인지과학, 제 15 권 제 2 호, 2004년

LSA 모형에서 다의어 의미의 표상*

Representation of ambiguous word in Latent Semantic Analysis

이 태 헌** 김 청 택***
(Tae-hun Lee) (Cheongtag Kim)

요 약 잠재의미분석은 단어 의미를 동일한 맥락 (문장/문서) 하에서 동시에 제시되는 단어들의 공기성(co-occurence)으로 정의한다. 이 분석에서 한 단어는 맥락들을 대표하는 축들로 구성된 다차원 상의 한 점으로 표상되며, 단어 의미는 각 단어가 맥락 속에서 등장한 빈도로 정의된다. 이 다차원 의미공간은 SVD를 통하여 차원이 축소되어 추상된 의미를 표상한다. 이 연구는 다의어의 표상이 가능하도록 LSA를 발전시켰다. 제안된 LSA는 축에 대한 해석이 가능하도록 축의 회전을 도입하였으며 다의어 표상을 가능하게 하였다. 시뮬레이션에서는, 먼저 LSA에 의해 산출된 단어-맥락 빈도표에서 다의어를 포함하고 있는 문서들만을 재수집한 다음 문서들을 다의어 의미별로 분류하였다. 두 번째 단계에서는 다의어의 특정의미에 대한 표상을 분류된 단어-맥락 빈도표에서 비해당 의미에 대한 맥락들을 제거한 후 LSA를 적용하여 구성하였다. 시뮬레이션 결과는 다의어의 의미들을 LSA가 표상할 수 있음을 보여주었다. 이는 축회전을 포함한 LSA가 다의어 다중의미를 표상할 수 있고 실용적인 측면에서 웹검색 엔진에도 적용될 수 있음을 시사한다.

주제어 잠재의미분석, 축회전, 다의어

Abstract Latent Semantic Analysis (LSA Landauer & Dumais, 1997) is a technique to represent the meanings of words using co-occurrence information of words appearing in the same context, which is usually a sentence or a document. In LSA, a word is represented as a point in multidimensional space where each axis represents a context, and a word's meaning is determined by its frequency in each context. The space is reduced by singular value decomposition (SVD). The present study elaborates upon LSA for use of representation of ambiguous words. The proposed LSA applies rotation of axes in the document space which makes possible to interpret the meaning of axes. A simulation study was conducted to illustrate the performance of LSA in representation of ambiguous words. In the simulation, first, the texts which contain an ambiguous word were extracted and LSA with rotation was performed. By comparing loading matrix, we categorized the texts according to meanings. The first meaning of an ambiguous word was represented by LSA with the matrix excluding the vectors for the other meanings. The other meanings were also represented in the same way. The simulation showed that this way of representation of an ambiguous word can identify the meanings of the word. This result suggest that LSA with axis rotation can be applied to representation of ambiguous words. We discussed that the use of rotation makes it possible to represent multiple meanings of ambiguous words, and this technique can be applied in the area of web searching.

Keywords Latent Semantic Analysis, Rotation, Ambiguous word

언어자극의 의미는 주어지는 맥락(context)에 따라 달라진다. 예컨대 "신문이 망했다"라는 문장과 "신문이 젖었다"라는 문장에서 신문은 서로 다른 의미를 참조한다. 이와 같이 단어의 의미는 맥락에 따라 결정되며 단어가 개별적으로 존재할 때는 그 의미의 진위를 판단할 수 없으며 화자의 의도도 판단할 수 없다는 데는 이론이 없다. 그러나 맥락은 언어현상이나 심리현상을 설명하기 위해 빈번하게 사용되지만 (예, Simpson, 1981; Seidenberg, Tanenhuas, Leiman, & Bienkowski, 1982; Tobossi, 1991; Labov, 1973) 맥락 자체에 대한 조작적 정의나 연구들을 찾기는 쉽지 않다. 대체적으로 맥락은 어떤 현상을 설명하기 위한 설명개념으로 사용되어 왔지만 맥락 그 자체를 설명하려는 시도는 많지 않았다. 본 논문에서는 맥락을 특정 자극(단어)과

* 논문은 2002년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2002-074-HS1002)
** 서울대학교 심리학과
*** 서울대학교 심리학과, 서울대학교 인지과학협동과정
교신저자 : 김청택, ctkim@snu.ac.kr
서울시 관악구 서울대학교 심리학과

수반하여 발생하는 자극 집합이라고 조작적으로 정의하고 이렇게 정의된 맥락이 언어현상을 설명할 수 있는지를 연구하고자 한다.

언어의 의미가 수반성에 의해 결정된다는 주장은 비록 잠정적이고 불완전하지만 언어 습득의 과정을 살펴보면 그리 설득력이 없지는 않다. 우선 Quine(1960)의 gavagai 문제를 생각해 보자. 토끼가 달려가는 장면을 어떤 아이가 보고 있는데 옆에 있던 어른이 "gavagai!"라고 외쳤다. 이 아이는 gavagai가 무엇을 의미한다고 생각하겠는가? 토끼의 귀, 하얀색, 달리기, 혹은 그 상황에서의 어떤 다른 것? 무수히 많은 가능성이 존재할 것이다. 하지만 어른이 다른 맥락에서 'gavagai'라는 단어를 사용할 때 항상 토끼가 존재하며, 토끼가 없을 때에는 이 말을 사용하지 않는다면 문제는 달라진다. 아이는 gavagai가 토끼와 관련된 단어임을 알아 챌 수 있는 것이다. 위의 예는 단어의 의미를 파악하는데 있어서 맥락이 매우 중요한 역할을 한다는 철학적 의미론과 맥을 같이하며, 언어 습득 역시 비슷한 과정을 거쳐 이루어 질 수 있다고 가정해 볼 수 있다. 단어와 맥락의 수반성에 근거하여 초보적인 언어 습득이 이루어 질 수 있다는 가정에 대한 검토는 인지과정을 이해하는데 중요한 의미를 지닐 것이다.

본 논문에서는 단어와 맥락의 수반성이 곧 그 단어의 용례(usage)로 정의된다고 가정한다. 즉 특정 단어가 어떤 맥락에서 사용되고 어떤 맥락에서는 사용되지 않는지에 대한 정보가 새로운 언어를 습득할 때 매우 중요하다는 것이다. 본 연구에서는 이러한 수반성을 구현하기 위하여 잠재의미분석(Latent Semantic Analysis, LSA)을 응용하였다. 단어를 조작적으로 정의된 맥락 즉 단어의 용례에 의해서 표상하고, 단어의 용례에 대한 정보를 이용하여 의미 표상을 모사할 수 있는지를 살펴보고자 하였다.

잠재의미분석

잠재의미분석에서는 단어의 의미가 각 맥락(문장) 속에서 제시된 빈도 집합으로 정의된다. 한 단어는 벡터로 표상되며 벡터의 각 구성분자는 맥락(문장)을 대표하고 개별 단어가 각 맥락에서 출현한 빈도로 기재된다. 예컨대 12개의 단어들이 9개의 맥락(문장) 속에서 제시되는 상황을 고려해 보자. 이때 A라는 단어가 각각의 문장에서 0, 1, 1, 2, 0, 0, 0, 0, 0 번 제시되었다면 A라는 단어는 9차원 공간상에서 (0,1,1,2,0,0,0,0,0)의 벡터로, 즉 하나의 점으로 표상된다. 12개의 모든 단어를 차례로 쌓아서 하나의 행렬로 작성하면 아래와 같은 행렬 X가 구성된다.

$X =$

	C1	C2	C3	C4	C5	C6	C7	C8	C9
W1	1	0	0	1	0	0	0	0	0
W2	1	0	1	0	0	0	0	0	0
W3	1	1	0	0	0	0	0	0	0
W4	0	1	1	0	1	0	0	0	0
W5	0	1	1	2	0	0	0	0	0
W6	0	1	0	0	1	0	0	0	0
W7	0	1	0	0	1	0	0	0	0
W8	0	0	1	1	0	0	0	0	0
W9	0	1	0	0	0	0	0	0	1
W10	0	0	0	0	0	1	1	1	0
W11	0	0	0	0	0	0	1	1	1
W12	0	0	0	0	0	0	0	1	1

행렬 X는 우리가 경험한 모든 언어 사례들의 집합을 나타낸다. 여기서 단어들 간의 유사성은 단어벡터들간의 코사인 혹은 상관계수로 정의된다. LSA에서는 경험한 언어자료 집합(X)을 그대로 사용하는 것이 아니라 추상화하여 사용한다. 즉 행렬 선형 대수학에서 잘 알려진 singular value decomposition (이하 SVD) 정리를 이용하여 9개의 차원으로 표상된 단어들의 공간을 예컨대 2개의 차원으로 축소하여 표상한다. 차원을 축소하면 축소하기 전에는 출현하지 않았던 의미속성들이 출현하게 만드는 장점을 지닌다. 예컨대 Landauer와 Dumais (1997)의 예에서 보면 human과 user라는 단어들은 X 행렬에서 살펴보면 한 번도 같은 맥락에서 출현한 적이 없으므로 일차 상관은 0이지만, 차원을 축소하면 두 단어는 높은 상관을 보인다. 원 차원에서 관찰되지 않았던 두 단어의 유사성이 축소차원에서 관찰되는 이유는 human, user가 각각과 동시에 출현한 다른 단어들을 매개로 한 고차적인 상관을 맺고 있는데 차원을 축소하는 것이 이러한 고차적인 상관을 탐지할 수 있게 하기 때문이다.

LSA 모형은 실제로 방대한 자료를 이용하여 인간의 인지 기능을 성공적으로 모방할 수 있음을 보여주었다. Landauer와 Dumais (1997)는 백과사전에 나와 있는 총 30,473개의 문장들과 적어도 두 개 이상의 문장들에서 출현한 60,768단어를 이용하여 300차원 상에서 LSA를 적용하여 동의어 테스트를 실시하였다. 그들은 TOEFL에서 사용한 문제들, 즉 하나의 단어에 대한 동의어를 네 개의 보기 중에서 고르는 문제들을 풀게 하였다. LSA에서는 주제어로 제시된 단어와 보기로 제시된 단어들간의 상관값을 축소된 차원상에서 계산하여 그 상관값을 가장 높게 만드

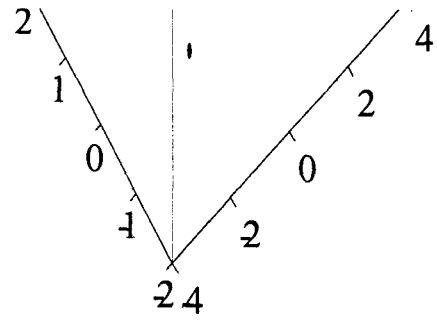
는 보기 단어를 정답으로 선택하는 절차를 이용하였다. 그 결과 총 80 문제 중 64.4%의 정답률을 보였다. 이는 영어를 모국어로 하지 않는 외국인이 미국 대학에 지원하기 위해 실시한 TOEFL 시험 자료에 축적된 동의어 테스트의 평균 정답률인 64.5%에 필적할 만한 결과였다.

LSA의 확장

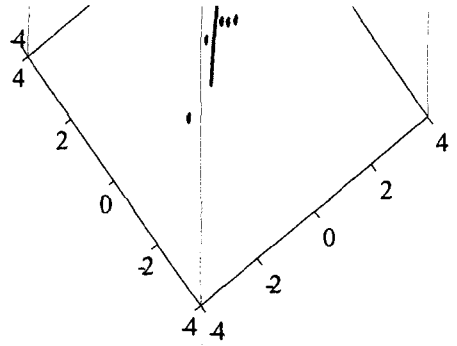
LSA 모형이 인간의 언어 습득 혹은 이해 기능을 모방하고 그에 필적하는 결과를 산출해 낼 수 있지만 LSA가 수학적이고 기계적인 처리과정을 통해 의미를 표상하기 때문에 LSA를 통하여 언어처리과정을 연구한다든지 LSA의 처리결과를 인지과학적으로 해석하는 것은 용이하지 않다. 예컨대 차원을 축소하는 과정이 실제 인간의 인지과정에서 이루어지는 것인지 알 수 없으며, 차원이 의미하는 바를 해석하는 것이 용이하지 않다. 따라서 단어 혹은 맥락이 지니고 있는 빈도들의 집합 즉 벡터의 의미가 파악되지 않는다. 이러한 점은 수행을 중요시하는 공학적 접근법에서는 문제가 되지 않지만 처리과정을 중요시하는 인지적 접근법에서는 만족스럽지 않다. 본 연구에서는 LSA를 확장하여 각 축의 의미 즉 차원의 의미를 해석할 수 있는 방법을 제안하고자 한다. 이를 위해 LSA 모형의 핵심인 SVD에 근거하여 차원을 축소한 다음, 단어의 다수 벡터성분이 0의 값을 가지도록 즉 단어가 소수의 축에 의해서만 기술될 수 있도록 축을 회전시켰다. 이러한 축회전은 축에 대한 해석을 가능하게 한다. 또한 이러한 기법을 이용하여 하나의 단어가 여러 의미를 지니는 다의어의 의미표상에 대한 하나의 모형을 제시하고자 하였다.

SVD에 따르면 모든 행렬은 세 개의 행렬로 분해할 수 있다 (Eckart & Young, 1936). 위의 예에서 X라는 행렬은 U, D, V라는 3개의 새로운 행렬로 분해될 수 있다. 이렇게 분해된 행렬의 곱을 원래 행렬의 기본 구조(basic structure)라 부른다. 이 기본 구조에 대한 이해가 LSA의 이해의 근간이 된다. 우선 행을 중심으로 기본 구조를 살펴보자. X 행렬은 총 12개의 행으로 이루어져 있으며 각각은 9개의 원소들로 이루어져 있다. X 행렬은 총 12개의 행벡터를 지니고 있는 셈이 된다. 이 12개의 행벡터는 9차원 상에서 12개의 점들로 표현될 수가 있다. 앞에서 사용된 X 행렬은 9차원으로서 기하학적 제시가 불가능하기 때문에 이해를 돕기 위하여 가상적인 3차원 데이터를 이용한다.

세 개의 맥락(문장)에 50개의 단어들이 출현한 행렬 X를 가정하면 (그림 1)과 같이 이러한 맥락 혹은 단어들은 3차원 상에서 50 개의 점들로 표시된다. 이들은 세로로 곧추선 타원 모양으로 군집해 있으며 이는 곧 이 벡터들이 특정 패턴을 이루고 있음을 의미한다. (그림 1)에서 벡터들의 군집은 (1,0,0), (0,1,0), (0,0,1)의 세 단위 벡터에 의해 정의된 직교하는 세 축에 의하여 표상되고 각 점의 좌표값이 곧 행벡터이다. 그런데 이 벡터들의 군집을 (그림 1)에 나타난 세 개의 축을 기준으로만 표현해야할 이유는 없다. 만약 벡터들의 군집이 특정 패턴을 나타낼 경우 그 패턴을 가장 잘 기술해 주는 새로운 축을 생성한다면 자료의 해석이 더욱 쉬워질 수 있기 때문이다.



(그림 1) 50개의 단어들이 세 개의 맥락(차원) 상에서 표상된 예



(그림 2) 세 개의 축으로 재표상된 단어들

(그림 2)는 새로운 3개의 축을 표시하고 있다. 우선 세 개의 새로운 축 역시 직교하고 있음을 볼 수 있으며, 가장 굵은 선으로 표시된 제 1축이 자료의 분산을 가장 많이 설

명할 수 있다. 두 번째로 굵은 선으로 표시된 제 2축은 두 번째로 자료의 분산을 많이 설명하고 있으며, 마지막 제 3축은 세 번째로 분산을 많이 설명하고 있다. 축을 위와 같이 새롭게 변형시켰을 때의 가장 큰 장점은 자료의 차원을 축소할 수 있는 방안을 제공한다는 것이다. 만약 3차원을 2차원으로 축소하고자 한다면, 자료의 변산성을 가장 잘 설명하지 못하는 제 3축을 탈락시켜 2차원을 구성하면 된다. SVD의 맥락에서는 X라는 행렬이 U, D, V'이라는 3개의 행렬로 분해되는데 V의 행벡터가 X의 행벡터의 새로운 축[1]으로서 기능하게 되는 것이다. 그리고 U의 행벡터는 원자료의 행벡터를 새로운 축에서 계산된 새로운 좌표값으로 표현하게 되며, D행렬의 대각 원소들은 새로운 축에서의 좌표값들의 분산을 표현하게 된다. SVD에서 차원 축소하는 과정은 예컨대 3차원에서 2차원으로 축소하는 과정은 다음과 같다. U와 V에서 첫 번째와 두 번째 행만을 추출하여 U^*와 V^*를 구성하고 D에서 첫 번째와 두 번째 대각원소만을 추출하여 D^*를 구성한 다음, $U^*D^*V^{*'}$을 구하면 2차원으로 축소한 공간의 표상이 된다.

그렇다면 차원의 축소는 어떤 경우에 필요하며 그 의미는 무엇인가? 데이터가 (그림 2)에서와 같이 군집을 이루고 있다면 데이터의 산포가 가장 넓은 방향으로 그어져 있는 제 1축이 대부분의 정보를 담고 있게 된다. 만약 자료가 제 1축을 중심으로 더욱 응집되어 있다면 자료를 기술하는데 굳이 3개의 축을 사용해야할 이유가 줄어들게 된다. 극단적 예일 수도 있으나, 자료가 제 1 축을 중심으로 거의 벗어나 있지 않은 경우 원래의 3개의 차원에서는 드러나지 않던 선형적 패턴을 차원을 축소함으로써 탐지할 수 있게 되고 나머지 두 개의 축은 필요 없다는 것도 확인할 수 있게 된다. 이는 3차원일 경우 기하학적으로 확인이 가능할 뿐만 아니라 D행렬의 2개의 원소가 거의 0에 가깝다는 것, 즉 나머지 두 개의 새로운 축에서 데이터의 분산이 0에 가깝다는 것으로도 확인할 수 있다. 만약 원자료 행렬의 벡터들이 상호 관련이 없이 거의 구(sphere)에 가까운 형태로 분포하고 있다면 차원의 축소는 불가능하다. 그러나 이러한 원자료는 사실 존재하기 힘들며, 상호 관련이 전혀 없는 데이터를 분석하는 일은 거의 없으므로 우리의 관심사가 아니다.

앞에서 기술한 LSA 모형은 X행렬의 기본 구조 정보를 이용하여 차원을 적절히 축소하고 새로운 공간에서 단어를 표상한다. 이러한 차원의 축소를 통해 단어-맥락 수반성 행렬에서 드러나지 않았던 단어 벡터의 패턴을 새롭게 생성할 수 있고 이렇게 추상화된 단어들 간의 관계성을 탐지할 수 있다. 그러나 이러한 기법은 축에 대한 해석을 가능하지 않게 한다. 그 이유는 단어들의 위치를 고정시킨 다음 축을 회전시키더라도 단어들의 상대적인 위치는 변화되지 않을 것이고, 따라서 무한히 많은 축에 의해서 축소된 차원이 표상될 수 있기 때문이다. 수리적으로 표현하면 축소된 축 상에서 재생된 X를 $\hat{X}$이라 하면 $\hat{X}$은 $\hat{X} = UDV'$으로 분해될 수 있고 임의의 직교행렬 T에 대하여 ($TT'=I$)

$$\hat{X} = UDV' = ZV' = ZTT'V \quad 단\ ZV = UD$$

가 성립함으로 $Z^* = ZT$로 두고 $V^* = VT$로 두면 새로운 축 V^*은 원래 축인 V를 사용했을 때와 동일한 $\hat{X}$을 산출한다. 즉 SVD에서 동일한 $\hat{X}$을 산출하는 축들은 유일하지가 않고 무한히 많이 존재한다. LSA에서는 축의 정보를 사용하지 않고 차원 축소에 의해 산출된 $\hat{X}$만을 사용하기 때문에 축에 대한 해석은 관심의 대상이 아니었다.

본 연구에서는 축회전 기법을 적용하여 축을 해석하고자 하였다. 또한 이러한 축회전을 도입하여 다의어에 대한 표상을 시뮬레이션하였다. 축의 비결정성의 문제는 심리측정학의 요인분석기법에서 오랫동안 연구되었으며 여러 가지 기법이 개발되어 있다. Thurstone(1947)의 단순구조에서부터 발달된 회전기법은 목적에 따라 서로 다른 기준을 제안하였으나, 가장 중심이 되는 개념은 각 변수는 하나 혹은 소수의 요인들에만 높은 요인부하량을 보이면 다른 요인들과는 0에 가까운 요인부하량을 보이도록 변환 함수 T를 정하는 방식이다. 이런 방식으로 요인부하량을 결정하면 특정한 변수는 하나 혹은 소수의 요인들에 의해서만 예언되기 때문에 설명이 용이하다. 이러한 축회전 기법을 LSA에 적용시키면 변수는 단어로 대치되고 요인은 의미성분으로 대치된다. 즉 한 단어의 의미는 소수의 의미성분들의 선형함수에 의하여 결정된다. 따라서 축회전을 사용하여 새로운 방식으로 축 즉 의미성분을 구성하면 한 단어가 최소한의 의미성분에 의하여 설명되어 단어에 대한 의미의 분석에 도움이 된다. 회전하는 방식을 결정짓는데 있어서 필요한 또 하나의 가정은 요인(의미성분)들간의 상관을 가정하는지의 여부이다. 상관이 없다고 가정하면 직교회전이 되며 상관이 있다고 가정하면 사경회전이 된다. 자세한 회전에 대한 개관은 Browne(2001)을 참조하라.

시뮬레이션: LSA 모형을 응용한 다의어 의미 해소

기존의 LSA 모형은 추론 과정, 특히 동의어 판단, 에세이 평가 등의 영역에서 인간의 수행을 유사하게 모사할 수 있었다. 그러나 LSA 모형은 어떤 단어가 맥락에 따라 여

1) 수학적 용어로는 정규직교기저(orthonormal basis)라 한다.

러 가지 의미들을 지닐 수 있음에도 불구하고 특정 단어를 오직 하나의 벡터로만 표상함으로써 다의어 의미 해소라는 인간의 중요한 언어처리능력을 반영하지 못하는 한계가 있었다. 본 연구에서는 한 단어가 지니는 여러 가지 의미들을 상이한 벡터로 표상하도록 LSA 모형을 개선하고 궁극적으로는 맥락에 따른 다의어의 의미 해소를 시도하였다.

LSA 모형이 지니는 한계 즉, 한 단어는 오직 하나의 벡터로만 표상된다는 점을 극복하기 위해서는 단어가 지니는 다의적인 의미에 따라 각기 개별적인 벡터 표상이 가능하여야 한다. 이러한 단어의 다중벡터표상을 가능하도록 하기 위해서 본 연구에서는 행렬을 단순히 기본 구조를 분해하고 차원을 축소하여 재결합하는 LSA 모형을 확장하여 특정 단어가 지닌 의미의 수만큼 다중벡터표상을 형성하도록 하는 모형을 제안하였다. 다만 여기서는 의미(meaning)와 의의(sense)를 구별하지 않고 의미는 의의의 하위개념으로 보았다. 즉 다의어의 의미에 대한 표상과 서론 예에서 제시되는 신문에 대한 의미(의의)에 대한 표상 방식이 동일한 방식에 의하여 이루어진다는 것이다.

모형: 다의어를 하나의 표상으로 형성하지 않고 의미별로 형성하기 위하여 다음의 두 단계에 의하여 표상을 형성하였다. 첫 번째 단계에서는 전체 단어-문서 행렬에서 표적 다의어 단어를 포함하는 문서 벡터들만 추출하여 새로운 단어-문서 행렬 (X)을 구성한 뒤 기본 구조로 분해한다 ($X = UDV'$). 이때 표적 단어의 의미를 구분할 수 있도록 문서(맥락)를 분류하기 위해 U행렬에서 단어의 의미 수 (r)만큼의 열벡터를 추출하여 차원을 축소하였다. 이렇게 축소된 차원에서 r 차원의 의미를 구분하기 위하여 축을 회전하였다. 이는 수학적으로 다음과 같이 표현된다.

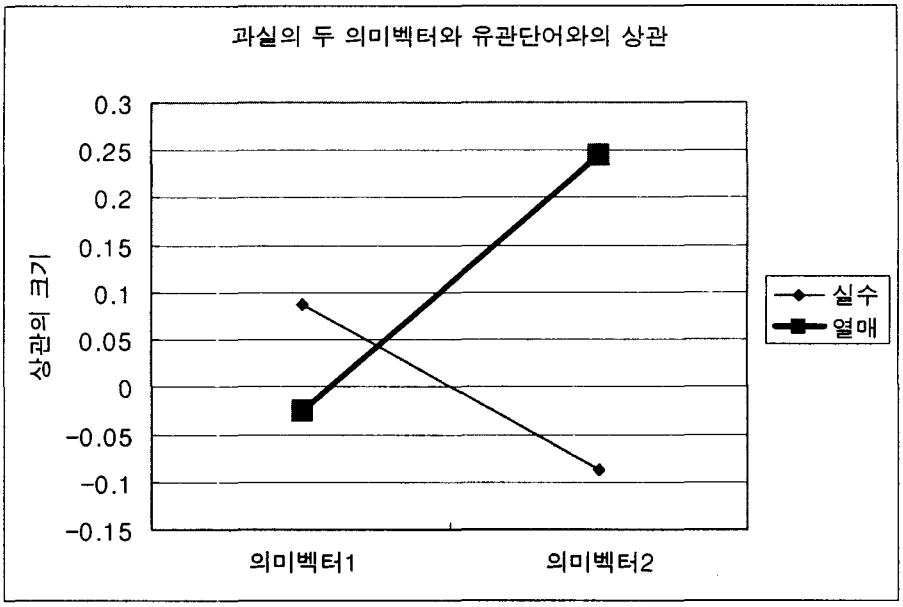
$$X = U_r D_r V_r' + \delta = (U_r T)(T^{-1} D_r V_r') + \delta$$

이 때 $T^{-1}D_rV_r'$ 행렬에는 문서들이 표적 단어의 의미별로 분류되어 있다 (김청택과 이태헌, 2002). 즉 첫 번째 행은 첫 번째 의미를 두 번째 행은 두 번째 의미를, r번째 행은 r번째 의미를 나타낸다. 각 행이 나타내는 의미는 해당 행에서 $T^{-1}D_rV_r'$ 의 부하량이 큰 문서를 추출하면 추론할 수 있다. 표 1은 이러한 문서들 중 가장 대표적인 문서를 나열하고 있다. 이때 회전을 담당하는 T행렬은 Jennrich(2001, 2002)의 Basic Singular Value(BSV) 알고리듬을 이용하여 결정하였다. BSV 알고리듬의 기본 논리는 $T^{-1}D_rV_r'$ 행렬의 행 또는 열의 분산을 최대화하여 각 행 또는 열이 가능한 적은 수의 영이 아닌 원소를 가지도록 함으로써 $T^{-1}D_rV_r'$ 행렬이 단순한 구조를 가지도록 하는 방식이다. 두 번째 단계에서는 다의어를 의미별로 표상을 형성하는 과정이다. 첫 번째 의미에 대한 표상은 다음과 같다. 전체 단어-문서 행렬에서 분류된 문서들 중 첫 번째 이외의 행에서 높은 값을 가지고 있고 첫 번째 행에서는 낮은 값을 가지고 있는 문서들을 제외한 뒤 새로운 행렬을 구성하고 이 행렬을 다시 기본구조로 분해하여 차원을 축소하고 재결합하였다. 이 행렬 상에서의 단어의 벡터는 다의어 의미의 첫 번째 의미에 대한 벡터 표상으로 간주할 수 있다. 그 나머지 의미도 위와 같은 동일한 과정을 거쳐 새로운 벡터를 형성함으로써 추출할 수 있다. 이 모형이 예언하는 바는 다의어가 이러한 방식으로 각 의미별로 개별적인 벡터로 표상된다는 것이다. 이를 검증할 수 있는 방법은 각 개별 벡터와 다의어의 의미들과 유관한 단어들의 연합정도를 산출하여 각 벡터들이 다의어의 하나의 의미성분과 높은 상관을 보이는지를 관찰하는 것이다.

자료: 본 연구에서 사용한 자료는 파스칼 전자백과사전에 있는 10,334개의 문서와 이들 문서에 등장하는 22,417개의 단어들을 수반표(contingency table) 형태로 구성한 22,417 × 10,334 행렬이었다. 이 수반표에서 각 단어와 수반한 문서빈도 분포에 대한 엔트로피를 계산하여 수반표의 빈도들을 각 단어에 해당하는 엔트로피로 나눈 값을 이후의 분석에 사용하였다. 이러한 변환을 한 이유는 기능어(functional words)에 대한 비중을 낮추고 단어빈도효과를 반영하기 위해서였다.

시뮬레이션 절차: 다의어 의미 표상 과정을 구체적으로 기술하면 다음과 같다. <표 1>에서 제시되어 있는 바와 같이 결정, 경기, 관리, 과실, 고문 등의 다의어가 사용되었다. 예컨대 과실이라는 단어의 의미를 해소하기 위해서 우선 단어-문서 행렬에서 과실이라는 단어가 포함된 열만을 추출하여 새로운 행렬을 구성하였다. 이렇게 구성된 행렬은 관심 다의어를 지닌 문서만을 포함한 행렬이 된다. 이러한 방식으로 새롭게 구성된 단어-행렬 문서를 주제별로 분류한다(김청택과 이태헌, 2002). 이 때 문서들은 과실이 두 개의 의미를 가지고 있기 때문에 두개의 범주로 분류된다. 하나의 범주에는 과실(過失)의 의미를 지닌 맥락들이 묶여져 있고, 다른 범주에는 과실(果實)의 의미를 지니는 맥락들이 묶여져 있다. 본 연구에서 분석된 다른 단어들 즉, 결정, 경기, 관리, 고문의 단어를 포함하는 문서들을 분류한 결과, 표 1에서 제시된 바와 같이 해당 다의어의 의미를 해소할 수 있는 맥락별로 문서들이 묶여져 있다.

다음 단계는 과실(過失)이라는 단어의 의미벡터와 과실(果實)이라는 의미벡터를 분리하는 것이다. 우선 과실(過失)이라는 의미 벡터를 생성하기 위해서 전체 단어-문서 행렬에서 첫 번째 단계에서 분류해 놓았던 문서들 중 두



(그림 3) 과실의 두 의미 벡터와 관련의미의 상관정도

번째 범주에 해당하는 문서들을 제외한 뒤 이 행렬을 다시 기본 구조로 분해하여 차원을 축소하고 재결합하였다. 이 때 과실이라는 단어의 행벡터는 과실(過失)의 의미를 지닌 벡터가 되는 것이다. 과실(果實) 의미벡터도 마찬가지로 전체 단어-문서 행렬에서 첫 번째 범주의 문서를 제외한 뒤 차원을 축소하여 재결합한 행렬의 결정 행벡터를 이용하여 표상하였다.

결과 및 논의

LSA에 의한 다의어 표상모형이 예언하는 바는 모형에서 제시된 방식으로 다의어에 대하여 의미별로 개별적인 표상이 형성된다는 것이다. 이를 검증하기 위하여 각 개별 벡터와 다의어가 지닐 수 있는 의미들과의 연합정도를 산출하여 각 벡터들이 다의어의 하나의 의미성분과 높은 상

<표 1> 각 의미성분을 대표하는 문장들

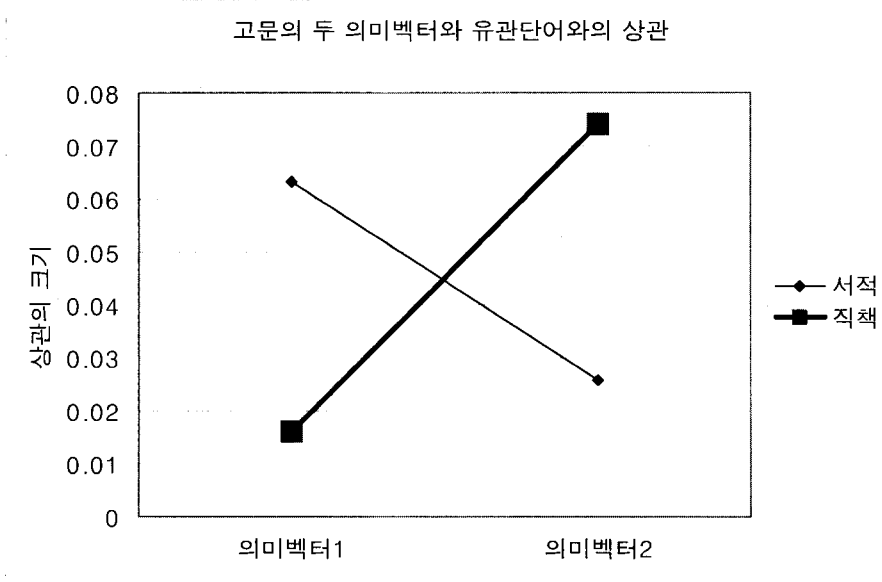
결정	決定	자기의 가능성을 반성적 자유에 기초해서 결정하는 사정을…
	結晶	냉각하면 결정화되기 쉽고 겨울철에는 얼어서 결정상태로 되기 때문에……
경기	競技	아시아 경기대회의 전신이라고 할 수 있는 …
	景氣	일정한 주기의 경기 순환이 반복됨을 통계적으로…
	京畿	경기 소리의 명창들이 …
관리	管理	국제경영으로서의 자본·인력·기술·관리 면에서의 …
	官吏	고려·조선 시대에 관리의 부지런함과 게으름 …
과실	過失	자기의 과실로 볼 수 없는 한 …
	果實	아행성이며 과실 외에 벌·개미·흰개미 등 곤충을 좋아하고 ……
고문	顧問	한성외교단 및 한국정부고문으로 중책을 맡았던 인물들에 ……
	古文	중국 한 나라 때 공자의 옛집 벽 속에서 나온 고문으로 된 경전 ……

관을 보이며 다른 성분과는 낮은 상관을 보이는지를 관찰하였다.

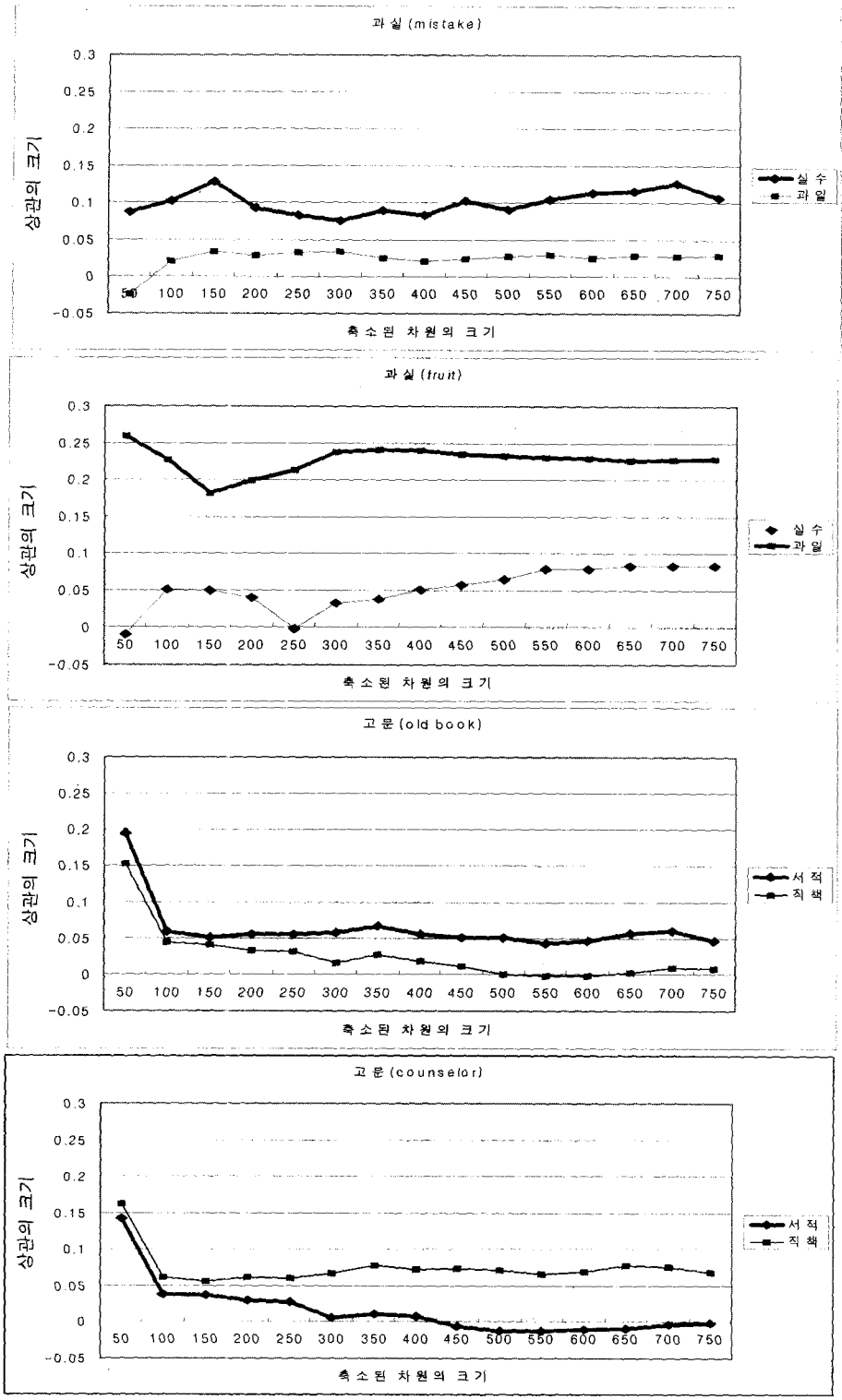
여기에서 제시된 LSA 방식에 의하여 의미별로 분리 표상된 두 개의 과실 벡터와 다의어와 유관한 의미를 지는 두 개의 단어 벡터 (열매 벡터와 실수(mistake) 벡터)의 상관을 구하였다. 첫 벡터는 실수 벡터와 높은 상관을 보였으나 열매 벡터와는 낮은 상관을 보였고, 두 번째 벡터는 열매 벡터와 높은 상관을 보였으나 실수 벡터와는 낮은 상관을 보였다 (그림 3). 이는 첫 번째 벡터는 과실(過失)이라는 의미를 표상하고 두 번째 벡터는 과실(果實)이라는 의미를 표상함을 의미한다. 다의어 고문에 대한 결과는 (그림 4)에 제시되어 있으며, 다른 다의어에 대한 결과도 동일한 패턴을 보였다. 이러한 결과는 LSA에 의하여 다의어에 대한 의미를 분리하여 표상할 수 있음을 시사한다.

<표 1>은 각 성분에서 부하량이 가장 높은, 즉 $T^{-1}D_rV_r$의 행렬에서 각 열에서 높은 값을 가지고 있는 문장들을 찾아서 나열한 것이다. 이러한 문장들에서 각 의미성분이 의미하는 바를 추론할 수 있는데 의미성분(요인)은 상관의 결과에서 예측된 바와 일치하였다.

(그림 5)는 다의어의 각 의미벡터와 관련 유사어와의 상관값이 축소한 차원의 수에 따라 변화하는 형태를 보여준다. 가장 두드러진 특징은 축소된 차원의 크기와 관계없이 다의어에 대한 의미표상들은 유관의미들과는 높은 상관을 가지며 다의어의 다른 의미들과는 낮은 상관을 가진다는 것이다. 이러한 결과는 시뮬레이션의 측면에서 차원의 수를 결정하는 것이 다의어 의미 표상에 결정적인 역할을 하는 것으로 보이지는 않음을 시사한다. 다만 단어들에 따라서 다의어에 대한 의미를 표상하는데 적절한 차원이 존재하는 것으로 보인다. (그림 5)에서 고문의 경우는 차원이 아주 낮으면 다의어 의미들 간의 변별력이 낮아지는 경향을 보이지만 과실의 경우는 차원이 낮으면 다의어 의미들 간의 변별력이 높아지는 경향을 보인다. 차원의 축소를 주어진 맥락을 추상화하는 과정으로 해석하면 이는 다의어의 의미가 구분되는 추상화의 차원에 따라 다를 수 있음을 의미한다. 유사한 의미를 구분하여 표상하기 위해서는 더 많은 차원이 필요할 것이고 아주 상이한 의미들을 구분하기 위해서는 적은 차원상의 표상이 충분할 수도 있다. 이러한 논의는 다의어의 연구에서 얻은 결과를 하나의 사전 항목(entry)에 여러 개의 의의(sense)가 포함된 경우로 확장하기 위해서도 필요하다. 차원을 많이 포함시켜 맥락에 의한 변별수준을 강화시키면 의의를 구분하여 표상하는 절차에도 다의어를 구분하는 절차를 적용시킬 수 있을 것이다. 그러나 이러한 논의는 현재 자료로는 증명될 수 없고, 보다 심층적인 연구가 필요할 것으로 보인다.



(그림 4) 고문의 두 의미 벡터와 관련의미의 상관정도



(그림 5) 차원에 따른 다의어 벡터들과 관련의미의 상관의 크기

이 연구에서는 LSA에서 다의어의 표상을 형성하는 모형을 제시하고 이 모형에 의하여 형성된 다의어 표상의 타당성을 검증하였다. 사실 모든 단어는 여러 의미로 구분되어 표상될 수 있다. 다의어 의미의 다중 표상뿐만 아니라 책과 같은 단어도 대학교라는 맥락에서 제시되었을 때와 책방이라는 맥락에서 제시될 때 다른 의의(sense)를 가질 수 있다. 이러한 다중 의미에 대한 표상방식으로 LSA에 의한 표상을 제안하였다. LSA에서는 단어의 의미를 그 단어와 함께 제시된 맥락에 의해 결정하였다. 즉, 단어들은 하나의 커다란 저장고에 각 맥락에 따른 수반성표의 형태로 저장되어 있고 주어진 맥락에 따라 적절한 수준의 추상화를 통하여 적절한 의미를 추출해 낸다는 것이다. 만약에 한 단어에 대한 세밀한 수준의 개념형성이 필요한 경우에는 많은 차원을 지닌 표상을 형성하고, 보다 상위의 개념을 형성하는 경우에는 적은 차원을 지닌 표상을 형성할 수 있을 것이다. 현 결과로는 이러한 주장을 직접적으로 지지하는 증거를 얻지 못하였지만 이러한 표상의 가능성을 간접적으로 시사하고 있다.

인간이 다의적 의미를 맥락에 따라 해소하는 과정에 대한 연구는 많으나 (예, Swinney, 1979; Simpson, & Kreuger, 1991), 연구의 결과물들이 형식화되어 시스템을 통해 구현하기에는 다소 부족한 점이 있다. 본 연구의 결과는 인간의 의미 해소 과정을 모사할 수 있는 모형을 개발했다는 데 그 의의가 있다. 즉 기존 LSA 모형에서는 불가능했던 의미별 다중 벡터 표상을 가능하도록 함으로써 시스템이 다의어를 단어별로 표상하지 않고 의미별로 표상할 수 있는 방안을 제시한 것이다. 이러한 결과는 정보 검색 시스템에 활용할 수 있다. 즉 다의어를 해소하는 과정에서 얻어진 문서 분류 결과는 해당 검색어를 포함하는 문서들을 검색어의 의미별로 분류하여 제공하는 검색 시스템에 적용될 수 있다. 이러한 시스템을 자동화하는데 걸림돌이 되는 것은 차원을 자동적으로 정할 수 있어야 한다는 것이다. 그러나 요인분석 등에서 많이 사용되는 고유치(eigen value)기준이나 스크리(scree) 검사를 이용하면 자동화될 수 있다. 다만 이러한 시스템은 새로운 정보가 입력될 때마다 새로운 표상을 형성해야 되므로 실시간으로 구현하는 것은 현 단계에서는 다소 무리가 있다. 증진적인(incremental) SVD등의 기법의 개발 등이 필요하다.

참고 문헌

김청택, 이태헌 (2002). 뇌와 인지모형: 잠재의미분석을 통한 문서분류. **한국심리학회지: 실험 및 인지, 14**, 309-319.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*, 111-150.

Eckart, C & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1*, 211-218.

Jennrich, R. I. (2001). A simple general procedure for orthogonal rotation. *Psychometrika, 66*, 289-306.

Jennrich, R. I. (2002). A simple general method for oblique rotation. *Psychometrika, 67, 7-19.*

Labov, W. (1973). The boundaries of words and their meanings. In C. J. Bailey and R. W. Shuy (eds.) *New ways of analyzing variation in English.* Washington, D. C.: Georgetown University Press.

Landauer, T. K. & Dumais S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Quine, W. V. O. (1960). *Word and object.* Cambridge: The MIT Press.

Seidenberg, M.S., Tanenhuas, M. K., Leiman, J. M. & Bienkowski,M. (1982). Automatic acces of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology, 14*, 489-537.

Simpson, G. B. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Verbal Learning and Verbal Behavior, 20*, 120-136.

Simpson, G. B., & Kreuger, M. A. (1991). Selective access of homograph meanings in sentence context. *Journal of Memory and Language, 30*, 627-643.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645-659.

Thurstone, L. L. (1947). *Multiple Factor Analysis.* Chicago: University of Chicago Press.

Tobossi, P. (1991). Understanding words in context. In G. B. Simpson(Ed.), *Understanding word and sentence.* Amsterdam: North-Holland.

접 수	2004년 5월 24일
게재승인	2004년 6월 4일