논문 모집 공고에서의 정보 추출을 위한 2 단계 은닉 마코프 모델

김정현[1] , 박성배[2] , 이상조[3]
경북대학교 언어정보연구실[1 2 3]
jhkim[1]@sejong.knu.ac.kr {seongbae[2], sjlee[3]}@knu.ac.kr

Two-Phase Hidden Markov Models for Call-for-Paper Information Extraction

Jeonghyun Kim[1], Seong-Bae Park[2], Sang-Jo Lee [3]
Dept. of Computer Engineering, Kyungpook National University[1 2 3]

요 약

본 논문은 은닉 마코프 모델(hidden Markov Model: HMM)을 2 단계로 적용하여 논문 모집 공고(Call-for-Paper: CFP)에서 필요한 정보를 추출하는 방법을 제안한다. HMM은 순차적인 흐름의 정보를 담고 있는 데이터를 잘 설명할 수 있으며 CFP가 담고 있는 정보에는 순서가 있기 때문에, CFP를 HMM으로 설명할 수 있다. 하지만, 문서를 전체적으로(global) 파악하는 HMM만으로는 정보의 정확한 경계를 파악할 수 없다. 따라서 첫 번째 단계로 CFP 문서에서 구(phrase) 단위를 구성하는 단어의 열에 대한 HMMs을 통해 국부적으로(local) 정보의 경계와 대강의 종류를 파악한다. 그리고 두 번째 단계에서 전체적인 문서의 내용 흐름에 근거하여 구축된 HMM을 이용하여 그 정보가 세부적으로 어떤 종류의 정보인지 정한다. PASCAL challenge에서 제공받은 CFP 말뭉치에 대한 첫번째 단계의 실험 결과, 0.60의 재현률과 0.61의 정확률을 보였으며, 정확률과 재현률을 바탕으로 F-measure를 측정한 결과 0.60이였다.

1. 서 론

인터넷이 발달함에 따라 온라인으로 접근이 가능한 정보의 양이 폭발적으로 늘고 있으며 정보에 대한 접근도 점점 더 쉬워지고 있다. 즉, 신문이나 학술 잡지, 또는 학위 논문 등의 각종 자료들이 디지털화되어 온라인 상에서의 문서로 서비스되고 있다. 하지만, 이러한 정보의 폭발적 증가는 정보 과부하(information overload)[1]를 초래하여 오히려 정보 이용자들이 모든 정보를 소화하기 힘들게 만들고 있다. 실례로, 이용자가 필요하던 불필요하던 상관하지 않고 검색을 통해 모든 정보를 제공하는 월드 와이드 웹(World Wide Web: WWW)에서, 정보 이용자는 자신에게 꼭 필요한 정보를 얻기 위해서는 반드시 직접 문서를 읽어야만 한다. 이에 따라, 최근에 와서는 검색 결과에서 사용자가 필요로 하는 정보를 자동으로 추출하는 시스템, 정보추출(information extraction)에 대한 필요성이 매우 높아지고 있다.

특히, 최근에는 여러 학술 분야의 저널이나 학술회의에 대한 논문 모집 공고(Call-for-Paper: CFP)가 e-mail이나 웹 페이지에 공고되는 형태로 게시되고 있다. 매년 수많은 CFP가 연구자들에게 보내어지고 있으므로, 이 CFP에서 꼭 필요한 정보만 추출하고자 하는 요구가 증가하고 있다. 예를 들면, 그림 1과 같은 CFP가 있을 때, 이 중에 많은 연구자들이 알고 싶어하는, 또 필요로 하는 정보는 많은 경우에 그림에서 밑줄로 표시한 학회 장소, 학회 기간, 논문 마감일, 논문 심사 결과 발표일 등이다.

Call for Paper
DESIGN AND MANAGEMENT OF DATA WAREHOUSE (DMDW99)

Heidelberg, Germany June 14-15, 1999
DMDW99 will be held in Heidelberg, Germany June 14-15, 1999.
DMDW99 will be co-located with COLT-1999 (July 1-4) and AI-1999 (July 8-11) at the Banff Park Lodge.

Important Dates
Submissions due Thursday, 5 February 1999
Author notifications sent Wednesday, 24 March 1999

그림 1. 논문 모집 공고(Call-for-Paper: CFP)의 예.

정보 추출은(information extraction)은 미리 필요한 정보에 대한 템플릿(templet)을 만든 후, 대용량의 자료들로부터 관심분야의 정보만을 인식하여 이 템플릿을 채우는 작업을 말한다. 본 논문에서는 CFP에서 추출하고자 하는 정보에 대해 템플릿을 만든 후, CFP에서 관심분야의 정보를 추출하여 그 템플릿을 채워 넣고자 한다. 전통적인 정보 추출 시스템의 기본 구조는 자연언어 분석 기법을 그 기반으로 한다. 즉, 대부분의 정보 추출 시스템은, 태깅, 부분 파싱, 의미 분석, 담화 분석으로 구성된다[2]. 이런 전통적인 방법은 자연언어 이해(natural language understanding)의 방법을 근간으로 텍스트에서 적합한 부분을 효과적으로 찾아서 추가적인 처리를 함으로써 정보 추출 시스템을 구성한다.

자연언어 이해에 기반한 정보 추출 방법과는 달리 기계 학습 알고리즘을 사용하는 방법이 있다. 기계 학습 알고리즘은 충분히 많은 양의 학습 데이터로부터 특정 기능을 수행하는 일반적인 지식을 습득한다. 기계 학습 알고리즘 중 대표적으로 통계적 이론에 기반한 HMM(hidden Markov Model)을 들 수 있다. HMM은 시간의 흐름에 따른 상태(state)의 천이(transition)를 확률적으로 결정하기 때문에 CFP와 같이 순차적인 흐름의 정보를 담고 있는 데이터를 잘 설명할 수 있다. 직관적으로 CFP는 하나의 HMM으로 모델링 할 수 있다. 즉, 문서를 모델링하는 HMM을 만들고 이를 이용하여 주어진 관찰 열(sequence)에 대하여 최적의 상태 열을 찾는 Viterbi 알고리즘을 이용하여 정보를 추출하는 것이다. 이와 같이, 하나의 문서에 대해 모델을 만드는 방법에서, HMM은 CFP에 나타나는 정보에 순서가 있음을 파악할 수 있었지만[3], 정보의 정확한 영역은 알 수가 없었다.

따라서, 우리는 정보의 정확한 영역을 알아낼 수 있는 HMM과 대강의 흐름을 파악하는 HMM을 2단계로 적용하는 방법을 제안한다. 첫 번째 단계로 구(phrase) 단위를 구성하는 단어의 열에 대한 HMMs을 통해, 지역적으로(local) 정보의 경계와 대강의 종류를 파악한다. 그리고 두 번째 단계에서 전체적인 문서의 내용 흐름에 근거하여 구축된, HMM을 이용하여 그 정보가 세부적으로 어떤 종류의 정보인지 정한다.

본 논문에서는 CFP에서의 정보 추출 시스템을 구축하는 것을 목적으로 한다. 그리고 그 기법으로 HMM을 이용하는데 2장에서 왜 HMM을 이용하는 것이 CFP의 정보 추출에 유리한지 알아본다. 3장에서는 CFP에서의 정보 추출 작업을 구체적으로 명시하고, 4장에서는 문제에 2 단계로 HMM을 적용하여 어떻게 모델링 할 수 있는가에 대하여 기술하였다. 마지막으로 5장과 6장에서는 구현한 시스템의 성능과 그 결과를 분석하며 결론을 맺는다.

2. CFP 정보 추출에서 은닉 마코프 모델의 이점

정보 추출 기술에서의 HMM은 문서의 유형에 종속되지 않는 범용 정보 추출 시스템에서 보다, 시간의 흐름에 따라 일반적 순서가 있는 문서에서의 정보 추출에 강점이 있다. 이는 HMM이 모델을 구성하고 있는 상태(state)들간의 천이(transition)와 각 상태에서 발생하는 관찰 기호(symbol)를 시간의 흐름에 따라 발생하는 데이터의 확률 분포에 의해 결정하기 때문이다. 이런 HMM은 CFP의 특징을 잘 설명할 수 있다. 대부분의 CFP는 학회의 이름과 개최지에 대해서 먼저 언급한 후, 학회의 세부 일정을 설명하는 날짜 정보나 관련 홈페이지 URL에 대해 기술한다. 그리고 학회의 이름과 개최지 설명에 자주 사용되는 단어가 있으며, 날짜 정보를 제시하는 부분에 자주 사용되는 단어가 정해져 있다. 예를 들면 월(月)을 나타내는 June, July와 같은 단어가 그것이다. 물론 모든 경우의 CFP가 이와 같은 규칙을 따르지는 않지만, 충분한 양의 학습 데이터를 기반으로 하여, CFP의 내용에 대한 흐름과 그 내용에서 자주 사용되는 단어들에 대하여

HMM 모델을 구축할 수 있다.

본 논문은 Baum-Welch 알고리즘을 이용하여 학습된 구(phrase)를 모델링한 HMMs에서 Forword 알고리즘을 이용하여 첫 번째 단계를 수행한다. 마찬가지로, Baum-Welch 알고리즘을 이용하여 문서를 모델링한 HMM에 Viterbi 알고리즘을 적용하여 두 번째 단계를 수행한다. Forword 알고리즘은 주어진 기호의 열이 어떤 HMM에서 발생할 확률을 구하는 알고리즘이다. 즉, 어떤 구가 여러 HMMs중 어떤 HMM에서 발생할 가능성이 가장 높은지를 결정하는데 쓰일 수 있다. 두 번째 단계에서 이용되는 Viterbi 알고리즘은 주어진 관찰 기호의 열이 발생할 가장 가능성 있는 상태(state)의 열을 구하는 알고리즘이다. 여기에서 관찰 기호의 열은 단어의 열이고, 상태는 관심있는 정보의 클래스(class)이다. 문서 전체에 대하여 HMM을 구축한 다음, Viterbi 알고리즘을 이용하여 주어진 단어의 열에 대한 클래스의 레이블을 붙일 수 있다.

3. CFP 정보 추출 작업 명세

우리가 CFP(Call-for-Paper)에서 추출하고자 하는 정보들은 아래의 표 1과 같다. 총 10 가지의 정보를 추출하는데 이는 다시 5개의 범주(category)로 나누어 진다. 5개의 범주는 Names, Acronyms, Workshop Location, Homepages, Dates정보이다. 각각의 범주는 Workshop의 세부 정보인지 Conference의 세부 정보인지, 또는 어떤 종류의 날짜인지에 따라 나누어지게 되며 총 10개의 정보 템플릿을 구성하게 된다.

범주	세부 사항	약어
Names	Workshop	NW
	Conference	NC
Acronyms	Workshop	AW
	Conference	AC
Workshop Location		WL
Homepages	Workshop	HW
	Conference	HC
Dates	Submission	DS
	Notification	DN
	Camera-ready copy	DC

표 1. 추출하고자 하는 정보.

각 범주를 구성할 수 있는 단어열(word sequence)은 각기 다른 특징을 가지고 있기 때문에 정보를 추출할 때 그 방법을 달리하여야 할 필요가 있다. 다시 말해서, 각각의 범주는 다른 특징이 나타나므로 개별적으로 HMM을 구축하는 것이 효과적이다. 아래의 표 2는 범주에 따른 대표적인 문자열을 보여준다. 표 2에서 보는 것과 같이 각각의 범주를 구성하는 단어들은 특징이 있다. 예를 들어, Workshop Location을 구성하는 단어는 지명이나 나라의 이름이 대부분이고, Dates를 구성하는 단어에는 숫자나 월(月)을 나타내는 영단어가 주로 쓰인다.

범주	Text source
N	Workshop on AGENT-ORIENTED INFORMATION SYSTEMS
A	AOIS@CAiSE' 99
WL	Brescia, ITALY
H	http://www.dexa.org
D	3-7 September 2001

표 2. 범주에 따른 text source. N은 Names, A는 Acronyms, WL은 Workshop Location, H는 Homepages, D는 Dates를 나타내는 머리글자(initial)이다.

4. 정보 추출을 위한 2 단계 은닉 마코프 모델

4.1 시스템 구조

우리는 CFP에서 정보를 추출하기 위하여, P-HMMs(Phrase HMMs)과 D-HMM(Document HMM)을 2단계로 동작시킨다. D-HMM은 CFP에서 정보가 나타나는 순서에 대한 흐름은 파악할 수 있으나, 정보의 경계 영역에 대해서 정확하게 파악할 수 없다. 따라서, P-HMMs을 이용하여 정보의 정확한 경계를 파악하고자 한다.

D-HMM은 CFP문서를 전체적으로 파악하고, P-HMMs는 지역적으로 파악하는데 목적이 있다. 특히, P-HMMs는 하나의 구(phrase)가 어떤 정보인지 파악하는데 목적이 있다. 그러나 P-HMMs는 해당 정보가 세부적으로 어떠한 정보인지 알아내는 것에는 한계가 있다. 예를 들어, '3-7 September 2001' 이 Dates 정보라는 것은 알아 낼 수 있으나, 그것이 Submission, Notification, Camera-ready copy 중 어느 것인지는 알 수 없다. 따라서, 우리는 첫 번째 단계로 P-HMMs을 이용하여 해당 구가 어떤 범주인지 파악하고, 그 다음 두 번째 단계로 세부적으로 어떤 종류인지는 전체 CFP에 나타나는 정보

의 흐름과 주변 단어 정보를 고려하는 D-HMM을 이용하여 알아낸다.

4.2 첫 번째 단계 - Phrase HMMs

기본적으로 하나의 HMM은 구(phrase)를 생성한다. 정보의 경계는 같은 구 내에 존재하는 단어들의 사이에 그어지지 않기 때문에 구 단위로 HMM을 만들면 같은 구 내의 단어 사이를 정보의 경계로 정하게 되는 잘못을 막을 수 있다. 때문에 P-HMMs에서는 각 범주를 구성하는 단어(word)를 HMM의 출력 기호(output symbol)로 본다. Homepages 범주를 제외한 나머지 범주는 각각 HMM을 구축하여 HMMs를 만든다. Homepages 범주는 정보 형식이 명확하기 때문에 P-HMMs로 파악하지 않고 URL을 검출하는 규칙으로 파악한다. 그림 2에 P-HMMs 중 하나인 Names HMM의 구조(topology)를 보였다. Names HMM은 Names 범주를 구성하게 되는 하나의 구를 생성할 수 있다. 다시 말해서, Names HMM은 아래의 그림 2에서 보는 바와 같이 'THE MACHINE LEARNING' 이라는 Names 범주를 구성하는 구를 만들어 낸다.

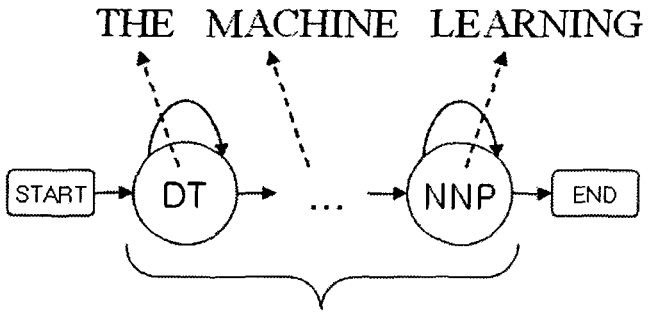

그림 2. Names HMM의 topology. 'THE MACHINE LEARNING' 이라는 Names 범주의 일부인 구를 출력 기호로 생성해 낸다. A(Acronyms), WL(Workshop Location), D(Dates)에 대한 HMM의 구조(topology)도 동일하다.

우리는 Names HMM의 구조를 디자인하기 위해 구를 구성하는 단어의 품사 태그를 이용하였다. 그러나, Names 구를 구성하게 되는 단어의 모든 품사를 이용하면 모델의 상태의 수가 지나치게 많아지게 된다. 이를 해결하기 위해, Bayesian model merging[4]을 이용하여 상태의 수를 줄이고 모델의 천이 구조(topology)를 최대한 단순화 하였다. 이는 HMM이 local maxima에 빠지는 것을 방지할 수 있다. 동일한 방법으로 WL(Workshop Location), Acronyms, Dates HMM의 구조도 정한다.

위와 같은 방법으로 구축된 HMMs에서 Forward 알고리즘을 이용하여 첫 번째 단계를 수행한다. 그림 3에서와 같이 'THE MACHINE LEARNING' 이란 Names의 일부를 구성하는 구(phrase)를 HMMs의 각각의 HMM에 대하여 Forward 알고리즘으로 검증하였을 때, Names HMM에서 가장 높은 확률을 보이므로 'OF DATA WAREHOUSE' 는 Names 범주로 정할 수 있다.

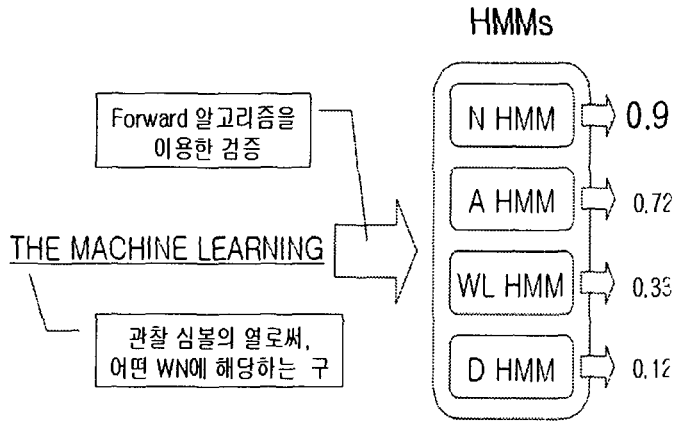

그림 3. Forward 알고리즘을 이용한 검증.

4.3 두 번째 단계 - Document HMM

D-HMM은 그림 4의 하단에서 보는 바와 같이 CFP 문서를 전체적으로 모델링하는 HMM이다. D-HMM을 구성하는 상태(state)는 관심있는 정보의 클래스(class)이고, 기호(symbol)는 첫 번째 단계의 P-HMMs에서 검증한 정보의 범주에 대한 레이블(label)이다. 첫 번째 단계에서 알아낸 정보의 범주에 대해 세부 클래스를 정하기 위하여 D-HMM을 사용한다. 예를 들어, 첫 번째 단계에서 Acronyms 범주(category)로 결정된 구에 대해서, 그것의 세부 클래스가 AW인지 AC인지 판단하는 일은 문서에 대한 HMM인 D-HMM을 이용하여 결정한다. 즉, 구 단위의 레이블을 기호의 열로 보고 그림 4의 하단에 그려진 HMM의 형태(topology)로 모델을 학습하고, Viterbi 알고리즘을 이용하여 주어진 범주에 대한 레이블이 세부적으로 어떤 정보의 클래스인

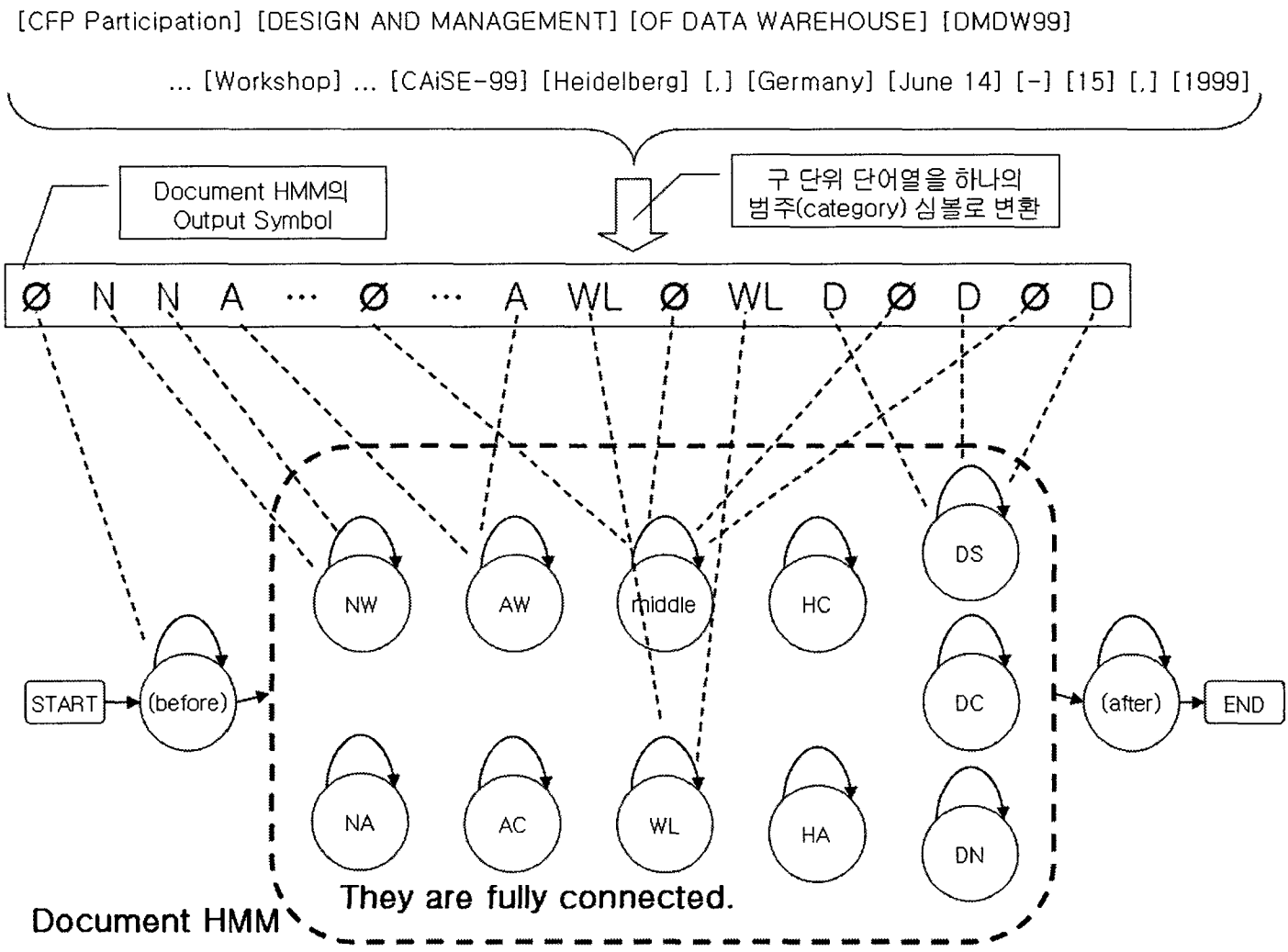

그림 4. Pharse HMM과 Document HMM의 동시 작용. 주어진 CFP의 단어 열(sequence)에 대하여, 구 단위로 묶은 다음, 각 구에 대해서 evaluation을 시행한다. 그리고 동시에 문서 전체에 대한 HMM인 D-HMM을 통하여 범주가 세부적으로 어떤 정보인지 파악한다.

지 최종적으로 결정한다.

그림 4의 D-HMM을 구성하는 상태를 보면, 정보 범주의 세부 클래스를 결정하는데 이용되지 않는 WL, HW, HC 상태를 볼 수 있다. 이는 전체의 흐름을 파악하기 위해서 실제 이용하지는 않지만 모델의 형태(topology)를 구성 하는 상태로서는 필요하다.

4.4 P-HMMs과 D-HMM의 적용

P-HMMs과 D-HMM의 적용을 설명하기 위하여, 그림 4에서의 'CFP Participation DESIGN AND MANAGEMENT OF DATA WAREHOUSE DMDW99 ...' 이라는 CFP 단어의 열을 예로 들어 설명한다. 먼저 전처리 단계로 주어진 단어의 열을 구 단위로 묶어, 구 단위의 열을 만든다. 그림 4에서 각괄호(bracket)로 묶은 단어의 열은 전체 단어의 열을 구 단위로 chunking한 것이다. 그 다음 첫 번째 단계로 열을 구성하는 각 구는 P-HMMs을 이용한 검증을 통하여, 어떤 정보 범주인지 알아낸다. 즉, chunking한 후 각 구에 대해 미리 학습해 놓은 HMMs에 대해서 검증을 실시하여 검증 결과 가장 높은 확률을 가지는 모델이 해당 구의 정보 범주 레이블이 된다. 그림 4에서 주어진 단어의 열은 Ø(null), N(Names), A(Acronym)등의 기호로 변환된다. 아래의 그림 4에서 박스로 표시된 영역이 범주 레이블의 열이다.

두 번째 단계는 D-HMM이 이용된다. D-HMM은 각각의 정보 범주 기호가 어떤 상태에서 출현하였는지 Viterbi 알고리즘을 이용하여 알아낸다. D-HMM에서의 관찰 열, 즉 정보 범주 기호열에 대한 최적의 상태 열은 해당 구가 어떤 정보의 클래스인지 나타내는 레이블의 열이 된다. 즉, 'CFP Participation' 은 관심 있는 영역의 정보가 아닌 Null 레이블이 되고, 'DESIGN AND MANAGEMENT' 은 NC 레이블을 가지게 된다.

5. 실 험 및 결 과 분 석

5.1 데이터 집합 및 실험 방법

실험에 이용된 CFP 문서는 PASCAL Challenge[5]에서 제공받은 것으로 웹에서 수집된 컴퓨터 과학, 생물학, 심리학에 관한 워크숍(workshop)과 학회(conference)의 CFP들이다. 문서의 총 개수는 400개이며, 이중에 390개의 문서는 P-HMMs의 학습에 사용하였으며, 나머지 10개의 문서로 첫 번째 단계의 성능을 테스트 하였다.

5.2 실험 결과

범주	측 정 방 법		
	정확률	재현률	F-measure
Names	0.28	0.55	0.35
Acronyms	0.48	0.67	0.56
WL	0.67	0.33	0.42
Date	0.63	0.48	0.52
NULL	0.98	0.95	0.97
평균	0.61	0.60	0.60

표 3. 실험 결과

표 3에서 Names 범주의 F-measure는 0.35로 모든 범주들 중 가장 낮은 수치이다. Names 범주에 대한 F-measure가 가장 낮은 이유는 Names 범주에 대한 HMM과 NULL 범주에 대한 HMM이 형태가 유사하기 때문이다. 즉, Names 범주에서 자주 사용되는 단어의 열은 NULL 범주에서도 발견된다. 예로, 품사가 전치사인 단어는 Names 범주와 NULL 범주 모두에서 자주 쓰인다. 따라서, 전치사들은 Names 범주에 속해 있지만 NULL 범주로 결정되는 빈도가 높은 편이다. 이는 후처리(post processing)을 통하여 해결 할 수 있다.

6. 결 론 및 향 후 과 제

CFP는 정형화된 형식은 없지만, 내용의 출현 순서에 따른 흐름이 존재한다. 따라서, 시간의 순서에 따라 들어오는 순차적인 데이터를 해석하는데 강점을 가진 HMM이 CFP를 이해하는데 적합하다. D-HMM은 전체적으로 문서가 가진 흐름을 쫓아가고, P-HMM으로 지역적으로 정보 영역의 경계를 인식한다.

CFP를 HMM으로 모델링 하는데 있어서, 주요한 이슈(issue)는 얼마나 모델을 간소화 할 수 있나에 있다. HMM에서 모델의 간소화는 출력 기호를 줄이고, 상태의 수를 줄이는 것이다. 따라서, 우리는 전체 CFP를 구단위로 쪼개어 P-HMMs으로 모델링 하여 각각의 P-HMM은 보다 적은 상태와 출력 기호를 가지게 하였다. 이렇게 구 단위로 HMM의 단위를 작게 잡으면 규칙에 의존하지 않아도 해당 구가 어떤 정보인지 인식할 수 있다. 그리고 D-HMM은 단어 자체를 출력 기호로 보지 않고, P-HMMs로 파악된 범주의 레이블을 출력 기호로 이용함으로써 모델의 기호의 수를 크게 줄일 수 있다.

1 단계, 즉 P-HMMs을 이용한 실험에서 CFP의 각 정보 영역을 구성하는 구에 특징이 있음을 알 수 있었다. 그러나 일절 규칙에 의존하지 않고, HMM을 이용한 통계적 정보에만 의존하다 보니 규칙으로 쉽게 찾을 수 있는 정보를 파악하지 못하는 결과를 초래하였다. 따라서, 전처리나 후처리를 통하여 규칙을 적절히 이용한다면 더욱 성능을 향상 시킬 수 있을 것으로 기대된다.

참고 문헌

[1] P. Maes, "Agents that Reduce Work and Information Overload," Communications of the ACM, Vol. 37, No. 7, pp. 31-40, 1994.

[2] Riloff, E., "Information Extraction as a Stepping Stone toward Story Understanding," *Understanding Language Understanding: Computational Models of Reading*, MIT Press, 1999.

[3] 김정현, 박성배 , 이상조, "은닉 마코프 모델을 이용한 Call-for-Paper 에서의 정보 추출," *인지과학회(HCI) 학술대회 발표논문집*, Vol 1, pp. 967-972, 2005

[4] Andreas Stolcke, "Bayesian Learning of Probabilistic Language Models," *Ph.D. thesis, University of California, Berkeley, CA*, 1994.

[5] http://nlp.shef.ac.uk/pascal/