EM 알고리즘을 이용한 전문용어의 자동 추출

오종훈°, 김재호, 최기선
한국과학기술원 전자전산학과/전문용어언어공학연구센터
{rovellia°, jjaeh, kschoi}@world.kaist.ac.kr

Automatic Term Recognition Through EM Algorithm

Jong-Hoon Oh°, Jae-Ho Kim, Key-Sun Choi
Department of EECS
Korea Advanced Institute of Science and Technology/KORTERM

요 약

전문용어란 전문분야의 개념이 언어적으로 표현된 형태이다. 전문분야마다 분야 특성적인 개념이 사용되므로, 전문용어는 전문분야를 특성화하는 단위로 사용된다. 따라서 전문분야문서에 대한 자연언어처리에서 전문용어를 효과적으로 처리하는 것은 매우 중요하다. 전문용어 추출은 분야 특성적인 전문용어를 해당 분야 문서에서 파악하는 작업을 말한다. 본 논문에서는 기계학습방법을 이용한 전문용어 자동 추출 기법을 제안한다. 본 논문의 기법은 전문분야 사전과 전문분야 문서를 이용하여 문서에서 나타나는 전문용어의 특성을 파악하고 이를 이용하여 전문용어를 추출한다. 본 논문의 기법은 70,000단어 수준의 영어 의학분야 300개 문서에 대하여 약 77%의 정확률로 전문용어를 추출하였다.

1. 서론

전문용어란 전문분야의 개념이 언어적으로 표현된 형태이다[1]. 전문분야마다 분야 특성적인 개념이 사용되므로, 전문용어는 전문분야를 특성화하는 단위로 사용된다. 따라서 전문분야문서에 대한 자연언어처리에서 전문용어를 효과적으로 처리하는 것은 매우 중요하다.

전문용어 추출은 분야 특성적인 전문용어를 해당 분야 문서에서 파악하는 작업을 말한다. 지금까지 전문용어 추출에 대한 많은 연구가 있어 왔다[2,3,4]. 이들 연구들은 크게 두 단계를 거쳐 전문용어를 추출하였다. 첫 번째 단계는 문서에서 전문용어가 될 수 있는 전문용어후보를 추출하는 '언어적 필터링' 단계이다. 언어적 필터링 단계에서는 부분구문정보를 이용하여 전문용어후보를 추출한다. 대부분의 기존연구에서는 전문용어를 두 단어 이상으로 구성된 명사구로 한정하고, 전문분야 문서에서 나타나는 명사구를 전문용어후보로 추출하였다. 전문용어 추출의 두 번째 단계는 전문용어후보 중 전문용어를 추출하는 '통계적 필터링' 단계이다. 언어적 필터링이 부분구문정보와 같은 언어적 정보를 사용하는 것과는 달리 통계적 필터링 단계에서는 전문용어후보의 문서내 통계적 특성을 사용하여 필터링을 수행한다. 기존 연구에서의 통계적 필터링은 전문용어후보에 대한 순위화 작업으로 정의된다. 전문용어의 일반적인 특성을 점수함수로 모델링하고, 점수함수로 전문용어후보들에 점수를 부여하였다. 기존의 연구들은 전문용어 후보의 빈도수, 전문용어후보 간의 내포관계, 전문용어후보의 문맥정보 등을 이용하여 점수함수를 모델링 하였다.

하지만 기존의 방법들은 크게 네 가지 한계점을 가진다. 첫째, 기존의 방법은 두 단어 이상으로 구성된 전문용어만을 추출 대상으로 한다. 대부분의 전문용어가 이러한 형태로 나타나지만, 하나의 단어로 구성된 전문용어도 많다. 특히, 의학분야의 경우 하나의 단어로 구성된 전문용어도 많은 비중을 차지한다. 따라서 하나의 단어로 구성된 전문용어도 추출 대상이 되어야 한다. 둘째, 기존의 방법은 전문용어 추출 결과를 점수함수에 의해 순위화된 전문용어후보로 나타낸다. 이는 순위화된 전문용어후보 중 전문용어와 비전문용어의 구분을 전문가의 개입에 의해 수행해야 함을 의미한다. 물론 순위화된 전문용어추출 결과에서 전문용어를 특정 점수 한계값 이상의 전문용어 후보로 한정할 수도 있다. 하지만 문서집합별, 전문분야별로 점수의 분포가 다양하게 나타나기 때문에 점수의 한계값에 대한 일반적인 적용이 어렵다. 또한, 점수의 한계값 설정에도 전문가의 개입이 필요하다. 이러한 이유로 기존의 전문용어 추출 기법은 반자동적인 전문용어추출 기법이라는 한계가 있다. 셋째, 대부분의 기존 연구들은 해당분야에 이미 존재하는 전문용어 정보를 사용하지 않고 전문용어를 추출한다. 새로운 전문용어는 기존의 전문용어를 기반으로 생성되는 경우가 많기 때문에 기존의 전문용어의 정보는 전문용어 추출에 매우 중요하다[1]. 전문분야 사전은 기존의 전문용어 정보를 제공하는 자원으로 사용할 수 있다. 전문용어추출에서 기계가독형 사전이 사용되기 어려웠던 것은 사전을 구축하는 데 있어 상당한 노력이 필요했기 때문이다. 하지만 기계가독형 언어자원을 구축하기 위한 도구들의 점진적인 개발은 전문용어추출 분야에 전문분야 사전이 점차적으로 사용하는 계기를 마련하였다. 넷째, 기존의 연구들은 일반적인 전문용어의 특성에 기반한 점수함수로 전문용어를 추출한다. 따라서 분야에 특성적인 전문용어의 정보를 효과적으로 파악하는데 그 한계가 있다. 즉, 전문분야마다 또는 전문분야 문서마다 전문용어가 나타나는 양상이 다르기 때문에, 전문용어를 효과적으로 추출하기 위해서는 문서에서 나타나는 전문용어의 특성을 파악하는 것이 중요하다.

본 논문에서는 이러한 문제점들을 해결하기 위하여 기계학습방법을 이용한 전문용어 자동 추출 기법을 제안한다. 본 논문의 기법은 두 단어 이상의 전문용어 뿐만 아니라 한단어로 구성된 전문용어를 추출 대상으로 한다. 또한, 전문분야 사전과 전문분야 문서를 이용하여 문서에서 나타나는 전문용어의 특성을 파악하고, 학습을 통하여 전문용어를 추출한다. 본 논문에서는 해당 분야의 전문용어가 같은 분야의 문서에서 비슷한 문맥을 가지고 나타난다는 사실에 기반하여 대상함수를 설계하여 전문용어를 추출한다. 예를 들어, 의학분야 문서에서 전문용어는 다음과 같은 패턴을 가지고 나타난다. - '<A> affect <B>' - [***Bcl-2 gene*** affects ***glioma cell viability***], [***TNF-alpha*** affects ***STAT1 activation***], [***A molecular component*** affects ***methylation patterns***]. 이 예제에서 'affect'의 주어와 목적어는 전문용어이며, 'affect'를 술어로서 공유한다. 본 논문에서는 이러한 패턴을 문맥패턴이라고 정의하고, 이를 이용하여 전문용어를 추출한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 기법에 대하여 설명한다. 3장에서는 실험과 실험결과에 대하여 기술하고 4장에서는 결론을 맺는다

2. EM 알고리즘을 통한 전문용어 추출 기법

2.1 시스템 구조

본 논문의 기법은 언어적 필터링 과정을 포함하는 전처리 단계와 통계적 필터링 과정을 포함하는 학습단계로 구성된다. 전처리 단계에서는 문서에 대한 구문분석을 수행하고 구문분석 결과에서 명사구를 전문용어후보로 인식한다. 전문용어후보에 대하여 정의된 구문관계를 추출한다. 전문용어의 문맥패턴을 추출하기 위하여 7가지 구문관계[1]를 정의하였다. 전처리 과정의 결과는 학습단계에서 학습을 위한 요소로 사용된다. 학습단계에서는 전처리 단계에서 추출된 전문용어후보와 전문용어후보의 구문관계를 기반으로 전문용어추출을 위한 파라메터를 학습한다. 학습된 파라메터를 이용하여 전문용어를 추출한다. 그림 1은 시스템의 전체구조를 나타낸다.

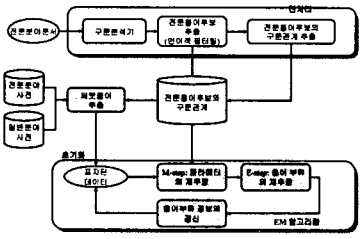

그림 1. 전문용어 추출 시스템 구조

2.2 EM 알고리즘을 이용한 파라메터의 학습

본 논문에서는 통계적 필터링 과정을 전문용어후보의 분류문제로 변환하여 해결한다. 즉, 주어진 전문용어 후보 t_i에 대하여 전문용어후보의 부류 c_j로 분류하는 문제로 변환할 수 있다. 이를 Naive Bayesian 분류기[5]를 사용하여 모델링하면 식 (1)과 같이 나타내어진다.

$$c^* = \arg\max_{c_j} p(c_j \mid t_i) = \arg\max_{c_j} \frac{p(c_j)p(t_i \mid c_j)}{p(t_i)} \quad (1)$$

전문용어 후보 t_i는 t_i의 구문관계로 $t_i=<sr_{ti,1}, sr_{ti,2}, \ldots, sr_{ti,n}>$와 같이 표현될 수 있다. 여기에서 t_i의 n번째 구문관계를 $sr_{ti,n}$로 나타낸다. $sr_{ti,n}$는 $sr_{ti,n}$에 참여하는 단어 $w_{ti,n}$와 $w'_{ti,n}$ 그리고 두 단어사이의 구문관계 $r_{ti,n}$에 의해 $sr_{ti,n}=<w_{ti,m}, r_{ti,m}, w'_{ti,n}>$로 표현된다. 이러한 전문용어 후보의 표현방법에 의하여, $p(t_i|c_j)$는 식 (2)와 같이 전개된다. 식 (2)의 간단화를 위하여 구문관계 sr은 용어부류 c_j가 주어졌을 때 조건부 독립임을 가정한다[5]. 식 (2)는 조건부독립 가정에 의해 식 (3)과 같이 $p(sr_{ti,k}|c_j)$의 곱의 연쇄로 나타낼 수 있다. 따라서 학습과정에서 $p(t_i|c_j)$의 학습은 $p(sr_{ti,k}|c_j)$의 학습을 의미한다. 본 논문에서는 EM알고리즘을 이용하여 $p(sr_{ti,k}|c_j)$와 $p(c_j)$를 학습하여 전문용어후보를 분류한다.

$$P(t_i \mid c_j) = P(< sr_{ti,1}, sr_{ti,2}, \ldots, sr_{tin} > \mid c_j) \quad (2)$$

$$P(< sr_{ti,1}, sr_{ti,2}, \ldots, sr_{tin} > \mid c_j) = \prod_{k=1}^{n} P(sr_{ti,k} \mid c_j) \quad (3)$$

$$\hat{\theta}^{k+1}_{sr_{ti,k}|c_j} \equiv P(sr_{ti,k} \mid c_j; \hat{z}^{k+1}) \quad (4)$$

$$= P(< w_{ti,k}, r_{ti,k}, w'_{ti,k} > \mid c_j; \hat{z}^{k+1})$$

$$\hat{\theta}^{k+1}_{c_j} \equiv P(c_j \mid \hat{z}^{k+1}) = \frac{1+\sum_{i=1}^{|T|} p(t_i \mid c_j; \hat{z}^{k+1})}{|C|+|T|} \quad (5)$$

$$P(< w_{ti,k}, r_{ti,k}, w'_{ti,k} > \mid c_j; \hat{z}^{k+1}) \quad (6)$$

$$\approx P_{MLE}(r_{ti,k} \mid c_j; \hat{z}^{k+1}) \times P_{MLE}(w_{ti,k} \mid r_{ti,k}, c_j; \hat{z}^{k+1})$$

$$\times P_{MLE}(w'_{ti,k} \mid r_{ti,k}, c_j; \hat{z}^{k+1})$$

$$c^* = \arg\max_{c_j} \frac{\theta_{c_j} \times \prod_{k=1} \theta_{sr_{ti,k}|c_j}}{\sum_i \left(\theta_{c_i} \times \prod_{k=1} \theta_{sr_{ti,k}|c_i} \right)} \quad (7)$$

$p(sr_{ti,k}|c_j)$와 $p(c_j)$를 효과적으로 학습하기 위해서는 대량의 정답이 표지된 학습데이터가 필요하다. 하지만 전문용어추출을 위한 정답이 표지된 학습데이터가 부족하기 때문에, 본 논문에서는 정답이 표지되지 않은 전문분야 코퍼스와 사전만을 이용하여 학습을 수행한다. 전문분야 코퍼스에서 전문용어 사전에 등재된 용어가 출현할 경우, 전문용어로 판단하고 일반분야 사전에 등재된 용어가 출현할 경우 비전문용어로 판별한다. 사전에 의해 판별된 전문용어와 비전문용어 정보를 씨앗 정보로 분류대상함수의 초기 파라메터를 학습한다. 씨앗정보는 전문분야 코퍼스 상에 나타나는 모든 전문용어와 비전문용어를 고려한 것이 아닌 사전에 등재된 용어만을 관찰 대상으로 했기 때문에 불완전한 관찰 데이터라고 할 수 있다. 학습은 불완전한 관찰 데이터로부터 최우도추정값을 찾는 효과적인 알고리즘인 EM알고리즘에 의해 수행된다. EM알고리즘을 이용한 파라메터 학습[5,6,7]은 다음과 같이 표현된다. z를 용어부류가 표지된 전문용어 후보의 집합이라 정의하면, E-step과 M-step의 과정을 통하여 파라메터가 학습된다.

E-step : k단계의 M-step에서 학습한 파라메터 θ^k를 이용하여 z^{k+1}를 추정한다.

Set $\hat{z}^{(k+1)} = \arg\max_z p(z \mid \hat{\theta}^{(k)})$

M-step : $k+1$단계의 E-step에서 추정한 z^{k+1}를 이용하여 파라메터 θ^{k+1} 를 학습한다

Set $\hat{\theta}^{(k+1)} = \arg\max_\theta p(\theta \mid \hat{z}^{(k+1)})$

EM알고리즘 학습과정에서 학습되는 파라메터 θ는 $\theta=\{\theta_{srti|cj}=p(sr_{ti,k}|c_j), \theta_{cj}=p(c_j)\}$와 같이 표현된다. 여기에서 학습에 의해 추정된 파라메터, θ는 $\hat{\theta}$와 같이 표현된다. 각 파라메터 θ는 식 (4), (5)와 같이 추정된다. 식 (4)에서 수식의 간단화를 위하여 $r_{ti,k}$가 주어졌을 때, $w_{ti,k}$와 $w'_{ti,k}$는 조건부 독립임을 가정한다[8]. 조건부 독립가정에 의해 $p(<w_{ti,k}, r_{ti,k}, w'_{ti,k}>|c_j)$는 식 (6)과 같이 전개된다. 식 (6)에서 P_{MLE}는 최우도추정법에 추정되는 확률값이다. 식 (1), (4), (5)에 의하여 전문용어후보를 분류하며, 식 (1)은 EM 알고리즘에 의해 학습된 θ에 의해 식 (7)과 같이 표현된다.

3. 실험

3.1 실험데이터

실험에 사용한 문서는 생물의학분야 영어 논문 초록을 포함하는 MEDLINE을 사용하였다. MEDLINE은 4,600개의 생물의학분야 논문지의 논문초록으로 구성되어 있다. 본 논문에서는 이들 중 분자 생물학분야로 실험의 범위를 한정하여 3,000문장 70,000단어 수준의 300개 문서를 대상으로 실험을 수행하였다. 전문용어 사전은 의학분야 전문용어사전 'UMLS Specialist lexicon'을 사용하였으며, 일반분야 사전으로는 Brill 태거의 명사사전을 사용하였다.

실험은 EM알고리즘을 이용한 학습단계별 평가를 위한 실험과 기존연구와의 비교 실험을 수행하였다. 학습단계별 평가에서는 학습을 수행하면서 전문용어 분류의 성능 평가를 위하여 식 (15)의 정확률, 커버율, F1값을 사용한다[9]. 정확률은 분류된 전문용어후보 중 올바르게 분류된 전문용어후보의 수의 비율을 의미하고, 커버율은 전체 전문용어후보 중 분류된 전문용어후보의 수의 비율을 나타낸다. F1값은 정확률과 커버율을 통합적으로 나타내는 평가 기준이다.

[1] 1) head 2) adj-mod-head 3) verb-subj- head 4) verb-obj-head 5) head-of-head 6) head- prep-head 7) verb-prep-head. 'I have a blue paencil'은 다음과 같은 세가지 문맥패턴을 포함한다. <have subj I>, <have obj pencil>, <blue mod pencil>.

$$정확률 = \frac{올바르게분류된전문용어후보의수}{분류된전문용어후보의수}$$

$$커버율 = \frac{분류된전문용어후보의수}{전체전문용어후보의수}$$

$$F1값 = \frac{2\times정확률\times(정확률\times커버율)}{정확률+(정확률\times커버율)} \quad (8)$$

기존연구와의 비교실험에서는 제안하는 전문용어추출 기법과 [2,3]의 연구와의 비교평가를 수행한다. 기존 연구의 전문용어추출 결과가 전문용어후보에 대한 용어부류를 할당한 것이 아니라 점수함수에 의해 순위화된 전문용어후보를 나타내므로 직접적인 비교가 어렵다. 제안하는 기법과의 비교를 위하여, 본 논문에서는 제안하는 전문용어추출결과를 전문용어후보가 전문용어로 분류될 확률값으로 순위화하여 평가한다. 기존방법과의 성능을 비교하기 위하여 정보검색분야에서 사용되는 11-포인트 평균 정확률 (11-point average preceision)을 사용하였다[9]. 11-포인트 평균 정확률은 재현율이 0%, 10%, 20%, 30%, ..., 90%, 100% 지점일 때의 정확률을 계산한 뒤, 11개 지점의 정확률을 합하고 이를 평균하여 나타내어진다. 따라서, 각 재현율의 지점에서 높은 정확률을 보일 경우 11-포인트 평균 정확률이 높게 나타난다. 이는 상위에 적합한 용어가 많이 존재할수록 높은 11-포인트 평균 정확률을 얻을 수 있음을 의미한다. 본 논문에서는 11-포인트 평균 정확률을 구하기 위하여, 전문용어 후보 9,600여개 중 전문용어로 판별되는 6,930개의 후보를 모두 추출하였을 때 재현율이 100%라고 가정한다. 그리고 재현율이 0% 지점에서 100%의 정확률을 가진다고 가정한다[9]. 이를 기준으로 재현율 0%~100% 지점을 찾아 해당 지점에서의 정확률을 계산한다.

3.2 실험결과

3.2.1 전문용어 추출 실험 결과

표 1. 전문용어 추출 실험 결과

학습단계	정확률	커버율	F1값
초기	89.16%	23.25%	33.64%
40	77.66%	95.11%	75.71%
100	77.58%	97.96%	76.78%
최종	77.49%	100.00%	77.49%

표 1은 학습단계별 전문용어 추출 실험결과를 나타낸다. 표에서 '초기' 학습단계에서는 사전정보만을 이용하여 전문용어를 추출한 결과를 나타내며, 높은 정확률과 낮은 커버율을 나타낸다. 이는 사전만으로 파악할 수 있는 전문용어와 비전문용어가 한정되어 있음을 나타낸다. 본 논문의 기법은 학습을 통하여 사전에 등재되어 있지 않는 전문용어를 정확히 찾아내는 것을 목표로 한다. 따라서 학습과정에서 정확률의 감소를 최소화하면서 커버율을 높이는데 목적이 있다. 실험결과는 15.5%의 정확률 감소로 76%의 커버율을 높여 본 논문의 기법이 효과적으로 전문용어를 추출함을 알 수 있다. 본 논문의 기법은 9,600개의 전문용어후보를 추출하였으며, 이중 약 77%의 전문용어후보에 대한 용어부류를 올바르게 할당함을 알 수 있다.

3.2.2 기존 연구와의 비교 실험

표 2. 기존 연구와의 비교 실험 결과

	[2]	[3]	제안방법	이상적 시스템
11pt-avg	77.61%	83.49%	89.13%	100.00%

표 2는 기존의 기법과 본 논문의 기법을 비교 평가한 실험 결과이다. 각 시스템의 전체적인 성능을 비교 평가하기 위하여 정보검색 분야에서 사용되는 11포인트 평균 정확률로 평가하였다. 표 2에서 이상적인 시스템은 100%의 11포인트 평균 정확률을 나타냄을 알 수 있다. 따라서 11포인트 평균 정확률이 100%에 가까울수록 전문용어를 효과적으로 추출함을 알 수 있다. 본 논문의 기법은 Justeson 등[2]의 기법보다 14.84%의 성능향상을 보이며, Frantzi 등[3]의 기법보다 6.75%의 성능향상을 나타낸다.

3.3 오류분석

오류 중 많은 부분은 약어와 사전에 의한 오류에 의해서 나타났다. 전문분야 문서에서 약어는 주로 전문용어이며, 약어와 원형태가 혼용되어 사용된다. 그런데, 원형태가 전문용어로 판별되더라도 약어가 전문용어로 판별되지 않는 경우가 많았다. 이는 약어의 문맥정보가 부족하여 올바른 판단이 어렵기 때문으로 분석된다. 사전에 의한 오류는 대부분 1어절로 구성된 용어에서 나타났다. 전문용어 중 일반분야 사전에만 등재된 경우 학습과정의 노이즈로 작용하여 오류를 야기하였다. 예를 들어, 'scavenger', 'tranduction'은 일반분야에서 각각 '청소도구', '변환'이라는 의미를 가지지만, 의학분야에서는 각각 '헤모글로빈', '형질도입'이라는 의미를 가지는 전문용어이다. 향후 이들 오류에 대한 보완이 필요할 것으로 생각된다.

4. 결론

본 논문에서는 EM알고리즘을 이용한 전문용어추출기법을 제안하였다. 본 논문에서는 전문용어추출에서 통계적 필터링 문제를 전문용어후보에 대한 분류문제로 변환하여 전문용어를 추출하였다. 본 논문의 기법은 전문분야 사전과 일반분야사전의 표제어를 씨앗정보로 사용하여 전문용어추출을 위한 파라메터를 학습하였다. 실험결과 본 논문의 기법은 약 77%의 정확률로 전문용어를 추출하였으며, 기존의 연구보다 우수한 성능을 나타냄을 알 수 있었다.

향후연구로 전문용어추출 기법을 향상시키는 방법에 대한 연구를 수행할 것이다. 성능향상 기법에는 약어 등을 자동으로 추출하는 연구와 의미정보를 이용한 구문관계의 일반화에 대한 연구가 있다. 또한 전문용어를 효과적으로 추출하기 위하여, 형태소 변이에 의한 용어의 변이에 대한 연구도 추가로 수행되어야 할 것이다.

감사의 글

본 연구는 한국과학기술기획평가원 국책연구개발사업 (M1-0107-00-0018)과 한국과학재단 특성장려연구사업 (R21-2003-000-10042-0)의 지원으로 수행되었습니다

참고문헌

[1] Sager, J.C. (1997), "Section 1.2.1 Term formation", in Handbook of terminology management Vol.1, John Benjamins publishing company

[2] Justeson, J.S. and S.M. Katz (1995) Technical terminology : some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(1) pp. 9-27

[3] Frantzi, K.T. and S.Ananiadou (1999) The C-value/NC-value domain independent method for multi-word term extraction. Journal of Natural Language Processing, 6(3) pp. 145-180

[4] Maynard D. and Sophia Ananiadou. (2000) TRUCKS: a model for automatic term recognition, Journal of Natural Language Processing, December

[5] Mitchell T. M., (1997) Machine learning, *New-York: McGraw-Hill*

[6] Nigam K., Andrew McCallum, Sebastian Thrun and Tom Mitchell. (2000). Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning, 39(2/3). pp. 103-134.

[7] Jones Rosie, Andrew McCallum, Kamal Nigam, Ellen Riloff, (1999). Bootstrapping for Text learning Tasks. IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications

[8] Lin D., (1998). Automatic Retrieval and Clustering of Similar Words. *COLING-ACL98*, Montreal, Canada.

[9] Ricardo B-Y. and Berthier R-N., "Mordern Information Retrieval", ACM-Press New York and Addison-Wesley, 1999