제36회 한국정보처리학회 추계학술발표대회 논문집 제18권 제2호 (2011. 11)

용어 정규화 방법

황명권, 정도헌, 성원경

한국과학기술정보연구원 정보유통본부 정보기술연구실
e-mail:mg.hwang@gmail.com, heon@kisti.re.kr

A Method for Term Normalization

Myunggwon Hwang, Do-Heon Jeong
Dept of Information Technology Research, Korea Institute of Science and Technology Information (KISTI)

요 약

자연어 처리에서 큰 걸림돌 중의 하나는 용어의 표현 다양성이라 할 수 있다. 용어들은 시제, 단수/복수 형태, 경우에 따라서는 동일한 의미의 다른 용어로 대체되어 사용될 수 있으며, 이러한 용어의 사용은 동일한 의미를 다르게 해석하는 원인이 되기도 한다. 이에 본 연구에서는 다양한 형태의 용어들을 하나의 표준화된 형태로 정규화 하는 방법을 제안한다.

1. 서론

용어들은 그 표현 형태가 다양하다. 특히 두 개 이상의 단어로 구성되어 있는 전문 용어(기술 용어)의 경우는, 일반 용어에 비해 더욱 많은 변형(Variants)을 갖는다. 일반 용어의 경우는 시제나 단/복수에 따라 달라지는 것이 대부분이지만, 전문 용어의 경우는 동일 의미의 다른 용어로 대체 그리고 전치사와 함께 사용되어 단어의 재배열 또한 추가적으로 발생할 수 있다.

자연어 처리 분야에서 정제된 용어의 수집 과정이 중요하지만, 다른 형태로 표현된 동일한 의미의 용어들을 군집화 하는 것은 쉬운 일이 아니다. 이에 반해 현재까지의 용어 정규화에 대한 연구는 일부[1, 2]에서만 수행되었을 뿐, 그에 대한 관심이 미비한 상태이다. 이에 본 연구에서는 대용량의 문서 집합에서 추출된 기술 용어 집합에서 표준화된 형태의 용어들을 선정하고, 이 용어들의 다양한 표현 형태를 파악하는 방법을 제안한다. 또한 간단한 실험과 평가를 통하여 본 연구의 가능성을 제시하도록 한다.

본 연구의 구성은 다음과 같다. 2장에서 본 연구의 핵심인 용어 정규화 방법을 기술하고 이에 대한 평가를 간단히 다룬다. 3장에서 가능성 및 향후 연구와 함께 마무리한다.

2. 용어 정규화 방법

본 연구의 구성은 (그림 1)과 같으며, 본 연구는 용어의 정규화 과정에만 집중되고 있기 때문에 기술용어 집합을 형성하는 “Term Extractor" 부분은 설명하지 않는다. 용어 정규화를 위한 본 연구는 표준화된 용어 집합과 변형된 용어 후보를 형성하는 과정, Co-occurrence 단어들을 이용하여 문맥정보를 형성하는 과정, 그리고 타겟 용어(변형된 용어)의 후보 정답 용어(표준 용어) 형성을 통한 유사도 측정 방법으로 구성된다. 본 장에서는 각 모듈에 대해 예제와 함께 간단히 설명하도록 한다.

2.1 표준 용어와 변형된 용어 집합 형성

NDSL[1] 논문 초록 집합에서 추출한 기술용어들을 위키피디아(이하 위키)와 매칭하는 과정을 거친다. 위키는 전세계 전문가들의 집단 지성을 기반으로 형성된 백과사전이기 때문에, 위키에 정의된 용어들은 표준화된 용어, 그렇지 않은 것은 변형된 용어로 간주한다. 89,231개로 구성된 초기 용어 집합에서 10,684와 78,547개 용어로 구성된 표준 집합과 변형 집합을 각각 구성할 수 있었다.

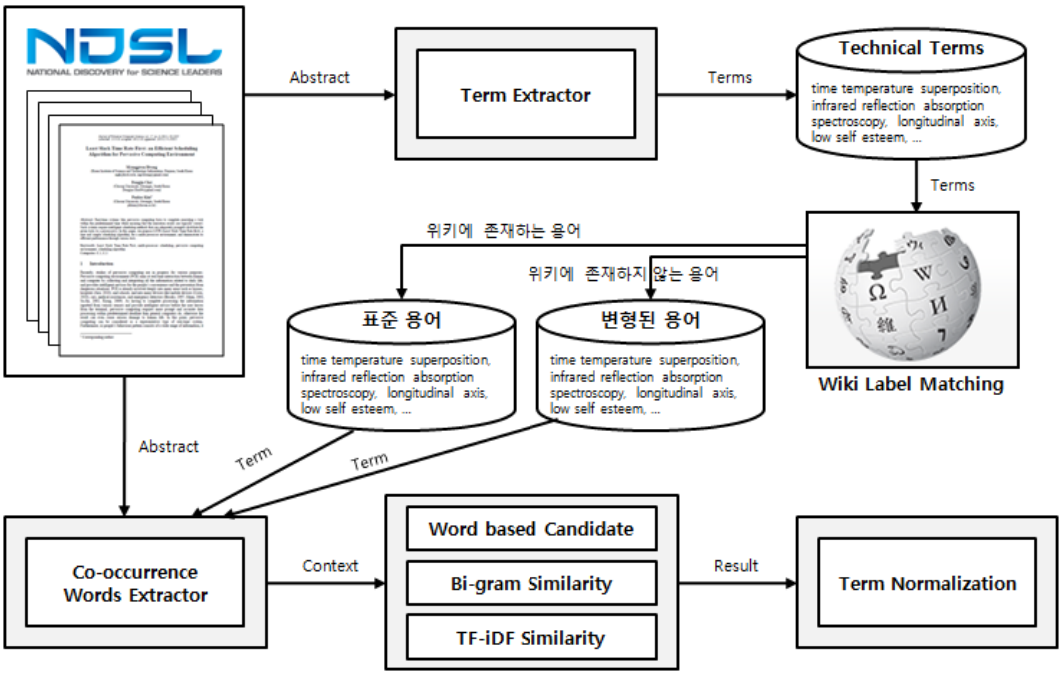
2.2 문맥 정보 형성

변형 용어의 표준 형태를 파악하기 위해 본 연구에서는 각 용어들과 동일한 문장에서 함께 출현한 단어들을 문맥 정보로 활용한다. NDSL 초록 전체에서 같은 모양의 기술용어는 의미가 같은 것으로 간주하며, 그와 함께 출현한 모든 명사들을 추출하여 구성한다.

2.3 타겟 용어의 후보 정답 용어 집합 형성

타겟(변형) 용어의 후보 정답(표준 용어) 집합을 구성하기 위해 기술용어를 구성하는 단어를 비교한다. 만약 타겟 용어를 구성하는 단어가 표준 용어의 것과 하나 이상 일치한다면 그 표준 용어를 후보 정답 용어 집합에 추

1) 국가과학기술정보센터 NDSL (National Discovery for Science Leaders): http://www.ndsl.kr/index.do



(그림 1) 용어 정규화를 위한 전체 과정

가한다. 이때 후보 집합의 규모가 커지는 것을 막기 위해 스톱 워드(Stop Word)들은 비교하지 않는다. <표 1>은 그 예를 보이고 있다.

<표 1> 타겟 용어의 후보 정답 집합 예

타겟 용어	object oriented data model
후보 정답 집합	agent based **model**, automatic **data** processing, austin **model** 1, agent **oriented** programming, aspect **oriented** programming, aspect **oriented** software development, ashkin teller **model**, breast imaging reporting and **data** system, community **oriented** policing, component **oriented** programming, ...

2.4 유사도 측정 방법

앞의 과정에서 변형 용어와 표준 용어를 구분하고, 타겟 용어의 후보 정답 집합을 형성하였으며, 각 용어에 대한 문맥 정보를 형성하였다. 타겟 용어와 후보 정답 집합 사이의 긴밀성을 측정하기 위해 두가지를 고려한다. 첫째는 용어를 구성하는 단어의 유사성 측정이며, 둘째는 두 용어의 문맥 정보 유사도를 측정하는 것이다.

단어 유사성 측정은 두 용어의 모습이 어느 정도 비슷한지를 측정하는 것이며, 이를 위해 문자 기반 bi-gram (단어를 구성하는 문자들 중 연속하는 2개)들을 비교한다. 용어를 구성하는 단어와 단어를 구성하는 bi-gram을 다음과 같이 표현한다.

$$term = \{w_i, 1 \le i \le n\}$$
$$bigram_w = \{b_j, 1 \le j \le |w|-1\}$$

여기서 *term*은 용어, *w*는 용어를 구성하는 단어, 그리고 *n*은 용어를 구성하는 단어의 개수, *b*는 단어를 구성하는 bi-gram, 그리고 bi-gram의 수는 단어를 구성하는 문자보다 1이 적음을 의미한다. 수식 (1)은 두 단어 사이의 유사도 측정을 표현하고 있다.

$$s(w_k, w_l) = \frac{2 \times P(bigram_{w_k} \cap bigram_{w_l})}{|w_k| + |w_l| - 2} \quad (1)$$

수식 (1)을 바탕으로 용어들 사이의 bi-gram 유사도는 수식 (2)를 이용하여 측정한다.

$$s(term_p, term_q) = \frac{2 \times \sum \arg\max(s(w_{pk}, w_{ql}))}{|term_p| + |term_q|} \quad (2)$$

수식 (2)에서 w_{pk}는 용어(*term*) *p*를 구성하는 단어 *k*를 의미한다. <표 2>는 단어 유사도의 예를 보이고 있다.

<표 2> bi-gram 기반 단어 유사도 측정 예

용어1	용어2	유사도
virus neutralisation	equine infectious anemia virus	0.394
visna maedi virus	maedi visna virus	1.000

<표 2>에서 보이는 것과 같이, 용어유사도는 용어를 구성하는 단어의 배치가 다르더라도 높은 값을 측정할 수 있는 강점이 있다.

본 연구에서는 단어 유사도와 문맥 정보 유사도를 함께 고려하는데, 문맥 정보의 유사도 측정을 위해 TF-IDF (Term Frequency-Inverse Document Frequency)를 이용하여, 문맥 정보를 구성하는 단어들의 가중치를 측정 한다 (본 논문에서 TF-IDF에 대한 설명은 하지 않는다).

용어 정규화를 위한 전체 유사도는 단어 유사도와 문맥 정보 유사도의 곱 연산을 통해 얻는다.

2.5 실험 결과

본 연구의 가능성 평가를 위해 타겟 용어 171개를 임의로 선정하고, 단어유사도와 문맥 정보 유사도를 함께 이용한 전체 유사도 측정에서, 최대값을 갖는 후보만을 이용하여 평가하였다. 171개 중에 40개의 용어만이 올바르게 정규화된 것으로 확인되었으며, 단어 유사도와 전체 유사도의 임계값에 따라 성능이 향상하는 것으로 확인되었다. <표 3>은 단어유사도 0.5이상일 때, 전체유사도 임계값에 따른 평가 결과를 보이고 있다.

<표 3> 단어 유사도 0.5이상, 전체 유사도 임계값에 따른 평가 결과

TH	TF	X	O	Pre.	Rec.	F1
0.1	47	14	33	0.702	0.846	0.767
0.2	34	7	27	0.794	0.692	0.74
0.3	24	5	19	0.792	0.487	0.603
0.4	20	3	17	0.850	0.436	0.576
0.5	18	1	17	0.944	0.436	0.596
0.6	13	0	13	1.000	0.333	0.5
0.7	13	0	13	1.000	0.333	0.5
0.8	12	0	12	1.000	0.308	0.471
0.9	11	0	11	1.000	0.282	0.44
1	10	0	10	1.000	0.256	0.408

TF: 단어 유사도 0.5와 TH에 따라 남아있는 데이터 수, TH: 전체유사도 임계값, Pre.: 정확도, Rec.: 재현율.

<표 3>에서와 같이 용어 유사도 0.5와 전체 유사도 0.1일 때 F1 기준 가장 높은 성능을 보임을 확인할 수 있다. 본 평가에 의해 논문에 제안하는 방법이 용어 정규화에 희망적인 결과를 제시할 수 있음을 보인다.

3. 결론

본 논문에서 자연어 처리의 걸림돌인 동일한 의미의 다양한 표현 형태를 하나의 표준화된 용어로 정규화하는 방법을 제안하였다. 용어 정규화를 위해 용어를 구성하는 단어의 bi-gram을 이용한 단어 유사도와 용어의 문맥 정보를 이용한 TF-IDF를 이용하였으며, 실험에 의해 단어 유사도 0.5 그리고 전체유사도 0.1 이상일 때 가장 우수한 결과를 보임을 확인하였다.

하지만 버전 정보를 포함하는 용어의 경우는 본 연구에서 다양한 에러를 발생함을 알 수 있었다. 예를 들어, 'Vitamin C'와 'Vitamin D'의 경우는 단어 유사도 0.5를 가짐과 동시에 문맥 정보 유사도에서도 높은 값을 갖게 된다. 지속적인 연구를 통해 이러한 문제에 대해 해결점을 제시하도록 하겠다.

참고문헌

[1] James Dowdal, Fabio Rinaldi, Fidelia Ibekwe-SanJuan, and Eric SanJuan, "Complex Structuring of Term Variants for Question Answering," In Proceedings of the ACM Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Vol. 18, pp. 1-8, 2003.

[2] Fidelia Ibekwe-Sanjuan, "Terminological Variation, a Means of Identifying Research Topics from Texts," In Proceedings of International Conference on Computational Linguistics, Vol. 1, pp. 564-570, 1998.