HMM 모델을 이용한 의료 문서 대상 고차원 개념 태깅

장혜주○ 송사광 맹성현
한국정보통신대학교
{hjjang○, smallj, myaeng}@icu.ac.kr

High Level Semantic Tagging in Clinical Documents Using a HMM Model

Hyeju Jang0, Sa Kwang Song, Sung Hyon Myaeng
Information and Communications University

요 약

본 논문에서는 의료임상 문서의 구절(phrase)를 대상으로 고차원 개념의 정보를 태깅하는 시맨틱 태깅 시스템를 제안하고 있다. 시스템은 의사들이 기록한 임상 기록으로부터 정보를 추출한다. 태깅은 UMLS와 POS, 약어 태깅이 된 문서를 대상으로 HMM 모델에 의거하여 이루어지게 된다. 태깅된 결과는 의료 상에서의 경험적 지식을 추출하는데 이용되어 의사들의 의사 결정을 지원하게 된다.

1. 서 론

경험이 중요한 의료 분야에서 의사에 의해 쓰여진 환자기록은 가치를 따질 수 없이 귀중한 정보이다. 만약 의사가 환자 기록으로부터 유용한 정보를 쉽게 찾을 수 있다면, 그 정보는 현재 환자의 문제를 다루고 치료하는 데에 이용될 수 있다. 즉, 의료 분야에서의 과거의 정보는 미래의 진료 방식을 결정하는 데 큰 역할을 하게 된다.

예를 들어, 만성적인 질병을 치료하는 경우, 그 동안 환자가 보여 온 증상, 치료, 효과에 대한 과거의 기록은 의사들이 그 환자의 질병을 다루는 여러 가지 방법에 대해 더욱 잘 이해하도록 해 준다. 그 결과, 그 정보들은 그 다음 행해져야 할 치료에 대한 방향 결정에 도움을 주게 된다.

더욱이, 요즘에는 의료 상황을 컴퓨터로 기록·보관하는 병원이 늘고 있다. 이렇듯, 컴퓨터로 인식 가능한 형태의 문서의 증가는 광대한 양의 의료 정보를 언어학적으로 그리고 통계적으로 분석하여 이용할 수 있도록 해 준다. 상당한 양의 의료 문서에 내포되어 있는 지식이 자동적인 방법에 의해 추출·이용될 수 있다.

본 논문에서는 의료 정보 추적 시스템에서 의료 임상 문서를 대상으로 고차원적인 개념의 시맨틱 태그를 다는 태깅 시스템에 관하여 서술한다. 본 시스템에서 태그는 '증상', '처방', '성능'과 같이 의료 문서의 각 구절(phrase)이 담고 있는 정보의 종류가 된다. 이러한 태그를 통해 추적 시스템은 의사들이 알고 싶어하는 과거 케이스의 검색을 할 수 있게 된다. 본 태깅 시스템은 시맨틱 태깅을 위해 기존의 어휘 및 의학 용어 자원과 Hidden Markov Model(HMM)[1]을 사용한다.

본 연구가 갖는 의의는 크게 두 가지로 요약할 수 있다. 실용적인 관점에서 볼 때 이 연구는 의사들이 과거의 환자 기록에 포함되어 있는 지식을 이용하여 그들의 간접 경험을 늘릴 수 있는 기회를 제공한다. 기술적인 관점에서, 본 시스템은 의료 임상 문서를 대상으로 syntactic level의 분석이나 단순히 단어의 종류에 대한 의미 정보를 제공하기보다 구절을 단위로 구절이 뜻하는 의학적 개념을 정보로서 제공하려는 시도를 한다. 또한, 본 시스템은 robust tagging을 위하여 학습셋 문서에 나타나지 않는 구절(unknown phrase)의 의미를 추측하는 방법을 제공한다.

2. 관련 연구

[10]과 [13]은 의료 도메인의 sublanguage의 특성에 대하여 언급하였다. 그들에 따르면 의료 기록은 의사 외의 보통 사람들은 이해 못 하는 수많은 전문 용어와 약어 그리고 기호를 포함하고 있다. 예를 들어서, 위를 향하는 화살표 기호는 종종 '증가'라는 뜻을 의미한다. 그리고 'sl'는 'slight'의 약어이다.

이제까지의 보편적이고 전통적인 POS 태깅 시스템에서는 가장 알맞은 태그를 찾기 위해 HMM 모델을 사용하였다 [2]. 일부 시스템들은 HMM 모델을 강화시키기 위하여 특정 기능을 더 붙여서 사용하였다. [3]와 [4]는 HMM 모델을 이용한 POS 태깅 시스템에 각각 ambiguity class와 euivalence class라는 개념을 도입하였다. 본 시스템 역시 단어 그 자체를 사용하기보다 적당한 클래스로 그룹지어 이용하는 equivalence class의 개념을 개조하여 이용한다.

현재까지의 의료 분야의 태깅 시스템은 주로 어휘 레벨의 syntactic 또는 시맨틱 태깅을 하였다. [5]과 [6] 은 Unitified Medical Language System (UMLS)을 이용하여 의학 용어에 대하여 어휘적으로 의미 태깅을 하였다. 그리고 [7]과 [8]는 단어에 대한 POS 태거를 만들었다. 반면에 본 시스템은 의미 단위라 할 수 있는 구절에 대한 태깅을 시도한다.

의료 기록으로부터 정보를 추출하는 여러 시스템 역시 있어 왔다[9, 10, 11]. [9]와 [10]은 6개의 포맷 타입을 정의하여 의료 기록에서 많은 정보를 포맷에 맞추어 분류하였다. [13]은 의료 문서를 대상으로 약어 disambiguation에 대한 연구를 수행하였다.

3. 목적 태그

본 태깅 시스템의 목적은 CDA 문서 대상 의료 정보 추적 시스템을 위하여 의료 임상 문서에 시맨틱 태깅을 하는 것이다. 본 연구는 의사가 관심 있어 하는 질문 리스트(서울대에서 제공)에 대한 답을 추적하는 것을 궁극적인 목적으로 하고 있고 특히, 그 중

'X가 Y의 치료에 어떻게 쓰이나'와 'Y의 상황에서 X의 효과가 무엇인가'라는 두 가지의 질문에 대한 답을 단기 목표로 삼고 있다. 여기에서 X는{Medical Device, Biomedical or Dental Material, Food, Thepapeutic or Preventive Procedure}, Y는 {Finding, Sign or Symptom, Disease or Syndrome}으로 각각 대체될 수 있다. 그 두 질문에 대답하기 위하여, 본 연구에서는 서울대학교 병원으로부터 제공 받은 Clinical Data Architecture (CDA) 문서[12]에서 각 환자의 과거 치료 기록인 '입원후경과'부분을 지식 소스로, 질문에 답할 수 있도록 해당되는 구절에 적당한 시맨틱 태그를 할당한다. 따라서, 많은 질문에 답하기 위해서는 많은 태그가 필요하지만, 본 시스템에서는 '증상', '처방', '효과'를 목적 태그로 하고 있다.

본 시스템의 목적 태그는 여타의 다른 태깅 시스템과 본 시스템을 구별짓는 중요한 요소 중 하나이다. 왜냐하면 목적 태그 자체가 고차원 개념이기 때문이다. 그리하여 단어의 문법적 역할과 의학적 의미를 태깅하는 POS, UMLS 태그와는 달리, 본 시스템의 목적 태그는 사용자에게 직접 서비스하는 어플리케이션 시스템에 의해 바로 활용될 수있다. 목적 태그는 어플리케이션 시스템의 목적에 따라 바뀔 수 있다. 학습셋 문서만 바꿔 주면 다른 목적 태그를 가진 태깅 시스템으로 확장될 수 있다.

4. HMM 모델을 이용한 시맨틱 태깅 시스템 구성

시맨틱 태그를 구절에 할당하는 시스템을 위해서는 여러 가지 방법이 있을 수 있다. 본 연구에서 제안하는 방법은 사람이 무언가를 기록할 때는 특정 순서를 따른다는 사실에 기인하였다. 예를 들어, 보통 원인 뒤에는 결과가 오게 되고 사건은 시간 순서로 기술되게 된다. 본 연구는 CDA 문서의 narrative 데이터가 그러한 암묵적인 순서를 따르고 있다고 가정한다.

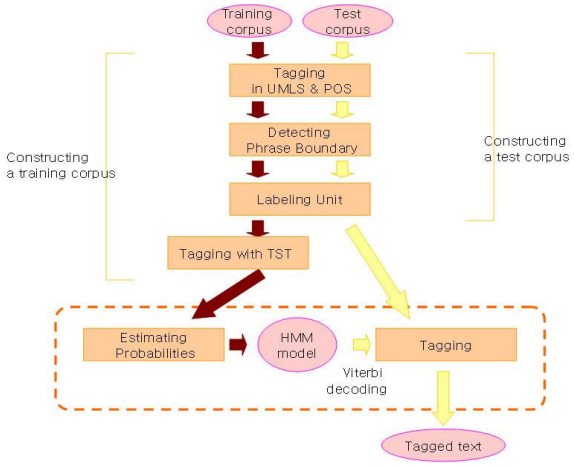
다른 POS 태거가 문법 규칙이나 패턴, 자주 나오는 순서를 상황정보(context)로 하여 HMM 모델을 이용하는 것처럼 본 연구에서는 의료 임상 문서의 순서 측면을 모델링하기 위하여 HMM 모델을 채택하였다. 그러나, 본 연구에서는 문서에 한글과 영어 단어들이 혼재해 있기 때문에 언어의 문법적인 요소를 충분히 활용할 수 없다. 하지만, 사람은 특정 순서로 메모를 하는 경향을 갖고 있다는 사실에 입각하여 HMM을 이용한 시맨틱 태깅 시스템을 제안한다.

HMM를 이용한 태거의 시스템의 구성은 [그림 1]과 같다. 전체는 크게 학습 단계와 태깅 단계로 나뉜다.

4.1. 학습 단계

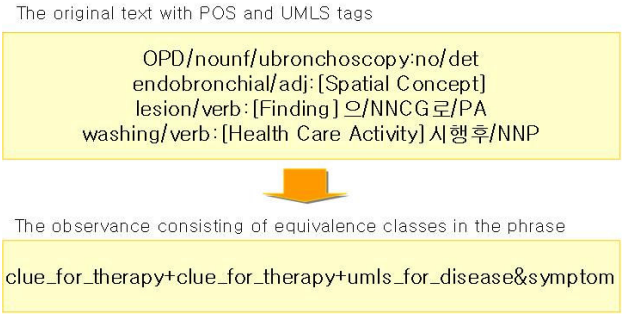
우선, 문서에 POS 태깅과 UMLS 태깅을 수행하여 단어의 기본적인 문법 역할과 의학 용어일 경우 어떤 종류의 용어인지를 분명하게 한다. 그 후, POS, UMLS 태깅된 정보와 기호와 수치 표현을 이용하여 약어 disambiguation을 시행한다[13]. 이는 본 연구에서의 목적인 시맨틱 태깅을 하기 위해서는 문서의 분석이 다소 필요하기 때문에 수행하는 전초 작업이라 할 수 있다.

기초가 되는 태깅 작업이 끝난 후에는 문서를 구절로 구분하여 각 구절을 단위로 한다. 이 단계에서 구분된 구절을 대상으로 각 구절에 태깅을 수행하게 되는 것이다. 구절은 POS 태그셋 중 "/EFC", "/EFN", "/NNCV", "후/NNCG", "후/NNP", "함/NNP", "/EFF"을 이용하여 구분하였다. 태깅의 단위를 단어가 아닌 구절로 한 이유는 이 연구의 목표가 고차원적인 개념을 태깅해 주는 것이기 때문에 태깅의 대상으로 의미를 가진 단위가 필요하기 때문이다. 그리고, CDA 문서의 의사 기록 부분에는 문장이 불분명하게 나타나기 때문에 의미상 주어와 용언으로 이루어진 구절을 대상으로 하였다.



[그림 1] 태깅 시스템의 전체 구성

그 다음 과정은 전 단계에서 구분된 구절 안의 단어가 담고 있는 정보를 equivalence class와 연결시켜 그 구절을 대표하는 패턴을 만드는 것이다. equivalence class는 단어와 관련된 일반화된 범주들의 집합이다. 이는 데이터 희소성의 문제를 해결해 주고 본 시스템에서는 구절이 갖고 있는 여러 가지 정보를 효과적으로 표현해 주는 방법이기도 하다. [표 1]에서는 구절을 이루는 단어를 위한 equivalence class의 종류를, [그림 2]에서는 구절과 equivalence class의 연결을 보여 주고 있다.



[그림 2] 구절과 equivalence class의 연결

equivalence class와의 연결이 끝난 후에는 학습셋을 구축하기 위하여 수동으로 목적 태그를 구절마다 표시해 준다. 수동 태깅을 마지막으로 학습셋이 완성되게 된다.

그 후에는 학습셋을 대상으로 빈도수(frequency) 계산을 통하여 HMM 모델에서 필요한 전이 확률과 발산 확률을 구함으로써 학습셋에 맞는 HMM 모델을 수립하게 된다. 수동 태깅이 된 학습셋이 없어도 된다는 Baum-Welch algorithm을 통한 학습의 장점이 있지만 태깅의 정확도가 빈도수 계산을 통한 학습보다 현저하게 떨어진다는 연구 결과[14,15]가 있기에 본 연구에서는 이용하지 않았다.

4.2. 태깅 단계

태깅이 필요한 문서를 대상으로 UMLS, POS, 약어 태깅와 equivalence class와 연결시키는 단계까지를 학습 단계에서와 같은 방법으로 수행한다. 그 후에는 학습 단계에서 수립된 HMM 모델에 기반하여 Viterbi 알고리즘[16]을 이용하여 가장 알맞은 태

그 sequence를 찾아내어 태깅을 한다. 이 때 학습셋에 나타나지 않았던 구절 패턴(unknown phrase)의 태그는 unknown phrase의 패턴과 기존 학습셋에 나타났던 패턴의 유사도를 통해 예측하는 방법을 사용하였다.

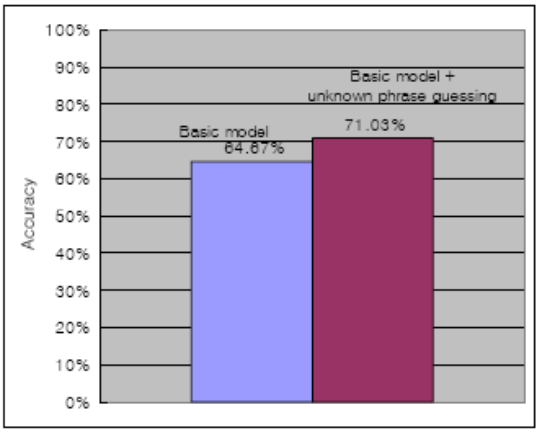
[표 1] Equivalence class의 종류

. Equivalence classes on words

UMLS tag for cause	Biomedical or Dental Material, Food
UMLS tag for disease or symptom	Finding, Sign or Symptom, Disease or Syndrome, Neoplastic Process
UMLS tag for therapy	Diagnostic Procedure, Food, Medical Device, Therapeutic or Preventive Procedure
Clue word for therapy	처방(prescription), 복용(administer medicine), 시행(operation), 후(after), 이후(later), 사용(use), 증량(increase), 수술(surgery), 중단(discontinue)
Clue word for symptom	발열(having fever), 관찰(observe)
Clue word for performance	호전(improvement), 감소(decrease), 상승(rise), 정상(normal), 발생(occurrence), 변화(change)
unknown	neither clue word nor UMLS tag

5. 성능 평가 및 분석

시스템을 객관적으로 평가하기 위해서 서울대학교 병원으로부터 제공 받은 CDA 문서 중 300 문서의 '입원후 경과' 섹션으로 수동으로 태깅을 하여 문서셋을 구축하였다. 이 중 200개는 학습셋, 100개가 평가셋이며 학습셋은 1187개의 태깅 대상이 되는 구절로 이루어져 있고, 평가셋은 601개로 이루어져 있다. 본 시스템을 평가하기 위해 사용한 평가지수(measure)은 총태그 수에 대한 정확한 태그 수의 비율이다. [그림 3]은 기본적인 모델로 측정한 시스템 성능과 unknown phrase에 대한 처리가 추가된 시스템의 성능을 각각 보여 주고 있다. 기본적인 모델에서보다 unknown phrase에 대한 처리가 추가된 시스템의 성능이 6% 정도 향상된 것을 볼 수 있다.



[그림 4] 시스템 성능

비록 시스템의 성능이 기대했던 것만큼 좋지는 않지만 시스템 개선의 방향을 이번 실험 분석을 통하여 다음과 같이 정할 수 있었다.

(1) 목적 태그셋을 확장하여 문서 내에서의 전이 확률을 좀더 정확하게 만들고자 한다.

(2) HMM 모델에서 가장 적절한 초기 확률을 찾는다.

(3) 단순한 unknown phrase 추측 방법을 보다 지능적으로 발전시킨다.

(4) 최대한 수동 태깅된 데이터를 많이 확보한다.

6. 결론 및 향후 연구 방향

본 논문에서는 HMM 모델에 기반한 의료 문서에 대한 시맨틱 태깅 시스템을 제안하였다. 고차원 개념의 의미 태깅 결과는 의료 정보 추적 시스템과 같은 응용 소프트웨어에서 바로 사용될 수 있는 정보라는 점에서 의의가 있다. 향후에는 의료 문서의 또 다른 중요한 정보원인 기호와 수치 표현을 각 구절을 나타내는 데 반영하고 좀더 다양한 태그를 선정하여 의료 문서에 포함돼 있는 귀중한 정보를 보다 정확하고 유용하게 추출할 계획이다.

<Acknowledgement>

본 연구는 보건복지부 Korea Health 21 R&D 프로젝트 (02-PJ1-PG6-HI03-0004) 지원으로 수행되었음.

7. 참고 문헌

[1] L.R.Rabiner et all, "An Introduction to Hidden Markov Models", IEEE ASSP Magazine, 1986

[2] Linda Van Guilder, "Automated Part of Speech Tagging:A Brief Overview", Handout for LING361, 1995

[3] Julian Kupiec, "Robust part-of-speech tagging using a hidden Markov model", Computer Speech and Language, pp. 225-242, 1992.

[4] Doug Cutting et all, "A Practical Part-of-Speech Tagger", In Proceedings of the 3rd ACL, pp.133-140, 1992

[5] Patrick Ruch, "MEDTAG: Tag-like Semantics for Medical Document Indexing", In Proceedings of AMIA'99, pp.35-42

[6] Stephen B. Johnson, "A Semantic Lexicon for Medical Language Processing", J Am Med Inform Assoc. 1999 MayJun; 6(3): 205-218

[7] Udo Hahn, "Tagging Medical Documents with High Accuracy", Pacific Rim International Conference on Artificial Intelligence Auckland, Newzealand , pp. 852-861, 2004

[8] Hans Paulussen, "DILEMMA-2: A Lemmatizer-Tagger for Medical Abstracts", In Proceeings of ANLP, pp.141-146, 1992

[9] Carol Friedman, "Automatic Structuring of Sublanguage Information," London: IEA, 1986, pp. 85-102.

[10] Emile C. Chi et all, "Processing Free-text Input to Obtain a Database of Medical Information", In Proceedings of the 8th Annual ACM-SIGIR Conference, 1985

[11] Udo Hahn, "Automatic Knowledge Acquisition from Medical Texts", In Proceedings of the 1996 AMIA Annual Fall Symposium, pp.383-387, 1996

[12] What is CDA?: http://www.h17.org.au/CDA.htm#CDA

[13] Sa Kwang, Song, "Abbreviation Disambiguation Using Semantic Abstraction of Symbols and Numeric Terms,", IEEE NLP-KE, 2005

[14] David Elworthy, "Does Baum-Welch Re-estimation Help Taggers?", Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 1989

[15] Bernard Merialdo, "Tagging English Text with a Probabilistic Model", Computational Linguistics 20.2, pp155-172, 1994

[16] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm", IEEE Transactions of Information Theory 13, pp 260-269, 19