구성정보와 문맥정보를 이용한 전문용어의 전문성 측정 방법

(Determining the Specificity of Terms using Compositional and Contextual Information)

류 법 모 † 배 선 미 †† 최 기 선 †††
(Pum-Mo Ryu) (Sun-Mee Bae) (Key-Sun Choi)

요 약 어떤 용어가 전문적인 개념을 많이 내포하고 있을 때 전문성이 높다고 말한다. 본 논문에서는 용어의 내부 구성정보와 외부 문맥정보를 이용하여 정보이론에 기반한 방법으로 전문용어가 내포하는 전문성을 정량적으로 계산하는 방법을 제안한다. 용어의 전문성은 용어간 상하위어 관계 설정에서 중요한 필요조건으로 사용될 수 있다. 제안한 방법은 전문용어의 내부 구성정보를 이용하는 방법, 문맥정보를 이용하는 방법 그리고 두 정보를 모두 이용하는 방법으로 나눈다. 구성정보를 이용하는 방법에서는 전문용어를 구성하는 단어의 빈도수, 가중치, 바이그램, 내부 수식구조 등을 이용하고, 문맥정보를 이용하는 방법에서는 전문용어를 수식하는 단어들의 분포를 이용한다. 본 논문에서 제안한 방법은 분야에 독립적으로 적용될 수 있고, 전문용어 생성 절차에 대한 특징을 잘 반영할 수 있는 장점이 있다. MeSH 트리에 포함된 질병 이름의 전문성 값을 계산한 뒤 상위어의 전문성 값과 비교한 결과 82.0%의 정확률을 보였다.

키워드 : 용어의 전문성, 상하위 관계, 전문용어, 정보이론, 코퍼스, 구성정보, 문맥정보

Abstract A term with more domain specific information has higher level of term specificity. We propose new specificity calculation methods of terms based on information theoretic measures using compositional and contextual information. Specificity of terms is a kind of necessary conditions in term hierarchy construction task. The methods use based on compositional and contextual information of terms. The compositional information includes frequency, tf·idf, bigram and internal structure of the terms. The contextual information of a term includes the probabilistic distribution of modifiers of terms. The proposed methods can be applied to other domains without extra procedures. Experiments showed very promising result with the precision of 82.0% when applied to the terms in MeSH thesaurus.

Key words : Term specificity, Hypernymy, Terminology, Information theory, Corpus, Compositional information, Contextual information

1. 서 론

사회가 빠른 속도로 발전하면서 새로운 전문분야가 지속적으로 만들어지고 있으며, 기존의 전문분야도 시대에 따라 성격이 변하고 있다. 지금까지 대부분의 전문분야 지식은 해당 분야 전문가들이 관리하고 있다. 그러나 이 방법은 빠르게 변화되는 지식을 효율적으로 반영하기 어려운 단점이 있기 때문에 자동으로 전문분야 지식을 관리는 방법이 활발히 연구되고 있다. 전문용어는 전문분야의 개념이 언어적으로 표현된 형태이다[1]. 따라서 전문용어는 전문분야 지식의 기본 단위로 사용되고 있으며, 전문용어 관리는 전문분야 지식 관리에서 핵심적인 부분을 차지한다.

용어의 전문성(specificity)은 용어가 포함하는 전문적인 정보의 양을 정량적으로 표현한 것이다. 어떤 용어가 도메인 전문적인 정보를 많이 포함하고 있을 때 전문성이 높고, 반대로 일상적인 용어일수록 전문성이 낮다고 가정한다. 본 연구에서는 용어의 구성정보와 문맥정보를 이용하여 주어진 도메인 D에서 사용되는 용어 t의 전문성을 식 (1)과 같이 실수로 표현하는 방법을 제안한다.

† 학생회원 : 한국과학기술원 전산학과
pmryu@world.kaist.ac.kr
†† 비 회 원 : 한국과학기술원 인문사회과학부 교수
sunmee@kaist.ac.kr
††† 종신회원 : 한국과학기술원 전산학과 교수
kschoi@cs.kaist.ac.kr
논문접수 : 2004년 4월 20일
심사완료 : 2006년 5월 12일

표 1 MeSH[1] 트리의 일부분. 노드 번호는 용어 사이의 계층구조를 나타낸다.

노드 번호	용어
C18.452.297	diabetes mellitus (당뇨병)
C18.452.297.267	insulin-dependent diabetes mellitus (인슐린 의존형 당뇨병)
C18.452.297.267.960	Wolfram syndrome (볼프람 증후군)

$$Spec(t \mid D) \in R^{+} \quad (1)$$

전문분야 개념은 자신을 다른 개념들과 구분시킬 수 있는 고유한 특징 집합을 가진다. 비슷한 특징 집합을 가지는 개념들은 유사한 의미를 표현한다. 어떤 개념을 표현하는 특징 집합에 새로운 특징을 추가하여 더 전문적인 개념을 만들 수 있다. 일반적으로 기존의 개념 X와 X에 새로운 특징을 추가하여 생긴 개념 Y 사이에는 상하위 관계가 성립된다. 즉 X는 Y의 상위 개념이고, X의 특징 집합은 Y의 특징 집합의 부분집합이다[2]. 전문분야 개념이 전문용어로 표현될 때 다음과 같은 두 가지 특징을 관찰할 수 있다. 첫째, 기존의 전문용어에 새로운 특징을 추가하는 수식어를 부가하여 더 전문적인 개념을 표현하는 용어가 만들어진다. 예를 들어 표 1에서 "insulin-dependent diabetes mellitus"는 "diabetes mellitus"에 "insulin-dependent"라는 수식어가 부가되어 만들어진 더 전문적인 용어이다. 이 방법으로 생성된 전문용어는 추가된 수식어의 전문성만큼 전체 용어의 전문성이 증가한다. 이 경우에는 용어의 구성단어들이 용어의 특징을 표현하는 정보로 사용된다. 둘째, 기존 전문용어의 구성단어와 전혀 다른 단어를 이용하여 더 전문적인 개념을 표현하는 경우가 있다. 예를 들어 표 1에서 "Wolfram syndrome"은 상위어 "insulin- dependent diabetes mellitus"의 구성단어와 전혀 다른 단어들로 구성되어 있다. 이 경우에는 용어의 문맥정보가 용어의 특징을 표현하는 정보로 사용된다. 따라서 본 연구에서는 전문용어의 전문성 결정에서 용어의 구성정보와 문맥정보가 중요한 정보가 된다는 가정을 기반으로 용어의 전문성을 측정하는 방법을 제안하고, 기존의 용어 계층구조에 포함된 용어들을 대상으로 제안한 방법의 유효성을 평가한다.

용어의 전문성은 또한 용어간 상하위 관계 자동 설정 과정에서 적용될 수 있다. 전문적인 용어일수록 구체적인 개념을 표현하며 용어 계층구조에서 하위에 나타나는 경향이 있기 때문에, 용어의 전문성은 용어간 상하위 관계를 결정하기 위한 하나의 필요조건으로 사용할 수 있다. 주어진 도메인 D의 전문용어로 구성된 용어 계층구조 H_D에서 용어 t_1이 용어 t_2의 상위어인 경우 t_1의 전문성은 t_2의 전문성보다 낮다. 그림 1에서와 같이 두 용어 t_1과 t_2가 의미적으로 충분히 유사하고, t_1의 전문성이 t_2의 전문성보다 작을 경우, t_1이 t_2의 상위어가 될 가능성이 높다. 그러나 용어의 전문성은 상하위어 관계 표현을 위한 충분조건은 되지 못한다. 예를 들어 그림 1에서 t_1의 전문성이 t_3의 전문성보다 작지만 의미적으로 유사하지 않기 때문에 두 용어 사이에 상하위어 관계가 성립할 가능성은 낮다.

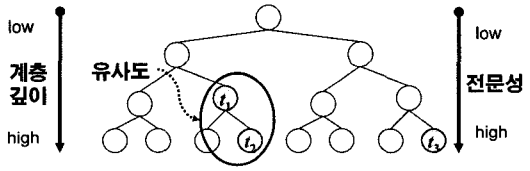

그림 1 전문 분야 용어 계층구조 H_D에서 용어의 전문성과 용어간 유사도. 두 용어 사이의 거리가 가까울수록 유사도가 높고, 용어의 계층 깊이가 깊을수록 전문성이 높다.

본 논문은 다음과 같이 구성된다. 2장에서는 용어의 구성정보와 문맥정보를 이용하여 정보이론에 기반한 방법으로 용어의 전문성을 측정하는 방법을 설명하며, 3장에서는 제안한 방법에 대한 실험과 평가가 소개되고, 마지막으로 4장에서는 결론 및 향후 연구를 소개한다.

2. 전문성 계산 방법

이 장에서는 용어의 구성정보와 문맥정보를 정보이론에 기반한 방법으로 정량화하는 방법을 설명한다.

정보이론에서는 정보량을 "불확실성" 또는 "놀라움"의 개념으로 설명한다. 출현 확률이 낮은 메시지가 채널의 출력에서 나타나기 전에는 "불확실성"이 높다고 이야기한다. "불확실성"이 높은 메시지가 실제로 나타난 경우 "놀라움"의 정도는 커지고, 그 메시지를 표현하기 위한 비트수는 다른 출력에 비해 길어진다. 따라서 그 메시지의 정보량은 높아진다[3]. 도메인 D와 관련된 코퍼스에서 나타나는 용어들이 어떤 채널의 출력에서 관찰되는 일련의 메시지라고 가정하면, 용어 t가 관찰되는 사건 x의 정보량 $I(x)$를 코퍼스의 각종 통계정보를 이용하여 계산할 수 있다. 그리고 $I(x)$ 를 식 (2)와 같이 용어 t

1) 미국 의학도서관(NLM, National Library of Medicine)에서 관리하는 의학용어 리스트이다. 용어들을 주제어라고 부르며, 주제어 사이의 상하위 관계 트리도 제공한다. 본 논문에서는 MeSH 2003 버전을 사용하였다. (http://www.nlm.nih.gov/mesh/)

의 전문성 $Spec(t|D)$으로 사용한다.

$$Spec(t \mid D) \approx I(x) \tag{2}$$

이 경우, 정보량 $I(x)$는 식 (3), (4), (5)와 같은 성질을 가진다.

$$I(x)=0, \quad p(x)=1 \text{ 일 때} \tag{3}$$

코퍼스에서 나타날 확률이 1인 용어 t가 실제 코퍼스에서 출현할 경우 얻을 수 있는 정보량은 없다.

$$I(x)\geq 0, \quad 0\leq p(x)\leq 1 \text{ 일 때} \tag{4}$$

용어 t가 코퍼스에서 나타날 경우, 정보의 손실을 초래하는 경우는 없다. 즉 코퍼스에서 나타나는 모든 용어는 정보량을 계산할 수 있으며, 0 이상의 값을 가진다.

$$I(x_i) > I(x_j), \quad p(x_i) \leq p(x_j) \text{ 일 때} \tag{5}$$

용어 t_i가 t_j 보다 코퍼스에서 나타날 확률이 낮을 때, 실제 코퍼스에서 t_i가 나타날 경우, 얻을 수 있는 정보량이 t_j가 나타날 경우 얻을 수 있는 정보량보다 많다. 즉 코퍼스에서 출현 확률이 낮은 용어일수록 정보량이 많아지고 전문성이 높아진다.

다음 장에서는 식 (2)의 $I(x)$를 계산하는 여러 가지 방법을 자세히 설명한다. 2.1장에서는 용어의 내부 구성 정보를 이용하는 방법, 2.2장에서는 용어의 문맥정보를 이용하는 방법, 2.3장에서는 두 가지 정보를 모두 이용하는 방법을 설명한다.

2.1 구성정보 기반 계산 방법(방법 1)

구성정보를 이용한 방법은 기존의 용어에 개념을 제한하는 단어를 추가하여 새로운 용어를 만드는 신조어 생성 특징을 반영하는 전문성 계산 방법이다. 구성 단어의 특징으로는 구성 단어의 출현 빈도수, 구성 단어의 가중치, 그리고 구성 단어의 바이그램 정보 등이 있다. 또한 용어의 구성 단어들이 수식어-피수식어 관계를 가지면서 상호 의존적이라는 정보를 추가적으로 이용한다.

2.1.1 구성단어의 특징을 이용한 계산 방법

한 용어를 구성하는 각각의 단어에 그 용어의 특징들이 분할되어 저장되어 있다는 가정을 하고, 각 구성단어의 특징을 정량화하여 전체 용어의 전문성을 계산한다. 이 계산 방법을 위하여 용어 t는 식 (6)과 같이 여러 개의 단어로 구성되어 있다고 가정한다.

$$t = w_1 w_2 ... w_m \tag{6}$$

여기에서 t는 한 개의 용어이고, w_i $(1\leq i\leq m)$는 t를 구성하는 단위 단어를 나타낸다. 예를 들어 "gestational diabetes mellitus"(임신당뇨병)는 세 개의 단위 단어 "gestational", "diabetes", "mellitus"로 구성된다. 용어를 구성하는 각 단어들이 서로 독립적이라고 가정을 하면 식 (2)의 $I(x)$는 식 (7)과 같이 각 구성단어들의 정보량의 합으로 정의된다.

$$Spec(t \mid D) = I(x) = -\sum_{i=1}^{m} p(x_i)\log p(x_i) \tag{7}$$

여기에서 $p(x_i)$는 단어 w_i가 코퍼스에서 나타나는 사건(x_i)의 확률을 나타낸다. 따라서 $p(x_i)$를 추정하면 해당 용어의 전문성을 계산할 수 있다. 다음은 $p(x_i)$를 추정하기 위한 3가지 정보를 차례로 설명한다.

정보 1. 구성단어의 출현 빈도수

이 방법에서는 채널의 출력에서 관찰될 확률이 낮은 단어가 실제로 관찰된 경우 높은 정보량을 가진다는 정보이론의 기본적인 아이디어를 따른다. 즉 코퍼스에서 출현확률이 낮은 단어들로 구성된 용어가 더 전문적이라는 가정에 기반한다. 발생 빈도수가 높은 단어는 여러 개의 전문용어에 공통적으로 나타나는 일반적인 단어이기 때문에, 자신을 포함하는 전문용어의 특징을 차별화시킬 수 있는 능력이 낮다. 반대로 발생 빈도수가 낮은 단어들은 적은 수의 전문용어에만 포함되기 때문에, 자신을 포함하는 전문용어의 특징을 차별화시킬 수 있는 능력이 높다. 예를 들어 MeSH 트리에서 다음의 두 용어를 생각해 보자.

- "inborn metabolic brain disease" (C18.452.100.100, 선천성 대사성 뇌질환)
- "Refsum disease" (C18.452.100.100.680.760, 레프섬병)

상위어 "inborn metabolic brain disease"를 구성하는 단어 "inborn", "metabolic", "brain"의 빈도수는 각각 1,296회, 34,407회, 18,735회[2]이고, 하위어 "Refsum disease"를 구성하는 단어 "Refsum"의 빈도수는 13회이다. 따라서 "Refsum"은 "inborn", "metabolic", "brain"에 비하여 다른 용어에 나타날 확률이 낮기 때문에 자신을 포함하는 용어를 차별화시키는 역할을 하므로 "Refsum disease"가 "inborn metabolic brain disease"보다 높은 전문성 값을 가진다. 이 가정에서 식 (7)의 $P(x_i)$는 식 (8)과 같이 추정한다.

$$p(x_i) \approx p_{MLE}(w_i) = freq(w_i) / \sum_j freq(w_j) \tag{8}$$

여기에서 $freq(w)$는 전체 코퍼스에서 단어 w의 빈도수를 나타낸다.

전문용어 자동 인식과 관련된 연구에서는 용어의 빈도수가 높을수록 전문용어일 가능성이 높다고 가정하고, 빈도수가 높은 전문용어 후보에 높은 점수를 부여하였다[4,5]. 그러나 이 방법에서는 전문용어의 빈도수가 아니고 전문용어를 구성하는 단어의 빈도수를 이용한다는 점에서 기존의 전문용어 인식 논문에서 제안하였던 방법과 차이가 있다.

2) 단어의 빈도수는 통계정보를 추출하기 위하여 사용된 코퍼스의 종류에 따라 다르다. 여기에서 제시한 단어의 빈도수는 3장의 실험방법에서 설명하는 코퍼스에서 추출한 값이다.

정보 2. 구성단어의 가중치

정보검색에서는 단어 빈도수(term frequency: tf)에 문서 빈도수의 역수(inverted document frequency: idf)를 곱한 tf·idf 를 색인어의 가중치 계산에 가장 널리 사용한다[6]. 단어 t 의 tf·idf 값은 식 (9)와 같이 계산된다.

$$tf \cdot idf(t) = \begin{cases} (1+\log tf(t))\log\dfrac{N}{df(t)} & if\ \ tf(t) \geq 1 \\ 0 & if\ \ tf(t) = 0 \end{cases} \quad (9)$$

여기에서 N 은 전체 문서의 개수를 나타낸다. 빈도수가 높으면서 제한된 문서에 집중적으로 나타나는 단어가 높은 가중치를 가진다. 가중치가 높은 단어는 특정 문서를 다른 문서와 차별화시키는 대표적인 단어의 역할을 하기 때문에 전문적인 정보를 많이 포함하고 있다고 할 수 있다. 따라서 용어 t 에 가중치가 높은 단어들이 많이 포함된 경우 전문성이 높다고 가정한다. 용어를 구성하는 모든 단위 단어들이 독립적으로 나타난다는 가정을 하면 식 (7)의 $P(x_i)$ 는 식 (8)과 같이 추정된다.

$$p(x_i) \approx p_{MLE}(w_i) = 1 - \frac{tf \cdot idf(w_i)}{\sum_j tf \cdot idf(w_j)} \quad (10)$$

이 식에서는 가중치 값이 높은 단어일수록 낮은 $P(x_i)$ 를 가진다.

정보 3. 구성단어 바이그램 확률

이 방법은 용어를 구성하는 단어들이 바로 앞 단어에만 영향을 받고, 코퍼스에서 인접해서 나타날 확률이 낮은 단어 쌍이 포함된 용어의 전문성이 높아진다는 가정을 기반으로 한다. 코퍼스에서 인접해서 나타날 확률이 낮은 단어 쌍은 제한된 용어에만 나타나기 때문에 자신을 포함하는 용어의 특징을 대표할 수 있다. 이 방법은 어떤 용어의 특징을 표현하기 위하여 각각의 구성단어들이 독립적인 역할을 하는 지, 아니면 여러 개의 단어들이 집합적으로 역할을 하는 지를 판단하기 위하여 도입되었다. 이 가정에서 식 (7)의 $P(x_i)$는 식 (9)와 같이 추정된다.

$$p(x_i) \approx \begin{cases} p_{MLE}(w_i) = freq(w_i)/\sum_j freq(w_j) & i=1 \text{ 인 경우} \\ p_{MLE}(w_i \mid w_{i-1}) = freq(w_{i-1}w_i)/\sum_j freq(w_{i-1}w_j) & i>1 \text{ 인 경우} \end{cases} \quad (11)$$

여기에서 $P_{MLE}(w_1)$ 은 t 를 구성하는 단어 중 첫 번째 단어가 나타날 확률을 나타낸다. $P_{MLE}(w_i|w_{i-1})$은 t에서 $i-1$ 번째 위치에 단어 w_{i-1}가 나타났을 때 i 번째 위치에 단어 w_i가 나타날 확률을 나타낸다. $freq(w_1w_2)$는 단어 w_1 과 w_2가 코퍼스에서 인접하여 주어진 순서대로 나타나는 빈도수를 나타낸다.

2.1.2 구성 단어간 수식 관계를 이용한 계산 방법

전문용어는 복합명사로 표현되는 경우가 많기 때문에 전문용어 내부의 수식 구조를 알 수 있으면 상대적으로 정확한 전문성 값을 계산할 수 있다. 2.1.1장에서는 모든 구성 단어를 독립적이라고 가정하였지만, 이 장에서는 구성 단어 사이에 수식어-피수식어 관계가 있다고 가정하고, 이 관계를 이용하여 용어의 전문성을 계산한다. 즉 전문용어에서 기반명사와 수식어를 분리하여 전문성 값을 독립적으로 계산한 뒤, 두 전문성 값의 합을 전체 용어의 전문성 값으로 사용한다. 이 방법으로 계산된 전문성은 기반 명사의 전문성보다 항상 큰 값을 가지는 장점이 있다. 그러나 전문용어 구성 단어들 사이의 정확한 수식구조를 분석하기 어렵다는 단점이 있다. 따라서 본 연구에서는 전문용어 사이의 내포 관계를 이용한 단순화된 수식구조를 이용한다. 용어 X가 다른 용어 Y의 일부로 포함되면 X는 Y에 내포되었다고 정의한다[4]. 예를 들어, 용어 "diabetes mellitus"는 용어 "insulin dependent diabetes mellitus"에 내포된다고 말한다.

두 개의 용어 X와 Y가 동일한 분류를 나타내는 용어이고, Y가 $Mod \cdot X$와 같은 형태로 X를 내포하고 있을 경우, X는 기반 용어이고 Mod는 X의 수식어라고 정의한다. 이 경우 $Spec(Y|D) > Spec(X|D)$ 관계가 성립한다. 위의 예에서 "diabetes mellitus"와 "insulin dependent diabetes mellitus"는 모두 질병 이름이고, "diabetes mellitus"가 "insulin dependent diabetes mellitus"에 내포하기 때문에 "diabetes mellitus"는 기반 용어이고, "insulin dependent"는 수식어이다. 한 개의 용어에 여러 개의 용어가 내포될 경우 길이가 가장 긴 용어를 기반 용어로 선택한다. 예를 들어 세 개의 용어 "neuropathy"(신경병증), "amyloid neuropathy"(아밀로이드 신경병증), "familiar amyloid neuropathy"(가족성 아밀로이드 신경병증)에서 "familiar amyloid neuropathy"는 다른 두 개의 용어 모두를 내포하지만 길이가 더 긴 "amyloid neuropathy"가 기반용어이고, "familiar"가 수식어이다.

수식관계를 이용한 용어의 전문성은 식 (12)와 같이 정의된다.

$$Spec(Y \mid D) = Spec(X \mid D) + \alpha \cdot Spec(Mod \mid D) \quad (12)$$

여기에서 $Spec(X|D)$, $Spec(Mod|D)$ 는 2.1.1 장에서 제안한 3가지 정보 중에서 한 가지를 선택하여 계산한다. 단 두 개의 전문성 값 모두 동일한 정보를 사용하여 계산한다. α 는 0과 1 사이의 값을 가지며, $Spec(Y|D)$가 지나치게 커지는 것을 방지하기 위하여 사용한다.

일반적으로 내포 관계를 이루는 두 용어에서 내포되는 용어는 내포하는 용어의 상위어가 된다. 따라서 이 방법으로 전문성을 계산하면 하위어는 상위어보다 항상 높은 전문성 값을 가지기 때문에 본 연구의 가정과 일치한다.

2.2 문맥정보 기반 계산 방법(방법 2)

상하위어 관계를 가지는 두 용어를 구성하는 단어들이 상이할 경우는 기존의 용어에 수식어를 추가하여 새 용어를 만드는 신조어 생성 특징에 위배된다. 따라서 방법 1만을 이용하여 이 현상을 설명하기 매우 어렵다. 이 장에서는 이 단점을 보완하기 위하여 용어의 문맥정보를 이용하여 전문성을 계산하는 방법을 설명한다. 코퍼스에서 어떤 용어를 중심으로 주위에 나타나는 단어들의 분포를 문맥정보라고 한다. 용어와 공기하는 단어들의 분포, 용어를 논항으로 가지는 술어의 분포, 용어를 수식하는 수식어의 분포 등이 문맥정보로 사용될 수 있다. 일반적으로 일상적인 용어일수록 다른 단어의 수식을 받을 확률이 높고, 전문적인 용어일수록 용어 내부에 많은 정보를 내포하고 있기 때문에 다른 단어의 수식을 받을 가능성이 낮다[7]. 따라서 용어를 수식하는 단어들의 분포를 전문성 계산을 위한 문맥정보로 사용한다. [7]에서는 코퍼스에서 추출한 용어의 최우측 전방 수식어의 분포만을 문맥정보로 이용한 경우 가장 좋은 실험결과를 보였다. 그러나 전문용어일수록 다른 단어의 수식을 받는 경우가 적기 때문에 통계적으로 충분한 문맥정보를 추출하는 작업이 매우 중요하다. 따라서 주어진 전문용어가 나타나는 문장을 의존 구조 파서[3]를 이용하여 분석한 뒤, 그 용어의 수식어를 추출하여 문맥정보로 이용한다. 그림 2에서 "developed", "established" 두 개의 단어가 "diabetes mellitus"를 수식한다. 따라서 "diabetes mellitus"의 수식어 집합에서 "develop"와 "establish"의 빈도수를 1씩 증가시킨다.

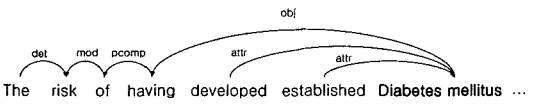

그림 2 문장 "The risk of having developed established diabetes mellitus …"의 의존 구조

용어 t를 수식하는 단어들의 분포를 이용하여 계산된 엔트로피를 식 (13)과 같이 계산한다.

$$H_{mod}(t) = -\sum_{i} p(mod_i,t)\log p(mod_i,t) \quad (13)$$

여기에서 $p(mod_i,t)$는 mod_i가 t를 수식할 확률을 나타내고, 식 (14)와 같이 추정된다.

$$p_{MLE}(mod_i,t) = freq(mod_i,t)/\sum_{j} freq(mod_j,t) \quad (14)$$

3) 본 연구에서는 영어 구문 분석을 위하여 Conexor functional dependency parser (http://www.conexor.fi)를 사용하였다. 이 파서에서 사용하고 있는 많은 구문 관계 중에서 "mod" (postmodifier), "attr" (attributive nominal) 관계를 사용하여 각 용어의 수식어를 추출하였다.

여기에서 $freq(mod_i,t)$는 전체 코퍼스에서 mod_i가 t를 수식하는 회수를 나타낸다. 식 (13)에서 계산된 엔트로피는 모든 (mod_i,t) 쌍의 평균 정보량을 나타낸다. 전문적인 용어일수록 수식어의 분포가 단순하기 때문에 낮은 엔트로피를 가지고, 일상적인 용어일수록 수식어가 복잡하기 때문에 높은 엔트로피를 가진다. 따라서 전문적인 용어일수록 높은 정보량을 가지도록 하기 위하여, 식 (15)와 같이 최고 엔트로피에서 그 용어의 엔트로피 값을 뺀 값을 그 용어의 정보량으로 정의하고, 식 (2)의 $I(x)$에 대응시킨다.

$$Spec(t \mid D) = I(x) \approx \max_{1\le i\le n} H_{mod}(t_i) - H_{mod}(t_k) \quad (15)$$

이 계산 방법은 용어 자체 또는 그 용어의 수식어가 코퍼스에서 나타나지 않는 경우에 전문성을 계산할 수 없는 단점이 있다.

2.3 구성/문맥정보 기반 계산 방법(방법 3)

용어를 구성하는 단어들이 상위어를 구성하는 단어들과 전혀 다를 경우, 내부 구성정보를 이용하여 얻어진 전문성 값을 상위어의 전문성 값과 객관적으로 비교하기 어렵다. 한편 실험 코퍼스에서 충분한 문맥정보를 구할 수가 없는 용어들은 문맥정보를 이용하여 정확한 전문성 값을 계산할 수 없다. 두 방법의 단점을 극복하기 위하여 식 (2)의 $I(x)$를 식 (16)과 같이 두 방법을 혼합하여 계산할 수 있다.

$$Spec(t \mid D) = I(x) \approx \frac{1}{\gamma(\frac{1}{I_{Comp}(x)}) + (1-\gamma)(\frac{1}{I_{Ctx}(x)})} \quad (16)$$

여기에서 $I_{Comp}(x)$와 $I_{Ctx}(x)$는 각각 t의 구성정보를 이용한 정보량과 문맥정보를 이용한 정보량을 0과 1사이의 값으로 정규화한 값이다. $\gamma(0 \le \gamma \le 1)$는 두 값의 가중치를 나타내고, $\gamma = 0.5$인 경우는 두 값의 조화평균이다. 따라서 두 값이 공통적으로 높은 값을 가질 경우에 높은 전문성 값을 가진다.

3. 실험 및 평가

3.1 시스템 구성

전체 시스템의 구성은 그림 3과 같이 세 개의 모듈로 구성되고 각각의 기능은 다음과 같다.

- **전문용어 관리자:** 통계정보 추출 또는 전문성 계산의 대상이 되는 전문용어 리스트를 관리한다. 용어의 전문성 값이 계산된 후 평가기준에 맞게 평가한다.
- **통계정보 관리자:** 코퍼스에서 전문성 계산을 위한 통계정보를 추출하고, 전문성 관리자에게 통계 정보를 서비스한다.
- **전문성 관리자:** 용어의 전문성 값을 논문에서 제안한 다양한 방법으로 계산한다.

표 2 실험 대상 용어 하위 트리의 요약 정보

항목	값	예제
용어 수	436	
트리의 최대 깊이	7	"Tay-Sachs disease" (C18.452.100.100.435.825.300.300.840, 테이삭스 병)
용어의 평균 구성 단어	2.22	
최대 구성단어 용어의 단어 수	5	"carbamoyl-phosphate synthesis I deficiency disease" (C18.452.100.100.162, 카르바밀 인산 신테시스 I 결핍증)
상위어를 내포하는 용어 수	62	

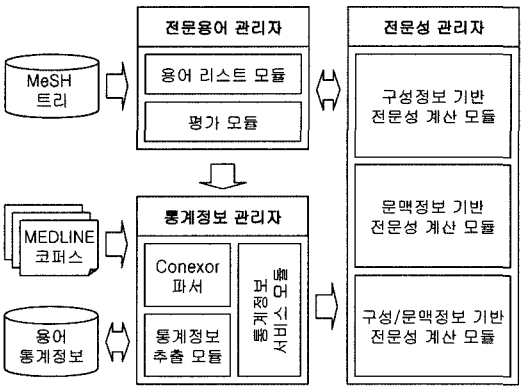

그림 3 전체 시스템 구성도

3.2 실험방법 및 평가기준

제안 방법의 유효성을 측정하기 위하여 기존의 용어 계층구조에서 상하위어 관계를 가지는 용어 사이의 전문성 값을 비교하였다. MeSH 트리 중에서 "metabolic diseases"(C18.452, 대사성 질환)를 루트 노드로 가지는 하위 트리에 포함된 용어 436개를 대상으로 전문성 계산 방법을 실험하였다. 이 하위 트리의 특징은 표 2에 정리되어 있다. 용어 436개를 검색어로 사용하여 MEDLINE[4] 데이타베이스에서 170,000개의 논문 요약문(약 20,000,000 단어)을 추출하였다. 추출된 요약문에서 제목과 요약 부분을 Conexor 파서로 분석한 뒤 다음과 같은 통계 정보를 추출하였다.

- 전문용어의 빈도수, tf·idf, 전문용어가 포함된 문서의 빈도수
- 전문용어의 수식어 분포
- 전문용어 구성단어의 빈도수, tf·idf, 구성단어가 포함된 문서의 빈도수
- 전문용어 구성단어의 바이그램 정보

적용율(coverage)과 정확률(precision)을 이용하여 제안한 방법을 평가한다. 적용율은 식 (17)과 같이 주어진 방법으로 전문성 값을 계산할 수 있는 용어의 비율로 정의된다. 방법 2에서는 코퍼스에서 해당 용어가 나타나지 않는 경우 전문성 값을 계산할 수 없기 때문에 적용율이 낮아진다. 이와 반대로, 방법 1은 전체 구성단어 중 일부 단어만 코퍼스에서 나타나도 전문성 값을 계산할 수 있기 때문에 적용율이 높다.

$$\text{적용율} = \frac{\text{주어진 방법으로 전문성을 계산할 수 있는 전문용어의 수}}{\text{전체 전문용어의 수}} \quad (17)$$

정확률은 식 (18)와 같이 전문성 값을 비교할 수 있는 모든 부모-자식 관계 중에서 올바른 전문성 값을 가지는 관계의 비율로 정의된다.

$$\text{정확률} = \frac{\text{올바른 전문성 값을 가지는 } R(parent,child) \text{ 개수}}{\text{트리에서 전체 } R(parent,child) \text{ 개수}} \quad (18)$$

여기에서 $R(parent, child)$는 부모-자식 관계를 가지는 용어 쌍 중에서 두 용어 모두 전문성 값을 가지는 관계를 나타낸다. 이 용어 쌍에서 상위어의 전문성이 하위어의 전문성보다 낮을 경우 올바른 전문성 값을 가진다고 말한다. 예를 들어, 그림 4에서 두 개의 용어 "metabolic diseases"와 "diabetes mellitus" 모두 전문성 값을 가지고 있는 경우 R("metabolic diseases", "diabetes mellitus") 관계가 성립하고, $Spec$("metabolic diseases"|disease)< $Spec$("diabetes mellitus"|disease)인 경우 이 관계가 올바른 전문성 값을 가진다고 판단한다. 용어의 상하위 관계를 두 가지 유형으로 나누어 정확률을 계산하였다. 유형 I은 상위어가 하위어에 내포된 경우이고, 유형 II는 그렇지 않은 경우이다. 전체 관계 중 유형 I은 62개이고, 유형 II는 374개이다. 유형 I은 2.1.2장의 용어 내부 수식 구조를 이용하는 방법을 사용하면 두 용어의 전문성 값은 항상 올바른 관계를 가진다.

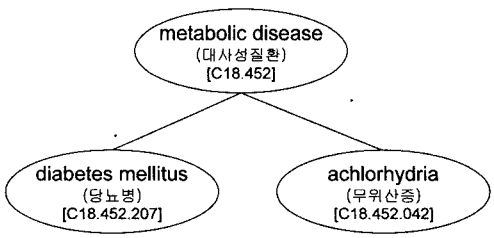

그림 4 MeSH 트리의 일부분

4) MEDLINE은 미국 National Library of Medicine(NLM)에서 관리하는 의료분야 서적 데이타베이스이다.(http://www.nlm.nih.gov/pubs/factsheets/medline.html)

먼저 정확률의 상한선(upper bound)를 알아보기 위하여 종합병원 내과 전문의와 전공의 10명에게 436개의 용어를 부모 노드의 용어와 함께 제시하고 더 전문적인 용어를 선택하는 실험을 실시하였다. "metabolic diseases"는 내과와 가장 관련이 있는 분야이다. 테스트 결과에서 유형 I, 유형 II에 대해서 각각 평균 정확률 96.6%와 86.4%를 보였고, 전체 관계에 대해서는 평균 정확률 87.4%를 보였다. 이 결과들이 이 논문에서 제안한 방법들로 얻을 수 있는 정확률의 상한선이라고 판단된다. 유형 I은 간단한 규칙으로 판단이 가능하지만 정확률이 100%가 되지 않은 것은 테스트에 참가한 사람의 실수라고 추정된다.

3.3 실험결과 및 분석

표 3과 같이 방법 1, 방법 2, 방법 3으로 용어의 전문성 값을 각각 계산한 뒤 평가하였다. 방법 1에서는 빈도수, 가중치, 바이그램 정보를 이용한 경우와 각각의 경우에 수식구조 정보를 이용한 경우를 나누어서 실험하였다. 방법 3은 방법 1과 방법 2에서 가장 좋은 결과를 보인 두 가지 방법을 혼합하였다. 또한 전문용어를 구성단어 단위로 나누지 않고 용어 자체의 빈도수와 가중치(tf·idf)를 이용하여 전문성을 계산하는 방법을 추가로 실험하였다. 이 추가 실험의 목적은 전문성 계산에서 구성단어 단위의 정보를 이용하는 경우와, 용어 전체 단위의 정보를 이용하는 경우를 비교하는 것이었다.

실험 결과 방법 1에서는 구성단어의 가중치와 용어의 수식구조 정보를 이용한 경우 정확률 78.9%, 적용율 100%로 가장 좋은 성능을 보였다. 방법 1에서 구성단어의 빈도수와 바이그램 정보를 이용하는 경우에도 수식구조 정보를 같이 이용하면 모두 좋은 성능을 보였다. 그림 5는 용어의 내부 수식구조를 이용하여 전문성을 계산할 때 수식어 가중치 의 변화에 따른 정확률의 변화를 보여준다. 이 그래프는 어떤 수식어가 기존 용어와 결합할 때 그 수식어의 정보량 중 일부분만큼만 전체 전문성 증가에 반영된다는 사실을 설명한다. 즉 수식어의 특징 집합과 기존 용어의 특징 집합 사이에 교집합이 있을 경우, 새 용어의 전문성은 두 단어의 전문성의 합에서 교집합만큼 줄어든다는 사실을 간접적으로 보여주고 있다.

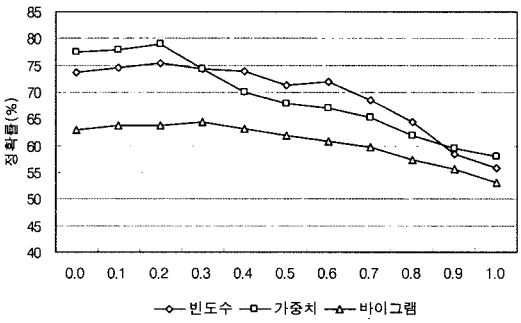

그림 5 방법 1에서 α값의 변화에 따른 정확률의 변화

표 3 용어의 전문성 실험 결과 (%)

구분		정확률			적용율
		유형 I	유형 II	전체	
전문가 평가(평균)		96.6 (41.55/43)	86.4 (339.45/393)	87.4 (381/436)	
용어 빈도수		57.9 (22/38)	61.0 (130/213)	60.6 (152/251)	89.5 (390/436)
용어 가중치		52.6 (20/38)	59.2 (126/213)	58.2 (146/251)	89.5 (390/436)
구성정보 (방법1)	빈도수	0.37 (16/43)	72.5 (285/393)	69.0 (301/436)	100.0 (436/436)
	빈도수+수식구조 (α= 0.2)	100.0 (43/43)	72.8 (286/393)	75.5 (329/436)	100.0 (436/436)
	가중치	44.2 (19/43)	75.3 (296/393)	72.2 (315/436)	100.0 (436/436)
	가중치+수식구조 (α= 0.2)	100.0 (43/43)	76.6 (301/393)	78.9 (344/436)	100.0 (436/436)
	바이그램	37.2 (16/43)	59.5 (234/393)	57.3 (250/436)	100.0 (436/436)
	바이그램+수식구조 (α= 0.3)	100.0 (43/43)	60.6 (238/393)	64.4 (281/436)	100.0 (436/436)
문맥정보(mod cnt>1) (방법 2)		90.0 (18/20)	66.4 (75/113)	70.0 (93/133)	70.2 (306/436)
구성정보+문맥정보 (방법 3) (가중치 + 수식구조, γ=0.8)		95.0 (19/20)	79.6 (90/113)	82.0 (109/133)	70.2 (306/436)

한편, 용어 전체를 이용한 방법 또는 구성단어의 바이그램 정보를 이용한 방법보다 구성단어를 독립적으로 이용한 방법이 더 좋은 성능을 보였다. 이 결과는 용어를 구성하는 각각의 단어들이 용어의 전체 특징 집합을 분할하여 가지고 있는 경향이 강하다는 사실을 설명한다. 즉 전문적인 개념은 기존의 개념에 새로운 특징을 추가하여 생기는 경우가 많고, 이 개념을 전문용어로 표현할 때 기존의 용어에 추가되는 특징을 나타내는 단어를 수식어로 사용하는 경우가 많다는 이 논문의 가정을 뒷받침한다.

방법 2에서는 수식어의 빈도수가 2 이상인 경우에 정확률 70.0%, 적용율 70.2%로 가장 좋은 성능을 보였다. 빈도수 기준을 높이면 충분한 수식어를 얻지 못하는 단점이 있고, 그 반대의 경우는 각 용어들이 비슷한 수식어들을 가지게 되어 변별력이 낮아지는 단점이 있다. 이 방법은 방법 1의 용어 구성단어의 빈도수와 가중치를 이용하는 방법보다 낮은 성능을 보였다. 그 이유는 전문용어는 그 자체로 충분한 정보를 가지고 있고, 일반 용어와는 달리 다른 단어의 수식을 받는 경우가 적기 때문에 코퍼스에서 충분한 문맥정보를 얻을 수 없기 때문이라고 추측된다.

방법 1과 방법 2에서 가장 좋은 성능을 나타낸 두 가지 방법을 혼합한 실험(방법 3)에서는 식 (16)에서 γ =0.8 인 경우에 정확률 82.0%, 적용율 70.2%의 성능을 보였다. 이 방법은 전체 실험 중 가장 높은 정확률을 보였지만, 방법 2에서 전문성 값을 계산하지 못하는 용어들은 이 방법에서 제외하였기 때문에 낮은 적용율을 보였다. γ=0.8 인 경우 가장 높은 정확률을 나타낸 것은 전문용어는 용어의 내부 구성정보가 문맥정보보다 더 중요하다는 사실을 설명한다. 그림 6은 방법 3에서 γ값의 변화에 따른 정확률의 변화를 보여준다. γ=1.0은 용어의 구성정보만 사용한 경우이고, γ=0은 용어의 문맥정보만 사용한 경우이다. 정확률 82.0%는 상한선 87.4%에 상당히 근접한 결과로 판단된다.

같은 용어 쌍에 대하여 방법 1, 방법 2의 결과와 방법 3의 결과를 비교하면 표 4와 같다. 방법 3은 표 3에서와 같이 방법 1과 방법 2의 결과 중 각각 가장 좋은 두 결과를 선택하여 결합하였다. 방법 1, 방법 2에서 올바른 전문성 관계를 가진 용어 쌍은 방법 3에서도 모두 올바른 전문성 관계를 가졌다. 방법 1에서만 올바른 전문성 관계를 가지는 용어 쌍은 모두 방법 3에서 올바른 전문성 관계를 가졌고, 방법 2에서만 올바른 전문성 관계를 가진 용어 쌍에서 6.7% 만이 방법 3에서 올바른 전문성 관계를 가졌다. 이 두 가지 결과는 전문 분야 용어의 전문성 계산에서는 용어의 구성정보가 더 중요하다는 사실을 다시 한 번 더 설명한다. 두 방법에서 모두 올바르지 않은 전문성 관계를 가지는 용어 쌍은 방법 3에서 모두 올바르지 않는 전문성 관계를 가졌다. 결과적으로 방법 1에서 올바른 전문성 관계를 가지는 용어 쌍은 방법 3에서도 모두 올바른 전문성 관계를 가졌고, 추가적으로 방법 2에서만 올바른 전문성 관계를 가지는 용어 쌍 중 일부분이 방법 3에서 올바른 전문성 관계를 가진다.

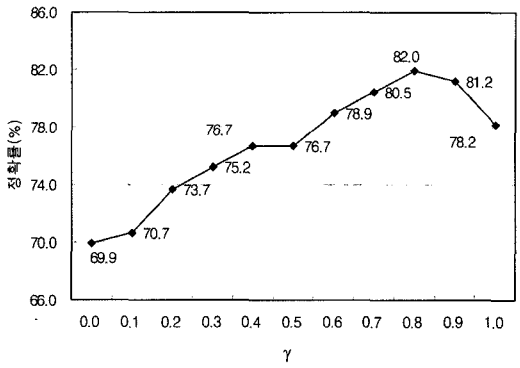

그림 6 방법 3에서 γ값 변화에 따른 정확률 변화

표 5는 방법 1의 오류를 방법 2의 결과를 이용하여 보정한 예를 보여준다. 방법 1에서는 상위어의 전문성이 더 높지만 방법 2와 방법 3에서는 하위어의 전문성이 더 높다. 이 예에서는 방법 1의 결과에서 두 용어의 전문성의 차이가 충분히 작기 때문에 방법 2의 결과를 이용하여 보정할 수 있었다.

전체적인 결과 분석에서 용어의 내부 구성정보가 전문성 계산에서 중요한 역할을 하고 있음을 알 수 있었다. 상대적으로 문맥정보를 이용한 방법은 자체적으로도 낮은 정확률과 적용율을 보였고, 혼합한 방법에서도 보

표 4 방법 1, 2, 3의 전문성 계산 결과 비교 (용어 쌍의 개수)

방법 1 (가중치+수식구조, α=0.2)	**방법 2** (mod cnt>1)	**방법 3** (γ=0.8)		**계**
		Correct	**Incorrect**	
Correct	Correct	71	0	71
Correct	Incorrect	36	0	36
Incorrect	Correct	2	14	16
Incorrect	Incorrect	0	10	10
계		109	24	133

표 5 방법 1과 방법 2의 결과가 결합하여 방법 3에서 올바른 전문성 관계를 계산한 예

		전문성 (방법 1)	전문성 (방법 2)	전문성 (방법 3)
상위어	calcinosis (석회침착증)	43.100	2.269	0.105
하위어	Crest syndrome (CREST 증후군)	42.619	3.395	0.106
평가		Incorrect	Correct	Correct

조적인 역할만 수행하였다. 따라서 문맥정보의 정확률을 높이는 방법에 대한 추가적인 연구가 필요하다.

시스템의 성능을 향상시키기 위하여 실험과정에서 향후 개선하여야 할 부분이 몇 가지 있었다. 첫째, MeSH 트리의 상위 또는 중간 노드가 질병의 분류를 나타내는 경우를 구분할 필요가 있다. 예를 들어 "acid-base imbalance"는 흔히 사용하는 질병의 이름이 아니고 "산염기 평형 이상(酸鹽基 平衡 異常)"이라는 질병의 분류 이름을 나타내기 때문에 코퍼스에서 하위어보다 상대적으로 출현빈도수가 낮다.[5] 따라서 하위어보다 높은 전문성 값을 가지는 오류가 발생한다. 질병의 분류와 질병의 이름의 구분은 전문가의 판단에 의존하여야 한다. 둘째, 전문용어의 이형태를 고려하지 않아서 정확한 통계정보를 추출하지 못한 경우가 많았다. 예를 들어 "diabetes mellitus"와 "diabetes"는 같은 의미를 가지지만 실험에서 서로 다른 용어로 인식하는 문제가 있었다. 이 문제 또한 전문가의 판단에 따르거나 전문 용어 사전을 참조하여야 한다. 향후 용어의 이형태를 함께 이용하여 통계정보를 추출할 필요가 있다. 셋째, 전문용어 조어법 분석이 가능하면 성능을 높일 수 있다. 방법 2에 의한 오류도 코퍼스에서 추출한 문맥정보가 상하위어 관계에 대한 가정과 일치하지 않아서 발생한다. 예를 들면 "nephrocalcinosis" (C18.452.174.130.560, 신석회침착증)가 "calcinosis" (C18.452.174.130, 석회침착증)의 하위어인 경우, 조어법 분석을 통하여 "nephrocalcinosis"가 수식어 "nephro"와 기반단어 "calcinosis"로 구성된다는 사실을 파악하면, 2.1.2에서 설명한 수식구조를 이용한 전문성 계산방법을 적용할 수 있다. 그러나 전문용어의 조어법은 분야마다 특징이 다르기 때문에 분야별로 별도의 조어법 분석이 필요하다.

3.4. 기존 연구와의 차이점

용어의 전문성 측정 방법과 관련된 연구는 정보 검색 분야에서 시스템의 정확률을 높이기 위하여 분야의 특징을 대표하는 색인어 추출과 관련된 연구에서 주로 연구되었다. Aizawa[8]와 Wong[9]은 용어의 전문성을 정보 이론에 기반한 방법으로 측정하였다. 이 연구들은 정보검색 시스템에서 많이 사용되는 용어의 가중치 계산 방법을 수학적으로 해석하려고 시도하였다. 문서 또는 전체 코퍼스에서 용어의 빈도수를 이용하여 용어의 가중치를 계산하였다. 전문분야 용어를 가정하지 않았기 때문에 용어의 구성정보와 문맥정보를 이용하지 않았다는 점에서 본 연구와의 차이점이 있다.

용어간 계층관계 설정을 위한 연구에서도 용어의 전문성이 논의되었다. Caraballo[7]는 본 연구의 방법 2와 유사하게 전문적인 정보를 많이 포함한 명사일수록 코퍼스에서 나타날 때 다른 수식어의 수식을 받는 경우가 적고, 반대로 일상적인 명사일수록 수식어의 수식을 받는 경우가 많다는 가정을 기반으로 하였다. 따라서 수식어의 엔트로피가 높을수록 다양한 수식어를 가지기 때문에 일반적인 명사이고, 엔트로피가 낮을수록 전문적인 명사라고 판단하였다. 이 연구는 일반 명사들의 전문성을 측정하였기 때문에 전문용어와 달리 비교적 풍부한 수식어를 코퍼스에서 수집할 수 있었다. 따라서 구문 분석과정을 거치지 않고, 각 명사들의 가장 오른쪽 전방 수식어(rightmost prenominal modifier)만 추출하여 엔트로피를 계산하였다. 또한 이 연구는 대상 명사가 대부분 단일 단어로 구성되어 있기 때문에 문맥정보만을 이용하여 전문성을 계산하였다.

4. 결론 및 향후 연구

본 논문에서는 용어가 전문적인 정보를 많이 포함할수록 전문성이 높다고 가정하고, 용어의 구성정보와 문맥정보를 이용하여 용어의 전문성의 정도를 정량적으로 계산하는 방법을 제안하였다. 제안한 방법은 적용 분야의 특징적인 정보를 이용하지 않기 때문에 다른 분야에 쉽게 적용할 수 있는 장점이 있다.

실험에서 용어의 내부 구성정보를 이용하는 방법, 문맥정보를 이용하는 방법, 그리고 두 가지 방법을 조합한 방법으로 용어의 전문성을 계산하였고, 의학용어 분류체계인 MeSH 트리에 적용하여 평가하였다. 실험결과 용어의 구성정보와 문맥정보를 함께 사용한 경우 가장 높은 정확률(82.0%)를 보였다.

향후 제안한 방법이 용어의 이형태를 고려할 수 있도록 하는 방법과, 전문용어 조어법 분석을 통하여 단어 내부에 포함된 정보도 추출할 수 있는 방법에 대한 연

5) "acid-base imbalance"의 빈도수는 85회이고, 하위어인 "acidosis"와 "alkalosis"는 각각 10,192회, 2,394회 나타났다.

구도 필요하다. 또한 용어 구성단위의 의미정보를 이용하는 정교한 모델을 개발할 예정이다. 마지막으로 제안한 방법을 용어간 자동 계층 관계 설정에 적용할 계획이다.

참 고 문 헌

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

[2] ISO 704, "Terminology work-Principle and methods," ISO 704 Second Edition, 2000.

[3] T.M. Cover & J.A. Tomas, Elements of Information Theory, New York: John Wiley and Sons Inc., 1991.

[4] Katerina Frantzi, Sophia Anahiadou, Hideki Mima, "Automatic recognition of multi-word terms: the C-value/NC-value method," Journal of Digital Libraries, Vol. 3, Num 2, pp. 115-130, 2000.

[5] 오종훈, 이경순, 최기선, "분야간 유사도와 통계기법을 이용한 전문용어의 자동 추출", 정보과학회논문지: 소프트웨어 및 응용 제29권 제1호, pp. 258-269, 2002.

[6] Christopher D. Manning and Hinrich Schutze, "Foundations of Statistical Natural Language Processing," The MIT Press, 1999, p. 543.

[7] Sharon A. Caraballo and Eugene Charniak, "Determining the Specificity of Nouns from Text," in the Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 63-70, 1999.

[8] A. Aizawa, An information-theoretic perspective of tf-idf measures, Journal of Information Processing and management Vol. 39, 2003.

[9] S.K.M Wong and Y.Y. Yao, An Information-Theoretic Measure of Term Specificity, Journal of the American Society for Information Science, Vol. 43, Num. 1, 1992.

류 법 모

1995년 경북대학교 컴퓨터공학과 졸업(학사). 1997년 포항공과대학교 대학원 컴퓨터공학과 졸업(석사). 2000년~현재 한국과학기술원 전산학과 박사과정. 1997년~1999년 한국전자통신연구원(ETRI) 자연어처리연구실 연구원. 1999년~2004년 (주)케이포엠 기술연구소 연구원. 관심분야는 자연언어처리, 온톨로지 학습

배 선 미

1992년 이화여자대학교 불어불문학과 졸업(학사). 1994년 이화여자대학교 대학원 불어불문학과 졸업(석사). 1997년 이화여자대학교 대학원 불어불문학과 수료(박사). 1998년 미국 산호세 주립대학 전산학과 수학. 2002년 프랑스 마른라발레 대학교 대학원 전산학과 전산언어학 전공(박사). 2003년 프랑스 마른라발레 대학교 Gaspard Monge 연구소 박사후 연구원. 2003년~2005년 한국과학기술원 정보전자연구소 BK 21 박사후 연구원. 2006년~현재 한국과학기술원 인문사회과학부 연구교수. 관심분야는 전산형태론, 전산통사론, 시소러스, 전문용어

최 기 선

1978년 서울대학교 자연과학대학 수학과 졸업(학사). 1980년 한국과학기술원 전산학과 졸업(석사). 1986년 한국과학기술원 전산학과 졸업(박사). 1987년~1988년 일본 NEC C&C 정보연구소 연구원. 1988년~현재 한국과학기술원 전산학과 교수. 1997년~1998년 미국 스탠포드대학 CSLI 객원교수. 2002년~2003년 일본 NHK 방송기술연구소 초빙연구원. 2006년~현재 한국인지과학회 회장. 2003년~현재 국가지정 언어자원특수소재은행장 http://bola.kaist.ac.kr. 2002년~현재 ISO/TC37/SC4 언어자원관리표준 Secretary. 2002년~현재 TermNet 회장. 2000년~현재 ACM TALIP, IJCPOL 편집위원, IAMT council member. 1998년~현재 전문용어언어공학연구센터 http://korterm.or.kr/. 관심분야는 온톨로지, 텍스트마이닝, 인공두뇌, 지식획득, 창의계산론, 언어공학, 시맨틱웹