용례 벡터와 웹 자원을 이용한 전문용어 용례의 추출 및 순위화

정하용, 최기선
한국과학기술원 전자전산학과 전산학전공
hymanse@world.kaist.ac.kr, kschoi@cs.kaist.ac.kr

Extraction and Ranking of Term Usages using Usage Vector and Web Resources

Ha-Yong Jung, Key-Sun Choi
Division of Computer Science, Dept. of Electrical Engineering and Computer Science, KAIST

요 약

전문용어의 용례는 일반용어의 용례와 다르게 의미를 드러내는 것이 중요하다.. 또한 사전 및 시소러스와 같은 자원이 부족하다는 특징이 있다. 본 논문에서 우리는 전문용어의 용례를 벡터를 이용한 표현을 통해 더 정량적으로 나타내는 방법을 제안했다. 또한 전문용어의 자료부족 문제를 극복하기 위해 대체적 자원으로 웹을 이용하는 것을 제안했다. 실험 결과, 제안한 시스템은 기존의 시스템에 비해 최대 30%의 성능 향상을 이룰 수 있었다.. 게다가 제안한 시스템에의 추출된 전문용어의 용례는 다른 자연어처리 응용을 위한 보완적 자원으로서의 가능성을 보여줬다.

1. 서 론

용례란 말이나 글이 실제로 쓰이는 예로서, 단어가 의미에 따라 특정 문맥 하에서 실제로 어떻게 쓰이는지를 나타내는 문장이다. 즉, 용례는 문장에 드러난 문맥을 통해서 단어의 의미와 쓰임새를 보여준다고 할 수 있다.

용례 추출 및 순위화는 말뭉치와 대상용어가 주어졌을 때, 주어진 말뭉치에서 대상용어의 좋은 용례들을 자동적으로 선별하는 작업이라고 할 수 있다. 이 때, 어떤 용례를 좋은 용례로 볼 것인지에 대해서는 용례 추출 및 순위화의 목적과 대상용어의 종류를 고려해야 한다.

용례 추출 및 순위화의 목적은 크게 직접적인 목적과 간접적인 목적으로 나누어 볼 수 있다. 우선, 직접적인 목적은 추출된 용례를 통해서 대상용어의 의미와 쓰임새를 파악하는 것으로서, 이것은 용례 본연의 목적이라고 할 수 있다. 정의문이 기존 개념인 상위어와 의미 특질소를 통해 대상용어의 의미를 연역적으로 파악하게끔 하는 것과 다르게, 용례는 대상용어가 실제로 어떤 문맥 하에서 어떤 단어들과 함께 쓰이는지를 통해 대상용어의 의미와 쓰임새를 귀납적으로 파악할 수 있게 한다.

그리고 간접적인 목적은 추출된 용례를 자원으로서 사용하는 것이다. 용례는 사전이나 시소러스, 온톨로지 등의 구축에 정의문과 함께 가장 기초가 되는 자원이라 할 수 있다.. 특히 신조어나 전문용어의 경우 정의문이 없는 경우가 많기 때문에, 좋은 용례를 선별해서 구축하는 것이 더욱 중요하다. 이외에도 좋은 용례는 정의문 추출이나 인과관계 추출, 단어 의미구분과 같은 자연어처리 응용 연구에도 좋은 자원이 될 수 있다.

한편, 용례 추출 및 순위화의 대상용어는 크게 일반용어와 전문용어로 나눌 수 있다.. 일반용어는 일상 생활에서 많이 쓰이는 어휘로서, 일반적으로 복잡하지 않은 의미를 나타내지만, 의미가 비슷한 유의어를 가지는 경우가 많다. 그리고 대부분의 일반용어는 이미 사전에 등재되어 있어서 정의문과 유의어, 용례 등을 쉽게 얻을 수 있다.. 또한 일반용어의 용례는 일반 소설, 수필, 교과서 등에서 추출되기 때문에 비교적 쉬운 문장구조와 어휘들로 구성된다.. 따라서 일반용어의 용례는 대상용어의 의미를 드러내는 것도 중요하지만 쓰임새를 드러내는 것이 더욱 중요하다. 특히 비슷한 의미를 가지는 유의어들 사이에서의 미묘한 의미적 차이를 분명히 드러낼수록 좋은 용례라고 생각할 수 있다.

일반용어의 용례 추출 및 순위화는 (김장희,

2005)에서 연구되었는데, 전술한 바와 같이 일반용어는 말뭉치도 다양하고, 사전과 시소러스 등과 같은 자원도 풍부하다. 따라서 기존의 연구는 자원을 충분히 활용하여 용례들을 순위화 하였다.

전문용어는 일반용어와 대립되는 전문분야의 어휘로서, 복잡하고 전문적인 의미를 나타내는 경우가 많다. 일반용어와는 다르게 사전에 등재되어 있지 않은 용어가 많기에 정의문, 유의어, 용례 등을 얻기가 쉽지 않다. 또한 전문용어는 전문분야에서만 주로 사용되는 어휘이기 때문에 전문서적이나 특허, 논문 등에서 용례를 추출할 수 있고, 복잡한 문장구조와 어휘를 가지는 경우가 많다. 전문용어는 전문분야의 의미전달에 핵심적인 역할을 하지만, 사전에 등재되어 있지도 않고 널리 쓰이지도 않기에 그 의미를 알기가 어려운 경우가 많다. 따라서 전문용어의 용례는 대상용어의 의미를 드러내는 것이 가장 중요하다고 볼 수 있다.

본 연구는 이와 같은 전문용어의 용례 추출 및 순위화에 관해서 다루고 있다. 전문용어의 용례 추출 및 순위화는 일반용어의 용례추출 및 순위화와 다르게 대상용어의 의미를 드러내는 것이 가장 중요하다. 또한 사전과 시소러스 등의 자원이 부족한 전문용어의 특징도 고려되어야 한다.

따라서 본 연구는 전문용어의 의미를 잘 드러낼 수 있는 용례를 좋은 용례로 가정하고, 이것을 효과적으로 표현하기 위해 벡터를 이용했다. 용례를 통해서 의미를 드러낼 수 있는 가장 효과적인 방법은 용례가 대상용어의 가장 전형적인 문맥을 포함하는 것이다. 이를 위해 전형적인 문맥을 선별하여 벡터화 하고, 그와 가장 유사한 용례를 찾는 접근 방법을 취했다. 그리고 자원이 부족한 전문용어의 특징을 고려하여 부족한 자원을 보완할 수 있는 대체적 자원으로서 웹을 이용하였다.

2. 관련 연구

용례를 자동으로 추출하기 위한 연구는 여러 곳에서 시도되어 왔다. 일반적인 용례 추출에 관한 연구는 임의적으로 조작된 용례가 아닌 실제 사용되는 용어의 사용 예를 추출하기 위한 연구로서, 말뭉치의 색인을 통한 빠른 용례 추출을 그 목적으로 한다. 대표적으로 영어를 위한 시스템인 Word Smith, Monoconc등이 개발되었으며, 한국어를 위한 시스템으로는 KCP, 글잡이 등이 개발되었다. 이들은 기본적으로 용어에 대하여 말뭉치를 색인하고 원하는 용어에대한 용례를 말뭉치로부터 빠르게 추출 해 내는 기능을 가지고 있다.

하지만 단순하게 모든 용례를 추출하기 때문에 너무 많은 용례가 추출된다는 단점이 있다. 이를 보완하기 위해 해당 용어의 주변 문맥에 대한 검색(KCP) 이나 패턴을통한 검색(글잡이) 등을 제공하기도 했지만, 여전히 문맥이나 패턴이 일치하는 용례를 찾는 단순한검색이라는 한계를 가지고 있었다.

한편, 좋은 용례에대한 고려는 (김장희, 2005)에서 최초로 이루어졌는데, 용례의 순위화를 위하여 좋은 용례의 기준을 전형성, 가독성, 차별성으로 정하고, 각각의 기준으로 용례를 정량화하여 순위화하는 시도를 하였다. 용례의 순위화 실험은 크게 일반용어와 전문용어에 대하여 이루어졌지만 전체적인 시스템의 설계가 일반용어를 기준으로 만들어져 전문용어에 그대로 적용하기에는 어려움이 있다. 또한 용례를 순위화하기 위해서는 크게 형태소 분석된 말뭉치와 사전, 그리고 시소러스 등의 자원을 필요로 했다.

각각의 기준을 좀 더 자세하게 살펴보면, 우선 전형성은 용례의 전형적인 정도를 나타내는 기준으로서 본 논문의 접근 방법과가장 유사한 기준이다. 전형성은 크게 공기정보와 사전 및 시소러스를 이용해 계산되었는데, 대상 용어의 용례에 등장하는 단어가 공기정보나 사전 혹은 시소러스에 나타나면 점수를 가산해주는 방법을 취했다. 전형성을 이용한 용례의 순위화는 가독성을 이용한 용례의 순위화에 비해 낮은 성능을 보였다. 또한 전형성 내부적으로도 사전과 시소러스에 정보를 이용한 순위화의 결과가 공기정보를 이용한 순위화의 결과보다 좋은 성능을 보였다.

두 번째 기준인 가독성은 용례의 읽기 쉬운 정도를 나타내는 기준으로서 용례에 가독성을 해치는 내포문 등의 문장구조나 사전에 등장하지 않는 단어가 등장하면 감점을 하는 방법을 사용했다. 가독성을 이용한 순위화는 일반용어를 대상으로 한 실험에서 가장 좋은 성능을 보였다.

마지막 기준인 차별성은 대상 용어의 용례가 대상 용어의 유사어와 얼마나 차별적인지를 나타내는 기준으로서 사전의 정의문을 이용해 유사어를 찾은 뒤 대상용어와 유사어에 공통적으로 존재하지 않는 공기 정보와 사전 정보를 이용해전형성과 마찬가지 방법으로 순위화했다.

(김장희, 2005)의 연구는 전술한 바와 같이 자원에 상당히 의존적이다. 실제로 시스템은 사전이 존재할 때에만 차별성 점수를 계산할 수 있으며, 전형성은 사전과 시소러스에 상당히 의존적이다. 이처럼 자원에 의존적인 용례 순위화는 자원이 충분한 일반용어의 용례 추출 및 순위화 에서는

큰 문제가 되지 않을 수 있지만 한정적인 자원을 가진 전문용어의 용례 추출 및 순위화에는 문제가 될 수 있다.

3. 접근 방법

3.1. 용례의 표현

전문용어 용례의 추출은 한정된 자원 하에서 좋은 용례를 찾는 문제이다. 이를 해결하기 위해서 문제를 작은 문제로 나누면 세 부분으로 나뉠 수 있는데, 우선 용례를 어떻게 정형화해서 표현할 것인지를 정해야 하는 문제가 있다. 그리고 어떤 용례가 좋은 용례인지를 정해야 하는 문제가 있고, 마지막으로 그러한 좋은 용례를 어떻게 찾을 것인지의 문제가 있다. 특히 이 문제들을 해결하려 할 때 제한된 자원이 충분히 고려되어야 한다.

본 연구에서는 용례를 정형적으로 나타내기 위해 용례를 벡터로 표현하는 방법을 사용했다. 특정 용어의 용례는 문맥을 통해서 용어의 의미를 드러낸다. 따라서 문맥에 등장하는 단어들을 벡터화해서 표현하면 용례를 정량적으로 잘 나타낼 수 있을 것이다. 본 연구에서는 이 벡터를 용례 벡터로 부르기로 한다.

두 번째로, 좋은 용례는 해당 용어의 의미를 가장 잘 드러내는 용례라고 볼 수 있다. 그리고 해당 용어의 의미를 가장 잘 드러내기 위해서는 그 용어의 용례 들 중에서 가장 전형적인 문맥을 지니고 있어야 한다. 따라서 좋은 용례는 용어의 가장 전형적인 문맥을 벡터화해야 하고 본 연구에서는 이것을 중심 벡터라 부르기로 한다.

이와 같이 용례와 좋은 용례를 각각 용례 벡터와 중심 벡터로 나타내면 좋은 용례를 찾는 문제는 주어진 대상 용어의 용례 벡터들 중에서 가장 전형적인 문맥과 유사한 벡터를 찾는 문제이고, 결국 용례 벡터와 중심 벡터간의 유사도 비교를 통해 해결 할 수 있다.

3.2. 용례 벡터

용례는 용례에 등장하는 공기 정보들을 축으로 삼고 각 공기 정보들의 빈도를 그 값으로 취함으로써 벡터화할 수 있다. 즉, 하나의 용례의 문맥을 그 용례에 함께 등장하는 공기 정보들로 간주한 것이다. 여기에서 공기 정보란 대상 용어와 함께 등장하는 단어를 의미하는데, 통사정보 만 을 담고 있는 기능어(조사나 어미) 정보를 제외하고, 의미정보를 가지고 있는 명사, 동사, 형용사를 의미한다.

이와 같이 용례를 벡터화 했을 때, 용례 벡터는 그 용례의 문맥을 잘 나타낼 수 있고, 용례를 정량적으로 표현할 수 있다. 용례 벡터는 수식 1과 같이 표현한다.

$$\overrightarrow{U_i} = (w_{i1}, w_{i2}, \ldots, w_{in})$$

수식 1 용례 벡터

수식 1에서 w_{ik}는 대상 용어 t의 i번째 용례 U_i 내에 등장하는 k번째 공기 정보의 빈도수이다.

한편 좋은 용례는 가장 전형적인 문맥을 벡터화해서 나타낼 수 있는데, 이를 위해 가장 전형적인 문맥을 나타낼 중심 단어들을 선별해야 한다. 중심 단어는 해당 용어의 용례 전체에서 용어와 함께 등장하는 공기 정보들 중 공기 정보의 중요도가 중요도 평균 보다 표준편차 이상 큰 공기 정보들만 선택하는 방법을 사용했다. 이 때 중요도는 다양한 방법을 사용할 수 있지만, 본 연구에서는 상호정보(Mutual Information)를 이용했다. 공기정보의 중요도는 대상 용어와 공기 정보에 따라 크게 달라질 수 있는데, 중요도가 평균보다 표준편차 이상 큰 공기 정보만을 중심 단어로 선별함으로써, 중요도만 정확히 평가된다면 중심 벡터는 대상 용어의 공기 정보 중 의미 있는 문맥만을 반영할 수 있다.

특히 용례의 순위화가 중심 벡터와의 유사도 비교를 통해 이루어지기 때문에, 이 중심 벡터를 어떻게 구성하느냐에 따라 순위화의 결과와 순위화의 의미가 크게 변할 수 있다. 중심 벡터는 수식 2와 같이 나타낼 수 있다.

$$\overrightarrow{C_t} = (cw_1, cw_2, \ldots, cw_n)$$

수식 2 중심 벡터

수식 2에서 cw_k는 대상 용어 t의 k번째 중심 단어의 가중치이다.

벡터로 정형화된 용례의 순위화는 용례 벡터와 중심 벡터간의 유사도 비교를 통해서 이루어진다. 즉, 중심 벡터와 유사도가 클수록 좋은 용례로 간주하는 것이다. 두 벡터간의 유사도 비교는 다양한 방법을 통해 구할 수 있지만, 본 연구에서는 코사인 유사도를 사용한다.

대상 용어 t의 중심 벡터가 $\overrightarrow{C_t} = (a_1, a_2, \ldots, a_n)$로 표 현되고, 용어 t의 i번째 용례 벡터가 $\overrightarrow{U_i} = (b_1, b_2, \ldots, b_n)$로 표현될 때, 두 벡터간의 코사인 유사도는 수식 3과 같다

$$sim(\overrightarrow{C_t}, \overrightarrow{U_i}) = \frac{\overrightarrow{C_t} \cdot \overrightarrow{U_i}}{\left|\overrightarrow{C_t}\right|\left|\overrightarrow{U_i}\right|} = \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \times \sqrt{\sum_{i=1}^{n} b_i^2}}$$

수식 3 코사인 유사도

코사인 유사도의 값은 0과 1사이로 수렴하는데, 두 벡터의 유사도 값이 1에 가까울수록 유사도가 큰 것이고, 유사도가 1인 경우는 두 벡터의 완전한 일치를, 0인 경우는 두 벡터간의 완전한 불일치를 의미한다.

3.3. 웹 자원

앞 절에서 언급한 바와 같이 공기 정보의 중요도는 상호 정보에 의해 평가될 수 있다. 하지만 이러한 통계적인 방법을 통해 중요도를 평가하는 데에는 상당 부분 한계가 있는데, 그것은 다음과 같은 것들로서 특히 전문용어의 특성에 의한 부분이 크다.

우선, 자료 부족 문제가 있다. 앞서 언급했듯이 공기 빈도가 낮은 단어는 일반적으로 중요도가 낮은 공기 정보이다. 하지만 그 중에서도 공기 비율이 높은 단어는 중요한 공기 정보이기 때문에 높은 중요도 점수를 받게 된다. 하지만 그럼에도 불구하고 공기 빈도가 너무 낮은 단어들은 중요도 점수를 결정하는데 문제가 생긴다. 예를 들어 아주 작은 공기 빈도를 가지지만 원래 거의 등장하지 않는 단어는 상대적으로 너무 높은 상호 정보 값을 가지게 된다. 이것은 상호 정보 방법의 대표적인 문제로 널리 알려져 있다(Manning and Schutze, 1999).

또한 자료 왜곡 문제도 존재한다. 전문분야에는 많은 전문용어가 존재하고 또한 계속해서 새롭게 등장하기 때문에 같은 의미를 가지는 용어라도 다르게 쓰이는 경우가 많다. 실제로 하나의 특정 문서 안에서 언제나 대상 용어와 함께 나타나는 공기 정보들이 존재하지만, 많은 경우 그 공기 정보는 그 문서 안에서만 사용되고 있어서 실제로는 대상 용어의 전형적인 문맥과 거리가 먼 경우가 많았다.

사실 이와 같은 자료 부족이나 자료 왜곡의 문제들은 어떠한 곳에서도 발생할 수 있는 문제이지만, 전문용어의 용례 추출에 있어서는 그 문제가 더욱 크다. 왜냐하면 전문 용어의 경우 사전과 시소러스 등의 자원이 거의 없기 때문에 보완적으로 사용할 수 있는 자원이 없다. 뿐만 아니라 전문 용어는 전문 분야 내에서도 그 빈도수가 낮은 편에 속해서 그나마 사용할 수 있는 자원인 공기 정보 조 차도 적을 뿐만 아니라 추출된 공기정보의 빈도수도 낮아서 통계적 유용성을 분별하기 어렵다.

실제로 주어진 자원 하에서 통계적 정보만을 이용해 공기정보의 중요도를 평가했을 때, 높은 중요도를 가진 많은 공기 정보들이 특정 문서에서만 많이 등장하는 잘못된 공기정보였고, 반대로 낮은 중요도를 가진 공기 정보들 중에 중요한 공기정보가 포함되어 있는 경우가 많았다.

본 연구는 이와 같은 어려움을 극복하기 위해 보완적인 자원으로서 웹의 이용을 제안한다.

구체적으로 웹을 어떻게 자원으로 이용할지에 대해서는 많은 연구가 있어왔다. 본 연구에서는 웹에서 추출할 수 있는 여러 가지 정보 중 웹 문서의 빈도수 정보에 기반한 상호 정보를 이용한다. 웹 기반 상호 정보는 수식 4와 같이 구할 수 있는데, 식의 값을 구하는 데 기반이 되는 검색엔진은 Google을 사용한다.

$$\mathrm{WMI}(t, w_i) = \log_2 N \frac{hits(t, w_i)}{hits(t) \cdot hits(w_i)}$$

수식 4 웹 기반 상호정보

수식 4에서 t는 대상 용어, w는 공기 단어이고 N은 Google에 의해 인덱스 된 전체 문서의 개수이다. 그리고 hits(t,w)는 t와 w를 함께 Google의 쿼리로 주었을 때 반환된 문서 개수이다. 마찬가지로 hits(t)와 hits(w)는 각각 t와 w를 Google의 쿼리로 주었을 때 반환된 문서 개수이다.

수식 4는 (Turney, 2001)에 의해 처음 도입된 것으로서 그 이후 (Baroni, 2003) 등에 의해 많이 이용되어 왔고, 여러 가지 응용 연구를 통해 말뭉치에서 추출한 상호 정보보다 좋은 성능을 보인다는 것이 입증되어 왔다. 특히 상호 정보는 빈도수가 너무 작을 때는 과잉 추정되는 경향이 있는데 (Manning and Schutze, 1999), 많은 양의 웹 문서 수는 상호 정보의 이러한 문제를 극복할 수 있다.

4. 실험 및 평가

4.1. 실험 환경

실험에서 사용한 말뭉치는 전산학 분야의 특허 문서 300개 문서를 대상으로 하였다. 말뭉치는 총 19,821개의 단어와29,533개의 문장으로 구성되어 있다. 용례의 추출 및 순위화에 사용한 공기 정보는 말뭉치에서 추출하였다.

대상 용어는 출현빈도가 20회 이상 150회 미

만인 단어들 중 적어도 20개 이상의 문장에서 나타나는 용어 38개를 선택하였다. 대상 용어의 예는 표 1과 같다.

표 1실험 대상 용어의 예

대상용어	출현 빈도	대상 용어	출현 빈도
SGML	34	엔트로피	33
온톨로지	39	전 자 상 거래	49
디코더	105	불용어	108

좋은 용례의 정답 집합을 구축하는 것은 많은 시간과 비용을 필요로 한다. 특히 좋은 용례를 선택하는 작업은 그것이 용례 인지 아닌지를 결정하는 문제가 아니라 그 용례가 다른 용례들보다 좋은가 혹은 나쁜가의 문제이기 때문에 상당히 주관적이 될 수 있다. 실제 정답 집합의 구축과정에서도 평가자에 따라 좋은 용례를 선정함에 있어 다른 경향이 나타났다. 또한 전문용어의 좋은 용례를 선별하기 위해서는 전문 용어의 분야에 관한 지식도 필요하다.

실험에서 사용된 정답 집합의 구축은 전산학과 석사과정과 박사과정에 재학중인 대학원생들에 의해 이루어졌으며, 3인의 평가자 중 2인 이상이 좋은 용례라고 선택한 용례를 정답으로 간주하였다.. 평가 방법은 자연언어처리 분야에 널리 사용되는 정확률, 재현율, F-measure를 사용했다.

4.2. 실험 결과

4.2.1. 벡터화의 효과

벡터화의 효과를 평가하기 위해서 특허분야 문서를 대상으로 제안하는 시스템의 성능을 (김장희, 2005)의 시스템과 비교하였다. 특히 이 실험에서 제안하는 시스템은 웹 자원 등을 이용하지 않고 벡터화만을 사용하였다. 그리고 비교 대상인 시스템은 각각 전형성을 이용한 순위화 결과와 가독성을 이용한 순위화를 수행하였다.. 직접적인 비교대상은 전형성을 이용한 순위화이며, 실험 결과는 표 2와 같다.

표 2 기존 시스템과의 성능 비교

용례 추출 방법		정확률	재현율	F-measure
(김장희, 2005)	전형성	0.2684	0.3923	0.3188
	가독성	0.2316	0.3385	0.2750
제안 시스템		0.3026	0.4423	0.3594

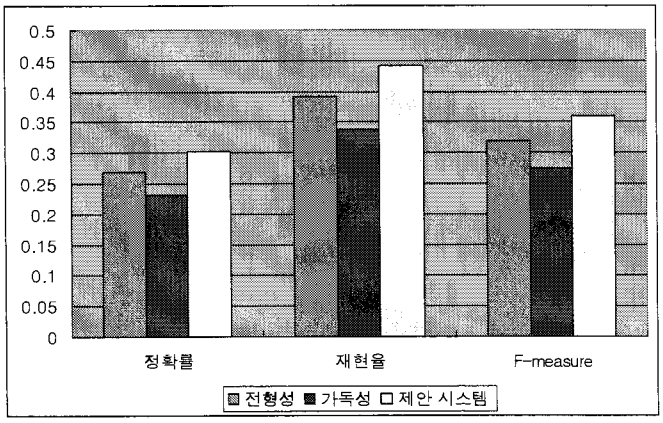

그림 1 기존 시스템과의 성능 비교

결과를 통해 알 수 있듯이 벡터화는 용례의 순위화에 좋 은 영향을 끼쳐서, 벡터화를 수행하지 않고 가산적인 점수를 부여한 기존 시스템의 결과보다 좋은 성능을 보였다.. 이와 같은 결과는 실험 대상 문서와 대상 용어가 전문 분야의 전문 용어이기 때문인 것으로 파악된다.

특히 이것은 기존 시스템의 전형성을 이용한 결과와 가독성을 이용한 결과의 성능 차이에서 뚜렷하게 드러나는데, 일반 분야에서 일반 용어를 대상으로 실험을 하였을 경우, 가독성을 이용한 순위화가 전형성을 이용한 순위화보다 좋은 성능을 보였던 것과 반대로(김장희, 2005), 이 실험에서는 기존 시스템을 이용한 결과에서도 전형성을 이용한 순위화가 가독성을 이용한 순위화 보다 좋은 성능을 보인다.

이와 같은 결과는 전문 용어에 있어서 좋은 용례란 용례의 가독성 보다 용례의 전형성을 통한 의미 파악이 더 중요하다는 것을 말해준다. 특히 제안하는 시스템이 가장 좋은 결과를 보이는 것은 제안하는 시스템에서 사용한 벡터화를 통한 방법이 이 같은 전형성을 기존의 방법보다 잘 반영한다는 것을 의미한다.

4.2.2. 웹 자원의 효과

웹 자원의 효과를 평가하기 위해 웹 자원을 사용한 경우와 사용하지 않은 경우를 비교 평가 하였다.. 웹 자원을 이용하는 구체적인 과정은 다음과 같다.. 우선 실험 문서 집합에서 대상 용어의 공기 정보를 추출한다. 추출한 공기정보 중 의미있는 공기 정보를 선택하기 위해, 추출한 공기 정보가 웹에서 사용되는 빈도와 추출한 공기 정보와 대상 용어 간의 공기 빈도를 Google을 이용하

여 추정한다.

이전 실험과 마찬가지로 웹에서 추정된 빈도 정보 역시 다양한 방법으로 평가 될 수 있고 각 방법에 따른 정확률을 웹 자원을 사용하지 않은 결과와 비교해서 표 3에 나타냈다.

표 1 웹 자원의 효과 평가

	공기빈도	상호정보	가중상호정보
말뭉치만 이용	0.2921	0.2895	0.3026
웹 자원 이용	0.3447	0.3526	0.3342

실험 결과를 통해 웹의 빈도 정보를 사용한 시스템의 정확률이 말뭉치의 빈도 정보만을 사용한 시스템의 정확률에 비해 약 5%가량 높은 것을 확인할 수 있다. 특히 웹 자원을 이용한 시스템의 성능은 어떤 방법을 사용해서 공기정보를 선택하더라도 말뭉치만 이용한 시스템의 성능보다 좋았다. 이것은 웹의 자원으로서의 유용성을 뒷받침해 줄 수 있는 결과로서, 웹의 정보는 통제할 수 없기에 정확하게 원하는 의미 혹은 원하는 분야의 정보만을 추출할 수는 없지만, 그러한 정보라도 용어의 전형적인 특징을 파악하는데 큰 도움이 된다는 것을 의미한다. 즉, 통제가 안되기 때문에 잘못된 정보를 가져올 가능성도 있지만, 의미 있는 정보를 가져올 가능성이 더 크기에 전체적으로는 시스템의 성능에 긍정적인 영향을 끼치는 것이다.

한편 웹 자원을 이용한 시스템은 앞 절에서 다뤘던 말뭉치만을 이용한 시스템의 결과와는 반대로 상호 정보를 사용해서 공기 정보를 선택했을 때의 성능이 가장 좋았다.. 이것은 웹 자원의 크기에 기인한 효과로서, 빈도수 정보가 부족할 때는 상호정보가 정보를 왜곡시킬 수 있지만 빈도수 정보가 충분할 때는 효과적인 공기 정보의 선택방법이 될 수 있음을 의미한다. 웹의 빈도 정보는 그 값이 충분히 크기 때문에 상호 정보가 좋은 성능을 보이는 것이다.

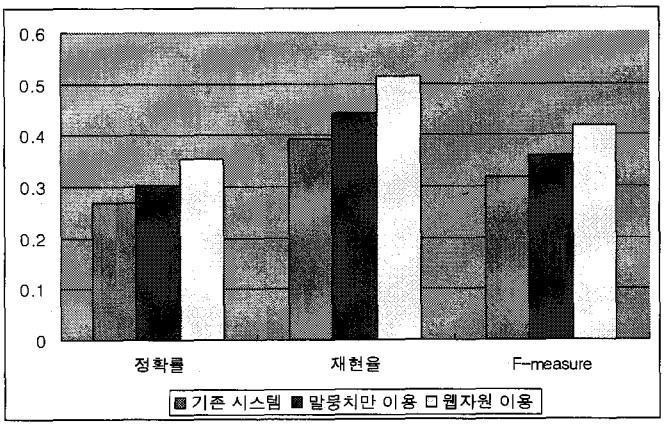

그림 2 각 시스템 성능 비교

마지막으로 (김장희, 2005)의 기존 시스템과 말뭉치만을 이용한 제안 시스템, 그리고 웹 자원을 이용한 제안 시스템간의 성능 비교를 그림 2에 나타내었다. 이것은 각 시스템의 가장 좋은 성능을 나타낸 것으로서, 웹 자원을 이용한 제안 시스템이 기존 시스템에 비해 30% 정도의 성능이 향상되었음을 확인할 수 있다.

5. 결론 및 향후 연구

본 논문에서는 용례를 벡터화하고 웹을 자원으로 이용하는 용례 추출 및 순위화 시스템을 구축하고, 한국어 특허문서에 적용해 보았다.

전문용어의 용례와 좋은 용례를 정량적으로 나타내기 위해 용례의 공기 정보를 이용하여 각각을 용례 벡터와 중심 벡터로 표현하는 방법을 제시하였다. 이 때, 벡터를 구성하는 공기 정보를 효과적으로 추출하기 위해 추출 범위와 품사를 제한하였으며, 상호정보로 공기 정보의 중요도를 평가하였다.

한편, 전문용어는 자원이 한정되어 있고, 말뭉치에서 출현 빈도가 낮아 공기 정보도 부족하다는 특징이 있는 데, 이로 인해 발생하는 자료 부족과 자료 왜곡 문제를 해결하기 위해 보완적인 자원으로 웹의 빈도 정보를 이용하는 방법을 제안하였다..

전산학 분야의 특허문서에 실험한 결과, 벡터화의 효과로 (김장희, 2005)의 기존 시스템에 비해 10% 가량의 성능을 향상 시킬 수 있었고, 최종적으로 웹 자원까지 이용하여 30% 가량의 성능을 향상시킬 수 있었다.

본 연구의 결과로 전문용어 용례의 특징이 일반용어 용례의 특징과 다르다는 것과 벡터화를 통한 좀 더 정확한 정량화가 가능하다는 것을 알 수 있었다. 또한 이 같은 벡터화를 위해서는 의미있는 공기정보의 추출이 중요하고 언제나 같은 중요도 평가 방법이 좋을 수 없다는 것을 알 수

있었다. 마지막으로, 자료 부족 문제를 보완하는 대체 자원으로서 웹의 효과를 확인할 수 있었다. 향후 연구로, 좀 더 정확한 순위화를 위해 공기 정보뿐만이 아닌 구조적 정보와 같은 자질을 사용한 용례 벡터의 구성에 대한 연구가 필요하다.. 또한 공기 빈도 자원으로서가 아닌 용례의 말뭉치로서 웹의 활용에 대한 연구와 순위화된 용례를 이용하는 응용에 대한 연구가 필요하다.

참고 문헌

김장희. 2005. 단어의 의미와 쓰임새를 위한 한국어 용례 추출 시스템. 한국과학기술원 석사학위논문.

김재호, 배선미, 최기선. 2004. 의학 전문용어의 정의문 자동 추출. 제 31회 정보과학회 학술대회..

신사임. 2002. 문맥 벡터와 상의어 정보를 이용한 한국어 명사의 의미구분. 한국과학기술원 석사학위논문..

최기선, 송영빈 외. 2000. 전문용어연구1. 홍릉과학출판사

한글학회. 1997. 우리말 큰사전.. 어문각.

C. Manning and H. Schutze. 1999. "Foundations of Statistical Natural Language Processing" MIT Press.

F. Keller and M. Lapata . 2003. "Using the web to obtain frequencies for unseen bigrams" Computational Linguistics 2003

F. Smadja. 1993. "Retrieving Collocations from Text: Xtract" Computational Linguistics 19(1).

Hang C., Min-Yen K., Tat-Seng C., Jing X. 2005. "A Comparative Study on Sentence Retrieval for Definitional Question Answering"

M. Baroni and S. Vegnaduzzo. 2004. "Identifying Subjective Adjectives through Web-based Mutual Information" KONVENS 2004

M. Baroni and S. Bisi. 2004. "Using cooccurrence statistics and the web to discover synonyms in a technical language"

P. Turney. 2001. "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL" ECML 2001, pp491-502