음절 바이그램과 CRFs를 이용한 의학 전문 용어 추출

송수민, 신준수, 김학수
강원대학교 컴퓨터정보통신전공
e-mail:macdowelly@gmail.com, nlpsjs@kangwon.ac.kr, nlpdrkim@kangwon.ac.kr

Biomedical Terminology Extraction using Syllable Bigram and CRFs

Soo-Min Song, Junsoo Shin, Harksoo Kim
Program of Computer and Communications Engineering,
Kangwon National University

요 약

웹(Web)상에 전문용어를 포함한 문서가 증가함에 따라 전문용어를 자동으로 추출하는 연구가 계속해서 이루어지고 있다. 기존 연구에서는 전문용어를 추출하는 단계에서 대부분 형태소 분석기를 이용한다. 그러나 전문용어의 특성으로 인해 형태소 분석 단계에서 오분석 되는 경우가 발생한다. 이러한 문제를 해결하기 위해서 본 논문에서는 음절 바이그램과 CRFs(Conditional Random Fields)를 이용하여 의학 전문 용어를 추출하는 방법을 제안한다. 네이버 지식인의 의사 답변 문서 2000개로부터 5-fold cross validation을 이용하여 실험하였다. 실험 결과 정확률은 평균 68.91%, 재현율은 평균 71.25%로 나타났으며 F-measure는 70.06%로 나타났다.

1. 서론

인터넷이 발전함에 따라 웹상에 다양한 주제의 문서들의 수가 증가하고 있다. 이 중 특정 분야의 전문적인 내용을 담고 있는 문서들도 크게 늘어나고 있다. 그림 1은 네이버 지식iN의 의료 상담 카테고리로 하루에도 수천 개의 의료관련 문서들이 생성되고 있다.

그림 1. 지식iN 의료상담

이와 같이 전문 용어를 사용하는 문서가 크게 늘어남에 따라 전문 용어를 자동으로 추출하고자 하는 연구가 활발하게 이루어지고 있다. 기존의 전문 용어 추출 연구에서는 형태소 분석기를 이용하여 자질을 추출하고 이를 대상으로 전문 용어를 판단하려는 연구가 있었다. 그러나 전문 용어는 일반적인 코퍼스에서는 거의 출현하지 않는다. 그렇기 때문에 형태소 분석 단계에서 오분석 되어 에러파급 효과가 커질 수 있다. 형태소 분석 단계에서 생기는 오분석 문제를 줄이기 위해서는 계속해서 등장하는 신조어 및 전문 용어에 대한 코퍼스 관리가 필요하다. 그러나 이러한 방법은 비용이 많이 든다는 단점이 있다. 이러한 문제를 해결하기 위해서 본 논문에서는 음절 바이그램 자질을 바탕으로 CRFs(Conditional Random Fields)를 이용한 의학 전문 용어 추출 시스템을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서 제안하는 시스템에 대해서 설명하고 3장에서 시스템의 성능 평가 및 결과를 기술하고 4장에서 결론 및 향후연구 과제에 대해서 기술한다.

2. 음절 바이그램 기반의 의학 전문 용어 추출

본 논문에서 제안하는 시스템의 순서도는 그림 2와 같다.

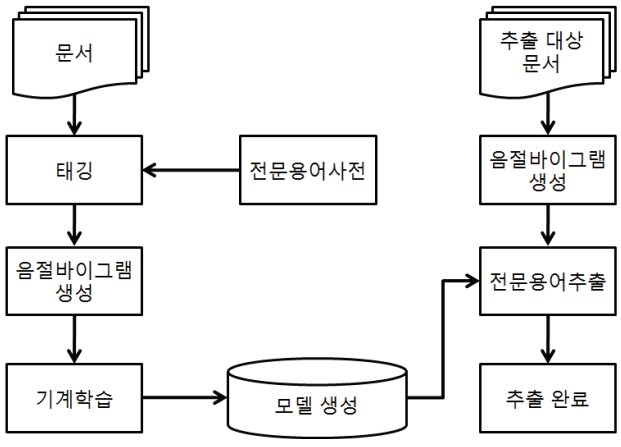

그림 2. 의학 전문 용어 추출 시스템의 순서도

시스템은 학습과정과 추출과정으로 구성되어있다. 태깅 단계에서는 음절 기반으로 BIO 태깅을 수행한다. B는 전문 용어의 시작 음절을 의미하며, I는 전문 용어의 중간 또는 끝 음절을 나타내고, O는 전문 용어가 아닌 일반적인 음절을 나타내는 태그를 붙이는 것이 BIO태깅이다. 자질 추출 단계에서는 음절 바이그램을 생성한다. 이를 기반으로 CRFs 기법의 기계학습을 한다.

2.1 태깅

전문 용어는 일반적인 사전이나 말뭉치에서 출현하지 않는 경우가 많다. 그렇기 때문에 형태소 분석기를 이용하게 되면 전문 용어의 분석 결과에서 원래 단어가 가진 형태를 변형하는 오류가 발생할 수 있다. 이러한 문제를 해결하기 위해서 본 논문에서는 음절을 기반으로 B, I, O 태깅을 한다. 전문 용어 판단은 '필수의학용어집'를 기반으로 하였다. 그림 3은 B, I, O 태깅의 예이다.

원본	동상후의 응급처치는 깨끗한 생리식염수일수록 좋구요
BIO 태깅	동 상 후 의 x 응 급 처 치 는 x 깨 끗 한 B I O O O B I I I O O O O O
	생 리 식 염 수 일 수 록 x 좋 구 요 B I I I I O O O O O O O

그림 3. B, I, O 태깅

2.2 음절 바이그램 생성

BIO 태깅 후 음절 바이그램을 생성한다. 음절 유니그램을 사용하게 되면 음절 내의 정보량이 부족하기 때문에 높은 성능을 기대하기 어렵다. 이러한 문제를 해결하기 위해서 음절 바이그램의 자질을 추출한다. 그림 4는 음절 바이그램을 생성하는 예이다.

원본	동상후의 응급처치는
BIO 태깅	동상 상후 후의 의x x응 응급 급처 처치 치는 B I O O O B I I I

그림 4. 음절 바이그램 생성

2.3 CRFs 학습

CRFs는 레이블링 문제에서 높은 성능을 보이고 있는 기계학습 방법이다. CRFs는 다양한 입력 노드의 값이 주어졌을 때 지정된 출력 노드의 조건부 확률값을 계산하기 위한 무방향성 그래프 모델이다. CRFs의 주요 장점은 HMM(Hidden Markov Model)의 단점인 독립 가정을 완화시키는 효과가 있다는 것과 MEMM(Maximum Entropy Markov Model)의 단점인 레이블 편향 문제(label bias problem)를 극복할 수 있다는 것이다. 이러한 이유에서 최근 자연어처리 분야에서 많이 사용되는 통계기반의 기계학습 모델 중의 하나이다. 그림 5는 그림 4의 일부를 CRFs 모델로 도식화한 것이다. 그림 5에서 Y_i는 B, I, O의 레이블을 나타내고 X_i는 음절 바이그램 자질을 나타낸다.

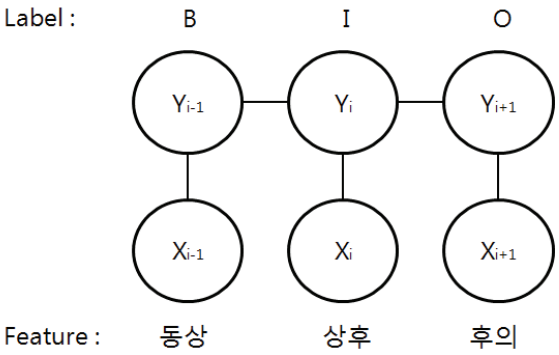

그림 5. CRFs

3. 실험 및 결과

본 논문에서 제안하는 시스템을 평가하기 위해서 지식iN 의료 상담의 20 카테고리에서 의사가 직접 답변한 문서 2000개를 수집하였다. 평가는 정확률, 재현율을 측정하였으며 이를 기반으로 F-measure 값을 계산하였다. 정확률은 시스템에서 판단한 전문 용어 중 정확하게 판단한 전문 용어의 비율을 나타내고, 재현율은 실제 정답 전문 용어 중 시스템이 정확하게 판단한 전문 용어의 비율을 나타낸다. F-measure는 정확률과 재현율을 함께 나타내는 평가 기준으로 사용하였다.

$$\text{정확률} = \frac{\text{올바르게 판단된 전문 용어의 수}}{\text{시스템이 판단한 전문 용어의 수}} \quad \text{(식 1)}$$

$$\text{재현율} = \frac{\text{올바르게 판단된 전문 용어의 수}}{\text{정답 전문 용어의 수}} \quad \text{(식 2)}$$

$$F-mesure = \frac{2 \times \text{정확률} \times \text{재현율}}{\text{정확률} + \text{재현율}} \quad \text{(식 3)}$$

5-fold cross validation으로 실험하였으며 정확률, 재현율의 결과는 표 1과 같다.

표 1. 실험 결과

	1	2	3	4	5	평균
정확률	0.6284	0.6979	0.7060	0.6961	0.7169	0.6891
재현율	0.7056	0.7166	0.7110	0.7116	0.7175	0.7125

정확률과 재현율이 비슷한 성능을 보이는 것을 확인 할 수 있다. 정확률과 재현율을 함께 나타내는 F-measure는 0.7006으로 나타났다. 한글 문서를 대상으로 전문 용어를 추출하는 연구는 많지 않기 때문에 직접적인 비교가 어렵

다. 간접적인 비교를 위해 전문용어 추출시스템[1]의 결과와 성능 비교를 하였다. 표 2는 전문용어 추출시스템[1]과의 성능 비교 결과이다.

표2. 성능 비교

	정확률	재현율	F-measure
[1] 시스템	0.591	0.684	0.6341
제안 시스템	0.689	0.712	0.7006

[1]의 연구 결과와 비교하였을 때 더 높은 성능을 내는 것을 확인할 수 있다. 그러나 사전, 실험 대상이 전혀 다르기 때문에 직접적으로 비교는 불가능하지만 간접적으로 본 논문에서 제안하는 시스템의 성능을 평가할 수 있다. 본 논문에서는 비교적 간단한 방법을 제안하였다. 음절 바이그램만을 이용하여 신뢰 있는 수준의 성능을 나타내는 것을 확인 할 수 있다.

4. 결론 및 향후과제

본 논문에서는 음절 바이그램 기반의 전문 용어 추출 방법을 제안하였다. 형태소 분석 단계의 오분석을 해결하기 위해 음절 바이그램을 자질로 사용하였으며, CRFs 기반의 기계학습을 이용하였다. 그 결과 간단한 방법으로 정확률 0.6891, 재현율 0.7125, F-measure 0.7006로 나타났다. 직접적인 시스템 비교는 어렵지만 표2와 같이 간접적인 비교 결과 신뢰할 만한 성능을 나타냄을 확인하였다.

향후과제는 다음과 같다. 본 논문에서는 음절 바이그램만을 자질로 사용하였는데, 더 많은 정보를 포함하기 위해서 앞 뒤의 어절을 함께 자질로 사용하는 방법을 연구할 계획이다. 또한 추출한 전문 용어의 범주를 판단하는 연구를 계속해서 진행할 예정이다.

참고문헌

[1] 박정오, 황도삼, “전문용어 추출시스템”, 한국정보과학회 2000년도 봄 학술발표논문집, 제7권 제1호, pp. 381~383, 2000.

[2] 오종훈, 최기선, “정보통합을 통한 생물/의학 분야 전문용어의 자동 추출”, 한국정보과학회 2004년도 가을 학술발표 논문집, 제31권 제2호(Ⅰ), pp. 775~777, 2004.

[3] 오종훈, 최기선, “기계학습에 기반한 생의학분야 전문용어의 자동 인식”, 정보과학회논문지, 제33권 제8호, pp. 718~729, 2006.

[4] 배영준, 최호섭, 옥철영, “백과사전 기반 전문용어 태깅 시스템”, 한국정보과학회 언어공학연구회 학술발표 논문집, 2005.

[5] 김재호, 배선미, 신효식, 최기선, “의학 전문용어의 정의문 자동 추출”, 한국정보과학회 2004년도 봄 학술발표 논문집, 제31권 제1호(B), pp. 922~924, 2004.