7]과 거의 비슷하며, 일반 최대 엔트로피 모델이나 결정트리보다 훨씬 좋은 성능을 보였다. 심지어, 부스팅을 하지 않았을 때에도 이 두 학습 알고리즘보다 좋은 성능인 81.8%의 성능을 보였다.

그림 2 는 전치사 접속 문제에 있어서의 능동 학습의 효과를 보여준다. 이 그림의 'Active Learning'은 능동 학습 기법을 사용하였을 때의 정확도 변화를 보이고 있고, 'Passive Learning'은 학습 예제를 주어진 순서대로 학습하였을 때의 정확도 변화를 보이고 있다. 이 두 라인 사이의 차이가 능동 학습 기법의 효과이다. 즉, 정확도 80% 근처에서 정확도가 더 이상 좋아지지 않을 때까지 사용된 학습 예제의 수는 능동 학습이 9,000 개 정도인 데 비해, 수동학습은 무려 14,500 개 정도를 사용하였다.

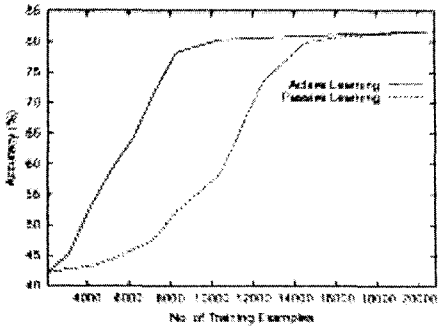

그림 2. 전치사 접속 문제에 있어서 능동학습의 효과.

5. 결 론

본 논문에서는 최대 엔트로피 부스팅 모델을 전치사 접속 모호성 해소 문제에 적용하였다. Wall Street Journal 말뭉치에 대한 실험 결과, 이 모델은 84.3%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] Seong-Bae Park and Byoung-Tak Zhang, "A Boosted Maximum Entropy Model for Learning Text Chunking," In *Proceedings of the 19th International Conference on Machine Learning*, pp. 482-489, 2002.

[2] 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.

[3] 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.

[4] K. Tumer and J. Ghosh, "Analysis of Decision Boundaries in Linearly Combined Neural Classifiers," *Pattern Recognition*, Vol. 29, 341-348, 1996.

[5] 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, 194.

[6] P. Pantel and D. Lin, "An Unsupervised 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.

[7] 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.