전문용어사전 구축을 위한 전문용어 추출 및 순위화

구희관*, 정한민**, 이병희**, 성원경**
*과학기술연합대학원대학교 응용정보과학전공
**한국과학기술정보연구원, 차세대정보시스템연구실
e-mail: {hkkoo, jhm, bhlee, wksung}@kisti.re.kr

Term Extraction and Ranking for Building Term Dictionary

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

요 약

전문용어는 특정 분야의 전문가 사이에서 통용되는 표현 매체이며, 일반용어에 비해 생성과 소멸의 주기가 짧은 특징을 가지고 있다. 이런 특징 때문에 일반용어 사전구축과 달리 전문용어 사전을 구축하기 위해서는 신속한 대응전략이 필요하다. 이를 위해 본 논문에서는 전문용어 사전 구축을 위한 다음과 같은 두 단계의 과정을 제안한다. 우선 형태소 후처리와 결합규칙을 이용하여 1,200만 어절의 신문 말뭉치로부터 단일어 10만과 복합어 30만의 용어후보를 추출하고, 고빈도 용어 후보 6만개를 선별해 용어지배지수(Term Dominance Value)라는 개념을 도입하여 전문용어를 선정한다. 실험을 통해 용어지배지수 순위와 누적빈도순위 및 최근연도 순위를 비교한 결과 본 논문에서 제안한 용어지배지수가 전문용어 활용도를 나타내는 훌륭한 지표역할을 할 수 있음을 확인할 수 있었다.

1. 서론

각 해당 분야의 최소 지식 표현 매체를 전문용어라 한다. 그러므로 지식 및 정보와 관련된 연구는 모두 전문용어를 떠나서는 생각할 수 없다[3]. 전문용어를 추출할 때, 전문용어 후보를 말뭉치로부터 추출하고, 전문용어 여부를 판별하기위해 일반적으로 빈도 정보를 이용해서 순위를 나타낸다.

전문용어 후보 순위화 방법은 전문용어 사전을 어떤 목적으로 만들 것인지에 따라 이용하는 요소가 달라진다. 전문용어는 전문용어가 속한 학문, 산업분야, 그리고 사회현상과 밀접한 관계를 가지고 있다. 또한, 타 분야로의 전이에 따라 정의와 성격도 당연히 변화한다. 전문용어를 관리적인 측면에서 본다면 용어사용 빈도 변화를 어떻게 사용할 것인가를 고려해야 한다.

이를 위해 본 논문에서는 복합어를 위한 결합규칙과 후처리를 통해 전문용어 후보를 추출하고 용어지배지수라는 지표를 이용하여 순위화하는 방법을 제안한다.

용어지배지수는 주어진 기간 동안 용어의 상대적 활용도를 나타내는 값이고, 사용빈도가 상대적으로 높아져가는 경향의 전문용어들이 우선적으로 사전에 추가할 수 있게 하는 근거를 제공한다. 즉, 전문용어 흐름에 대한 생명주기(life-cycle)를 분석 할 수 있는 중요한 지표가 된다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구, 3장에서는 추출 및 순위화, 4장에서는 실험 및 결과에 대해 기술하고, 5장에서는 결론과 향후 연구 방향을 논의한다.

2. 관련 연구

박종오[1]에서 사용된 전문용어 추출 시스템은 전문용어 추출을 위해 명사의 용어 추출 구문 패턴과 문서정보를 이용해서 전문용어 후보를 추출하였다. 이 추출시스템이 후보를 추출하는 방법은 단순히 명사와 명사의 조합과 외래어와 명사와의 조합 방법만

을 사용하여 전문용어후보를 추출한다. 또 다른 연구로는 이 후보선정 과정 중에 명사에 대한 예외처리 명사를 사어(死語)로 규정하여 제한하려고 하는 연구도 있었다[2].

FASTR[4]은 후보추출 후에 클러스터링 방법을 사용해서 전문용어 후보를 선별한다. 클러스터링 방법은 추출된 전문용어 후보에 대하여 여성형과 남성형의 명사가 다른 형태를 보이는 외국어의 경우, 해당 명사의 단일어 형태를 추출하는 경우에는 적합하였다. 그러나 명사형이 동일한 한국어에서는 이 방법을 적용하기 힘들며 이 클러스터링 방법은 단순히 용어 후보를 검출하는 방법으로 우선순위나 용어변화 정보를 생성하지 않는다.

TFB[5]는 웹기반 말뭉치를 이용하여 전문용어를 추출하고, 사전에 어휘를 추가하는 시스템이다. 그러나 용어 사용빈도 변화를 반영할 수 없다는 약점을 가진다.

전문용어 선정작업 중 말뭉치에서 산출해 낸 통계치를 이용하여 판별하는 방법은 매우 유용한 도구들이다[6]. 그러나 단순히 빈도수를 이용하므로 최근빈도변화를 고려하지 않는 문제점을 가지고 있다.

전문용어 후보 순위화 방법은 다양한 요소를 이용하여 순위화를 하게 된다. 전문용어는 전문용어가 속한 분야와 밀접한 관계를 가지고 있고 빠르게 변화하기 때문에 여러 요소 중에서 용어사용의 빈도변화를 반영해야 한다. 이러한 문제점을 해결하기 위해서 본 논문에서는 두 가지 단계를 이용한 방법을 제안하고 실험하였다.

3. 추출 및 순위화

본 논문에서 제안하는 두 단계 중 첫 단계는 형태소 후처리와 결합규칙을 적용한 복합어 추출하는 것이고, 두 번째 단계는 용어지배지수를 이용한 순위화를 하는 것이다.

후처리 규칙에는 3가지 경우를 고려한다. 전문용어 후보군에서 명사 성격을 제한하여 후보에서 해당 명사를 예외품사로 처리하는 경우, 하나의 명사를 명사들로 분리해야 경우와 그리고 명사와 명사를 하나의 명사로 이어주는 경우로 나누어서 사용한다.

<표 1>은 형태소 후처리와 결합규칙을 적용한 복합어 추출의 4가지 결합규칙을 보여준다. 명사간의 결합에는 명사조합, 파생어미에 대한 결합을 설명하는 파생조합, 어절 간 결합을 대상으로 하는 어절조합, 관형어구를 제거하고 결합을 하는 복합후치사 제거를 보여준다.

복합어 후보생성 결합규칙
{비서술명사|서술명사|적절한 명사|외래어|수사} * {비서술명사|서술명사|적절한 명사|외래어|단위명사}
{비서술명사|서술명사} * {명사형 파생어미|적합한 파생명사형 어미}
하나의 어절 안에서 어절 사이에 표현을 제외하고 명사조합과 같은 패턴
{비서술명사|서술명사|해당명사|외래어|의존명사}* {관형어구}* {비서술명사|서술명사|해당명사| 외래어|의존명사}

<표 1> 복합어 후보생성 결합규칙

전문용어 후보들의 순위화에 이용하는 용어별 용어지배지수(*TDV*)를 계산하는 방법은 다음과 같다.

$$TDV_t = \left(\sum_{i=first\_year}^{last\_year}(NTF_{ti} - ANTF_t) * (YW_i)^2\right) * (YF_t)$$

위의 수식에서 NTF_{ti}는 정규화 용어의 출현빈도이며, YW_t는 연도 별 가중치 및 YF_t는 연도 별로 용어가 출현한 횟수를 의미한다.

TF_{ti} 값은 연도마다 발생한 용어의 빈도수를 말한다. YW_i 값은 해당년도의 가중치 값인데 관찰시작 연도인 1994년의 가중치 값을 0.1로 시작해서 0.1씩 더한 값을 가중치로 사용한다.

용어지배지수 식에서 정규화 NTF_{ti}를 구하는 방법은 해당년도 용어의 출현빈도에 연도별 말뭉치의 크기가 제일 작은 값을 해당년도 말뭉치의 크기로 나눈 값이다. 이를 식으로 표현하면 다음과 같다.

$$NTF_{ti} = TF_{ti} \times \frac{SmallestCorpusSize}{YearCorpusSize_i}$$

용어지배지수 식에서 $ANTF_t$ (정규화 평균용어 빈도)는 다음과 같이 표현된다.

$$ANTF_t = \frac{\sum_{i=first\_year}^{last\_year}(NTF_{ti}))}{YF_t}$$

위의 식에서 YF_t 값은 관찰 연도 중 발생한 용어 발생 연도수를 나타내며, NTF_{ti}를 구하는 수식을 활용하여 $ANTF_t$의 결과값을 구할 수 있다.

4. 실험 및 결과

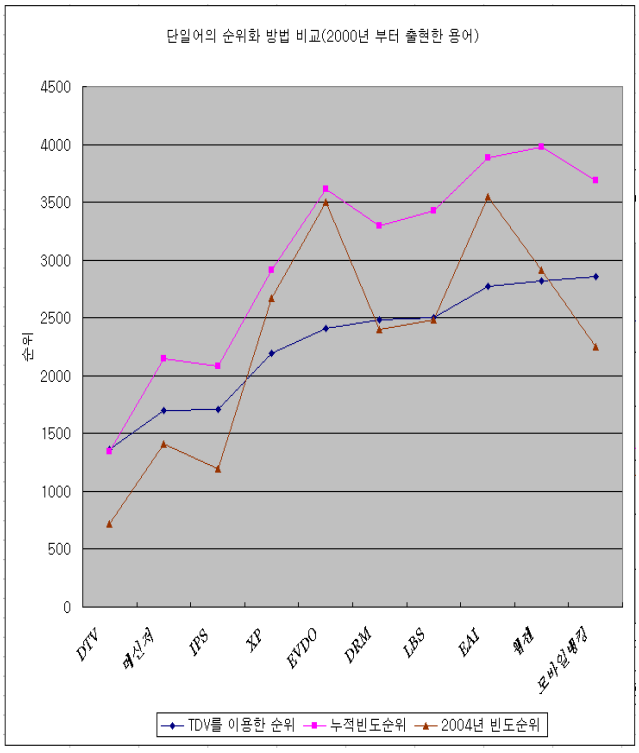
본 논문에서는 웹 로봇을 이용하여 수집된 신문 말

뭉치를 이용하여 실험하였다. 용어에 빈도변화정보를 추출하기 위해 연도 별로 기사를 구별한다. 1994년에서 2004년까지 말뭉치를 생성하고 각각 단일명사와 복합어를 추출한다.

이 과정에서 말뭉치를 형태소 분석을 하고, 품사태깅한 결과를 4가지 결합규칙을 적용하여 연도 별 전문용어 후보를 추출한다. 그리고 각각 용어별 용어지배지수를 연도 별로 구한 후, 합산하여 순위목록을 작성한다.

전체 1,200만 어절의 신문 말뭉치에서 단일명사 99,642개, 복합명사 304,055개의 전문용어후보를 추출하였다. 이중 고빈도 용어 후보 60,000개를 이용해 용어출현 누적빈도 순위, 가장 최근 연도 사용빈도 순위, 용어지배지수 순위와 비교한다.

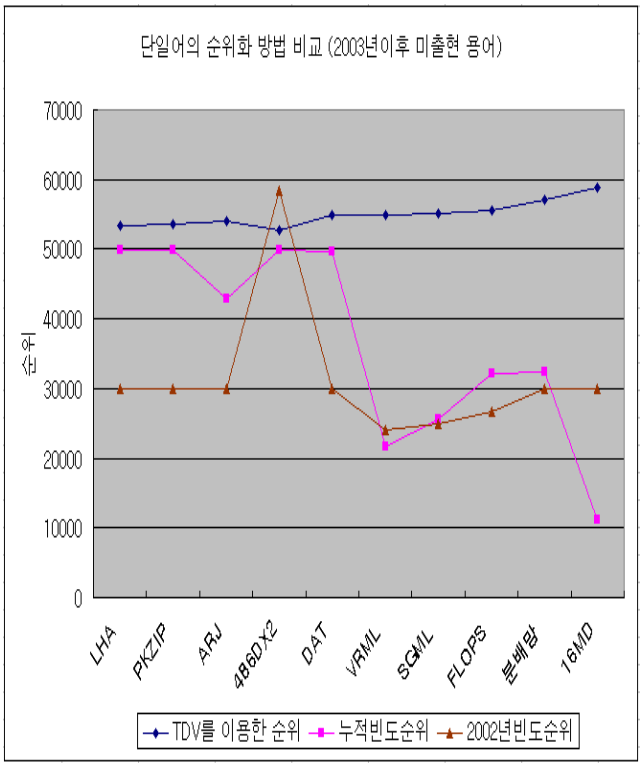
각 용어후보에 대한 용어지배지수는 양수, 0과 음수 값으로 나타난다. 양수값은 전문용어 후보가 연도 별 말뭉치 내에서 사용빈도가 증가되는 경향을 보여준다. 0의 값을 가지는 용어후보는 사용빈도 변화가 거의 없다. 또한, 음수로 나타나는 후보는 사용추세가 감소중이다. 결과적으로 0을 기준으로 용어후보들이 넓게 분포하며, 사용빈도가 높은 용어일수록 전체 분포표에서 양끝에 위치한다.



<그림 1> 단일어의 순위화 방법 비교
(2000년부터 출현한 용어)

<그림 1>은 2000년 이후 나타난 용어후보 10개를 대상으로 2004년도 용어순위와 전체 말뭉치 사용빈도를 계산한 순위를 비교한 것을 그래프로 표현한다. 최근 사용이 증가하는 용어후보들은 전체 누적빈도 순위보다 용어지배지수 순위는 높게 나타난다.

<그림 2>는 2003년 이후 미출현 용어 10개를 대상으로 한다. 전체 빈도는 높더라도 최근 사용빈도가 감소하는 전문용어는 누적빈도 순위를 이용한 방법에 비해 순위가 매우 낮게 나타난다. 그래서 누적빈도 순위 목록에서는 꽤 높은 순위를 보이는 용어후보일지라도 용어지배지수 순위에서는 매우 낮게 나타난다.

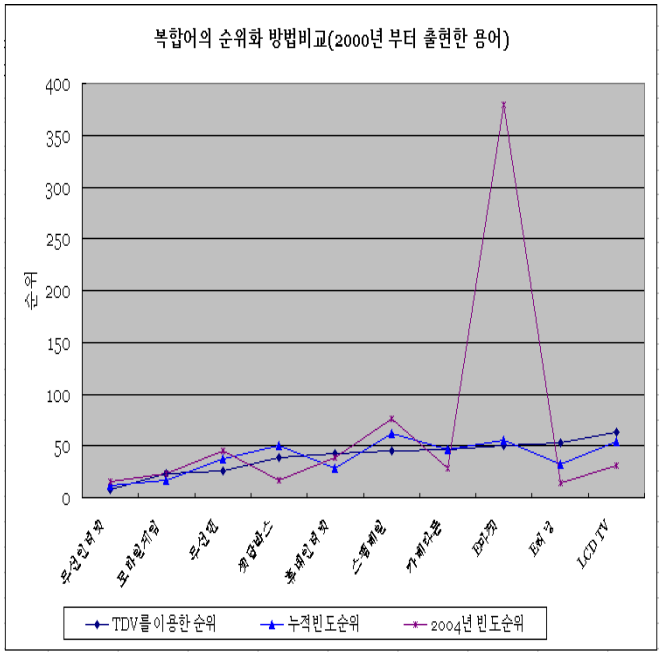


<그림 2> 단일어의 순위화 방법 비교
(2003년 이후 미출현 용어)

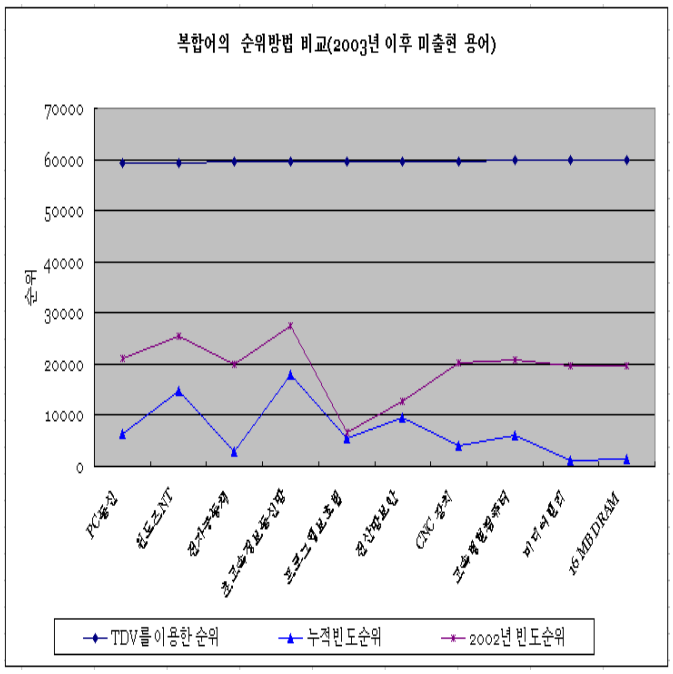
<그림 3>은 복합어 순위들을 보여준다. 단일어 경우와 유사하지만 용어지배지수 순위가 최근에 사용되는 경향이 증가하는 용어후보라는 것을 확인할 수 있다.

<그림 4>의 전체 형태는 <그림 2> 와 유사한 형태를 보여준다. <그림 4>는 용어지배지수 순위가 누적빈도순위보다 적절하게 용어의 빈도변화를 잘 드러내는 것을 보여준다.

요약해보면, 최근 연도에 대한 빈도순위는 전체 용어에 대한 단편적인 정보를 제공한다. 용어지배지수 순위는 누적 빈도 순위보다 용어 선정작업에 용어후



<그림 3> 복합어의 순위화 방법 비교
(2000년부터 출현한 용어)



<그림 4> 복합어의 순위화 방법 비교
(2003년 이후 미출현 용어)

보들에 활용도에 대한 정보를 제공할 수 있다.

그래서 용어지배지수 순위를 만들면 전체 말뭉치나 최근 연도의 사용빈도 순위를 보다 용어의 변화를 수용하는 값으로 순위를 만들어내고 그 순위에서 높게 나타나는 용어일수록 활용도가 높아 우선적으로 검증해야 한다는 것을 확실히 보여준다.

5. 결론

지금까지 본 논문에서는 형태소 후처리와 4가지 결합규칙을 사용한 용어 추출방법과 용어지배지수라는 개념을 이용한 순위화 방법을 제안하고 실험을 통해 용어지배지수 순위와 누적빈도순위 및 최근 연도 순위를 비교하였다.

본 논문이 제안한 용어지배지수는 용어뿐만이 아니라 용어 관련된 분야의 변화를 표현하는데 활용될 수 있을 것이다. 향후 용어지배지수를 활용하여 일반화한 전문용어주기를 만들고, 용어들이 전문용어 주기 중 어느 단계에 해당하는 가를 규명하는 방법에 대한 연구가 이뤄져야 될 것이다.

참고문헌

[1] 박종오, 황도삼, “전문용어 추출시스템”, 한국정보과학회 춘계학술대회, 2000.

[2] 이종인, 한광록, 양승현, 김영섬, “한국어 명사의 시소러스 구축을 위한 시스템 설계 및 구현”, 한국정보처리학회 논문지 A. Vol. 06, No. 2, 1999.

[3] 최기선, 송영빈, "전문용어연구1“, 홍릉과학출판사, 2000.

[4] Didier Bourigault and Christian Jacquemin, “Term Extraction + Term Clustering”, Proceedings of the 9th conference on European chapter of the Association for Computational Linguistics, 1999.

[5] Stefan Evert, Ulrich Heid, Bettina Säuberlich, Esther Debus-Gregor, and Werner Scholze-Stubenrecht, “Supporting Corpus-based Dictionary Updating”, Proceedings of the 11th Euralex International Congress, 2004.

[6] Hiroshi Nakagawa and Tatsunori Mori, “Automatic Term Recognition Based on Statistics of Compound Nouns and Their Components”, Terminology, Vol. 9, No. 2, 2003.