S-HMM을 이용한 텍스트 정보추출

엄재홍[0] 장병탁
서 울 대 학 교 컴 퓨 터 공 학 부
{jheom, btzhang}@bi.snu.ac.kr

Information extraction with S-HMM from textual data

Jae-Hong Eom *Byoung-Tak Zhang*
School of Computer Science and Engineering, Seoul National University

요 약

본 논문에서는 패턴이나 음성데이터와 같이 순차적 데이터를 인식하는데 널리 사용되어온 모델로서, 일련의 순차적인 성질을 내포하고있는 데이터를 다루는 문제에 적합하다고 할 수 있는 HMM을 이용하여 정보추출 문제를 다룬다. 기본적으로는 통상적인 HMM 사용법을 따르나 모델의 구조를 정함에 있어서 HMM을 사용할 때는 주로 목적에 맞는 HMM의 구조를 수동으로 구성하고 모델 내부의 확률 파라미터 값을 학습시켰던 데 반해, 본 논문에서는 데이터의 전처리 정보를 이용하여 초기에 추상적으로 설정한 모델이 학습을 통해서 점차 구체화되어 가는 자기 구성 은닉마르코프 모델(S-HMM)을 제시하여 사용한다. 제시된 방법은 CFP(Call for Paper)등의 텍스트 데이터에 대한 실험에서 기존 방식을 사용한 HMM보다 향상된 결과를 보여준다.

1. 서론

인터넷이 발달함에 따라 많은 양의 정보들에 대한 온라인 접근이 그 어느 때보다도 쉬워지고 있으며, 신문이나 잡지 학술논문 등의 각종 정보들이 디지털화되어 온라인상 문서의 형태로도 존재하게 되었다. 이러한 정보의 증가로 인한 정보 과부하(information overload)는 사용자들로 하여금 모든 정보를 소화하기 힘들게 만들고 있다. 이에 정보검색 분야가 온라인 문서의 증가와 함께 활발히 연구되어 사용자들이 인터넷 상에서 필요한 문서를 찾는데 많은 도움을 주고 있다. 또한, 많은 수의 문서를 접근하는 데에서 오는 어려움을 덜기 위하여 사용자들의 관심도를 학습해 사용자의 개입 없이 정보를 여과하여주는 여러 가지 에이전트 시스템들도 등장하고 있다. 하지만 늘어나는 문서를 효율적으로 검색한 후, 또는 원하는 분야의 문서를 에이전트 시스템이 여과(filtering)하여 준 후에는, 결국 사용자가 문서 내에서 찾으려고 하는 특정 필드의 정보가 있게 마련인데 이것은 검색 결과나 여과의 결과물이 다른 문서와 마찬가지로 방대한 경우 사용자들을 곤란하게 만든다. 이에 정보검색이나 여과의 결과물에서 원하는 필드를 자동으로 추출하는 시스템이나 이를 위한 방법에 관한 정보추출(information extraction) 연구의 필요성이 대두되게 되었다[1].

정보추출에 대한 필요성이 널리 인식된 후, 은닉마르코프모델이나 문법규칙(grammar rule), 기계학습(machine learning), 여러 가지 자연언어처리 기법 등의 방법을 이용한 정보추출에 관한 여러 연구가 진행되었다. 기존에 음성인식에 널리 사용되어 오던 은닉마르코프모델이 정보검색이나 정보추출 분야에서도 사용되기 시작하였다. 이러한 예로는 문서가 내부구조로 가지고있는 주제별 구조를 은닉마르코프모델로 모델링하여 사용자가 찾으려고 하는 정보와 어느 정도 유사한 문서인지를 판단하여 정보검색을 수행하는 HMM 정보검색 시스템[7,9]을 들 수 있다. 또한, 정보추출 방법의 예로 HMM을 사용하지는 않았지만 형식화된 간단한 온라인 세미나 안내 자료의 시간, 장소, 연사 등과 같은 각각의 주요 필드에 대하여 여러 가지 기계학습 기법들을 사용하여 구성한 학습자(learner)들을 학습시켜 이들의 예측 결과를 regression을 이용하여 병합하여 필드의 추출을 수행하는 다중기법(multistrategy)을 사용한 정보추출 방법을 들 수 있다[6].

이러한 여러 가지 방법들 중에서 은닉마르코프모델은 순차적 성질을 가지는 데이터를 인식하는데 널리 사용되어온 모델이다. 여러 문서들 중에서 학회참가요청(Call for Papers) 문서나 세미나 안내문서, 또는 논문과 같은 문서들은 어느 정도 정형화된 형식을 따르고있기 때문에 문서 내부적으로 어떤 순차적 성질을 가진 문서로 고려할 수 있다. 이 때문에, 음성 데이터와 같이 명확한 순차적 성질을 가지는데 응용된 은닉마르코프모델을 위와 같은 문서에 사용하여 정보추출의 문제를 해결할 수 있는 것이다. 그러나 텍스트 문서에 응용된 기존의 은닉마르코프모델은 초기 모델의 세부 구조를 모델 설계시점에서 직접 지정해야만 했으며, 시스템의 모델 구조는 이때 정적으로 정해지게 되었다. 이렇게 됨으로써 처리하는 데이터와, 필드의 특징에 따른 적합한 모델 구조를 학습할 수 없었으며 단지 초기에 정해진 제한적 모델 구조의 확률값 조정을 통하여 데이터의 특징에 대한 학습을 수행할 수 있었다.

본 논문에서는 스스로 초기에 정해진 모델의 구조를 변경할 수 있는 은닉마르코프모델을(Self-Organizing Hidden Markov Model: S-HMM) 사용하여 정보추출의 문제를 다룬다. S-HMM은 모델 구조의 제한을 두지 않은 HMM방법으로 전처리로 얻어진 데이터의 특징을 이용하여 모델이 스스로 구조를 학습하는 HMM을 이용하여 정보추출의 문제를 해결하는 방법을 제안한다.

2. HMM과 정보추출

2.1 은닉 마르코프 모델 (HMMs)

HMM은 관찰이 불가능한 미지(hidden)의 확률론적 과정(stochastic process)을 관찰이 가능한 기호(symbol)를 발생시키는 다른 확률론적 과정을 통하여 모형화(modeling) 하는 이중 확률론적 과정으로 다음과 같이 2개의 상태집합과 3개의 확률 집합으로 구성된다.

- 은닉상태집합(hidden state set): 마르코프 프로세스에 의해 설명되는 상태들의 집합
- 관찰가능 상태집합(observable state set): 외형적으로 눈에 보이는 전이 상태들의 집합
- π벡터: 특정 은닉 상태가 시간 $t=1$일 때의 모델의 확률
- 상태전이 행렬: 이전의 은닉상태에서 현재의 은닉상태로의 전이 확률을 나타내는 것으로 모델 내부의 은닉상태들 간의 전이 확률을 나타내는 행렬
- 관찰확률 행렬: 특정 은닉상태에서의 관찰 가능한 각각의 상태들에 대한 확률을 나타내는 행렬

HMM은 상태 q_t의 의존이 바로 이전의 상태에만 의존한다는 1차 마르코프 가정(first order Markov assumption)을 이용하여 시간 t이후의 모든 상태에 대하여 그 의존성은 아래와 같이 일반화 될 수 있다.

$$P(q_t = j \mid q_{t-1} = i) = P(q_{t+l} = j \mid q_{t+l-1} = i)$$

이를 이용하여 은닉마르코프모델 λ는 다음과 같은 (Π, A, B)의 3가지 요소로 정의할 수 있다.

- $\Pi = (\pi_i)$, $\pi_i = P(q_1 = i)$, $1 \le i \le N$
  상태들의 초기 확률값을 표현하는 벡터
- $A = (a_{ij})$, $a_{ij} = P(q_t = j \mid q_{t-1} = i)$, $1 \le i, j \le N$
  상태들간의 전이 확률 행렬을 표현
- $B = (b_j(k))$, $b_j(k) = P(o_t = v_k \mid q_t = j)$, $1 \le k \le M$, $1 \le j \le N$: 관찰확률 행렬을 나타내며 $P(o_i(k) \mid q_j)$로 표현

이렇게 정의된 HMM은 첫 번째로, 관찰되어진 관측열 $O = (o_1, o_2, \ldots, o_T)$와 모델 $\lambda = (\Pi, A, B)$에 대하여 주어진 HMM에서 관찰되어진 순서의 확률 $P(O \mid \lambda)$를 계산하는 확률 추정(probability estimation)의 문제와 두 번째로, 관찰된 관측열 $O = (o_1, o_2, \ldots, o_T)$와 모델 λ에 대하여 최적의 상태순서 $q = (q_1, q_2, \ldots, q_T)$ 순서를 생성할 확률이 가장 높은 은닉상태들 간의 순서를 찾는 최적 상태 순서의 결정(optimal sequence) 문제, 그리고 마지막으로 관찰된 관측열 $O = (o_1, o_2, \ldots, o_T)$에 대하여 $P(O \mid \lambda)$를 최대로 하는 모델 $\lambda = (\Pi, A, B)$의 매개변수(parameter)를 결정하는 매개변수 추정(parameter estimation) 문제를 해결함으로써 모델을 사용할 수 있게 된다. 본 논문에서도 통상적인 HMM사용법을 이용하여 모델을 구성하였다[3,5,6].

2.2 정보 추출

HMM을 이용하여 정보 추출을 수행하기 위해서 HMM의 각 요소들을 다음과 같이 실험 텍스트데이터의 각 요소들로 고려하였다.

우선, HMM 모델의 관찰열 O는, 출현한 키워드가 속한 단어집합의 번호벡터로 고려하였다. 여기에서 단어집합의 번호벡터는, 모델에 특정 단어가 입력으로 들어오는 경우 이 단어가 미리 정의된 어떤 집합에 속하는지를 판단하여 단어나 단어들이 출현할 때 이를 집합의 번호로 표현한 것을 의미한다. 단어가 속한 단어집합은 실험 전에 규칙기반 전처리 과정을 통하여 각 실험 데이터에 대하여 규칙들을 추출하였다 (설명은 지면관계상 생략함). 단어집합을 정의하는 데에는 해당 필드가 출현한 부분을 기준으로 전후 3단어까지 고려하였다. 다음으로, 모델의 은닉 상태는 추출하려는 필드의 개수만큼 정의하고 상태간의 전이를 위한 중간 상태를 고려하기 위해서 추출하려는 필드의 수+1개의 중간단계 상태를 고려하였다. 즉, 각 필드의 모델에 대해서 제공되는 초기 추상 모델 상태의 수는 다음과 같이 된다.

$$\#\ of\ Initial\ Model\ States = 2 \times (\#\ of\ Target\ Fields) + 1$$

위의 식에서 표현된 상태의 수는 초기 추상모델의 상태의 수를 나타내는 것으로, 훈련을 통해서 구성된 최종 모델은 일반적으로 위의 식보다 적은 상태 개수를 갖는다.

```
<NL>
<paragraph><sentence> SECOND CALL FOR PAPERS </sentence></paragraph>
<NL>
<paragraph><sentence> FIFTH ANNUAL INTERNATIONAL CONFERENCE ON
COMPUTATIONAL MOLECULAR BIOLOGY </sentence></paragraph>
<NL>
<sentence><paragraph>(<c_name>RECOMB 2001</c_name>)</sentence></paragraph>
<NL>
<paragraph><date>April 21-24, 2001</date></paragraph>
<paragraph><location>Montreal, Canada</location></paragraph>
<NL>
<paragraph><sentence>Organized by <NL> Centre de recherches mathTatiques
<NL> Universite de Montreal <NL></sentence></paragraph>
<NL>
<paragraph><sentence>Sponsored by <NL>
Association for Computing Machinery (ACM-SIGACT)</sentence></paragraph>
```

그림 1. CFP 데이터의 일부분 예 (SGML형식의 태그 포함)

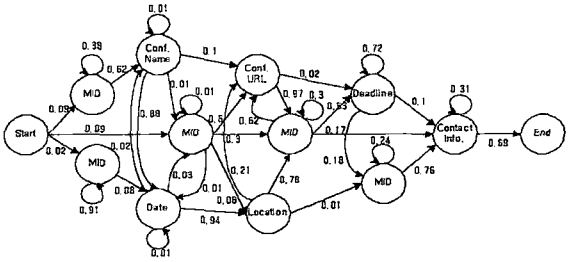

그림 2. CFP 데이터에 대하여 구성된 초기 모델(λ_0)의 예

그림1과 같은 데이터와 전처리를 통하여 구성한 규칙집을 이용하여 구성된 모델의 그림 2와 같은 모델의 구조는 아래와 같은 방법으로 모델의 상태수를 최소화 하도록 구조를 변경한다.

Step 0. Do data preprocessing and generate a field generation rule R for each field.
  τ = initial model state by given equation.
Step 1. Construct initial model λ_0 with internal state τ and rule R.
Step 2. If ($nState_{min} \le nState_\lambda < nState_{max}$) Goto Step 5.
Step 3. Compute field distance of model λ_0.
Step 4. if (exist(state pair within state distance θ)) then
  Merge nearest two states(one pair).
  $nState_\lambda \leftarrow nState_\lambda$ - 1, Goto Step 2.
  else Goto Step 5.
Step 5. Compute $P(O \mid \lambda)$ with observation
  $O = (o_1, o_2, o_2, \cdots, o_T)$ and model $\lambda = (\pi, A, B)$.
Step 6. Estimate optimal model parameters with the standard EM algorithm (Baum-Welch)
Step 7. *Find the optimal state sequence*
  $q = (q_1, q_2, q_2, \cdots, q_T)$ with given observation
  $O = (o_1, o_2, o_2, \cdots, o_T)$ and model λ.

그림 1. S-HMM의 동작 의사코드(pseudo-code)

3. 실험 및 결과

CFP데이터는 각 학회의 웹페이지나 메일링리스트로 안내되는 학회 논문제출 안내 문서를 직접 수집하여 구성한 텍스트 데이터이다. 데이터는 크게 학회의 이름, 개최지, 날짜, 장소, 학회에서 다루는 주제(topic), 위원회 명단, 논문제출 스케줄, 연락처 등을 포함하고 있다. 본 논문에서는 실험에 사용하기 위해서 CFP데이터를 그림 1과 같이 각 필드에 대하여 SGML형식으로 태깅(tagging)을 하였다. 데이터에 포함된 이와 같은 SGML형식의 태그는 모델 훈련에 사용되었다. 실험에서 사용한 CFP데이터는 컴퓨터과학(Computer Science)분야 학회의 CFP 문서 400개와 생물학 분야의 학회 CFP문서 200개로 총 600개의 서로 다른 학회에 대한 CFP문서로 구성되어 있으며, 전체 데이터의 크기는 6717K바이트이다.

CFP데이터는 SGML형식의 태그를 이용하여 전처리(label)하여 그림 1과 같이 구성하였다. 그림에서 "<NL>" 개행 문자(new line character)를 표현한다. 실험에서는 전체 600개의 CFP문서 중에서 임의로 선택한 250개의 CFP문서를 훈련데이터로 사용하였고, 나머지 350개를 이용하여 훈련된 모델의 성능을 검증하는데 사용하였다.

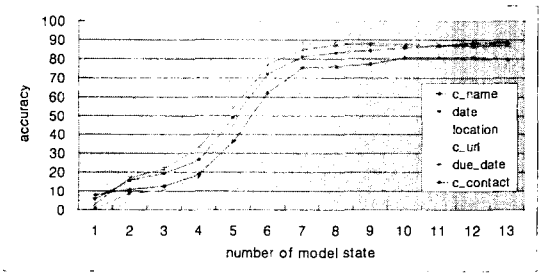

그림 3. 고정모델 HMM을 사용한 CFP데이터 정보추출

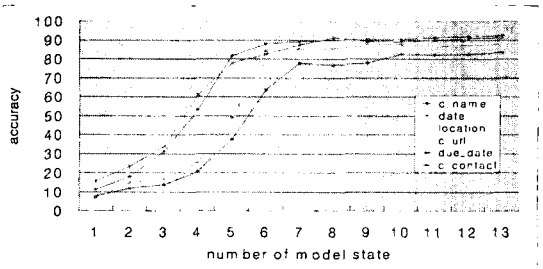

그림 4. S-HMM을 사용한 CFP데이터 정보추출

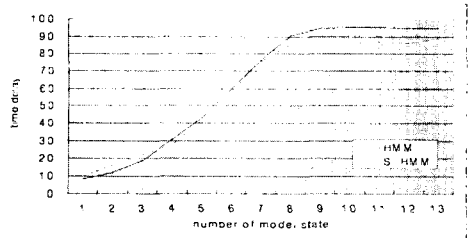

그림 5. HMM과 S-HMM의 time delay

그림 3~5는 고정모델 HMM과 자기구성 HMM을 사용한 CFP 데이터에 대한 정보추출의 결과를 모델이 가지는 상태의 수의 변화에 따른 성능변화로 표현한 것이다. 그림 5는 두 방법이 주어진 모든 데이터에 대하여 정보추출 작업을 완료할 때까지의 소요시간을 각각 상대적으로 표현한 것이다.

그림 3과 4을 비교해보면 동일한 CFP 데이터에 대해서 완전히 모델의 상태를 고정한 경우(그림 3)보다 초기에 어느 정도 모델의 상태 조정을 허용한 경우(그림 4)에 월등히 추출 성능이 높아지는 몇몇 필드들을 볼 수 있다. 이것은 추출할 필드에 대한 모델의 학습이 자기구성 HMM을 사용한 경우에 고정상태를 사용한 경우보다 더 잘 이루어졌다는 것을 나타낸다고 볼 수 있다. 또한 그림 3~5를 보면 전반적으로 SHMM이 더 적은 상태(state)를 가지고 목표 정확도에 도달하는 것을 볼 수 있다.

4. 결론 및 향후 과제

본 논문에서는 S-HMM을 이용한 정보추출을 학회참가안내 문서를 모아 구성한 Call-For-Papers 데이터에 대하여 특정 필드를 추출하도록 하여 실험을 수행하였다. S-HMM을 이용하여 초기의 모델을 규칙정보를 이용하여 변형해 나가도록 함으로써 데이터의 특징에 보다 적합한 구조를 스스로 찾을 수 있었다. 또한 훈련시간이 다소 증가하기는 하였지만 훈련 후에 정보추출에 소요된 시간 면에서는 약 12% 정도의 속도 향상 효과와 4%정도의 정확도 향상의 효과를 얻을 수 있었다. 단, 규칙 생성을위한 전처리 단계의 시간을 모두 고려하면 이러한 개선폭은 조금 낮아지는데 이는 앞으로 개선해야 할 과제라 할 수 있다. 또한, XML과 같이 데이터 자체가 풍부한 부가정보를 제공하는 경우에 대한 효율적인 data driven 접근 방법도 고려해 보아야 할 것이다.

감사의 글

본 연구는 과학기술부 뇌신경정보학 사업 (BrainTech), 교육부 BK21-IT 프로그램 및 첨단기술 연구센터 (AITrc)에 의하여 일부 지원되었음.

참고문헌

[1] Leek, T. R., "Information extraction using hidden Markov models", Master's thesis, UC San Diego, 1996.

[2] Seymore, K., McCallum, A., and Rosenfeld, R., "Learning hidden Markov model structure for information extraction", *AAAI'99 Workshop on Machine Learning for Information Extraction*, pp. 37-42, 1999.

[3] Rabiner, L. R., "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of IEEE*, Vol. 77, No. 2, February 1989.

[4] Muslea, I., Minton, S., and Knoblock, C., "STALKER: Learning extraction rules for semistructured, Web-based information sources", *AAAI'98*, 1998.

[5] Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *Journal of Royal Statistic Society B*, Vol. 39, pp. 1-38, 1977.

[6] Freitag, D., McCallum, A., "Information extraction with HMM structures learned by stochastic Optimization", *AAAI-2000*, 2000.

[7] Cohen, W, "A Web-based Information System that Reasons with Structured Collections of Text," *Proceedings of Second International Conference on Autonomous Agents*, pp. 400-407, 1998.

[8] Repository of Test Domains for Information Extraction, http://www.isi.edu/~muslea/RISE/repository.html.

[9] David, R. H. Miller, Tim Leek, and Richard M. Schwartz, "A hidden Markov model information retrieval system", *Proceedings on the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 214-221, 1999.