http://dx.doi.org/10.3745/KIPSTD.2011.18D.5.329

과학 기술 문헌 분석을 위한 기계학습 기반 범용 전문용어 인식 시스템

최 윤 수[†] · 송 사 광[††] · 전 홍 우[†] · 정 창 후[†] · 최 성 필[†]

요 약

문헌에서의 전문용어 인식 연구는 정보검색, 정보추출, 시맨틱 웹, 질의응답 분야 등의 연구를 위한 선행 연구로서, 지금까지 대부분 특정 분야, 특히 생의학 분야에서 집중되어 연구되어 왔다. 그러나 기존 연구들이 특정 도메인 또는 문헌 내부 통계 정보를 활용함으로써 범용적인 전문용어 인식에 한계점을 보여 왔기 때문에, 본 연구에서는 웹 검색 결과와 사전, 후보용어의 문형 특징 등을 활용하는 기계 학습 기반 범용 전문용어 인식 방법을 제안하였다. 제안한 방법을 문헌의 지역 통계 정보를 사용하는 방법(C-value)과 비교 실험하여 80.8%의 F-값으로 6.5%의 성능향상을 보였다. 다양한 응집도 자질들을 접목한 두 번째 실험에서는 Normalized Google Distance 방법과 접목한 방식이 F-값 81.8%의 성능으로 최고의 성능을 나타냈다. 기계 학습 방법으로는 로지스틱 회귀분석, C4.5, SVMs 등을 적용하였는데, 일반적으로 이진 분류에 좋은 성능을 보이는 SVMs과 로지스틱 회귀분석 방법보다 결정 트리 방식의 C4.5가 전반적으로 좋은 성능을 보였다.

키워드 : 전문용어 인식, 텍스트 마이닝, 기계 학습, 정보 추출

Terminology Recognition System based on Machine Learning for Scientific Document Analysis

Yun-Soo Choi[†] · Sa-Kwang Song[††] · Hong-Woo Chun[†] · Chang-Hoo Jeong[†] · Sung-Pil Choi[†]

ABSTRACT

Terminology recognition system which is a preceding research for text mining, information extraction, information retrieval, semantic web, and question-answering has been intensively studied in limited range of domains, especially in bio-medical domain. We propose a domain independent terminology recognition system based on machine learning method using dictionary, syntactic features, and Web search results, since the previous works revealed limitation on applying their approaches to general domain because their resources were domain specific. We achieved F-score 80.8 and 6.5% improvement after comparing the proposed approach with the related approach, C-value, which has been widely used and is based on local domain frequencies. In the second experiment with various combinations of unithood features, the method combined with NGD(Normalized Google Distance) showed the best performance of 81.8 on F-score. We applied three machine learning methods such as Logistic regression, C4.5, and SVMs, and got the best score from the decision tree method, C4.5.

Keywords : Terminology Recognition, Text Mining, Machine Learning, Information Extraction

1. 서 론

텍스트 마이닝은 비정형 문헌으로부터 잠재적인 유용한 패턴을 발견하는 과정으로, 비정형 과학 기술 문헌의 증가에 따라 점점 중요한 분야로 인식되고 있다. 특히 특허 문헌은 정부기관, 산업계 등의 정책 결정자들을 위한 가치 있는 연구 결과들을 포함하는 중요한 문헌으로, 전 세계적으로 매년 200여만 건의 새로운 특허가 출원되고 있으며, 이들의 70~80%는 공개되지 않은 신규 기술 정보이다. 특허 문헌은 문서의 길이가 길고, 많은 전문용어를 포함하고 있어, 분석을 위해 상당한 노력과 전문성을 요구하므로, 특허문헌을 자동으로 분석하기 위한 텍스트 마이닝 기술이 크게 요구되고 있다[18].

이러한 과학 기술 문헌에 대한 전문용어 인식 기술은 텍스트 마이닝의 기반이 되는 연구로서 특히 생의학 분야에서 많은 연구가 수행되어 왔다[9][11][13]. 즉, 지금까지 전문용어 자동인식에 대한 연구는 고려되고 있는 특정 분야에 대한 문서 집합들을 대상으로 문서내의 통계(지역통계)정보를

† 정 회 원 : 한국과학기술정보연구원 선임연구원
†† 정 회 원 : 한국과학기술정보연구원 선임연구원(교신저자)
논문접수 : 2011년 6월 27일
수 정 일 : 1차 2011년 7월 26일, 2차 2011년 8월 12일
심사완료 : 2011년 8월 17일

기반으로 한 기법들이 주로 연구되었고, 기계학습 기법을 이용하여 통계적 기법들을 보완하는 방법들이 제안되어 왔다[14][16]. 일반적으로 전문용어 자동 인식 과정은 과학 기술 문헌에 대한 언어학적 구문분석을 통하여 전문용어 후보 집합들을 추출하는 작업과, 추출된 용어 후보 집합들에 대하여 다양한 통계적 방법을 이용하여 각 후보 용어들에 전문성을 부여하는 작업으로 구분된다.

대부분의 통계적 방법들은 문헌 내의 후보용어들의 출현빈도에 기반 하여 후보 용어의 전문성을 측정하므로 다음과 같은 문제점을 가지고 있다. 첫째, 수집된 대상 문서집합의 규모 및 특성에 의해 전문용어 인식 성능이 영향을 받는다. 예를 들어, Joachim Wermeter (2005)[10]의 실험 문서 집합에서 비전문 용어인 "t cell response"가 2,410회, 전문용어인 "long terminal repeat"은 434회 출현하였는데, 출현빈도가 더 높은 "t cell response"가 전문용어인 "long terminal repeat"에 비해 전문용어로 인식될 가능성이 높다. 둘째, 특정 분야의 전문용어 인식에 초점이 맞추어져 있어, 다양한 분야를 포함하는 과학 기술 문헌을 처리하기 어렵다.

따라서 본 논문에서는 대상 문서집합의 규모와 분야에 영향을 받지 않고, 특허, 논문, 보고서 등의 다양한 과학 기술 문헌으로부터 전문용어를 인식하기 위하여, 전문용어 사전과 웹 검색 결과를 기반으로 특정 분야에 종속적이지 않은 전역 통계 정보를 활용한 전문용어 후보에 대한 가중치를 할당하는 범용 전문용어 인식 방법을 제안하고, 대상 문헌집합에서 추출한 지역 통계 정보를 사용하는 방법과 비교한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로써 다양한 통계적 전문용어 인식 방법에 대하여 기술하고, 3장에서는 본 연구에서 제안하는 구체적인 전문용어 인식 시스템에 대해 상세히 설명한다. 4장에서는 실험방법 및 실험결과에 대하여 기술하고 5장에서는 결론을 맺는다.

2. 관련 연구

전문용어 자동인식에 관한 연구는 후보용어들을 추출하는 '후보 용어 추출(candidate term extraction)' 작업과 '용어 가중치 할당(termhood assignment)' 작업으로 구분된다. 일부 연구의 경우 후보용어 추출하는 과정을 거치지 않고 기계학습 기법을 활용해 전문용어를 추출하려는 연구[6,8]가 있지만, 대부분의 경우 후보추출 단계와 추출된 후보 중 전문용어 가중치를 할당하는 단계를 분리하여 성능을 향상 시키려 하였다.

먼저, 후보용어 리스트는 일반적으로 문장 내의 품사 패턴을 이용하여 추출하는데, <표 1>은 이전 연구들에서 용어 후보 추출작업에 사용된 후보용어 추출패턴들을 나타낸다. Daille et. al.(1994)[1]와 Dagan and Church(1994)[7]는 매우 제한적인 후보용어들을 추출하는 필터를 사용하였는데, 이 필터는 용어가 될 확률이 높은 패턴들만을 후보로 추출하므로 시스템의 전체 정확률(precision)을 증가시키지만 다양한 용어 형태들을 포괄하지 못하기 때문에 재현율(recall)을 감소시키는 경향이 있다. Justeson and Katz(1995)[9]는 다양한 패턴들을 수용하는 개방된 필터를 제안하였고, 이 필터는 앞의 필터와는 반대로 재현율을 증가시키지만 정확률을 감소시킨다. Franzi and Anaiadou(1997)[11]은 이 두 필터를 조합하여 <표 1>의 마지막 행에 제시된 것처럼, 정확률과 재현율을 조화시키기 위한 새로운 필터를 제안하였다.

다음으로, 위와 같은 후보 용어 추출 패턴을 활용해 추출된 각 후보 용어에 용어 가중치를 할당하는 통계기반 및 기계학습 기반 관련연구에 대해 설명한다.

• 통계기반 관련연구

- Mutual Information

Church and Hanks(1990)은 두 단어 간의 연관성을 계산하기 위하여, MI(Mutual Information)을 사용하였다[2].

$$MI(a,b) = \log_2 \frac{P(x,y)}{P(x)P(y)} \quad (1)$$

여기에서, x와 y는 용어후보를 구성하는 단어이고, $P(x)$는 단어 x의 출현확률, $P(y)$는 단어 y의 출현확률, $P(x,y)$는 단어 x와 y가 함께 출현하는 확률을 말한다.

- Log Likelihood-Ratio

Dunning T.(1993)은 소규모 문헌 집합 내에서 낮은 빈도로 출현하는 용어들에 대해 적합한 가중치를 할당하기 위하여 Likelihood-Ratio를 사용하였다[4].

〈표 1〉 후보용어 추출 패턴 예

저자(년도)	후보용어 추출 패턴	비고
Daille et. al. (1994)	Adj N N1 N2	closed filter
Dagan and Church (1994)	N+	closed filter
Justeson and Katz (1995)	((Adj|N)+|(((Adj|N)*(NPrep)?)(Adj|N)*))N	open filter
Frantzi, Ananiadou and H. Mima (2000)	N+N (Adj|Noun)+N ((Adj|N)+|((Adj|N)* (NounPrep)?)(Adj|N)*)N	명사구+전치사구를 수용

$$\begin{aligned}\lambda(x,y) = {} & a\log(a) + b\log(b) + c\log(c) + d\log(d) \\ & - (a+b)\log(a+b) - (a+c)\log(a+c) \\ & - (b+d)\log(b+d) - (c+d)\log(c+d) \\ & + (a+b+c+d)\log(a+b+c+d)\end{aligned} \tag{2}$$

여기에서, a는 x와 y가 함께 출현하는 경우, b는 x만 출현한 경우, c는 y만 출현한 경우, d는 x와 y가 둘 다 출현하지 않는 경우를 말한다.

- Dice Coefficient

F. Smadja et. al.(1996)은 병렬 코퍼스 내에서 용어들의 응집도(Unithood)를 계산하기 위하여 Dice Coefficient를 사용하였다[5].

$$Dice(x,y) = \frac{2\sum_i (x_i \bullet y_i)}{\sum_i x_i + \sum_i y_i} \tag{3}$$

여기에서 x와 y는 각각 문헌 내의 용어이고, $\sum_i x_i$는 문헌 내의 x의 출현빈도, $\sum_i y_i$는 y의 출현빈도를, $\sum_i x_i \bullet y_i$는 두 용어가 함께 출현하는 빈도수를 나타낸다.

- Normalized Google Distance

NGD(Normalized Google Distance)는 웹 검색건수를 기반으로 두 단어에 대한 응집도를 계산하는 방식으로 R. Cilibrasi and P. Vitanyi (2007)에 의해 소개되었다[15]. 두 단어의 독립적인 출현빈도와 동시 출현빈도를 이용하여 응집도를 부여한다.

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}} \tag{4}$$

여기에서, M은 단어 x와 y를 포함하는 전체 건수, $f(x)$와 $f(y)$는 검색 단어 x와 y의 검색 건수, $f(x,y)$는 x와 y를 모두 포함하는 검색 건수를 말한다. $f(x), f(y) > 0$ 이고 $f(x,y) = 0$ 인 경우, $NGD(x,y) = \infty$ 가 되어, 두 단어의 응집도가 전혀 없음을 보여주고, $f(x) = f(y) = f(x,y)$ 인 경우, $NGD(x,y) = 0$ 이 되어, 두 단어가 항상 함께 나타나는 최고 응집도를 보여 준다.

- C-value

Frantzi and Ananiadou (2000)은 출현빈도기반의 가중치 부여 방식에 대한 문제점을 개선하기 위해, 다어절 용어 후보에서 내포된 부분 문자열의 전문성을 측정하기 위해 C-value를 제안하였다[11].

$$C\text{-}value(a) = \begin{cases} \log_2|a| \bullet f(a) \\ \quad ;a\text{가 내포되는 용어가 없는 경우} \\ \log_2|a| \bullet \left(f(a) - \dfrac{1}{P(T_a)} \sum_{b \in T_a} f(b)\right) \\ \quad ;\text{그렇지 않은 경우} \end{cases} \tag{5}$$

여기에서 a는 후보 문자열, $f(a)$는 문서집합 내에서 a의 출현빈도, $|a|$ 는 문자열의 길이(공백으로 구분되는 단어수), $f(b)$는 a를 내포하는 후보 문자열의 출현빈도, T_a는 a를 포함하는 용어집합, $P(T_a)$는 T_a에 포함되는 용어의 수이다. 예를 들어, "soft contact lens"라는 실제 안과학의 전문용어에 대해 살펴보면, "soft contact" 과 "contact lens "의 두 부분문자열도 전문용어의 후보로 추출되는데, 이때 "contact lens"는 단독으로 자주 출현하거나 다른 용어들에 내포되어 출현하는 횟수가 많기 때문에 높은 가중치를 할당받는 반면, "soft contact"은 "soft contact lens" 외의 다른 용어에 내포되어 출현하는 횟수가 거의 없기 때문에 낮은 가중치를 할당받는다.

• 기계학습 기반 관련연구

- 로지스틱 회귀 분석(Logistic Regression Analysis)

로지스틱 회귀 분석은 독립적인 여러 자질들을 입력으로 사용하여 이항확률을 계산하기 위해 사용되는 일반화된 선형모델로서, Qing T. Zeng, Tony Tse, et. al. (2007)에 의해 의학 분야 용어추출을 위해 사용되었다[16].

$$f(z) = \frac{1}{1+e^{-z}},$$
$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k \tag{6}$$

식(6)에서 x_i는 자질들의 값이고, β_i는 학습 집합에서 학습 과정을 통하여 습득한 계수(Regression Coefficient)이다.

- C4.5

C4.5는 Ross Quinlan에 의해 개발된 결정트리(Decision tree)를 생성하기 위한 알고리즘으로, 주어진 자질들을 이용하여 데이터를 분류하는데 사용하기 위한 기계학습 방법[14]이다. 학습 집합 $S = s_1, s_2, \ldots s_n$ 인 경우, 학습 집합 내의 표본 s_i는 $s_i = x_1, x_2, \ldots x_m$ 로 표현된다. 여기에서 x_i는 표본 s_i의 자질들을 나타내며, C4.5는 학습을 통해 표본 s_i의 여러 자질 중에서 가장 정확히 분류를 수행하는 최적의 자질을 선택하여 이용한다.

- SVMs(Support Vector Machines)

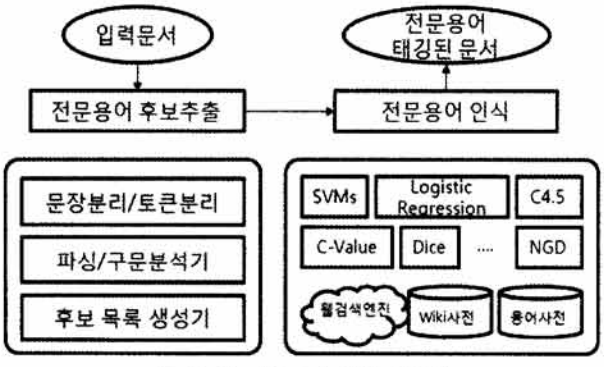
지지벡터기계는 비선형 매핑함수를 이용하여 학습 데이터의 샘플 공간을 선형 초평면(Hyperplane)이 만들어지는 고

차원 특징 공간으로 매핑하고, 인식오류를 최소화하는 최적 초평면을 찾는 기계학습방법으로, 희소자질 집합으로부터 효과적인 학습이 가능한 이진분류를 위해 널리 사용되고 있고, 많은 경우에서 매우 좋은 성능을 나타내고 있다[3].

전문용어인식을 위해 기계학습모델을 사용하기 위해서는 전문용어의 특징을 나타내는 여러 자질들이 필요하다. 생의학분야는 BioLexicon[1], MeSH[2], UMLS[3] 등 관련 용어사전들이 풍부하고 그리스 라틴어원의 용어들을 많이 포함하고 있어 기계학습모델을 적용하기가 비교적 수월하지만, 범용분야의 과학 기술 문헌에 대한 전문용어 인식 시스템에 적용하기에는 역부족이다. 이러한 이유로, 본 논문에서는 다양한 통계적 방법으로부터 계산된 통계 정보를 기계학습을 위한 자질로 사용한다.

3. 전문용어 인식 시스템

기존 연구에서 대상 문헌집합으로부터 용어의 전문성을 측정하는 다양한 통계적 방법들은 특정분야에 대한 전문용어를 인식하는데 초점이 맞추어져 있다. 본 논문에서는 다양한 분야의 과학 기술 문헌에 대한 범용적인 전문용어 인식을 위해, 전문용어 사전정보와 웹 검색 결과를 기반으로 한 통계 정보를 자질로 사용하여 범용적인 전문용어 인식을 수행하기 위한 기계학습 방법을 제안한다.



(그림 1) 전문용어 인식 시스템

(그림 1)은 본 연구에서 제안하는 전문용어 인식 시스템의 구조도로써 크게 두 부분으로 구성되는데, 하나 이상의 문헌을 입력으로 받아 문장 단위로 각 문장에서 전문용어 후보를 추출하는 과정과, 추출된 후보 용어에 가중치를 할당하고, 할당된 가중치를 기준으로 전문용어인지를 판단하는 과정으로 구성된다. 전자인 후보 용어 추출을 위해서, 구문 분석에 기반한 패턴 기반 후보용어 추출 방법을 사용하였다. 본 논문은 단순 명사구 뿐만 아니라 전치사, 접속사도 포함하는 명사구 패턴도 포함하였다. 후자인 전문용어 판단 과정에서 전문용어 후보들에 대한 전문용어 가중치를 부여하기 위하여, 사전 포함 여부, 웹 검색엔진을 통하여 수집된 다양한 전역 통계 정보, 그리고 문헌 집합으로부터 추출된 지역 통계 정보[4]를 자질로 하여, 3가지 기계 학습 방법(지지 벡터 기계(SVMs), 규칙 기반 결정나무(C4.5), 로지스틱 회귀 방법)을 이용하여 후보용어에 대한 전문성 가중치를 계산한다.

〈표 2〉 후보 용어 추출과 가중치 부여 알고리즘

$d \rightarrow s_0, s_1, \ldots, s_m$
for(i=0; i < m; i++) {
  $s_i \rightarrow t_0, t_1, \ldots, t_n$
  for(j = 0; j < n; j++) {
    $f_k \leftarrow t_j$에 대한 자질 추출(3.2절의 문형특징, 사전정보, 구글 정규거리 등)
    $w_j \leftarrow ML(f_0, f_1, \ldots, f_k)$;
  }
}

여기에서
d는 입력 텍스트
s_i는 d로부터 분리된 문장
t_j는 s_i로부터 추출된 용어 후보
f_k는 기계학습을 위한 3.2절의 8가지 자질들
ML은 지지벡터기계, 규칙 기반 결정나무, 로지스틱 회귀
w_j는 용어 후보 t_j에 대해 계산된 전문용어 가중치

<표 2>는 입력 텍스트로부터 전문 용어 후보들을 추출하고, 추출된 후보에 가중치를 할당하는 전체적인 알고리즘이다. 각 단계별로 자세한 내용은 3.1절과 3.2절에서 설명한다.

3.1 전문용어 후보추출

전문용어 후보들을 추출하기 위하여, 입력된 텍스트들을 문장들로 분리하고, 각 문장들을 구성하는 단어들의 품사 정보를 획득하기 위하여 Enju 파서[5]를 이용하여 파싱을 수행한다. 파싱 결과로부터 획득된 품사 정보를 기반으로 <표 3>에 나타난 후보용어 추출 패턴을 적용하여 후보용어들을 추출한다. 기존 연구들에서 사용하였던 언어적 필터에 전치사 또는 접속사를 포함하는 명사구를 포함하여 "B and T cell" 같은 형태도 후보로 추출하도록 하였다.

<표 4>는 한 문장에 대한 파싱 결과 및 후보 용어 추출 예를 보여준다. <표 4>의 '후보용어' 행에서 4개의 후보 용어가 추출되었는데, 1,3,4번째 후보는 0번 품사 패턴에 의해 추출된 단순 명사구이고, 2번째 후보는 1번 품사 패턴에 의해 추출된 전치사구를 포함한 명사구 후보용어이다. 이렇게

1) http://www.ebi.ac.uk/Rebholz-srv/BioLexicon/biolexicon.html
2) http://www.nlm.nih.gov/mesh/meshhome.html
3) http://www.nlm.nih.gov/research/umls/

4) 본 연구는 용어사전과 웹 검색결과에 의존한 일반 도메인에서의 전문용어 추출 방법을 제안하지만, 문헌의 지역 정보를 활용하는 기존 시스템과 비교하기 위하여 문서집합으로부터 추출된 정보를 포함한다.
5) 동경대학교 Tsujii 연구실에서 개발된 파서(http://www-tsujii.is.s.u-tokyo.ac.jp/enju/)

〈표 3〉 후보용어 추출에 사용된 품사 패턴

패턴 ID	후보용어 패턴 (정규 표현식)	비고
0	(JJ+|VB[G|N]+)*(NN[S|P]?+)+	명사구 예) gene expression
1	((JJ+)|(VB[G|N]))*(NN[S|P]?+)+(IN)(PR|$|DT)?(JJ+|VB[G|N])*(NN[S|P]?+)+ (Adj | NN)? NN+(IN)	명사구 + 전치사구 예) expression of polypeptide
2	((JJ+)|(VB[G|N]+))+(CC)((JJ+)|(VB[G|N]+))*(NN[S|P]?+)+	접속사를 포함하는 명사구 예) B and T cell
3	(NN[S|P]?+)+(CC)((JJ+)|(VB[G|N]+))*(NN[S|P]?+)+	

〈표 4〉 후보용어 추출 예

문장: Thermodynamic properties of arsenic sulfides studied by EMF measurements.

파싱결과

일련번호	단어	단어 원형	품사(단어)	품사(단어 원형)
0	Thermodynamic	thermodynamic	JJ	JJ
1	properties	property	NNS	NN
2	of	of	IN	IN
3	arsenic	arsenic	JJ	JJ
4	sulfides	sulfide	NNS	NN
5	studied	study	VBN	VB
6	by	by	IN	IN
7	EMF	emf	NNP	NNP
8	measurements	measurement	NNS	NN

후보용어

일련번호	후보용어	패턴 ID
1	Thermodynamic properties	0
2	Thermodynamic properties of arsenic sulfides	1
3	arsenic sulfides	0
4	EMF measurements	0

추출된 후보용어 리스트들이 전문용어로 태깅될 수 있는지, 그 가치를 평가하기 위해 3.2절에서는 후보용어로부터 추출될 수 있는 다양한 특징 자질을 소개하고 기계 학습 기반의 평가 방법을 설명한다.

3.2 전문용어 인식을 위한 기계학습 방법

본 논문의 전문용어 인식 과정은 추출된 후보 용어에 대한 다양한 통계적 수치들을 수집하고, 수집된 값들을 기계학습 모델의 자질벡터로 사용하여, 후보용어에 대한 가중치를 부여하고 전문용어 해당 여부를 이진 분류하는 기계학습 방법을 사용한다. 기계학습 모델의 훈련을 위해 선택된 자질벡터는 크게 후보 용어 특징 정보, 사전 정보, C-value, 그리고 웹 검색 기반 통계 정보를 사용하였으며 상세한 설명은 아래와 같다.

① 후보용어 문형 특징 (S: Syntactic Feature) (<표 3>참조)
- 후보 추출 패턴 유형: 후보 용어 추출 시 적용된 패턴 ID
- 후보 용어 형태 유형: 복합 명사구, 형용사+명사구 등의 구성 유형 정보
- 후보 용어 단어 수: 후보 용어를 구성하는 단어의 수

② 사전정보 (D: Dictionary) (<표 5>참조)
- 분야별 사전과 통합사전 등재 여부
- 위키 사전 등재 여부는 검색 엔진의 검색결과 중 상위 10내에 해당 후보용어의 위키 페이지 링크 포함여부로 결정[6]

③ 개별 단어 웹 빈도수 (W: Web Frequency)
- 후보 용어에 포함된 각 단어의 웹 검색 및 위키 빈도수 정보를 포함
- 후보 용어에 포함된 단어 중 최대 빈도수, 최소 빈도수, 빈도수 합계, 빈도수 평균, 위키에 포함된 단어 수

④ C-value (C)
기계학습의 자질 중 유일하게 대상 학습 집합으로부터 획득되는 지역 통계 정보 (웹 검색결과를 이용한 전역 통계 정보와 비교하기 위하여 사용됨)

6) 위키 사전은 웹 검색엔진을 이용한 변형/유사어 검색 등을 통한 확장매치를 활용하기 위하여 직접 구축하지 않고 실시간 웹 검색을 이용함

〈표 5〉 과학 기술 전문용어 사전

분야(Domain)	출처	내용	건수
생의학	미국 국립의학 도서관	MeSH Terminology[7]	166,358
통합	KISTI 내부	NDSL 저자 키워드	120,718
통합	KISIT 내부	STEAK[8]	1,302,776
통합	Wikipeda	Wikipedia 표제어	3,664,000
합계	5,253,852		

⑤ 구글 정규거리(F1: Normalized Google Distance)
⑥ 상호정보(F2: Mutual Information)
⑦ 로그 우도 비(F3: Log-likelihood Ratio)
⑧ 다이스 상관계수(F4: Dice Coefficient)

①~③의 세부 자질들은, 일차로 후보용어로부터 추출 가능한 모든 자질들을 추출한 후, Weka[17] 도구의 정보 이득(Information Gain) 기반 자질선택 방법을 이용해 일정 임계값 이상의 자질들만 필터링하여 정리된 것으로, 실험 결과에 영향을 미치지 않는 자질들을 제거하여 성능향상을 꾀했다.

③,⑤,⑥,⑦,⑧ 등의 특징 정보는 웹 통계 정보를 기반으로 하고 있는데, 이들은 Bing[9] 검색을 사용하여 수집된다. 이러한 웹 검색 결과의 활용은, 웹의 특성상 특정 분야에 종속되지 않고, 실시간으로 변화하는 기술의 흐름을 동적으로 반영할 수 있다는 장점이 있어, 대상 문헌의 지역적 통계 정보나 후보어절 주위의 문맥정보 활용 등만으로는 습득해 내기 어려운 정보를 보완할 수 있게 한다. 웹을 통한 정보 추출은 크게 두 가지로 요약된다. 하나는 주어진 질의에 대한 검색결과 문서수이고, 다른 하나는 검색된 N개의 결과 중에 위키 검색결과가 포함되어 있는지에 대한 정보이다. 검색 시 제공되는 질의어는 다어절인 경우를 고려하여 Bing 검색옵션을 사용하여 정확 매칭(Exact Matching)만을 고려하였다. 웹 검색의 단점인 속도 저하 문제는 다중 동시접속 방법을 통한 효율성 제고 방법과 한번 검색된 결과를 재활용하는 캐싱 방법을 사용했으며, 유효기간을 두고 주기적으로 업데이트하여 최신 정보를 유지하도록 하였다.

④는 대상 문헌 집합 내에서 추출 가능한 용어 통계 정보를 활용한 대표적인 용어 전문성 측정 방법이다. ⑤~⑧의 용어 응집도는 두 어절이 얼마나 잦은 빈도로 함께 사용되어지는 지를 표현한 것으로, 다어절 전문용어 추출 과정에서 하나의 전문용어로 간주할 수 있는지 그 정도를 표현하는 방식이다. 예를 들어 A B C D 4개 단어로 구성된 어절에서 중심어(Head word) D를 기준으로 A+B+C+D, B+C+D, C+D 3가지 후보 용어가 추출된 경우, 웹 통계 정보를 기반으로 각 후보용어의 응집도를 계산하여 최고 응집도를 갖는 경우를 전문용어로 추출하게 된다. 이러한 응집도 계산 방법은 '전문용어'와 '수식어+전문용어'의 차이를 구분할 수 있는 중요한 정보를 제공한다. 즉 전문용어를 수식하는 다양한 변형 중에서 핵심적인 전문용어만을 추출하는 효과가 있다.

⑥,⑦,⑧은 ⑤에서 NGD를 구하기 위한 절차와 동일하게 수행하여 웹 기반 통계 정보를 활용하여 해당 용어 가중치를 계산한다. 예를 들어, 단어 A, B가 후보로 추출된 용어를 구성하는 단어인 경우, 단어 A, B가 개별적으로 쓰인 웹 문서 빈도와 단어 A, B가 함께 쓰인 웹 문서 빈도를 기준으로 위 공식에 대입하여 얻어진 수치가 된다.

이와 같은 방식으로 추출된 다양한 자질을 활용해, 각 후보용어가 전문용어인지 아닌지를 분류하는 방법으로 2장에서 설명한 로지스틱 회귀분석(Logistic Regression Analysis), C4.5(Decision Tree), 지지벡터기계(Support Vector Machines, SVMs) 등의 기계학습 기반 이진분류 방법을 적용하였다. 참고로, 본 실험에서 SVMs의 커널함수는 Linear, Polynomial, Radial Basis Function, Sigmoid를 사용하였다. 임의의 두 학습 벡터 xi와 xj에 대한 커널 함수는 다음과 같다.

- Linear: $K(x_i,x_j) = x_i^T x_j$
- Polynomial: $K(x_i,x_j) = (\gamma x_i^T x_j + r)^d$, $\gamma > 0$
- Radial basis function(RBF): $K(x_i,x_j) = \exp(-\gamma \| x_i - x_j \|^2)$, $\gamma > 0$
- Sigmoid: $K(x_i,x_j) = \tanh(\gamma x_i^T x_j + r)$

이때, 커널 파라미터는 학습 시 오류 항목에 대한 벌점(penalty) 파라미터인 C와 γ,r,d 등이 사용된다. 본 논문에서는 위 3가지 기계학습방법에 대하여 실험을 수행하여, 각 방법에 대한 성능을 평가하였다.

4. 실 험

4.1 실험방법

범용적인 전문용어 인식률을 측정하기 위하여, KISTI[10]가 보유하고 있는 과학 기술 문헌 중, 해외논문 300건과, 미국특허 문헌 100건을 수집하여 학습 집합을 구축하였다. 논문은 초록만을 포함하므로 평균 1.95KB 크기이고, 특허는 청구항까지 포함하여 평균 6.28KB 이었고, 논문과 특허에서 인식되는 전문용어 수의 균형을 맞추기 위하여 1:3의 비율

7) http://www.nlm.nih.gov/mesh/meshhome.html
8) KISTI 다국어 자동의미색인 시스템(S&T Terminology system for the Evaluation and Analysis of Knowledge)
9) http://www.bing.com/
10) KISTI(Korea Institute of Science and Technology Information), http://www.kisti.re.kr

〈표 6〉 다양한 개체 인식 특징별 기계학습 개체 인식 결과

		D	C	DC	DS	DW	DWS	DWSC
로지스틱 회귀	정확률	57.6%	72.1%	75.4%	78.2%	78.9%	79.1%	79.2%
	재현율	75.9%	75.9%	76.2%	78.2%	79.7%	79.8%	79.9%
	F-값	**65.5%**	**74.0%**	75.8%	**78.2%**	79.3%	79.4%	79.5%
C4.5	정확률	57.6%	69.6%	75.1%	78.7%	81.4%	81.2%	81.1%
	재현율	75.9%	75.9%	76.8%	77.8%	80.2%	80.7%	81.4%
	F-값	**65.5%**	72.6%	**75.9%**	**78.2%**	**80.8%**	**80.9%**	**81.2%**
SVMs	정확률	57.6%	69.0%	74.8%	78.7%	79.8%	80.6%	77.9%
	재현율	75.9%	75.8%	76.7%	77.8%	79.8%	80.2%	79.5%
	F-값	**65.5%**	72.2%	75.7%	**78.2%**	79.8%	80.4%	78.6%

로 학습 집합 건수를 선정하였다. 학습 집합은 전문가 2인에 의해 수작업으로 전문용어에 대한 태깅작업을 수행하였고, 결과적으로 7,118개의 전문용어가 추출되었다.

실험 방법으로, 먼저 3.1절에서 설명한 것과 같이 명사구 뿐만 아니라 접속사 및 전치사를 포함한 명사구 형식의 다양한 후보 용어를 추출하였고, 각 추출된 후보용어를 3가지 기계 학습된 알고리즘 기반으로 이진 분류를 수행하여 전문용어 판별 정확도를 확인하는 방식으로 수행되었다.

1차 실험으로, 3.2에서 설명된 기계학습 자질 중에서 웹기반 단어 응집도 자질들(⑤,⑥,⑦,⑧)을 제외한 나머지 ①,②,③,④ 자질들만을 조합하여 로지스틱 회귀, C4.5, SVMs의 성능을 비교하였다. 이때, 로지스틱 회귀와 C4.5 방법은 각각 Weka[17]의 Logistic 모듈과 J48 모듈을 사용하였고, SVMs 도구는 libSVM[12]을 활용하여 수행하였다. libSVM은 Linear, RBF, Polynomial, Sigmoid 4가지 커널을 제공하고 있어서, 이 4개 커널에 대해 각각 실험을 수행했으나, 일괄적으로 RBF커널의 결과가 상대적으로 우수하여 본 실험에서는 RBF커널의 결과만을 포함하였다. RBF커널의 두 파라미터(C: 오류항목에 대한 penalty 파라미터, γ (Gamma):$1/(2*\sigma^2)$) 튜닝을 위해 libSVM에서 제공하는 "Loose grid search" 방식[11]을 적용하여 최적의 파라미터 선택 작업을 수행하였다. 예를 들어, 파라미터 C와 γ값을 $C=2^{-5},2^{-3},...,2^{15}$, $\gamma=2^{-15},2^{-13},...,2^{9}$ 사이에서 변경해 가며 최적의 C와 γ값의 범위를 찾은 후 C와 γ값의 범위를 좁혀서(예를 들어, $C=2^{3},2^{-3.25},...,2^{-5}$, $\gamma=2^{-7},2^{-6.5},...,2^{-5}$) 다시 수행하는 방식으로 최적의 파라미터 C와 γ값을 찾았다.

2차 실험은 문헌 내부의 용어빈도 정보를 활용하여 전문용어 추출하는 방식인 C-value와의 비교실험을 수행 하였다. 1차 실험결과에서 C-value(C) 조합 중 가장 높은 성능을 보인 자질조합과 C-value(C)를 포함하지 않은 조합 중 가장 높은 성능을 보인 자질조합을 대상으로, 웹 기반 응집도 자질들을 조합하여 성능을 측정하였다.

1차 및 2차 실험평가를 위하여 기존 연구들[10][11]에서 일반적으로 사용하는 정확률, 재현율, F-값을 사용한다. 정확률은 시스템에서 전문용어로 인식된 것 중 정답 비율을 의미하고, 재현율은 문서에 태깅된 모든 전문용어 중 인식된 정답 비율을 나타낸다. F-값은 정확률과 재현율을 통합적으로 나타내는 평가 기준이다. 일반적으로 사전 기반 전문용어 인식 시스템은 높은 정확률을 나타내는 반면 재현율이 낮은 단점을 갖고 있다. 따라서 정확한 성능평가를 위해 이 두 가지 평가 척도를 함께 고려하는 게 일반적이다.

$$\text{정확률}(preision) = \frac{\text{전문용어 정답 수}}{\text{시스템에서 인식한 전문용어의 수}}$$

$$\text{재현율}(recall) = \frac{\text{전문용어 정답 수}}{\text{문서에 태깅된 전체 전문용어 수}}$$

$$F\text{-값}(F\text{-}score) = \frac{2 \times precision \times recall}{precision + recall} \quad (7)$$

기존 전문용어 인식 결과는 특정 도메인에서 수행된 결과이기 때문에, 본 논문에서 제안하는 범용 전문용어 인식 방법과 직접적인 비교가 어렵다. 따라서 비교 실험을 위해, 전문용어 인식에 일반적으로 많이 사용되는 지역 통계 정보 기반 C-value 방법과 웹 검색 기반 전역 통계 정보 활용 방법을 비교하는 방식으로 핵심 실험을 수행하였다.

4.2 실험 결과

아래의 각 실험 결과들은 10-fold 교차 검증 방식으로 수행하였다. <표 6>은 1차 실험의 결과를 나타낸 것으로, 각 자질별 기계학습 방법의 실험 결과를 정확률/재현율/F-값으로 구분해서 표현하였다. 1차 실험의 주된 목적은 사전, C-value, 웹 정보 활용 방법 간의 성능을 비교하기 위한 것이다.

먼저 사전(사전+위키 사전)만을 적용한 경우(D) 65.5%의 F-값으로 높지 않은 성능을 보였다. 보통 사전 기반 인식 성능을 높이기 위해 사전의 항목을 늘리게 되는데, 특정 도메인에 대한 전문용어 인식 시스템인 경우는 어느 정도 가능하나 범용 도메인에 적용하기 위한 시스템인 경우는 사전

11) http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

〈표 7〉 응집도 정보 추가에 따른 성능 비교

		DWSCF1	DWSF1	DWSF2	DWSF3	DWSF4
로지스틱 회귀	정확률	80.4%	80.6%	79.6%	79.1%	79.8%
	재현율	80.6%	80.7%	80.4%	79.8%	79.9%
	F-값	80.4%	80.6%	80.0%	79.4%	79.8%
C4.5	정확률	81.8%	82.4%	82.5%	82.1%	80.0%
	재현율	81.3%	81.2%	80.9%	81.2%	80.0%
	F-값	**81.5%**	**81.8%**	**81.7%**	**81.6%**	**80.0%**
SVMs	정확률	79.4%	81.9%	81.3%	79.3%	68.9%
	재현율	80.6%	81.1%	81.1%	79.8%	75.4%
	F-값	80.0%	81.5%	81.2%	79.5%	72.0%

수집/관리의 어려움을 고려할 때 바람직하지 않은 방법이라 할 수 있다. 대신 사전기반 방법에 다양한 통계적인 방법을 접목하여 활용하고 있는데, 일반적으로 많이 사용되는 통계 기반 전문용어 인식 방법이 C-value 방법이다. 이 방법을 적용한 경우(C)는 최고 74.0% F-값으로 본 실험에서도 상대적으로 높은 성능을 보였다. 사전(D)과 C-value(C) 방법을 함께 적용한 경우(DC)는 76.0% F-값으로, C-value(C)만을 적용한 경우보다 약간 향상된 결과를 보이고 있으나, 사전(D)과 후보용어 형태 특징(S)을 함께 적용한 경우(DS)가 사전(D)+C-value(C) 혼합 방법(DC)보다 모든 기계학습 방법에서 향상된 결과를 볼 수 있었다. 사전(D)을 웹 빈도수(W)와 접목한 경우(DW), 사전(D)+C-value(C)와 비교하여 최고 6.5% 상승한 80.8% 값을 얻을 수 있었다. 이는 문헌 내부의 후보용어의 통계 정보를 활용한 방식(C-value)보다 웹에서 추출 가능한 통계 정보를 활용하는 방식이 높은 성능을 보인다는데 그 의미를 찾을 수 있다.

지역 통계 정보를 사용하는 C-value의 경우에 대상 문헌 집합의 규모[12]에 영향을 받기 때문에, 전역 통계 정보인 웹 통계 정보를 사용하는 경우가 좋은 성능을 보이는 것으로 판단된다. 대상 문서집합이 소규모이거나, 다양한 분야를 포함하는 전문용어 인식 시스템의 경우는 본 논문에서 제안하는 웹을 활용한 전역 통계 정보를 사용하는 것이 적합하다고 할 수 있다.

마지막으로, 모든 정보를 통합한 방식인 사전(D)+웹 빈도수(W)+후보 용어 형태 특징(S)+C-value(C)의 경우(DWSC)는 81.2%의 값으로 사전(D)+C-value(C)의 경우에 비해 7% 상승한 결과를 보였다. 특히, 대부분의 경우 로지스틱 회귀 방법이나 SVMs보다 결정 트리 방법론인 C4.5의 성능이 좋게 나타나는 것을 확인할 수 있었다.

2차 실험의 주목적은 다양한 응집도 계산 방식의 비교이다. <표 7>은 2차 실험 결과를 요약한 결과를 보여준다. <표 6>에서 C-value를 포함하여 최대 성능을 보인 DWSC와 C-value를 포함하지 않고 최대 성능을 보인 DWS에 웹 통계를 이용한 용어의 응집도(F1~F4) 특징을 추가하여 성능을 평가한 결과이다.

DWSC부분에서는 DWSCF1(<표 7>의 첫 번째 열)이 다른 DWSCF2~4의 결과보다 높은 성능을 보였다. 그리하여 DWSCF1과 C-value를 포함하지 않은 DWS와 F1~F4를 조합한 결과를 비교하였다. 하지만 DWS와 응집도(F1-F3) 특징을 함께 추가한 것 방법에서 높은 성능을 확인할 수 있었다. 이는 DWSCF에서 사용한 C-value(C)가 문헌 내에서의 후보용어의 통계 정보를 활용한 반면, DWSF 조합은 웹 정보와 문형 정보만을 활용했다는 차이가 있다. 즉, 웹 통계 정보를 활용한 전문용어 인식 방법이 문헌 내부 통계 정보를 활용한 방법을 대체할 수 있다는 결론이다. 또한, 실험 결과를 기반으로 가장 우수한 응집도 계산 방식은 Normalized Google Distance(F1)을 이용한 방법임을 확인할 수 있었다.

이 실험에서도 일반적으로 이진 분류에서 강점을 드러내고 있는 SVMs 방법과 비교해서 높은(비록 큰 차이는 아니었지만) 성능을 보인 모델은 결정 트리 방법인 C4.5 방법이었다. 결정 트리 방법은 학습시간 효율이나 적용 방법 편리성 등에서 다른 알고리즘보다 효율적이라는 점에서 범용 개체 추출 시스템에 적합하다고 할 수 있다.

5. 결 론

본 논문에서는 웹 검색 결과를 기반으로 하는 범용 과학기술 전문용어 인식기법을 제안하였다. 대상 문헌집합의 특정 분야에 종속되는 전문용어 인식방법을 탈피하여, 대상 문헌집합의 지역 통계 정보 외에 용어사전과 웹 검색결과의 전역 통계 정보를 기계학습모델의 자질로 활용하였다. 세 가지 기계학습모델, 로지스틱 회귀, C4.5, SVMs를 이용하여 성능을 비교하였고, C4.5 알고리즘을 적용한 방법에서 F-값 81.8%인 최고 성능을 나타냈다.

이 실험결과는 문헌의 지역 통계 정보인 C-value의 사용 없이, 웹 정보만을 이용하여 높은 전문용어 인식 성능을 확보함을 보여주고 있어, 본 논문에서 제시한 전문용어 인식

12) 본 논문에서 지역 통계 정보를 추출하기 위해 사용된 대상 문헌 집합은 학습 집합 300건을 포함하여, 10,439건이다.

시스템이 특정 분야에 한정적이지 않은 범용 전문용어인식으로 활용될 수 있음을 보여주었다. 향후 범용전문용어 인식 성능을 추가 향상하기 위하여, 웹 자원 중의 하나인 위키사전에 대한 분야 분류값을 활용한 연구를 수행할 예정이다.

참 고 문 헌

[1] Beatrice Daille, Eric Gaussier, and Jean-Marc Lange, "Towards Automatic Extraction of Monolingual and Bilingual Terminology. COLING-94, 1994.

[2] Church, K. & Hanks. P, "Word association norms, mutual information, and lexicography," Computational Linguistics, Vol.16, No.1, pp.22-29, 1990.

[3] Corinna Cortes and V. Vapnik, "Support-Vector Networks", Machine Learning, Vol.20, No.3, pp-273-297, 1995.

[4] Dunning, T. "Accurate methods for the statistics of surprise and coincidence," Computational Linguistics, Vol.19, No.1, pp.61-74, 1993.

[5] F. Smadja, K. R. McKeown, and V. Hatzivassiloglou, "Translating collocations for bilingual lexicons: A statistical approach", Computational Linguistics, Vol.22, No.1, pp.1-38, 1996.

[6] G. Zhou, J. Zhang, J. Su, D. Shen and C. Tan, "Recognizing names in biomedical texts: a machine learning approach," Bioinformatics, Vol.20, No.7, pp.1178-1190, 2004.

[7] Ido Dagan and Kenneth W. Church, "Termight: Identifying and translating technical terminology," ANLP, pp.34-40, 1994.

[8] J. Kazama, T. Makino, Y. Ohta, J. Tsujii, "Tuning support vector machines for biomedical named entity recognition," Proceedings of the ACL-02 workshop on NLP in the biomedical domain, Vol.3, pp.1-8, 2002.

[9] Justeson, J.S. and S.M. Katz, "Technical terminology : some lingustic propertis and an algorithm for identification in text," Natural Language Engineering, Vol.1, No.1, pp.9-27, 1995.

[10] Joachim Wermter and Udo Hahn, "Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms," HLT'05 Proceedings of the conference on Human Language Technology and Empirical Methods in NLP, 2005.

[11] K. Frantzi and S. Ananiadou and Hideki Mima, "Automatic recognition of multi-word terms: the C-value/NC-value method," International Journal on Digital Libraries, Vol.3, No.2, pp.115-130, 2000.

[12] LIBSVM - A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[13] Nakagawa, Hiroshi and Tatsunori Mori, "Automatic term recognition based on statistics of compound nouns and their components," Terminology, Vol.9, No.2, pp.201-219, 2003.

[14] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[15] Rudi Cilibrasi and Paul Vitanyi, "The Google Similarity Distance," IEEE Trans. Knowledge and Data Engineering, Vol.19, No.3, pp.370 - 383, 2007.

[16] Qing T. Zeng, Tony Tse, et. al., "Term identification methods for consumer health vocabulary development," Journal of medical Internet research, Vol.9, No.1, 2007.

[17] WEKA - Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

[18] Y. Tseng, C. Lin, Y. Lin, "Text mining techniques for patent analysis," Information Processing and Management, Vol.43, No.5, pp.1216-1247, 2007.

최 윤 수

e-mail : armian@kisti.re.kr

1993년 충남대학교 컴퓨터공학과(학사)

1995년 충남대학교 컴퓨터공학과(석사)

1995년~현 재 한국과학기술정보연구원 선임연구원

관심분야 : 정보검색, 텍스트마이닝

송 사 광

e-mail : esmallj@kisti.re.kr

1997년 충남대학교 통계학과(학사)

1999년 충남대학교 컴퓨터과학과(석사)

2011년 한국과학기술원 잔산학과(박사)

2005년~2010년 한국전자통신연구원 바이오인포매틱스팀 연구원

2010년~현 재 한국과학기술정보연구원 선임연구원

관심분야 : 텍스트마이닝, 자연어처리, 정보검색, 시맨틱 웹

전 홍 우

e-mail : hw.chun@kisti.re.kr

2002년 고려대학교 컴퓨터학과(학사)

2004년 고려대학교 컴퓨터학과(석사)

2007년 일본 동경대학교 컴퓨터과학(박사)

2007년~2008년 Japan National Institute of Advanced Industrial Science and Technology (AIST), Japan Biological Information Research Center (JBIRC), 박사후과정

2008년~2009년 Japan Research Organization of Information Systems, Database Center for Life Science, Project researcher

2009년~현 재 한국과학기술정보연구원 선임연구원

관심분야 : 자연어처리, 기계학습

정 창 후

e-mail : chjeong@kisti.re.kr
1999년 충남대학교 컴퓨터과학과(학사)
2002년 충남대학교 컴퓨터과학과(석사)
2003년~현 재 한국과학기술정보연구원 선임연구원
관심분야 : 정보검색 및 추출, 텍스트마이닝

최 성 필

e-mail : spchoi@kisti.re.kr
1996년 부산대학교 전자계산학과(학사)
1998년 부산대학교 전자계산학과(석사)
2009년 한국과학기술원 정보통신공학과 (박사 수료)
1998년~현 재 한국과학기술정보연구원 선임연구원
관심분야 : 기계학습, 정보검색, 자연어처리, 정보추출, 텍스트마이닝