능동 학습을 이용한 학습 데이터 반자동 구축

이창기, 허정, 왕지현, 이충희, 오효정, 장명길, 이영직
한국전자통신연구원
음성언어정보연구센터
{leeck,jeonghur,jhwang,forever,ohj,mgjang,ylee}@etri.re.kr

Semi-automatic Construction of Training Data using Active Learning

Changki Lee, Jeong Hur, Ji Hyun Wang, Chung Hee Lee, Hyo-jung Oh,
Myung Gil Jang, YoungJik Lee
Electronics and Telecommunications Research Institute
Knowledge Mining Laboratory

요 약

본 논문은 정보검색, 정보추출, 번역, 자연어처리 등의 작업을 위한 통계적 방법론에서 필요한 학습 데이터 구축을 효율적으로 하기 위한 학습 데이터 반자동 구축 장치 및 그 방법에 대하여 기술한다.

본 논문에서는 학습 데이터 구축양을 줄이기 위해서 능동 학습을 이용한다. 또한 최근 각광 받고 있는 Conditional Random Fields(CRF)를 능동학습에 이용하기 위해서 CRF 를 이용한 Confidence measure 를 정의한다.

Keyword : Active Learning, Machine Learning, Conditional Random Fields, Maximum Entropy

1. 서론

정보검색, 정보추출, 번역, 자연어처리 등의 작업등에 최근 통계적 방법론이 많이 사용되고 있다. 이러한 통계적 방법론은 각 작업에 따른 학습 데이터 구축을 필요로 하고, 학습 데이터 구축의 양이 많을수록 높은 성능을 낸다. 그러나 학습 데이터의 구축은 많은 인력과 시간이 필요 하기 때문에 학습 데이터 부족 문제가 자주 발생한다.

본 논문에서는 학습 데이터 구축을 효율적으로 하기 위해서 Conditional Random Fields(CRF)를 이용하여 자동으로 태그를 부착하여 학습 데이터 후보를 생성하고 신뢰점수를 계산하고, 능동 학습을(Active learning) 이용하여 학습 데이터 후보를 선택하여 학습 데이터 구축의 비용 및 시간을 줄이기 위한 방법을 기술한다.

2. 관련 연구

학습 데이터 부족 문제를 해결하기 위해서 기존의 방법들은 크게 세 가지 방법론으로 분류 된다. 첫 번째 방법론은 기계 학습을 이용하여 자동 태깅 기능을 지원하는 워크벤치를 사용하는 방법이다. 이는 자동 태깅을 지원하는 점은 본 논문과 비슷하지만 학습 데이터 후보 선택이나 오류 수정된 학습 데이터를 재 사용하여 학습 데이터를 증강시켜 점진적으로 자동 태깅 성능을 높여주는 기능은 지원하지 않는다. 두 번째 방법은 Bootstrapping 이나 Co-training 같은 방법으로 학습 데이터를 자동으로 태깅하여 학습 데이터를 증강시키는 방법이다 [1]. 이는 본 논문과 기능은 비슷하나 자동 태깅 결과의 오류 수정이나 학습 데이터 후보 선택 등의 기능은 지원하지 않는다. 또한 자동 태깅 결과에 오류가 포함되어 있기 때문에 성능이 좋지 않다는 단점이 있다. 세 번째 방법론

은 능동 학습(Active learning) 방법이다 [2]. 이 방법은 최적의 학습 데이터 후보를 선택하여 적은 양의 학습 데이터를 구축해도 높은 성능을 내도록 한다. 본 논문이 여기에 속하며, 본 논문에서는 최근 각광을 받고 있는 Conditional Random Fields(CRF)를 능동학습에 적용한다.

3. Conditional Random Fields

최근 품사 태깅, 청킹, 개체명 인식 등과 같은 Sequence Labeling 문제에 Conditional Random Fields(CRF)가 많이 연구되고 있으며 기존의 HMM 및 MEMM 등에 비해 높은 성능을 보이고 있다 [3][4][5]. CRF 는 방향이 없는 그래픽 모델(undirected graphical model)로써, 관측 가능한 입력 데이터 열이 주어지고, 이에 따른 결과 열의 조건부 확률을 계산하는데 사용된다. 특히 선형 구조의 CRF (linear chain structured CRF)가 주로 쓰이고 있다 [3].

o = <o_1, o_2, …o_T>는 관측 가능한 입력 데이터 열이라 하고, **s** = <s_1,s_2,…s_T>를 결과 열이라 하면, 주어진 입력 데이터 열 **o**에 대한 결과 열 **s**가 나올 조건부 확률은 다음과 같이 정의된다.

$$P_\Lambda(\mathbf{s}\mid\mathbf{o}) = \frac{1}{Z_o}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(s_{t-1}, s_t, \mathbf{o}, t)\right), \quad \textbf{(1)}$$

위 식에서, Zo는 모든 결과 열의 확률값의 합이 1 이 되도록 만드는 normalization factor이고, $f_k(s_{t-1},s_t,o,t)$는 feature 함수로 0 과 1 의 값을 반환한다. λ_k는 각 feature 함수의 weight로써 이 값은 학습에 의해 결정된다.

Zo 는 모든 모든 결과 열의 합으로써, 다음과 같이 정의된다.

$$Z_o = \sum_{s\in S^T}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(s_{t-1}, s_t, \mathbf{o}, t)\right). \quad \textbf{(2)}$$

입력 데이터 열의 길이에 따른 결과 열의 수는 기하급수적으로(exponential) 증가하게 되는데, 선형 구조의 CRF(linear chain structured CRF)인 경우 HMM 의 forward-backward 알고리즘을 이용하여 확률 값을 효율적으로 계산할 수 있다 [3].

CRF의 각 feature의 weight {λ…}는 학습 데이터에 있는 입력 데이터 열과 결과 열의 conditional log-likelihood 값을 최대화 시키도록 결정된다. D={<**o,l**>$^{(1)}$,…,<**o,l**>$^{(N)}$}를 학습 데이터에 있는 입력 데이터 열(o)과 결과 열(l)이라 할 때, conditional log-likelihood 값은 다음과 같이 정의된다.

$$L_\Lambda = \sum_{j=1}^{N}\log\left(P_\Lambda(\mathbf{l}^{(j)}\mid\mathbf{o}^{(j)})\right) - \sum_{k}\frac{\lambda_k^2}{2\sigma^2} \quad \textbf{(3)}$$

위 식에서, 두 번째 부분은 데이터 부족으로 인한 문제를 해결하기 위한 것이다.

Conditional log-likelihood 값을 최대화 시키는 feature weight {λ…}를 구하기 위해 GIS 와 L-BFGS 같은 방법이 있는데, L-BFGS 가 좀더 효율적이라고 밝혀졌다 [5].

L-BFGS 는 black-box optimization 모듈로써, 입력으로 최적화할 함수의 미분 함수를 받는다. CRF 의 학습의 경우에는 최적화할 함수가 수식 3 번인 Conditional log-likelihood 함수가 되고, 이의 미분 함수는 다음과 같다.

$$\frac{\delta L}{\delta\lambda_k} = \left(\sum_{j=1}^{N} C_k(\mathbf{s}^{(j)}, \mathbf{o}^{(j)})\right) - \left(\sum_{j=1}^{N}\sum_{s} P_\Lambda(\mathbf{s}\mid\mathbf{o}^{(j)}) C_k(\mathbf{s}, \mathbf{o}^{(j)})\right) - \frac{\lambda_k}{\sigma^2} \quad \textbf{(4)}$$

위 식에서 C_k(**s,o**)는 feature 함수 f_k의 값의 1 인 경우의 회수이다.

4. CRF 를 이용한 능동 학습

능동 학습(Active learning)이란, 좀더 적은 양의 학습데이터로 학습하여 높은 성능을 내기 위해서 학습 데이터를 선택하여 사용하는 것이다 []. 본 논문에서 사용하는 능동 학습 알고리즘은 다음과 같다.

1. 초기 seed 학습 데이터로 학습
2. 현재 학습 모델로 태깅이 어려운 문장을 골라서 사용자에게 추천 (Confidence 값이 낮은 것을 선택)
3. 사용자가 추천 문장을 태깅하여 학습 데이터에 추가
4. 변경된 학습 데이터로 학습
5. 2 번으로 이동

Fig. 1. 능동 학습 알고리즘.

본 논문에서는 최근 각광을 받고 있는 CRF 를 기본 기계 학습기(base learner)로 사용한다. 또한 능동 학습을 위해 필요한 CRF 의 Confidence 값을 다음과 같이 정의한다.

- CRF-Prob: 한 문장의 입력 데이터 열(o)와 결과 열(s)가 주어 졌을 때, 현재 모델에서 best 결과 열 s 가 생성될 확률. 다음과 같이 정의된다.

$$Conf_{prob}(\mathbf{o}) = P_{\Lambda}(\mathbf{s}|\mathbf{o}).$$

- CRF-Entropy: 입력 데이터 열(o)의 N-best 결과들의 확률값의 entropy 에 -1 을 곱한 값. 다음과 같이 정의 된다.

$$Conf_{entropy}(\mathbf{o}) = \sum_{s \in N-best} P_{\Lambda}(\mathbf{s}|\mathbf{o}) \log P_{\Lambda}(\mathbf{s}|\mathbf{o}).$$

- CRF-KL: 입력 데이터 열(o)의 N-best 결과들의 확률값의 Kullback-Leibler divergence (D(*||*)) 값. 다음과 같이 정의된다.

$$Conf_{KL}(\mathbf{o}) = \frac{1}{N} \sum_{s \in N-best} D(P_{\Lambda}(\mathbf{s}|\mathbf{o}) \| P_{avg})$$

$$P_{avg} = \frac{1}{N} \sum_{s \in N-best} P_{\Lambda}(\mathbf{s}|\mathbf{o})$$

5. 실험 및 결과

앞에서 정의한 능동학습을 위한 CRF 의 세가지 Confidence measure 에 대해서 어느 것이 성능이 높은지를 알기 위해서 각 measure 별로 능동학습을 수행하였다. 능동학습을 수행 대상은 정보 추출 분야이다.

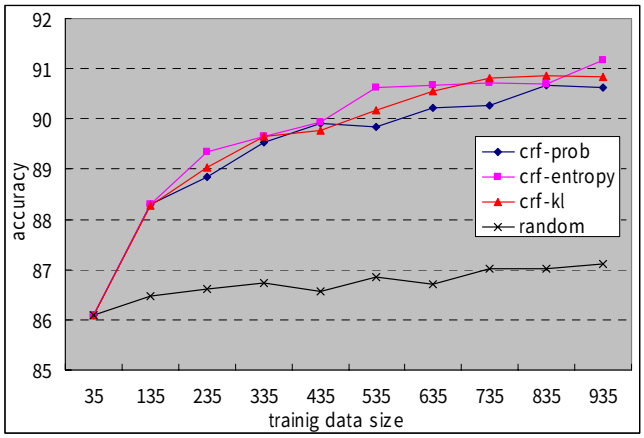

Fig. 2. 능동 학습 결과.

그림 2 는 정보 추출 분야에서 각 Confidence measure 별 능동 학습을 사용한 결과이다. 세로축은 정보 추출의 정확도(accuracy) 이고, 가로축은 학습 데이터의 크기이다. Random 은 능동 학습을 이용하지 않고 임의로 학습 데이터를 선택해서 태깅하여 학습 데이터에 추가할 경우이다. 3 개의 Confidence measure 모두 Random 보다 높은 성능을 보여, 능동 학습이 임의 선택보다 적은 양의 학습 데이터로 높은 성능을 보임을 알 수 있다. Confidence measure 로 CRF-Entropy 를 사용했을 경우가 전반적으로 성능이 좋으며, CRF-KL 을 사용했을 때가 가장 안정적으로 성능이 상승했다. 그러나 CRF-Entropy 와 CRF-KL 은 N-best 결과를 사용하기 때문에 속도가 느려지는 단점이 있다. 속도를 고려했을 경우에는 CRF-Prob 가 가장 빠르고, CRF-Entropy 와 CRF-KL 의 성능에 큰 차이가 없으므로 CRF-Prob 를 선택하는 것이 좋음을 알 수 있다.

그림 3 은 CRF 와 능동 학습을 이용한 학습 데이터 반자동 구축기의 사용자 인터페이스이다. 학습 데이터 반자동 구축기는 CRF 와 능동 학습을 이용하여 학습 데이터 후보를 선정한 후, 사용자에게 자동으로 태깅된 결과를 보여주고, 사용자가 오류를 수정하면, 이를 다시 학습 데이터에 추가하여, 이 과정을 반복하게 하게 된다.

6. 결론

본 논문에서는 학습 데이터 구축을 효율적으로 하기 위해서 능동학습을 이용한 학습 데이터 반자동 구축 도구를 구축하였다. 능동 학습의 기본 학습기로는 최근 각광 받고 있는 CRF 를 이용하였다. CRF 를 능동학습에 이용하기 위하여 세가지 confidence measure 를 정의하고 각각에 대하여 능동학습 실험을 수행하였다. 실험 결과로부터 능동 학습이 임의 선택보다 적은 양의 학습 데이터로 높은 성능을 보임을 알 수 있었다. 성능은 CRF-Entropy 가 가장 좋았고, 안정성은 CRF-KL 이 가장 좋았으며, 수행 속도는 CRF-Prob 가 가장 빠름을 알 수 있었다.

References

1. Michael Collins, Yoram Singer. *Unsupervised Models for Named Entity Classification*, EMNLP, 1999.
2. Gokhan Tur, et al., Steven J. Maiorano. *Active Learning for Spoken Language Unerstanding*, ICASSP, 2003.
3. J. Lafferty, A. McCallum, and F. Pereira. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. ICML, 2001.
4. David Pinto, A. McCallum, Xing Wei, and W. Bruce Croft. *Table Extraction Using Conditional Random Fields*. SIGIR, 2003.
5. S. Fei, F. Pereira. *Shallow Parsing with Conditional Random Fields*, HLT & NAACL, 2003.

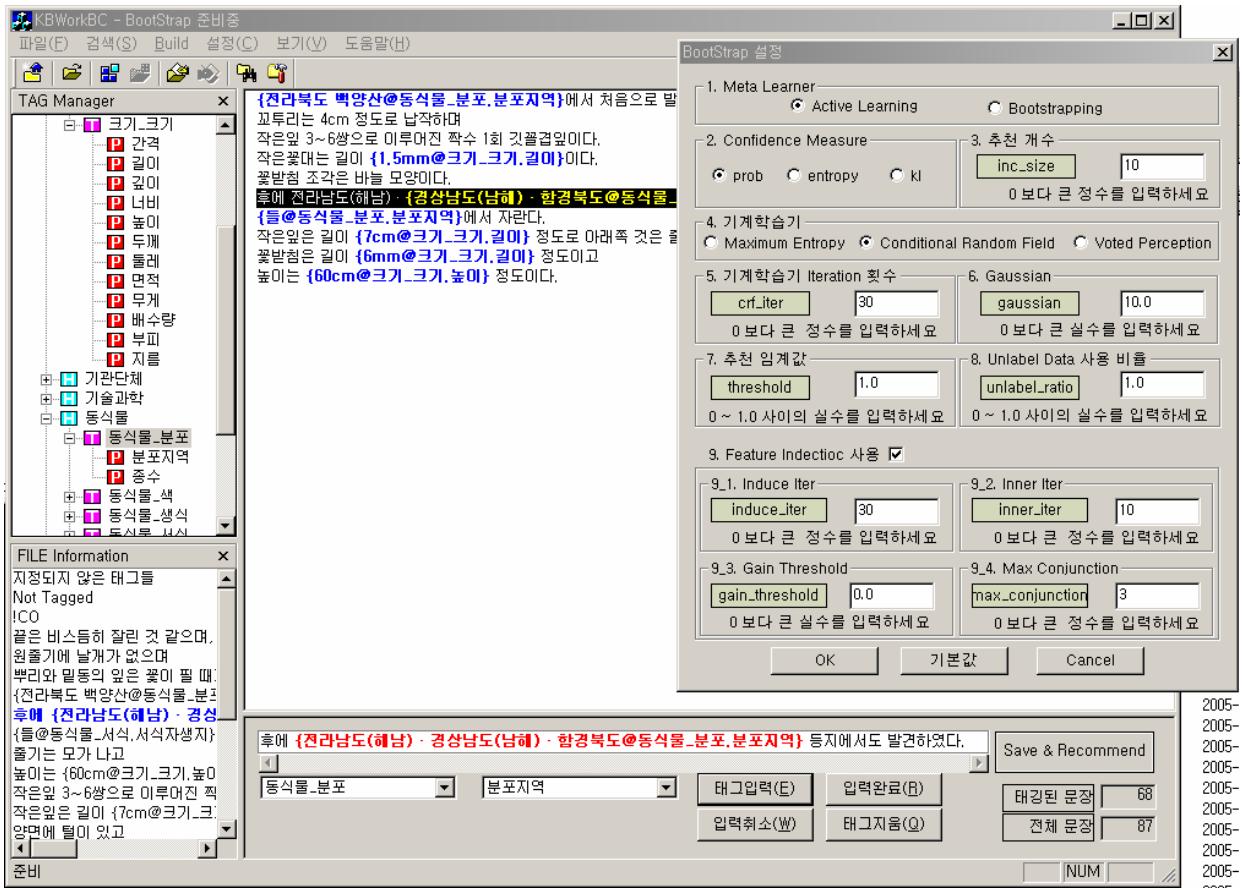

Fig. 3. 학습 데이터 반자동 구축 도구