공학석사학위논문

은닉마르코프모델을 이용한 정보추출

Information Extraction Using Hidden Markov Models

2001년 2월

서울대학교 대학원

컴퓨터공학과

엄 재 홍

은닉마르코프모델을 이용한 정보추출

Information Extraction Using Hidden Markov Models

지도 교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함

2000년 10월

서울대학교 대학원

컴퓨터공학과

엄 재 홍

엄재홍의 공학석사 학위논문을 인준함

2000년 12월

위원장 김영택 印

부위원장 장병탁 印

위 원 김형주 印

초 록

은닉마르코프모델(Hidden Markov Model, 이하 HMM)은 일종의 오토마타(automata)로서 모델을 구성하고있는 상태들간의 전이가 특정한 확률값을 통하여 이루어진다. 이 모델은 음성데이터와 같은 순차적 데이터를 인식하는데 널리 사용되어온 모델로서 일련의 순차적인 성질을 내포하고있는 데이터를 다루는 문제에 잘 적용될 수 있다. 본 논문에서는 이러한 은닉마르코프모델을 이용하여 정보추출 문제를 해결하는 방법을 연구한다.

정보추출을 위하여 본 논문에서는 left-to-right구조에 기반한 HMM모델을 사용한다. 또한, 기존에 HMM을 사용할 때는 주로 목적에 맞는 HMM을 수동으로 구성하고 모델 내부의 확률 파라미터 값을 학습시켰던 데 반해서 본 논문에서는 이전의 방식들과는 달리 초기에 추상적으로 설정한 모델이 학습을 통해서 점차 구체화되어 가는 자기구성 은닉마르코프모델(Self-Organizing Hidden Markov Model: S-HMM)을 제시하여 사용하였다.

본 논문에서는 S-HMM을 이용한 정보추출을 학회참가안내 문서를 모아 구성한 Call-For-Papers 데이터와, CMU에서 학내 세미나 안내자료로 구성된 데이터, 그리고 LA지역의 레스토랑을 소개하는 자료(http://www.laweekly.com/)로 구성된 데이터에 대하여 특정 필드를 추출하도록 하여 실험을 수행하였다. S-HMM을 이용하여 초기에 주어진 추상 모델을 스스로 구성해 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.

주요어: 정보 추출, 은닉마르코프모델, 모델구조 학습, 자기구성 은닉마르코프모델

학　번: 99419-524

제목 차례

표 차례

그림 차례

1. 서론

1.1 연구 배경

최근 인터넷이 발달함에 따라 많은 양의 정보들에 대한 온라인 접근이 그어느 때보다도 쉬워지고 있으며, 신문이나 잡지 학술논문 등의 각종 정보들이 디지털화되어 온라인상 문서의 형태로 존재하게 되었다. 이러한 정보의 증가로 인한 정보 과부하(information overload)는 사용자들로 하여금 모든 정보를 소화하기 힘들게 만들고 있다[Maes, 1994]. 이에 정보검색 분야가 온라인 문서의 증가와 함께 활발히 연구되어 사용자들이 인터넷 상에서 필요한 문서를 찾는데 많은 도움을 주고 있다. 또한, 많은 수의 문서를 접근하는데에서 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 연구의 필요성이 대두되게 되었다.

정보추출에 대한 필요성이 널리 인식된 후, 은닉마르코프모델이나 문법규칙(grammar rule), 기계학습(machine learning), 여러 가지 자연언어처리 기법 등의 방법을 이용한 정보추출에 과한 여러 연구가 진행되었다.

기존에 음성인식에 널리 사용되어 오던 은닉마르코프모델이 정보검색이나

정보추출 분야에서도 사용되기 시작하였다. 예를 들면, 문서가 내부구조로 가지고있는 주제별 구조를 은닉마르코프모델로 모델링하여 사용자가 찾으려고 하는 정보와 어느 정도 유사한 문서인지를 판단하여 정보검색을 수행하는 HMM 정보검색 시스템[David, 1999]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.

이러한 여러 가지 방법들 중에서 은닉마르코프모델은 음성데이터와 같은 순차적 데이터를 인식하는데 널리 사용되어온 모델로서 일련의 순차적인 성질을 내포하고있는 데이터를 다루는 문제에 잘 적용될 수 있다. 여러 가지 정보를 담고있는 문서들 중에서 학회참가요청(Call for Papers) 문서나 세미나 안내문서, 또는 논문과 같은 문서들은 어느 정도 지정된 형식을 따르고있기 때문에 내부적으로 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.

1.2 관련 연구

초기에 은닉마르코프모델이 응용된 분야는 음성인식(speech recognition) 분야이다. 은닉마코프모델은 높은 인식률과 편리한 학습성으로 음성인식에 가장 널리 쓰이는 방법으로서 음성의 시간적 변화를 모델링하는 천이확률과 스펙트럼 변화를 모델링하는 출력확률로 구성된다. 또한, 입력패턴과 참조 패턴간의 유사도를 이용하기보다는 주어진 모델과의 확률적인 추정값을 사용하여 모델의 유사도를 계산하는데, 전통적인 음성인식에서는 각 음소에 대한 정적 HMM모델링을 이용하였으며, 입력되는 음성 데이터를 전처리하여 해당 입력에 근사한 출력을 생성하는 은닉 상태집합의 경로를 Viterbi 알고리즘을 사용하여 찾아 시간적 순서로 입력되는 음성을 인식하였다[Rabiner, 1989].

이후 HMM의 다양한 응용들이 있었는데 이중에서 정보추출에 응용된 초기의 예로는 Leek의 연구를 들 수 있다. Leek은 HMM을 이용하여 생물의학 텍스트에서 유전자의 위치를 찾는데 응용하였다[Leek, 1997]. Leek의 HMM의 모델은 생물의학 텍스트라는 분야에 적합하도록 각 분야에 대한 언어 모델링(language modeling)을 각 상태에 반영시켜 수동으로 정교하게 구성된 정적모델 구조를 가졌다. 또한 Leek은 언어의 문법(syntax)을 표현하기 위해서 망 구조(network topology)를 사용하였으며, 각 상태는 유니그램(unigram) 언어모델을 기반으로 구성하였다. 알려지지 않은 토큰(token)들은 공백 상태(gap state)를 두어 처리하였다. Leek의 연구는 전통적으로 음성인식에 사용되어온 HMM을 비교적 성공적으로 정보 추출에 응용한 초기의 예라고 할 수 있다.

은닉마르코프모델을 이용한 정보추출에 관한 또 다른 초기 연구로는 Bikel의 연구를 들 수 있다. Bikel 등은 HMM을 이용하여 MUC-6(Message Understanding Conference)의 텍스트 데이터에서 고유명사(name entity)를 추출하는 정보추출 시스템 Nymble을 만들었다[Bikel et al., 1997]. Nymble은 추

출하려고 하는 모든 필드를 Seymore[Seymore et al., 1999]와 같이 하나의 HMM으로 구성하였다. 하지만, 모델의 구조 학습의 어려움을 극복하기 위해서 목적 상태마다 하나의 상태를 두고 상태들 간에는 완전히 연결된(fully connected) 상태전이를 구성하였다. 또한, 상태에서 뿐만 아니라 최종 출력 확률함수에 대해서도 성긴 훈련데이터(sparse training data)를 이용하여 출력확률(emission probabilities)과 전이확률(transition probabilities)을 고려하였다. 이것은 보다 확실한 모델 파라미터를 얻기 위해서 모델 축소(shrinkage)의 형태를 사용한 것으로 볼 수 있다. 또한, 이 방법에서는 최종 출력을 다르게 하는 확률들의 분포를 평균화하여 사용하였으며, EM 알고리즘을 통한 복잡한 최적 혼합 가중치(optimal mixture weight)를 학습하지는 않았다.

또한 정보검색에 HMM이 응용된 경우도 있다. David[1999] 등의 연구가 이것인데, 여기서는 문서가 문서내부에 포함하고있는 각각의 주제부분들이 이루는 구조를 HMM을 통하여 인식하였다. 여기서 사용된 HMM은 동일 모델에서 다중 단어생성을 다룰 수 있도록 설계된 모델이다. David 등이 제안한 시스템은 TREC(Text REtrieval Conference)중 TREC-6, TREC-7 ad hoc retrieval 문제에서 기존의 *tf · idf*를 기반으로 한 시스템보다 월등한 성능을 나타내었다 [David et al., 1999]. 이 방법에서는 HMM의 틀에 블라인드 피드백(blind feedback)을 사용하였으며, bigram을 고려하기 위한 모델도 추가적으로 고려하였다.

위에서와는 달리 HMM은 어떤 데이터를 모델링하는 것 이외에도 사용자의 프로파일(profile)을 모델링하는 데에도 응용되었다. Lane[Lane 1990]은 HMM을 사용하여 어떠한 시스템을 사용하는 사용자들을 모델링하여 사용자의 비정상성(anormality)을 탐지하는데 사용하였다. Lane은 시스템 사용자 각각에 대하여 정적 구조를 가지는 HMM 모델을 구성하여 해당 사용자의 비정상성을 탐지하도록 하였다. 구성된 모델을 훈련데이터를 통한 파리미터 조정을 이용하여 사용자의 속성을 학습하였다. Lane은 이 연구에서 모델의 은닉상태의 수가

적절한 수위를 유지할 때 성능에 보다 긍정적인 영향을 주며, 지나치게 많은 상태는 효율성을 오히려 저하시킨다는 것을 실험적으로 보였다.

HMM을 사용하여 정보추출의 문제를 다루면서 모델 내부의 상태들간의 전이 구조에 대한 연구도 이루어졌다. Seymore 등[Semore et al., 1999]은 HMM의 상태간의 전이 구조를 위한 연구를 수행한 대표적인 예라 할 수 있다. Seymore 등은 논문의 머릿글이나 참고문헌과 같이 어느 정도 양식을 가지고 밀집되어있는 몇 개의 필드들을 추출하기 위하여 하나의 HMM을 사용하였다. 따라서, 다수의 필드들이 하나의 HMM으로 표현이 가능하였으며, 관찰된 필드들이나 예상되는 필드들의 순서가 대부분 HMM을 통하여 포착될 수 있었다. 하지만, 이 경우 세부적인 모델의 구조를 학습하지 않았기 때문에 필드구조의 지나친 단순화가 문제가 되며, 모델 구조에 대한 학습이 추가적으로 필요하다.

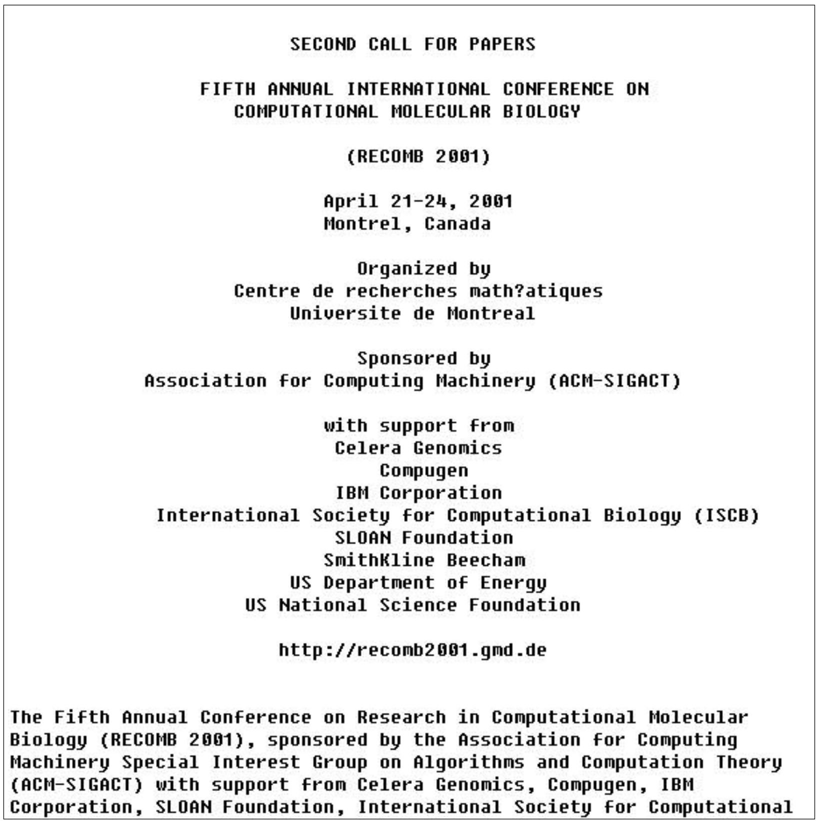
Freitag[Freitag, 1998]는 HMM을 사용하지는 않았지만 온라인 세미나 안내자료를 대상으로 정보 추출을 수행하였다. 여기에서는 기계학습의 여러 가지 방법들로 구성된 학습자들의 예측 결과를 회귀(regression)을 이용하여 결합해 최종 결과를 내는 다중전략(multistrategy) 방법을 사용하여 온라인 세미나 안내자료에서 세미나 장소, 세미나 시간 등의 정보를 추출하였다.

의미와 패턴정보만을 사용하여 정보추출을 다룬 연구도 있었다. Riloff[Riloff, 1999]는 사전(dictionary)을 위한 의미어휘(semantic lexicon)와 추출패턴(extraction pattern)을 동시에 생성해 내는 시스템을 구성하였는데, 이 방법에서 사용한 다단계 상호 부스트래핑(Multi-level mutual bootstrapping)은 초기에 주어진 기본단어(seed word)를 사용하여 동일 부류의 단어를 학습하는 방식이며, 학습한 결과를 다시 정보 추출에 사용하였다. 이를 이용하여 웹 문서에서의 정보 추출과 뉴스기사에서 테러에 관한 정보를 추출하였다.

1.3 연구 내용

본 연구에서는 스스로 초기에 정해진 모델의 구조를 변경할 수 있는 은닉마르코프모델을(Self-Organizing Hidden Markov Model) 사용하여 정보추출의 문제를 다룬다. 은닉마르코프모델의 파라미터는 EM(Expectation-Maximization) 알고리즘[Dempster et al., 1997]을 이용하여 학습이 가능하다. 여기서는 모델 구조의 제한을 두지 않은 HMM방법으로 데이터의 특징에 따라 모델이 스스로 구조를 학습하는 HMM을 이용하여 정보추출의 문제를 해결하는 방법을 제안한다.

은닉마르코프모델의 구조는 추출하려고 하는 정보를 가지고있는 데이터의 필드가 가지는 특징을 반영하여 구성하는 것이 바람직하다. 이것은, 초기에 구성된 모델이 불필요한 상태를 포함하게 되는 경우 모델의 구조는 이후 변경이 불가능하기 때문에 전체 상태간의 전이 행렬의 크기를 최소화하고 모델의 파라미터를 학습하는데 소요되는 시간을 줄이는 것을 방해하는 요소로 작용된다. 예를 들어 그림 1과 같은 텍스트 필드 구조를 가지는 학회 참가안내 CFP데이터를 고려해 보도록 하자. 그림 1에서 볼 수 있는 학회이름, 개최일, 개최지, 참고 URL과 같은 여러 가지 필드 중에서 개최일, 개최지와 같은 필드는 필드가 나타나기 전에 실마리가 되는 특별한 키워드가 있는 경우도 있지만, 이는 데이터의 특징에 따라 형식상에서 조금씩 차이가 발생할 수 있다. 그림 1에서 볼 수 있듯이 개최일, 개최지와 같은 필드가 함께 나타나면서 전 후에 공백문자를 통해 구분되는 것과 같은 경우, 특별히 다른 키워드를 인식하는 상태를 가지는 모델의 상태는 더욱 축소될 수 있다. 즉, 데이터가 공백문자 이외에 다른 키워드를 내포하는 경우가 상대적으로 많은 경우가 아니라면 모델의 상태 수를 줄여 속도를 최대화하는 것이 바람직하다.



SECOND CALL FOR PAPERS

FIFTH ANNUAL INTERNATIONAL CONFERENCE ON
COMPUTATIONAL MOLECULAR BIOLOGY

(RECOMB 2001)

April 21-24, 2001
Montrel, Canada

Organized by
Centre de recherches math?atiques
Universite de Montreal

Sponsored by
Association for Computing Machinery (ACM-SIGACT)

with support from
Celera Genomics
Compugen
IBM Corporation
International Society for Computational Biology (ISCB)
SLOAN Foundation
SmithKline Beecham
US Department of Energy
US National Science Foundation

http://recomb2001.gmd.de

The Fifth Annual Conference on Research in Computational Molecular Biology (RECOMB 2001), sponsored by the Association for Computing Machinery Special Interest Group on Algorithms and Computation Theory (ACM-SIGACT) with support from Celera Genomics, Compugen, IBM Corporation, SLOAN Foundation, International Society for Computational

그림 1. RECOMB 2001 학회의 CFP 데이터의 일부

정적인 모델 구조를 가지는 HMM의 단점은 앞에서 살펴본 바와 같이 데이터가 간단한 구조를 가지는 경우에도 초기에 훈련데이터로부터 구성된 모델의 구조를 끝까지 유지해야하는 데에서 오는 시간비용의 증대이다. 이에 본 논문에서는 은닉마르코프모델의 모델 구조를 초기 추상 모델에서부터 데이터를 이용한 학습을 통하여 스스로 조직화하도록 함으로써 데이터에 적합한 모델을 구설 할 수 있도록 하였다. 또한, 모델의 최소 및 최대의 상태개수 변화에 따라 고정 HMM과 자기구성 HMM을 이용한 각각의 방법에 따라 어떠한 성능을 보이는지 정보추출의 문제에 대하여 실험적으로 비교하였다.

1.4 논문의 구성

본 논문의 구성은 다음과 같다. 2장에서는 논문에서 사용된 은닉마르코프모델과 모델의 정의와 여기에서 사용되는 기본 알고리즘으로 전향 알고리즘(Forward Algorithm)과 전향-후향 알고리즘(Forward-Backward Algorithm)에 대하여 살펴보고, 모델 학습 알고리즘인 EM알고리즘과 Viterbi 알고리즘에 대하여 살펴본다. 3장에서는 정보추출의 문제를 위하여 은닉마르코프모델을 적용한 방법에 대하여 학회참가안내 문서를 모아 구성한 CFP(Call-For-Papers) 데이터를 기준으로 설명하고, 4장에서는 제안된 방법을 정보추출 문제에 적용하여 그 성능을 평가한다. 그리고 마지막으로 5장에서는 실험 결과를 토대로 결론을 맺는다.

2. 은닉마르코프모델 (Hidden Markov Model)

2.1 은닉마르코프모델의 소개

HMM은 관찰이 불가능한 미지(hidden)의 확률론적 과정(stochastic process)을 관찰이 가능한 기호(symbol)를 발생시키는 다른 확률론적 과정을 통하여 모형화(modeling)하는 이중의 확률론적 과정이다. 예를 들어 우리가 쉽게 관찰할 수 있는 음파나 공기의 압력, 또는 조음기관의 위치 등을 이용하여 두뇌의 활동을 모형화 하는 것을 생각해 볼 수 있다. 이러한 HMM은 다음과 같이 2개의 상태집합과 3개의 확률 집합으로 구성되는 5개의 요소를 갖는다.

· 은닉상태집합(hidden state set)
마르코프 프로세스에 의해서 설명되는 상태들의 집합

· 관찰가능 상태집합(observable state set)
외형적으로 눈에 보이는 전이 상태들의 집합

· π 벡터
특정 은닉 상태가 시간 $t=1$일 때 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 나타내

는 행렬

이와 같이, HMM은 관찰 가능한 상태들과 은닉상태들간의 확률적 관계를 이용하여 확장된 표준 마르코프 프로세스라 볼 수 있다. HMM은 상태들 간의 전이 관계를 다음과 같은 가정을 통해 규정한다.

$$P(q_t = j \mid q_{t-1} = i, q_{t-2} = k, \ldots) = P(q_t = j \mid q_{t-1} = i) \tag{1}$$

$$P(q_t = j \mid q_{t-1} = i) = P(q_{t+l} = j \mid q_{t+l-1} = i) \tag{2}$$

식 (1)은 마르코프 모델의 상태 전이는 이전의 상태들과는 무관하며, 단지 바로 전 단계의 상태에 의해서만 다음상태의 결정이 영향을 받는다는 것을 의미한다. 즉, 시간 t에서의 상태 q_t가 j로 결정되는 것은 시간 $t-1$의 상태 q_{t-1}에 따라 결정된다는 것을 나타낸다. 이는 1차 마르코프 가정(first order Markov assumption)라 불리며 HMM의 기본 이론을 구성하는 요소로 식 (2)와 같이 시간 t 이후의 모든 상태에 대하여 일반화 될 수 있다.

모델의 상태들이 출력하는 기호에 대한 모델 관찰열(observation symbol sequence)은 다음의 식 (3)과 같이 표현될 수 있다.

$$O = O_1, O_2, \ldots, O_{T-1}, O_T \tag{3}$$

여기서 T는 관찰열의 지속시간을 나타내는 시간이다. 이후 N을 은닉 상태들의 개수라 하고, M을 관찰 기호의 개수를 표현한다.

$$Q = \{q_1, q_2, \ldots, q_n\} \tag{4}$$

위의 식 (4)는 모델 내부에 있는 각각의 상태를 나타낸다.

$$V = \{v_1, v_2, ..., v_m\} \tag{5}$$

마찬가지로 식 (5)는 상태에서 관찰이 가능한 기호들의 집합을 의미한다.

은닉마르코프모델(λ)은 다음과 같은 (Π, A, B)의 3가지 요소로 정의된다.

- $\Pi = (\pi_i),\ \pi_i = P(q_1 = i),\quad 1 \le i \le N$

  상태들의 초기 확률값을 표현하는 벡터

- $A = (a_{ij}),\ a_{ij} = P(q_t = j | q_{t-1} = i),\quad 1 \le i, j \le N$

  상태들간의 전이 확률 행렬을 표현

- $B = (b_j(k)),\ b_j(k) = P(o_t = v_k | q_t = j),\quad 1 \le k \le M,\ 1 \le j \le N$

  관찰확률 행렬을 나타내며 $P(o_i(k) | q_j)$와 같이 표현된다.

위에서 상태전이 행렬 A는 다음과 같이 구성된다.

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2N} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{iN} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{Nj} & \cdots & a_{NN} \end{pmatrix} \tag{6}$$

식 (6)이 가지는 각 원소는 식 (7)과 같이 정의되며, a_{ij}는 식 (8), (9)와 같은 제약 조건을 갖는다.

$$a_{ij} = P(q_t = j \mid q_{t-1} = i) \qquad 1 \le i, j \le N \tag{7}$$

$$a_{ij} \ge 0, \qquad \forall i, j \tag{8}$$

$$\sum_{j=1}^{N} a_{ij} = 1, \qquad \forall i \tag{9}$$

여기에서 A, B 두 행렬은 시간에 독립적이라고 가정하는데, 실제적인 문제에 있어서 이것은 마르코프 모델이 가지는 비현실적 가정이 된다. 하지만, HMM은 하나의 벡터와 두 개의 행렬을 이용하여 실제의 시스템을 충분히 모델링할 수 있기 때문에 앞에서와 같은 가정을 통한 근사 모델링으로도 널리 사용된다.

이렇게 HMM을 정의하고 나면 실제로 모델을 사용하기 위해서는 다음과 같은 3가지 문제를 해결해야만 한다.

1. 확률추정(probability estimation)의 문제

관찰되어진 관측열 $O = (o_1, o_2, ..., o_T)$와 모델 $\lambda = (\Pi, A, B)$에 대하여 주어진 HMM에서 관찰되어진 순서의 확률 $P(O|\lambda)$를 계산하는 문제.

2. 최적 상태 순서의 결정(optimal sequence) 문제

관찰된 관측열 $O = (o_1, o_2, ..., o_T)$와 모델 λ에 대하여 최적의 상태순서 $q = (q_1, q_2, ..., q_T)$ 순서를 생성할 확률이 가장 높은 은닉상태들 간의 순서를 찾는 문제.

3. 매개변수 추정(parameter estimation) 문제

관찰된 관측열 $O = (o_1, o_2, ..., o_T)$에 대하여 $P(O|\lambda)$를 최대로 하는 모델 $\lambda = (\Pi, A, B)$의 매개변수(parameter)를 결정하는 문제로 다음과 같이 표현될 수 있다.

$$\max_{\lambda}\{P(O\,|\lambda)\} \tag{10}$$

문제 1의 경우 은닉상태(hidden state)를 고려해야 하기 때문에 확률을 추정하는 문제가 복잡해진다. 문제 2의 경우는 데이터를 가장 잘 설명하는 상태 순서를 결정하기 위해서 '최적'의 기준을 정하는 것이 필요하다. 문제 3은 HMM의 학습과 관계가 있다.

HMM을 사용하기 위하여 해결해야 하는 위와 같은 3가지 문제를 해결하기 위해서 몇몇 방법들이 널리 쓰여왔다. 확률평가 문제(문제 1)의 경우 전향 알고리즘(forward algorithm)과 후향 알고리즘(backward algorithm)을 이용하여 해결이 가능하며 이에 대해서는 2.1.2절에서 다루고 있다. 최적의 상태 순서의 경로를 결정하는 문제(문제 2)는 일반적으로 동적 프로그래밍 기법의 하나인 Viterbi 알고리즘을 이용하여 결정한다. 이에 대해서는 2.1.3절에서 다룬다. 마지막으로, 모델의 확률을 최대로 하는 매개변수의 추정은 Baum-Welch 알고리즘을 사용하여 수행할 수 있으며 이는 2.1.4절에서 다루고 있다.

2.2 모델의 확률 추정

전향 알고리즘 (Forward Algorithm)

전향 알고리즘은 은닉마르코프모델을 사용하려고 할 때, 모델 $\lambda = (\Pi, A, B)$ **에 대해서 관찰되어진 관측열** $O = (o_1, o_2, ..., o_T)$**의 확률** $P(o_1, o_2, ..., o_T | \lambda)$**을 구하는데 사용된다. 관측열의 확률은 간단히 다음과 같은 직접적인 알고리즘을 이용한 접근 방법으로 구할 수도 있다.**

Algorithm - Let $q = (q_1, q_2, ..., q_T)$ be a state sequence. - Assume the observations are independent: $P(O \mid q, \lambda) = \prod_{t=1}^{T} P(o_t \mid q_t, \lambda)$ $= b_{q1}(o_1) b_{q2}(o_2) \cdots b_{qT}(o_T)$ - Probability of a particular state sequence is: $P(q \mid \lambda) = \pi_{q1} a_{q1q2} a_{q2q3} \cdots a_{qT-1qT}$ - Also, $P(O, q \mid \lambda) = P(O \mid q, \lambda) P(q \mid \lambda)$ - Enumerate paths and sum probabilities: $P(O \mid \lambda) = \sum_{q} P(O \mid q, \lambda) P(q \mid \lambda)$

표 1. **직접적 모델 확률의 계산**

그런데 위의 표 1**과 같은 방법으로 확률을 구하기 위해서는** N^T**개의 상태 순서와** $O(T)$**의 복잡도를 가지는 계산을 수행해야 한다. 즉, 알고리즘을 수행**

하기 위해서 $O(TN^T)$의 계산 복잡도를 극복해야 하는 것이다. 이는 상대적으로 시간지연 값 T가 큰 실제적인 응용에서는, T에 대한 지수함수로 계산 복잡도가 증가하는 문제로 인해서 사용하기에 부적합하다.

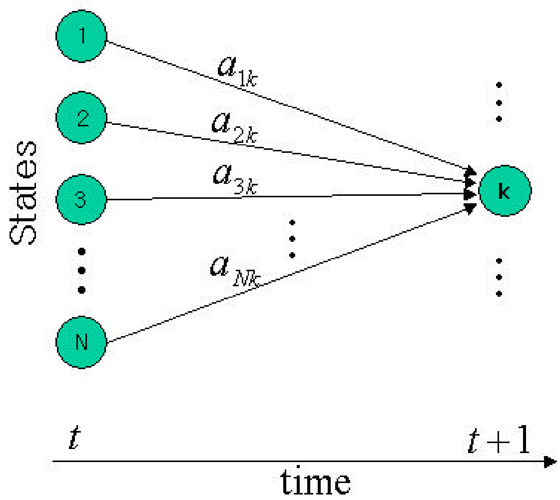


그림 2. 전향 진행 (Forward Procedure)의 직관적인 표현

앞에서 살펴본 것처럼 직접적인 접근 방법은 그 계산의 복잡도 때문에 실제적 응용이 불가능하다. 따라서 일반적으로는 다음과 같은 전향 알고리즘을 사용하여 확률을 정하게 된다.

우선 전향변수(forward variable) $\alpha_t(i)$를 다음과 같이 정의한다.

$$\alpha_t(i) = P(o_1, o_2, \cdots, o_t, q_t = i \mid \lambda) \qquad (11)$$

식 (11)에서 정의한 $\alpha_t(i)$는 상태 q_t가 i에 있을 때 부분상태 열 $(o_1, o_2, \cdots, o_t)$에 대한 관찰 확률을 의미한다. 이후 표 (2)와 같은 귀납 알고리즘을 사용하여 모델의 관찰확률을 구하게 된다.

이렇게 계산을 하게되면 앞에서 살펴보았던 직접적인 접근법 보다 계산의 복잡도가 $O(N^2T)$로 감소되어 모델에서 사용하는 시간지연 T에 의해서는 선형적인 영향을 받기 때문에 직접적 방법에서의 지수적 계산 복잡도를 극복할 수 있게 된다.

- Induction

1. Initialization:

$$\alpha_1(i) = \pi_i b_i(o_1)$$

2. Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1})$$

3. Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

표 2. 전향 알고리즘의 관찰확률 계산알고리즘

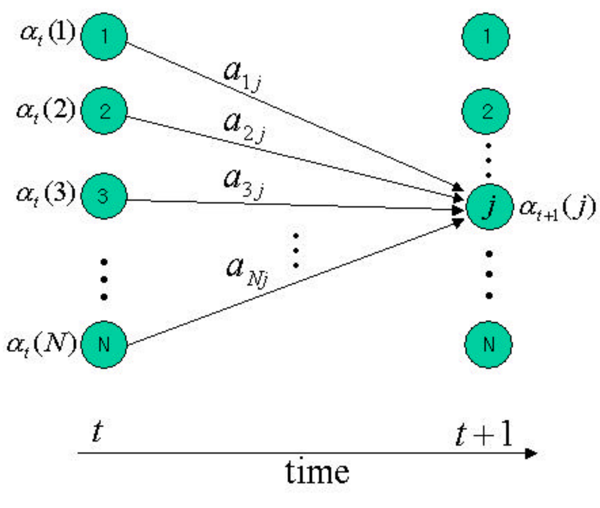


그림 3. 전향 알고리즘의 진행

전향 알고리즘은 시간 $t=1$일 때부터 $t=T$일 때까지의 확률벡터를 구하여 저장 전향변수 α에 유지시킨다.

후향 알고리즘 (Backward Algorithm)

후향 알고리즘에서는 전향 알고리즘에서와 유사하게 후향변수(backward variable)를 정의하여 사용한다. 후향변수 $\beta_t(i)$는 다음과 같이 정의한다.

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, \cdots, o_T \mid q_t = i, \lambda) \tag{12}$$

식 (12)에서 정의한 $\beta_t(i)$는 상태 q_t가 i에 있을 때 부분상태 열 $(o_{t+1}, o_{t+2}, \cdots, o_T)$에 대한 관찰 확률을 의미한다. 이후 표 (3)과 같은 귀납 알고리즘을 사용하여 역 방향에서의 모델의 관찰확률을 구하게 된다.

- Induction
  1. Initialization:

     $\beta_T(i) = 1$

  2. Induction:

     $\beta_t(i) = \sum_{=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$

     $1 \le i \le N,$

     $t = T-1, \cdots, 1$

표 3. 후향 알고리즘의 관찰확률 계산알고리즘

후향알고리즘에서는 전향알고리즘과 반대로 후향변수 β에 시간 $t=T$일 때부터 $t=1$일 때까지의 역으로 확률벡터를 구하여 저장한다.

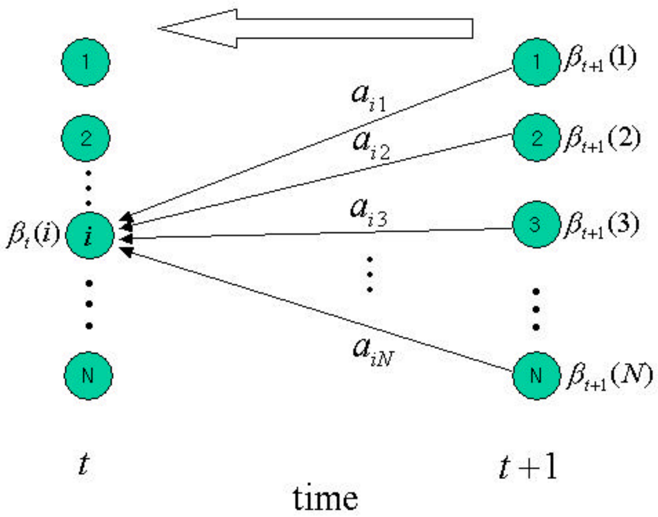


그림 4. 후향 알고리즘의 진행

2.3 최적의 상태 열(state sequence)의 결정

일단, 관측열에 대하여 최대 확률을 가지는 적합한 모델을 선택한 후에는 관측열을 생성하는 모델의 최적 상태 순서를 결정하여야 한다. 일반적으로 모델의 최적 상태를 결정하는 데에는 동적 프로그래밍(dynamic programming)방법 중의 하나인 Viterbi 알고리즘을 사용한다.

비터비알고리즘 (Viterbi algorithm)

Viterbi 알고리즘은 주어진 관측열 O와 모델 λ를 가지고 관측열 O를 최대 확률로 생성하는 상태순서(state sequence) $(q_1, \cdots, q_t, \cdots, q_T)$를 결정하는데 사용된다. 관측열 O와 모델 λ를 기초로 관측열을 생성하는 확률은 식 (13)과 같

이 표현할 수 있다.

$$P(q_1, q_2, \cdots, q_T \mid O, \lambda) \tag{13}$$

아래의 그림 5는 시간 t 에서부터 시간 $t+1$로의 상태 전이들 중에서 최대 확률로 상태 전이를 하는 상태들의 순서를 결정해 나아가는 하나의 단계를 표현하고 있다.

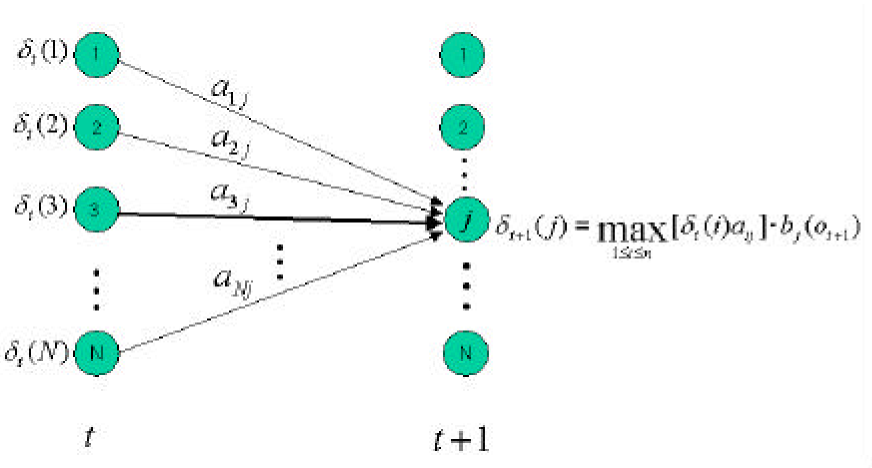


그림 5. Viterbi 알고리즘을 이용한 최적의 상태 열 탐색

그림 5에서 $\delta_t(j)$는 상태 j로 끝나는 경로들 중에서 최대 확률을 가지는 경로에 대한 확률값을 표현하고 있는 것으로 식 (14)와 같이 표현된다.

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P(q_1, q_2, \cdots, q_t = i, o_1, o_2, \cdots, o_t \mid \lambda) \tag{14}$$

$$\delta_{t+1}(j) = \max_i [\delta_t(i)\, a_{ij}] \cdot b_j(o_{t+1}) \tag{15}$$

귀납에 의하여 식 (14)는 식 (15)와 같이 확장될 수 있으며, 이를 이용하면 시간 t 뿐만 아니라 이후의 $t+1$ 에 대해서도 최대 확률을 가지는 상태들의 순서를 구할 수 있다.

- Initialization

$$\delta_1(i) = \pi_i b_i(o_1), \quad 1 \le i \le N$$

$$\psi_1(i) = 0$$

- Recursion

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(o_t)$$

$$\psi_t(j) = \underset{1 \le i \le N}{\arg\max} [\delta_{t-1}(i) a_{ij}]$$

$$2 \le t \le T, \quad 1 \le j \le N$$

- Termination

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

$$q_T^* = \underset{1 \le i \le N}{\arg\max} [\delta_T(i)]$$

- Path (state sequence) backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \cdots, 1$$

표 4. Viterbi 알고리즘

표 4에서 $\psi_t(i)$는 시간 t에서 상태 i로 전이하는 최적 경로를 유지하는 변수이다. $\psi_t(i)$는 $\psi_t(j) = \underset{1 \le i \le N}{\arg\max} [\delta_{t-1}(i) a_{ij}]$와 같이 이전 상태($t-1$)까지의 최대 확률을 가지는 경로 δ_{t-1}와 시간 t에서의 상태 j로의 전이 행렬을 이용하여 최대 확률을 가지는 상태 경로를 계산한다.

2.4 모델의 매개변수 추정

관측열에 대하여 최대 확률을 가지는 적합한 모델을 선택하고 이 모델의 내부 상태들 중에서 최적의 상태 순서를 결정한 후에는, 관찰된 관측열 $O=(o_1, o_2, ..., o_T)$에 대하여 $P(O|\lambda)$를 최대로 하는 모델 $\lambda=(\Pi, A, B)$의 매개변수(parameter)를 결정하여야 한다(문제 3). 모델의 매개변수를 결정하는 문제는 모델의 복잡성 때문에 분석적(analytic) 방법으로는 어렵다. 이 때문에 일반적으로 모델 파라미터의 학습에는 EM 알고리즘이라고도 불리는 Baum-Welch 알고리즘을 사용한다.

EM (Baum-Welch) 알고리즘

Baum-Welch 알고리즘은 초기모델(λ_0)을 구성하고, 초기모델과 관찰열 O를 기반으로 새로운 모델(λ)을 구성한다. 이렇게 하여 새로 구성된 모델과 이전 모델이 관찰열을 생성하는 확률의 차이가 특정값 이상이 될 때까지 모델의 매개변수를 변경하면서 새로운 모델을 생성해 나아간다. 이를 알고리즘 형식으로 표현하면 표 5와 같다.

Step 1. Let initial model be λ_0.
Step 2. Compute new λ based on λ_0 and observation O.
Step 3. If $\log P(O\|\lambda) - \log P(O\|\lambda_0) < DELTA$ then stop.
Step 4. Else set $\lambda_0 \leftarrow \lambda$ and goto step 2.

표 5. EM (Baum-Welch) 알고리즘

또한, Baum-Welch 알고리즘에서는 다음과 같은 두 개의 새로운 변수를 추가적으로 정의하여 사용한다. 식 (16)은 시간 t에서 상태 i에 있고, 시간 $t+1$에서 상태 j에 있을 확률을 표현한다. 마찬가지 방법으로, 식 (17)은 시간 t에서 주어진 관찰열을 가지고 상태 i에 있을 확률을 나타낸다.

$$\xi(i,j) = \frac{\alpha_t(i)\, a_{ij}\, b_j(o_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} \tag{16}$$

$$= \frac{\alpha_t(i)\, a_{ij}\, b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)\, a_{ij}\, b_j(o_{t+1})\beta_{t+1}(j)}$$

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \tag{17}$$

위의 식 (16)을 $\sum_{t=1}^{T-1}\xi(i,j)$와 같이 하면 관찰열 O에서 상태 i에서 상태 j로 전이한 횟수의 기댓값이 된다. 또한, 식 (17)을 이용하면 $\sum_{t=1}^{T}\gamma_t(i)$와 같이 합을 취하여 관찰열 O에서 상태 i가 방문된 횟수에 대한 기댓값을 구할 수 있다.

식 (16)에 들어있는 α와 β는, 앞절의 전향-후향(forward-backward)알고리즘에서 설명한 각각의 변수로 그림 6과 같이 전향알고리즘과 후향알고리즘을 함께 사용하여 계산할 수 있다.

EM알고리즘은 계산된 변수값을 이용하여 모델의 매개변수를 재추정(reestimate) 하는데, 모델 변수의 갱신 규칙은 식 (18) (20)과 같다.

$$\widehat{\pi_i} = \gamma_1(i) \tag{18}$$

$$\widehat{a_{ij}} = \frac{\sum\xi_t(i,j)}{\sum\gamma_t(i)} \tag{19}$$

$$\widehat{b_j}(k) = \frac{\sum_{t, o_t = k} \gamma_t(j)}{\sum_t \gamma_t(j)} \tag{20}$$

여기에서, 식 (18)의 우변은 시간 $t=1$일 때, 상태 i에서의 기대빈도 수를 의미한다. 식 (19)의 분모는 상태 i에서의 전이에 대한 기댓값을 나타내고, 분자는 상태 i에서 상태 j로 전이되는 회수에 대한 기댓값을 의미한다. 식 (20)의 분모는 상태 j의 방문 회수에 대한 기댓값을, 분자는 기호 k를 가지고 상태 j를 방문하는 회수에 대한 기댓값을 의미한다.

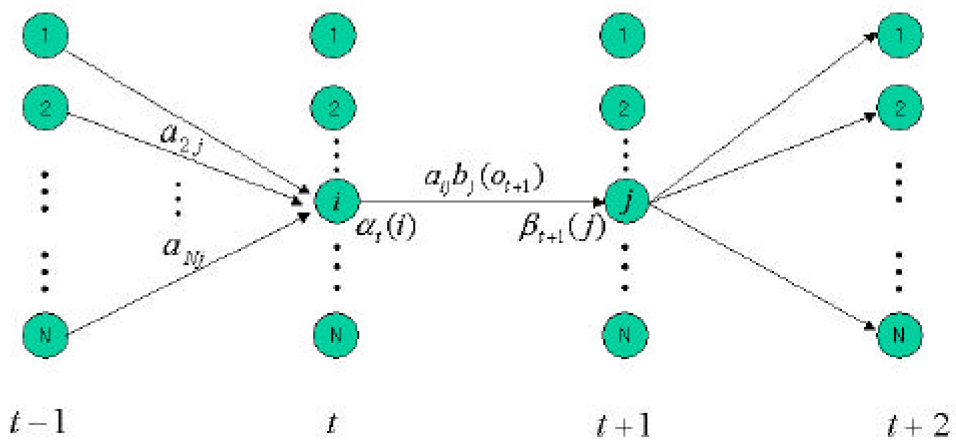


그림 6. EM 알고리즘에서의 전향, 후향 변수의 계산

3. 은닉마르코프모델을 이용한 정보추출

3.1 정보추출을 위한 은닉마르코프모델

정보추출에 은닉마르코프모델을 적용하기 위해서는, 먼저 HMM의 각 모델 파라미터를 사용하려는 텍스트 데이터에 사상(mapping)시켜야 한다. 또한, 사용하는 텍스트 데이터에서 추출하려고 하는 필드에 대한 모델을 구성하는 작업도 필요하다. 본 장에서는 텍스트 데이터의 각 요소들을 HMM의 파라미터들에 어떻게 대응시켰는지, 그리고 이를 이용하여 모델을 구성하는 것에 대하여 설명하도록 한다.

3.1.1 정보추출에 대한 HMM의 적용

본 논문에서는, 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 전처리 과정을 통하여 각 실험 데이터에 대하여 아래의 표 6 9와 같이 정의하였다. 단어집합을 정의하는 데에는 해당 필드가 출현한 부분을 기준으로 전후 3단어씩을 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.

$$Number\ of\ Initial\ Model\ States\ =\ 2\times(Number\ of\ Target\ Fields)\ +\ 1 \qquad (21)$$

위의 식 (21)에서 표현된 상태의 수는 초기 추상모델의 상태의 수를 나타내는 것으로, 훈련을 통해서 구성된 최종 모델은 식 (21)의 것보다 적은 상태를 갖는다.

아래의 표 6 9는 본 논문에서 사용한 3개의 실험데이터 각각의 추출필드와 이를 위한 단어집합을 사전 정의한 것이다. 정의된 단어집합은 실험의 훈련단계에서 단어규칙을 추가하거나 삭제하는 과정을 통하여 수정되어질 수 있도록 하였다.

필드 이름	단어규칙	실제 예
학회 이름	{n-NAME_SYM + year}, {n-NAME_SYM + "'" + year}, {n-NAME_SYM + "-" + year}, {n-NAME_SYM + " " + year}, etc. NAME_SYM := Capital character | Begin with Capital Character year := 4-digit | 2-digit $2\leq n\leq 15$	ICONIP 2000 SIGIR'2001 PRICAI'01 KDD01 EC-Web 2000 CL2000 GECCO 2000 CEC 2001
학회 날짜	{day + month + "," + year}, {day + "," + month + "," + year}, {day + th + D_SYM + day + "th + month + year}, {day + D_SYM + day + "," + year}, {month + day + D_SYM + day + "," + year}, etc. D_SYM := "-" | "to"	23 September 2000 24th to 28th July, 2000 8-12 July 1999 August 20-23, 2000

표 6. CFP(Call-For-Papers) 데이터의 각 필드별 단어규칙 집합

필드 이름	단어규칙	실제 예
학회 장소	{PLACE + <NL> + CITY + "," + COUNTRY}, {PLACE + "," + CITY + "," + COUNTRY}, etc. NL := new line character	Riviera Hotel Las Vegas, Nevada USA
참고 URL	{URL_MARK + URL}, etc. URL_MARK := "http://" | "www."	http://www.genetic-algorithm.org/GECCO2000/gecco2000mainpage.htm
논문 제출 마감일	{DUE_SYM + <NL> + date}, {DUE_SYM + ":" + date}, {DUE_SYM + date}, etc. DUE_SYM := "Important Dates" | "Deadline" | "Due Dates" | "Submission Deadline" | "Paper Submission" "Dates" | "Schedule" | "Conference Schedule" | “CALENDAR"	Important Dates December 20, 1999: Deadline for the submissions of the proposals. SUBMISSION DEADLINE: January 26, 2000
학회 연락처	{CON_SYM1 + ":" + number}, {CON_SYM2 + ":" + e-mail}, etc. CON_SYM1 := "Phone" | "Tel" | "Fax" CON_SYM2 := "E-MAIL"	Phone: 650-328-3123 FAX: 650-321-4457 E-MAIL: gecco@aaai.org

표 7. CFP(Call-For-Papers) **데이터의 각 필드별 단어규칙 집합** (**계속**)

표 6**의 학회 날짜와 같은 필드는 실제 데이터집합 에서는 일반적으로 사용하는 방법 외에도 다양한 형식으로 표현되고있는 예도 일부 발견되었다**.

필드 이름	단어규칙	실제 예
세미나 주제	{TOPIC_SYM + ":" + topic}, etc. TOPIC_SYM := "Topic" + "Title"	Topic: SKVORETZ Seminar
세미나 날짜	{DATE_SYM + ":" + date}, etc. DATE_SYM := "Dates" date := day + "-" + month + "-" + +year day := 1-digit | 2-digit month := month year := 4-digit | 2-digit	Dates: 4-May-95 Dates: 15-April-1997
세미나 장소	{LOC_SYM + location}, etc. LOC_SYM := "in" | "at"	5:30 in PH 223D. at Porter Hall A18C.
연사 이름	{PER_SYM + name}, etc. PER_SYM := "Mr." | "Dr." | "President"	DR. WILLIAM FISH MR. Jill Fain Lehman
세미나 시작시간	{TIME_SYM + ":" + time}, etc. TIME_SYM := "Time" time := number + ":" + number	3:30-5:00
세미나 끝 시간	{"-" + time}, etc. time := number + ":" + number	3:30-5:00

표 8. CMU 세미나 안내 데이터의 각 필드별 단어규칙 집합

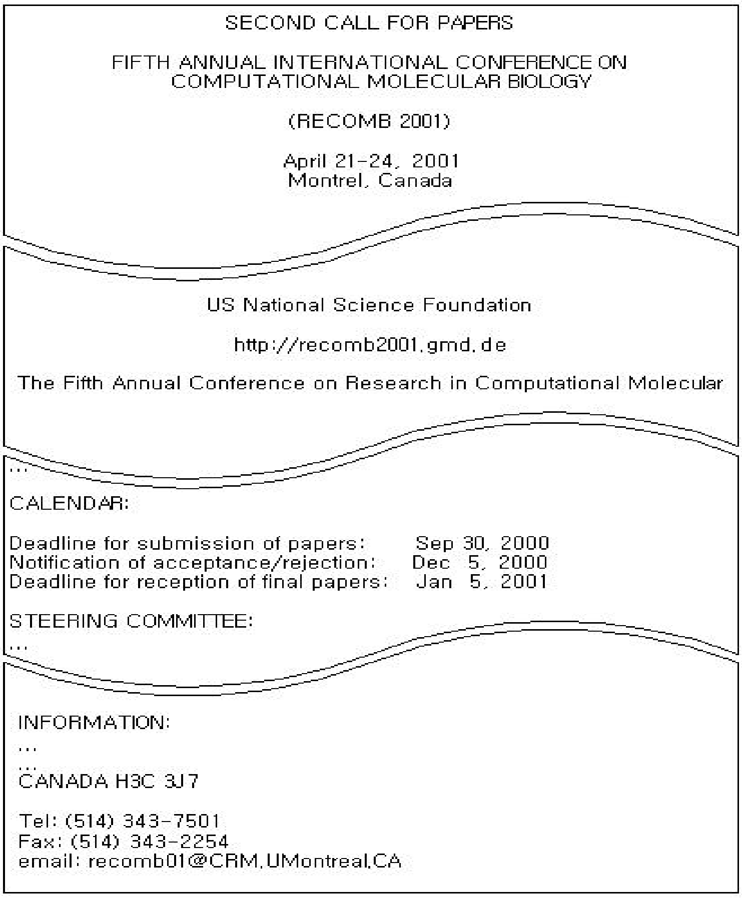
실제 CMU 세미나 안내 데이터에서는 위의 표 8에서 보이고있는 필드들 모두가 각 데이터에 존재하지는 않았다. 일부 데이터의 경우 세미나 종료시간에 관한 언급이 없었으며, 또 다른 데이터에서는 연사의 이름이 생략되어 있는 경우도 있었다. 예를 들어 특정 그룹세미나 또는 정기 그룹 세미나의 경우 특정한 연사가 명시되지 않는 예가 존재하는 것을 확인하였다.

필드 이름	단어규칙	실제 예
레스토랑 이름	{NAME_SYM + name + RST_SYM}, etc. NAME_SYM := "Named after" RST_SYM := "RESTAURANT"	EL FLORIDITA RESTAURANT 1253 N. Vine St., L.A. (213) 871-8612 …
주소	{STREET NO + STREET + "," + CITY}, etc.	1253 N. Vine St., L.A.
연락처	{"(" + number + ")" + number + "-" + number}, etc.	(213) 871-8612
평론	{SECTION_SYM + Review + SECTION_SYM}, etc. SECTION_SYM := new line | new line + tab	Named after Hemingway-'s favorite hangout … weekend reservations su-ggested.
신용카드	{n-CD + card + SECTION_SYM}, etc. CD = {card + ","} card := AE | CB | V | MC | DC | DIC | BC $0 \le n \le 7$	AE, CB, DC, DIS, MC, V.

표 9. LA 주간 레스토랑 소개 데이터의 각 필드별 단어규칙 집합

시스템에서는 위의 표 6 9와 같이 정의된 규칙을 기반으로 각 데이터를 처리하면서 입력으로 들어오는 단어집합을 고려하여 HMM의 관찰열 O를 단어집합별 번호를 출력함으로써 구성한다. 예를 들어, 입력으로 들어온 데이터가 다음의 그림 7과 같다고 하면, 시스템은 이와 같은 CFP데이터를 읽어나가면서 표 6 7과 같이 정의된 CFP데이터의 필드별 번호를 고려하여 다음과 같이 관찰열 O를 생성한다(편의상 표 6 7의 각 필드의 번호를 1 6까지 순서대로 할당하고 중간단계를 영어 소문자로 할당하였다고 가정하자. S:시작, E:끝).

$$O = (o_1, o_2, \cdots, o_T) = \quad S\ 1\ a\ 2\ b\ 3\ c\ 4\ d\ 5\ e\ 6\ E \qquad (22)$$



SECOND CALL FOR PAPERS

FIFTH ANNUAL INTERNATIONAL CONFERENCE ON
COMPUTATIONAL MOLECULAR BIOLOGY

(RECOMB 2001)

April 21-24, 2001
Montrel, Canada

US National Science Foundation

http://recomb2001.gmd.de

The Fifth Annual Conference on Research in Computational Molecular

...

CALENDAR:

Deadline for submission of papers: Sep 30, 2000
Notification of acceptance/rejection: Dec 5, 2000
Deadline for reception of final papers: Jan 5, 2001

STEERING COMMITTEE:

...

INFORMATION:

...

...

CANADA H3C 3J7

Tel: (514) 343-7501
Fax: (514) 343-2254
email: recomb01@CRM.UMontreal.CA

그림 7. CFP **데이터의 위치별 각 필드 출현 예제**

3.1.2 필드 인식을 위한 HMM의 구조와 초기 HMM구조의 변경

본 논문에서는, **각 데이터별로** HMM**의 모델을 구성하였다. 모델은 기본적으로 식** (21)**에서와 같이 은닉 상태를 갖는다. 식** (21)**의 상태 수는 모델의 시작상태(** S **)와 종료상태(** E **)를 포함한 것이다. 데이터별로 구성된 모델은 각 추출 필드를 표현하는 상태와 상태간의 전이 중간 단계를 표현하는 중간상태들로 구성된다. 기본적으로 각 필드들은 하나의 상태와 중간전이상태로 표현되고있**

는 것이다. 이렇게 모델을 구성하고 나면 일반적으로 사용되는 고정 상태를 가지는 HMM보다 많은 상태를 가지는 모델이 구성된다. 일반적으로 CFP데이터의 경우, 고정 상태의 HMM으로 구성하게 되면 그림 8과 같이 각 필드를 하나의 상태로 표현하는 모델을 구성하게된다[Seymore et al., 1999]. 하지만, 데이터 필드의 특성을 보다 자세히 고려하면 몇 개의 필드를 HMM모델의 하나의 상태로 표현할 수 있다. 만일 이렇게 하여, 1개 이상의 필드를 하나의 상태로 표현할 수 있다면, 모델의 인식 정확도를 유지하면서도 최종 모델의 상태 수를 줄일 수 있게 된다. 모델의 상태수가 줄게되면 전향-후향 알고리즘(forward-backward algorithm)이나 Viterbi알고리즘, 그리고 EM알고리즘을 이용하여 모델파라미터를 최적화하는 과정에서 고려해야 하는 경우의 수가 급격하게 줄게 되어 정보추출의 전반적 속도향상을 기대할 수 있게 된다.

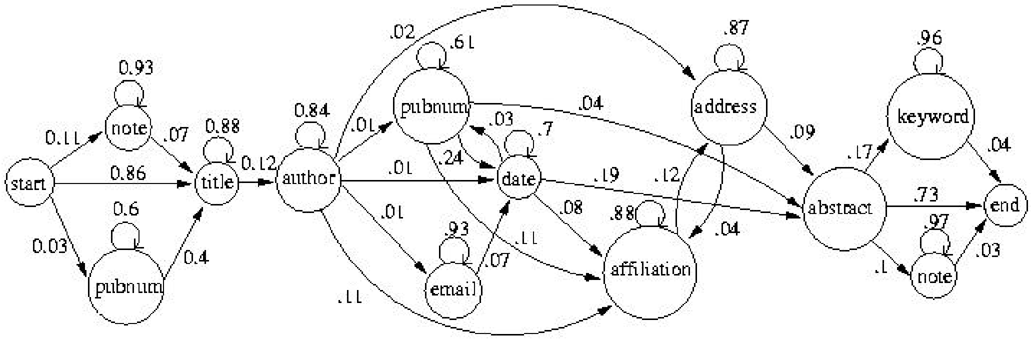


그림 8. 각 추출 필드별로 하나의 상태를 가지는 HMM의 예[Seymore et al., 1999]

이에 본 논문에서는, 초기에는 다소 많은 상태로 구성된 모델을 이용하지만 데이터의 추출 필드가 가지는 단어 규칙들을 학습하여, 초기 상태를 감소시켜 최종 모델을 구성하는 HMM을 제안한다.

그림 9는 그림 7과 같은 문서에 대하여 구성된 모델의 예를 나타내고 있다. 그림 9는 초기에 구성된 모델 λ_0를 표현한 것이다. 모델 λ_0는 각각의 추출 필드를 위한 상태들과 필드별 전이를 위한 중간 상태들(MID), 그리고 시작과 끝

상태들을 유지한다.

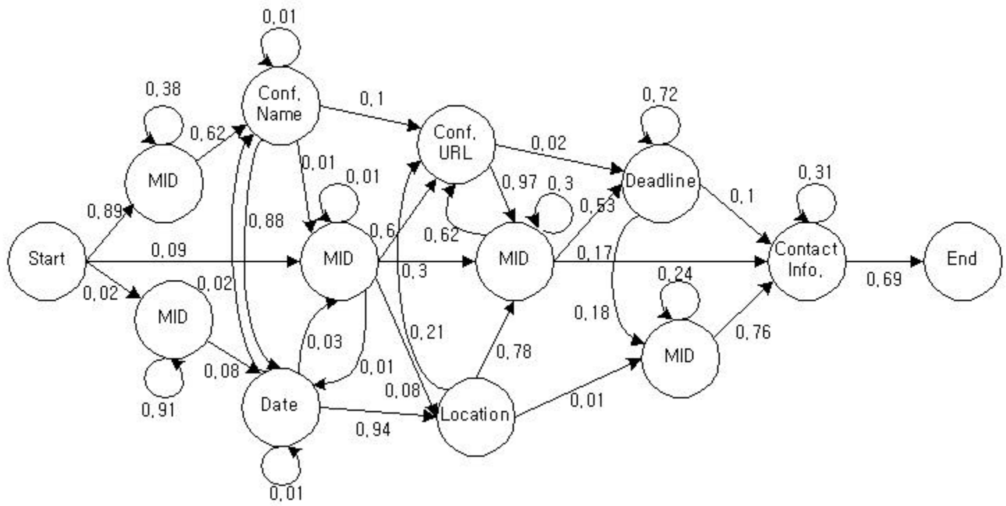


그림 9. 그림 7과 같은 CFP 데이터에 대하여 구성된 초기 모델(λ_0)의 예

위의 그림 9와 같이 구성된 모델에서 Date 필드의 경우를 보면 다음 상태인 Location 필드로 중간단계인 MID를 거치지 않고 매우 높은 확률로 진행하고 있다. 이렇게 구성된 필드들은 실제 데이터에서 주로 매우 인접하여 나타나는 것을 의미한다. 따라서 이러한 두 상태는 차후 상태병합 과정에서 하나의 상태로 합쳐지게 된다. 위의 그림 9를 기준으로 보면 결국 중간단계 없이도 매우 높은 확률로 서로의 상태가 연결되어진 Data와 Location, Conf. Name을 표현하는 상태들이 합쳐지게 된다.

본 논문에서는 앞에서 설명한 것과 같이 각각의 필드를 표현하는 두 상태가 최종적으로 하나의 상태로 병합되는 경우 병합되어 생성된 상태는 HMM의 관찰열을 생성할 때 병합되기 이전의 상태처럼 관찰열 O를 생성하도록 하였다.

아래의 표 10은 본 논문에서 제안한 자기구성 은닉마르코프모델의 알고리즘에 대한 의사코드(pseudo code)를 나타낸다.

Step 0. Do data preprocessing and generate a field generation
rule R for each field.
τ = initial model state by equation (21).
Step 1. Construct initial model λ_0 with internal state τ and
rule R.
Step 2. If ($nState_{\min}$ <= $nState_{\lambda}$ < $nState_{\max}$) 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 $P(O|\lambda)$ 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 EM
algorithm in Table 5.
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.

표 10. 자기구성 HMM의 동작 의사코드(pseudo-code)

표 10의 Step 3에서 상태 i와 j사이의 거리는, 두 상태를 연결하는 가장 짧은 경로의 상태전이 확률의 합의 역수로 계산된다. 상태들간의 전이확률이 지정된 특정 값 θ를 넘지 않는 경우 두 상태간의 거리는 매우 큰 값으로 고려되어 상태간의 병합 가능성이 차단된다. 이러한 상태 병합 과정은 전체 상태들 중에서 지정된 임계치 θ를 넘는 쌍이 더 이상 존재하지 않을 때까지 계속 이루어진다.

3.2 HMM을 이용한 정보추출 시스템의 설계 및 구현

추출 시스템은 아래의 그림 10과 같은 구조를 가진다. 그림 10의 "Field Rule Generator"는 앞에서 설명한 표 6 9와 같은 규칙을 전처리기로부터 받아 유지한다. 생성된 규칙은 이후 초기모델 λ_0의 상태를 감소시켜 모델의 복잡도를 감소시키는데 사용된다. 아래의 그림 10은 본 논문에서 구성한 정보추출 시스템의 전체 구성도를 나타낸다.

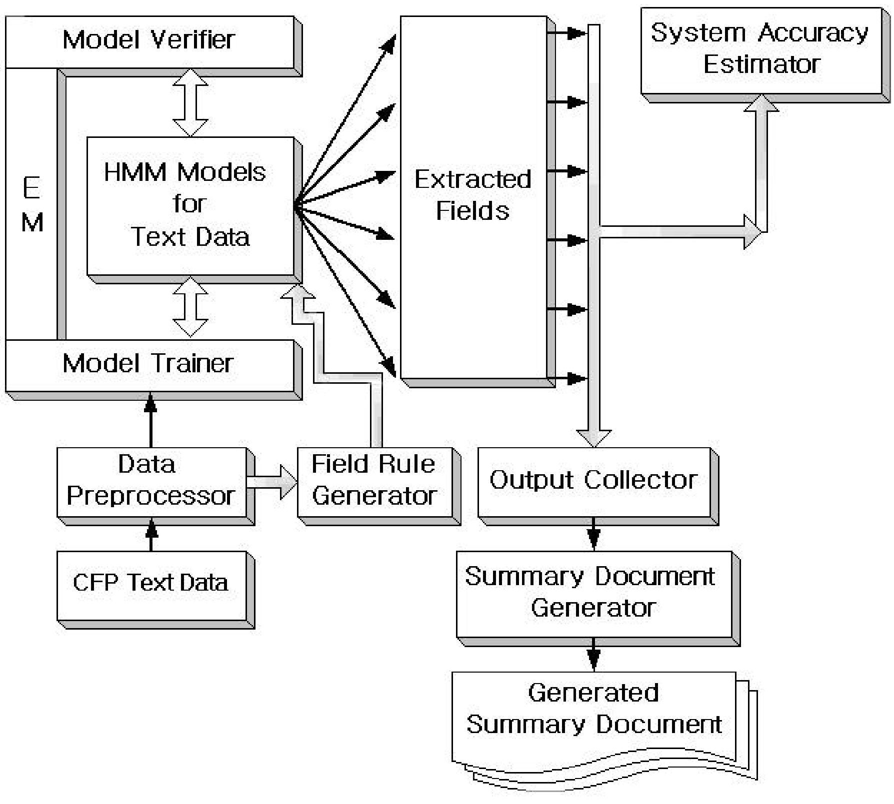


그림 10. 정보추출 시스템의 전체 구성도

본 논문에서 설계한 HMM알고리즘을 이용한 CFP 문서 정보추출 시스템은 크게 위의 그림 10과 같은 컴포넌트 구조를 갖는데, 그림 10의 CFP Text

Data는 훈련 데이터와 검사 데이터를 의미한다. 입력된 데이터에 대하여 전처리기는 필드 출현 규칙을 구성하고 이를 이용하여 다시 입력데이터를 모델의 입력으로 사용할 관찰열로 변환한다.

변환되어 입력된 관찰열은 모델 훈련 과정의 경우 모델 훈련모듈을 통해서 모델을 훈련하는데 사용된다. 모델은 각각의 추출 필드에 대하여 구성되어 있으며 입력되어 들어오는 훈련 데이터를 통하여 모델의 파라미터들이 학습된다. 이후 검사 데이터를 입력으로 하여 추출을 수행하는 경우, 입력 데이터에 대하여 각각의 모델의 인식 확률을 이용하여 필드를 인식한다. 모델에 의하여 인식된 필드는 추출필드를 수집·정리하는 단계를 거쳐 추출필드를 요약해주는 모듈의 입력으로 들어가게 된다. 추출필드 요약 모듈은 추출된 결과를 이용하여 HTML형태의 요약 레포트 파일을 생성한다. 생성된 요약 문서는 이후 메일로 추출 결과를 전송하거나 인터넷을 통하여 추출 결과를 접근할 수 있는 시스템을 구성하는데 응용될 수 있도록 구성되었다.

4. 실험 및 결과

4.1 학회 안내 데이터(Call-For-Paper Data)

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 표 11과 같은 필드에 대하여 SGML형식으로 태깅(tagging)을 하였다. 데이터에 포함된 이와 같은 SGML형식의 태그는 모델 훈련에 사용되었다. 실험에서 사용한 CFP데이터는 컴퓨터과학(Computer Science)분야 학회(Conference)의 CFP문서 200개와 생물학 분야의 학회 CFP문서 100로 총 300개의 서로 다른 학회에 대한 CFP문서로 구성되어 있으며, 전체 데이터의 크기는 2051K바이트이다.

논문의 첫번째 절에 나와있는 그림 1은 RECOMB 2001의 CFP데이터의 일부를 표현한 것이다. 그림 1과 같은 CFP데이터는 표 11의 SGML 형식의 태그를 이용하여 전처리하여 그림 11과 같이 구성하였다. 그림 11의 "<NL>" 개행문자(new line character)를 표현한다.

실험에서는 전체 300개의 CFP문서 중에서 임의로 선택한 200개의 CFP문서를 훈련데이터로 사용하였고, 나머지 100개를 이용하여 훈련된 모델의 성능을 검증하는데 사용하였다.

```
<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>Montrel, Canada</location></paragraph>
<NL>
<paragraph><sentence>Organized by <NL> Centre de recherches math?atiques
<NL> Universite de Montreal <NL></sentence></paragraph>
<NL>
<paragraph><sentence>Sponsored by <NL>
Association for Computing Machinery (ACM-SIGACT)</sentence></paragraph>
<NL>
<paragraph><sentence>with support from <NL> Celera Genomics <NL>
Compugen <NL> IBM Corporation <NL>
International Society for Computational Biology (ISCB) <NL>
SLOAN Foundation <NL>
SmithKline Beecham <NL>
US Department of Energy <NL>
US National Science Foundation </sentence></paragraph><NL>
<NL>
<c_url>http://recomb2001.gmd.de</c_url>
<NL><NL>
<paragraph><sentence>The Fifth Annual Conference on Research in
Computational Molecular <NL> Biology (<c_name>RECOMB 2001</c_name>),
sponsored by the Association for Computing <NL>
Machinery Special Interest Group on Algorithms and Computation Theory <NL>
(ACM-SIGACT) with support from Celera Genomics, Compugen, IBM <NL>
```

그림 11. CFP 데이터의 예 (SGML형식의 태그 포함)

필드 이름	설 명	필드 구분 태그(Tag)
학회 이름	학회 이름을 표시 필드	<c_name> … </c_name>
학회 날짜	학회의 개최일 표시 필드	<date> … </date>
학회 개최장소	학회 개최장소 표시 필드	<location> … </location>
학회 참고 URL	학회의 웹 페이지 주소	<c_url> … </c_url>
논문제출 마감일	논문이나 초록 제출 마감일	<due_date> … </due_date>
학회 연락처	전화, 팩스, 전자우편과 같은 연락처	<c_contact> … </c_contact>
단락의 표시	CFP 데이터 내에서 단락의 시작과 끝을 표시한다.	<paragraph> … </pagragraph>
문장의 표시	CFP 데이터 내에서 문장의 시작과 끝을 표시한다.	<sentence> … </sentence>

표 11. CFP 학회 안내 데이터에 포함된 필드들

4.2 CMU 세미나 안내 데이터

CMU 세미나 안내 데이터는 CMU(Carnegie Mellon University)에서 정보추출을 위한 테스트 자료로 구축하여 둔 데이터집합[CMU Data]의 일부로, 정보추출을 위한 연구를 위하여 그룹 내에서 공지되는 온라인 세미나 안내 문서를 학습데이터의 형식으로 전처리하여 구성한 데이터이다. 총 데이터의 크기는 1149K바이트이며, 전체 488개의 세미나에 대한 온라인 공지 문서를 포함하고 있다. 각 세미나 공지 자료에는 아래 표 X와 같은 필드가 포함되어 있으며 필드를 명시하기 위하여 데이터 내부는 SGML형식의 태그로 필드를 표시한다. 각 필드를 명시적으로 표시하기 이하여 사용된 SGML형식의 태그는 필드 모델의 훈련에 사용된다. *표시가 있는 필드는 본 실험을 위하여 추가적으로 추가한 필드를 나타낸다. **표시가 있는 필드는 데이터에 SGML형식의 태그로는 들어있기는 하지만, 추출 필드를 표시하는 것이 아니라 문장이나 단락과 같은 문서의 구조에 대한 정보를 제공하는 필드를 나타낸다. 이러한 필드들은 차후에 정보추출을 위한 대상으로 사용되며 데이터에 따라서는 특정 필드가 존재하지 않는 경우도 있다. 그림 12은 실제 데이터가 포함하고 있는 내용을 나타내고 있으며, 그림 13은 그림 12의 데이터를 표 12에 있는 SGML 형식의 태그를 이용하여 태깅(tagging)한 데이터의 실제 원본을 나타낸다.

```
<0.2.5.95.11.00.22.cd01+@andrew.cmu.edu.0>
Type:      cmu.andrew.academic.sds.seminars
Topic:     SKVORETZ Seminar
Dates:     4-May-95
Time:      4:00 - 5:30
PostedBy: Carole Deaunovich on 2-May-95 at 11:00 from andrew.cmu.edu
Abstract:

Professor John Skvoretz, U. of South Carolina, Columbia, will present
a seminar entitled "Embedded Commitment," on Thursday, May 4th from
4-5:30 in PH 223D.
```

그림 12. CMU 세미나 안내 데이터 (안내 내용)

```
<0.2.5.95.11.00.22.cd01+@andrew.cmu.edu.0>
Type:    cmu.andrew.academic.sds.seminars
Topic:   SKVORETZ Seminar
Dates:   4-May-95
Time:    <stime>4:00</stime> - <etime>5:30</etime>
PostedBy: Carole Deaunovich on 2-May-95 at 11:00 from andrew.cmu.edu
Abstract:

<paragraph><sentence><speaker>Professor John Skvoretz</speaker>, U. of South Carolina, Columbia, will present
a seminar entitled "Embedded Commitment," on Thursday, May 4th from
<stime>4</stime>-<etime>5:30</etime> in <location>PH 223D</sentence></location>.</paragraph>
```

그림 13. CMU 세미나 안내 데이터 (SGML형식의 태그 포함)

필드 이름	설 명	필드 구분 태그(Tag)
세미나 주제(Topic)*	세미나의 제목이나 주제를 나타내는 필드	<topic> … </topic>
세미나 날짜*	세미나의 날짜를 나타내는 것으로 “25-Oct-93”과 같은 형식으로 표시된다.	<date> … </date>
세미나 장소	세미나 개최지를 표시하는 필드로 주로 Room이나 Hall을 표시한다.	<location> … </location>
세미나 발표 연사의 이름	세미나 발표 연사를 표시한다.	<speaker> … </speaker>
세미나 시작시간	세미나의 시작시간을 표시한다.	<stime> … </stime>
세미나 끝 시간	세미나의 예상 종료 시간을 표시한다.	<etime> … </etime>
단락의 표시**	세미나 안내자료 내에서 단락의 시작과 끝을 표시한다.	<paragraph> … </pagragraph>
문장의 표시**	세미나 안내자료 내에서 문장의 시작과 끝을 표시한다.	<sentence> … </sentence>

표 12. CMU 온라인 세미나 안내 데이터에 포함된 필드들

CMU 온라인 세미나 안내 데이터를 이용한 실험에서는 전체 488개의 세미

나 안내자료 중에서 250개를 훈련데이터로 사용하였고, 나머지 238개의 데이터를 이용하여 훈련된 모델의 성능을 검증하였다. 훈련데이터의 선택은 전체 데이터에서 임의로 추출하는 방법을 사용하였으며, 훈련은 표 12에 명시되어 있는, 데이터의 필드를 명시하여 주는 SGML형식의 태그와 단락이나 문장을 표시하는 태그를 함께 사용하여 각 필드별로 수행하였다.

4.3 LA 주간 레스토랑 소개 데이터

LA 주간 레스토랑 안내 데이터는 LA Weekly(http://www.laweekly.com/)라는 사이트에서 매주 근교의 식당을 소개하는 글을 실은 것을 모아 구성한 웹 자료이다. 원본 데이터는 HTML형식으로 구성되어 있으며 435K바이트의 크기로 총 200개의 식당에 대한 안내를 담고 있다. 각 안내문서는 그림 14와 같이 구성되어 있으며, 식당이름, 주소, 전화번호, 식당에 대한 비평기사, 그리고 사용 가능한 신용카드의 종류가 명시되어 있고 본 실험에서는 각 필드에 대하여 정보 추출을 수행하였다.

EL FLORIDITA RESTAURANT
1253 N. Vine St., L.A.
(213) 871-8612

Named after Hemingway's favorite hangout in Havana, El Floridita offers food, fun and festivities. Try the yellow-rice paella complete with lobster, mussels and shrimp ($19.50) or the tender roasted pork marinated in a savory lemon garlic sauce ($10.95). Wash it all down with a Mojito, the popular Cuban drink made from mint leaves, lime juice and sugar. Live bands, including salsa, entertain Thursday through Saturday, and Monday nights. Prices vary on weekends. Lunch and dinner seven days. Full bar; takeout; delivery; valet parking; weekend reservations suggested. AE, CB, DC, DIS, MC, V.

그림 14. LA Weekly 데이터

실험은 전체 200개의 데이터 중에서 100개를 훈련데이터로 사용하고 나머지 100개를 검사 데이터로 사용하였다. 훈련데이터의 선택은 다른 실험에서와 마찬가지로 임의선택 하였다. 또한, 표 9에 나와있는 필드들에 대하여 SGML형식의 태그를 이용하여 원래의 HTML문서 데이터를 태깅 하였다.

4.4 결과 분석

그림 15 16은 고정모델 HMM과 자기구성 HMM을 사용한 CFP 데이터에 대한 정보추출의 결과를 모델이 가지는 상태의 수의 변화에 따른 성능변화로 표현한 것이다. 그림 17은 각각 두 방법이 주어진 모든 데이터에 대하여 정보추출 작업을 완료할 때까지의 소요시간을 상대적으로 표현한 것이다.

그림 15와 16을 비교해보면 동일한 CFP 데이터에 대해서 완전히 모델의 상태를 고정한 경우(그림 15)보다 초기에 어느 정도 모델의 상태 조정을 허용한

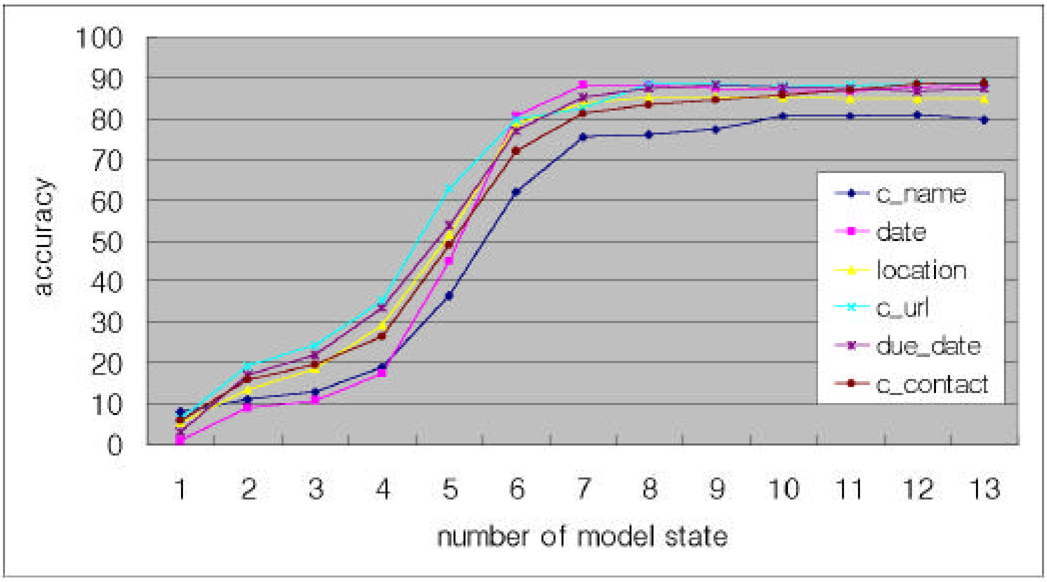


그림 15. 고정모델 HMM을 사용한 CFP데이터 정보추출

경우(그림 16)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.

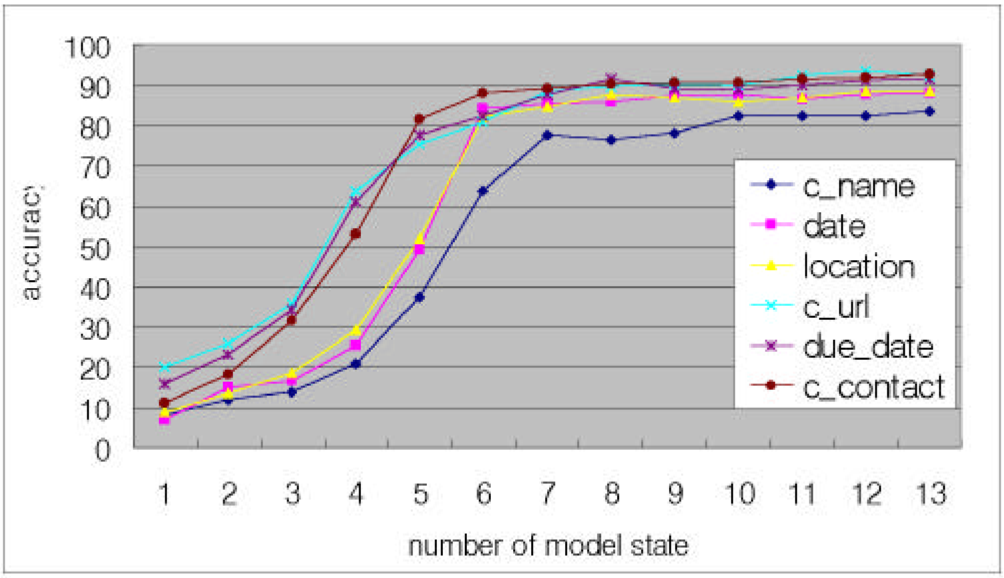


그림 16. 자기구성 HMM을 사용한 CFP데이터 정보추출

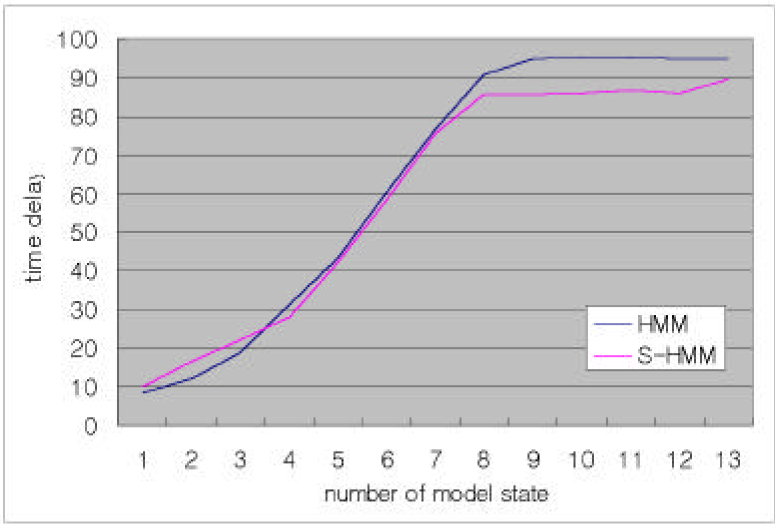


그림 17. HMM과 S-HMM의 시간지연 - CFP 데이터

그림 17은 자기구성 HMM의 처리 속도가 생태를 완전히 고정한 HMM보다 10%정도 향상되었다는 것을 보여주고 있다.

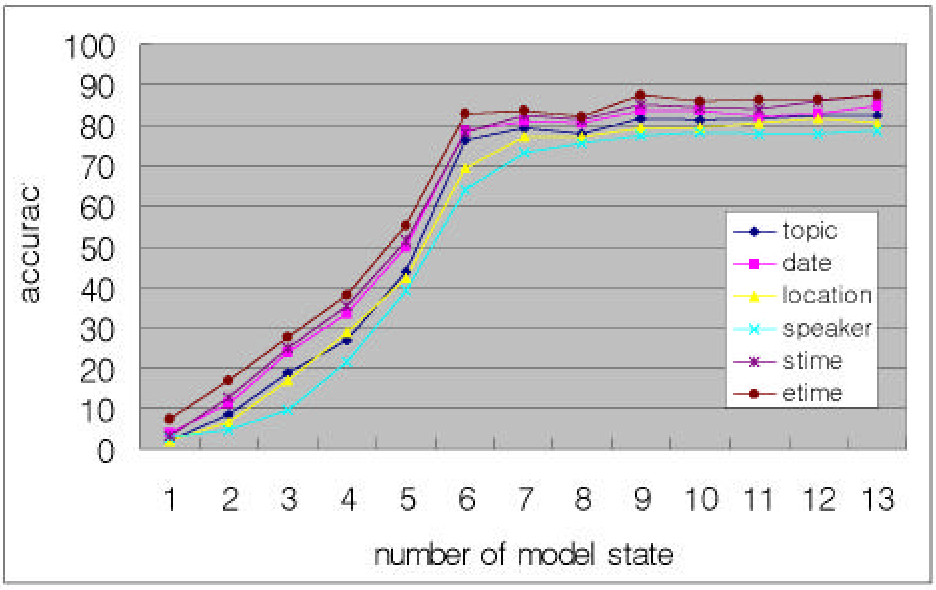


그림 18. 고정모델 HMM을 사용한 CMU 세미나 안내 데이터 정보추출

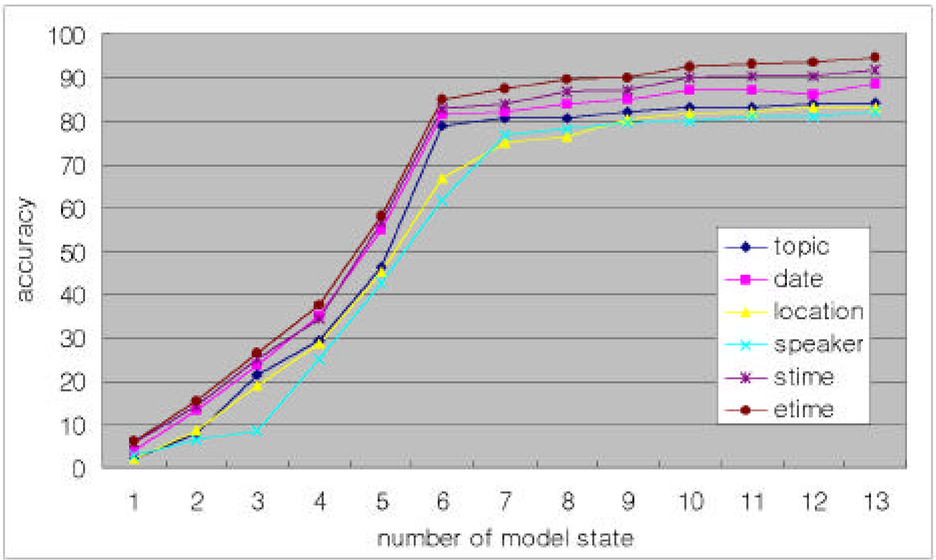


그림 19. 자기구성 HMM을 사용한 CMU 세미나 안내 데이터 정보추출

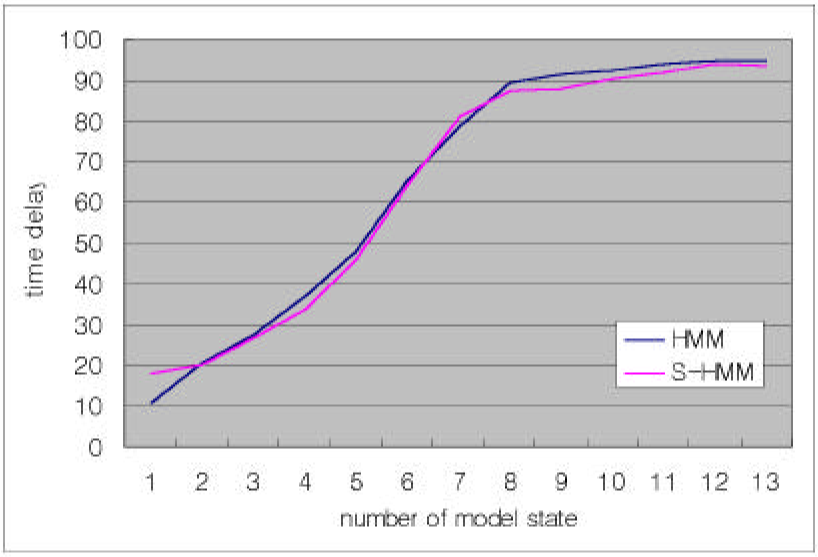


그림 20. HMM**과** S-HMM**의 시간지연** - CMU **세미나 데이터**

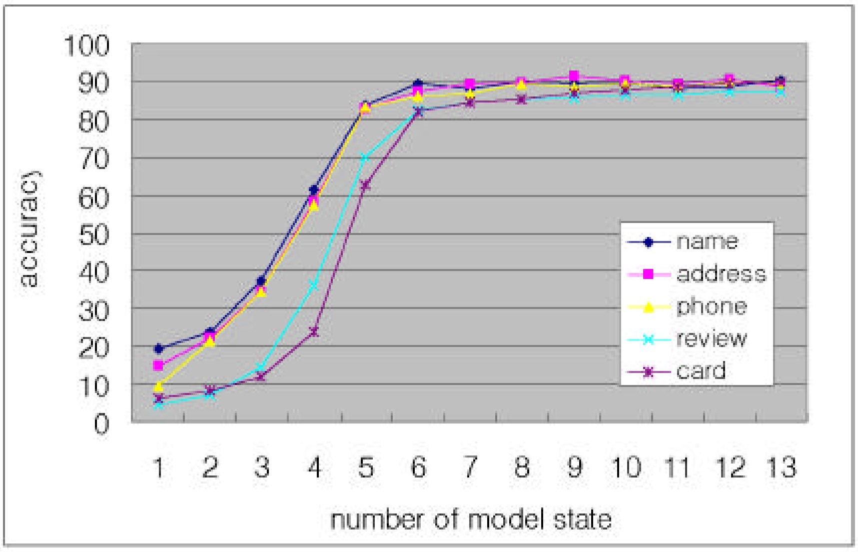


그림 21. **고정모델** HMM**을 사용한** LA Weekly **데이터 정보추출**

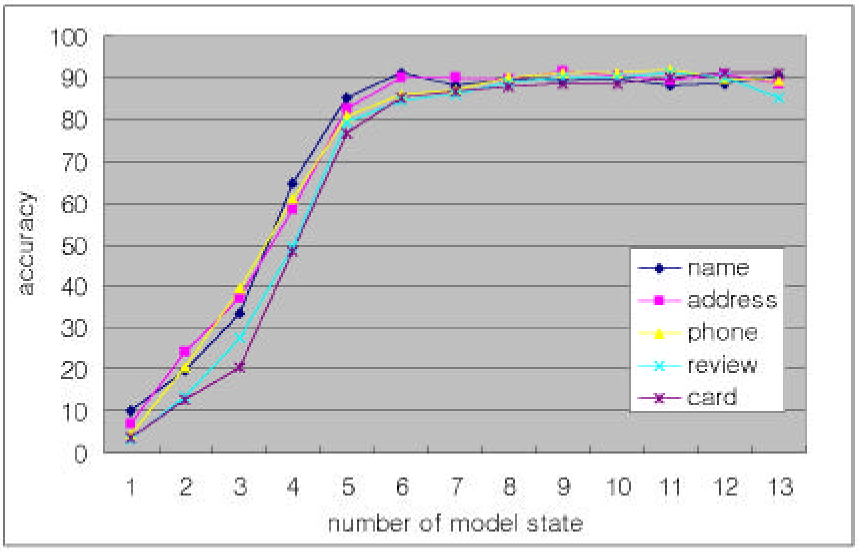


그림 22. 자기구성 HMM을 사용한 LA Weekly 데이터 정보추출

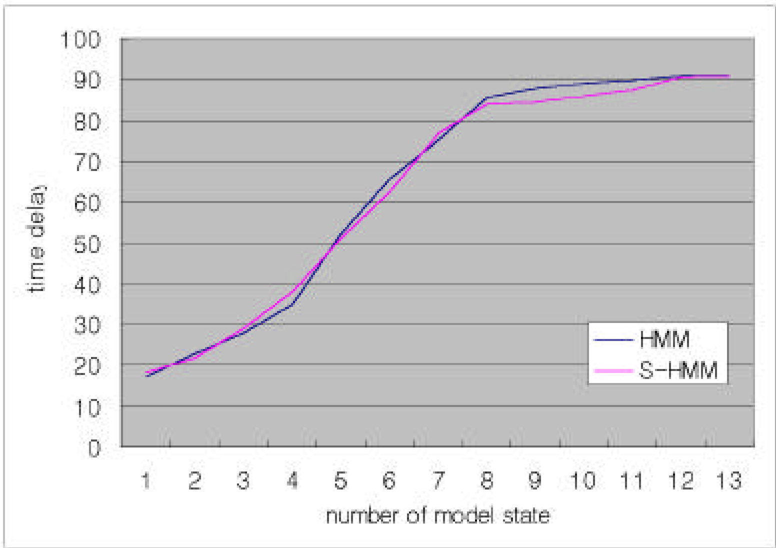


그림 23. HMM과 S-HMM의 시간지연 - LA Weekly 데이터

그림 18 19은 CMU 세미나 안내 데이터에 대한 정보추출의 결과를 나타내고 있다. 그림에서 볼 수 있듯이 상태 고정 HMM과 자기구성 HMM 모두 유

사한 성능을 보이고 있다. 하지만, 자기구성 HMM은 그림 20과 같이 시스템의 시간지연 면에서 상태고정 HMM을 앞섰다.

그림 21 22는 비교적 쉬운 데이터인 LA Weekly 데이터에 대한 정보추출 결과를 나타내고 있다. 필드들이 비교적 쉬운 형태로 데이터에 존재하기 때문에 전체적인 추출 정확도는 다른 실험에 비하여 높게 나타난다. 그림 23은 실험에 사용한 두 가지 HMM모델의 시간지연을 나타내고있다.

5. 결론

본 논문에서는 모델의 상태가 초기에 고정되는 기존의 HMM 응용과는 달리, 전처리 단계에서 데이터를 통하여 얻은 규칙을 기반으로 초기 모델의 구조를 구성하고 구성된 모델간의 거리를 계산하여 병합 가능한 모델의 상태를 하나로 묶어줌으로써 정확도를 어느 정도 유지하면서도 시스템 전체의 속도를 향상시키는 방법을 제안하였다.

자기구성 은닉마르코프모델의 경우 모델의 상태가 초기에 고정되는 경우와 달리 데이터의 규칙에 기반하여 초기 모델을 잡을 수 있어, 데이터에서 추출 필드가 연속하여 나오는 경우 이를 표현하는 모델의 각 상태들을 병합할 수 있도록 해준다. 이는 결국 시스템 전체의 시간지연을 감소시켜주는 효과를 다져오며, 두 개의 필드를 하나의 상태로 인식한다 하더라도 연속되는 필드를 하나로 표현한 것이기 때문에 추출 정확도는 그대로 유지할 수 있다.

본 논문에서는 상태를 고정시켜둔 HMM보다 어느 정도 데이터에 대한 지식을 사용하여 모델의 재구성을 허용하는 자기구성 HMM이 정보추출의 도메인에 대하여 보다 효율적임을 3개의 데이터 집합에 대한 실험을 통하여 보여주었다.

하지만, 본 논문에서 사용한 것과 같이 단순한 통계적 학습만을 이용하는 것은 훈련데이터와 차이가 있는 실제 검사 데이터에 대하여 매우 위험한 결과를 가져온다. 이에 따라서 정보추출을 수행하려고 하는 대상 문서에 대한 보다 자세하고 충분한 사전지식의 활용이 필요할 것이다. 예를 들어 XML로 구성된 텍스트 문서를 다루는 경우 XML이 지원하는 각종 태그정보를 활용하여 추출할 필드의 출현 규칙이나 필드의 속성을 학습할 수 있도록 할 수 있을 것이다. 또한 추가적으로 기본적인 자연언어처리 기술을 사용함으로써도 추출의 전반

적인 성능을 향상시킬 수 있을 것이다. 이러한 사전 지식의 활용은 시스템에서 모델의 입력으로 사용되는 관찰열을 생성하는데 있어서 보다 적은 에러를 가지는 관찰열을 생성할 수 있도록 해준다.

또한 본 연구를 수행할 때에는 실험에 사용되는 데이터를 직접 수집을 하였는데 이러한 데이터 수집의 경우 지능형 인터넷 에이전트와 같은 자동화된 시스템을 구축하여 수집 과정을 자동화할 수도 있을 것이다. 또한, 학습엔진이나 추출을 담당하는 부분을 기능별로 가능한 한 독립적으로 모듈화 하여 다른 분야로의 응용에 적합하도록 하는 연구가 필요하다.

참고 문헌

[Baldi et al, 1998] Baldi, P. and Brunak, S., *BIOINFORMATICS - The Machine Learning Approach*, MIT Press, 1998.

[Bikel et al., 1997] Bikel, D. M., Miller, S., Schwartz, R., and Weischedel, R., “Nymble: a high-performance learning”, *Proceesings of ANLP-97*, pp. 194-201, 1997.

[Cherkassky et al., 1998] Cherkassky, V. and Mulier, F. *LEARNING FROM DATA - Concepts, Theory, and Methods*", Wiley-Interscience, 1998.

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

[Cohen, 1998] 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.

[David et al., 1999] 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.

[Dempster et al., 1997] 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.

[Freitag, 1997] Freitag, D., "Using grammatical inference to improve precision in information extraction", *Working Notes of the ICML-97 Workshop on Automata Induction, Grammatical Inference, and Language Acquisition*, 1997.

[Freitag, 1998] Freitag, D., "Multistrategy learning for information extraction," *15th International Conference on Machine Learning*, 1998.

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

[Kushmerick et al., 1997] Kushmerick, N., Weld, D., and Doorenbos, R., "Wrapper Induction for Information Extraction," *International Joint Conference on Artificial Intelligent*, pp. 729-735, 1997.

[Lane, 1999] Lane, T., "Hidden Markov models for human/computer interface modeling", *In Proceedings of the IJCAI-99 Workshop on Learning about Users*, pp. 35-44, 1999.

[Leek, 1997] Leek, T. R., "Information extraction using hidden Markov models", Master's thesis, UC San Diego, 1996.

[Maes, 1994] Maes, P., "Agents that reduce work and information overload", *Communications of the ACM*, Vol. 37, No. 7, pp. 31-40, 1994.

[Mitchell, 1997] Mitchell, T. M., *Machine Learning*, McGraw-Hill, 1997.

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

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

[Seymore et al., 1999] 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.

Abstract

HMM is a kind of automata and it's internal state transitions are decided with some probabilistic values. HMM is used widely for an application with temporal data of sequential characteristics, for example, speech data. This thesis presents a new effective method for building HMM structure for information extraction tasks.

For information extraction tasks, we used HMM based on left-to-right structure. Traditional HMM are used with pre-constructed static model structure and trained its model parameter after model construction. We present here a new HMM called S-HMM (Self-Organizing Hidden Markov Model) that constructs it's structure with the rules that are obtained from training dataset.

In this paper, we used S-HMM for information extraction tasks from Call-For-Papers data, CMU online seminar announcement data, and LA restaurants review and recommendation data (http://www.laweekly.com/). We construct model structure using S-HMM from initial abstract model structure to more detailed structure with the set of rules learned from training data. We could find more appropriate structure with this set of rules. The experimental results show improved average extraction accuracy of 12% increase in average extraction speed in comparison with fixed-state HMM.

Keywords: Information Extraction, Hidden Markov Model, Model Structure Learning, Self-Organizing Hidden Markov Model

감사의 글

부모님께 감사드립니다. 공부한다는 핑계로 제대로 자주 찾아뵙지도 못했는데 항상 너그러이 이해해주시던 어머니, 그리고 아버님께 감사드립니다. 효도를 아직 시작도 하지 못하였는데 고생만 하시다가 먼저 가버리신 어머님께 늦게나마 눈물을 머금고 이제야 사랑한다는 말씀 전하고 싶습니다. "어머니 사랑합니다." 그리고 곁에서 항상 든든한 지지자로 남아있어 준 기철이, 기완이, 그리고 경인이 누님께도 고마운 마음을 전합니다.

모자란 것이 많은 저를 지도해 주시느라 고생하신 장병탁 교수님께 감사드립니다. 아직 많은 것을 배우지는 못했지만 무엇을 배워야 할지를 생각하게 해주셨습니다.

연구실의 여러분들께도 감사의 말씀을 전하고 싶습니다. 우선 같이 연구실 생활을 하며 항상 밝게 웃어주던 선이형, 인영이누나 그리고 태진이형, 현구형, 승준이 그리고 늦게 연구실에 합류한 희주, 모두 즐거운 연구실 생활을 할 수 있도록 여러모로 제게 도움을 주었습니다. 지금은 먼저 졸업하시고 안 계시지만 제가 모르는 것을 친절히 잘 가르쳐준 영우형, 광주형, 동호형, 상윤이형께도 감사의 마음을 전하고싶습니다. 많은 것을 배울 수 있도록 항상 멋진 모습을 보여준 동기들인 박사장 상욱이, 우리들의 바니 형주, 과묵한 돌고래 호식이, 항상 열심인 유환이에게도 고마운 마음을 전합니다. 그리고 연구실 선배님들인 동연이형, 수용이형, 정집이형, 성동이형, 유섭이형, 성배형, 정호형, 종우형, 장민이형, 규백이형과 고향 선배인 영용이형 그리고 기타 미처 지면을 통해 언급하지 못한 많은 분들께 감사의 말씀을 드립니다.

그리고 언제나 저를 옆에서 지켜주고 바로잡아준 연정이에게 진실한 고마움을 전합니다.