기계학습에 기반한 생의학분야 전문용어의 자동 인식

(Machine-Learning Based Biomedical Term Recognition)

오 종 훈 † 최 기 선 ††
(Jong-Hoon Oh) (Key-Sun Choi)

요 약 일정 분야의 문서들에서 그 분야 특징을 반영하는 전문용어를 자동으로 인식하는 연구에 대한 관심이 증가하고 있다. '전문용어 인식'은 문서에서 전문용어가 될 수 있는 언어적 단위를 파악하는 '용어추출'과정과 '용어추출'과정에서 얻어진 용어목록 중 해당분야의 전문용어를 고르는 '전문용어 선택' 과정으로 구성된다. '전문용어 선택'과정은 용어목록을 전문용어의 특징에 따라 순위화한 후 타당한 전문용어를 파악하는 작업으로 정의된다. 따라서 전문용어 선택 문제는 용어목록의 순위화 작업과 순위화된 목록에서 전문용어와 비전문용어 간의 경계를 인식하는 작업으로 정의된다. 기존의 전문용어 선택 기법은 주로 용어의 빈도수 등과 같은 통계적 특징만을 이용하였다. 하지만 통계적 특징만으로는 효과적으로 전문용어를 선택하기 어렵다. 본 논문의 논제는 전문용어 선택에서 다양한 전문용어의 특징을 고려하고 이들 중 전문용어 선택에서 효과적인 특징을 찾으려는 것이다. 순위화 문제는 다양한 전문용어 특징을 도출하고 이들을 기계학습방법으로 통합하여 해결한다. 경계인식 문제는 전문용어와 비전문용어의 이진 분류 문제로 정의하고 기계학습방법으로 해결한다. 본 논문의 기법은 경계인식측면에서 78~86%의 정확률과 87%~90%의 재현율을 나타내었으며, 순위화 측면에서 89%~92%의 11포인트 평균정확률을 나타내었다. 또한 기존 연구보다 최고 26%의 성능향상을 보였다.

키워드 : 전문용어 인식, 기계학습, 생의학, 전문용어, 분류

Abstract There has been increasing interest in automatic term recognition (ATR), which recognizes technical terms for given domain specific texts. ATR is composed of 'term extraction', which extracts candidates of technical terms and 'term selection' which decides whether terms in a term list derived from 'term extraction' are technical terms or not. 'term selection' is a process to rank a term list depending on features of technical term and to find the boundary between technical term and general term. The previous works just use statistical features of terms for 'term selection'. However, there are limitations on effectively selecting technical terms among a term list using the statistical feature. The objective of this paper is to find effective features for 'term selection' by considering various aspects of technical terms. In order to solve the ranking problem, we derive various features of technical terms and combine the features using machine-learning algorithms. For solving the boundary finding problem, we define it as a binary classification problem which classifies a term in a term list into technical term and general term. Experiments show that our method records 78~86% precision and 87%~90% recall in boundary finding, and 89%~92% 11-point precision in ranking. Moreover, our method shows higher performance than the previous work's about 26% in maximum.

Key words : Automatic term recognition, Machine learning, Biomedicine, Terminology, Classification

1. 서 론

전문용어란 전문분야의 개념을 표현하기 위한 언어적 기호이다. 전문분야마다 분야 특징적인 개념이 사용되기 때문에 전문용어는 전문분야를 특성화하는 단위로 사용된다. 따라서 전문분야 문서처리에서 전문용어를 효과적으로 파악하는 것은 매우 중요하다. 하지만 기술의 발전

† 학생회원 : Expert researcher, Computational Linguistics Group, NICT(情報通信研究機構), Japan
rovellia@nict.go.jp
†† 종신회원 : 한국과학기술원 전산학과 교수
kschoi@world.kaist.ac.kr
논문접수 : 2005년 6월 22일
심사완료 : 2006년 7월 12일

으로 인해 전문분야 개념과 이를 지칭하는 전문용어가 지속적으로 새롭게 만들어지기 때문에 사전에만 의존한 전문용어 파악에는 한계가 있다. 이러한 이유로 전문분야 문서에서 해당 분야 전문용어를 자동으로 파악하는 '**전문용어 인식**'에 대한 연구가 활발히 진행되어 왔다[1-12].

기존의 전문용어 인식 연구들은 문서에서 전문용어가 될 수 있는 언어적 단위를 파악하는 '용어 추출'과정과 용어 중 올바른 전문용어를 고르는 '전문용어 선택'과정으로 구성된다. 전문용어 인식 결과는 순위화된 용어 목록으로 표현된다. 순위화된 용어 목록에서 순위는 해당 분야의 전문용어가 될 가능성을 나타내는 잣대이다. 이상적인 전문용어 인식 체계는 주어진 전문분야와 문서에 대하여 전문분야에서의 전문성에 따라 순위화된 용어 목록을 제시하여야 하며, 전문용어 선택 척도에 의해 용어 목록에서 비전문용어를 효과적으로 걸러내어야 한다[13]. 즉, 이상적인 전문용어 인식 기법은 용어 목록에 대한 올바른 순위화와 순위화된 용어 목록에서 전문용어와 비전문용어 간의 올바른 경계정보를 제공하여야 한다. 하지만 기존의 전문용어 인식 기법에서 전문용어 선택은 단순 통계적 정보만을 사용하였기 때문에 효과적인 순위화와 올바른 경계 정보를 제공하지 못하였다. 본 논문에서는 기존의 전문용어 인식 기법의 문제점을 해결하기 위하여 다양한 전문용어 특징과 기계학습에 기반한 전문용어 선택 기법을 제안한다.

효과적인 용어 목록의 순위화는 전문용어의 다양한 특징을 고려하여야 한다. 본 논문에서는 생의학분야에서 효과적으로 용어 목록을 순위화하기 위하여 **철자적, 형태소적, 어휘적, 구문적, 사전적 특징**이라는 다섯 가지 특징을 사용하였다. 그리고 이들 특징들을 효과적으로 통합하기 위하여 기계학습기법을 사용하였다. 각 특징은 생의학분야 전문용어와 비전문용어간의 대비되는 철자, 형태소, 어휘, 구문, 사전 수준에서의 차이점을 표현하기 위해 사용된다. 예를 들어, 생의학분야 전문용어는 비전문용어에 비해 *gene, protein*과 같은 어휘를 중심어로 가지는 경우가 많다.

경계인식 문제를 해결하기 위하여 '전문용어 선택'과정을 '**이진 분류**'문제로 변환하여 해결하였다. 즉, 용어 목록 중의 용어를 전문용어와 비전문용어로 이진 분류하는 작업으로 정의된다. 기존의 전문용어 인식기법[1-12]은 각 기법이 사용한 점수함수로 순위화된 용어 목록은 제시하였지만 전문용어와 비전문용어의 경계를 명확하게 제시하지 못하였다. 기존 연구들에서는 각 기법이 용어에 부여한 점수와 이에 대한 특정 임계값이 전문용어와 비전문용어를 분류하는 기준이 된다. 하지만 기존 연구에서 사용한 점수함수는 문서 크기와 문서 특징에 의존적이기 때문에 임계값만으로는 효과적인 분류 기준이 되기 어렵다. 즉, 특정 문서에서 추출한 순위화된 용어 목록의 경계인식에 효과적인 임계값이 다른 크기와 다른 특징을 가진 문서로부터 추출한 순위화된 용어 목록의 경계인식에는 효과적이지 않는 경우가 많다. 따라서 문서 크기나 문서 특징에 독립적인 경계인식기법이 필요하다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 전문용어 인식 기법에 대하여 기술하고, 3장에서는 본 논문의 기법에 대하여 설명한다. 4장에서는 실험과 그 결과에 대하여 설명하고 5장에서는 결론을 맺는다.

2. 기존 연구

기존의 전문용어 인식 기법은 '**용어 추출**'과정과 '**전문용어 선택**'과정으로 구성된다. 용어 추출과정에서는 명사구를 추출하기 위한 구문규칙을 용어 추출 규칙으로 정의하고, 이를 이용하여 용어를 추출한다. 예를 들어, [5]에서 사용한 용어 추출 규칙은 다음과 같다.

$$(A|N)+N$$

여기에서, A는 관형사, N은 명사를 각각 나타낸다.

용어 추출과정의 결과인 용어 목록은 전문용어 선택 과정에서 여러 가지 통계적 정보에 의해 모형화된 점수함수에 의해 순위화 된다. 기존의 전문용어 선택 기법은 사용한 점수함수의 특징에 따라 크게 빈도수기반[8], 용어간 내포관계 기반[5], 용어의 공기 용어 기반[5,9], 명사결합정보기반 방법[10]으로 분류된다.

빈도수기반 방법[8]은 용어의 문서 내 빈도수를 이용하여 식 (1)과 같이 점수함수를 모형화하고 용어 목록을 순위화하였다. 식 (1)의 기법은 해당 분야 문서에서 자주 나타나는 용어는 전문용어일 가능성이 높다라는 가정에 기반한다.

$$term\_score(t_i) = freq(t_i) \quad (1)$$

여기에서 t_i는 용어 목록의 i번째 용어를, $freq(t_i)$는 t_i의 문서 내 빈도수를 각각 나타낸다.

용어간 내포관계 기반 방법[5]에서는 용어의 빈도수 뿐만 아니라 용어 간의 내포관계를 이용하여 C-value라는 점수함수를 모형화하였다. C-value는 t_i가 문서에서 자주 나타날수록, 길이가 길수록, t_i를 내포하는 용어가 적을수록 높은 값을 가진다. 용어 t_i가 용어 t_j를 내포한다는 것은 t_j가 t_i에 포함된다는 것을 의미한다. 예를 들어, *Bcl-2 gene*은 *gene*을 내포한다. t_i가 내포되지 않은 완전한 형태로 자주 나타난다는 것은 그 자체로 해당분야 문서에서 하나의 개념을 지칭하는 표현임을 나타낸다. 따라서 C-value에서는 내포되지 않는 용어가 높은 점수를 갖는다. 하지만 생의학 분야의 용어 *gene*

과 같이 많은 용어에 내포가 되더라도 그 분야 전문용어로 사용되는 경우가 있다. C-value에서는 길이가 긴 용어에 높은 점수를 부여한다. 이는 길이가 길수록 수식어가 많아져 보다 세부적인 개념을 지칭할 가능성이 높다는 가정에 기반한다. 하지만 particular cell line, other cell line의 particular, other와 같은 수식어에 의해서는 용어의 전문성이 낮아지는 경우도 있다[13].

$$C\text{-}value(t_i)=\begin{cases}\log_2|t_i|\cdot freq(t_i); & t_i\text{가 내포되는 용어가 없는 경우}\\ \log_2|t_i|\cdot\left(freq(t_i)-\frac{1}{P(T_{t_i})}\sum_{b\in T_{t_i}}freq(b)\right); & \text{그렇지 않으면}\end{cases} \quad (2)$$

여기에서 t_i는 i번째 용어를, $freq(t_i)$는 t_i의 문서내 빈도수를, $|t_i|$는 t_i의 구성요소의 수를, T_{ti}는 t_i를 내포하는 용어 집합을, $P(T_{ti})$는 T_{ti}에 포함되는 용어의 개수를 각각 나타낸다.

문맥 기반 방법에는 Frantzi[5]가 제안한 NC-value와 Maynard[9]가 제안한 SNC-value가 있다. NC-value는 용어의 문맥정보를 나타내는 점수함수 $CF(t_i)$와 C-value를 선형적으로 결합하여 모형화된다. $CF(t_i)$는 t_i의 좌우 문맥에 나타나는 어휘집합 $W=\{w_1,\cdots,w_n\}$내 어휘들의 문맥가중치로 표현되며, 문맥가중치는 t_i와 w_j공기관계에 의해 계산된다[5]. SNC-value에서는 용어의 문맥 내 의미정보를 $IW(t_i)$라는 점수함수로 모형화하고 NC-value와 선형적으로 결합하였다. $IW(t_i)$는 용어 t_i의 경계에 나타나는 어휘의 품사 정보에 기반한 경계품사가중치와 t_i의 의미정보와 t_i의 문맥 내에 나타나는 용어 t_j의 의미정보에 의해 계산되는 의미유사도 가중치에 의해 계산된다. 의미유사도 가중치 계산을 위해 의미망에서의 의미간 거리 정보를 사용하였다[9].

$$NC-value(t_i)=0.8\times C-value(t_i)+0.2\times CF(t_i)$$
$$SNC-value(t_i)=0.4\times C-value(t_i)+0.1\times CF(t_i)+0.5\times IW(t_i) \quad (3)$$

Nakagawa는 "용어의 구성요소로 사용되는 명사간의 결합 강도가 높을수록 전문용어가 될 가능성이 높다"라는 가정에 기반하여 명사결합정보기반 방법인 GM과 FGM이라는 점수함수를 제안하였다. GM과 FGM은 용어의 구성요소인 명사들이 그 분야 문서에서 사용되는 양상을 점수함수로 모형화한 것이다. GM, FGM은 분야 문서에서 특정 명사의 왼쪽과 오른쪽에 나타나는 명사의 종류 및 빈도수를 이용하여 식 (4)와 같이 표현된다. FGM은 용어 t_i의 GM(t_i)값과 빈도수 $f(t_i)$의 곱으로 표현된다. 여기에서 용어 t_i는 L개의 구성요소로 이루어지며, 분야 문서에서 N_j의 왼쪽에 나타나는 명사의 개수는 $LN(N_j)$, 오른쪽에 나타나는 개수는 $RN(N_j)$라고 정의된다.

$$GM(t_i)=\left(\prod_{j=1}^{L}(LN(N_j)+1)\times(RN(N_j)+1)\right)^{1/2L}$$
$$FGM(t_i)=GM(t_i)\times f(t_i)$$
$$where\ t_i=\{N_1,\ldots,N_L\} \quad (4)$$

기존의 전문용어 인식 연구들[1,2,3,4,5,6,7,8,9,10,11,12]은 주로 전문용어 선택과정에 연구의 초점을 맞추고 있다. 전문용어의 선택과정은 용어에 대한 순위화 과정으로 생각할 수 있으며, 순위화 과정에 사용하는 점수함수들은 식 (5)와 같이 점수함수들의 선형 결합 (linear interpolation)으로 일반화 된다. 여기에서 $score(t_a)$는 전문용어후보 t_a에 대한 순위화 함수이며, sf_i는 각 특성에 따른 점수함수를, β_i는 sf_i에 대한 가중치를 각각 나타낸다.

$$score(t_a)=\beta_1\times sf_1(t_a)+..+\beta_n\times sf_n(t_a) \quad (5)$$

이상적인 전문용어 인식 기법은 용어 목록에 대한 올바른 순위화와 순위화된 용어 목록에서 전문용어와 비전문용어 간의 올바른 경계정보를 제공하여야 한다. 하지만 기존의 전문용어 인식 기법에서 전문용어 선택은 대부분 단순 통계적 정보만을 사용하였기 때문에 효과적인 순위화와 올바른 경계 정보를 제공하지 못하였다. 또한 기존의 전문용어 인식기법은 각 기법이 사용한 점수함수로 순위화된 용어 목록은 제시하였지만 전문용어와 비전문용어의 경계를 명확하게 제시하지 못하였다. 본 논문에서는 기존의 전문용어 인식 기법의 문제점을 해결하기 위하여 다양한 전문용어 특성과 기계학습에 기반한 전문용어 선택 기법을 제안한다.

3. 기계학습을 통한 전문용어 인식

효과적인 용어 목록의 순위화는 전문용어의 다양한 특성을 고려하여야 한다. 본 논문에서는 생의학분야에서 효과적으로 용어 목록을 순위화하기 위하여 철자적, 형태소적, 어휘적, 구문적, 사전적 특성이라는 다섯 가지 특성을 사용하였다. 그리고 이들 특성들을 효과적으로 통합하기 위하여 기계학습기법을 사용하였다. 각 특성은 생의학분야 전문용어와 비전문용어간의 대비되는 철자, 형태소, 어휘, 구문, 사전 수준에서의 차이점을 표현하기 위해 사용된다. 예를 들어, 생의학분야 전문용어는 비전문용어에 비해 다음과 같은 철자, 형태소, 어휘, 구문, 사전적 특성을 보인다. 형태적 수준에서는 기호/숫자 등을 포함하는 경우가 많으며, 형태소 수준에서는 '-cyte', '-itis'와 같은 그리스/라틴 어원의 형태소를 포함하는 경우가 많다. 또한 어휘적 수준에서는 'gene', 'protein'과 같은 어휘를 중심어로 가지는 경우가 많고, 구문적 수준에서는 'activate', 'affect', 'bind'와 같은 동사의 주어나 목적어로 사용되는 경우가 많다. 사전적 수준에서는 생의학분야 전문용어가 기존 생의학 전문분야 사전에 나타나는 표제어로 구성되는 경우가 많다. 경계인식 문제를 해결하기 위하여 "전문용어 선택"과정을 "이진 분류"문제로 변환하여 해결하였다. 즉, 용어 목록 중의

용어를 전문용어와 비전문용어로 이진 분류하는 작업으로 정의된다.

3.1절에서는 제안하는 전문용어 인식 기법의 전체적인 시스템 구조도를 설명하고 3.2절에서는 전문용어 선택에서 사용하는 다양한 특성에 대하여 설명한다. 3.3절에서는 이진분류를 이용한 기계학습기반 전문용어 선택 기법을 통하여 용어목록의 용어에서 전문용어를 파악하는 과정을 설명한다.

3.1 시스템 구조도

그림 1은 본 논문에서 제안하는 다양한 전문용어 특징과 기계학습에 기반한 전문용어 인식 기법의 구조도를 나타낸다. 본 논문의 기법은 기존의 전문용어 인식기법과 마찬가지로 '**용어 추출**' 단계와 '이진 분류'에 기반한 '**전문용어 선택**' 단계로 구성된다.

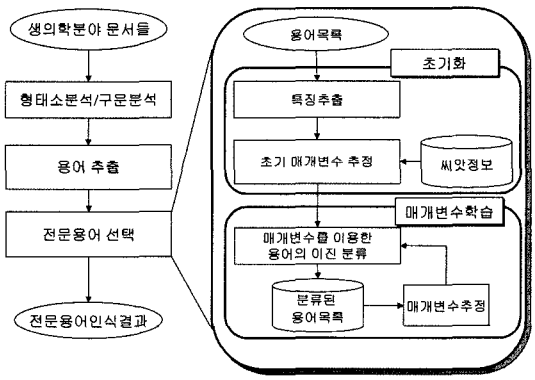

그림 1 전문용어 인식 체계도

용어 추출단계에서는 구문분석된 생의학분야 영어문서내의 용어를 구문규칙 (품사 나열 규칙)을 이용하여 추출한다. 전문용어의 많은 부분이 명사구 형태로 나타나기 때문에 명사구에 대한 구문규칙을 이용하여 용어를 추출한다. 본 논문에서 추출한 명사구의 형태는 다음과 같은 정규식으로 표현된다.

$$(A|N)^* N$$

여기에서 A는 관형사를, N은 명사를 각각 나타낸다. 정의된 정규식으로 *EGF receptor, focal adhesion kinase*와 같은 명사구를 파악할 수 있다.

명사구는 관형사와 명사의 나열로서 표현되며, 수식어와 중심어로 구성된다. 그런데, 수식어가 전문분야의 특징을 나타내지 못하는 일반적인 단어일 경우 용어목록의 품질을 떨어뜨리는 요인이 되므로, 이를 적절히 처리해야 한다. 예를 들어 "*cell line*"에 대하여, 분야를 특성화하지 못하는 *particular, some, other*와 같은 수식어로 인하여 *particular cell line, other cell line, some cell lines*과 같이 부적절한 용어가 추출된다. *particular, some, other*와 같은 수식어들은 제거되어야 하며, 이를 통하여 *cell line*이라는 하나의 용어만이 추출되게 된다. 본 논문에서는 일반분야와 전문분야 문서에서 나타나는 수식어의 상대빈도수를 기반으로 용어에서 불필요한 수식어를 제거하였다.

추출된 용어는 전문용어 선택 단계에서 전문용어와 비전문용어로 이진분류(binary classification)된다. 용어의 이진분류는 기계학습을 통하여 이루어지며, '초기화(initialization)'단계와 '매개변수학습 (parameter learning)' 단계로 구성된다(표 1 참조). 초기화 단계에서는 소량의 씨앗정보(seed information)을 이용하여 초기매개변수를 추정한다. 씨앗정보는 전문용어와 비전문용어가 표지된 용어들의 집합이다. 본 논문에서는 생의학분야 전문용어에 대한 씨앗정보로 생의학분야 사전을 사용하였으며, 비전문용어에 대한 씨앗정보로 일반분야 사전을 사용하였다. 초기화 단계에서는 용어가 각 분야 사전에 등재되어 있는 경우 전문용어 또는 비전문용어로 분류하고 이들 분류정보를 이용하여 초기매개변수($parameter_0$)를 추정한다.

표 1 용어의 이진분류 알고리즘

```
#초기화 단계
CTC0 ← 초기분류기 (씨앗정보, TC);
parameter0 ← 매개변수 추정기(CTC0);

#매개변수 학습 단계
i=0;
While(i == 0 || parameteri != parameteri-1){
        i=i+1;
        CTCi ← 분류기(parameteri-1);
        parameteri ←매개변수 추정기(CTCi);
}

여기에서
   TC는 용어 목록
   CTCi는 i번째 학습단계에서 분류된 용어 목록
   parameteri는 i번째 학습단계에서 추정한 매개변수
```

매개변수학습 단계에서는 용어의 이진 분류를 위한 매개변수를 학습한다. 학습은 매개변수의 변화가 없을 때까지 반복된다. i번째 학습단계에서의 학습과정은 다음과 같다. 우선 $i-1$번째 학습단계에서 추정된 매개변수($parameter_{i-1}$)를 이용하여 용어를 이진 분류한다. 분류된 용어 집합 CTC_i는 $parameter_i$를 추정하는데 사용된다.

전문용어 선택 성능은 "어떤 매개변수를 사용할 것인가 혹은 어떤 특징을 사용할 것인가"와 "그 매개변수 혹은 특징을 이용하여 어떻게 용어를 선택할 것인가"에 의해 결정된다. 3.2절에서는 본 논문에서 사용한 매개변수에 대해 정의하고, 3.3절에서는 전문용어 선택을 위해 사용된 기계학습방법에 대하여 설명한다.

3.2 전문용어 선택을 위한 특징

본 논문에서는 전문용어 선택을 위한 특징으로 어휘(lexicon) 특징을 비롯하여, 구문(syntax), 철자(orthography), 사전(dictionary), 형태소(morpheme)의 특징을 이용하였다.

첫째, 어휘특징이란 용어를 구성하는 어휘의 공통성을 나타낸다. 용어의 중심어와 그 수식어는 용어의 개념을 해석하는데 중요한 역할을 한다. 예를 들어, 용어들이 같은 중심어를 가진다면 그 용어들은 유사한 개념을 지칭하거나 개념적으로 상호 연관성을 가진다(예 insulin **receptor**와 estrogen **receptor**). 또한 용어 t_i에 수식어를 추가하여 용어 t_j를 구성하면 일반적으로 t_j는 t_i보다 세부적인 개념을 지칭한다(예 **estrogen receptor**와 leucocytic **estrogen receptor**). 이러한 특징은 분야 개념을 지칭하는 전문용어를 효과적으로 표현하는 특징이므로 본 논문에서는 용어를 구성하는 중심어와 중심어에 대한 수식어를 어휘 특징으로 정의하고 LF(t_i)로 표기한다. 예를 들어, LF(*insulin receptor*) = {M: insulin, H: receptor}이다. 여기에서 M은 수식어, H는 중심어를 나타낸다.

둘째, 구문 특징에 대하여 살펴보자. *bind*, *affect*, *activate*와 같은 동사는 생의학분야 문서에서 분야 지식을 표현하는데 많이 사용된다. 분야 지식은 주로 전문용어간의 연관관계에 의해 표현되므로 이들 동사를 파악하는 것은 분야 지식을 표현하는 용어를 추출하는데 도움이 된다. 예를 들어, 다음의 문장을 보자. "retinoic acid *activates* interferon regulatory factor-1 gene expression", "shed receptor *binds* interleukin-2", "tumour suppressor gene *affects* E2F-mediated regulation"에서 'activate', 'bind', 'affect'는 전문분야 지식을 표현하는데 사용되며 해당동사의 주어 및 목적어는 전문분야 개념을 지칭하는 전문용어이다[3,14-20]. 구문특성은 전문분야 문서에서 자주 사용되는 괄호표현으로 연결되는 관계를 포함한다. 생의학 분야에서 괄호는 주로 다음과 같은 세가지 경우에 사용된다[6].

1) "약어를 정의하기 위해 사용되는 경우"
   - estrogen receptor (ER)
   - GABA (gamma-aminobutyric acid)
2) "동의어나 상위어 같은 의미적 관계를 나타내기 위해 사용되는 경우"
   - natural toxin (i.e., aflatoxin)
   - an inactive HRas protein (RasN17)
3) "참조 또는 수치를 표현하기 위해 사용되는 경우"
   - here by using a recently developed ultrasensitive HPLC technique (Sakhi et al. J. Chromatogr. A. 828:451-460, 1998)
   - CGRP failed to inhibit glucose-stimulated (16.7 mM)

이 중, 1)과 2)에서 괄호 안과 괄호 밖의 전문용어들 간에는 약어나 의미적 상관관계가 존재한다. 반면 3)에서는 괄호 안의 표현이 명사구로 표현되지 않기 때문에 전문용어가 될 가능성이 적다. 본 논문에서는 1)과 2)의 형태의 괄호로 표현되는 전문용어 간의 관계를 괄호관계로 정의하여 파악한다. 구문특성인 주어-동사 구문관계, 목적어-동사 구문관계, 괄호관계는 전문용어 인식에 중요한 특징으로 사용할 수 있으며, 본 논문에서는 이들을 구문 특징으로 정의하고 SF(t_i)로 표기한다. 예를 들어, "shed receptor binds interleukin-2"에서 SF(shed receptor)={S: bind}, SF(interleukin-2)={O: bind}. 여기에서 S는 주어-동사 관계를 O는 목적어-동사 관계를 나타낸다.

세 번째 특징은 철자특징이다. 단백질이나 유전자를 나타내는 전문용어의 명명법 (naming convention)으로 인하여 생의학 분야 전문용어는 비전문용어에 대비되는 철자적 특징을 나타낸다. Fukuda[6]는 단백질을 나타내는 생의학분야 전문용어를 철자적 특징에 따라 세 가지로 분류하고 새로운 단백질을 나타내는 전문용어는 대부분 1), 2)의 형태로 나타난다고 보고하였다.

1) 대문자/소문자, 숫자, 기호로 구성된 단일단어 전문용어 (Nef, p53, Vav)
2) 대문자/소문자, 숫자, 기호로 구성된 다중단어 전문용어 (IL-1 responsive kinase)
3) 소문자로만 구성된 단일단어 전문용어 (actin, tublin, insulin)

즉, 새로운 단백질을 나타내는 전문용어는 소문자로만 구성되지 않고 대문자/소문자/숫자/특수기호 등이 함께 사용되는 경향이 있다(예 ADE12, PKA, SH3, p53)[6]. 이러한 철자적 특징은 생의학분야 용어를 추출하는 특징으로 사용될 수 있으며 본 논문에서는 알파벳/숫자 관련 특징[1]과 기호관련 특징[2]을 정의하고 이를 철자적 특징 OF(t_i)로 표기한다. 예를 들어, OF(*ADE12*) = {'대문자포함', '숫자포함', '자모숫자', '대문자로 시작'}이다.

네 번째로 사전적 특징 TF(t_i)를 살펴보자. 새로운 개념의 전문용어를 만드는 가장 일반적인 방법은 기존의 전문용어에 수식어를 추가하는 방법이다[21]. 따라서 기존의 전문용어 정보는 새로운 전문용어를 파악하는데 도움이 된다. 사전적 특징은 중심어와 수식어를 전문분야사전과 일반분야사전을 이용하여 파악한다. 본 논문에서는

1) '모두 대문자' (e.g. DNA), '모두 소문자' (e.g. *motif*), '자모숫자' (e.g. *p.53*), '대문자 포함' (e.g. *c-Rel*), '대문자로 시작' (*Egfr*), '숫자' (e.g. *12, IV*) 등
2) '이음표 포함', 'apostrophe 포함', '괄호포함' 등

사전적 특징을 *BM, BH, GM, GH*으로 표현한다. 여기에서 *GM, GH*는 각각 일반분야 용어인 수식어, 일반분야 용어인 중심어를 각각 나타내며, *BM, BH*는 각각 생의학분야 용어인 수식어, 생의학분야 용어인 중심어를 나타낸다. 예를 들어, 용어 *cell surface T-cell receptor*에서 *cell surface*와 *T-cell receptor*가 기존의 생의학분야 전문용어일 경우(생의학분야 전문용어사전에 등재되어 있을 경우) TF(*cell surface T-cell receptor*)=(BM: *cell surface*, BH: *T-cell receptor*)로 표현된다.

다섯 번째로서 형태소적 특징을 살펴보자. 생의학분야 용어들은 그리스/라틴어원의 용어들이 많다. 이러한 용어들은 일반적으로 접두사, 어근, 접미사와 같은 형태소로 구성되며, 고유의 생의학분야 개념을 지칭한다[22]. 예를 들어 접두사 'leuko-', 접두사 'neuro-', 접미사 '-cyte', 접미사 '-itis'는 각각 '하얀색 (白色)', '신경', '세포', '통증'이라는 개념을 지칭한다. 그리고 이들 형태소로 구성된 'leukocyte (leuko+cyte)'는 '백혈구'라는 개념을 지칭하며, 'neuritis (neur(o)+itis)'는 '신경통'이라는 개념을 지칭한다. 따라서 용어들이 같은 그리스/라틴어원의 형태소를 포함한다면 해당 용어들은 개념적으로 연관관계를 가진다고 본다. 예를 들어 leukocyte(white blood cell), lymphocyte(a small white blood cell), thymocyte(lymphocyte within the thymus), hepatocyte (epithelial cell of liver), phagocyte(a cell that is capable of phagocytosis)는 '-cyte'를 접미사로 가지며 다양한 '세포'를 표현하는 생의학분야 용어이다. 본 논문에서는 그리스/라틴어원의 형태소들을 형태소적 특징으로 정의하고 MF(t_i)로 표기한다. 예를 들어 MF(*leuko-cyte*) = {leuko-, -cyte}이다.

이와 같은 다섯 가지 특징은 $LF(t_i)$, $SF(t_i)$, $OF(t_i)$, $TF(t_i)$, $MF(t_i)$라는 특징 집합으로 식 (5)와 같이 용어를 표현하는 잣대로 사용된다. 여기에서 $LF(t_i)$, $SF(t_i)$, $OF(t_i)$, $TF(t_i)$, $MF(t_i)$는 각각 t_i의 어휘적, 구문적, 철자적, 사전적, 형태소적 특징을 나타낸다. $LF(t_i)$, $OF(t_i)$, $TF(t_i)$, $MF(t_i)$는 용어를 구성하는 성분에 대한 특징이므로 용어의 내적특징(內的特徵)이며, $SF(t_i)$는 용어 외부에 있는 동사와의 구문관계를 나타내므로 외적특징(外的特徵)이다.

다섯 가지 특징들은 전문용어의 서로 다른 측면을 나타내기 때문에 전문용어선택을 위해서는 이들 특징들을 효과적으로 통합하는 방법이 필요하다. 3.3절에서는 기계학습방법을 이용한 방법론을 제시한다.

3.3 이진분류 기계학습을 통한 전문용어의 선택

이진분류 기계학습의 목적은 각각의 특징들이 전문용어와 비전문용어를 구분하는데 얼마나 효과적인지를 파악하기 위한 것이다. 본 논문에서는 Naïve Bayesian 분류기, 결정목록(Decision List), 지지벡터기계(Support Vector machine), 최대엔트로피모델(Maximum Entropy Model)을 이용하여 용어 목록의 용어를 이진분류한다. 이진분류는 각 용어 t_i에 대하여 가장 적합한 용어 유형 c_j (c_0=전문용어 혹은 c_1=비전문용어)를 할당하는 문제이다. 모든 용어는 3.2절에서 기술한 특징들에 의해 $\boldsymbol{f(t_i)=\{f_1(t_i),\cdots,f_m(t_i)\}}$와 같이 표현된다. 여기에서 m은 t_i에 나타나는 특징들의 총 수를 나타낸다.

Naive Bayesian 분류기에 의한 용어의 이진 분류는 주어진 용어 t_i에 대하여 확률 $p(c_j|t_i)$를 최대로 하는 용어유형 c_j를 파악하는 문제로 정의된다[23]. 이는 식 (6)과 같이 표현된다.

$$p(c_j|t_i)=\frac{p(c_j)\times p(t_i|c_j)}{p(t_i)}=\frac{p(c_j)\times p(f_1(t_i),\ldots,f_n(t_i)|c_j)}{\sum_{c_l\in C}p(c_l)\times p(f_1(t_i),\ldots,f_n(t_i)|c_l)} \quad (6)$$

여기에서 $p(c_j)$는 용어유형 c_j의 사전확률(a priori probability)을 나타내며 t_i는 $f(t_i)$로 표현된다. $p(f_1(t_i),\cdots f_n(t_i)|c_j)$의 추정에서 naïve Bayesian 분류기는 특징간의 독립가정을 사용한다. 독립가정에 의해 식 (6)은 식 (7)과 같이 곱의 열로 표현된다[24,25]. 식 (7)에서 매개변수 $p(f_k(t_i)|c_j)$와 $p(c_j)$는 기대최대화 알고리즘(EM 알고리즘)에 의해 추정된다.

$$p(c_j|t_i)=\frac{p(c_j)\times\prod_{k=1}^{n}p(f_k(t_i)|c_j)}{\sum_{c_l\in C}\left[p(c_l)\times\prod_{k=1}^{n}p(f_k(t_i)|c_l)\right]} \quad (7)$$

두 번째 기계학습방법은 결정목록(decision lists: DL)이다. 결정목록은 귀납적 규칙학습방법의 하나이다. 특정 특징에 대하여 특정 유형을 할당하는 결정목록은 학습을 통하여 반복적으로 구축되며, 학습을 통해 해당 결정목록의 신뢰도가 결정된다[26-28]. 결정목록 기계학습방법은 결정목록에 있는 각 항목이 결정하려는 유형에 기여하는 신뢰도를 가지고 있다. 즉, 신뢰도는 결정목록이 용어를 특정 용어유형으로 분류할 때 얼마나 효과적인지를 나타내는 척도이다. 주어진 t_i의 특징 $f_k(t_i)$와 용어 유형 c_j에 대하여 $f_k(t_i)$에 의해 t_i를 c_j로 분류하는 결정목록 $DL(f_k(t_i), c_j)$의 신뢰도는 식 (8)의 로그우도(log-likelihood)로 계산된다. 만약 $DL(f_k(t_i), c_j)$의 신뢰도 값이 0보다 작은 값이 되면 해당 결정목록은 t_i를 c_j로 분류할 때 사용되지 않는다. 결정목록을 이용한 용어의 이진 분류에서는 용어 t_i에 의해 생성되는 결정목록 중 신뢰도가 가장 높은 결정 목록만이 t_i를 이진분류하는데 사용된다.

$$\log\left(\frac{p(c_j|f_k(t_i))}{\sum_{l\neq j}p(c_l|f_k(t_i))}\right) \quad (8)$$

학습된 결정목록은 조건부 규칙으로 표현된다. 표 2는 본 논문에서 사용한 결정목록의 예를 나타낸다. 예를 들어 용어 "NF-kappa B receptor", "NF-kappa B activation", "NF-kappa B inhibitor"는 DL_1에 의해 c_0으로 분류되고, "play a significant role", "play an important role", "play a role"에서의 용어 "significant role", "important role", "role"은 DL_6에 의해 c_1으로 분류된다.

세 번째 기계학습방법은 지지벡터기계(SVM)이다. SVM은 이진분류를 수행하는 기계학습방법으로 모든 학습데이타는 특징으로 표현되는 벡터공간에 표현된다[29-31]. SVM에서의 학습은 벡터공간 내에서 분류경계(hyperplane)와 가장 가까운 거리에 있는 학습데이타(지지벡터)와의 최소거리를 최대화시키는 것을 목표로 한다[31]. 즉 최대여백 분류경계(maximal margin hyperplane)를 찾는 것을 목표로 한다. SVM에서는 지지벡터만이 분류경계를 구성하는데 사용된다. 분류경계는 커널함수(kernel function)로 표현되며, 커널함수는 벡터공간에서의 벡터간의 스칼라 곱(dot product)으로 표현된다. SVM에서 가장 널리 사용되는 커널함수로는 선형커널(linear kernel), 다항커널(polynomial kernel), RBF커널(RBF kernel)이 있다. 이 중 선형커널은 가장 빠르고 간단한 커널함수로서 많은 SVM기반 응용에서 강력한 성능을 나타내었다[31,32]. 본 논문에서도 *SVM light*[32]에 구현된 선형커널함수를 이용하여 지지벡터기계를 학습하고 용어의 이진 분류를 수행하였다.

t_i의 특징 벡터를 $x_i=<f_1(t_i),\cdots f_m(t_i)>$로 정의하면 선형커널함수에 기반한 용어의 이진 분류는 식 (9)와 같이 표현된다. 학습데이타는 $S=\{(x_1,y_1), (x_2, y_2),\cdots,(x_n, y_n)\}$로 표현된다. 여기에서 $y_i \in Y=\{+1, -1\}$는 출력공간 (용어 유형: +1은 전문용어, -1은 비전문용어)을 나타낸다. x_i의 가중치 $\alpha_i \neq 0$인 경우 x_i를 지지벡터(support vector)라 정의한다[31,32]. 지지벡터기계를 이용한 용어의 이진 분류를 도식화하면 그림 2와 같다.

$$h(\mathbf{x}) = sign\{w\cdot x + b\}, w = \sum \alpha_i y_i x_i,\ \alpha_i \geq 0 \quad (9)$$

네 번째 기계학습방법으로 최대 엔트로피 모델(maximum entropy model)을 실험하였다. 최대 엔트로피 모델은 비동형 정보(heterogeneous information)를 효과적으로 통합하는 확률적 모델링 기법이다. 최대엔트로피 모델에서 확률 사건은 하위사건으로 세분화되고, 하위사건으로부터 도출될 수 있는 중복되는 특징을 이용하여 유연한 확률적 모형화가 가능하다[33,34]. 확률 사건 $ev=<te,he>$로 표현되며, 여기에서 te는 대상 사건(target event)을 나타내고, he는 문맥사건(history event)을 나타낸다. 이진분류에서 대상사건은 이진분류의 결과인 용어 유형이 되며, 문맥사건은 용어 유형을 결정하기 위해 필요한 특징들을 나타낸다. 특징 함수 $feature_i(ev)$는 이진값(binary value)을 가지는 함수로서 특징 함수가 특정조건을 만족시키면 활성화되어 $feature_i(ev)=1$의 값을 가지고, 그렇지 않을 경우 0의 값을 가진다[33,34]. ev는 활성화되는 특징함수($feature_i(ev)$)의 집합으로 표현된다.

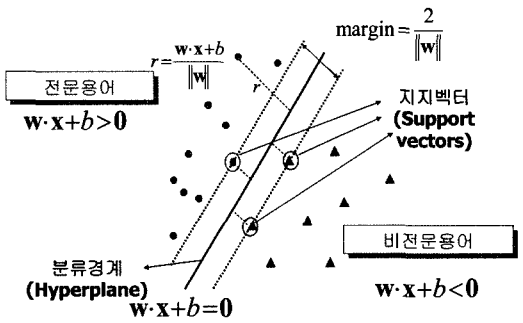

그림 2 지지벡터기계를 이용한 이진분류의 도식화

최대엔트로피모델 p_M은 로그선형모델(log-linear model)로서 사건 ev에 대한 조건부확률을 식 (10)과 같이 추정한다. 여기에서 $\tau(he)$는 he로부터 도출할 수 있는 te의 집합을 나타내고, β_i는 특징 함수 $feature_i(ev)$에 대한 가중치를 나타내며, Z_h는 정규화 계수이다. 최대엔트로피모델은 주어진 특징함수와 학습데이타를 이용하여 우도(likelihood)를 최대화하는 확률분포를 도출한다.

$$p_M(te|he) = \frac{1}{Z_h}\prod_i \beta_i^{feature_i(te,he)};\ Z_h = \sum_{te'\in\tau(he)}\prod_i \beta_i^{feature_i(te',he)} \quad (10)$$

용어의 이진 분류에 사용된 특징 함수는 식 (11)과

표 2 결정목록의 예

$DL(f_k(t_i), c_j)$	신뢰도	$f_k(t_i)$	c_j
DL_1	14.82	*NF-kappa*가 t_i에서 수식어	c_0 = 전문용어
DL_2	14.22	*t-cell*이 t_i에서 중심어	c_0 = 전문용어
DL_3	14.18	*activate*의 목적어가 t_i	c_0 = 전문용어
DL_4	14.64	*result*가 t_i에서 중심어	c_1 = 비전문용어
DL_5	13.18	*important*가 t_i에서 수식어	c_1 = 비전문용어
DL_6	13.27	*play*의 목적어가 t_i	c_1 = 비전문용어

같이 표현된다. 특징 함수의 활성화 조건은 결정목록과 유사하게 조건부 규칙으로 표현된다. 특징 함수의 활성화 조건은 *he*와 *te*의 조합으로 표현된다. 예를 들어 식 (11)의 $feature_j(te, he)$는 *he*가 '*t-cell*이 t_i의 중심어'이고 *te*가 전문용어일 때 활성화된다.

$$feature_i(te,he)=\begin{cases}1 & \text{if } he='activate\text{의 목적어가 } t_i' \\ & \text{and } te=\text{전문용어} \\ 0 & \text{otherwise}\end{cases}$$

$$feature_j(te,he)=\begin{cases}1 & \text{if } he='t-cell\text{이 } t_i\text{의 중심어}' \\ & \text{and } te=\text{전문용어} \\ 0 & \text{otherwise}\end{cases}$$

$$feature_k(te,he)=\begin{cases}1 & \text{if } he='result\text{가 } t_i\text{의 중심어}' \\ & \text{and } te=\text{비전문용어} \\ 0 & \text{otherwise}\end{cases} \quad (11)$$

본 논문에서는 최대엔트로피모델 기반 용어의 이진 분류를 위하여 Maximum entropy modeling toolkit [35]을 이용하였다.

4. 실험

4.1 실험데이타

본 논문에서는 영어 생의학분야 전문용어를 추출하기 위하여 생의학분야 영어 논문 초록을 포함하는 GENIA 코퍼스[36]를 사용하였다. 사용된 코퍼스는 GENIA corpus 2.01이며, 총 670여 개 문서를 포함하고 있다. 기계학습방법에 기반한 용어 분류의 초기학습을 위해 생의학분야 전문용어사전인 *UMLS Specialist lexicon*과 *UMLS Metathesaurus*[37]을 사용하였으며 일반분야 사전으로는 Brill 태거[38]의 명사사전을 사용하였다.

실험은 경계인식 실험과 순위화 실험을 수행하였다. 경계인식 실험에서는 각 기계학습방법에 따른 실험, 특징에 따른 실험을 수행하였으며, 순위화 실험에서는 기존 연구와의 비교 실험을 수행하였다. 기존연구와 비교 실험을 순위화 성능 평가 실험에 한정시킨 것은 기존 연구들이 경계 인식에 대한 결과를 명확히 제시하지 못하기 때문이다. 경계인식 실험의 평가를 위하여 식 (12)의 정확률, 재현율, F-값을 사용한다[39]. 정확률은 주어진 전문분야에 대한 용어라고 판단된 것 중 옳은 비율을 의미하고, 재현율은 문서에 나타나는 모든 전문용어 중 인식된 전문용어의 비율을 나타낸다. F-값은 정확률과 재현율을 통합적으로 나타내는 평가 기준이다.

$$\text{정확률}=\frac{\text{올바르게 분류된 용어의 수}}{\text{분류된 용어의 수}}$$

$$\text{재현율}=\frac{\text{인식된 전문용어의 수}}{\text{문서에 나타난 전문용어의 수}}$$

$$\text{F}-\text{값}=\frac{2\times\text{정확률}\times\text{재현율}}{\text{정확률}+\text{재현율}} \quad (12)$$

순위화 실험에서는 제안한 전문용어 인식법과 [5,8,10]의 기법을 비교하였다. 기존 연구의 전문용어 인식 결과는 용어에 용어유형을 할당한 것이 아니라 점수함수에 의해 순위화된 용어 목록이므로 용어목록을 전문용어와 비전문용어로 분류하는 본 논문의 기법과 직접적인 비교가 어렵다. 비교평가를 위하여 본 논문에서 제안한 기법의 전문용어 인식 결과도 순위화된 용어 목록으로 변환하였다. 제안한 기법에 의해 생성된 전문용어 인식 결과는 여러 가지 기계학습방법에 의해 전문용어로 판단될 확률이나 이진분류의 신뢰도를 이용하여 순위화된 용어 목록으로 변환되었다. 성능평가는 정보검색분야에서 사용하는 11포인트 평균정확률을 이용하였다[39]. 11포인트 평균 정확률은 재현율이 0%, 10%, 20%, 30%, ..., 90%, 100% 지점일 때의 정확률을 계산한 뒤, 11개 지점의 정확률을 평균하여 나타낸다. 따라서, 각 재현율의 지점에서 높은 정확률을 보일 경우 11포인트 평균 정확률이 높게 나타난다. 이는 순위화된 용어 목록의 상위에 전문용어가 많이 존재할수록 높은 11포인트 평균 정확률을 얻을 수 있음을 의미한다. 본 논문에서는 11포인트 평균 정확률을 구하기 위하여, 용어 16,730여 개 중 전문용어로 판별되는 11,320개의 용어를 모두 인식하였을 때 재현율이 100%라고 가정한다. 그리고 재현율이 0% 지점에서 100%의 정확률을 가진다고 가정한다[39]. 이를 기준으로 재현율 0%~100% 지점을 찾아 해당 지점에서의 정확률을 계산한다. 이는 식 (13)과 같이 표현된다.

$$11pt-avg=\frac{1}{11}\times\sum_{i=0}^{10}precision(recall_{i\times0.1}) \quad (13)$$

또한 기존의 연구에서 전문용어 인식의 평가방법으로 사용한 식 (14)의 부분정확률(segment precision)을 이용하여 비교 평가하였다[5,8,10]. 부분 정확률은 순위화된 용어 목록에서 상위와 하위에서의 전문용어의 분포를 나타낸다. 따라서 전문용어 인식법이 생성한 순위화된 용어 목록에서의 전문용어와 비전문용어 간의 경계를 유추하는 척도로 사용될 수 있다.

$$\text{부분정확률}=\frac{\text{부분내 전문용어수}}{\text{부분내 용어의 수}} \quad (14)$$

4.2 경계인식 실험결과

4.2.1 기계학습방법에 따른 전문용어 인식 성능

표 3은 기계학습방법에 따른 전문용어 인식 성능을 나타낸다. 표에서 *Baseline*은 사전만을 이용하여 전문용어를 추출한 결과를 나타내고, *NBC, DL, SVM, MEM*은 각각 *Naïve Bayesian* 분류기, 결정목록, 지지벡터기계, 최대엔트로피모델을 이용하여 용어를 이진 분류하였을 때의 성능을 나타낸다.

표 3 기계학습방법에 따른 실험 결과

	정확률	재현율	F-값
Baseline	95.05%	28.34%	43.66%
NBC	81.57%	87.37%	84.37%
DL	78.42%	89.67%	83.67%
SVM	85.52%	87.79%	86.64%
MEM	86.59%	90.24%	88.38%

실험결과에서 *Baseline*은 높은 정확률과 낮은 재현율을 나타낸다. 이는 사전을이용할 경우 비교적 정확하게 전문용어를 추출할 수 있지만, 파악할 수 있는 전문용어와 비전문용어의 범위가 매우 한정되어 있다는 것을 나타낸다. 본 논문의 기법은 학습을 통하여 사전에 등재되어 있지 않는 전문용어를 문서에서 자동으로 찾아내는 것을 목표로 한다. 즉, 사전만으로 파악하지 못한 전문용어를 사전에 등재되어 있는 전문용어와 해당 전문용어가 문서에서 나타나는 양상을 파악하여 전문용어를 추출한다. 따라서 정확률의 감소를 최소화하면서 재현율을 높이는데 목적이 있다. 전체적으로 본 논문의 기법은 9~17%의 정확률 감소로 59~72%의 재현율 향상을 나타냄을 알 수 있다.

기계학습방법별 성능 측면에서는 지지벡터 기계와 최대엔트로피 모델을 이용한 경우가 높은 성능을 나타내며, Naïve Bayesian 분류기와 결정목록을 이용한 경우가 비교적 낮은 성능을 나타낸다. Naïve Bayesian 분류기의 성능이 비교적 낮은 이유는 특징간의 독립을 가정하기 때문으로 분석된다. 결정목록의 성능이 비교적 낮은 이유는 용어의 특정 특징에 의존하여 분류를 수행하기 때문으로 분석된다. 즉 용어의 특정 특징을 이용한 결정목록이 잘못된 결과를 도출할 경우, 해당 용어의 다른 특징을 고려하지 못하여 오류를 바로잡기 힘들기 때문이다.

4.2.2 특징에 따른 전문용어 인식 성능

표 4는 특징에 따른 전문용어인식 성능을 나타낸다. 표 4에서 OF, MF, LF, SF, DF는 각각 철자, 형태소, 어휘, 구문, 사전 특징을 나타낸다. 실험에서는 하나의 특징만을 이용한 경우와 네 가지 특징을 이용한 경우 그리고 모든 특징을 이용한 경우의 성능을 비교 평가하였다. 하나의 특징만을 이용한 경우와 네 가지 특징을 이용한 전문용어 인식 성능 평가는 각 특징이 실제 전문용어 선택에 기여하는 정도를 파악하기 위한 것이며, 모든 특징을 사용한 경우는 각 분류기에서 모든 특징을 사용할 경우와 그렇지 않았을 경우의 성능을 비교하기 위한 것이다.

실험결과에서 하나의 특징만을 이용한 F-값 결과는 DF가 모든 분류기에서 일관되게 가장 높은 성능을 나타내었다. 그리고 하나의 특징만을 사용했을 때 가장 낮은 성능을 나타내는 경우는 분류기마다 다르게 나타나는데 NBC에 의해서는 LF가 DL에 의해서는 SF가 SVM과 MEM에 의해서는 MF가 가장 낮은 성능을 나타내었다. 네 가지 특징을 이용한 경우에는 LF나 DF를 사용하지 않은 경우가 가장 낮은 성능을 나타내며, MF나 SF를 사용하지 않은 경우가 가장 높은 성능을 나타내었다. 그리고 모든 특징을 사용할 경우 가장 좋은 성능을 나타내며, 이를 통하여 다섯 가지 특징이 효과적으로 통합되어 전문용어 선택에 사용되었음을 알 수 있다.

4.3 순위화 실험 결과

본 절에서는 기존 연구와 본 논문에서 제안한 기법과의 성능비교를 수행한다. 비교대상이 되는 기존 연구는 *Freq*[8], *C-value*[5], *FGM*[10]이다. 표 5는 기존 연구와의 비교실험 결과를 11포인트 평균 정확률로 평가한 결과이다.

표 5에서 IDEAL은 이상적 전문용어 인식 체계를 나타내며 모든 전문용어가 순위화된 용어 목록의 상위에 존재하기 때문에 100%의 11포인트 평균 정확률을 나타낸다. 따라서 11포인트 평균정확률이 100%에 가까울수록 효과적인 전문용어 인식 체계라고 평가할 수 있다.

표 4 특징에 따른 실험 결과

OF	MF	LF	SF	DF	NBC	DL	SVM	MEM
√					77.36%	67.79%	79.53%	83.95%
	√				77.29%	77.27%	77.90%	82.21%
		√			70.60%	79.15%	81.02%	86.94%
			√		74.30%	44.14%	85.44%	82.80%
				√	79.06%	79.96%	86.44%	82.80%
√	√	√	√		82.47%	84.84%	81.51%	86.88%
√	√	√		√	83.02%	85.15%	86.63%	88.64%
√	√		√	√	79.04%	84.84%	86.34%	88.60%
√		√	√	√	83.84%	85.10%	86.49%	88.67%
	√	√	√	√	82.25%	84.96%	86.62%	88.60%
√	√	√	√	√	84.34%	86.25%	86.25%	89.92%

기존 연구들의 11포인트 평균 정확률은 약 73%~77%로 IDEAL의 성능과 많은 격차를 나타낸다. 또한 순위화된 용어 목록의 상위로 판단되는 20% 재현율 지점에서조차 비교적 낮은 성능을 나타낸다. 이는 단순 통계적 방법에 의한 용어목록의 순위화기법은 전문용어와 비전문용어를 구분하는데 한계가 있음을 나타낸다. 이와 반대로 본 논문에서 제안한 기법의 11포인트 평균정확률은 약 89%~92%로 IDEAL의 성능에 근접해 있음을 알 수 있다. 기존 기법과 달리 80% 재현율 지점에서도 90%에 가까운 성능을 나타낸다. 이를 통해 본 논문의 기법이 기존 연구에 비해 약 16~26%의 성능향상을 나타냄을 알 수 있다.

기존 연구와 본 연구에서 제안한 기법에 의해 순위화된 용어 목록에서 전문용어의 분포를 살펴보기 위해 부분정확률을 이용하여 비교평가를 수행하였다. 부분 정확률은 순위화된 용어 목록을 상위에서 하위까지 몇 개의 부분으로 등분하였을 때 각 부분에서의 전문용어포함비율을 나타낸다. 본 논문에서는 순위화된 용어목록를 5개의 부분으로 등분하였다. 따라서 부분 1, 2, 3, 4, 5는 각각 상위 20%, 20%~40%, 40%~60%, 60%~80%, 80%~100%를 나타낸다.

표 6은 기존 연구의 부분 정확률과 제안기법의 부분 정확률을 나타낸다. 결과에서 전체적으로 부분 1에서 부분 5로 갈수록 전문용어 포함 비율이 낮아짐을 알 수 있다. 즉, 순위화된 용어 목록의 상위에는 전문용어의 밀도가 높으며, 하위에는 비전문용어의 밀도가 높다는 것을 알 수 있다. IDEAL 시스템은 순위화된 용어목록의 상위에 모든 전문용어를 위치시키기 때문에 부분 1, 2, 3에서는 100%의 부분 정확률을 나타내며, 부분 4에서는 42%의 부분정확률을 나타낸다. 여기에서 IDEAL 시스템은 부분 4에서 전문용어와 비전문용어의 경계를 설정한다는 것을 알 수 있다.

표 5와 6을 통하여 11포인트 평균정확률과 부분정확률은 밀접한 관계가 있다는 것을 알 수 있다. 전문용어 인식 체계가 상위에 보다 많은 전문용어를 위치시킬수록(부분 1,2의 부분정확률이 높을수록), 11포인트 평균정확률이 높을 뿐만 아니라 각 재현율 지점에서의 정확률도 높음을 알 수 있다. 따라서 부분 1과 2에서 낮은 부분 정확률을 나타내고, 부분 5에서 높은 부분 정확률을 나타내는 기존 연구들은 결과적으로 낮은 11포인트 평균정확률을 나타낸다. 반대로 본 논문에서 제안한 기법은 부분 1, 2에서 비교적 높은 부분정확률을 나타내며, 부분 5에서는 낮은 부분정확률을 나타낸다. 또한 부분 1에서 부분 5에 걸쳐 IDEAL 시스템과 유사한 전문용어 분포를 나타낸다.

실험을 통하여 본 논문의 기법이 기존 연구에 비해 효과적임을 알 수 있었다. 또한 본 논문의 기법에 의해 추출된 전문용어의 품질이 우수함을 알 수 있었다.

5. 결 론

본 논문에서는 다양한 전문용어의 특징과 기계학습에 기반한 전문용어 인식 기법을 제안하였다. 본 논문의 기법은 '용어 추출'과 '전문용어 선택'을 통하여 전문용어를 인식하였다. 기존의 전문용어 인식법의 문제점인 순위화 문제와 경계인식 문제를 효과적으로 해결하였다. 순위화 문제는 용어의 철자적, 형태소적, 어휘적, 구문적, 사전적 특징을 이용하였으며, 기계학습방법으로 이들을 효과적으로 통합하였다. 용어 목록에서의 전문용어와 비전문용어간 경계인식문제는 기계학습에 기반한 이진분류법으로 해결하였다. 본 논문의 기법은 경계인식측

표 5 기존 연구와의 비교실험 결과: 11포인트 평균정확률

재현율	Freq	C-value	FGM	NBC	DL	SVM	MEM	IDEAL
0%	100%	100%	100%	100%	100%	100%	100%	100%
20%	78.3%	75.1%	81.4%	93.6%	96.4%	94.6%	94.7%	100%
40%	75.1%	71.7%	75.7%	92.9%	92.7%	92.2%	95.2%	100%
60%	70.4%	66.4%	73.4%	92.7%	88.5%	90.9%	94.9%	100%
80%	68.8%	66.5%	70.0%	89.9%	83.3%	88.9%	93.2%	100%
100%	66.4%	66.4%	66.4%	66.5%	66.5%	67.3%	66.4%	100%
11pt-avg	74.6%	72.8%	**76.8%**	90.6%	88.8%	90.2%	**92.3%**	100%

표 6 기존 연구와의 비교실험결과: 부분정확률

Segment	Freq	C-val	FGM	NBC	DL	SVM	MEM	IDEAL
1	78.75%	73.97%	79.62%	92.92%	95.38%	93.65%	**95.60%**	100%
2	70.31%	66.98%	70.28%	93.21%	82.66%	89.16%	**94.47%**	100%
3	63.39%	58.38%	68.21%	83.05%	74.58%	83.92%	**87.94%**	100%
4	61.93%	66.59%	61.00%	53.14%	68.27%	60.09%	46.05%	42%
5	57.60%	66.11%	52.88%	9.86%	11.06%	9.43%	8.31%	0%

면에서 78~86%의 정확률과 87%~90%의 재현율을 나타내었으며, 순위화 측면에서 89%~92%의 11포인트 평균정확률을 나타내었다. 또한 기존 연구보다 최고 26%의 성능향상을 보였다..

향후 전문용어 인식법의 성능향상을 위하여 의미정보를 특징으로 고려하는 전문용어 인식법[9]에 대한 연구가 추가적으로 수행되어야 할 것이다. 또한 메모리기반 기계학습과 같은 다른 기계학습방법을 이용한 전문용어 인식에 대한 연구도 병행될 예정이다.

참 고 문 헌

[1] Ananiadou, S., A Methodology for Automatic Term Recognition, In Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, pp. 1034-1038, 1994.

[2] Bourigault D., LEXTER, a Terminology Extraction Software for Knowledge Acquisition from Texts, *9th Knowledge Acquisition for Knowledge-Based Systems Workshop*, 1995.

[3] Blaschke C, Andrade MA, Ouzounis C and Valencia A., Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions. *ISMB99*, 60-67, 1999.

[4] Damerau F. (1993) Generating and Evaluating Domain-Oriented Multi-Word Terms from Texts. *Information Processing & Management*, 29(4), 433-448, 1993.

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

[6] Fukuda, K. and A Tamura and T Tsunoda and T Takagi, Toward information extraction: identifying protein names from biological papers. In Proceeding of the Pacific Symposium on Biocomputing (PSB98), 707-718, 1998.

[7] Jacquemin, C., Judith L.K. and Evelyne, T., "Expansion of Muti-word Terms for indexing and Retrieval Using Morphology and Syntax," 35th Annual Meeting of the Association for Computational Linguistics, pp. 24-30, 1997.

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

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

[10] Nakagawa, Tataunori Mori, Automatic Term Recognition based on Statistics of Compound Nouns, In Proceeding of the First Workshop on Computational Terminology Computerm02 in COLING02, pp. 29-35, 2002.

[11] Oh Jong-Hoon, Juho Lee, Kyung-Soon Lee, Key-Sun Choi, "Japanese Term Extraction Using Dictionary Hierarchy and Machine Translation System," Terminology, 6(2), John Benjamins Publishing Company, pp. 287-311, 2000.

[12] 오종훈, 이경순, 최기선, "분야간 유사도와 통계기법을 이용한 전문용어의 자동 추출", 정보과학회 논문지, 제29권 제3,4호, pp. 258-269, 2002.

[13] Pum-Mo Ryu, Key-Sun Choi, "Determining the Specificity of Terms based on Information Theoretic Measures," Proceedings of CompuTerm 2004, 3rd International Workshop on Computational Terminology, Coling 2004, pp.87-90, 2004.

[14] Friedman, C., Kra, P., Yu, H., Krauthammer, M. and Rzhetsky, A., GENIES: a natural-language processing systems for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17, S74-S82, 2001.

[15] Jessen, T.-K., Laegreid, A., Komorowski, J. and Hovig, E, A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.*, 28, 21-28, 2001.

[16] Ono T., H. Hishigaki, A. Tanigami, T. Takagi, Automated Extraction of Information on Protein-Protein Interactions from the Biological Literature. *Bioinformatics*, 17:155-161, 2001.

[17] Rindflesch, T.C., L.Tanabe, J.N.Weinstein, and L.Hunter, Edgar: Extraction of drugs, genes, and relations from the biomedical literature. *In Proc. Pacific Symposium on Biocomputing*, pages 514-525, 2000.

[18] Thomas, J., Milward, D., Ouzounis, C., Pulman, S. and Carroll, M., Automatic extraction of protein interactions from scientific abstracts. *PSB'00*, 541-551, 2000.

[19] Yakushiji Akane, Y. Tateisi, Y. Miyao, and J. Tsujii, Event extraction from biomedical papers using a full parser, In proceedings of PSB01, 408-419, 2001.

[20] Yandell, M.D. and Majoros, W.H., Genomics and natural language processing. Nat.Rev. Genet., 3, 601-610, 2002.

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

[22] Janson Barbara and Med Cohen, Medical Terminology: An Illustrated Guide, Lippincott Williams & Wilkins, 2003.

[23] Lewis D.D., Naïve (Bayes) at forty: The independence assumption in information retrieval. In ECML-98, 1998.

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

[25] Nigam K., Andrew McCallum, Sebastian Thrun

and Tom Mitchell, Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning, 39(2/3). pp. 103-134, 2000.
[26] Mooney R.J. and M.E. Califf., Induction of First-Order Decision Lists: Results on Learning the Past Tense of English Verbs, Artificial Intelligence Research, Vol. 3, pp. 1-24, 1995.
[27] Rivest, Ronald L., Learning Decision Lists, Machine Learning, 2(3), pp. 229-246, 1987.
[28] Yarowsky, D., Unsupervised Word Sense Disambiguation Rivalling Supervised Methods, In Proceeding of the 33rd Annual Meeting of the Association for Computational Linguistics, pp. 189-196, 1995.
[29] Burges. C.J.C., A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):1-47, 1998.
[30] Cristianini N. and J. Shawe-Taylor, An Introduction to Support Vector Machines. Cambridge University Press, 2000.
[31] Vapnik. V., Statistical Learning Theory. John Wiley & Sons, 1998.
[32] Joachims Thorsten, Learning to Classify Text Using Support Vector Machines, Kluwer, 2002.
[33] Berger A., S. Della Pietra, and V. Della Pietra, A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39-71, 1996.
[34] Miyao, Yusuke and Jun'ichi Tsujii, Maximum Entropy Estimation for Feature Forests. *In the Proceedings of Human Language Technology Conference* (HLT 2002), 2002.
[35] Zhang Le., Maximum Entropy Modeling Toolkit for Python and C++, http://www.nlplab.cn/zhangle/, 2004.
[36] Ohta and Yuka Tateisi and Hideki Mima and Jun'ichi Tsujii, GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain *In Proceeding of he Human Language Technology Conference*, 2002.
[37] NLM,, Unified Medical Knowledge System (UMLS), 2003.
[38] Brill, E., Transformation-Based error-driven learning and natural language processing: a case study in part of speech tagging. *Computational Linguistics*, 1995.
[39] Salton, G. and McGill, M., Introduction to Modern Information Retrieval, New-York: McGraw-Hill, 1983.

오 종 훈

1998년 성균관대학교 정보공학과 졸업(학사). 2000년 한국과학기술원 전산학과 졸업(공학석사). 2005년 한국과학기술원 전산학과 졸업(공학석사). 2005년~현재 Expert researcher, Computational Linguistics Group, NICT(情報通信研究機構), Japan. 관심분야는 자연언어처리, 기계번역, 정보검색, 전문용어 등

최 기 선

정보과학회논문지 : 소프트웨어 및 응용
제 33 권 제 7 호 참조