인지과학, 제23권 제3호
Korean Journal of Cognitive Science
2012, Vol. 23, No. 3, 295～322.

벡터 공간 모델과 HAL에 기초한 단어 의미 유사성 군집*

김 동 성†
고려대학교

본 연구에서는 벡터 공간 모델과 HAL (Hyperspace Analog to Language)을 적용해서 단어 의미 유사성을 군집한다. 일정한 크기의 문맥을 통해서 단어 간의 상관성을 측정하는 HAL을 도입하고(Lund and Burgess 1996), 상관성 측정에서 고빈도와 저빈도에 다르게 측정되는 왜곡을 줄이기 위해서 벡터 공간 모델을 적용해서 단어 쌍의 코사인 유사도를 측정하였다(Salton et al. 1975, Widdows 2004). HAL과 벡터 공간 모델로 만들어지는 공간은 다차원이므로, 차원을 축소하기 위해서 PCA (Principal Component Analysis)와 SVD (Singular Value Decomposition)를 적용하였다. 유사성 군집을 위해서 비감독 방식과 감독 방식을 적용하였는데, 비감독 방식에는 클러스터링을 감독 방식에는 SVM (Support Vector Machine), 나이브 베이즈 구분자(Naive Bayes Classifier), 최대 엔트로피(Maximum Entropy) 방식을 적용하였다. 이 연구는 언어학적 측면에서 Harris (1954), Firth (1957)의 분포 가설(Distributional Hypothesis)을 활용한 의미 유사도를 측정하였으며, 심리언어학적 측면에서 의미 기억을 설명하기 위한 모델로 벡터 공간 모델과 HAL을 결합하였으며, 전산적 언어 처리 관점에서 기계학습 방식 중 감독 기반과 비감독 기반을 적용하였다.

주제어 : 분포 가설, 벡터 공간 모델, 기계학습, HAL, 심리언어학, 클러스터링, 다차원 축소, 코퍼스언어학

---

* 논문에 대한 심사를 담당한 세 분의 심사자들에게 감사를 드린다. 올바른 지적이 있어서 논문의 논지와 내용이 향상될 수 있었다. 논문에서 발견되는 모든 오류는 필자의 몫임을 밝혀둔다.
† 고려대학교 언어정보연구소, 연구세부분야: 언어모델링, E-mail: dsk202@korea.ac.kr

도 입

본 연구에서는 코퍼스의 문맥을 다시 해석하여 일정한 크기의 문맥에 따라 만들어지는 벡터 공간 모델(Vector Space Model)과 HAL(Hyperspace Analog to Language)을 적용해서 단어 간의 상관성을 토대로 유사한 의미를 가진 단어들을 군집하고자 한다. 이 연구의 관심은 크게 세 분야에 걸쳐있는데, 언어학, 심리언어학, 정보검색과 같은 전산적 언어 처리 분야이다.

분포 가설(Distributional Hypothesis)에[1, 2] 따르면 모든 언어학적 정보는 분포하는 환경에 따라 결정된다. 단어의 의미는 주변 문맥에 따라 결정되므로, 문맥을 분석하면 해당 단어의 의미를 알 수 있고 단어 간의 유사성을 측정할 수 있다[3]. 따라서 분포 가설에 기초한 단어 의미 분류가 어떠한 통계적 모델로 설명이 가능한지가 언어학적 연구의 관심이다.

심리언어학적으로 의미 기억(semantic memory)에 대한 연구는 오랜 전통을 갖고 있다. 문맥 측정과 개념 축 설정을 통한 피실험자 실험을 통해서 의미 기억을 추적할 수 있으나, 여러 횟수의 실험이 필요하고 피실험자의 직관에 근거하므로 객관성과 정확성 등에 의문이 제기되며 많은 자원(resource)이 소모된다[4]. HAL은 일련의 텍스트에서 얻어진 어휘 공기(lexical co-occurrence)가 의미 공간을 설명할 수 있다는 점에 기초해서, 어휘 공기 행렬로 문맥을 설명한다. 또한 HAL은 코퍼스에 근거하므로, 여러 자원 소모를 억제하고 객관성과 정확성을 담보한다[4]. 그러나, HAL은 어휘 공기 행렬을 활용한 모델이므로, 의미 군집을 위한 자세한 방법론의 제시가 필요하다. 이러한 문제에 대해서 본 연구는 벡터 공간 모델을 결합한 HAL을 제시한다.

벡터 공간 모델은 정보 검색의 입장에서 제시된 모델로 단어의 의미 공간을 정규화(normalized) 측정을 통해서 탐색한다[5]. HAL과 결합해서 의미를 분류하는 작업은 단어 의미들의 차원에 대한 고려가 필요하다[6]. 개별 단어가 하나의 차원이라면 너무 많은 차원이 계산되어야 하므로, 차원 축소와 이를 통한 의미 군집이 얼마나 정확한지에 대한 검증이 필요하다. 군집에는 크게 비감독 데이터 분류(non-supervised data classification)나 감독 데이터 분류(supervised data classification)가 있는데, 본 연구에서는 두 가지 방식을 다 활용하였다.

연구를 간략하게 설명하면 다음과 같다. 먼저 일정한 크기의 문맥을 통해서 단어 간의 상관성을 행렬로 측정하는 HAL을 활용하였다. 상관성 측정에서 고빈도와 저빈도에 다르게 측정되는 왜곡을 줄이기 위해서 벡터 공간 모델을 적용해서 단어 쌍의 코사인 유사도를 측정하였다[5, 6. 7]. HAL과 벡터 공간 모델로 만들어지는 공간은 다차원이므로, 차원을 축소하기 위해서 PCA (Principal Component Analysis)와 SVD (Singular Value Decomposition)를 적용하였다. 의미 분류를 위해서는 비감독 방식과 감독 방식을 적용하였는데, 비감독 방식에는 클러스터링을 감독 방식에는 SVM (Support Vector Machine), 최대 엔트로피(Maximum Entropy) 방식과 나이브 베이즈 구분자(Naive Bayes Classifier)를 적용하였다.

HAL과 연관된 기존 연구

인간의 의미 기억이 어떻게 구성되는가는 심리언어학의 오랜 주제로, 의미의 기억은 개념과 연관되어 있으며, 의미에 대한 정보로 축적된다. 또한 인간 의미 기억은 언어 사용을 통해서 가능하므로, 언어 사용을 통해서 어떠한 방식으로 의미 기억이 이루어지는지 많은 논의가 있었다.

단어 의미 군집은 의미 기억과 연관된 의미 공간을 군집하는데, A라는 단어와 B라는 단어가 다른 의미라면 서로 다른 의미 축으로 구성되는 의미 공간으로 나뉘게 된다. 이를 입증하기 위해서는 주어진 단어들이 주어진 의미 공간에서 어떻게 분포하는지에 대해서 심리언어 실험을 통해서 살펴보아야 한다. 문제는 이러한 실험은 시간, 인력, 비용 등의 많은 자원이 필요한 작업이다.

어휘 공기는 의미 공간 구성의 기초된다고 주장되는데[8], 이를 확장하면 실험자 실험보다 언어 자료인 코퍼스를 사용하는 것이 의미 공간을 찾기 위해서는 더 적절하다는 결론에 도달한다. 또한 실제 실험에서 소모되는 자원을 줄이려는 측면에서 코퍼스를 활용해서 필요한 의미 공간을 찾아내고, 구축하는 작업이 중요하게 되었다[4].

코퍼스에서 의미 공간을 찾아내는 작업에서는 해당 단어가 속한 공간에서 빈도수치를 토대로 통계적 측정을 한다. 다음과 같이 여러 통계 모델들이 제시되었다.

- LSA (Latent Semantic Analysis) [9]
- Probabilistic Topic Model [10]
- COALS (Correlated Occurrence Analog to Lexical Semantics) [11]
- Log-Likelihood [12]
- Log-Odds [13]
- HAL [4]
- PMI (Pointwise Mutual Information) [14]
- Vector Space Model [5]
- EM (Expectation Maximization) based Clustering [7]

여러 통계적 모델 중 본 연구에서 채택한 것은 HAL과 벡터 공간 모델이다. HAL은 어휘 공기 패턴을 측정하고, 거리를 측정해서 분포를 상대적 공간에서 측정이 가능하다. 그러나, 상대적 거리 측정 때문에 구성되는 의미에 따라 어떤 어휘는 너무 가깝고, 어떤 어휘들은 너무 멀게 측정된다. 따라서 절대적인 기준치에 의해서 공평하게 일률적으로 측정할 수 있는 방식이 필요한데, 수학적으로 정규화 개념의 적용이 필요하다[6]. 정규화가 적용된 모델은 벡터 공간 모델로, 본 연구에서는 HAL의 방식을 적용하였지만 단어들 간의 거리 측정 방식은 벡터 공간 모델에 기초하였다.

실험자 실험에서 발견되는 공간도 너무 많아서 이를 조정하는 작업도 어렵지만 [4], 코퍼스에서 발견되는 의미 공간은 헤아릴 수 없이 많아서, 공간을 줄이는 통계적 작업이 필요하다. 공간을 여러 개의 벡터 행렬로 전환해서 그 안에 존재하는 값들을 분해하는 SVD 방식을 활용해서 의미 공간의 축인 개념들을 구분한다[7]. 그 외에 여러 통계적 방식이 가능한데, 본 연구에서는 SVD와 더불어 공분산 (covariance) 행렬을 활용하는 PCA 방식도 적용하였다.

의미 공간을 찾아내는 작업은 의미의 축들을 찾아내는 작업이며, 의미의 축은 개념이 된다. 벡터 공간 모델과 HAL에 기초해서 개념을 자동으로 추론하는 비감독 기반으로 개념을 추출한다[15]. 본 연구에서는 비감독 기반 이외에 감독 기반의 작업으로서 단어들 간의 의미 공간이 일련의 정제된 자료에서 제시한 개념들의 공간과 일치하는지를 통계적으로 측정하였다. 연구에서 활용한 것은 세종 전자 사전

의 의미 분류인 '대상 부류(classes d'objets)'와 선정된 일련의 단어들의 군집과 비교하였다.[1] 분류 모델은 SVM, 나이브 베이즈 구분자, 최대 엔트로피를 사용해서 해당 분류 기준과 얼마나 통계적으로 유사성이 있는지를 측정하였다.

HAL과 벡터 공간 모델

HAL

HAL은 일정한 문맥 단위로 빈도를 추출한다. 여기서 문맥은 일정한 단어 배열의 길이를 말하는데, 일정한 단어의 길이에 따라서 하나의 문맥을 설정한다. 코퍼스 언어학의 용어로 윈도우(window)를 적용해서 문맥을 설정한다. 예를 들어서 잘라내는 단어의 단위가 10이면 윈도우가 10인 문맥이 설정된다. 일정한 크기의 윈도우로 잘라진 텍스트에서 어휘들 간의 공기 빈도를 측정하면 HAL의 모델이 완성이 된다. 가령 단어 집합 {w1, w2, w3, w4, w5, w6, w7} 단어들이 표 1에서 제시된 빈도와 같이 문맥 1, 문맥 2, 문맥 3에서 출현하였다고 가정해 보자.

여기서, 두 단어 간에 공기하는 빈도를 측정하면 표 2와 같은 공기 행렬이 작성된다. 이 행렬은 {w1, w2, w3, w4, w5, w6, w7} 단어 집합의 원소인 단어들의 유사성을 나타내는 HAL 모델이 된다.

HAL에서 산출되는 코퍼스를 활용해서 단어 간의 유사성을 측정한 결과는 단어 간의 연관성 실험과 연관된 인간 피실험자의 심리언어적 실험결과와 동일하다고 주장된다[8]. HAL이 심리적, 언어적 현상에 대한 설명을 제공하는 것은 인간 심리 계산 모델과 유사하다고 주장되었다. 이러한 관점에서 인간의 인지 심리 모델을 관찰하기 위한 방식으로도 HAL이 제안되었는데, 제1언어 습득의 측면과[17], 의미점화(semantic priming)의 측면에서[18] 논의되었다. 또한 정보검색, 자연언어 처리 분야에서도 활용된다[19].

---

1) '대상 부류'는 세종 전자 사전에 사용된 의미 분류 체계로 G. Gross에 의해서 제안된 어휘 의미 분류 체계이다[16]. 정보처리 및 분류 시스템 관점으로 살펴보면 이 체계는 소규모 온톨로지(Ontology)이다.

표 1. 문맥 출현빈도

	문맥1	문맥2	문맥3
w1	0	0	4
w2	2	4	0
w3	0	2	2
w4	2	0	0
w5	1	0	0
w6	0	3	0
w7	0	1	2

표 2. HAL 모델

	w1	w2	w3	w4	w5	w6	w7
w1	4	0	2	0	0	0	2
w2	0	6	2	2	1	3	1
w3	4	4	4	0	0	3	3
w4	0	2	0	2	1	0	0
w5	0	2	0	2	1	0	0
w6	0	4	2	0	0	3	1
w7	4	4	4	0	0	3	3

벡터 공간 모델

HAL은 여러 단어들의 상대적 빈도들을 측정할 수 있는 장점이 있지만, 일정하게 정규화해서 측정하지 않는다. 하나의 단어를 선택할 경우에 상대 빈도가 높은 어휘들은 유클리드 거리가 너무 가깝게 측정되기 때문에 유사성이 높은 어휘로 측정되고, 반대로 상대 빈도가 낮은 어휘는 유사성이 거의 없는 것으로 나타난다[6]. 이러한 문제점은 여러 단어들 간의 거리 왜곡으로 나타난다[5].

따라서, 상대적 거리로 인한 거리 왜곡을 없애야 하는데, 일정한 기준이나 측정

도구로 정규화 되어야 한다. 따라서 좌표의 원점에서부터 측정하는 방식으로 거리를 정규화하는 벡터 공간 모델이 제시되었다[5]. 벡터 공간 모델은 단어의 벡터를 측정하는 방식으로 정규 벡터(또는 길이 벡터)는 원점에서 자신까지의 거리로 측정된다. 예를 들어서 벡터 a의 정규 벡터는 원점에서부터 유클리드 거리 (1)로 측정한다.

(1) $\|a\| = \sqrt{a \bullet a}$

여기서 자기 자신을 정규벡터로 나누면 $\frac{a}{\|a\|}$ 이 되는데, 이것을 $\hat{a}$ 표기하며, 원점에서부터 지름이 1인 단위 벡터가 된다. 이를 활용해서 모든 상대 빈도로 측정되는 벡터들을 원점에서부터 지름이 1인 정규화된 벡터로 치환하고, 단어 간의 유사도를 나타내는 단어 간의 상관도는 (2)의 코사인 유사도를 통해서 측정한다.

(2) $\cos(a,b) = \frac{a \bullet b}{\|a\| \|b\|}$

두 단어들 각각이 단위 벡터로 정규화되어 있기 때문에 두 단어 간의 거리도 정규화 되므로, 코사인 유사도 측정법은 모든 단어 쌍들을 각각의 개별 거리가 아닌 원점에 출발하는 길이가 1인 단위 벡터의 방향성으로 측정한다[6, 7].

표 1의 단어 집합 {w1, w2, w3, w4, w5, w6, w7}의 출현빈도 행렬은 벡터 공간 모델로 표 1의 정규 벡터는 (1)을 적용해서 표 3과 같이 나타난다.

행렬 연산에서 두 행렬 간은 곱의 합으로 연산되는데,[2] (2)의 코사인 유사도를 적용해서 연산하면 표 2의 {w1, w2, w3, w4, w5, w6, w7} 단어 간 HAL 모델에 적용하면 표 4와 같이 산출된다.[3]

---

2) 두 행렬 $a = (a_1, a_2, a_3, ..., a_n)$와 $b = (b_1, b_2, b_3, ..., b_n)$의 연산 행렬 R^n은 $a \bullet b = a_1b_1 + a_2b_2 + a_3b_3 + ... + a_nb_n$ 와 같이 측정된다.

3) 예를 들어서 w5, w4의 코사인 유사도는 $\cos(w5, w4) = 0.47 + 0 + 0 = 0.47$ 와 같이 측정된다.

표 3. 정규화된 행렬

	문맥1	문맥2	문맥3
w1	0	0	1
w2	0.44	0.89	0
w3	0	0.7	0.7
w4	1	0	0
w5	1	0	0
w6	0	1	0
w7	0	0.44	0.89

표 4. 벡터 공간 모델을 적용한 HAL 모델

	w1	w2	w3	w4	w5	w6	w7
w1	1	0	0.7	0	0	0	0.89
w2	0	1	0.63	0.44	0.44	0.89	0.4
w3	0.7	0.63	1	0	0	0.7	0.91
w4	0	0.44	0	1	1	0	0
w5	0	0.44	0	1	1	0	0
w6	0	0.89	0.7	0	0	1	0.44
w7	0.89	0.4	0.91	0	0	0.44	1

정규화된 거리로 측정하는 벡터 공간 모델은 HAL 모델의 상대적 거리로 인한 왜곡된 유사성 측정의 문제를 해결한다.[4] 또한 두 단어 쌍 간의 코사인 유사도 측정을 통해서 단어 쌍의 거리 측정도 정규화 방식으로 측정한다. 각기 하나의 단어는 각각 하나의 차원에서 각 단어와의 유사성을 나타내게 된다. 예를 들어서 w2는

4) [5]에서 제안한 것은 문서 벡터에서 발견되는 용어 빈도(TF: Term Frequency)와 문서 빈도의 역수(IDF: Inversed Document Frequency)를 활용한 연산이다. 여기서 IDF를 직접적으로 활용하지는 않았지만 일정한 크기의 윈도우에서 측정된 빈도에 근거하였으므로, 문서 벡터에 근거한 [5]의 제안과 일치한다[6].

하나의 차원으로 단어 쌍 {(w2, w1), (w2, w2), (w2, w3), (w2, w4), (w2, w5), (w2, w6), (w2, w7)}이 생성되고, 각각의 코사인 유사도는 {0, 1, 0.63, 0.44, 0.44, 0.89, 0.4}로 측정된다. 스스로의 단어쌍인 (w2, w2)를 제외하고, (w2, w6)의 단어 쌍이 가장 유사도가 높다. 다르게 해석하면 w2 차원의 단어 쌍들은 w2의 의미 공간을 의미하고 이 공간에서 각각 측정되는 단어 쌍들은 해당 문맥의 의미 공간을 말한다.

여기서 문제가 되는 주어진 단어 쌍만큼의 많은 공간이 생성되므로, 차원을 줄이는 해석이 필요하다. 만약 n개의 단어가 존재한다면, 하나의 단어의 유사도는 하나의 차원이므로 n×n개의 차원이 생성이 된다. 따라서 차원이 너무 많게 되므로 해석이 가능한 차원으로 줄여야 한다.

실험 및 결과

실험 데이터

실험은 유사 의미 군집을 대상으로 하므로, 의미 분류된 코퍼스를 사용하였다. 사용한 코퍼스는 의미 분석이 된 세종 코퍼스 9백만 어절이며, 의미 분석은 표준국어대사전에 나타난 의미목록 번호인 어깨번호에 따라서 이루어졌다. 실험의 대상이 되는 어휘는 의미 분석된 세종 코퍼스에 기초로 해서 국립국어원에서 발표한 한국어 학습용 기본 어휘 6,000 단어 중 명사에 해당하는 어휘 중 등급이 제일 높은 어휘 224개를 대상으로 하였다. 이 어휘 정보에는 빈도와 세종의미분석 코퍼스와 일치하는 어깨번호, 의미, 등급이 표기되어 있다.

표 5. 한국어 학습용 기본 어휘

단어	천만어절 빈도	품사	의미	등급
가게	1195	명사		A
가구04	7434	명사	家具	B
가입	4397	명사		C
...	...	...	...	...

224개의 단어를 대상으로 어휘의 빈도를 5,000 이상인 경우만을 대상으로 세종 전자 사전에 수록된 의미 분류를 조사하였다. 목록에 수록된 어휘 중 세종 전자 사전에 수록되지 않은 어휘는 제외하였다. 또한 어깨번호에 해당하는 어휘의미를 세종 전자 사전에서 찾아보았는데, 예를 들어서 '가구04'는 어깨번호가 04의 경우인데, 세종 전자 사전에 의미 번호 04가 수록되어 있어야 하므로, 전자 사전을 찾아보고 04번이 있는지를 검사하였다. 어깨번호가 수록되지 않은 어휘는 조사 목록에서 제외하고 최종적으로 76개의 어휘를 선출하였다.

세종 전자 사전에는 단어의 의미를 '대상 부류'라는 의미 분류 기준에 따라서 분류하였는데, 술어 형식의 단어가 취하는 논항의 의미에 따른 분류 방식이다[20]. 이 분류 방식에 따르면 여러 의미 영역은 단계적으로 분화해서 계층적으로 분류된다[20]. 세종 전자 사전 체언부에 기술된 것을 살펴보면, 표 4와 같이 표준국어대사전의 표제어와 어깨번호에 일치하게 단어를 수록하였다. 이에 따라서 단어 의미에 대한 분류를 하였다.[5)]

세종 전자 사전의 의미 분류의 예를 보면 다음과 같다. 표 5의 '가게'의 경우에 세종 전자 사전에 수록된 의미는 '상업건물'이고, '상업건물'의 전체 의미 분류는 'ALL → 지상장소 → 건물 → 상업건물'로 되어 있다.

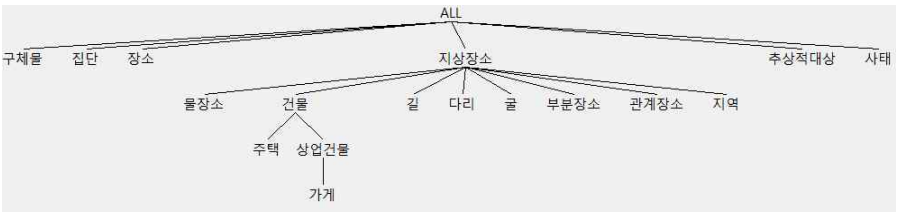

그림 1. 세종 전자 사전의 의미 분류

연구에서는 다음 절차에 따라서 자료를 추출하였다. 표 5의 단어들이 표준국어대사전의 표제어 단위이므로, 이와 같은 부분이 일치하는 세종 의미 코퍼스 9백만 어절을 대상으로 대상 단어들을 추출하고, 관련한 HAL과 벡터 공간 모델을 적용

5) 어깨번호는 동일한 의미를 갖는 단위이며, 동음이의어나 동철이의어 단위를 포함하고 있다. 어깨번호 아래 부분에는 다의적 성격의 의미를 분류하였다.

하였다. 또한 세종 전자 사전에서 실제 표제어 단위가 일치하는지에 따르는지를 확인하였다.[6)]

이전 소절에서 논의한 HAL과 벡터 공간 모델을 적용해서 일정한 윈도우 크기로 코퍼스를 잘라내서 이 어휘들을 대상으로 빈도를 추출하였다. 윈도우는 25, 15, 10, 5로 네 가지 경우로 나누어서 설정하였다. 이를 토대로 표 4와 같이 각각의 단어 쌍에 대한 코사인 유사도를 측정하였다. 최종적으로 추출된 것은 표 4와 같은 코사인 유사도 측정표로 5,776개의 단어 간 코사인 유사도 행렬이다. 개별 어휘를 하나의 차원으로 해석하면 76개의 차원이 만들어지게 된다. 76개의 다차원을 축소하기 위해서 이전 소절에서 논의한 바와 같이 PCA와 SVD방식을 적용해서 차원을 1, 2차원으로 축소하여서 의미의 군집이 어떠한지를 살펴보았다. 또한 클러스터링과 SVM 분류기, 나이브 베이즈 구분자, 최대 엔트로피 방식을 사용해서 HAL과 벡터 공간 모델을 적용한 모델이 얼마나 세종 전자 사전과 정확한지도 살펴보았다.

실험 및 결과 분석[7)]

표 4와 같이 HAL과 벡터 공간 모델만을 적용한 행렬은 단어 쌍의 상관성을 나타내므로, 이를 활용해서 클러스터링을 할 수 있다. 상관성 행렬을 통한 유클리드 클러스터링 작업을 하면, 그림 3과 같다.[8)9)]

클러스터링 방식은 비감독 자료 분류 방식으로 자료만으로 자동으로 군집을 추

---

6) 예를 들어서 '감사08'은 고마움의 의미인데, 세종 전자 사전에는 감사업무 및 감사업무를 담당한 사람의 의미인 '감사12'가 포함되어 있다. '감사08'은 '행위'로 '감사12'는 '직위인간'으로 의미 분류가 되어있으므로, 의미 분류는 '감사08'에 해당하는 '행위'를 선택하였다.

7) 연구에서 활용한 통계 및 기계학습툴은 R 2.11.1과 R 2.15.1이다.

8) 클러스터링에는 다른 알고리즘에 따른 많은 방법론이 있다. 연구에서는 개별 단어 의미에 따른 클러스터링을 목적으로 하므로 계층적 클러스터링 방식을 활용하였다. 계층적 클러스터링 방식으로 상향식, 하향식을 시도하고, 측정 방식도 유클리드, 맨하탄 등 여러 가지를 시도해 보았으나, 유클리드 거리 측정에 의한 클러스터링의 성능이 제일 좋았다.

9) 상관성은 피어슨 상관성 검증을 하였다(df = 5575, t = 2.212, p-value = 0.027)

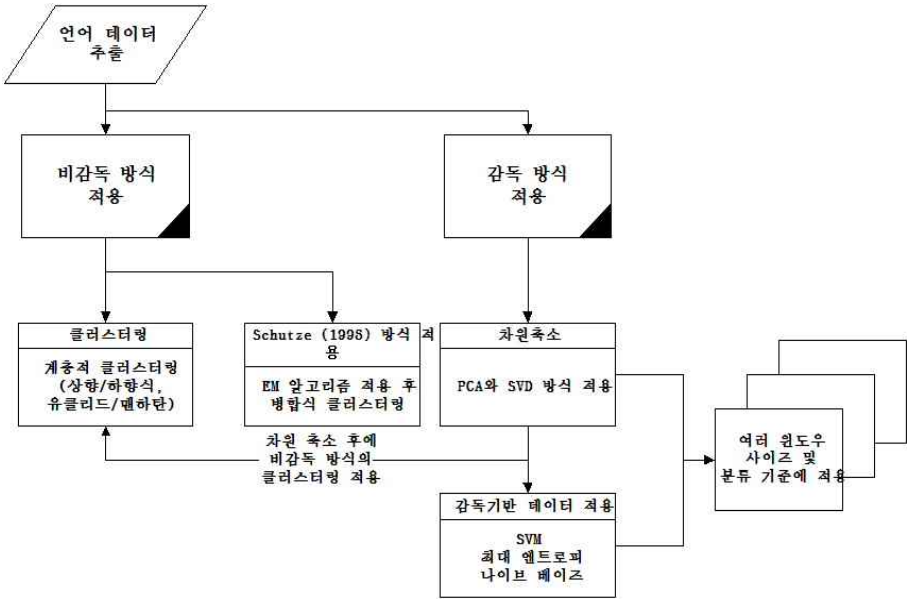

그림 2. 실험 흐름도

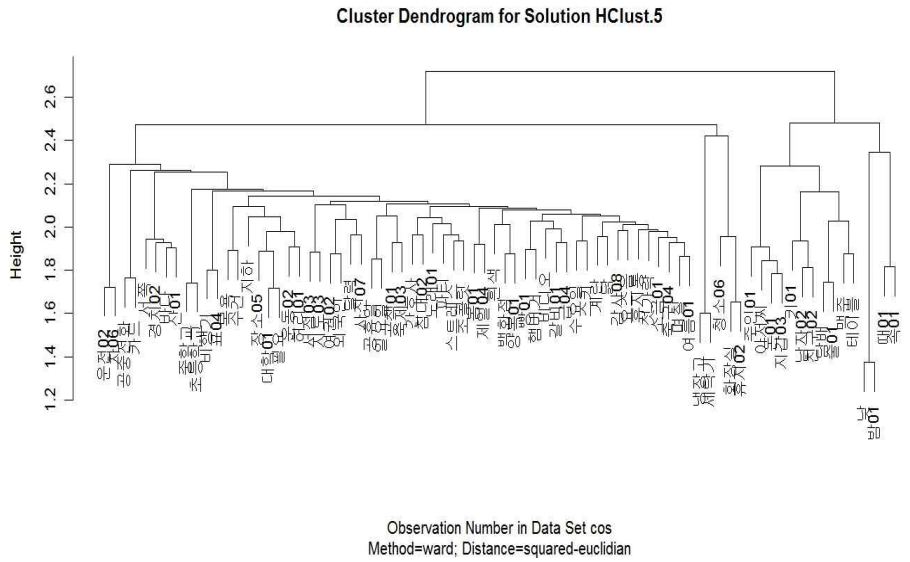

그림 3. 유클리드 거리에 의한 클러스터링

출해 낼 수 있다. 그림 3에서 계층화된 구조는 의미적 유사성이 있는 단어들이 모여 있는 것을 볼 수 있다. 예로 들어서 {운전, 차}, {중학교, 초등학교}, {비행기, 표}, {공휴일, 일요일}, {빵, 햄버거}, {냉장고, 세탁기}, {청소, 화장실, 휴지}, {낮, 밤}등의 단어군은 의미적으로 유사성이 높은 단어들의 집합이다. 그러나, 문제는 이러한 어휘들의 집합들이 다시 군집될 경우에 어떠한 의미 부류를 설명하는지가 확실하지 않다. 예를 들어서 {운전, 차}가 {공중전화, 카드}와 결합해서 {{운전, 차}, {공중전화, 카드}}의 클러스터에 포함되어 있는 경우에 {운전, 차, 공중전화, 카드}의 단어 군집이 무엇을 의미하는지 이해하기 어렵다. 다시 말하면, 어휘 군집들이 합쳐지면서 새로운 군집이 만들어지는데, 새로운 군집에서 단어들 간에 의미적 유사성이 적어지는 문제점이 발생한다.

[7]에서는 EM 알고리즘과[10] 카이제곱 독립성 검정을 활용해서 비감독 기반의 클러스터링 작업에서 유사도 군집의 효용성을 높이도록 하였다. 자동화된 비감독 방식의 클러스터링을 활용해서 의미 군집화를 시도하였기 때문에, 의미 분류나 비교를 위한 다른 자원이 필요 없는 장점이 있다[7]. [7]의 연구는 여러 측면에서 중요성이 있으므로 더 자세히 소개하면 다음과 같다. 의미 분류가 되어 있지 않은 원시 코퍼스 형태의 텍스트를 대상으로 벡터 공간 모델을 적용하였는데, 코사인 유사도에 기초한 용어 빈도와 문서 빈도의 역수를 각각 단어 벡터와 문맥 벡터로 지정하고 이를 연산한다. 이를 활용해서 해당 의미를 추정하는데, 여기서 문맥에 따른 무수히 많은 벡터가 생길 수 있으므로, EM 알고리즘과 병합식 클러스터링(agglomerative clustering) 방식을 조합해서 문맥을 줄여 나간다. 적절하게 줄여진 문맥을 대상으로 다시 SVD를 적용해서 차원을 줄여서 의미를 구분한다.

그러나 [7]의 실험이 많은 장점을 갖는 반면에 단점도 있다. 실험 결과가 통계적 유의성을 통해서 검증되는 것이지, 실제 환경에서 정확하게 적용된다는 것을 입증하기 위해서는 실험에 대한 정확한 재현이 필요하다. 또한 [7]에서 제시한 실험 데이터의 기준 정확도(baseline accuracy)가 너무 낮기 때문에 실험 결과가 유의성 검증

10) EM 알고리즘은 확률 모델에 관측이 불가능한 변수들이 포함되어 있는 경우 가우스 함수를 적용한 최대우도(maximum likelihood)를 활용해서 변수들의 기댓값을 최대화하는 예측 단계(Expectation-Step)와 기댓값을 최대화하는 최대화 단계(Maximization-Step)를 반복적으로 적용하는 과정이다. 자세한 부분은 [21]참조.

을 벗어날 확률이 매우 높다. 따라서 비감독 기반의 방식의 정확도가 실제 환경에서 가용할 정도의 효용성을 갖는지 논의될 필요성이 있다.

[7]의 방식은 의미 구분이 없는 원시 코퍼스를 대상으로 EM 알고리즘을 적용한 후에 적절히 의미 분류가 된 대상에 SVD를 적용하였다. 이와 다르게 본 연구의 대상은 의미 분류가 된 코퍼스를 대상으로 추출하였다. 따라서, [7]의 EM 알고리즘에 기초한 비감독 방식을 본 연구에 적용한다는 것은 논의의 의미가 없다.

다른 방식으로 [7]의 연구와 본 연구를 비교해 볼 수 있다. SVD의 적용 단계에는 어느 정도 의미 구분이 되어 있는 상태로, 본 연구에서 적용한 바와 같이 정확히 의미 구분된 상태가 아닐지라도 유사하거나 근접한 상태일 것이다. 이에 근거해서 SVD 방식을 본 연구에 적용하면 [7]에서 제시한 방식과 유사성이 있다. 연구에서는 SVD를 적용한 데이터에 감독 기반 데이터 분류 방식도 적용하여 어느 정도 정확한 의미 분류가 산출되는지도 비교하였다.

차원 축소 전에 [7]에서 제시한 EM 알고리즘을 적용한 후에 병합식 클러스터링을 적용하였다. [7]과 같이 코사인 유사도에 근거한 단어 벡터와 문서 빈도의 역수인 문맥 벡터를 연산하고[5], EM 알고리즘을 적용하였다. 클러스터링의 덴드로그램

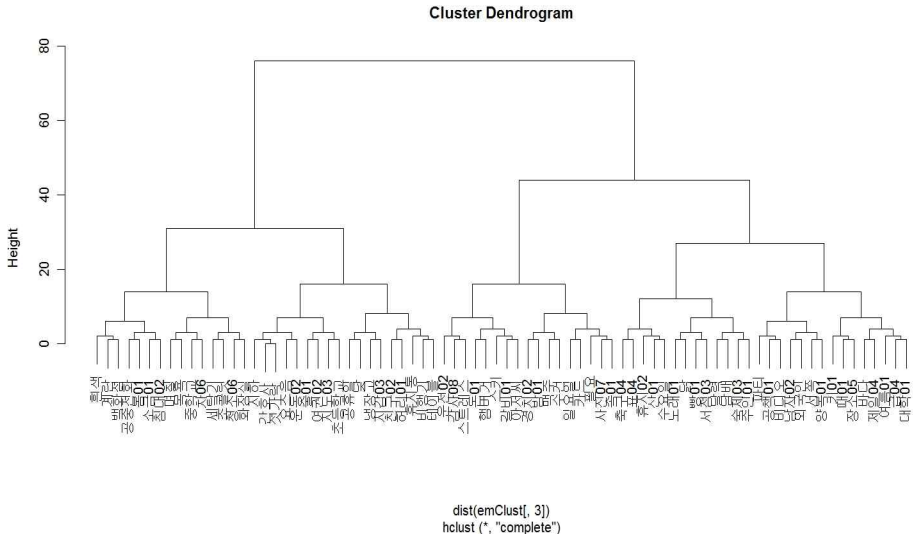

그림 4. EM 알고리즘 적용 후 병합식 클러스터링

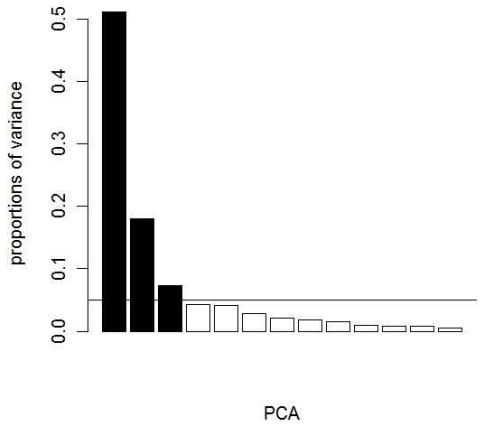

그림 5. PCA 차원들의 비율적 분산 수치

인 그림 4를 살펴보면 그림 3보다 더 부정확한 결과가 도출되었다. 예를 들어서 의미적으로 유사성이 적은 {냉장고, 지갑, 친구, 허리, 휴지통}이 같은 군집에 포함되어 있다. 이와 같이 그림 4가 부정확한 결과를 보이는 것은 본 연구에서 실행한 결과가 [7]에서 제시한 실험 결과와 다르다는 것을 보여준다. 특히 본 연구가 [7]과 서로 다른 점에서 출발하였기 때문에, 제시한 EM 알고리즘과 병합식 클러스터링을 결합한 방식이 좋은 결과를 보여주지 못한다. 따라서, 본 연구가 의미 구분된 데이터에 기초하기 때문에 [7]의 실험을 정확히 재현하기는 어려운 것으로 보인다.

다음으로는 감독 방식을 적용하기 위해서 차원 축소 방식인 PCA를 적용하였다. 먼저 76개의 차원들이 PCA 방식에 따라 변형되기 때문에, 통계적으로 가장 의미 있는 차원들을 찾아내야 한다. PCA 방식으로 전환한 차원들의 분산의 비율 수치를 살펴보면 그림 5와 같다. 여기서, 각 차원들의 분산의 비율 수치가 0.05이상이면 의미 있는 것으로 간주한다[22].[11] PCA의 경우 일반적으로 1, 2차 차원이 의미 있는 차원이 되기 때문에, 1, 2차 차원을 2차 평면에 투사하는 것이 가능하다. 그림 5에서 1~3차 차원들이 0.05이상인데, 이 중 1, 2차가 가장 높은 수치를 보인다. 그림 5에서는 76개 중 의미 있는 차원만을 제시하였다.

11) 여기서 분산 비율의 수치는 $\frac{\text{분산의 제곱}}{\text{분산의 제곱의 합}}$ 으로 구한다[22].

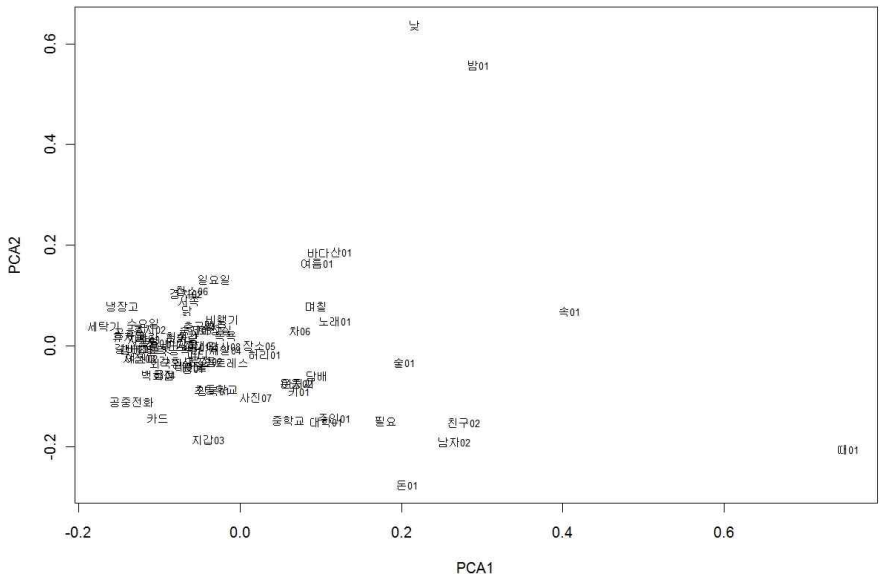

그림 6. PCA로 축소된 차원

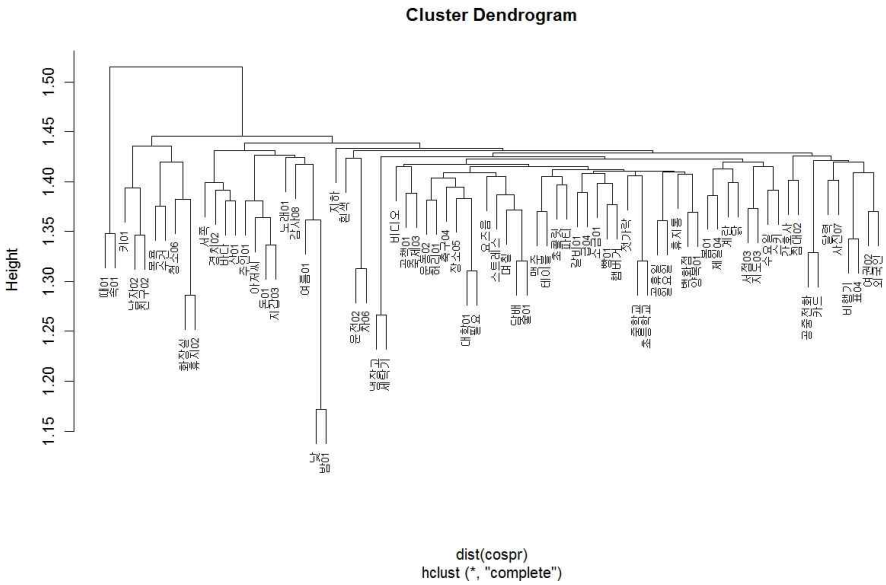

그림 7. PCA를 활용한 클러스터링

PCA 방식으로 차원을 1, 2차 계수를 적용해서 2차원으로 축소하고 투사하면 그림 6과 같다. 가까운 거리에 있는 어휘끼리는 의미적으로 유사성이 높으며, 반대로 먼 거리에 있는 어휘들끼리는 의미적 유사성이 적다. 여기서 의미적 유사성을 거리로 표현되었기 때문에 거리 관계를 통한 군집을 찾아낼 수 있으므로, 이를 기초로 클러스터링을 하면 그림 7과 같다. 클러스터링의 결과는 그림 3과 같은 문제점을 갖고 있다. 작은 군집들은 의미적 유사성이 높으나, 더 큰 군집으로 합쳐질 때 어떠한 의미적 유사성이 있는지 이해하기 어렵다. 따라서, PCA를 적용한 결과가 의미적으로 유사성이 있는 결과인지 감독 방식으로 측정해 보았다. 연구에서는 감독 방식들을 적용하기 위해서 의미 구분된 자료로 세종 전자 사전의 의미 분류를 적용해서 의미적 유사성 군집의 정도를 측정해 보았다.

세종 전자 사전의 의미 분류 방식을 소개하면 다음과 같다. 작업자들이 개별 어휘에서 발견되는 의미를 일일이 파악하고 분류하였는데, 계층적 의미 분류 방식을 적용하였다. 예를 들어서 '간호사'의 경우에는 '직업인간'으로 의미 분류되어 있는데, '직업인간'은 6단계에 걸쳐 설정된 의미 분류이다. 그림 7과 같은 체계가 관찰되는데, '구체물(1단계) → 구체자연물(2단계) → 생물(3단계) → 인간(4단계) → 역할인간(5단계) → 직업인간(6단계) → 간호사(실제어휘)'순이다. 6단계를 하나씩 감독 방식의 분류기를 적용해서 비교하고, 분류가 얼마나 정확한지 관찰하였다.

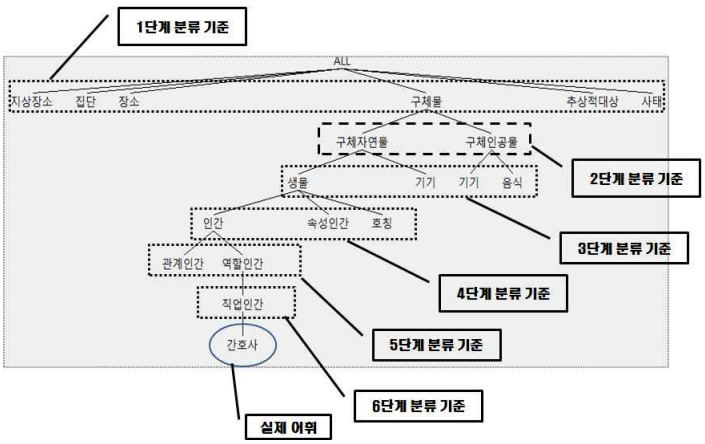

그림 8. 세종 전자 사전의 분류 기준

표 6. SVM 적용 결과

실제 분류 \ SVM 결과	구체물	사태	장소	집단	추상적대상
구체물	20	0	0	0	0
사태	1	10	0	0	0
장소	0	1	16	1	0
집단	0	0	0	1	0
추상적대상	1	0	1	0	15

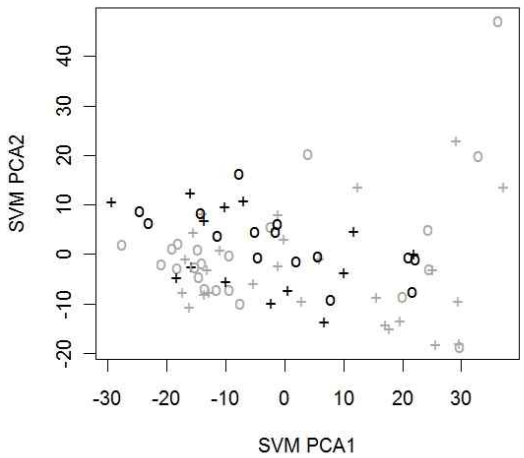

그림 9. 1단계 분류 기준을 적용한 SVM

SVM을 활용해서 '대상 부류' 체계의 1단계를 적용했을 때, 가장 정확도가 높았다. 76개의 어휘들은 1단계에 4개의 부류에 속하는데, '구체물, 사태, 장소, 추상적대상'이다. 각 어휘들은 SVM 분류기의 분류에 따라서 적절하게 분류되었는데, 그림 9와 같이 분류되고, 윈도우 10으로 적용했을 때 표 6과 같이 92.1%의 정확도를 보인다.

SVM은 지지 벡터(support vector)에 의해서 나타나는 공간의 분리 경계면(hyperplane)을 통해서 최적의 분리 작업을 하는데, 그림 9에서는 '+' 기호가 경계면에서 집중

되어 있다.[12] SVM이 예측한 정확도는 85.1%인데, 이 정확도가 통계적으로 유의미한지를 검증하기 위해서 표 6의 실제 정확도인 92.1%에 기준한 관찰치, SVM이 예측한 정확도 85.1%에 근거한 예측치를 토대로 카이제곱 검정을 적용하였다.[13] χ^2 =4.305, df = 1, p-value = 0.038로 통계적으로 유의미한 결과를 얻었다. 윈도우를 25, 15, 10, 5의 네 가지 경우에 SVM의 정확도와 실제 정확도를 살펴보면 그림 10과 같다. 윈도우 크기가 10일 때 가장 정확한 것으로 나타났다.

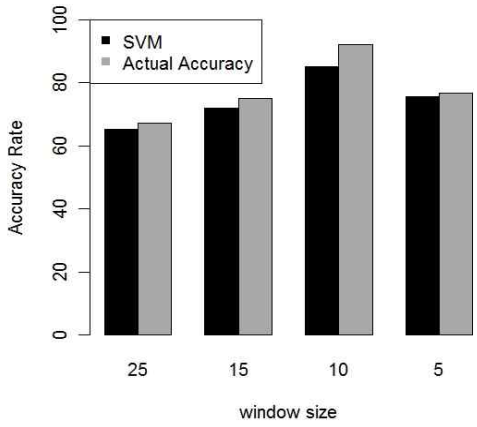

그림 10. 윈도우 크기와 정확도

정확도가 가장 높은 윈도우 10일 경우에 세종 전자 사전의 1에서 6단계의 분류기준을 적용했을 때, 그림 11과 같은 정확도가 산출되었다. 1단계일 경우에 가장 정확도가 높으며, 2~6단계는 정확도가 1단계에 비해서 상대적으로 낮다. 3단계는 분류기준이 25개로 가장 많고 6단계는 분류기준이 3개 정도이다. 또한 많은 단어들은 1~3단계까지만 분류되었고, 4~5단계는 분류되어 있지 않다. 모든 단어들이

12) SVM은 여러 종류의 커널(kernel)이 있는데, 연구에서 활용한 것은 C-classification이고 10차 교차 검증(10-fold cross-validation)을 하였다. 간단히 모델은 소개하면 다음과 같다. 커널은 $\min \frac{1}{2}\alpha^\top Q\alpha - e^\top \alpha$이고, 계수는 $0 \le \alpha_i \le C, i=1,...,l, y^\top \alpha = 0$ 로 정의된다.

13) 여기서 2×2 분할표이므로, 피어슨 카이제곱 검정(Pearson's Chi-Squared Test)에 예이츠의 수정(Yate's Continuity Correction)을 적용하였다.

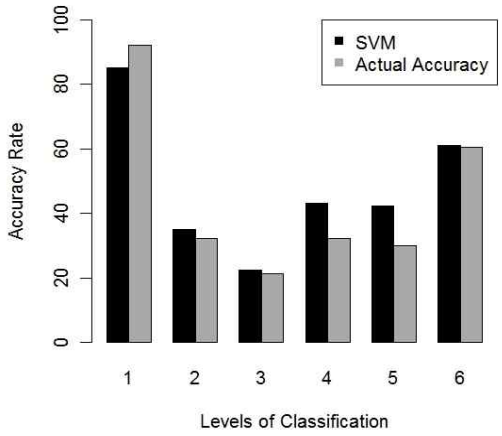

그림 11. 단계별 기준과 정확도

1단계는 분류되었고, 분류 기준의 수도 가장 적다. 1단계가 6단계보다 더 높은 분류 성능을 갖는 이유는 모든 단어들에 분류가 적용되어 있고, 반대로 6단계는 가장 적은 숫자의 단어에 분류가 적용되어 있기 때문이다. 결과적으로 분류 기준이 적고, 분류된 개수가 가장 많은 경우에 분류의 성능이 높다. 이것은 SVM이 통계적

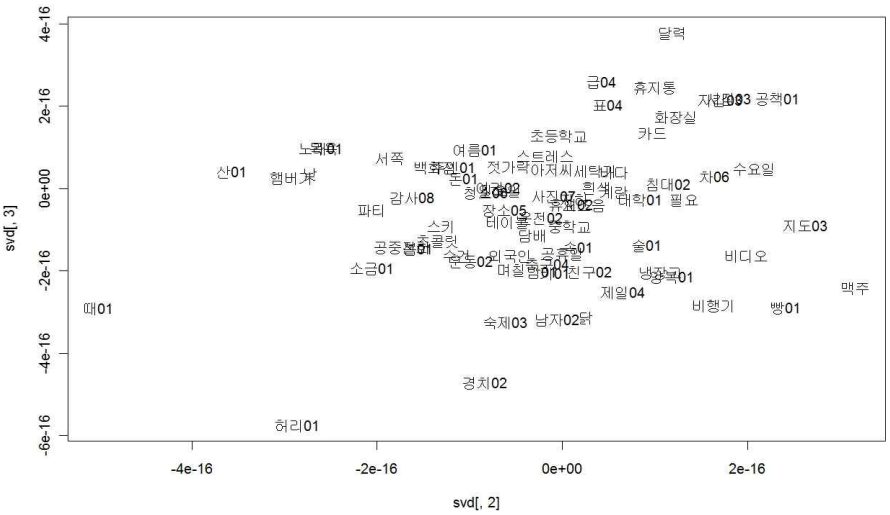

그림 12. SVD로 차원을 축소한 결과

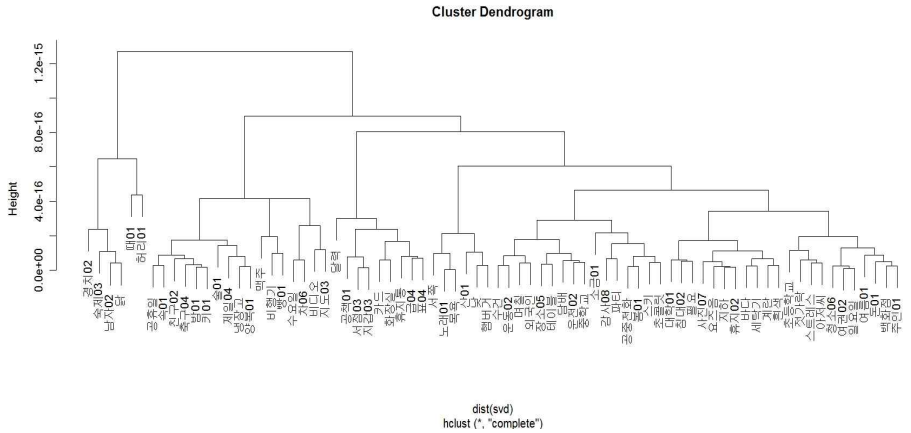

그림 13. SVD를 통한 클러스터링

이진분류(binary classification)에 근거하기 때문에, 분류 기준이 적은 경우와 분류된 데이터의 양이 가장 많은 경우에 성능이 높기 때문으로 해석된다.

SVD 방식을 적용하면 그림 12, 그림 13와 같이 각각 차원 축소와 클러스터링이 만들어진다. 여기서 그림 12는 SVD로 연산된 여러 차원 중 1, 2차 차원을 대상으로 나타낸 것이다. 그림 13을 살펴보면 SVD는 PCA에 비해서 더 부정확한 의미 유사성 군집을 보여준다. 예를 들어서, {햄버거, 낮, 산}과 같이 유사성이 없어 보이

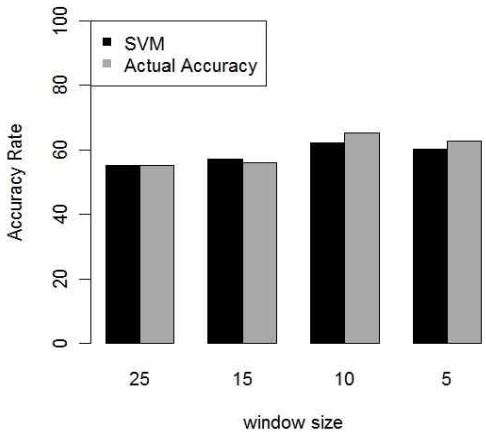

그림 14. SVD을 적용한 후에 SVM 적용한 결과

는 단어들의 군집이 나타난다.

그림 14는 윈도우 25, 15, 10, 5로 추출 후에 SVD 방식으로 축소한 차원으로 SVM 방식을 적용해서 정확도를 검증해본 결과이다. 윈도우가 10일 경우에 가장 정확도가 높았지만, PCA 방식에 비해서 정확도가 낮다. 윈도우가 10, 1단계 분류 기준을 적용한 SVD의 경우에 그림 3과 같은 각 차원의 비율적 분산 수치 비율이 가장 높은 1, 2차 차원도 0.15이상으로 낮은 유의성이 검출되고, SVM의 예측 정확도도 62.1%정도이다. 또한 이 수치를 대상으로 SVM의 예측치와 관찰치를 카이제곱 검정한 결과는 χ^2=0.641, df = 1, p-value = 0.423이어서 통계적 유의성을 가지지만, 신뢰성이 낮은 검증 결과가 산출되었다.

[21]에 따르면, SVD는 상대적 데이터 빈약성(relative data sparseness)이 발견되는 경우에 바람직한 차원 축소가 산출되지 않는다고 한다. 본 연구에서 발견되는 단어 공기 행렬에서 어떤 단어 쌍은 상대적으로 매우 높은 빈도를 보이고, 반대로 어떤 단어 쌍은 거의 공기하지 않는 데이터의 빈약성이 나타난다. 따라서, 이러한 문제로 인해서 SVD의 성능이 낮은 것으로 해석된다.

SVM은 데이터 마이닝, 기계학습 방식에 활용되는 통계적 방식을 사용하는 수학적 방식의 구분 모델이다. 이 이외에 여러 가능한 수학적 방식에 근거한 통계적 구분 모델이 가능한데, 본 연구에서는 나이브 베이즈 구분자, 최대 엔트로피 방식을 적용하였다. 각각을 소개하면 다음과 같다.

나이브 베이즈 구분자는 베이의 확률을 적용해서 확률 변수의 자질을 연산하는데, (3)에서 구분 범주 c는 자질 f의 조건부 확률로 연산된다.

$$(3)\quad P(c|f_1,...,f_n) = \frac{P(c)P(f_1,...,f_n|c)}{P(f_1,...,f_n)}$$

(3)에서 분모 $P(f_1,...,f_n)$를 상수로 가정하면, (4)와 같은 연산으로 (3)을 대체할 수 있다.

$$(4)\quad P(c|f_1,...,f_n) \propto P(c)\prod_{i=1}^{n} P(f_i|c)$$

따라서 개별 변수의 조건부 확률의 곱의 합으로 전체 연산을 할 수 있으며, 결정 모델은 (5)의 연산을 따른다.

(5) $classify(f_1,...,f_n) = \arg max P(c) \prod_{i=1}^{n} P(f_i|c)$

최대 엔트로피 방식은 제약 조건을 만족하는 여러 확률 분포 중 엔트로피가 최대가 되는 모델을 구성하는데, 최대 엔트로피는 (6)과 같이 구한다.

(6) $H(P) = -\sum_{x \in X, y \in Y} P(x,y) \log P(x,y)$

개체 x에 대해서 분류 범주 c를 예측하는 함수 $f(x,c)$는 일정한 조건이 성립되면 1, 아니면 0으로 설정한다. 구분 모델의 파라메터를 예측하는 방식은 (7)과 같다.

(7) $P(c|x) = \frac{1}{Z(c)} exp(\sum_{i=1}^{n} \lambda_i f_i(x,c))$

$= \frac{1}{Z(c)} \prod_{i=1}^{n} \mu_i^{f_i(c,x)}$

여기서, $\mu_i = e^{\lambda_i}$가 성립하는 분포 $Z(\lambda_1,....,\lambda_n)$를 가정하고, $\sum_{i=1}^{n} P(c_i|x_i) = 1$로 연산이 되도록 한다. 구분자 모델은 $P^* = \arg max H(P)$ 로, 최대치가 되는 분류 범주를 선택한다.

의미 군집의 의미가 높은 PCA 방식을 적용한 데이터에 나이브 베이즈 구분자와 최대 엔트로피 방식을 적용한 결과는 그림 15와 같다.

나이브 베이즈와 최대 엔트로피을 적용한 결과는 SVM을 적용한 결과와 유사성이 있다. 먼저, 세종 전자 사전의 1단계에서 가장 높은 정확도를 보이나,[14] 2단계

14) 나이브 베이즈 구분자는 95.3%, 최대 엔트로피는 92.1%의 정확도를 보인다.

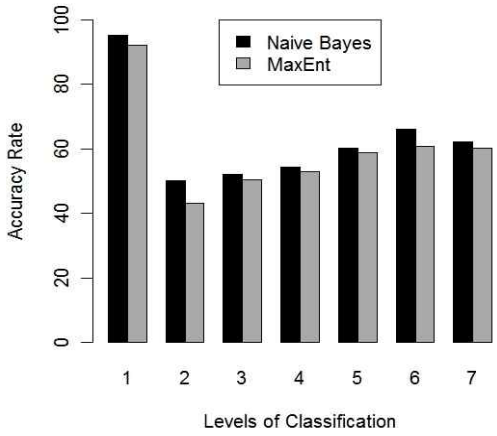

그림 15. 나이브 베이즈와 최대 엔트로피 방식의 정확도 비교

로 바뀌면서 정확도가 낮아진다. SVM과 비교해서 2~7단계에서 더 높은 정확도를 보이나, 정확도의 수준이 낮다. 나이브 베이즈의 정확도가 최대 엔트로피보다 조금 더 높으나, 거의 유사한 수준의 정확도를 보인다.

위의 결과를 통해서 볼 때, PCA를 적용한 결과는 SVM, 나이브 베이즈, 최대 엔트로피 구분 모델에서 모두 높은 수준의 정확도를 보인다. 따라서, PCA의 적용 결과가 의미적으로 유사성이 높은 단어들을 군집하는데 유용하다는 결론에 도달할 수 있다. 또한 PCA를 통해서 축소된 차원은 SVD를 통해서 축소된 차원에 비해서 더 좋은 결과를 보이는데, 차원 축소를 통한 왜곡이라는 문제를 PCA가 잘 해결하고 있다는 결론에도 도달한다.

결 론

본 논문은 HAL과 벡터 공간 모델을 활용해서 어휘의 상관성을 측정하고 유사성이 있는 단어들을 군집하였다. 제시한 모델은 코퍼스를 재해석하여 문맥에 따른 코퍼스를 구축하여 문맥에 기초한 공간을 분석한다. 여기서 다차원 공간이 산출되므로, 차원을 줄이기 위한 통계적 작업을 하였다. 또한 비감독과 감독 방식의 분류

를 통해서 의미 분류가 어느 정도 정확하게 군집하는지를 검증하였다.

연구에서는 한국어 학습용 어휘 중에서 빈도가 높고, 의미 분류가 세종 전자 사전에서 확인되는 76개의 기본 단어를 대상으로 하였다. 이 단어를 대상으로 상관성에 기초한 클러스터링을 통해서 비감독 데이터 분류를 시도하였는데, 군집에 대한 정확한 해석을 위해 다시 SVM을 적용한 감독 데이터 분류를 시도하였다. 여기서 다차원을 줄이기 위해서 PCA와 SVD를 적용하였는데, PCA의 경우에 더 정확한 결과가 도출되었다. 문맥 윈도우는 10일 경우에 가장 정확한 결과가 산출되었으며, 세종 전자 사전의 분류 체계는 1단계 분류에서 가장 높은 정확도의 의미 분류가 산출되었다.

일련의 실험을 통해서 얻어진 결과는 향후에 다음과 같은 방향에서 향상될 수 있다. [7]처럼 문맥 효과 측정을 토대로 비감독 데이터 분류가 필요하다. 비감독 방식은 감독 방식에 비해서 효용성이 떨어지나, 분류를 위한 기준 데이터가 필요 없는 장점이 있다. 연구를 향상 시키는 방식은 비감독 기반의 데이터 분류를 통해서 적절한 문맥에 대한 해석을 시도하는 것이다. 이를 위해서 다른 방식의 클러스터링이나 데이터 분류를 적용해야 할 것이다.

세종 전자 사전의 계층적 의미 구분 방식은 상층부는 상대적으로 적은 수의 의미 구분이 되고, 하층부는 상대적으로 많은 수의 의미 구분이 된다.[15] 연구 결과 SVM, 나이브 베이즈 구분자, 최대 엔트로피 방식에 근거한 구분에서도 1단계의 상층부 구분은 우수한 성능을 보이는데 반해서 그 이하의 단계인 2~7단계는 매우 낮은 성능을 보인다. 따라서, 이러한 많은 수의 의미 구분의 문제를 해결하기 위한 방식에 대한 연구도 향후 연구에 필요하다. 또한 다의어 구분이나 정확한 의미 구분에 대한 해석도 필요하다.

15) 심사자들 중 한 분이 이러한 문제점을 제기하였는데, 이 문제는 중요한 문제로 해결되어야 할 문제라고 여겨진다. 그러나, 본 논문에서는 차원 축소를 통한 단어 의미 유사성 군집의 문제를 다루고 있으므로, 논문의 연구 대상을 벗어난다고 생각한다. 이 논문에서는 향후 연구 대상으로 이 문제를 고민하고자 한다.

참고문헌

[1] Harris, Z. (1954), Distributional structure. *Word 10.23*, 146-162.

[2] Firth, J. R. (1957), A synopsis of linguistic theory 1930-1955. *In Studies in Linguistic Analysis,* 1-32. Oxford: Philological Society.

[3] McDonald, S., and Ramscar, M. (2001), Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In Proceedings of the 23rd Annual Conference of the Cognitive Science Society, 611-616.

[4] Lund, K. and Burgess, C. (1996), Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers, 28.2,* 203-208.

[5] Salton, G., A. Wong, and C. S. Yang (1975), A vector space model for automatic indexing, *Communications of the ACM, 18.11,* 613-620.

[6] Widdows, D. (2004), *Geometry and Meaning,* Center for the Study of Language and Information.

[7] Schutze, H. (1998), Automatic Word Sense Discrimination. *Journal of Computational Linguistics, 24.1,* 97-123.

[8] Burgess, C. and G. Cottrell (1995), Using high-dimensional semantic spaces derived from large text corpora, In Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society. 13-14.

[9] Landauer. T. and S. Dumais (1997), A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211-241.

[10] Steyvers, M. and T. Griffiths (2007), Combining background knowledge and learned topics. *Topics in Cognitive Science, 3.1,* 18-47.

[11] Rohde, D., L. Gonnerman and D. Palut (2005), An improved method of deriving word meaning from lexical cooccurrence. *Cognitive Psychology, 7,* 573-605.

[12] Dunning, T. (1993), Accurate methods for the statistics of surprise and coincidence. *Journal of Computational Linguistics, 19,* 61-74.

[13] Lowe, W. (2001), Towards a theory of semantic space I*n Proceedings of the 23rd*

Conference of the Cognitive Science Society.

[14] Church, K. and P. Hanks (1990), Word association norms, mutual information, and lexicography. *Journal of Computational Linguistics, 16,* 22-29.

[15] Song, D., P. Bruza and R. Cole (2004), Concept learning and information inferencing on a highdimensional semantic space, In ACM SIGIR 2004 Workshop on Mathematical /Formal Methods in Information Retrieval.

[16] 홍재성 (2007), **21세기 세종계획 전자사전 개발분과 연구보고서.** 국립국어원・문화관광부.

[17] Baroni, M., A. Lenci and L. Onnis (2007), Isa meets Iara: A fully incremental word space model for cognitively plausible simulations of semantic learning. In Proceedings of the 45th Meeting of the Association for Computational Linguistics.

[18] Jones, M., W. Kintsch and D. Mewhort (2006), High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55,* 534-552.

[19] Turney, P., and P. Pantel (2010), From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research, 37,* 141-188.

[20] 이성헌 (2001), 서술명사 기술을 위한 대상부류 개념의 활용. **프랑스어문교육, 12,** 129-149.

[21] Tan, P., M. Steinbach and V. Kumar (2006), Introduction to Data Mining, Pearson Education.

[22] Baayen, H. (2008), *Analyzing Linguistic Data.* Cambridge University Press.

1 차원고접수 : 2012. 4. 26
2 차원고접수 : 2012. 8. 14
최종게재승인 : 2012. 9. 17

(*Abstract*)

Word Sense Similarity Clustering Based on Vector Space Model and HAL

Dong-Sung Kim

Korea University

In this paper, we cluster similar word senses applying vector space model and HAL (Hyperspace Analog to Language). HAL measures corelation among words through a certain size of context (Lund and Burgess 1996). The similarity measurement between a word pair is cosine similarity based on the vector space model, which reduces distortion of space between high frequency words and low frequency words (Salton et al. 1975, Widdows 2004). We use PCA (Principal Component Analysis) and SVD (Singular Value Decomposition) to reduce a large amount of dimensions caused by similarity matrix. For sense similarity clustering, we adopt supervised and non-supervised learning methods. For non-supervised method, we use clustering. For supervised method, we use SVM (Support Vector Machine), Naive Bayes Classifier, and Maximum Entropy Method.

Keywords : Distributional Hypothesis, Vector Space Model, HAL, Supervised/Non-supervised Learning, Psycholinguistics, Clustering, Dimensionality Reduction, Corpus Linguistics