사전간 계층관계를 이용한 전문용어 자동 추출 기법

오종훈, 이경순, 최기선
전문용어언어공학센터/한국과학기술원

Automatic Terminology Recognition using the Dictionary Hierarchy

Jong-Hoon Oh, Kyung-Soon Lee, Key-Sun Choi
KORTERM, Korea Advanced Institute of Science and Technology

요 약

기존의 통계에 기반한 용어 자동 추출 기법 (Automatic Term Recognition)은 비교적 좋은 성능의 결과를 보여왔다. 하지만 전문용어 사전 등의 정보를 이용하여 성능의 향상을 이룰 수 있는 여지는 여전히 남아 있다. 본 논문에서는 이러한 근거에 기반하여 전문용어 간의 계층정보를 전문용어 사전을 통하여 구축하고 이를 이용하여 전문용어를 추출하는 방법을 제안하고자 한다. 본 논문이 제안하는 기법은 기존의 방법에 비해 좋은 성능을 나타내었다.

1. 서론

기존의 통계에 기반한 용어 자동 추출 기법 (Automatic Term Recognition: ATR) [5][6][9][12][14]은 비교적 좋은 성능의 결과를 보여왔다. 하지만 여러 다른 정보를 이용하여 성능의 향상을 이룰 수 있는 여지는 여전히 남아 있다. 예를 들어, 전문용어 사전은 이미 존재하는 전문용어에 대한 정보를 제공함으로써, 전문용어 추출의 성능을 향상시킬 수 있다. 지금까지 용어 추출분야에 있어 기계가독형 사전이 사용되기 어려웠던 것은 사전을 구축하는데 있어 상당한 노력이 필요했기 때문이다. 하지만 기계가독형 언어자원을 구축하기 위한 도구들의 점진적인 개발은 전문용어 분야에 이러한 사전을 이용할 수 있는 새로운 계기를 마련하고 있다.

하지만 사전에 등재되지 않은 미등록어로 인해 사전 그 자체만으로는 전문용어를 효율적으로 추출할 수 없으며, 여전히 자동적으로 전문용어를 추출할 수 있는 방법이 필요하다. 이러한 관점에서 전문용어사전은 전문용어 자동 추출 기법에 사용되는 기존 전문용어의 언어 자원으로서 사용될 수 있다. 예를 들어, 컴퓨터 분야 용어인 '분산 데이터베이스'는 기존의 용어인 '분산'과 '데이터베이스'에 의해 만들어졌다. 그리고 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 전자분야의 사전에는 존재하지만 컴퓨터 분야에서도 사용되는 전문 용어이다. 이러한 특성을 이용하기 위해서는 분야들간의 상호 연관성을 결정할 필요가 있다. 정보 검색 분야에서 사용되는 계층적 클러스터링 방법은 이러한 목적에 부합하는 방법을 제공한다. 계층적 클러스터링 방법을 통하여 사전간 (분야간)의 계층 관계를 구축할 수 있으며, 이를 통하여 분야간의 연관성을 유추할 수 있다. 예를 들어 전자 분야의 용어는 컴퓨터 분야의 용어와 밀접한 관계를 가진다는 것을 유추할 수 있다. 따라서 전자분야 전문용어 사전의 용어는 컴퓨터 분야의 용어가 될 확률이 높게 된다[8].

최근의 용어 자동추출[5][6][9][12]은 용어를 추출하는 데 있어 부분 구문정보 (shallow syntactic information)를 이용하였다. 이들 연구에서는, 세 단계의 과정을 거쳐 용어를 추출한다. 첫번째 단계에서는 부분 구문정보를 이용하여 명사구를 추출한다. 두 번째 단계에서는 통계적 정보를 이용하여 추출된 명사구에 대하여 점수를 부여한다. 세 번째 단계에서는 점수에 의해 순위가 정해지고, 지정된 임계값에 의해 상위의 용어를 추출한다. 하지만, 통계적 정보만으로는 용어가 반복적으로 자주 나타나지 않는 작은 양의 코퍼스나 특정한 분야의 코퍼스에 대해서는 좋은 성능을 기대하기 어렵다.

본 논문은 이러한 문제점을 해결하기 위하여 사전간의 계층구조를 이용한다. 그리고 통계적 정보를 이용하여 사전에 나타나지 않는 미등록어에 대한 처리도 고려한다. 마지

막으로 사전과 통계적 정보에 의해 부여된 가중치는 하나의 통합된 값으로 전문용어 추출에 이용된다.

2. 전체 시스템 구조

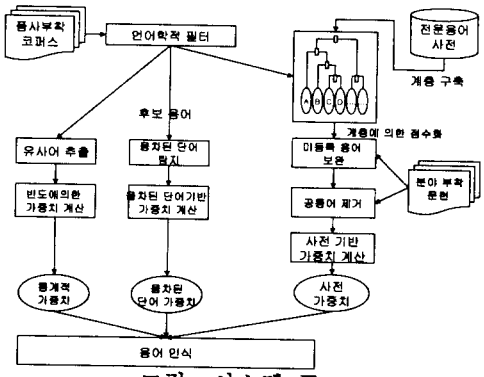

그림1 시스템 구조도

본 논문에서 제안하는 전문용어 추출 방법의 전체 과정은 그림 1에 나타나 있다. 본 논문에서 제안하는 방법은 네 단계의 과정으로 이루어진다. 첫 번째 단계는 전처리 단계로 클러스터링 기법에 의해 사전간의 계층관계가 구축된다. 두 번째 단계에서는 부분 구문정보[1]를 이용하여 명사구를 추출하며, 세 번째 단계에서는 추출된 명사구에 대하여 사전과 통계 정보를 이용하여 가중치가 주어진다. 사전에 의한 가중치 기법은 해당 명사구가 나타난 전문용어 사전의 개수에 기반하여 가중치를 부여한다. 그리고 사전에 수록되어 있지 않은 미등록어와 사전에 나타난 전문용어의 개념을 기술하기 위해 필요한 일반용어에 대한 고려를 위하여 분야정보가 표시된 문서를 이용한다. 명사구는 해당 문서에서 나타난 출현빈도와 음차 표기된 단어의 개수에 의하여 가중치가 부여된다. 네 번째 단계에서는 각각의 가중치가 하나의 값으로 통합되고 이를 이용하여 전문용어를 추출한다.

3. 사전간 계층관계

분야간의 상호 연관성은 전문용어를 추출 시 중요한 요소가 될 수 있다. 어떠한 분야의 전문용어를 추출할 때, 해당 분야 혹은 인접분야의 사전에 나타나는 단어와 전혀 다른 분야에서 나타나는 단어의 가중치와 의미는 다르다고 할 수 있다. 이러한 의미에서 사전 간의 계층관계를 분야간의 계층관계라고 생각할 수 있다. 본 장에서는 사전간의 계층관계를 구축하고 이를 이용하여 용어의 가중치를 결정하는 방법에 대하여 기술한다.

3.1 사전 계층관계를 위한 데이터

사전 간의 계층관계는 이중언어 사전 (영어-한국어)을 이용하여 구축한다. 사전은 과학기술분야의 57개 사전[2]을 이용한다. 그리고 모든 사전에 나타나지 않은 미등록어와 사전에 나타나는 일반용어를 보완하기 위하여 분야 정보가 포함된 ETRI-Kemong 문서집합을 이용하였다[7].

3.2 사전간 계층관계의 구축

사전 간의 계층관계를 구축하기 위해서 클러스터링 방법이 사용된다. 클러스터링 방법은 문서간의 유사성을 이용하여 분야 구조를 구성하는 통계적 기법으로 계층적 클러스터링과 비계층적 클러스터링이 있다[4]. 본 논문에서는 이러한 클러스터링 방법 중에서 계층적 클러스터링 방법을 사용하였으며, 계층적 클러스터링 방법 중 비교적 대칭적인 구조의 계층구조를 만들어 내며, 클러스터링 각 과정에서 전체 그룹 오류 합 (the total within-group error sum)의 증가가 최소화되는 방법 (Lorr, 1983)인 «상호 최근 인접이웃 알고리즘» (a reciprocal nearest neighbor algorithm)[16]을 사용하였다. 이 알고리즘의 수행과정은 다음과 같다.

1. 모든 개체 (사전)간의 유사도를 결정한다.
2. 가장 유사한 개체를 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 개체들을 $D_i = (x_{i1}, x_{i2}, \ldots, x_{iL})$와 같이 벡터로 나타낸다. 1단계에서 개체간 유사도는 유클리디안거리를 이용하여 계산한다. 2단계에서 상호 가장 유사한 개체는 상호 최근 인접 이웃

[1] 부분 구문 정보는 명사구를 추출하기 위한 규칙을 나타낸다.

[2] 사용된 사전의 분야는 농화학, 물리학, 생물학, 수학, 식품학, 기상학, 주조, 용접, 치의학, 의학, 전자공학, 전산학, 전기공학, 화학 등이다.

에 의해 결정된다. 주어진 개체 i와 j에 대하여 i와 가장 유사도가 높은 개체가 j이고, j와 가장 유사도가 높은 개체가 i일 때 이들 i와 j는 상호 최근 인접 이웃이라고 정의된다. 이러한 이유로 기술된 알고리즘이 상호 최근인접 이웃 알고리즘이라 불린다. 이 알고리즘을 이용하여 사전간의 계층관계는 그림2와 같이 나타내어진다. 그림 2에서 계층관계를 구성하고 있는 사전은 10개 분야의 사전이며, 전체 57개 분야의 사전으로 구성된 계층관계의 일부를 나타내고 있다.

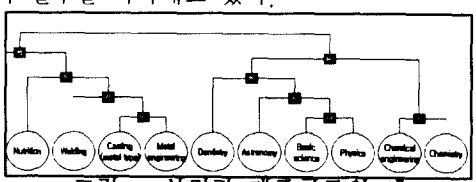

그림 2. 사전간 계층구조의 예

3.3 계층관계를 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 이러한 전제를 기반하여 계층을 이용한 연관성을 계산하고 이를 이용하여 용어의 가중치를 부여한다.

$$similarity_{ij} = \frac{2 \times Common_{ij}}{depth_i + depth_j} \qquad (3.3)$$

이를 위해 다음과 같은 3가지 단계의 과정이 필요하다.

1. 식 (3.3)을 이용하여 분야간 유사도를 계산한다[17].

여기서, $Depth_i$는 계층에서는 분야의 깊이 정보를, $Common_{ij}$는 분야$_i$와 분야$_j$ 간에 공유하는 가장 깊은 노드의 깊이 정보를 나타낸다. 식 (3.3)에서 계층의 노드 깊이는 계층의 루트로 부터 해당 노드까지의 거리를 나타낸다. - root의 깊이는 1로 가정한다. 예를 들어 그림 2에서 노드 C1과 C8의 부모노드를 루트 노드라 가정하였을 경우 '화학'과 '화학공학'의 유사도는 표 2와 같이 계산된다.

분야	$Depth_i$	$Common_{ij}$	$Similarity_{ij}$
화학	4	3	2*3/(4+4) =0.75
화학공학	4	3	2*3/(4+4) =0.75

Table 2. $similarity_{ij}$ 의 계산

2. 추출하고자 하는 분야와 용어가 나타난 사전의 분야와의 거리는 식 (3.4)에 의해 계산된다.

$$Score(term) = \frac{1}{N}\sum_{i=1}^{N} similarity_{ti} \qquad (3.4)$$

여기서 N은 용어가 나타난 사전의 개수, $Similarity_{ti}$는 식(3.3)에서 계산된 추출하고자 하는 분야와 용어가 나타난 사전의 분야와의 유사도를 나타낸다. 예를 들어, 그림 2에서 물리학 분야의 전문용어를 추출하고자 하고 후보 용어가 물리학, 화학, 천체학 사전에 나타난다면 용어의 가중치는 표 3과 같이 계산된다.

N	3
$similarity_{physics\text{-}chemistry}$	0.4
$similarity_{physics\text{-}physics}$	1
$similarity_{physics\text{-}astronomy}$	0.7
Score	2.1/3 = 0.7

Table 3. 식 (3.4)에 의해 부여된 가중치의 예

3. 사전 미등록어의 보완 : 분야정보가 부착된 문서집합의 이용.

$$DWeight(term) = (Score(term)+1) * \sqrt{\frac{\sum_{i=1}^{W} df_i}{W}} \qquad (3.5)$$

여기서, W는 용어 후보에 포함된 단어의 개수, df_i는 분야정보가 부착된 문서집합에서 나타난 단어의 분야 개수를 나타낸다. 사전에 나타나지 않은 미등록어와 전문분야 사전에 특별한 의미를 기술하기 위하여 사용되는 일반적인 단어들은 사전 정보만으로는 해결할 수 없다. 미등록어는 새로이 생성된 용어일 가능성이 높기 때문에 용어 추출에 중요한 정보가 될 수도 있으며, 반대로 일반용어는 전문용어로 인식되지 말아야 되는 용어이기 때문에 제거되어야 한다. 본 논문에서는 이를 위하여 분야정보가 부착된 문서집합[7]을 이용하여 사전 정보를 보완하였다. 후보 용어에 포함되어 있는 단어는 분야정보가 부착된 코퍼스에서 나타난 분야의 개수에 의해 그 전문성을 추정한다. 만약, 해당 용어의 단어가 여러 분야에 걸쳐 나타난다면, 일반용어가 될 가능성이 높으며, 특정 분야에 한정되어 나타난다면 이는 전문용어가 될 가능성이 높다.

본 논문에서는 사전간 계층정보에 부여된

용어의 가중치를 사전 가중치라 정의하고 DWeight라 나타낸다.

4. 통계적 방법

통계적 방법은 두 가지의 요소로 이루어진다. 첫번째 요소는 통계적 가중치라 나타내며, 용어들의 문서에 나타난 출현빈도에 기반하여 가중치가 부여된다. 두 번째 요소는 음차 표기 단어 가중치라 나타내며, 해당 용어가 포함하는 음차 표기된 외래어의 개수에 기반하여 가중치가 부여된다.

4.1 통계적 가중치: 빈도수에 기반한 가중치

통계적 가중치를 계산하기 위하여 문서에서 나타나는 괄호 표현에 의한 유사어쌍과 용어들의 문서에서의 출현빈도를 고려한다. 우선 유사어쌍은 다음과 같은 휴리스틱을 이용하여 추출한다.

주어진 괄호표현 A(B)에 대하여,

1. A와 B가 약어와 그 확장어의 쌍인지를 검사한다. 이를 위해 A와 B의 대문자를 비교하여 반 이상이 순서적으로 일치하면 약어쌍이라고 판단한다[10]. 예를 들어, 'ISO (International Standardization Organization)'의 'ISO'와 'International Standardization Organization'은 이러한 휴리스틱을 이용하여 약어쌍이라 판단된다.
2. A와 B가 번역쌍인지를 검사한다. 이를 위해 이중언어 사전을 이용한다.

만약, A와 B가 약어쌍이거나 번역쌍이면 A와 B를 유사어쌍이라 판단한다. 이러한 유사어 쌍을 추출한 뒤 통계적 가중치 SWeight는 식 (4.1)에 의해 계산된다.

$$SWeight(a) = \sum_{b \in Similar(a)} \left[\sqrt{|a|} \times \left(f(a) + \lambda \times \frac{\sum_{c \in Ta} f(c)}{C(Ta)} \right) \right] \qquad (4.1)$$

여기서, a는 후보 용어, |a|는 용어 a의 어절수, Ta는 용어 a를 내포하는 용어들의 집합, f(a)는 문서에서 용어 a의 출현빈도, C(Ta)는 용어 a를 내포하는 용어의 종류의 개수, λ는 용어 a가 문서에서 내포된 경우가 있으면 0, 그렇지 않으면 1을 각각 나타낸다.

식 (4.1)에서 내포관계는 다음과 같이 정의된다. A와 B를 용어라 하고, A가 B를 포함하면 A가 B를 내포한다고 정의한다. 예를 들어, '분산 데이터 베이스'와 '데이터 베이스'에서 '분산 데이터 베이스'는 '데이터 베이스'를 내포한다라고 말한다. 식 (4.1)은 유사어쌍의 통계적 가중치는 같아야 된다는 것과 용어가 내포되지 않은 경우일 수록 가중치가 높다라는 것을 내포하고 있다.

4.2 음차 표기 단어 가중치: 음차 표기된 외래어의 개수에 기반한 가중치

전문용어는 주로 음차 표기되는 경우가 많기 때문에 음차 표기된 용어는 주어진 분야의 용어를 추출하기 위한 중요한 단서가 될 수 있다. 그런데, 이러한 음차 표기된 외래어는 표기에 대한 표준이 있음에도 불구하고 사용자마다 그 형태가 달리 표기되기 때문에 사전에 수록되어 있지 않은 경우가 많다. 따라서 어떠한 용어가 음차 표기를 포함하는가를 사전에 의존해서 판단하는 것은 어려움이 있으며, 이를 자동적으로 추출하는 방법이 필요하다. 본 논문에서는 은닉 마르코프 모델을 이용한 외래어 자동 추출 모델 [3]을 이용하여 외래어를 자동적으로 추출하였다. 자동적 외래어 추출방법은 한국어와 외국언어가 음성학상으로 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 표기 된다. 그런데, 한국어에서 이들 자음들은 순수 한국어에서 자주 사용되지 않는 자음들이다. 이러한 특성은 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 추출하는데 중요한 정보로 사용될 수 있다.

은닉 마르코프 모델을 이용한 외래어 추출 모델은 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 '시스템'은 '시/F + 스/F + 템/F'라고 태그 될 수 있다. 외래어 자동 추출 모델에서는 자음정보를 POS (Part-of-Speech)태깅에서의 어휘정보와 같이 사용하였다. 식 (4.2)는 은닉 마르코프 모델을 이용한 외래어 추출모델을 나타낸 식이며, 식 (4.3)은 추출된 외래어에 따라 주어진 용어의 가중치를 할당하는 식이다. 식(4.3)을 Tweight라고 정의하고 전문용어 추출에 사용한다. 식(4.3)은 음차 표기된 외래어를 많이

포함할수록 전문용어일 가능성이 높다라는 의미를 내포하고 있다.

$$P(T \mid S)P(S) = p(t_1)p(t_2 \mid t_1)$$
$$\left[\prod_{i=3}^{n} p(t_i \mid t_{i-1}, t_{i-2})\right]\left[\prod_{i=1}^{n} p(s_i \mid t_i)\right] \quad (4.2)$$

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

$$TWeight(a) = \frac{trans(a)}{\sqrt{|a|}} \quad (4.3)$$

여기서 |a|는 용어 a의 어절 수를 trans(a)는 용어 a에서 음차 표기된 외래어를 포함하는 어절 수를 나타낸다.

5. 용어의 가중치

위에서 기술한 3가지 가중치는 식 (5.1)에 의하여 통합되어, TERMWeight라 정의되며 적절한 전문용어를 추출하는데 사용된다. 식 (5.1)에서 세 가지 각각의 가중치는 더해져서 TERMWeight로 정의된다.

$$TERMWeight(term) = DWeight(term) + SWeight(term) + TWeight(term) \quad (5.1)$$

6. 실험 및 평가

본 논문에서 제안한 전문용어 추출방법은 컴퓨터분야와 전기전자분야의 문서를 포함하는 정보검색 테스트 집합인 KT문서 집합을 사용하여 실험하였다. 본 논문에서는 이 중 컴퓨터분야의 문서만을 추출하여 실험에 사용하였다 [2]. KT문서 집합은 명사구를 추출하기 위한 품사정보를 얻기 위하여 POS 태거 [1]로 자동적으로 태깅되었다. 전문용어 추출에 대한 결과 평가는 DWeight만을 이용하여 전문용어를 추출하여, 본 논문이 제안하는 사전간의 계층관계가 전문용어 추출에 얼마나 유효한가를 살펴보고, 전체 시스템에 대한 평가를 기존 연구인 C-value 방법 [9]과 비교 평가함으로써 본 논문이 제안하는 기법의 효용성을 살펴보고자 한다.

6.1 평가 기준

두 명의 분야 전문가가 제안된 전문용어추출 방법에 의해 추출된 용어에 대한 평가를 하였으며, 두 명의 전문가가 평가한 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 포함된 올바른 전문용어의 비율을 나타내는 정확율로서 평가된다.

6.2 사전 가중치

사전 가중치는 추출하고자 하는 분야의 정보가 전문용어를 추출하는데 중요한 정보로 사용될 수 있다는 전제에 기반한다. 추출하고자 하는 분야와 이와 밀접하게 연관된 분야의 사전들에 수록되어 있는 용어들은 전문용어를 추출하는데 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.

	Top 10%	Bottom 10%
Term	94%	54.8%
Non-Term	6%	45.2%

Table 4. 사전 가중치만을 이용한 전문용어 추출 결과

표 4는 사전 가중치만으로 전문용어를 추출하였을 경우의 결과를 나타낸다. 표 4는 다음과 같이 해석될 수 있다. 사전 가중치에 의해 추출된 결과의 상위 10% 중 94%가 전문용어이며, 나머지 6%가 전문용어가 아니다. 또한 하위 10% 중 54.8%가 전문용어이며, 45.2%가 전문용어가 아니다. 이는 사전 가중치에 의해 추출된 전문용어가 하위에서보다 상위에서 보다 많은 비율로 나타난다고 볼 수 있다.

6.3 전체 시스템 평가

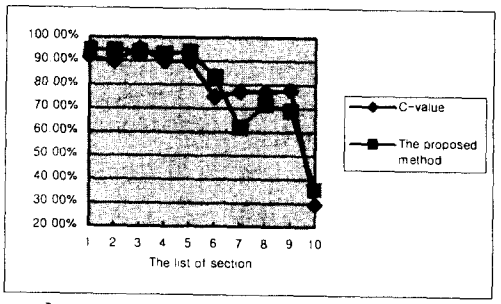

그림 3. C-value와 사전간 계층관계를 이용한 전문용어 추출 방법과의 비교

그림 3은 본 논문에서 제안한 전문용어 추출 기법과 기존의 연구와의 성능을 비교한 결과를 나타낸다. 결과의 비교는 각 방법에서 사용한 가중치에 의해 순위화 된 전체 전

문용어를 10개의 부분으로 나누어 비교하였다. 각 부분은 1290개의 단어들로 이루어 졌으며, 각각이 독립적으로 서로 비교되었다. 그림 3에 나타난 결과는 다음과 같이 해석될 수 있다. 결과의 첫번째 부분에서 제안된 방법은 기존 방법보다 높은 정확율을 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, 그림3에서 5번째 부분부터 10번째 부분까지의 정확율이 급격히 감소하는 것을 알 수 있는데 이는 대부분의 전문용어가 상위에 존재한다는 것을 나타낸다.

7. 결론

본 논문에서는 사전간의 계층관계를 이용한 전문용어 추출방법에 대하여 기술하였다. 사전간의 계층관계는 클러스터링 방법에 의해 구축되어 분야간의 관계를 유추하는 도구로서 사용되었다. 실험결과는 본 논문의 기법이 기존의 연구보다 좋은 성능을 나타내는 것을 알 수 있었으며 특히 본 논문의 기법이 보다 효율적으로 전문용어를 추출함을 알 수 있었다 (기존 연구에 비해 상위부분에 보다 많은 전문용어가 포함되어 있었다).

하지만 본 논문의 기법은 여전히 개선의 여지가 남아 있다. 명사가 아닌 전문용어 [13], 전문용어의 변형 [11], 문맥정보의 이용 [17] 등은 전문용어 추출의 성능을 향상 시킬 수 있는 여지를 가지고 있다. 그리고 본 논문의 기법의 효용성을 검증하기 위해서는 정보검색 시스템과 형태소 분석기와 같은 NLP 시스템에 적용하는 것이 필요하겠다.

참고 문헌

[1]강인호,김길창 (1998), *최대 엔트로피 모델을 이용한 한국어 품사 태깅*, 제 10 회 한글 및 한국어 정보처리 학술대회

[2]김성혁 외 (1994), *자동 색인기 성능 시험을 위한 Test Set 의 개발*, 정보관리학회지 제 11 권 1 호, 929-932

[3]오종훈, 최기선 (1999), *은닉 마르코프 모델을 이용한 과학기술문서에서의 외래어 자동 추출 모델*, 제 11 회 한글 및 한국어 처리 학회 논문집 pp. 137-141.

[4] Anderberg, M.R. (1973) *Cluster Analysis for Applications*. New York: Academic

[5] Bourigault, D. (1992) *Surface grammatical analysis for the extraction of terminological noun phrases*. In Proceedings of the 14th International Conference on Computational Linguistics, COLING'92 pp. 977-981.

[6] Dagan, I. and K. Church. (1995) *Termight: Identifying and translating technical terminology*. In Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, EACL'95, pp 34-40

[7] ETRI (1997) Etri-Kemong set

[8] Felber Helmut (1984) *Terminology Manual*, International Information Centre for Terminology (Infoterm)

[9] 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

[10] Hisamitsu, Toru and Yoshiki Niwa (1998) *Extraction of useful terms from parenthetical expressions by using simple rules and statistical measures*. In First Workshop on Computational Terminology Computerm'98, pp 36-42

[11] Jacquemin, C., Judith L.K. and Evelyne, T. (1997) *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

[12] 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

[13] Klavans, J. and Kan M.Y (1998) *Role of Verbs in Document Analysis*, In Proceedings of the 17th International Conference on Computational Linguistics, COLING'98 pp. 680-686.

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

[15] Lorr, M. (1983) *Cluster Analysis and Its Application*, Advances in Information System Science,8 , pp.169-192

[16] Murtagh, F. (1983) *A Survey of Recent Advances in Hierarchical Clustering Algorithms*, Computer Journal, 26, 354-359

[17] Maynard, D. and Ananiadou, S. (1998) *Acquiring Context Information for Term Disambiguation* In First Workshop on Computational Terminology Computerm'98, pp 86-90