한국지능시스템학회 논문지 2008, Vol. 18, No. 2, pp. 243-250

2단계 은닉 마코프 모델을 이용한 논문 모집 공고의 자동 요약

An Automatic Summarization of Call-For-Paper Documents Using a 2-Phase hidden Markov Model

김정현* · 박성배** · 이상조*** · 박세영***
Jeong Hyun Kim, Seong-Bae Park, Sang-Jo Lee and Se-Young Park

* 삼성전자 무선사업부
** 경북대학교 컴퓨터공학과 (교신저자)
*** 경북대학교 컴퓨터공학과

요 약

본 논문에서는 은닉 마코프 모델을 이용하여 논문 모집 공고에서 정보를 추출하는 시스템을 제안한다. 논문 모집 공고는 완전히 정형화된 형식을 가지지는 않지만, 내용의 출현 순서에 따른 흐름이 어느 정도 존재한다. 따라서 순차적인 데이터를 해석하는데 강점을 지닌 은닉 마코프 모델을 논문 모집 공고를 분석하는데 사용한다. 하지만, 논문 모집 공고를 은닉 마코프 모델로 직관적으로 모델링하면 정보 경계가 정확히 인식되지 않는 문제가 발생한다. 본 논문에서는 이 문제를 해결하기 위해 2-단계의 은닉 마코프 모델을 사용한다. 즉, 첫 번째 단계에서, 문서를 구로 모델링한 P-HMM(Phrase hidden Markov model)이 지역적으로 문서를 인식한다. 그리고 두 번째 단계에서 D-HMM(Document hidden Markov model)은 문서가 가진 전체적인 구조와 정보의 흐름을 파악한다. 웹에서 수집된 400개의 논문 모집 공고에 대한 실험 결과, F-measure 성능이 0.49를 보인다. 이는 직관적인 은닉 마코프 모델보다 F-measure로 0.15 정도 향상된 결과이다.

Abstract

This paper proposes a system which extracts necessary information from call-for-paper (CFP) documents using a hidden Markov model (HMM). Even though a CFP does not follow a strict form, there is, in general, a relatively-fixed sequence of information within most CFPs. Therefore, a hiden Markov model is adopted to analyze CFPs which has an advantage of processing consecutive data. However, when CFPs are intuitively modeled with a hidden Markov model, a problem arises that the boundaries of the information are not recognized accurately. In order to solve this problem, this paper proposes a two-phrase hidden Markov model. In the first step, the P-HMM (Phrase hidden Markov model) which models a document with phrases recognizes CFP documents locally. Then, the D-HMM (Document hidden Markov model) grasps the overall structure and information flow of the document. The experiments over 400 CFP documents grathered on Web result in 0.49 of F-score. This performance implies 0.15 of F-measure improvement over the HMM which is intuitively modeled.

Key Words : 정보 추출, 논문 모집 공고, 은닉 마코프 모델, 2단계 학습

1. 서 론

인터넷이 발달함에 따라 온라인으로 접근이 가능한 정보의 양이 폭발적으로 늘어나고 있으며, 정보에 대한 접근도 점점 더 쉬워지고 있다. 하지만, 이러한 정보의 폭발적 증가는 정보 과부하(information overloading)[1]를 초래하여, 오히려 정보 이용자들이 모든 정보를 소화하기 힘들게 만들고 있다. 따라서, 정보검색(information retrieval) 분야가 온라인 문서의 증가와 함께 활발히 연구되고 있다. 그렇지만, 정보검색만으로는 여전히 정보의 과다라는 문제를 풀지 못하고 있다. 즉, 늘어나는 문서를 효율적으로 검색할 수 있다고 하더라도, 정보 이용자에게 꼭 필요한 정보는 이용자가 직접 문서를 읽어 봄으로써만 얻을 수 있다. 특히, 월드와이드 웹(World Wide Web)과 같이 검색이나 정보여과(information filtering)의 결과가 방대한 경우에는 검색의 결과가 사용자에게 정보를 주기 보다는 불필요한 정보를 양산하는 문제가 발생한다. 이에 따라, 최근에는 검색 결과에서 사용자가 필요로 하는 정보를 자동으로 추출하는 시스템, 즉 정보추출(information extraction)에 대한 필요성이 대두되고 있다.

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 논문

접수일자 : 2007년 11월 18일
완료일자 : 2007년 12월 28일
본 연구는 한국과학재단 특정기초연구(R01-2006-000-11196-0)지원으로 수행되었음.

Call for Papers
The Twenty-first International Conference on Machine Learning
Banff Alberta Canada
4-8 July 2004

The Twenty-first International Conference on Machine Learning (ICML-2004) will be held in Banff Alberta Canada, 4-8 July 2004. The conference will bring together researchers to exchange ideas and report recent progress in the field of machine learning. ICML-2004 will be co-located with COLT-2004 (July 1-4) and UAI-2004 (July 8-11) at the Banff Park Lodge.

Topics for Submission
ICML-2004 invites submissions on substantial, original, and previously unpublished research on all aspects of machine learning research, including applications, techniques, theories, and connections to related fields of inquiry.

Important Dates

Abstracts due	Friday, 30 January 2004	*23:59:59 Apia, Samoa time*
Submissions due	Thursday, 5 February 2004	*23:59:59 Apia, Samoa time*
Author notifications sent	Wednesday, 24 March 2004	
Camera-ready copies of all accepted papers due (both accepted and conditionally accepted papers)	Wednesday, 21 April 2004	*23:59:59 Apia, Samoa time*

그림 1. 논문 모집 공고(CFP)의 예. 일반적으로 사각형 안에 있는 내용들이 독자의 관심 영역이 됨.
Fig. 1. An Example of Call-for-Papers. The areas boxed are in general the concern of CFP readers.

모집 공고가 연구자들에게 보내어지고 있으므로, 이 논문 모집 공고에서 꼭 필요한 정보만 추출하고자 하는 요구가 증대되고 있다. 예를 들면, 그림 1과 같은 논문 모집 공고가 있을 때, 이 중에서 많은 연구자들이 알고 싶은 정보는 대부분의 경우 아래와 같다.

- ▶ 학회 장소: Banff Alberta Canda
- ▶ 학회 기간: 4-8 July 2004
- ▶ 논문 마감일: Thursday, 5 February 2004
- ▶ 논문심사 결과발표: Wednesday, 24 March 2004
- ▶ 최종본 제출일: Wednesday, 21 April 2004

따라서, 논문 모집 공고로부터 이런 정보만 자동으로 추출하여 연구자들에게 제공하면 연구자가 관심있는 학회나 저널의 special issue에 대한 논문 모집 공고를 읽어보지 않아도 되며, 관련 있는 학회의 논문 모집 공고들을 모아서 이런 정보를 추출하면 한 눈에 어느 학회에 언제까지 논문을 제출해야 하는지 판단할 수 있다.

본 논문에서는 논문 모집 공고로부터 위와 같이 필요한 정보만 추출하기 위하여 은닉 마코프 모델에 기반을 둔 시스템을 제안한다. 논문 모집 공고를 은닉 마코프 모델로 직관적으로 모델링한 후 단순히 Viterbi 알고리듬[2]을 이용하여 정보추출을 수행하면 정보 경계가 정확히 인식되지 않는 문제가 발생한다. 이 문제를 해결하기 위해 본 논문에서는 2-단계의 은닉 마코프 모델을 제안한다. 첫 번째 단계에서, 문서를 구로 모델링한 P-HMM(Phrase hidden Markov model)이 지역적으로 문서를 인식한다. 그리고 두 번째 단계에서 D-HMM(Document hidden Markov model)이 문서가 가진 전체적인 구조와 정보의 흐름을 파악한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 관련 연구를 소개하고, 3장에서 2단계 은닉 마코프 모델을 이용한 논문 모집 공고 모델을 제안한다. 4장에서 이 모델에 대한 실험 결과를 제시하고, 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

문서로부터 필요한 정보를 추출하는 정보 추출 시스템의 기본 구조는 전통적으로 문법 규칙을 이용한 자연언어 분석에 그 기초를 두고 있다. 그리고, 또 다른 정보 추출의 방법으로는 최근에 활발한 연구가 진행되고 있는 기계학습(machine learning) 기법을 기반으로 한 방법이 있다. 자연언어 분석 기법에 기반한 정보 추출 시스템들은 대부분 구조적으로 태깅(tagging), 부분 파싱(partial parsing), 의미분석, 담화 분석으로 구성된다[3]. 이런 전통적인 방법은 자연언어 이해(natural language understanding)를 근간으로 텍스트에서 적합한 부분을 효과적으로 찾아 추가적인 처리를 함으로써 정보를 추출한다. 하지만, 이 방법은 문서의 유

형이나 특정 언어에 종속된다는 단점을 안고 있다.

기계학습에 근간을 둔 알고리듬은 충분히 많은 양의 데이터로부터 특정 기능을 수행하는 언어처리 규칙을 습득한다. 특히 기존에 음성 인식 분야에 널리 사용되어 오던 '은닉 마코프 모델(hidden Markov model: HMM)'이 정보 검색이나 다른 자연언어처리 분야에 활발히 응용되고 있다. 은닉 마코프 모델을 이용하여 모델링한 문서의 집합에서 사용자가 원하는 정보와 어느 정도 유사한 문서인지 판단하여 정보 검색을 수행하는 시스템[4]이나, 은닉 마코프 모델을 이용하여 단어의 품사를 결정하는 시스템[5] 등이 여기에 해당된다.

정보 추출과 관련된 은닉 마코프 모델에 대한 연구는 여러 연구자들에 의해 연구되었는데, 그 중 Seymore 등은 은닉 마코프 모델을 이용한 정보 추출 문제에서 모델 내부의 상태들 간의 천이(transition) 구조에 대한 연구를 수행하였다[6]. 그들은 어느 정도 양식을 가지고 밀집되어 있는 정보를 추출하기 위하여 하나의 은닉 마코프 모델을 사용하였는데, 학습된 은닉 마코프 모델의 구조를 최적화하기 위해 초기에 상세하게 구성된 상태를 여러 가지 합병 방법을 통하여 은닉 마코프 모델을 단순화하였다. 이는 모델 복잡도를 줄여 일반화 성능을 높이는 데 기여하였다.

정보추출에 관한 초기의 연구는 주로 Message Understanding Conference(MUC)에서 활성화되었는데[7], MUC 참가자들은 미리 정해진 주제에 대하여 정보 추출 시스템을 구축하고 MUC에서 개발한 공식 평가 프로그램으로 각자 시스템의 성능을 평가하였다. 이러한 연구 노력에 의해 정보 추출은 실세계의 텍스트 기반 응용 분야를 위한 필수적 기술로 인정되고 있다. 특히, 방대한 양의 문서가 존재하는 인터넷 기반의 응용 시스템에서 정보 추출은 유용한 것으로 인정되고 있다.

최근에는 PASCAL(Pattern Analysis, Statistical Modelling and Computational Learning)에서 제시된 정보 추출에 대한 도전(challenge)에 많은 연구자들이 참가하였다. PASCAL에서는 각종 기계학습 기법들이 정보 추출에서 각각 얼마나 뛰어난 성능을 발휘하는지 산정하고, 각 시스템들을 비교 평가하였다[8].

3. 논문 모집 공고의 정보 추출을 위한 2단계 은닉 마코프 모델

3.1. 논문 모집 공고의 추출 정보

논문 모집 공고에서 추출하고자 하는 정보들은 표 1과 같다. 총 10 가지의 정보를 추출하는데 이는 다시 5개의 범주로 나누어진다. 5개의 범주는 Names, Acronyms, Workshop Location, Homepages, Dates이다. 각각의 범주는 Workshop의 세부 정보인지 Conference의 세부 정보인지, 또는 어떤 종류의 날짜인지에 따라 나누어 지게 되며 총 10개의 정보 템플릿(template)을 구성한다.

각 범주를 구성할 수 있는 단어의 열(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 은닉 마코프 모델을 구축하는 것이 효과적이다. 표 2는 범주에 따른 대표적인 문자열을 보여준다. 예를 들어, Workshop Location을 구성하는 단어는 지명이나 나라의 이름이 대부분이고, Dates를 구성하는 단어에는 숫자나 월(月)을 나타내는 영어 단어가 주로 쓰인다.

표 1. 논문 모집 공고에서 추출하고자 하는 정보.
Table 1. Information to be extracted from CFP.

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 2. 각 범주에 대한 논문 모집 공고에서 추출한 정보의 예.
Table 2. An example information extracted from a CFP for each information category.

범주	범주의 예
Names	Workshop on AGENT-ORIENTED INFORMATION SYSTEMS
Acronyms	AOIS@CAiSE' 99
Workshop Location	Brescia, ITALY
Homepages	http://www.dexa.org
Dates	3-7 September 2001

3.2. 2단계 은닉 마코프 모델

직관적으로 논문 모집 공고를 모델링하는 방법은, 논문 모집 공고에서 정보가 나타나는 순서에 대한 흐름은 파악할 수 있으나 정보의 경계 영역에 대해서는 정확하게 파악할 수 없다. 논문 모집 공고에 대한 직관적인 은닉 마코프 모델의 문제점을 해결하기 위해 본 논문에서는 P-HMM(Phrase HMM)과 D-HMM(Document HMM)을 2 단계로 동작시키는 방법을 제안한다. 이 방법에서는 P-HMM을 이용하여 정보의 정확한 경계를 파악하고, 모델의 단순화를 위하여 2단계에서 문서의 구조를 파악하는 D-HMM을 적용한다.

3.2.1. P-HMM (Phrase HMM)

기본적으로 하나의 P-HMM은 구(phrase)를 생성한다. 정보의 경계는 같은 구 내에 존재하는 단어들의 사이에서는 만들어 지지 않기 때문에 구 단위로 은닉 마코프 모델을 만들면 같은 구 내의 단어 사이를 정보의 경계로 정하게 되는 오류를 막을 수 있다. 이를 위해 P-HMM에서는 각 범주를 구성하는 단어를 은닉 마코프 모델의 출력 심볼(output symbol)로 본다. Homepages 범주를 제외한 나머지 범주는 각각의 은닉 마코프 모델을 구축하여 은닉 마코프 모델들을 만든다. Homepages 범주는 정보 형식이 명확하기 때문에 P-HMMs로 파악하지 않고 URL을 검출하는 규칙으로 파

악한다.

그림 2에 P-HMM들 중 하나인 Names HMM의 구조(topology)를 보였다. Names HMM은 Names 범주를 구성하게 되는 하나의 구를 생성할 수 있다. 즉, Names HMM은 'THE MACHINE LEARNING'이라는 Names 범주를 구성하는 구를 만들어 낸다.

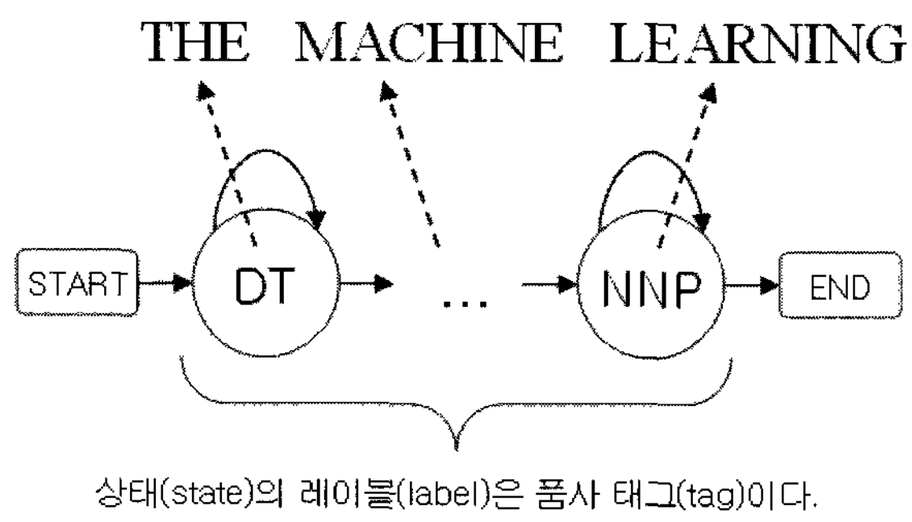


그림 2. Names HMM의 구조.
Fig. 2. The structure of Names HMM.

P-HMM의 구조를 작성할 때 구를 구성하는 단어의 모든 품사를 이용하면 상태의 수가 너무 많아지는 문제가 발생한다. 이를 해결하기 위해, Bayesian model merging[9]을 이용하여 상태의 수를 줄이고 모델의 천이 구조를 최대한 단순화 하였다. 이는 은닉 마코프 모델이 극대값(local maxima)에 빠지는 것을 방지할 수 있다.

위와 같은 방법으로 구축된 P-HMM들에서 전향 알고리듬(forward algorithm)을 이용하여 주어진 구의 종류를 결정한다. 그림 3이 이 과정을 도시하고 있다. 예를 들어, 'OF DATA WAREHOUSE'이란 구가 주어졌을 때, 이를 전체 P-HMM에 적용하여 각 P-HMM의 확률값을 얻는다. 주어진 구에 대해서, Names HMM이 가장 높은 확률을 출력하므로, 'OF DATA WAREHOUSE'는 Names 범주로 결정된다.

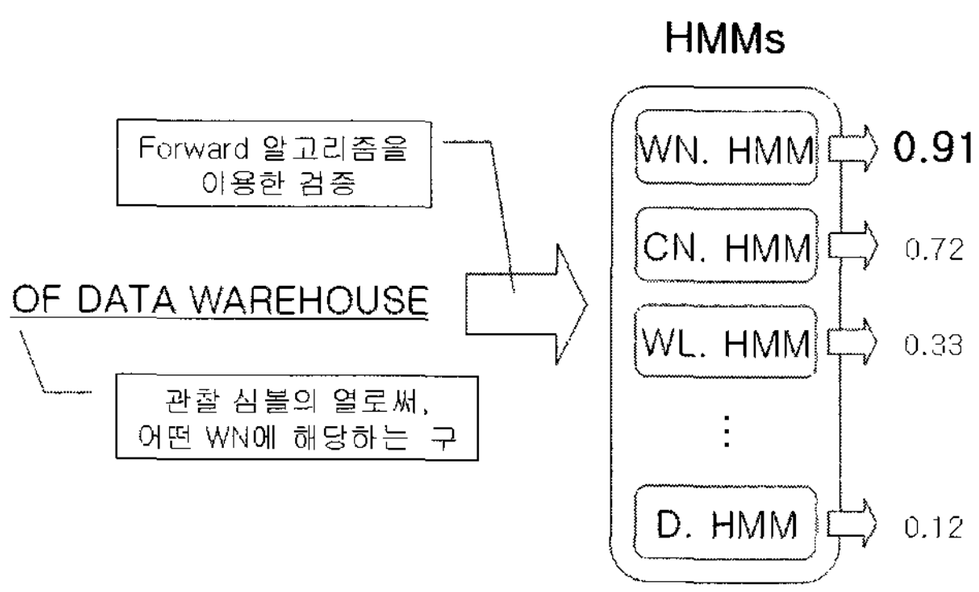


그림 3. 전향 알고리듬을 이용한 구의 클래스 분류.
Fig 3. Classification of phrases using the forward algorithm.

3.2.2. D-HMM(Document HMM)

D-HMM은 그림 4와 같이 논문 모집 공고 문서를 전체적으로 모델링하는 은닉 마코프 모델이다. D-HMM을 구성하는 상태(state)는 관심 있는 정보의 클래스(class)이고, 기호(symbol)는 P-HMM에서 검증한 정보의 범주에 대한 레이블(label)이다. 즉, 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 AC인지는 문서에 대한 은닉 마코프 모델인 D-HMM을 이용하여 결정한다.

구 단위의 레이블을 기호의 열로 보고 학습한 결과 그림 4의 구조(topology)를 가진 은닉 마코프 모델이 학습되었다. 이 그림은 비교적 의미적인 값을 가지는 천이만 표시하였다. 이렇게 학습된 모델에 대하여 Viterbi 알고리듬을 수행하여, 주어진 범주에 대한 레이블이 세부적으로 어떤 정보의 클래스인지 최종적으로 결정한다.

3.2.3. P-HMM과 D-HMM의 적용

P-HMMs과 D-HMM의 적용을 설명하기 위하여, 그림 5에서의 'CFP Participation DESIGN AND MANAGEMENT OF DATA WAREHOUSE DMDW99 …'이라는 논문 모집 공고 단어의 열을 예로 들어 설명한다. 먼저 전처리 단계로 주어진 단어의 열을 구 단위로 묶어, 구 단위의 열을 만든다. 그림 5에서 각괄호(bracket)로 묶은 단어의 열은 전체 단어의 열을 구 단위로 단위화(chunking)한 것이다. 그 다음 첫 번째 단계로 열을 구성하는 각 구는 P-HMM을 이용한 검증을 통하여, 어떤 정보 범주인지 알아낸다. 즉, 단위화(chunking)한 후 각 구에 대해 미리 학습해 놓은 은닉 마코프 모델들에 대해서 검증을 실시하여 검증 결과 가장 높은 확률을 가지는 모델이 해당 구의 정보 범주 레이블이 된다. 이 그림에서 주어진 단어의 열은 Ø(null), N(Names), A(Acronym)등의 기호로 변환된다. 박스로 표시된 영역이 범주 레이블의 열이다.

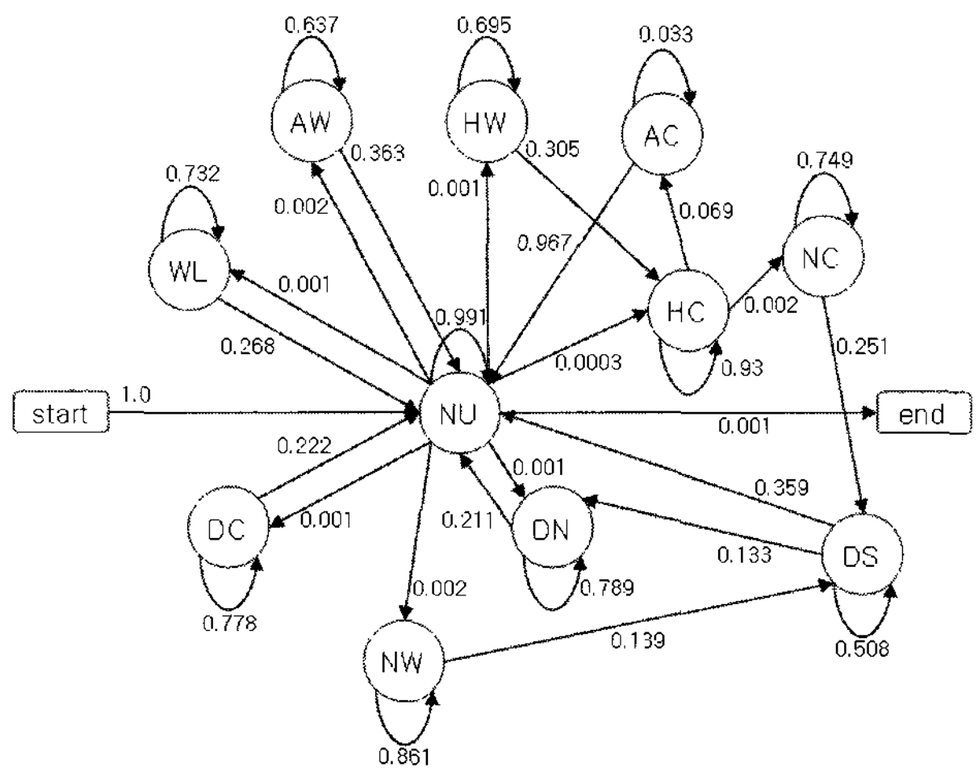


그림 4. D-HMM 구조.
Fig. 4. The topology of D-HMM.

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

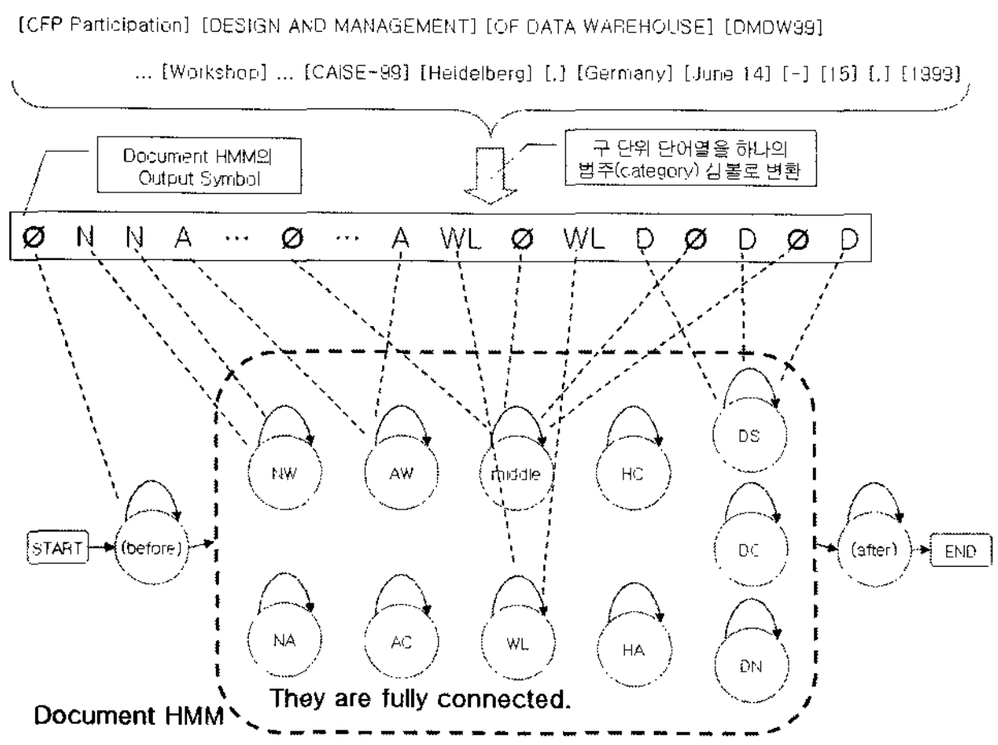


그림 5. P-HMMs와 D-HMM의 적용 예.
Fig. 5. An example of applying P-HMMs and D-HMM to information extraction.

4. 실 험

4.1. 실험 데이터

실험에 이용된 논문 모집 공고 데이터는 웹에서 수집된 컴퓨터 과학, 생물학, 심리학에 관한 워크숍(workshop)과 학회(conference)의 논문 모집 공고이다. 논문 모집 공고의 총 수는 400개로, 각 논문 모집 공고 문서는 그림 6과 같이 정답 레이블이 태깅되어 있다. 400개의 문서 중, 390개의 문서를 학습에 이용하였으며, 나머지 10개의 문서는 태그를 제거하여 시스템의 성능을 검증하는데 이용하였다.

<conferenceacronym>AAAI-2000</conferenceacronym>
<workshopname>WORKSHOP on CONSTRAINT DATABASES in AI</workshopname>
<workshoplocation>Austin, TX</workshoplocation>
<workshopdate>July 31, 2000</workshopdate>
FINAL CALL FOR PAPERS
The last few years have seen a growing interest in the use of
constraint databases for supporting several problems considered by AI
researchers. This workshop focuses on the use of constraint databases
for such problems.
Topics
Constraint-based agents.
Planning and scheduling with temporal constraints.

그림 6. 정보 태그가 달린 논문 모집 공고의 예
Fig. 6. An example of tagged CFP.

4.2. 비교 모델

본 논문에서 제안한 2단계 은닉 마코프 모델의 성능을 객관적으로 평가하기 위하여, 기본 모델을 제안한 후 2단계 은닉 마코프 모델과 비교한다. 본 논문에서 사용된 기본 모델은 직관적으로 설계된 은닉 마코프 모델이다. 이 모델은 추출하고자 하는 정보 태그를 상태로 삼고 논문 모집 공고에서 사용된 단어를 관측 심볼로 삼아 설계된 은닉 마코프 모델이다.

이론적으로 관측 심볼의 수는 실세계에 존재하는 단어의 수와 같다. 그러나 전체 단어를 모두 다루는 것은 현실적으로 불가능하므로, 본 논문에서는 학습용으로 사용된 논문 모집 공고 문서 390개에서 나타난 단어들만 관측 심볼로 간주하였다. 학습 데이터로 쓰인 논문 모집 공고에서 나타나는 단어들의 수는 문서 당 평균 약 600개이며, 학습용 논문 모집 공고 400개에 대해서 약 240,000 개의 단어가 관측 심볼로 존재한다. 여기에 스테밍(stemming) 과정과 중복 단어 제거 과정을 거치면 약 35,000 개의 단어만 남는다. 그러나 35,000 단어의 대부분은 어느 관심 정보 태그에도 속하지 않는 의미 없는 단어이며, 이런 단어의 대다수는 한 문서에서 나타나면 다른 대부분의 문서에서는 관측되지 않는다. 그리고 이런 단어들에 의한 데이터의 분포는 추론 시 대부분의 확률 값이 0이 되는 희소 문제를 야기한다. 따라서 데이터의 부족으로 나타나는 희소 문제를 해결하기 위하여 관찰 심볼을 관심 있는 영역에서 발생한 단어들로만 한정하였다.

4.3. 실험 결과

하나의 은닉 마코프 모델로 직관적으로 논문 모집 공고를 모델링한 경우, 문서 형식의 흐름을 따라가는 것은 대략적으로 가늠하다. 그러나 각각의 정보가 구체적으로 무엇이 될 것인지, 해당 정보의 정확한 시작과 끝 영역은 어디인지를 결정하는데 있어서는 상당히 미흡한 성능을 보였다. 직관적 모델링 방법에 대한 성능은 아래의 표 3과 같다.

표 3. 직관적 은닉 마코프 모델의 성능.
Table 3. The performance of intuitive HMM.

범주	약어	정확률 (%)	재현률 (%)	F-measure
Names	NW	0.19	0.35	0.25
	NC	0.22	0.40	0.28
Acronyms	AW	0.10	0.28	0.15
	AC	0.12	0.58	0.20
WL	WL	0.07	0.56	0.13
Homepages	HW	0.44	0.84	0.58
	HC	0.55	0.06	0.11
Dates	DS	0.60	0.69	0.64
	DN	0.69	0.65	0.67
	DC	0.57	0.36	0.44
평균		0.35	0.47	0.34

논문 모집 공고 문서에 포함된 각 구 단위의 단어 열을 P-HMM의 평가 문제를 통하여 성능을 측정한 결과는 표 4와 같다. 표 4에서 Names 범주의 F-measure는 0.35로 모든 범주들 중 가장 낮은 수치이다. Names 범주에 대한 F-measure가 가장 낮은 이유는 Names 범주에 대한 은닉 마코프 모델과 NULL 범주에 대한 은닉 마코프 모델의 형태가 유사하기 때문이다. 즉, Names 범주에서 자주 사용되는 단어의 열은 NULL 범주에서도 발견된다. 예로, 품사가 전치사인 단어는 Names 범주와 NULL 범주 어느 곳도 가리지 않고 모두 자주 쓰인다. 따라서 전치사들은 Names 범주에 속해 있지만 NULL 범주로 결정되는 빈도가 높은 편이다.

표 4. P-HMM 실험 결과.
Table 4. Experimental results for P-HMM.

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 5는 최종적으로 2단계 은닉 마코프 모델을 적용한 결과이다. 전체적인 F-measure는 0.49로 논문 모집 공고를 직관적으로 모델링한 방법을 이용하였을 때의 0.34보다 나은 성능을 보였다.

표 5. 2단계 은닉 마코프 적용 실험 결과.
Table 5. Experimental results of 2-phase HMM.

범주	약어	정확률	재현률	F-measure
Names	NW	0.65	0.24	0.35
	NC	0.77	0.34	0.47
Acronyms	AW	0.73	0.25	0.38
	AC	0.66	0.23	0.34
WL	WL	0.62	0.40	0.48
Homepages	HW	0.67	0.41	0.51
	HC	0.55	0.06	0.11
Dates	DS	0.76	0.63	0.69
	DN	0.87	0.77	0.81
	DC	0.84	0.66	0.74
평균		0.71	0.40	0.49

표 6은 본 논문에서 제시된 방법의 결과와 PASCAL challenge에 제시된 ITC-IRST(Istituto per la Ricerca Scientifica e Tecnologica) 시스템[10]의 정보 추출 결과를 비교한 표이다. ITC-IRST 시스템은 Support Vector Machines을 근간으로 하여 정보 추출 전 미리 정보를 가지고 있지 않는 단어를 필터링하여 보다 정확한 정보를 추출을 한다.

표 6에서 2-HMMs는 본 논문에서 제시한 방법의 결과이고, ITC는 ITC-IRST 시스템의 실험결과이다. 이 표에 따르면 ITC-IRST 시스템이 본 논문에서 제시된 방법보다 전반적인 성능이 조금 더 뛰어나다. 그러나 본 논문에서 제시된 실험 결과는 성능 향상을 위한 어떤 전처리나 후처리 작업을 하지 않은 채 측정된 것이다. 즉, 단지 정보 추출에 은닉 마코프 모델을 적용하는 방법만을 다르게 함으로써, 직관적인 모델링 방법보다 더 나은 성능을 볼 수 있었음을 보였다. 그러므로, 다른 시스템에서처럼 전처리나 후처리 작업을 통하면 거의 비슷한 성능을 보일 수 있을 것으로 기대된다.

표 6. ITC-IRST 시스템과의 성능 비교.
Table 6. Performance comparison against ITC-IRST.

범주	약어	정확률		재현률		F-measure	
		2-HMMs	ITC	2-HMMs	ITC	2-HMMs	ITC
Names	NW	0.65	0.83	0.24	0.50	0.35	0.63
	NC	0.77	0.81	0.34	0.34	0.47	0.48
Acronyms	AW	0.73	0.71	0.25	0.28	0.38	0.40
	AC	0.66	0.61	0.23	0.20	0.34	0.31
WL	WL	0.62	0.79	0.40	0.38	0.48	0.51
Home pages	HW	0.67	0.68	0.41	0.44	0.51	0.53
	HC	0.55	0.54	0.06	0.09	0.11	0.16
Dates	DS	0.76	0.83	0.63	0.63	0.69	0.72
	DN	0.87	0.92	0.77	0.78	0.81	0.84
	DC	0.84	0.92	0.66	0.68	0.74	0.78
평균		0.71	0.76	0.40	0.43	0.49	0.54

그림 7은 세 가지 방법, 즉 직관적인 은닉 마코프 모델 모델, 2단계 은닉 마코프 모델, ITC-IRST 방법을 F-measure를 사용하여 각 정보 영역 별로 비교한 그림이다. 위에서 설명한 바와 같이 전처리나 후처리와 같은 튜닝(tuning)을 한 ITC-IRST와 본 논문에서 제시된 2단계 은닉 마코프 모델 방법의 성능이 비슷한 것을 볼 수 있다. 워크숍의 이름(NW)을 추출하는 부분에서 비교적 성능의 차이가 남을 볼 수 있는데 이는 해당 영역의 적절한 전처리나 후처리로 시스템의 성능을 향상 시킬 수 있음을 증명하는 것이다. 나머지 영역에서도 적절한 전처리와 후처리를 하면 ITC-IRST의 성능을 능가할 수 있을 것으로 생각된다.

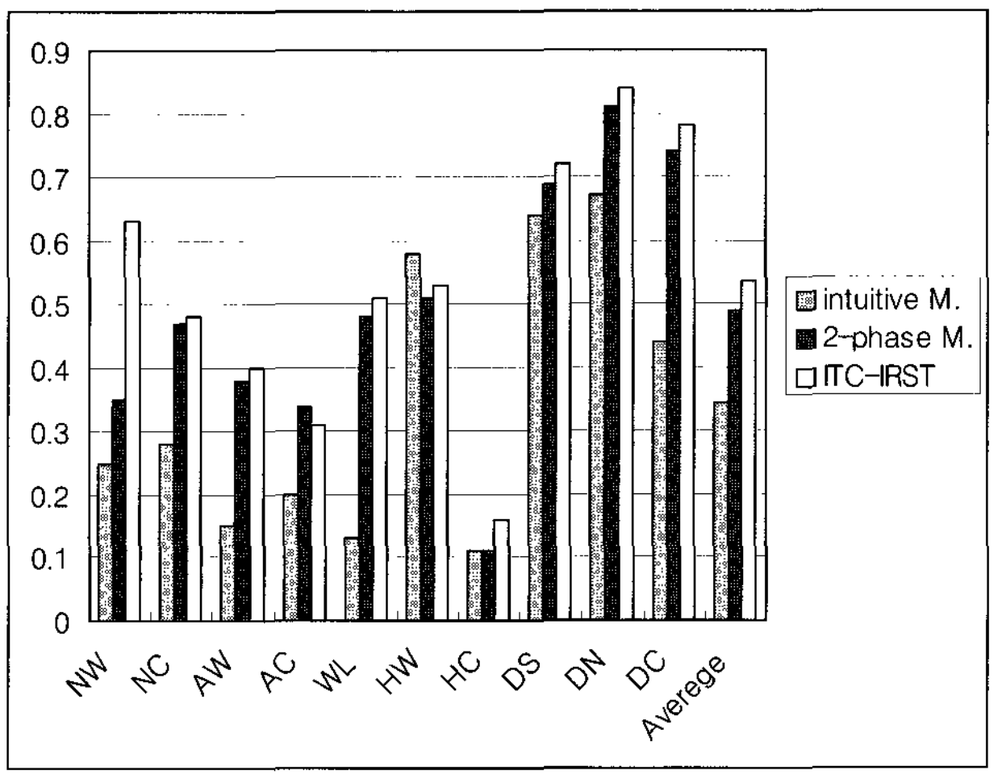


그림 7. 2단계 은닉 마코프 모델과 ITC-IRST 시스템의 F-measure 비교.
Fig. 7. F-measure comparison between 2-phase HMM and ITC-IRST system.

5. 결 론

본 논문에서는 은닉 마코프 모델을 이용하여 논문 모집

공고에서 필요한 정보를 추출하였다. 논문 모집 공고는 정형화된 형식은 없지만, 내용의 출현 순서에 따른 흐름이 존재한다. 따라서 시간의 순서에 따라 들어오는 순차적인 데이터를 해석하는데 강점을 가진 은닉 마코프 모델은 논문 모집 공고 문서를 모델링 하는데 적합하다. 본 논문에서는 논문 모집 공고를 은닉 마코프 모델로 직관적으로 모델링한 후, 단순히 Viterbi 알고리즘을 이용하여 정보 추출을 하는데서 나타나는 문제인, 정보 경계 인식의 문제를 해결하여 보다 나은 성능의 정보 추출 시스템을 제시하였다.

직관적 모델링 방법에 따른 문제를 해결하기 위해, 우리는 미리 정보 경계에 대해 구 단위로의 단위화(text chunking)을 수행하여, 나누어 질 수 있는 곳과 나누어 없는 곳을 분명히 하였다. 또한, 모델의 간소화를 위해서 2단계로 은닉 마코프 모델을 적용하였다. 먼저 1단계로, 정보 영역의 경계를 구분하여 부분 단어 열들로 모델링한 P-HMM은 지역적으로 문서의 정보를 인식하였다. 그리고 D-HMM은 전체적으로 문서가 가진 흐름을 파악하였다.

P-HMM을 이용한 실험에서 논문 모집 공고의 각 정보 영역을 구성하는 구는 고유의 특징이 있음을 알 수 있었다. 그러나 규칙에 전혀 의존하지 않고, 은닉 마코프 모델을 이용한 통계적 정보에만 의존하다 보니 규칙으로 쉽게 찾을 수 있는 정보를 파악하지 못하는 결과를 초래하였다. 이 1단계 오류는 2단계 D-HMM에도 반영되어 전체적인 시스템의 성능 저하에 기인하였다. 따라서 보다 세밀한 전처리나 후처리를 통하고 단어 규칙이나 문법 규칙을 적절히 이용한다면 더욱 성능을 향상 시킬 수 있을 것으로 기대된다.

참 고 문 헌

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

[2] A. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, Vol. 13, No. 2, pp. 260-267, 1967.

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

[4] D. Miller, T. Leek, and R. Schwartz, "A Hidden Markov Model Information Retrieval System," In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 214-221, 1999.

[5] C. Manning and H. Schuetze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.

[6] K. Seymore, A. McCallum, and R. Rosenfeld, "Learning Hidden Markov Model Structure for Information Extraction," In *Proceedings of AAAI '99 Workshop on Machine Learning for Information Extraction*, pp. 37-42, 1999.

[7] N. Chinchor, "Overview of MUC-7/MET-2," In *Proceedings of the 7th Message Understanding Conference*, 1998.

[8] N. Ireson, F. Ciravegna, M. Claiff, D. Freitag, N. Kushmerick, and A. Lavelli, "Evaluating Machine Learning for Information Extraction," In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 345-352, 2005.

[9] A. Stolcke, *Bayesian Learning of Probabilistic Language Models*, Ph.D Thesis, University of California, Berkeley, 1994.

[10] C. Giuliano, A. Gliozzo, A. Lavelli, and L. Romano, "Filtering Uninformative Words to Speed up IE: ITC-irst Participation in the PASCAL Challenge," *PASCAL Challenge*, 2005.

저 자 소 개

김정현(Kim, Jeong Hyun)
2004년 : 영남대학교 컴퓨터공학과 학사
2006년 : 경북대학교 컴퓨터공학과 석사
2006년 ~ 현재 : 삼성전자 정보통신총괄 무선사업부 연구원

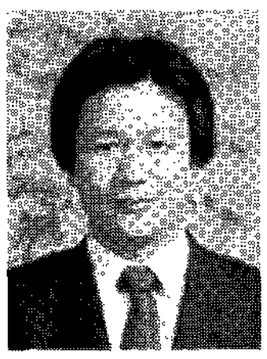
관심분야 : 기계학습, 자연언어처리, 정보검색

e-mail : rafy.kim@samsung.com

박성배(Park, Seong-Bae)
1994년 : 한국과학기술원 전산학과 학사
1996년 : 서울대학교 컴퓨터공학과 석사
2002년 : 서울대학교 전기컴퓨터공학부 박사
2004년 ~ 현재 : 경북대학교 컴퓨터공학과 교수

관심분야 : 기계학습, 자연언어처리, 바이오인포매틱스

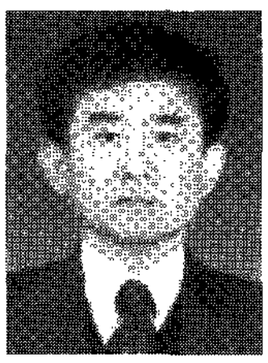
e-mail : seongbae@knu.ac.kr



이상조(Lee, Sang-Jo)
1974년 : 경북대학교 수학교육과 학사
1976년 : 한국과학기술원 전산학과 석사
1994년 : 서울대학교 컴퓨터공학과 박사
1976년 ~ 현재 : 경북대학교 컴퓨터공학과 교수

관심분야 : 자연언어처리, 정보검색, 기계번역

e-mail : sjlee@knu.ac.kr



박세영(Park, Se-Young)
1976년 : 경북대학교 전자공학과 학사
1982년 : 한국과학기술원 전산학과 석사
1989년 : 프랑스 파리 7대학 전산학과 박사
1982년~2000년 : 한국전자통신연구원 연구부장
2000년~2003년 : 주식회사 서치캐스트 대표이사
2003년~2005년 : 정보통신연구진흥원 전문위원, 정통부 디지털컨텐츠 S/W 분야 담당 PM
2005년~현재 : 경북대학교 컴퓨터공학과 교수

관심분야 : 시맨틱웹, 자연언어처리, 정보검색

e-mail : seyoung@knu.ac.kr