분야간 유사도와 통계기법을 이용한 전문용어의 자동 추출

(Automatic Term Recognition using Domain Similarity and Statistical Methods)

오 종 훈 † 이 경 순 †† 최 기 선 †††
(Jong-Hoon Oh) (Kyung-Soon Lee) (Key-Sun Choi)

요 약 지금까지 전문용어를 자동으로 추출 (Automatic Term Recognition: ATR)하기 위한 많은 연구들이 있어 왔다. 이들 연구들은 주로 문서 내의 용어의 빈도수와 같은 단순한 통계정보를 이용하여 전문용어를 추출하였다. 하지만 전문분야의 기계가독형 사전의 구축으로 인하여 전문용어를 추출하는 데 있어 전문분야 사전의 사용이 가능하게 되었다. 본 논문에서는 이러한 기계가독형 전문분야 사전들을 이용하여 사전 간의 계층관계를 구축하고 이를 이용하여 전문용어를 추출하는 방법을 제시한다. 또한 전문용어 사전에서 나타나지 않는 전문용어를 추출하기 위하여 용어의 빈도수, 외래어 및 외국어, 내포관계 등을 포함한 통계기법을 이용한다. 본 논문에서 제안하는 기법은 기존의 방법에 비해 좋은 성능을 나타내었다.

키워드 : 전문용어추출, 전문용어, 분야유사도, 사전간 계층관계, 통계기법

Abstract There have been many studies of automatic term recognition (ATR) and they have achieved good results. However, there are scopes to improve the performance of extracting terms still further by using the additional technical dictionaries. This paper focuses on the method for extracting terms using the hierarchy among technical dictionaries. Moreover, a statistical method based on frequencies, foreign words, and nested relations assists extracting terms which do not appear in dictionaries. Our method produces relatively good results for this task.

Key words : ATR, Term, Domain similarity, Dictioanry hierarchy, Statistical method

1. 서 론

지금까지 통계정보를 이용하여 용어를 자동으로 추출 (Automatic Term Recognition: ATR)하는 많은 연구들이 있어 왔다[1, 2, 3, 4, 5]. 이들 연구들이 비교적 좋은 성능을 보였지만, 전문용어 사전에 나타나는 기존의 전문용어 정보와 같은 여러 다른 정보를 이용하여 성능의 향상을 이룰 수 있는 여지는 여전히 남아 있다. 용어추출 분야에 있어 기계가독형 사전이 사용되기 어려웠던 것은 사전을 구축하는 데 있어 상당한 노력이 필요했기 때문이다. 하지만 기계가독형 언어자원을 구축하기 위한 도구들의 점진적인 개발은 전문용어추출 분야에 이러한 사전을 이용할 수 있는 새로운 계기를 마련하고 있다.

하지만, 전문용어는 계속적으로 생성되고 사전에 등재되지 않은 경우도 많기 때문에 전문용어사전 그 자체만으로는 전문용어를 효율적으로 추출할 수 없다. 따라서, 전문용어사전 정보 뿐만 아니라 문서내의 통계정보와 같은 용어의 정보도 여전히 전문용어를 자동적으로 추출하는 데 중요한 요소가 될 수 있다. 전문용어사전 정보는 전문용어 자동추출 기법에 사용되는 기존 전문용어의 언어자원으로서 사용될 수 있다. 예를 들어, 컴퓨터 분야 용어인 '분산 데이타베이스'는 기존의 용어인 '분산'과 '데이타베이스'에 의해 만들어졌다.

한 분야의 전문용어와 이를 지칭하는 개념은 관련된 다른 분야의 용어로부터 비롯된 것도 많다. 예를 들어, 전자 분야의 단어인 '지리 정보 시스템 (GIS : Geographical Information System)'은 전자 분야의 사

† 비 회 원 : 한국과학기술원 전산학과
rovellia@world.kaist.ac.kr
†† 비 회 원 : 일본 NII(National Institute of Informatics)
kslee@world.kaist.ac.kr
††† 종신회원 : 한국과학기술원 전산학과 교수
kschoi@world.kaist.ac.kr
논문접수 : 2001년 4월 17일
심사완료 : 2002년 1월 3일

전에만 존재하지만 컴퓨터 분야에서도 사용되는 전문용어이다. 이처럼 전문용어는 기존의 용어로부터 새로이 생성될 뿐만 아니라 유사한 분야의 용어를 이용하는 경우도 있기 때문에 전문용어를 효율적으로 추출하기 위해서는 이러한 전문분야들간의 상호 연관성을 고려할 필요가 있다. 본 논문에서는 정보 검색 분야에서 사용되는 계층적 클러스터링 방법을 이용하여 전문분야간의 관계를 구축하여 전문용어 추출에 이용하고자 한다. 계층적 클러스터링 방법을 이용해 사전간 (분야간)의 계층관계를 구축할 수 있으며, 이를 통하여 분야간의 연관성을 유추할 수 있다. 예를 들어, 전자 분야의 용어는 다른 분야에 비하여 컴퓨터 분야의 용어와 일치되는 수가 많기 때문에 계층적 클러스터링 방법에 의해 구축된 트리의 단말노드사이에서 관계를 가지게 된다. 이를 통하여 전자 분야와 컴퓨터 분야는 아주 밀접한 관계를 가진다는 것을 유추할 수 있다. 따라서, 다른 분야에 비해서 전자 분야 전문용어사전의 용어는 컴퓨터 분야의 용어가 될 확률이 높게 된다[6].

본 논문에서는 이러한 특성을 반영하여 특정 분야 문서에서 나타나는 해당 분야의 전문용어를 효율적으로 추출하는 방법론을 제안하고자 한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대하여 기술하고, 3장에서는 본 논문에서 제안하는 방법론들을 자세히 설명한다. 4장에서는 실험 및 결과를 제시하고 5장에서는 본 논문의 결론을 맺는다.

2. 관련연구

2.1 빈도수에 기반한 전문용어 추출[2, 4, 5]

빈도수에 기반한 전문용어 추출 연구는 자동용어추출(ATR) 분야에서 가장 일반적이고 간단한 방법으로 사용되며, 분야에 독립적이고 다른 자원이 필요 없다는 장점을 가진다. 이들 연구에서는 문서에 대하여 형태소 분석을 하고 언어필터 (linguistic filter)라고 하는 명사구에 대한 정규표현을 이용하여 문서에서 정규표현에 해당하는 명사구를 추출한다. 해당 명사구는 빈도수로 가중치가 부여되며 이를 이용해 전문용어를 추출한다. 예를 들어 [4]의 연구에서는 명사, 관형사, 전치사로 구성된 언어필터인 "((A|N)+| ((A|N)* (N P)?) (A|N)*)N"를 사용하였다. 여기서 A는 관형사, N은 명사, P는 전치사를 각각 나타낸다. 또한 식 (1)에 의해 해당 명사구에 대한 가중치를 계산하였다.

$$Score(a) = f(a) \quad (1)$$

여기서 a는 언어필터에 의해 추출된 명사구를 나타내고 $f(a)$는 a의 문서내 빈도수를 나타낸다.

이들 방법들은 문서에서 자주 나타나는 고정된 형태의 용어에 대하여 비교적 좋은 성능을 나타낸다. 하지만 문서에서 빈도수가 작게 나타나는 용어의 경우 제대로 추출하지 못하는 단점을 가지고 있다. 또한 한국어의 경우 띄어쓰기가 자유로워 '분산 데이타베이스'와 '분산데이타베이스' 같이 같은 용어라도 서로 다른 형태로 나타나기 때문에 올바른 결과를 기대하기 어렵다.

2.2 빈도수와 명사구간의 내포관계에 기반한 전문용어 추출[3]

[3]에서는 빈도수와 명사구 사이의 내포 (nested) 관계를 이용하여 전문용어를 추출하였다. 언어필터를 이용하여 추출한 후보 명사구에서 어떠한 명사구 A가 다른 명사구 B의 일부로 포함되면, A는 B에 내포되었다고 정의했다. 예를 들어 명사구 '데이타베이스'와 명사구 '분산데이타베이스'에 대하여 '데이타베이스'는 '분산데이타베이스'에 내포된다라고 말한다.

[3]에서는 길이가 긴 명사구이면서 내포되지 않은 명사구는 전문용어일 가능성이 높은 반면, 빈도수가 낮은 경우가 많기 때문에, 명사구의 길이와 빈도수와의 관계를 고려하여 해당 명사구의 가중치를 결정하였다. 또한, 길이가 짧은 명사구이면서 다른 명사구에 내포된 명사구는 전문용어일 가능성은 낮지만 그 자체로 높은 빈도수를 가지므로, 그 명사구를 내포한 명사구의 종류와 내포된 빈도수에 따라 해당 명사구의 가중치를 결정하였다. [3]에서는 이를 C-value라 정의하고 식 (2)와 같이 나타내었다.

$$C-value(\alpha) = \begin{cases} \log_2|\alpha| \cdot f(\alpha); \text{ if } \alpha \notin S_N \\ \log_2|\alpha| \cdot \left(f(\alpha) - \frac{1}{P(T_\alpha)} \sum_{\beta \in T\alpha} f(\beta) \right); \text{ if } \alpha \in S_N \end{cases} \quad (2)$$

여기서, a는 후보 명사구, S_N은 다른 명사구에 내포되는 명사구의 집합, $|a|$는 a의 길이, T_a는 명사구 a를 내포하는 명사구의 집합, $f(a)$는 문서에서의 a의 빈도수, $P(T_a)$는 명사구 a를 내포하는 명사구의 종류를 각각 나타낸다.

[3]에서 제안한 C-value는 용어의 길이를 2어절 이상으로 제한하여, 1어절의 전문용어를 추출할 수 없다는 문제점과 명사구간 내포관계의 적용에 있어 문제점을 가진다. 같은 빈도수를 가지는 내포된 두 명사구에 대하여 내포하는 명사구의 종류가 많은 명사구에 높은 가중치를 할당한다. 이는 전문용어를 구성하는 일반적인 명사에 높은 가중치가 부여되는 문제점을 가지게 된다. 예를 들어 일반적인 용어 '방법'은 '전문용어 추출방법', '페트리네트의 구성 방법' 등의 전문용어에 내포되는 경우가 많다. 따라서 [3]의 방법을 이용하면, 이러한 일반적인 용어가 높

은 가중치를 가져, 전문용어로 추출되는 경우가 발생한다.

기존의 전문용어 추출 기법들이 사용하는 빈도수나 명사구간의 내포관계만으로는 한국어 전문용어를 추출하는 데는 어려움이 있으며, 한국어에 맞는 전문용어 추출 기법이 필요하다. 본 논문에서는 기존의 기법의 문제점을 보완하기 위하여 사전에서 추출한 정보와 빈도수 및 외래어에 기반한 문서내 통계정보를 이용하여 한국어 전문용어를 추출하는 효과적인 방법을 제안하고자 한다.

3. 사전계층관계에 기반한 분야간 유사도와 통계기법을 이용한 전문용어 자동추출 기법

본 논문에서 제안하는 전문용어추출 방법의 전체 과정은 (그림 1)과 같다. 본 논문에서 제안하는 방법은 네 단계의 과정으로 이루어진다. 첫 번째 단계에서는 클러스터링 기법에 의해 사전간의 계층관계가 구축된다. 두 번째 단계에서는 부분 구문정보[1]를 이용하여 명사구를 추출하며, 세 번째 단계에서는 추출된 명사구에 대하여 가중치를 부여한다. 사전에 의한 가중치 기법은 해당 명사구가 나타난 전문용어 사전의 개수에 기반하여 가중치를 부여하며, 사전에 수록되어 있지 않은 미등록어와 일반용어에 대한 처리를 위하여 분야정보가 표시된 문서를 이용한다. 통계기법에 의한 가중치 기법은 명사구의 출현빈도와 내포관계 등을 이용한다. 음차 표기 단어 및 외국어를 이용한 가중치 기법에 의해서 주어진 후보 명사구는 음차 표기 외래어나 영어가 포함된 어절수에 의해 가중치가 정해진다. 네 번째 단계에서는 각각의 가중치 값을 하나의 값으로 통합하여 전문용어를 추출한다.

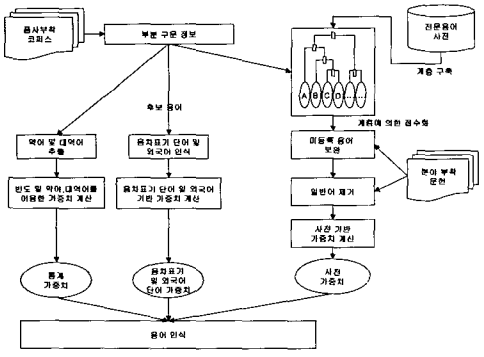


그림 1 전문용어 자동 추출 시스템 구조도

1) 부분 구문 정보는 명사구를 추출하기 위한 규칙을 나타낸다. 본 논문에서 사용한 부분 구문 규칙은 "Noun+(jcm? Noun+)"의 정규표현으로 나타낼 수 있다. 여기에서 Noun은 명사, jcm은 관형격 조사를 나타낸다.

3.1 사전간 계층관계를 이용한 용어의 가중치 계산

특정 분야의 전문용어는 유사한 다른 분야의 전문용어로부터 비롯된 경우가 많기 때문에 전문분야간의 상호 연관성은 전문용어를 추출할 때 중요한 요소가 될 수 있다. 이는 어떠한 분야의 전문용어를 추출할 때, 해당분야 혹은 인접분야의 사전에 나타나는 용어와 전혀 다른 분야에서 나타나는 용어를 구별한다는 것을 의미한다. 본 논문에서는 이러한 분야간의 상호 연관성을 구축하기 위하여 계층적 클러스터링 방법을 이용하고, 구축된 계층관계를 통해 사전간 (분야간) 상호 연관성을 유추하여 전문용어 추출에 사용한다. 본 장에서는 사전간의 계층관계를 구축하는 방법과 이를 이용하여 용어의 가중치를 결정하는 방법에 대하여 기술한다

3.1.1 사전간 계층관계 구축을 위한 데이타

사전간의 계층관계는 이중언어 사전 (영어-한국어)을 이용하여 구축한다. 사전은 과학기술분야의 57개 분야사전[2]을 이용한다. 그리고 모든 사전에 나타나지 않은 미등록어와 사전에 나타나는 일반용어를 처리하기 위하여 분야 정보가 표시된 ETRI-Kemong 문서집합[7]을 이용하였다.

3.1.2 사전간 계층관계 구성을 통한 분야간 유사도 계산

본 논문에서는 분야간의 유사도를 계산하기 위하여 사전간의 계층관계를 구축한다. 이를 위하여 정보검색분야에서 사용되는 클러스터링 방법을 사용한다. 클러스터링 방법은 문서간의 유사성을 이용하여 문서들간의 구조를 구성하는 통계적 기법으로 계층적 클러스터링과 비계층적 클러스터링이 있다[8]. 본 논문에서는 이러한 클러스터링 방법 중에서 계층적 클러스터링 방법을 사용하였다. 계층적 클러스터링 방법에 의해 구축된 문서들간의 구조는 트리형태를 가진다. 본 논문에서는 클러스터링 될 문서로 각 분야 사전을 사용하였다. 또한 각 분야 사전에 나타나는 표제어를 이용하여 클러스터링을 수행한다. 계층적 클러스터링 방법에 의해 구축된 트리에서 분야간 유사도는 트리내 각 분야 사전의 위치를 이용하여 계산한다. 그런데 트리 형태가 편향(skewed)된 형태로 구성되면, 클러스터간 관계나 분야와 클러스터간 관계가 많아져 분야간의 유사성을 올바르게 유추하기 힘들다. 따

2) 사용된 사전은 "한림원 과학기술분야용어집"으로 가정, 건축, 국토, 금속, 기계, 기초과학, 농공, 농기, 농생, 농화학. 대기, 물리, 산업공학, 생물, 설비, 섬유, 소방, 수문, 수산, 수의, 수학, 식품, 약학, 영약, 요업, 용접, 원예, 원자, 육수, 육종, 응용과학, 의학, 인쇄, 임학, 자동, 자원, 작물, 잠사, 전기, 전산, 전자, 조선, 주조, 지리, 지질, 천문, 체육, 축산, 치과, 토목, 토지비료, 통계, 통신, 항공, 해약, 화공, 화학의 57개 분야 약 446,500개 표제어를 포함한다.

라서 단말노드 (leaf node) 사이의 결합이 보다 많은 형태로 나타나는 대칭적인 형태의 트리를 구성할 필요가 있다.[3] 이를 위해 본 논문에서는 계층적 클러스터링 방법 중에서 비교적 대칭적인 계층구조를 만들어내는 [9] "상호 최근인접 이웃 알고리즘 (a reciprocal nearest neighbor algorithm)" [10]을 사용하였다.

계층 구조를 형성하기 위한 알고리즘의 수행과정은 다음과 같다.

1. 모든 사전간의 유사도를 결정한다.
2. 가장 유사한 개체[4]를 하나의 클러스터로 구성한다.
3. 2단계에서 구성된 새로운 클러스터와 다른 개체간 또는 이미 만들어진 클러스터간의 유사도를 재계산한다 (새로운 클러스터와의 유사도 외에 다른 개체간 유사도는 변하지 않는다.)
4. 모든 개체가 하나의 클러스터로 구성될 때까지 2단계와 3단계 과정을 반복한다.

상호 최근 인접 이웃 알고리즘에서는 모든 개체들이 $O_i=\{x_{i1}, x_{i2}, \ldots\ldots, x_{iL}\}$와 같은 벡터로 나타내어 진다: 여기서 O_i는 i번째 개체를 나타내며, x_{ij}은 i번째 개체 내에서의 j번째 용어를 나타낸다. 1단계에서 개체간 유사도는 유클리디안 거리 (Euclidian distance)를 이용하여 계산되고, 2단계에서 상호 가장 유사한 개체는 상호 최근 인접 이웃에 의해 결정된다. 주어진 개체 i와 j에 대하여 i와 가장 유사도가 높은 개체가 j이고, j와 가장 유사도가 높은 개체가 i일 때 이들 i와 j는 상호 최근인접 이웃 (reciprocal nearest neighbor)이라고 정의된다. 가장 유사한 개체는 두 개체가 통합되었을 때, 평균에 대한 그룹내 분산의 증가가 가장 작은 개체쌍이 된다. 주어진 두 개체 O_i와 O_j에 대하여 분산의 증가는 식 (3)과 (4)에 의해 나타내어진다.

$$I_{ij}=\frac{m_i \times m_j}{m_i+m_j} \times d_{ij}^2 \qquad (3)$$

$$d_{ij}^2=\sum_{k=1}^{L}(x_{ik}-x_{jk})^2 \qquad (4)$$

여기서, m_i는 개체 O_i내 개체 수를 나타내고, d_{ij}^2는 유클리디안 거리의 제곱을 나타낸다.

이 알고리즘을 이용하여 구축된 사전간의 계층관계는 (그림 2)와 같이 나타내어진다. (그림 2)에서 계층관계를 구성하고 있는 사전은 5개 분야의 사전이며, 전체 57개 분야의 사전으로 구성된 계층관계의 일부를 나타내고 있다.

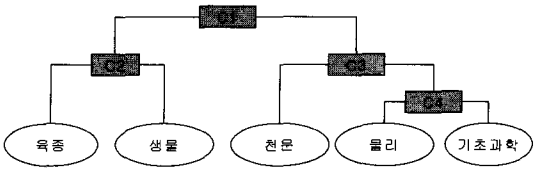


그림 2 구축된 사전간 계층관계의 예

3.1.3 분야간 유사도를 이용한 용어의 가중치 계산

계층관계를 이용한 용어의 가중치 부여는, 추출하고자 하는 분야의 사전에 나타나는 용어와 그 분야와 연관성이 많은 분야의 사전에 나타나는 용어는 전문용어 추출에 있어 긍정적인 지시자 (positive indicator)로 작용될 수 있으며, 해당 분야와 연관성이 적은 분야의 사전에 나타나는 용어는 부정적인 지시자 (negative indicator)로 작용될 수 있다는 전제에 기반한다. 본 논문에서는 클러스터링 방법에 의해 구축된 계층관계에서 분야간 상호 연관성을 계산하고 이러한 연관성에 기반하여 용어의 가중치를 부여한다. 이를 위해 다음과 같은 3가지 단계의 과정이 필요하다.

1. 식 (5)를 이용하여 분야간 유사도를 계산한다[11].

$$similarity_{ij}=\begin{cases}\dfrac{2}{(depth_i+depth_j-2\times Common_{ij})} & i\neq j\\ 2 & i=j\end{cases} \qquad (5)$$

여기서, $depth_i$는 사전간 계층에서 분야 i의 깊이 정보를, $Common_{ij}$는 사전간 계층에서 분야 i와 분야 j간에 공유하는 가장 깊은 노드의 깊이 정보를 나타낸다.

식 (5)에서 계층의 노드 깊이는 계층의 루트 (root)로부터 해당 노드까지의 거리를 나타낸다. - 루트의 깊이는 1로 가정한다. 예를 들어 (그림 2)에서 노드 C1을 루트 노드라 가정하였을 경우 '물리'와 '기초과학'의 유사도는 <표 1>과 같이 계산된다.

표 1 $Similarity_{ij}$의 계산

분야	물리	기초과학
2루트 (root)로부터의 경로	루트->C3->C4->물리	루트->C3->C4->기초과학
Depth	4	4
Common	3	
유사도 (Similarity)	2/(4+4-2*3) = 1	

3) 계층적 클러스터링 방법에 의해 구축된 트리에서 단말 노드 (leaf node)간의 결합이 많다는 것은 클러스터링간의 결합보다는 분야간에 결합이 더 많다는 것을 의미한다. 따라서 편향된 트리보다 대칭된 형태가 분야간의 관계를 유추하는 데 좋은 형태이다.

4) 여기서 개체는 사전뿐만 아니라 여러 사전이 하나로 묶여진 형태인 클러스터도 포함한다.

2. 추출하고자 하는 분야와 용어가 나타난 사전의 분야와의 거리는 식 (6)에 의해 계산된다.

$$Score(term,t)=\frac{|match|}{|term|}\times\frac{1}{N}\times\sum_{i=1}^{N}similarity_{ti} \quad (6)$$

여기서 N은 용어가 나타난 사전의 개수, t는 추출하고자 하는 분야 (target domain), $Similarity_{ti}$는 식 (5)에서 계산된 추출하고자 하는 분야와 용어가 나타난 사전의 분야와의 유사도, $|term|$은 주어진 용어의 어절수, $|match|$는 사전의 수록용어와 최장일치를 통하여 일치된 주어진 용어의 부분열(sub-string)의 어절 수를 각각 나타낸다. 따라서, 주어진 용어 전체가 사전의 수록용어와 일치 될 때에는 $|match| = |term|$이 된다.

식 (6)에서 주어진 용어가 추출하고자 하는 분야의 사전에만 나타나며, 주어진 용어 전체가 해당 분야 사전에 수록되어 있을 경우에는 가장 높은 가중치를 가지게 된다. 하지만, 용어가 가장 높은 가중치를 가지는 최적의 경우 외에도 나타날 수 있는 두 가지 경우가 있다. 첫 번째로 주어진 용어가 추출하고자 하는 분야의 사전에 나타나지 않고 다른 분야의 사전에만 나타날 경우이며, 두 번째로 주어진 용어가 추출하고자 하는 분야의 사전 뿐만 아니라 다른 분야의 사전에도 나타날 경우이다. 이러한 경우들에서는 식 (5)에 의해 계산된 분야간의 관계 ($similarity_{ti}$)를 이용하여, 주어진 용어에 가중치를 부여한다.

주어진 용어와 사전에 수록되어 있는 용어와의 비교는 전체용어일치기법(exact matching method)과 부분용어일치기법 (partial matching method)을 사용한다. 주어진 용어의 전체 형태가 어떠한 사전에 수록되어 있는 경우 전체용어일치기법을 사용하며, 그렇지 않을 경우에는 부분용어일치기법을 이용한다. 부분용어일치기법을 사용할 때 용어의 다음과 같은 특성을 이용한다. 일반적으로, 여러 단어로 구성된 용어 (Multi-word term)에 있어서 가장 중요한 의미를 가지고 있는 단어를 중심단어라 하며, 이는 대부분 용어의 끝에 위치한다. 또한 중심단어는 전체 용어의 의미를 용어의 다른 부분에 비해 잘 표현한다. 따라서 주어진 용어에서 전체 용어가 사전에 포함되어 있지 않다 하더라도 중심단어를 포함하는 부분열이 다른 부분보다 해당 단어의 의미를 보다 명확히 한다고 할 수 있다. 예를 들어, '오염된 방사능 원소'의 경우 중심단어를 포함하는 '방사능 원소'가 '오염된 방사능' 또는 '오염된 원소'보다는 '오염된 방사능 원소'의 의미를 보다 잘 표현한다고 할 수 있다. 이러한 특성을 이용하여, 본 논문에서 사용한 부분용어일치기법은 오른쪽에서 왼쪽으로의 최장일치기법 (right-to-left longest matching procedure)[5]을 사용한다[11].

표 2 식 (6)에 의해 부여된 '오염된 방사능 원소'에 대한 가중치의 예

N	3
t	물리분야
$similarity_{물리-생물}$	0.5
$similarity_{물리-물리}$	2
$similarity_{물리-기초과학}$	1
$\|term\|$: '오염된 방사능 원소'의 어절수	3
$\|match\|$: 부분용어일치된 '방사능 원소'의 어절수	2
$Score$('오염된 방사능 원소')	2/3*1/3*(0.5+2+1) =0.78

예를 들어 추출하고자 하는 분야가 물리이고, 주어진 용어가 '오염된 방사능 원소'라고 하였을 때, 주어진 용어가 어떠한 사전의 용어와도 전체용어일치가 되지 않은 용어이고, 부분용어일치에 의해 '방사능 원소'가 물리, 기초과학, 생물 분야 사전에 나타났다고 가정하자. 주어진 용어가 전체용어일치가 되지 않았지만 부분용어일치기법을 이용하여 식 (6)에 의해 가중치가 부여되며, 부여된 가중치는 <표 2>와 같이 계산된다.

3. 사전 미등록어의 보완: 분야정보가 부착된 문서집합의 이용

주어진 용어가 전체나 부분의 형태로 사전에 출현하지 않을 경우, 식 (6)에 기술한 방법으로는 용어의 가중치를 부여할 수 없다. 하지만 이러한 용어들은 해당 분야의 전문용어일 가능성이 있으므로, 이들에 대한 처리도 고려해야 한다. 본 논문에서는 이러한 경우의 용어를 처리하기 위하여 분야가 태그된 문서집합[7]을 이용한다. 용어를 구성하는 모든 단어에 대하여 분야 태그된 문서집합에 출현하는가의 여부를 판별할 수 있으며, 해당 단어가 얼마나 많은 분야의 문서에 나타났는가를 계산할 수 있다. 이러한 계산 결과를 통하여, 출현한 분야의 개수가 많을 경우 분야 변별성이 떨어지므로, 일반용어일 가능성이 높고, 출현한 분야의 개수가 작을 경우에는 분야 변별성이 높으므로, 전문용어일 가능성이 높다. 본 논문에서는 식 (7)을 사전계층관계를 이용한 사전 가중치라 하고 W_{Dic}으로 나타낸다.

5) 주어진 후보 용어에 대하여 용어의 오른쪽으로부터 사전에 수록된 용어와 일치되는 가장 긴 용어가 부분적으로 일치된 용어로 판정한다.

$$W_{Dic}(\alpha)=\begin{cases} score(\alpha) & if\ Dic(\alpha)=1 \\ \sqrt{\dfrac{W}{\sum_{i=1}^{W} dof_i}} & if\ Dic(\alpha)=0 \end{cases} \quad (7)$$

여기서, α는 주어진 후보 용어, $Dic(\alpha)$는 α가 부분이나 전체의 형태로 사전에 나타날 경우 1의 값을 그렇지 않을 경우 0의 값을 나타내는 함수, W는 용어 후보에 포함된 단어의 개수, dof_i는 분야정보가 부착된 문서 집합에서 나타난 단어의 분야 개수를 나타낸다.

3.2 문서내 통계정보를 이용한 용어의 가중치 계산

문서내 통계정보를 이용한 용어의 가중치 계산은 크게 두 가지 요소로 이루어진다. 첫 번째 요소는 통계 가중치 (W_{Stat})라 나타내며, 용어들이 문서에 나타난 출현빈도와 용어들 사이의 내포관계에 기반한 가중치이고, 두 번째 요소는 음차 표기 단어 및 외국어 가중치 (W_{Trl})라 나타내며, 해당 용어가 포함하는 음차 표기된 외래어 및 영어의 개수에 기반한 가중치이다.

3.2.1 통계 가중치: 용어의 문서내 빈도수와 용어간 내포관계에 기반한 가중치

통계 가중치를 계산하기 위하여 문서에서 나타나는 괄호 표현에 의한 대역어 쌍과 약어쌍, 그리고 용어들의 문서에서의 출현빈도와 용어들간의 내포관계를 고려한다. 우선 대역어쌍과 약어쌍은 다음과 같은 휴리스틱을 이용하여 추출한다.

주어진 괄호표현 A(B)에 대하여,

1. A와 B가 약어와 그 확장어의 쌍인지를 검사한다. 이를 위해 A와 B의 영어 대문자를 비교하여 반 이상이 순서적으로 일치하면 약어쌍이라고 판단한다[12]. 예를 들어, 'GIS(Geographical Information System)'의 'GIS'와 'Geographical Information System'은 약어쌍이라 판단된다.

2. A와 B가 대역어쌍인지를 검사한다. 이를 위해 이중언어 사전을 이용한다.

약어쌍과 대역어쌍을 추출한 후 식 (8)에 의해 통계 가중치 W_{Stat}를 계산한다.

$$W_{Stat}(\alpha)=\begin{cases} \sum_{\beta\in S(\alpha)\cup\{\alpha\}}\left[\sqrt{|\beta|}\times\left(f(\beta)+\dfrac{\sum_{\gamma\in T(\beta)} f(\gamma)}{C(T(\beta))}\right)\right] & if\ \beta\in S_N \\ \sum_{\beta\in S(\alpha)\cup\{\alpha\}}\left(\sqrt{|\beta|}\times f(\beta)\right) & if\ \beta\notin S_N \end{cases} \quad (8)$$

여기서, α는 후보 용어, S_N은 다른 명사구에 내포되는 명사구의 집합, $|\alpha|$는 용어 α의 어절수, $S(\alpha)$는 α의 약어쌍이거나 대역어쌍인 용어들의 집합, $T(\alpha)$는 용어 α를 내포하는 용어들의 집합, $f(\alpha)$는 문서에서 용어 α의 출현빈도, $C(T(\alpha))$는 용어 α를 내포하는 용어의 종류를 각각 나타낸다.

식 (8)에서 내포관계는 다음과 같이 정의된다. A와 B를 용어라 하고, A가 B를 포함하면 A가 B를 내포한다고 정의한다. 예를 들어, '이진 탐색 트리'와 '탐색 트리'에서 '이진 탐색 트리'는 '탐색 트리'를 내포한다라고 말한다. 식 (8)은 주어진 용어가 약어 및 대역어를 가질 경우 해당 용어의 통계정보 뿐만 아니라 약어나 대역어의 통계정보도 같이 계산된다. 또한 용어 α가 α를 포함하는 다른 용어를 만들 경우, α는 보다 높은 가중치를 가진다. 이는 해당분야에서 용어의 생산성(productivity of terms)이 높은 용어일수록 전문용어일 가능성이 높다는 것을 나타낸다. 또한, [3] 등과 같은 기존연구가 여러 어절로 구성된 용어만을 추출대상으로 한 것과는 달리 식 (8)은 여러 어절로 구성된 용어뿐만 아니라 단일어절로 된 용어를 추출할 수 있다. 이는 'GUI (Graphical User Interface)'에서의 'GUI'와 같이 약어의 경우 단일어절로 구성되어 있으며, 영어에서는 여러 어절로 나타나는 용어가 한국어에서는 단일어절의 용어로 번역되는 경우가 많기 때문에 단일어절로 구성된 경우를 고려하여야 한다. (e.g. 'distributed database' => '분산데이타베이스')

3.2.2 음차 표기 외래어 및 외국어 기반 가중치: 용어에 포함된 음차 표기된 외래어 단어와 외국어의 개수에 기반한 가중치

외국언어에 어원을 두는 전문용어는 주로 음차 표기되는 경우가 많기 때문에 음차 표기된 용어는 주어진 분야의 용어를 추출하기 위한 중요한 단서가 될 수 있다. 하지만 음차 표기된 외래어는 표기에 대한 표준이 있음에도 불구하고 사용자마다 달리 표기하기 때문에 사전에 수록되어 있지 않은 경우가 많다[13]. 어떠한 용어가 음차 표기된 단어를 포함하는가를 사전에 의존해서 판단하는 것은 어려움이 있으며, 이를 자동으로 추출하는 방법이 필요하다. 본 논문에서는 은닉 마르코프 모델을 이용한 외래어 자동추출 모델[14, 15]을 이용하여 외래어를 자동으로 추출하였다.

외래어 추출방법은 한국어와 외국언어가 음운학상으로 서로 다르기 때문에, 음차 표기된 외래어의 구성과 순수한국어의 구성은 서로 다르다는 전제에 기반한다. 특히 영어의 경우, 영어에서 자주 사용되는 자음인 'p', 't', 'c', 'f'는 각각 한국어 자음인 'ㅍ', 'ㅌ', 'ㅋ', 'ㅍ'로 음차 표기된다. 그런데, 한국어에서 이들 자음들은 순수 한국어에서 자주 사용되지 않는 자음들이다. 이러한 특

성은 외래어를 추출할 때 중요한 단서가 될 수 있다. 예를 들어, '시스템'이라는 단어에서 '템'은 순수한국어에서 자주 사용되지 않는 자음 'ㅌ'을 초성으로 사용하기 때문에 음차 표기된 외래어가 될 가능성이 높다. 어떠한 단어를 구성하는 각 음절의 자음 정보는 외래어를 추출하는 데 중요한 정보로 사용될 수 있다.

은닉 마르코프 모델을 이용한 외래어 추출모델은 주어진 단어의 각 음절이 순수한국어의 음절인지 음차 표기된 외래어의 음절인지를 결정한다. 이를 위해 순수한국어의 음절인 경우에는 'K'라는 태그를 할당하고, 음차 표기된 외래어의 음절인 경우에는 'F'라는 태그를 할당한다. 예를 들어 '시스템은'은 '시/F + 스/F + 템/F+은/K' 이라고 음절 태깅 (syllable-tagging)할 수 있다. 외래어 자동 추출 모델에서는 음절정보를 품사 (Part-of-Speech)태깅에서의 어휘정보와 같이 사용하였다. 식 (9)는 은닉 마르코프 모델을 이용한 외래어 추출모델을 나타낸 식이며, 식 (10)은 추출된 외래어에 따라 주어진 용어의 가중치를 할당하는 식이다. 식 (10)을 W_{Trl}라고 정의하고 전문용어 추출에 사용한다. 식 (10)은 음차 표기된 외래어를 많이 포함할수록 전문용어일 가능성이 높다라는 의미를 내포하고 있다.

$$P(T|S)P(S) = p(t_1)p(t_2|t_1)\left[\prod_{i=3}^{n} p(t_i|t_{i-1},t_{i-2})\right]\left[\prod_{i=1}^{n} p(t_i|s_i,s_{i-1},t_{i-1})\right] \quad (9)$$

여기서, s_i는 주어진 용어의 i번째 음절을 t_i는 주어진 용어의 i번째 음절의 태그 ('F' or 'K')를 나타낸다.

$$W_{Trl}(\alpha) = \frac{trans(\alpha)}{|\alpha|} \quad (10)$$

여기서 $|\alpha|$는 용어 α의 어절 수를, trans(α)는 용어 α에서 음차 표기된 외래어 및 외국어를 포함하는 어절 수를 나타낸다.

3.2.3 용어의 가중치

위에서 기술한 3가지 가중치는 식 (11)에 의하여 통합되어, W_{Term}이라 정의된다. 각각의 가중치 기법인 W_{Stat}, W_{Trl}, W_{Dic}은 서로 다른 정보에 기반하기 때문에 각 용어의 가중치가 포함하는 용어의 범위 또한 다르다. 따라서 각각의 가중치 기법만으로는 효율적으로 전문용어를 추출할 수 없기 때문에 식 (11)과 같이 각각의 가중치를 통합한다.

$$W_{Term}(\alpha) = \lambda_1 \times f(W_{Stat}(\alpha)) + \lambda_2 \times g(W_{Trl}(\alpha)) + \lambda_3 \times h(W_{Dic}(\alpha)) \quad (11)$$

여기서, α는 후보용어를 나타내며, f,g,h는 각 가중치를 정규화시켜주는 함수를 나타낸다. 또한 $\lambda_1, \lambda_2, \lambda_3$은 W_{Stat}, W_{Trl}, W_{Dic}에 대한 가중치이며, $\lambda_1+\lambda_2+\lambda_3=1$이다. 실험에 의해 $\lambda_1=0.3, \lambda_2=0.4, \lambda_3=0.3$으로 정해지며, $\lambda_1, \lambda_2, \lambda_3$에 대한 실험은 4.2.2장에 기술하였다.

본 논문에서는 식 (11)에 의해 통합된 가중치를 이용하여 '사전에 수록되어 있는 전문용어'와 '사전에 수록되어 있는 용어를 이용하여 새로이 생성된 용어' 뿐만 아니라 '사전에 수록되어 있지 않는 용어들'도 문서내에 나타나는 용어의 정보를 이용하여 전문용어를 자동으로 추출한다.

4. 실험 및 평가

컴퓨터분야와 전기전자 분야의 문서를 포함하는 정보검색 테스트 집합인 KT문서 집합[16]을 사용하여 컴퓨터 분야의 용어를 추출하는 실험을 수행하였다. 본 논문에서는 전체 4,413개 문서 중 컴퓨터분야의 논문의 초록을 포함하는 1,000개 문서 약 67,250어절을 사용하였다. 또한 명사구를 추출하기 위한 품사정보를 얻기 위하여 품사 태거[17]로 자동으로 태깅하였다.

부분 구문 분석에 의하여 추출된 전체 용어의 개수는 총 12,915개이며, 이 중 전문용어는 10,226개, 일반용어는 2,689개이다. <표 3>은 어절의 길이에 따른 KT문서 집합에 포함된 전문용어의 분포를 나타낸다. <표 3>에서 전체 전문용어 비율이 약 80%정도로 비교적 높게 나타나는데, 이는 문서 집합이 논문의 요약문으로 전문용어가 많이 포함되어 있기 때문으로 분석된다. 또한, 1어절 용어의 경우 2어절 이상의 용어보다 전문용어의 비율이 낮은 것을 알 수 있는데 이는 1어절의 경우 '논문', '방법'과 같은 일반용어가 2어절 이상의 용어보다 많기 때문으로 분석된다.

표 3 KT문서 집합에 나타난 어절별 전문용어의 분포

	1어절 용어	2어절 이상 용어	총계
전문용어	2394 (61.31%)	7832 (86.93%)	10,226 (79.18%)
일반용어	1511 (38.69%)	1178 (13.07%)	2,689 (20.82%)
총계	3905	9010	12,915

각 요소의 유용성과 정보의 통합이 전문용어 추출의 성능에 어떠한 영향을 미치는지를 알아보기 위하여 다음의 경우에 대하여 비교 실험하였다.

▶ 각 W_{Stat}, W_{Trl}, W_{Dic}들을 통합하지 않고 전문용어를 추출하는 경우

▶ 각 W_{Stat}, W_{Trl}, W_{Dic}에 대한 가중치($\lambda_1, \lambda_2, \lambda_3$)를 달리 했을 경우의 전문용어 추출 결과

- ▶ $W_{Stat}-W_{Trl}$를 통합하여 이용한 경우
- ▶ $W_{Stat}-W_{Dic}$를 통합하여 이용한 경우
- ▶ $W_{Trl}-W_{Dic}$를 통합하여 이용한 경우
- ▶ W_{Stat}, W_{Trl}, W_{Dic}를 통합하여 이용한 경우

또한 빈도수에 기반한 용어 추출 방법[4]과 C-value 방법[3]을 비교 평가함으로써 본 논문이 제안하는 기법의 효용성을 살펴보고자 한다.

4.1 평가 기준

두 명의 분야 전문가가 제안된 전문용어추출 방법에 의해 추출된 용어에 대한 평가를 하였으며, 평가된 결과에서 두 명 모두가 전문용어라고 판단한 경우에만 전문용어로 인정하였다. 이는 한 명이 이러한 평가작업을 수행할 경우에 나타나는 주관적 평가를 배제하기 위한 것이다. 결과는 전문용어 추출방법에 의해 추출된 전문용어 중에 전문용어라 판단된 용어의 비율을 나타내는 정확률 (precision)로서 평가된다. 이를 수식으로 나타내면 식 (12)와 같다.

$$\text{정확률} = \frac{\text{추출한 용어 중 전문용어의 개수}}{\text{추출한 용어의 개수}} \quad (12)$$

본 논문에서는 정확률을 평가하기 위하여 후보 용어들에 부여된 점수를 높은 순에서 낮은 순으로 정렬한 뒤 10개 부분으로 똑같이 나누어서 독립적으로 평가하였다. 따라서 10개 부분 중 상위에 존재하는 부분의 정확률은 높을수록, 하위에 존재하는 부분은 낮을수록 전문용어를 효과적으로 추출한다고 말할 수 있다[3].

여러 방법들의 전체적인 성능을 비교하기 위하여 정보검색분야에서 사용되는 11-포인트 평균 정확률 (11-point average preceision)을 사용하였다. 11-포인트 평균 정확률은 재현율이 0%, 10%, 20%, 30%, ..., 90%, 100% 지점일 때의 재현율에 따른 정확률을 계산한 뒤, 각 지점의 정확률을 합하고 이를 평균하여 나타내어진다. 따라서, 각 재현율의 지점에서의 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[18]. 본 논문에서는 11-포인트 평균 정확률을 구하기 위하여, <표 3>에서의 전문용어 후보 12,915개 중 전문용어로 판별되는 10,226개의 후보를 모두 추출하였을 때 재현율이 100%라고 가정한다. 이를 기준으로 재현율 0%~100% 지점을 찾아 해당 지점에서의 정확률을 계산한다.

4.2 W_{Stat}, W_{Trl}, W_{Dic}의 통합 여부에 따른 전문용어 추출 비교실험

4.2.1 각 가중치만 사용한 전문용어 추출 실험

<표 4>는 각 가중치만으로 전문용어를 추출하였을 때의 성능을 나타낸다.

표 4 각 가중치만을 사용한 전문용어 추출 결과

부분	W_{Stat}	W_{Trl}	W_{Dic}
1	89.61%	93.1%	88.37%
2	89.30%	89.53%	87.60%
3	83.10%	96.89%	86.28%
4	84.73%	92.87%	82.64%
5	90.93%	94.80%	83.80%
6	87.13%	71.62%	78.06%
7	78.14%	62.95%	81.86%
8	80.93%	60.23%	81.71%
9	71.40%	66.04%	78.53%
10	36.28%	63.49%	42.71%
11pt-avg	88.12%	90.11%	86.65%

실험 결과에서 W_{Trl}의 상위부분(부분 1~5)은 평균 약 93%의 정확률을 보이며, 하위 부분 (부분 7~10)은 평균 약 63%의 정확률을 보인다. 이는 W_{Trl}의 가중치기법 특성상 해당 용어에 음차표기된 외래어가 많을수록 높은 값을 부여하고 외래어를 포함하지 않는 용어에 대해서는 모두 일정한 가중치를 부여하기 때문으로 분석된다. 용어후보 중 외래어나 외국어가 포함된 용어는 7,034개로 전체 12,915개 후보의 약 54%를 차지하며, 7,034개 중 6,584개가 전문용어로 판별되어 외래어나 외국어가 포함된 용어가 전문용어가 될 경우가 약 93%로 나타난다. 따라서 W_{Trl}에 의해 추출된 부분 1에서 부분 5까지는 외래어나 영어를 포함하는 용어들이 나타나며 비교적 높은 성능을 보이는 반면, 용어에 외래어나 외국어가 없는 부분 7에서 부분 10까지는 일정한 정확률을 보인다.

W_{Stat}의 경우 해당 용어가 문서에서 높은 빈도수로 나타날수록, 대역쌍이나 약어쌍으로 판별되었을 경우 높은 값을 가지게 된다. 이러한 특성으로 인하여, W_{Stat}의 경우 'Graphic User Interface', 'GUI', 'Graphical User Interface', '그래픽 사용자 인터페이스'가 유사어 관계로서 상위에 위치한다. 하지만 W_{Stat}만으로는 추출할 수 없는 전문용어가 있을 뿐만 아니라, 문서에서 나타난 빈도수가 많은 용어를 전문용어로 추출하는 경우가 발생한다. 예를 들어, 'task scheduling'은 전문용어임에도 불구하고, 문서에서 나타난 빈도수가 작아 전문용어로 추출되지 못한다. 또한 '추출'은 전문용어가 아님에도

불구하고 문서에서 나타난 빈도수가 많아 상위에 위치한다. 따라서 W_{Stat}이 제대로 추출하지 못하는 부분을 W_{Trl}과 W_{Dic}을 이용하여 보완할 필요가 있다.

W_{Dic}은 추출하고자 하는 분야의 정보가 전문용어를 추출하는 데 중요한 정보로 사용될 수 있다는 전제에 기반한다. 추출하고자 하는 분야와 이와 밀접하게 연관된 분야의 사전들에 수록되어 있는 용어들은 전문용어를 추출하는 데 긍정적인 지시자로 작용할 수 있고, 추출하고자 하는 분야와 관계없는 분야의 사전에 나타나는 용어는 전문용어 추출에 부정적인 지시자로 작용할 수 있다. 사전간 계층관계는 이러한 분야간의 연관성을 유추하기 위하여 구성된다. 하지만 'Signalling Network Operations System'을 나타내는 'SIGNOS'와 같은 전문용어는 사전에 수록되어 있지 않아 미등록어로 낮은 순위를 가진다. 새로이 생성된 전문용어나 약어와 같은 전문용어사전의 미등록어에 대해서는 사전 정보만으로 전문용어 추출이 어렵다.

각 가중치 기법인 $W_{Stat}, W_{Trl}, W_{Dic}$은 서로 다른 정보에 기반하기 때문에 각 가중치가 포함하는 용어의 범위 또한 다르다. 본 논문에서는 이러한 각 가중치들을 상호 보완적으로 통합하여 전문용어 추출의 성능을 향상시키고자 한다.

4.2.2 각 $W_{Stat}, W_{Trl}, W_{Dic}$에 대한 가중치를 달리 했을 경우의 전문용어 추출 실험

본 장에서는 $W_{Stat}, W_{Trl}, W_{Dic}$에 대한 가중치 (식 (11)의 $\lambda_1,\lambda_2,\lambda_3$)를 달리 하여 전문용어 추출 실험을 하였다. $0.1\leq\lambda_1,\lambda_2,\lambda_3\leq0.9$ 범위 내에서 $\lambda_1,\lambda_2,\lambda_3$이 가질 수 있는 가능한 값[6]에 대하여 전문용어 추출 결과를 비교 평가한다. <표 5>는 $\lambda_1,\lambda_2,\lambda_3$에 따른 11-포인트 평균 정확률값이 높은 상위 5가지 경우의 $\lambda_1,\lambda_2,\lambda_3$의 값과 그 때의 11-포인트 평균 정확률값을 나타낸다. <표 5>에서 재현율이 0%일 경우, 정확률은 100%라고 가정한다 [18].

<표 5>에서 $\lambda_1=0.3,\lambda_2=0.4,\lambda_3=0.3$의 경우, 다른 가중치 조합에 비하여 11-포인트 평균 정확률이 높게 나타남을 알 수 있다. 따라서 본 논문에서는 11포인트 평균 정확률 값이 가장 높게 나타난 $\lambda_1=0.3$, $\lambda_2=0.4$, $\lambda_3=0.3$으로 $\lambda_1,\lambda_2,\lambda_3$의 값을 정하고 전문용어 추출 실험을 하였다.

6) 본 논문에서는 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9의 값에 대해서만 평가하였다.

표 5 $\lambda_1,\lambda_2,\lambda_3$에 따른 11-포인트 평균 정확률 값이 높은 상위 5개의 $\lambda_1,\lambda_2,\lambda_3$의 값과 그 때의 11-포인트 평균 정확률

재현율	$\lambda_1=0.3$ $\lambda_2=0.4$ $\lambda_3=0.3$	$\lambda_1=0.3$ $\lambda_2=0.3$ $\lambda_3=0.4$	$\lambda_1=0.3$ $\lambda_2=0.5$ $\lambda_3=0.2$	$\lambda_1=0.4$ $\lambda_2=0.3$ $\lambda_3=0.3$	$\lambda_1=0.2$ $\lambda_2=0.5$ $\lambda_3=0.3$
0%	100%	100%	100%	100%	100%
10%	97.71%	93.93%	94.02%	94.45%	92.41%
20%	94.15%	93.38%	93.12%	93.25%	93.12%
30%	92.54%	92.68%	92.49%	92.52%	92.49%
40%	93.31%	93.23%	93.44%	93.46%	93.48%
50%	93.38%	92.94%	93.38%	93.44%	93.37%
60%	93.60%	92.73%	93.63%	92.84%	93.56%
70%	92.08%	91.80%	91.86%	91.89%	91.82%
80%	89.59%	88.77%	89.07%	89.20%	88.72%
90%	87.20%	85.21%	85.56%	85.76%	85.03%
100%	79.28%	79.31%	79.28%	79.27%	79.31%
11pt -avg	92.08%	91.27%	91.44%	91.46%	91.21%

4.2.3 각 가중치의 통합에 따른 전문용어의 추출 실험

<표 6>은 각 가중치의 통합에 따른 전문용어의 추출 결과를 나타낸다.

표 6 $W_{Stat}, W_{Trl}, W_{Dic}$ 통합에 따른 전문용어 추출 결과

부분	$W_{Stat}-W_{Trl}$	$W_{Stat}-W_{Dic}$	$W_{Trl}-W_{Dic}$	W_{Term}
1	96.51%	89.46%	91.94%	96.59%
2	89.22%	87.52%	92.40%	93.64%
3	94.03%	87.52%	94.80%	93.72%
4	93.80%	85.35%	94.42%	93.95%
5	84.88%	84.65%	90.15%	89.38%
6	73.10%	82.25%	87.13%	84.42%
7	72.56%	77.60%	71.62%	74.03%
8	72.56%	78.76%	63.79%	75.50%
9	64.34%	78.37%	69.61%	63.18%
10	32.57%	40.08%	37.90%	27.44%
11pt-avg	91.07%	88.41%	91.05%	92.08%

실험 결과에서 W_{Term}의 11-포인트 평균 정확률이 $W_{Stat}-W_{Trl}$, $W_{Stat}-W_{Dic}$, $W_{Trl}-W_{Dic}$의 11-포인트 평균 정확률보다 높은 것을 알 수 있으며, 이를 통하여 W_{Term}에 의해 순위화된 용어들은 상위에는 많은 전문용어가 포함되어 있고, 하위에는 비교적 적은 전문용어가 포함됨을 알 수 있다.

<표 6>에서 $W_{Stat}-W_{Dic}$와 W_{Term}의 실험결과를 비교

하였을 때 $W_{Stat}-W_{Dic}$에 W_{Trl}이 통합된 W_{Term}의 성능이 높게 나타난다. 이러한 성능향상이 W_{Trl}에 의해 이루어 졌는가를 분석하기 위하여, $W_{Stat}-W_{Dic}$에 의해 추출된 전문용어 중 외래어나 외국어를 포함하는 것의 분포와 W_{Trl}을 포함하는 $W_{Stat}-W_{Trl}$, $W_{Trl}-W_{Dic}$, W_{Term}에 의해 추출된 전문용어 중 외래어나 외국어를 포함하는 것의 분포를 살펴 보았다. <표 7>은 각 부분별 외래어나 외국어를 포함하는 용어 후보의 개수를 나타낸다. 총 7,034개의 용어 후보가 외래어나 외국어를 포함하는 것으로 나타났으며, 이들 중 6,584개가 전문용어로 판별되었다. <표 7>에서 나타나듯이 $W_{Stat}-W_{Dic}$의 경우 상위 뿐만아니라 하위에도 외래어나 외국어를 포함하는 전문용어후보들이 많이 나타나는 것을 알 수 있다. 하지만, W_{Trl}을 이용한 결과($W_{Stat}-W_{Trl}$, $W_{Trl}-W_{Dic}$, W_{Term})에서는 부분 1에서 부분 7까지에서만 외래어나 외국어를 포함하는 전문용어 후보가 나타나는 것을 알 수 있으며, 외래어나 외국어를 포함하는 용어가 전문용어가 될 확률이 높기 때문에 상위뿐만 아니라 하위에서도 $W_{Stat}-W_{Dic}$을 이용한 경우보다 좀 더 좋은 성능을 나타냄을 알 수 있다.

표 7 각 방법에서의 외래어, 외국어를 포함하는 전문용어 후보의 분포

부분	$W_{Stat}-W_{Trl}$	$W_{Stat}-W_{Dic}$	$W_{Trl}-W_{Dic}$	W_{Term}
1	1,290	849	1,290	1,286
2	1,290	803	1,290	1,282
3	1,289	830	1,284	1,264
4	1,289	837	1,246	1,200
5	1,275	818	1,117	1,146
6	601	743	788	827
7	0	748	19	29
8	0	607	0	0
9	0	575	0	0
10	0	224	0	0
총계	7,034	7,034	7,034	7,034

본 장의 실험결과를 통하여 W_{Stat}, W_{Trl}, W_{Dic}이 상호 보완적으로 전문용어를 추출하는 데 유용한 정보로 사용됨을 알 수 있다.

4.3 기존 연구와의 비교실험

본 장에서는 제안한 전문용어 추출 기법과 기존의 연구와의 성능을 비교한 결과를 나타낸다.

<표 8>은 기존연구와의 비교실험 결과를 나타낸다. <표 8>에 나타난 결과는 다음과 같이 해석될 수 있다. 결과의 첫 번째 부분에서 3번째 부분까지 (상위 30%)에서 제안된 방법은 기존 방법[3][4]보다 높은 정확률을 보인다. 또한 본 논문에서 제안 방법이 상위 부분에서 기존연구보다 많은 전문용어가 포함되고 하위부분에서 기존연구보다 보다 적은 전문용어를 포함하는 양상을 보이기 때문에, 전문용어의 분포도 기존의 방법보다 좋은 결과를 보여준다. 이는 본 논문에서 제안한 방법에 의해 높은 가중치가 용어에 부여되면 해당 용어는 전문용어가 될 확률이 높다라는 의미를 내포한다. 또한, <표 8>에서 8번째 부분부터 10번째 부분까지의 정확률이 급격히 감소하는 것을 알 수 있는데 이는 대부분의 전문용어가 상위에 존재하고 하위에는 전문용어가 적게 나타남을 나타낸다. 본 논문의 기법은 상위 3부분에서 [3]보다 평균 12.60%, [4]보다 평균 29.00%의 성능향상을 보였으며, 하위 3부분에서는 [3]보다 평균 20.54%, [4]보다 평균 7.20%의 성능 향상을 나타내었다. 특히 상위에서 기존 연구보다 높은 성능을 나타내는데 이는 본 논문의 기법이 "B+트리", "HMM", "데이타베이스"와 같은 1어절의 전문용어를 효과적으로 처리하고, "논문", "방법"과 같은 1어절의 일반용어를 효과적으로 배제시키기 때문으로 분석된다.

또한 11-포인트 평균 정확률에 있어서도 본 논문의 기법이 기존 연구[3][4]보다 전체적으로 높은 성능을 나타냄을 알 수 있으며, 약 10%~13% 정도의 성능 향상을 보였다.

표 8 기존연구와의 비교 실험결과

부분	제안한 방법 (W_{Term})	C-value방법 [3]	빈도수기반방법 [4]
1	96.59%	72.17%	63.57%
2	93.64%	87.67%	77.98%
3	93.72%	92.33%	91.09%
4	93.95%	86.05%	92.87%
5	89.38%	91.78%	92.64%
6	84.42%	81.47%	90.78%
7	74.03%	71.55%	80.62%
8	75.50%	66.36%	79.46%
9	63.18%	73.88%	63.49%
10	27.44%	68.84%	59.07%
11-pt avg	92.08%	83.06%	81.20%

본 장에서는 전문용어 추출 성능을 기존 연구와 비교실험하였다. 본 논문에서 제시한 기법은 기존 연구보다 좋은 성능을 나타내었다.

4.4 오류 분석

실험 결과 다음과 같은 오류에 의해 전문용어가 추출되지 못하는 경우가 있었다.

첫 번째는 태깅오류이다. 예를 들어, 전문용어로 판별되지 않은 전문용어 후보 중에 '여러 개의 모노미디어 데이타'가 있었다. '여러 개의 모노미디어 데이타'의 경우, '여러'와 '개'가 각각 관형사와 의존명사로 태깅되어야 함에도 불구하고, 모두 일반명사로 태깅되어 '모노미디어 데이타'라는 전문용어를 추출하는 문제점이 발생하였다.

두 번째는 내포된 전문용어의 오류이다. 예를 들어, '기계번역 시스템'의 경우 '기계번역'과 '기계번역 시스템' 모두가 전문용어이지만 문서에서 '기계번역 시스템'이 나타날 경우, '기계번역 시스템'은 전문용어로 추출될 수 있지만, '기계번역'은 전문용어로 추출하지 못할 가능성이 있다.

세 번째는 모든 가중치가 낮은 경우 전문용어를 추출하지 못하는 경우가 발생한다. 예를 들어 "신경망의 학습"은 문서에서의 빈도수가 낮고, 음차표기된 외래어가 없다. 또한, 57개분야 사전에 "학습"만이 등재되어 있고, "학습"의 경우 "[기초과학], [생물], [의학], [전기], [전산], [전자]" 분야에 나타났다. 이러한 이유로 W_{Stat}, W_{Trl}, W_{Dic} 각각에 의해 낮은 가중치가 할당되어 전문용어를 효율적으로 추출하지 못하였다.

향후 이러한 태깅 오류, 내포된 전문용어 처리, 그리고 모든 가중치가 낮은 경우에 대한 보완이 필요하다고 하겠다.

5. 결 론

본 논문에서는 사전간의 계층관계를 이용한 분야 유사도와 빈도수와 음차표기된 외래어에 기반한 문서내 통계정보를 이용하여 전문용어를 추출하는 방법에 대하여 기술하였다. 본 논문에서는 사전간의 계층관계와 문서내 통계정보를 이용한 용어의 가중치를 계산하여 '사전에 수록된 전문용어'와 '사전에 수록된 용어를 이용하여 새로이 생성된 전문용어' 그리고 '사전에 수록되어 있지 않은 새로운 전문용어'를 자동적으로 추출하였다. 사전간의 계층관계는 클러스터링 방법에 의해 구축되고 구축된 사전 계층관계로부터 분야간의 유사도를 유추하여 전문용어를 추출하는 데 사용되었다. 문서내 통계정보에 의한 가중치는 빈도수에 기반한 방법과 음차 표기된 외래어와 외국어에 기반한 방법을 이용하였다. 빈도수를 이용한 방법은 괄호표현에 의해 나타나는 대역쌍과 약어쌍과 용어의 생산성을 용어의 빈도수와 결합하여 가중치를 부여하였다. 용어의 생산성은 내포관계로서 파악하고 용어의 생산성이 높을수록, 즉 새로운 전문용어를 생성하는 데 많이 사용되는 용어에 대하여 높은 가중치를 부여한다. 또한 음차 표기된 외래어와 외국어를 주어진 용어에서 추출하여 주어진 용어에서 많은 부분을 차지할수록 높은 가중치를 부여하였다. 본 논문에서는 효율적으로 전문용어를 추출하기 위하여 이들 가중치들 (사전 가중치, 빈도수에 기반한 가중치, 음차 표기된 외래어 및 외국어에 기반한 가중치)을 하나로 통합하여 사용하였다. 실험결과는 본 논문의 기법이 기존의 연구[3][4]보다 좋은 성능을 나타내었으며, 특히 본 논문의 기법이 보다 효율적으로 전문용어를 추출함을 알 수 있었다.

향후 연구로는 명사가 아닌 전문용어 [19], 전문용어를 구성하는 형태소의 변형 [20], 문맥정보의 이용 [11] 등을 이용한 전문용어 추출에 대한 연구가 필요하다. 또한, 본 논문의 기법의 효용성을 검증하기 위해서는 정보검색 시스템과 형태소 분석기와 같은 자연언어처리 시스템에 적용하는 것이 필요하다.

참 고 문 헌

[1] Bourigault, D., "Surface grammatical analysis for the extraction of terminological noun phrases," In Proceedings of the 14th International Conference on Computational Linguistics, COLING92, pp. 977-981, 1992.

[2] Dagan, I. and K. Church, "Termight: Identifying and translating technical terminology," In Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, EACL95, pp 34-40, 1995.

[3] Frantzi, K.T. and S.Ananiadou, "The C-value/NC-value domain independent method for multi-word term extraction," Journal of Natural Language Processing, Vol. 6, No.3, pp. 145-180, 1999.9.

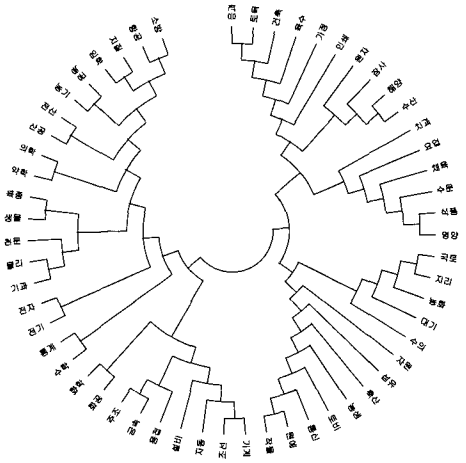
[4] Justeson, J.S. and S.M. Katz, "Technical terminology : some linguistic properties and an algorithm for identification in text," Natural Language Engineering, Vol.1, No.1, pp. 9-27, 1995.

[5] Lauriston, A., "Automatic Term Recognition: performance of Linguistic and Statistical Techniques," Ph.D. thesis, University of Manchester Institute of Science and Technology. 1996.

[6] Felber Helmut, Terminology Manual, International Information Centre for Terminology(Infoterm), 1984.

[7] ETRI, Etri-Kemong set, 1997.
[8] Anderberg, M.R, Cluster Analysis for Applications, New York: Academic, 1973
[9] Murtagh, F., "A Survey of Recent Advances in Hierarchical Clustering Algorithms," Computer Journal, Vol.26, pp. 354-359, 1983.
[10] Lorr, M., "Cluster Analysis and Its Application," Advances in Information System Science, Vol.8, pp.169-192, 1983.
[11] Maynard, D. and Ananiadou, S., "Acquiring Context Information for Term Disambiguation," In First Workshop on Computational Terminology Computerm98, pp 86-90, 1998.
[12] Hisamitsu, Toru and Yoshiki Niwa, "Extraction of useful terms from parenthetical expressions by using simple rules and statistical measures," In First Workshop on Computational Terminology Computerm98, pp 36-42, 1998.
[13] 이재성, 다국어 정보검색을 위한 영-한 음차 표기 및 복원 모델 박사학위 학위논문, 한국과학기술원 전산학과, 1999
[14] 오종훈, 최기선, "은닉 마르코프 모델을 이용한 과학기술문서에서의 외래어 자동 추출 모델", 제11회 한글 및 한국어 처리 학회 논문집 pp. 137-141, 1999.
[15] Oh, Jong-Hoon and Key-Sun Choi, "Automatic Extraction of Transliterated Foreign Words using Hidden Markov Model," ICCPOL'2001, pp. 433-438, 2001.
[16] 박영찬, 최기선, 김재군, 김영환, "한국어 정보검색을 위한 시험용 데이타 모음 2.0 개발", 1996년도 한국정보과학회 인공지능 연구회 춘계학술 대회, pp.59-65, 1996.
[17] 강인호, 김길창, "최대 엔트로피 모델을 이용한 한국어 품사 태깅", 제10회 한글 및 한국어 정보처리 학술대회, pp.9-14, 1998.
[18] Ricardo B-Y. and Berthier R-N., "Mordern Information Retrieval," ACM-Press New York and Addison-Wesley, 1999.
[19] Klavans, J. and Kan M.Y., "Role of Verbs in Document Analysis," In Proceedings of the 17th International Conference on Computational Linguistics, COLING98 pp. 680-686, 1998.
[20] 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.

부록. 사전계층관계 구성도



오 종 훈
1998년 성균관대학교 정보공학과 졸업 (학사). 2000년 한국과학기술원 전산학과 졸업 (공학석사). 2000년 ~ 현재 한국과학기술원 전산학과 박사과정. 관심분야는 자연언어처리, 전문용어, 정보검색 등

이 경 순
1994년 계명대학교 컴퓨터공학과 졸업 (학사). 1997년 한국과학기술원 전산학과 졸업 (공학석사). 2001년 한국과학기술원 전산학과 졸업 (공학박사). 2001년 ~ 현재 일본 NII(National Institute of Informatics) 연구원. 관심분야는 정보검색, 정보추출, 언어처리 등

최 기 선
1978년 서울대학교 수학과 졸업 (학사). 1980년 한국과학기술원 전산학과 졸업 (공학석사). 1986년 한국과학기술원 전산학과 졸업 (공학박사). 1985년 ~ 1986년 한국외국어대학교 전산학과 조교수. 1987년 ~ 1988년 일본 NEC C&C 정보연구소 초빙연구원. 1988년 ~ 현재 한국과학기술원 전산학과 교수. 1998년 ~ 현재 한국과학기술원 전문용어언어공학연구센터 소장. 관심분야는 자연언어처리, 기계번역, 정보검색, 전문용어 등