구성정보와 문맥정보를 이용한 용어의 전문성 측정 방법*

류법모° 배선미 최기선
한국과학기술원 전산학과
전문용어언어공학연구센터, 언어자원은행
{pmryu°,sbae,kschoi}@world.kaist.ac.kr

Determining the Specificity of Terms based on Compositional and Contextual Information

Pum-Mo Ryu° Sun-Mee Bae Key-Sun Choi
Dept. of CS, KAIST/KORTERM/BOLA

요 약

어떤 용어가 전문적인 개념을 많이 내포하고 있을 때 전문성(specificity)이 높다고 말한다. 본 논문에서는 정보이론에 기반한 방법으로 전문용어가 내포하는 전문성을 정량적으로 계산하는 방법을 제안한다. 제안한 방법은 전문용어의 구성정보를 이용하는 방법과 문맥정보를 이용하는 방법으로 나눈다. 구성정보를 이용하는 방법에서는 전문용어를 구성하는 단어의 빈도수, tf·idf 값, 내부 수식구조 등을 이용하고, 문맥정보를 이용하는 방법에서는 전문용어를 수식하는 단어들의 분포를 이용한다. 제안한 방법은 분야에 독립적으로 적용될 수 있고, 전문용어 생성 절차에 대한 특징을 잘 표현할 수 있는 장점이 있다. MeSH 트리에 포함된 질병명을 대상으로 실험한 결과 82.0%의 정확률을 보였다.

1 서론

새로운 전문분야가 지속적으로 만들어지고, 기존의 전문분야도 빠른 속도로 특성이 변화하면서 자동으로 전문분야 지식을 관리는 방법에 대한 연구가 활발히 진행되고 있다. 전문용어는 전문분야의 개념이 언어적으로 표현된 형태이므로, 전문분야 지식을 대표하는 단위로 사용된다[1].

어떤 용어의 전문성은 그 용어가 포함하고 있는 전문적인 정보의 정도를 정량적으로 표현한 것이다. 용어 X의 전문성 $Spec(X)$를 식 (1)과 같이 실수로 표현된다.

$$Spec(X) \in R^{+} \qquad (1)$$

용어의 전문성은 코퍼스에서 전문용어를 추출하거나 용어 사이의 계층관계를 설정할 때, 그리고 자동으로 구축된 용어 계층구조를 평가할 때 중요한 정보로 이용될 수 있다. 기존의 용어 사이의 상하위 관계를 설정하는 연구[2,3,4,5]와 용어의 전문성을 측정하는 연구[6]에서는 용어의 문맥 정보를 가장 중요한 정보로 이용하였다. 상대적으로 용어를 구성하는 단어의 특징을 분석하려는 노력은 거의 이루어지지 않았다. 전문용어는 일반용어보다 길이가 긴 복합명사로 표현되는 경우가 많고, 용어를 구성하는 단어들은 그 용어의 특징을 분할하여 내포하고 있기 때문에, 구성 단어들의 특징을 분석하면 전체 용어의 특징 중 많은 부분을 파악할 수 있다. 본 연구에서는 용어의 내부 구성정보와 문맥정보를 이용하여 용어의 전문성을 정량적으로 표현하는 방법을 제안한다. 제안한 방법은 분야에 종속적인 정보를 이용하지 않기 때문에 다른 분야 전문용어에 쉽게 적용할 수 있다.

이 논문의 나머지 부분은 다음과 같이 구성된다. 2장에서는 용어의 전문성 계산방법을 설명하고, 3장에서는 제안한 방법을 평가하고, 4장에서는 결론 및 향후 연구가 소개된다.

2 용어의 전문성 계산 방법

전문용어는 특정 전문분야에서 통용되는 개념을 어휘로 표현한 것이다. 전문분야 개념은 자신을 다른 개념들과 구분시킬 수 있는 고유한 특징들의 집합으로 표현된다. 또한 기존의 개념을 표현하는 특징 집합에 새로운 특징을 추가하여 더 전문적인 개념을 만들 수 있다. 일반적으로 기존의 개념 X와 X에 새로운 특징을 추가하여 생긴 개념 Y사이에는 상하위 관계가 성립된다. 즉 X는 Y의 상위 개념이고, X의 특징 집합은 Y의 특징 집합의 부분집합이다. 개념이 용어로 표현될 경우, 대응하는 용어에 상위 용어의 단어들이 포함될 수도 있고 전혀 다른 단어로만 표현될 수도 있다. 따라서 두 가지 경우를 모두 반영하는 전문성 계산 방법이 필요하다.

정보이론에서는 출현 확률이 낮은 메시지가 실제로 나타난 경우 "놀라움"의 정도는 커지고, 그 메시지를 표현하기 위한 비트수는 다른 출력에 비해 길어진다. 즉 그 메시지의 정보량이 높아진다[7]. 코퍼스에 나타난 용어들을 어떤 채널의 출력이라고 가정하면 코퍼스에서 추출한 각종 통계정보를 이용하여 용어의 정보량을 계산할 수 있고, 이 정보량을 용어의 전문성으로 이용할 수 있다. 구체적인 설명을 위하여 전문용어 집합 T를 식 (2)와 같이 정의한다.

$$T = \{t_k \mid 1 \le k \le n\} \qquad (2)$$

여기에서 t_k는 한 개의 전문용어를 나타내고, n은 전체 전문용어의 개수를 나타낸다. 다음 단계에서 전문용어 t_k가 코퍼스에서 나타나는 사건 x_k를 나타내는 랜덤변수 X를 식 (3)과 같이 정의한다.

$$X = \{x_k \mid 1 \le k \le n\}$$
$$p(x_k) = \mathrm{Prob}(X = x_k) \qquad (3)$$

여기에서 $p(x_k)$는 사건 x_k의 확률을 나타낸다. 사건 x_k의 정보량 $I(x_k)$는 식 (4)와 같이 정의되고, 용어 t_k의 전문성 값으로 사용될 수 있다. $p(x_k)$가 낮은 용어일수록 전문성이 높아진다.

$$Spec(t_k) \approx I(x_k) = -\log p(x_k) \qquad (4)$$

따라서 코퍼스에서 추출한 통계 정보를 이용하여 $p(x_k)$를 추정할 수 있으면 t_k의 전문성을 계산할 수 있다.

2.1 구성 정보를 이용한 계산

이 장에서는 기존의 용어에 수식어를 부가하여 생성되는 새로운 전문용어의 특징을 반영하는 전문성 계산 방법을 설명한다. 이 방법에서는 어떤 용어를 구성하는 각각의 단어에 그 용어의 특징들이 분할되어 저장되어 있다는 가정을 하고, 각 구성

* 이 논문은 과학기술부, 과학재단의 지원에 의하여 이루어짐.

단어의 특징을 정량화하여 전체 용어의 전문성을 계산한다.

2.1.1 구성 단어의 특징을 이용한 계산

구성단어를 이용한 용어의 전문성을 계산을 위하여 용어 t_k는 식 (5)와 같이 여러 개의 단어로 구성되어 있다고 가정한다.

$$t_k = t_{k,1}t_{k,2}..t_{k,m} \quad (5)$$

여기에서 $t_{k,i}$ ($1 \le i \le m$)는 t_k를 구성하는 단위 단어를 나타낸다. 각 단어들이 서로 독립적이라고 가정하면 식 (4)의 $I(x_k)$는 식 (6)과 같이 각 구성 단어들의 정보량의 평균값으로 정의된다.

$$I(x_k) = -\sum_{i=1}^{m}[p(x_{k,i})\log p(x_{k,i})] \quad (6)$$

여기에서 $p(x_{k,i})$는 단어 $t_{k,i}$가 코퍼스에서 나타나는 사건의 확률을 나타낸다. $p(x_{k,i})$를 추정하기 위한 2가지 방법을 설명한다.

방법 1. 구성 단어의 빈도수 이용

전문용어 자동 인식과 관련된 기존의 연구에서는 용어의 빈도수가 높을수록 전문용어일 가능성이 높다고 가정하였다[8,9]. 그러나 이 방법에서는 코퍼스에서 출현 확률이 낮은 단어를 포함하는 용어는 전문적일 가능성이 높다는 가정에 기반한다. 출현확률이 낮은 단어들은 적은 수의 전문용어에만 포함되기 때문에, 자신을 포함하는 전문용어의 특징을 차별화시킬 수 있는 능력이 높기 때문이다. 이 가정에서 식 (6)의 $P(x_{k,i})$는 식 (7)과 같이 추정된다.

$$p(x_{k,i}) \approx p_{MLE}(t_{k,i}) = freq(t_{k,i}) / \sum_j freq(w_j) \quad (7)$$

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

방법 2. 구성 단어의 tf·idf 이용

정보검색에서 단어의 가중치를 결정할 때 단어 빈도수(tf)에 문서 빈도수의 역수(idf)를 곱한 값 tf·idf를 많이 이용한다. 빈도수가 높으면서 제한된 문서에 집중적으로 나타나는 단어는 높은 tf·idf를 가진다. tf·idf가 높은 단어는 특정 문서를 다른 문서와 차별화시키는 대표적인 단어의 역할을 하기 때문에 전문적인 정보를 많이 포함하고 있다. 따라서 용어 t_k 에 tf·idf가 높은 단어들이 많이 포함된 경우, t_k 의 전문성이 높다고 가정한다. 용어를 구성하는 모든 단위 단어들이 독립적으로 나타난다는 가정을 하면 식 (6)의 $P(x_{k,i})$는 식 (8)과 같이 추정된다.

$$p(x_{k,i}) \approx p_{MLE}(t_{k,i}) = 1 - \frac{TF.IDF(t_{k,i})}{\sum_j TF.IDF(w_j)} \quad (8)$$

이 식에서는 tf·idf가 높은 단어일수록 낮은 확률값을 가진다.

2.1.2 수식 구조를 이용하는 계산

전문용어 내부의 구조를 분석하여 기반명사와 수식어를 분리할 수 있으면, 기반명사의 전문성을 먼저 계산한 뒤 수식어의 전문성을 이용하여 전체 용어의 전문성을 증가시킬 수 있다. 이 방법으로 계산된 전문성은 기반 명사의 전문성보다 항상 큰 값을 가지는 장점이 있다. 그러나 전문용어는 일종의 복합명사이기 때문에 내부 단어들 사이의 정확한 수식구조를 분석하기가 매우 어렵다. 이 방법에서는 전문용어 사이의 내포 관계를 이용한 단순화된 수식구조를 이용한다. 용어 X가 다른 용어 Y의 일부로 포함되면 X는 Y에 내포되었다고 정의한다[8]. 두 개의 용어 X와 Y가 동일한 분류에 포함된 용어이고, Y가 W_1XW_2와 같이 X를 내포하고 있을 경우, X는 기반 용어이고 W_1과 W_2는 X의 수식어라고 정의한다. 수식관계를 이용한 용어의 전문성은 식 (9)와 같다.

$$Spec(Y) = Spec(X) + \alpha \cdot Spec(W_1) + \beta \cdot Spec(W_2) \quad (9)$$

여기에서 $Spec(X)$, $Spec(W_1)$, $Spec(W_2)$는 2.1.1 장에서 제안한 방법 중에서 한 가지를 선택하여 계산한다. α와 β는 0과 1 사이의 값을 가지며, $Spec(Y)$가 지나치게 커지는 것을 방지하기 위하여 사용된다.

2.2 문맥정보를 이용한 계산

상하위어 관계를 가지는 두 용어에서 구성하는 단어들이 상이할 경우, 용어의 구성 정보만을 이용한 전문성 계산방법을 적용하기에는 문제점이 있다. 이 장에서는 이 단점을 보완하기 위하여 용어의 문맥정보를 이용하여 전문성을 계산하는 방법을 설명한다. 일반적으로 일상적인 용어일수록 다른 단어의 수식을 받을 확률이 높고, 전문적인 용어일수록 용어 내부에 많은 정보를 내포하고 있기 때문에 다른 단어의 수식을 받을 가능성이 낮다[6]. 따라서 용어의 수식어의 분포를 전문성 계산을 위한 문맥정보로 사용한다. 주어진 전문용어가 나타나는 문장의 의존 구조를 Conexor 파서[1]를 이용하여 분석한 뒤, 그 용어를 직접 수식하는 수식어들을 추출하여 문맥정보로 이용한다. 먼저 어떤 용어를 수식하는 단어의 분포에 대한 엔트로피를 식 (10)과 같이 계산한다.

$$H_{mod}(t_k) = -\sum_i [p(mod_{k,i}, t_k) * \log p(mod_{k,i}, t_k)] \quad (10)$$

여기에서 $p(mod_{k,i}, t_k)$는 $mod_{k,i}$가 t_k를 수식할 확률을 나타내고, 식 (11)과 같이 추정된다.

$$p_{MLE}(mod_{k,i}, t_k) = freq(mod_{k,i}, t_k) / \sum_j freq(mod_{k,j}, t_k) \quad (11)$$

여기에서 $freq(mod_{k,i}, t_k)$는 전체 코퍼스에서 $mod_{k,i}$가 t_k 를 수식하는 회수를 나타낸다. 식 (10)에서 계산된 엔트로피는 모든 $(mod_{k,j}, t_k)$ 쌍에 대한 평균 정보량을 나타낸다. 전문적인 용어일수록 수식어의 분포가 단순하기 때문에 낮은 엔트로피를 가지고, 일상적인 용어일수록 수식어의 분포가 복잡하기 때문에 높은 엔트로피를 가진다. 따라서 전문적인 용어일수록 높은 정보량을 가지도록 하기 위하여, 식 (12)와 같이 최고 엔트로피 값에서 그 용어의 엔트로피 값을 뺀 값을 그 용어의 정보량으로 정의하고 식 (4)의 $I(x_k)$에 대응시킨다.

$$I(x_k) \approx \underset{1 \le i \le n}{Max}(H_{mod}(t_i)) - H_{mod}(t_k) \quad (12)$$

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

2.3 내부정보와 문맥정보를 결합한 계산

2.1절과 2.2절에서 제안한 두 방법의 단점을 보완하기 위하여 식 (4)의 $I(x_k)$를 식 (13)과 같이 두 방법을 혼합하여 계산할 수 있다.

$$I(x_k) \approx \frac{1}{\gamma(\frac{1}{I_{Internal}(x_k)}) + (1-\gamma)(\frac{1}{I_{Context}(x_k)})} \quad (13)$$

여기에서 $I_{Internal}(x_k)$와 $I_{Context}(x_k)$는 각각 t_k의 내부 구조를 이용한 정보량과 문맥정보를 이용한 정보량을 0과 1사이의 값으로 정규화한 값이다. $\gamma (0 \le \gamma \le 1)$ 는 두 값의 가중치를 나타낸다.

[1] http://www.conexor.com

$\gamma = 0.5$ 인 경우는 두 값의 조화평균 값이다. 따라서 두 값이 공통적으로 높은 값을 가질 경우에 높은 전문성 값을 가진다.

3 실험 및 평가

MeSH[2] 트리에서 "Metabolic Diseases(C18.452)"를 루트 노드로 가지는 하위 트리에 포함된 전문용어 436개를 대상으로 제안한 전문성 계산 방법을 실험하였다. 대상 전문용어를 검색어로 사용하여 MEDLINE[3] 데이터베이스에서 170,000개의 논문 요약문(약20,000,000 단어)을 추출하였다. 추출된 논문 요약문을 Conexor 파서로 분석한 뒤 1) 전문용어의 빈도수, 문서 빈도수, tf·idf, 2) 전문용어의 수식어 분포 3) 전문용어 구성단어의 빈도수, 문서 빈도수, tf·idf를 추출하였다. 정확률(precision)과 적용율(coverage)을 이용하여 제안한 방법의 성능을 평가하였다. 정확률은 MeSH 트리에서 전문성 값을 비교할 수 있는 모든 부모-자식 관계 중에서 올바른 전문성 값을 가지는 관계의 비율로 정의된다. MeSH 트리에서 하위어의 전문성 값이 상위어의 전문성 값보다 높은 경우 제안한 전문성 계산 방법이 타당하다고 볼 수 있다. 적용율은 전체 용어 중 전문성을 계산할 수 있는 용어의 비율로 정의된다.

표1과 같이 구성 정보를 이용하는 방법, 문맥정보를 이용하는 방법, 두 방법을 혼합한 방법으로 나누어 용어의 전문성 값을 계산한 뒤 평가하였다. 구성정보를 이용하는 방법에서는 각각 수식구조 정보를 이용한 경우와 이용하지 않는 경우를 나누어서 실험하였다. 또한 용어 자체의 빈도수와 tf·idf 값을 이용한 실험도 추가적으로 실행하였다. 내부 구성정보를 이용한 방법 중에서는 구성단어의 tf·idf값과 용어의 수식구조 정보를 이용한 경우 정확률 78.9%, 적용율 100%로 가장 좋은 성능을 보였다. 이 결과는 tf·idf값이 단순한 빈도수 정보보다 더 유용하다는 사실과, 내포관계를 이용하여 용어 내부의 수식구조를 분석하는 방법이 유용하다는 사실을 설명한다. 또한 용어의 구성 정보를 이용할 경우 용어 자체의 빈도수와 tf·idf를 이용하는 경우보다 좋은 성능을 보였다. 문맥정보를 이용하는 방법이 용어 구성 단어의 빈도수와 tf·idf를 이용하는 방법보다 낮은 성능을 보인 이유는 전문용어는 그 자체로 많은 정보를 가지고 있기 때문에 코퍼스에서 충분한 수식어 정보를 얻을 수 없었기 때문이라고 판단된다. 용어의 내부 구조를 이용하는 방법과 문맥정보를 이용하는 방법에서 가장 좋은 성능을 나타낸 두 가지 방법을 결합한 실험에서는 정확률 82.0%, 적용율 70.2%의 성능을 보였다. 이 방법이 가장 높은 정확률을 보인 것은 두 가지 방법의 장점이 서로 상보적으로 작용했기 때문으로 판단된다.

가장 높은 정확률을 나타낸 실험에서 발생한 오류의 유형은 다음과 같다. 첫째, MeSH 트리의 중간 노드 중에서 질병명이 아니고 질병의 분류를 나타내는 경우가 있기 때문에 오류가 발생한다. 예를 들어 "Acid-base imbalance (C18.452.076)"는 흔히 사용하는 질병의 이름이 아니고 "산염기평형이상"이라는 질병의 분류 이름을 나타내기 때문에 실제 코퍼스에서 많이 나타나지 않는다. 따라서 하위어보다 높은 전문성 값을 가지는 오류가 발생한다. 둘째, 하위어보다 상위어의 tf값, idf값이 더 높지만 정확한 이유를 알 수는 없는 경우이다. 이 경우는 코퍼스의 양이 충분하지 않았기 때문일 수도 있고, 전문용어의 이형태를 고려하지 않았기 때문일 수도 있다. 셋째, 문맥정보를 이용한 방법에서도 코퍼스에서 추출한 문맥정보가 상하위어 관계에 대한 가정과 일치하지 않아서 오류가 발생한다. 예를 들면 "Nephrocalcinosis (C18.452.174.130.560)"가 상위어

구분			정확률	적용율
용어 빈도수			60.6	89.5
용어 tf·idf			58.2	89.5
구성 정보	구성단어 빈도수	빈도수	65.3	100.0
		빈도수 + 수식구조	76.8	100.0
	구성단어 tf·idf	tf·idf	63.4	100.0
		tf·idf + 수식구조	78.9	100.0
문맥 정보			70.0	70.2
구성+문맥(tf·idf + 수식구조, γ=0.8)			82.0	70.2

표1. 용어의 전문성 실험 결과 (%)

"Calcinosis (C18.452.174.130)"보다 더 다양한 수식어의 수식을 받기 때문에 더 낮은 전문성 값을 가진다. 이 경우는 조어법 분석을 통하여 "Nephrocalcinosis"가 수식어 "Nephro"와 기반단어 "Calcinosis"로 구성된다는 사실을 파악하면 수식구조를 이용한 전문성 계산방법을 적용할 수 있다.

4 결론 및 향후 연구

본 논문에서는 용어가 전문적인 정보를 많이 포함할수록 전문성이 높다고 가정하고, 용어의 구성 정보와 문맥정보를 이용하여 전문성의 정도를 정량적으로 계산하는 방법을 제안하였다. 실험에서 용어의 내부 구성 정보를 이용하는 방법, 문맥 정보를 이용하는 방법, 그리고 두 가지 방법을 조합한 방법으로 용어의 전문성을 계산하였고, 의학용어 분류체계인 MeSH 트리에 적용하여 평가하였다. 실험결과 두 가지 방법을 조합한 경우 가장 높은 정확률(82.0%)를 보였다.

향후 제안한 방법이 용어의 이형태를 고려할 수 있도록 수정되어야 하고, 전문용어 조어법 분석을 통하여 단어 내부에 포함된 정보도 추출할 수 있는 방법에 대한 연구도 필요하다. 마지막으로 제안한 방법을 전문용어 추출 등 응용 분야에 적용할 계획이다.

참고문헌

[1]J.C. Sager, "Handbook of Terminology Management Vol.1", John Benjamins publishing company, 1997

[2]M. A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora", in the Proc. of the Fourteenth International Conference on Computational Linguistics, 1992

[3]S.A. Caraballo, "Automatic construction of a hypernym-labeled noun hierarchy from text", in Proc. of ACL, 1999

[4]G. Grefenstette, "Explorations in Automatic Thesaurus Discovery", Kluwer Academic Publishers, 1994

[5]M. Sanderson and B. Croft, "Deriving concept hierarchies from text", in Proc. of the SIGIR, 1999

[6]S. A. Caraballo and E. Charniak, "Determining the Specificity of Nouns from Text", in the Proc. of the Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora, 1999

[7]S. Haykin, "Neural Network", IEEE Press, 1994, pp. 444

[8]K. Frantzi, S. Anahiadou, H. Mima, "Automatic recognition of multi-word terms: the C-value/NC-value method", Journal of Digital Libraries, Vol. 3, Num. 2, 2000

[9] 오종훈, 이경순, 최기선, "분야간 유사도와 통계기법을 이용한 전문용어의 자동 추출", 정보과학회논문지: 소프트웨어 및 응용 제29권 제1호, 2000

[2] Medical Subject Headings: 미국 National Library of Medicine(NLM)에서 관리하는 의료 분야 용어체계

[3] NLM에서 관리하는 의학/생물학분야 자료 데이터베이스