수식어구를 이용한 한국어 용어의 전문성 측정

*구희관° **정한민 **이병희 **성원경
*과학기술연합대학원대학교 응용정보과학
**한국과학기술정보연구원 차세대정보시스템연구실
{hkkoo°, jhm, bhlee, wksung}@kisti.re.kr

Measuring the Specificity of Korean Terms Using Modifiers

*Heekwan Koo° **Hanmin Jung **Byeong-Hee Lee **Won-Kyung Sung
*Practical Information Science, UST
**Information System Research Lab., KISTI

요 약

용어의 전문성은 전문용어의 판별 기준을 정하고 그 계층구조를 밝히는데 유용하다. 본 논문에서는 말뭉치로부터 추출한 한국어 용어의 전문성을 측정하는 효과적인 방법을 제안한다. 말뭉치에서 관형형 전성어미('ㄴ/은/는')가 부여된 전문용어와 함께 출현하는 수식어구는 일반명사의 수식어구보다 제한적인 형태로 나타난다. 이런 점에 착안하여 본 논문에서는 수식어구를 포함하는 문맥정보에 대해 엔트로피를 측정하여 용어의 전문성을 측정하였다. 이를 위해 한국어 수식어구를 분석하고 기존 전문성 측정 방법에서 간과되어진 수식어구 출현빈도를 고려하여 엔트로피를 상대적 비율로 계산함으로써 한국어에 적합한 전문성 측정을 하였다. 400만 어절의 신문 말뭉치에서 추출한 전문용어와 ETRI 시소러스를 이용하여 실험을 해 본 결과 본 논문에서 제안하는 한국어 용어 전문성 측정방법이 효과적임을 알 수 있었다.

1. 서 론

전문성은 용어에 대한 전문용어 판별 기준을 정하고 그 계층구조를 밝혀내는데 유용하다. 특히 전문용어의 전문성은 일반용어보다 수식어구의 형태가 제한적으로 나타나는 특징을 보인다[1][4][5].

옥철영[2]에서는 관형형 전성어미가 부여된 용언류를 피수식어인 명사에 대한 의미 속성이라 규정하고, 이것과 뒤의 명사에 대한 어휘 연관도를 측정하였다. 이러한 수식구조를 가지는 용어들은 논리형식 종류 중 상당부분을 차지하고 있기 때문에 이 정보를 유용하게 활용할 수 있다[3].

이 수식관계는 표 1에서와 같이 상당히 다양한 형태로 나타나게 되는데 문맥정보로 활용하기 위해서는 그 의미적인 경계를 적절하게 판단해야 한다. 이를 위해 본 논문에서는 한국어에 적합한 수식어구 자질과 전문성 측정방법을 제안한다.

2. 한국어 전문성 평가 방법

2.1 한국어 수식어구 분석

수식어구와 피수식어 간의 의미적 경계를 구하기 위해 수식어구를 형태소 결합구조별로 살펴보면 다음과 같다.

① 수식어구는 용언과 관형형 전성어미('ㄴ'/'은'/'는')가 결합된 형태를 가진다. "새로운 화장품" 이나 "높은 부가가치"등은 두 개의 형태소를 수식어구로 가지는 예이다.

② 세 개의 형태소로 이루어진 형태를 고려하면 추로 명사, 접사와 그리고 관형형 전성어미로 구성되며 (표1에서 95%), "유용한 재테크정보"나 "등장한 집권세력"과 같은 예가 있다. 이 두 가지 경우들은 수식어구가 피수식어와 직접적으로 의미적 관계를 가진다.

③ 수식어구의 용언과 어미결합 형태 앞에 주어-술어의 관계나 목적어-술어의 관계로 확장되는 경우는 의미적 수식어구 범위가 네 개 이상의 형태소를 포함한다.

표 1. 의미적 결합을 고려한 수식어구 형태

형태소5	형태소4	형태소3	형태소2	형태소1	계
				어미	15
			형용사	어미	13
			동사	어미	2
		비서술형명사	접사	어미	23
		비서술형명사	형용사	어미	1
	비서술형명사	비서술형명사	서술격조사	어미	7
	비서술형명사	조사	형용사	어미	3
	비서술형명사	조사	동사	어미	3
	비서술형명사	서술형명사	접사	어미	3
	서술형명사	비서술형명사	서술격조사	어미	3
	동사	어미	조사	어미	3
	비서술형명사	비서술형명사	형용사	어미	1
	비서술형명사	비서술형명사	동사	어미	1
	서술형명사	접사	어미	어미	1
	대명사	조사	동사	어미	1
	동사	어미	형용사	어미	1
비서술형명사	조사	서술형명사	접사	어미	6
비서술형명사	비서술형명사	조사	동사	어미	3
비서술형명사	서술형명사	조사	동사	어미	2
동사	어미	의존명사	형용사	어미	2
비서술형명사	조사	부사	동사	어미	1
비서술형명사	조사	비서술형명사	서술격조사	어미	1
비서술형명사	서술격조사	비서술형명사	서술격조사	어미	1
비서술형명사	서술격조사	서술형명사	접사	어미	1
비서술형명사	서술형명사	비서술형명사	서술격조사	어미	1
형용사	어미	서술형명사	접사	어미	1

이는 옥철영[2]에서 사용한 수식범위를 의미적으로 확장한 것이다. 수식어구 형태의 제한성을 정확히 측정하기 위해서는 수식어구 범위에 따른 실험을 통해 확인해 보아야 한다. 표 1은 2004년도 전자신문에서 추출한 100개 전문용어에 대한 수식어구의 의미적 경계 형태소별 구분을 보여준다.

2.2 용어의 전문성 측정

기존 영어를 대상으로 용어의 전문성을 측정하는 여러 연구가 있었다[4][5]. 그렇지만 한국어에 대해 수식어구 분석 및 이를 고려한 연구는 미흡하다. 본 절에서는 한국어 용어의 전문성 측정 기법을 제안하고자 한다.

한국어 용어t_k를 수식하는 n개의 형태소 분포에 대한 엔트로피를 측정하는 식은 다음과 같다.

$$Hmod(t_k, n) = -\sum_i P(\mathrm{mod}(n)_i, t_k) \times \log P(\mathrm{mod}(n)_i, t_k)] \quad (1)$$

여기에서 $P(mod(n)_i, t_k)$는 $mod(n)_i$가 t_k를 수식하는 확률이며 n의 크기에 따라 그 값이 변한다 [1][4]. 본 논문에서는 n을 가변 변수로 두어 한국어에 적합한 n값 (적용할 수식어구 범위)을 찾고자 한다.

$$P_{MLE}(\mathrm{mod}(n)_i, tk) = \frac{freq(\mathrm{mod}(n)_i, tk)}{\sum_j freq(\mathrm{mod}(n)_j, t_k)} \quad (2)$$

식 (2)는 [1]에서 정의한 확률값으로 본 논문에서도 그대로 적용한다.

$$Hpmod(t_k, n) = -\sum_i P(pmod(n)_i, t_k) \times \log P(pmod(n)_i, t_k)] \quad (3)$$

식 (3)은 수식어구 자질을 형태소가 아니라 품사로 정의한 것이다. $P(pmod(n)_i, t_k)$ 역시 식 (2)와 같은 방식으로 계산한다.

$$P_{동사\_어미}(t_k) = \frac{freq(pmod(동사\_어미, t_k))}{\sum_j freq(pmod_j, t_k)} \quad (4)$$

$$P_{접사-어미}(t_k) = \frac{freq(pmod(접사-어미, t_k))}{\sum_j freq(pmod_j, t_k)} \quad (5)$$

$$P_{형용사-어미}(t_k) = \frac{freq(pmod(형용사-어미, t_k))}{\sum_j freq(pmod_j, t_k)} \quad (6)$$

식 (4), (5), (6)은 특정 수식어구 조합이 나타날 확률이다.

적절한 한국어 용어 전문성 평가 방법을 찾기 위해 다음 장의 실험을 통해 상기 수식들을 평가한다. 특정용어에 대해 결과 값이 낮다면 수식어구 분포가 단순한 경우이므로 해당 용어의 전문성이 높다고 말을 할 수 있다.

3. 실험 및 평가

3.1 수식어구 전문성 측정 방법 실험 및 평가

2004년도 전자신문(400만 어절)에서 추출한 용어들 중 일반용어와 전문용어[1]를 각각 10개씩 선정했다. 이들에 대해 수식어구에 포함되는 형태소 수와 품사 수를 변화시키면서 (1~5개 까지)[2] 전문성을 측정하고, 이들 중 변별력이 높게 나타나는 것을 이후 3.2절의 실험에 사용한다. 다음은 본 실험에서 사용된 일반용어와 전문용어 목록이다.

●일반용어
주상복합, 겨울방학, 부가가치, 고정관념, 신용카드
국회의원, 생활습관, 고속철도, 근린상가, 건강식품

●전문용어
온라인게임, 줄기세포, 초고속인터넷, 디지털카메라
E메일, 인터넷전화, 캐주얼게임, 모바일게임
문자메시지, MP3플레이어

표 2. 다양한 전문성 측정 방법에 따른 일반용어와 전문용어의 변별력

	일반용어	전문용어	차이
$Hmod(t_k,1)$	0.369741535	0.307758477	0.061983058
$Hmod(t_k,2)$	0.897789742	0.808632326	0.089157416
$Hmod(t_k,3)$	1.178828996	0.979669562	0.199159434
$Hmod(t_k,4)$	1.060965884	0.969578198	0.091387686
$Hmod(t_k,5)$	1.081123074	0.992633273	0.088489801
$Hpmod(t_k,1)$	0.028087127	0.130055762	-0.101968635
$Hpmod(t_k,2)$	0.511560342	0.526008153	-0.014447811
$Hpmod(t_k,3)$	0.823263227	0.764205258	0.059057969
$P_{동사\_어미}(t_k)$	0.200393716	0.186660562	0.013733154
$P_{접사\_어미}(t_k)$	0.256370495	0.363681319	-0.107310824
$P_{형용사\_어미}(t_k)$	0.34476898	0.206098901	0.138670079

적절한 의미경계를 찾기 위한 평가방법의 실험 결과는 표 2와 같다. 표 2는 일반용어와 전문용어 목록에 대한 전문성 측정값의 평균을 보여준다. 이중 $Hmod(t_k,3)$이 한국어의 전문성 측정을 위해 가장 적합한 방법으로 보여진다. 전문성 측정 결과값의 추이를 살펴보면 $Hmod(t_k,3)$을 정점으로 하여 점차 감소하는 것을 알 수 있다.

표 1에서 수식어구 범위가 4이상인 경우 그 형태가 너무 다양해져서 오히려 변별력을 떨어뜨리는 결과를 가져오며, 범위가 2인 경우는 변별력을 가지기 힘든 획일적인 형태를 가진다. 본 실험에서도 $Hmod$, $Hpmod$가 모두 3일 경우가 가장 좋은 성능을 보이는 데, 특히 $Hmod(t_k,3)$이 최적의 성능을 보인다. $Hpmod$, $P(동사\_어미(t_k))$, $P(접사\_어미(t_k))$, $P(형용사_어미(t_k))$등은 변별력을 가지지 않는다.

1) 전문용어에 대한 판별을 위해 두 사람이 과학기술용어를 중심으로 동시 평가하며 모두 인정하는 경우를 기준으로 삼음.
2) 형태소의 경우 1~5까지, 품사의 경우 1~3까지

표 3. ETRI 시소러스를 이용한 용어 전문성 측정 결과 *Hmod(t_k,3)*기준

출현빈도	엔트로피 최대값	측정된 엔트로피	최대값 대비비율	용어	의미코드
74	1.869232	1.564569	0.837012	물체	030102
284	2.453318	2.060939	0.840062	물질	0301021904
44	1.643453	1.253348	0.762631	원소	03010219041U
48	1.681241	1.024963	0.609647	플루토늄	03010219041U0I
98	1.991226	1.370668	0.688354	우라늄	03010219041U1X01
219	2.340444	1.619435	0.691935	세포	0302070n
12	1.079181	0.276435	0.256152	단세포	0302070n0U
20	1.30103	0.819382	0.629795	체세포	0302070n0E
296	2.471292	2.147383	0.868931	물건	030101
11	1.041393	0.98666	0.947443	물자	03010173
45	1.653213	1.460855	0.883646	물품	0301017309
11	1.041393	1.041393	1	기구	030101730910
717	2.855519	2.350374	0.823099	컴퓨터	03010173091001010o02
276	2.440909	1.93902	0.794384	노트북	03010173091001010o020C

3.2 수식어구 선정 실험 및 평가

ETRI 시소러스와 전자신문에서 추출한 용어를 대상으로 일반용어와 전문용어에 대한 상대적 전문성 평가 결과를 보여준다. 대상 용어는 ETRI 시소러스에서 용어와 비교하여 커버리지가 높은 3개의 개념 클러스터에 대해 이전 실험에서 가장 성능이 좋았던 *Hmod(t_k,3)*로 실험을 하였다.

실험한 결과는 표 3과 같다. 엔트로피 최대값은 해당 용어가 출현빈도 기준으로 완전 분산된 경우 가질 수 있는 최대값이다. 다음 필드는 *Hmod(t_k,3)*로 측정된 엔트로피이다. 최대값 대비 비율은 이 둘 간의 차이를 비율로 표시한 것으로 이 수치가 작을수록 수식어 형태가 단순해짐을 보여준다. 의미코드는 두 캐릭터가 한 레벨을 의미한다. 예를 들어, 5레벨의 "물질"과 6레벨의 "원소"는 IS-A 관계를 가지는 부모-자식이다.

시소러스 상의 상위 레벨 용어들은 대부분 일반용어이므로 전문성 측정에 있어서 변별력이 없다. 그러나 "원소"를 비롯한 상위 레벨 용어들에 대해 "플라토늄"과 "우라늄"은 최대값 대비 비율이 차이를 보인다. 동일한 결과를 다른 두 개의 클러스터에서도 발견할 수 있다. "체세포"와 "단세포"의 상위레벨용어인 "세포"는 이들보다 높은 최대값 대비 비율을 보이며 "컴퓨터"와 "노트북"도 상위 레벨 용어들에 대해 같은 결과를 보인다.

엔트로피를 이용하는 방법은 용어에 대한 수식어구 출현 빈도수가 증가하면서 절대값이 커지기 때문에 절대값을 고려한 측정값의 보정이 필요하다. 출현 빈도수의 차이가 심한 용어들 간의 상대적인 전문성 측정에 이러한 방법이 필요하다. 이러한 보정 기법은 빈도수의 차이를 가지는 말뭉치 기반의 용어 추출 및 선정을 수행하는 모든 연구들에 적용이 가능하다.

4. 결론

본 논문은 관형형 전성어미가 부여된 수식어구와 함께 출현하는 전문용어에 대해 한국어에 적합한 전문성 측정방법을 제안하였다. 엔트로피 계산에 있어서 *Hmod(t_k,3)*가 한국어에 가장 적합한 평가방법임을 3.1절 실험 1을 통하여 보였다. 또한 3.2절 실험 2를 통해 용어간 수식어구 출현빈도 차이를 최대값 대비 비율로 보정함으로써 말뭉치 기반 연구에 적합한 전문성 측정방법을 제안했다.

향후 본 논문에서 제안한 전문성 측정을 실제 용어 선정을 위한 방법으로 적용하기 위해 커버리지 향상 기법을 중심으로 연구할 예정이다.

참고문헌

[1] 류법모, 배선미, 최기선, "구성정보와 문맥정보를 이용한 용어의 전문성 측정 방법", 정보과학회 춘계학술대회, 2004.

[2] 옥철영, 옥은주, 이왕우, "수식 관계 구문에서 공기제약 어휘간의 정보량 측정", 한글, 제255호, 2002.

[3] 장두성, 오종훈, 최기선, "이벤트 탐색을 사용하는 일정 영역 질의 응답 시스템의 구현, 제13회 한글 및 한국어 정보처리 학술대회, 2001.

[4] Pum-mo Ryu, "Determining the Specificity of Terms using Compositional and Contextual Information", Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 2004.

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