평균점에 대한 불일치의 합을 이용한 자동 단어 군집화†

이호, 서희철, 임해창
고려대학교 컴퓨터학과
서울시 성북구 안암동 5가 1번지
우: 136-701
leeho@nlp.korea.ac.kr, hcseo@nlp.korea.ac.kr, rim@nlp.korea.ac.kr

Automatic word clustering using total divergence to the average

Ho Lee, Hee-Chul Seo, Hae-Chang Rim
Natural Language Processing Lab.,
Dept. of Computer Science and Engineering, Korea Univ.

요약

본 논문에서는 단어들의 분포적 특성을 이용하여 자동으로 단어를 군집화(clustering) 하는 기법을 제시한다. 제안된 군집화 기법에서는 단어들 사이의 거리(distance)를 가상 공간상에 있는 두 단어의 평균점에 대한 불일치의 합(total divergence to the average)으로 측정하며 군집화 알고리즘으로는 최소 신장 트리(minimal spanning tree)를 이용한다. 본 논문에서는 이 기법에 대해 두 가지 실험을 수행한다. 첫 번째 실험은 코퍼스에서 상위 출현 빈도를 가지는 약 1,200 개의 명사들을 의미에 따라 군집화 하는 것이며 두 번째 실험은 이 논문에서 제시한 자동 군집화 방법의 성능을 객관적으로 평가하기 위한 것으로 가상 단어(pseudo word)에 대한 군집화이다. 실험 결과 이 방법은 가상 단어에 대해 약 91%의 군집화 정확도와(clustering precision)와 약 81%의 군집 순수도(cluster purity)를 나타내었다. 한편 두 번째 실험에서는 평균점에 대한 불일치의 합을 이용한 거리 측정에서 나타나는 문제점을 보완한 거리 측정 방법을 제시하였으며 이를 이용하여 가상 단어 군집화를 수행한 결과 군집화 정확도와 군집 순수도가 각각 약 96% 및 95%로 향상되었다.

1. 서론

코퍼스(corpus) 기반의 자연어 처리 기법에서 발생하는 가장 큰 문제점 중의 하나는 자료 부족(data sparseness)이다. 자료 부족 문제를 완화시키기 위한 대표적인 방법으로는 평탄화(smoothing) 기법, 부류 기반 모형(class based model), 유사도 기반 모형(similarity based model) 등이 있다[5].

이들 방법 중에서 부류 기반 모형은 각 단어에 대한 정보 대신에 그 단어가 속하는 부류 혹은 그 부류에 속한 단어들의 정보를 이용하는 방법이다. 이를 위해서는 WordNet과 같은 단어 분류법(taxonomy)이나 시소러스(thesaurus)가 필요하다. 그러나, WordNet이나 시소러스는 구축하는데 많은 노력이 필요할 뿐 아니라 이용하려는 의도에 따라 다른 기준에 의해 만들어진 단어 분류법이나 시소러스가 필요한 문제점이 있다. 이런 이유에서 주어진 특성에 따라 자동으로 유사한 단어들을 하나의 군집으로 묶어주는 자동 단어 군집화(automatic word clustering) 기법이 연구되어 왔다.

일반적으로 자동 단어 군집화 기법에서는 각 단어들에 대해 특징(feature)을 추출하고 이 특징들의 확률적 분포 특성에 따라 단어 사이의 유사도(similarity)나 거리(distance)를 측정하여 유사도가 높은(혹은 거리가 가까운) 단어들을 동일한 군집(cluster)에 할당하는 방식을 이용한다. 따라서 자동 단어 군집화 연구에서의 주된 쟁점으로는 크게 각각의 단어를 잘 표현해 줄 수 있는 특징 추출 과정과 특징 벡터(feature vector)를 이용하여 유사도를 측정하는 기준 및 유사도를 이용하여 단어들을 군집화 하는 알고리즘이 있다.

이 논문에서 제안하는 군집화 기법에서는 유사한 의미를 가지는 명사들을 같은 군집으로 묶어주는 것을 목표로 한다. 이를 위해 군집화 대상 명사가 출현한 문맥에서 의미적으로 밀접한 관계에 있는 단어들을 추출하여 특징 벡터를 구축한다. 두 명사 사이의 거리는 그 명사들의 특징 벡터에 대해 평균점에 대한 불일치의 합으로 계산되며 이는 최소 신장 트리에 기반을 둔 군집화 알고리즘에서 이용된다[1, 11].

† 이 논문은 1998년도 과학재단 핵심전문 연구과제 "다의어의 단어 의미 중의성 해결에 관한 연구" 지원에 의한 결과임

2. 자동 단어 군집화

이 장에서는 본 논문에서 이용하는 특징 벡터를 추출하는 방법과 평균점에 대한 불일치의 합을 이용하여 단어 사이의 유사도를 측정하는 방법 및 유사도를 이용하여 자동으로 단어를 군집화 하는 방법에 대해 소개한다.

2.1 특징 벡터

단어 사이의 유사도를 계산하기 위해서는 각 단어에 대한 특징 벡터를 추출하는 작업이 필요하다. 이때 사용되는 특징들은 군집화를 하는 의도에 따라 달라진다. 예를 들어 단어를 품사에 따라 군집화 하는 경우에는 단어의 품사적 특성을 잘 반영할 수 있는 특징들이 사용되어야 하며 단어를 의미에 따라 군집화 하는 경우에는 의미에 관련된 특징을 이용하여야 한다. 이 논문에서는 의미적 유사성에 따라 군집화 하는 것을 목표로 한다.

논문 [8]에서는 인간이 두 단어가 의미적으로 유사한지 판단할 때 그 단어들이 사용된 문맥의 유사성을 살핀다는 것을 증명하였다. 따라서 문맥에 나타나는 단어들을 특징으로 이용하면 단어를 의미에 따라 군집화 할 수 있다. 특히 명사와 그 명사를 지배하는 동사 사이에는 밀접한 관계가 있기 때문에 명사를 군집화 할 때에는 동사가, 동사를 군집화 할 때에는 명사가 좋은 특징이 될 수 있다. 이러한 성질을 이용하여 이 논문에서는 명사에 대한 군집화를 할 때 문맥에 나타나는 모든 단어들 대신 명사에 부착된 조사의 격과 그 명사를 지배하는 동사만을 이용하여 만들어진 **<격, 동사>** 쌍을 특징으로 사용한다.

각 명사에 대해 코퍼스로부터 모든 **<격, 동사>** 쌍에 대한 출현 빈도를 구해내기 위해서는 견고한 구문 분석기가 필요하지만 본 논문에서는 구문 분석 정보를 이용하지 않고 단문 분리가 가능한 문장들만 단문으로 분리한 다음 여기서 **<격, 동사>** 정보를 추출하여 군집화에 이용한다. 격을 결정하는 과정에서는 주격, 목적격, 보격 조사가 사용되었을 경우에는 격 정보를 그대로 사용하지만 부사격 조사의 경우에는 조사의 형태에 따라 세부적인 격을 정확히 판별할 수 있는 것들만 이용하고 나머지 부사격 조사나 보조사는 고려 대상에서 제외한다.

이 작업의 결과 추출된 군집화 대상 명사들에 대한 **<격, 동사>** 쌍의 출현 확률로 이루어진 확률 벡터는 평균점에 대한 불일치의 합을 이용하여 명사들 사이의 거리[1]를 계산하는데 사용된다.

2.2 유사도 측정 방법

확률 벡터가 주어졌을 때 두 단어 사이의 유사도를 측정하는 방법으로는 L_1 및 L_2 norm및 cosine coefficient 등과 같은 기하학적 거리 측정법[6], Kendall의 τ coefficient와 같은 통계적 거리 측정법[4], KL divergence [7]나 평균점에 대한 불일치의 합[1]과 같은 정보 이론 기반 거리 측정법 등이 있다.

명사의 집합 $W=\{w_1, w_2, \cdots, w_n\}$와 각 명사 w_i에 대한 **<격, 동사>** 쌍의 출현 확률 벡터 $\vec{P_i}=(p_{i1}, p_{i2}, \cdots, p_{im})$가 주어졌을 때 각 방법의 유사도 계산식은 다음과 같다(단, n은 명사의 종류 수, m은 특징의 종류 수, p_{ij}는 명사 w_i가 나타난 문맥에서 j번째 **<격, 동사>** 쌍 v_j가 나타나는 확률을 의미한다.).

■ L_1 **norm**

$$L_1(w_i, w_j) = \sum_{k=1}^{m} |p_{ik} - p_{jk}|$$

■ L_2 **norm**

$$L_2(w_i, w_j) = \|\vec{p_i} - \vec{p_j}\| = \sqrt{\sum_{k=1}^{m} (p_{ik} - p_{jk})^2}$$

■ **cosine coefficient**

$$\cos(w_i, w_j) = \frac{\sum_{k=1}^{m} p_{ik} p_{jk}}{\|\vec{p_i}\| \|\vec{p_j}\|}$$

■ τ **coefficient**

$$\tau(w_i, w_j) = \frac{2}{m(m-1)} \sum_{k=0}^{m} \sum_{l=k+1}^{m} \frac{p_{ik} - p_{il}}{|p_{ik} - p_{il}|} \frac{p_{jk} - p_{jl}}{|p_{jk} - p_{jl}|}$$

■ **KL divergence**

$$D(w_i \| w_j) = \sum_{k=1}^{m} p_{ik} \log \frac{p_{ik}}{p_{jk}}$$

■ **total divergence to the average**

$$\begin{aligned} & A(w_i \| w_j) \\ &= D\left(w_i \middle\| \frac{w_i + w_j}{2}\right) + D\left(w_j \middle\| \frac{w_i + w_j}{2}\right) \\ &= \sum_{k=1}^{m} \left\{ p_{ik} \log \frac{2p_{ik}}{p_{ik} + p_{jk}} + p_{jk} \log \frac{2p_{jk}}{p_{ik} + p_{jk}} \right\} \\ &= 2\log 2 - H(w_i) - H(w_j) + H(w_i + w_j) \\ &= [0, 2\log 2] \end{aligned}$$

이 논문에서는 이들 여러 가지 방법 중에서 평균

1) 단어 사이의 거리는 단어 사이의 유사도에 반비례하기 때문에 군집화를 할 때는 유사도 대신 거리를 사용할 수 있다.

점에 대한 불일치의 합을 이용하여 단어 사이의 거리를 계산한다. KL divergence와 비교할 때 이 방법은 몇 가지 장점을 가지고 있다. 첫째로, KL divergence를 이용하여 구해진 거리는 0~∞ 사이의 범위를 가지지만 평균점에 대한 불일치의 합에 의해 구해진 거리는 0~ 2log2 사이의 고정된 범위의 값만을 가진다. 둘째로, KL divergence에서는 p_{ik}의 값이 0일 경우 값을 계산할 수 없기 때문에 평탄화가 부가적으로 필요하다. 그러나 평균점에 대한 불일치의 합을 이용할 때에는 p_{ik}나 p_{jk}가 0일 경우에도 아무런 문제가 발생하지 않는다.

2.3 최소 신장 트리를 이용한 군집화

단어 사이의 유사도나 거리가 주어졌을 때 단어들을 군집화 하는 알고리즘에는 집적적 군집화(agglomerative clustering)[10], k-means 알고리즘[3], expectation maximization(EM) 알고리즘[2], k-nearest neighbor 알고리즘[9], 최소 신장 트리 알고리즘[11] 등이 있다.

이 논문에서는 이들 방법 중에서 최소 신장 트리를 이용한 군집화 방법을 이용한다. 최소 신장 트리를 이용한 군집화 알고리즘에서는 각 단어들을 완전 연결 그래프 상에서의 정점(node)으로, 단어 사이의 거리를 두 정점을 이어주는 간선(edge)의 가중치(weight)라 간주한다. 이 방법에서는 군집화 대상 단어들과 특징 벡터로부터 추출된 거리를 이용하여 최소 신장 트리를 생성한 다음 가중치가 가장 높은(즉, 거리가 가장 먼) $k-1$ 개의 간선을 제거하여 k 개의 군집을 생성한다. 최소 신장 트리를 이용한 군집화 방법은 유사도 재계산이 필요 없기 때문에 계산 복잡도가 낮은 장점을 가지고 있지만 하나의 거대한 군집이나, 단어 하나 만을 가지는 군집을 생성할 수 있는 단점을 가지고 있다. 본 논문에서는 이러한 문제점을 완화시키기 위해 군집의 최대 크기를 제한하였다.

3. 실험 및 평가

이 장에서는 2 장에서 소개된 유사도 계산 방법과 군집화 알고리즘에 의해 두 가지의 군집화 실험을 수행한다. 첫 번째 실험은 군집화 대상 명사가 나타난 문장에서 **<격, 동사>** 쌍이 나타날 조건부 확률을 이용하여 명사를 군집화 하는 것이며 두 번째 실험은 가상의 단어에 대한 군집화를 통해 제안된 군집화 기법의 성능을 평가하는 것이다.

3.1 실험 1 : 명사의 의미별 군집화

이 실험에서는 명사들을 의미에 따라 자동으로 군집화 한다. 일반적으로 의미가 유사한 명사들은 동사들과의 관계도 서로 유사하다. 그리고, 한국어에서는 명사에 부착된 조사의 격을 이용하여 명사와 동사 사이의 관계를 파악할 수 있기 때문에 **<격, 동사>** 쌍은 명사를 의미별로 군집화 하는 작업에서 특징으로 유용하다.

표 1. 명사의 자동 군집화 결과(일부)

군집	단어
2	가게 감옥 교실 댁 도서관 부엌 술집 시내 식당 화장실
4	가까이 근처 둘레 멀리 부근 사방 입구 주변 주위 지점
6	가방 봉투 사례 상자 술잔 시중 예 잔 주머니 짐
8	가스 기분 나이 물질 분자 색 색깔 성문 심정 철
10	가슴속 김가 눈앞 마음속 머릿속 사이 시야 양쪽 염두 중간
12	가을 겨울 밤 방학 봄 시절 여름 옛날 오후 하루
14	가족 딸 새끼 식구 아기 아들 자녀 자식 주민 환자
16	각도 관점 맥락 면 솜씨 시각 자세 차원 측면 태도
18	간 그때 당시 도중 반면 이후 초 초기 평소
20	강물 공기 눈물 땀 물 바닷물 소주 술 커피 피

이 실험에서는 약 800만 어절 크기의 코퍼스에서 출현 빈도가 100 회 이상인 1246 개의 명사들을 추출한 다음 2.1 절에서 설명된 방법으로 30 회 이상의 빈도를 가지는 9524 개의 **<격, 동사>** 쌍들을 추출하여 구성한 특징 벡터를 이용하여 군집화를 수행하였다. 이때 생성된 전체 군집의 최대 크기는 10 단어로, 최대 거리는 1.7log2로 제한하였다. 그 결과 235 개의 군집이 생성되었다.

표 1은 이 실험을 통해 얻어진 군집화 된 명사의 일부이다. 표 1을 살펴보면 대부분의 군집이 사람의 직관에 일치되는 것을 알 수 있다. 군집화가 잘못된 경우에도 추상 명사, 부사성 명사, 물질 명사 등은 대부분 같은 부류의 명사들과 군집을 이루고 있음을 볼 수 있다. 표 1의 실험 결과에서 군집 6에 '사례'나 '예', '시중'이 포함된 것은 군집 6에 포함된 명사들이 '들다'라는 동사의 목적어로 사용되는 특성이 강한 점이 크게 작용한 결과이다. 이때 '들다'의 의미는 목적어에 따라 여러 가지로 나뉘어질 수 있지만 현재 실험에 사용된 코퍼스에서는 의미 정보가 포함되어 있지 않기 때문에 이와 같은 오류가 발생한 것이며 이 오류는 의미 정보가 제공되면 해결될 수 있다.

표 2는 크기별 군집들의 개수를 나타낸다. 표 2를 보면 크기가 1 단어 혹은 10 단어인 군집이 많음을 볼 수 있는데 이는 최소 신장 트리를 이용한 방법의 특성 때문인 것으로 추측된다. 하지만 군집화 되지 않은 단어의 수가 104 개로 전체 단어의 10% 미만으로 적을 뿐 아니라 이들을 중심(centroid)까지의 거리가 가장 가까운 군집에 재할당 시킬 수도 있기 때문에 큰 문제가 되지는 않는다.

한편, 최소 신장 트리를 이용한 군집화 기법에서

표 2. 크기별 군집의 개수

군집 크기	군집 개수	군집 크기 × 군집 개수
1	104	104
2	12	24
3	4	12
4	3	12
5	3	15
6	1	6
7	3	21
8	3	24
9	2	18
10	101	1010

는 단어들이 하나의 거대 군집으로 집중되는 경우가 종종 발생하는데 이는 단어별 평균 유사도가 불균등하게 분포되어 있을 때 특히 잘 나타난다. 그림 1은 평균점에 대한 불일치의 합으로 거리를 측정하였을 때 단어의 빈도에 따른 그 단어와 다른 단어 사이의 평균 거리를 나타낸다. 그림 1에서 볼 수 있듯이 단어의 빈도가 높을수록 평균 거리가 낮음을 알 수 있다. 앞의 실험 환경을 군집의 최대 크기를 정의하지 않도록 변경했을 경우 전체적으로 거대한 몇 개의 군집이 생성되고 나머지 군집들은 대부분 크기가 한 단어인 결과를 얻었는데 이와 같은 현상은 그림 1에서 보여지듯이 거리 측정 방법의 특성에 그 원인이 있다.

3.2 실험 2 : 가상 단어의 군집화

군집화 기법의 성능을 평가할 때 군집화 결과 중에서 어떤 단어나 군집이 정확하고 어떤 것이 잘못 되었는지를 직접 조사하기는 어렵기 때문에 군집화 기법의 성능 평가는 주로 간접적인 방법으로 이루어진다.

군집화 기법의 성능 평가에 일반적으로 많이 사용되는 방법은 군집화에 의해 정보가 얼마나 손실되었는가를 살펴보는 것이다. 다시 말해, 단어 단위로 측정된 정보량과 군집으로 변환되어 측정된 정보량의 차이가 적을수록 더 좋은 성능을 가진 군집화 기법으로 평가할 수 있다. 하지만 이러한 평가 방법은 두 개 이상의 군집화 기법 사이

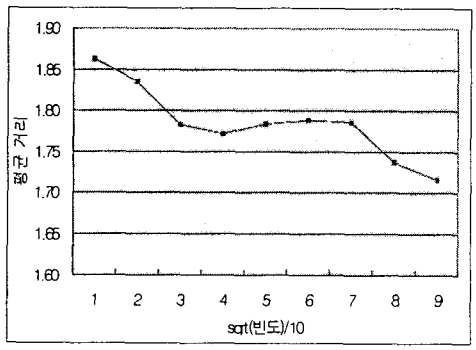

그림 1. 단어의 빈도와 평균 거리의 관계

의 상대적인 비교만 가능할 뿐만 아니라 정보의 손실양이 적다고 해서 바람직한 군집화 방법이라는 것도 보장하지 못한다. 이러한 이유 때문에 이 실험에서는 가상으로 생성한 단어를 군집화 함으로써 절대적인 평가를 수행하고자 한다.

이 실험에서 사용되는 가상 단어는 다음과 같이 생성된다. 먼저 코퍼스에서 출현 빈도가 1,000~3,000 범위에 속하는 단어를 선택한다. 다음으로 각 단어들에 대해 2~10 범위 내의 임의의 개수로 가상 단어를 생성시킨다. 예를 들어, '계획'이라는 단어가 3개의 가상 단어로 나누어진다면 '계획0', '계획1', '계획2'라는 가상 단어를 생성시킨다. 가상 단어가 생성되면 코퍼스에서 원래 단어가 출현한 각 경우에 대해 그 단어의 가상 단어들 중에 임의로 선택된 하나로 대체한다. 이 작업을 수행하고 나면 코퍼스로부터 가상 단어에 대한 특징 벡터를 추출할 수 있다. 따라서, 이후 과정에서는 실험 1에서의 방법과 동일하게 가상 단어에 대해 군집화를 수행하고 나서 동일한 단어로부터 생성된 가상 단어들이 같은 군집에 할당되는지를 조사하면 군집화의 성능을 알아볼 수 있다.

여기서 출현 빈도가 1,000~3,000 범위의 비교적 고빈도에 속하는 단어를 선택한 이유는 출현 빈도가 너무 낮은 단어일 경우 여러 개의 가상 단어로 나누었을 때 특징 벡터가 빈약하게 되어 군집화 평가에 부적절하기 때문이다.

이 논문에서는 가상 단어의 군집화 결과에 대해 군집화 정확도와 군집 순수도의 측면에서 평가를 수행한다. 군집화 정확도는 전체 단어에 대해 정확히 군집화 된 단어의 비율이며 군집 순수도는 전체 군집에 대해 올바른 군집의 비율이다. 여기서 정확히 군집화 된 단어는 주어진 단어로부터 생성된 모든 가상 단어가 하나의 군집에 속해 있는 경우를 의미하며 올바른 군집이란 하나의 군집을 구성하는 모든 가상 단어가 동일한 단어로부터 생성된 경우를 의미한다. 이 두 가지 척도를 수식으로 표현하면 다음과 같다.

$$\text{군집화 정확도} = \frac{\text{정확히 군집화된 단어의 수}}{\text{전체 단어의 수}} \times 100\%$$

$$\text{군집 순수도} = \frac{\text{올바른 군집의 수}}{\text{전체 군집의 수}} \times 100\%$$

출현 빈도가 1,000~3,000 범위에 속하는 79 개의 명사를 448 개의 가상 단어로 나눈 다음 전체 가상 단어를 79개의 군집으로 만들되 한 군집당 최대 단어의 수를 10 단어로 제한한 결과 91.13%의 군집화 정확도와 81.01%의 군집 순수도를 나타내었다.

표 3은 군집 순수도를 측정할 때 잘못된 군집으로 결정이 된 군집들을 나타낸다. 표 3에서 군집 2, 4, 7, 28, 36, 37, 41 등은 비록 동일한 단어로부터 생성된 가상 단어들만으로 이루어진 군집은 아니지만 사람이 판단하기에 서로 유사한 단어로 이루어진 군집임을 알 수 있다. 따라서 이 군집들까지 올바른 군집이라고 고려할 경우 군집 순수도는 90% 이상이 된다. 또한 이렇게 실제 유사한 단어가 하나의 군집으로 묶여진 결과 다른 가상

표 3. 잘못 생성된 군집

군집	구성 단어
2	거기0 거기1 거기2 거기3 거기4 거기5 여기0 여기1 여기2
4	결과0 결과1 결과2 현상0 현상1 현상2 현상3 현상4 현상5
6	계획0 계획1 계획2 내용0 내용1 내용2 내용3 내용4 내용5
7	고개0 고개1 고개2 고개3 고개4 고개5 고개6 머리0 머리1
8	과정0 과정1 상황0 상황1 상황2 상황3 상황4 상황5
9	관계0 관계1 관계2 성격0 성격1 의미0 의미1 의미2 의식2 의식6
11	기능0 기능1 무엇0 무엇1 무엇2 무엇3 무엇4 무엇5 무엇6 무엇7
23	방식0 방식1 방식2 방식3 방식4 방향0 방향1 방향2 방향3 방향4
28	사회0 사회1 사회2 세계0 세계1
30	생활0 생활1 생활2 생활3 생활4 생활5 역할0 역할1 역할2 역할3
31	서울0 서울1 학교0 학교1 학교2 학교3 학교4 학교5 학교6
36	아버지0 아버지1 아버지2 아버지3 아이0 아이1 어머니0 어머니1 어머니2
37	얘기0 얘기1 얘기2 얘기3 얘기4 얘기5 얘기6 이야기2 이야기7 이야기8
41	운동0 운동1 운동2 운동3 운동4 운동5 운동6 활동0 활동1 활동2
47	이유0 이유1 이유2 필요0 필요1 필요2 필요3 필요4

단어들의 군집이 두 개 혹은 그 이상으로 분리되어 군집화 정확도가 낮아진 것이므로 실제의 군집화 정확도는 실험 결과보다 더 높을 것으로 예상된다.

한편, 이 실험에서는 실험 1에서 언급했던 거리 측정 방법의 문제점을 보완하기 위한 세 가지 변환을 제시하고 이를 적용한 방법들의 성능을 기존의 거리 측정 방법의 성능과 비교였다. 이 실험에서 이용한 첫 번째 변환은 거리 측정 결과를 단어별로 정규 분포로 변환하여 합한 다음 이를 단어의 빈도의 제곱근으로 나누어주는 것이다. 정규 분포로 변환하면 단어의 빈도에 상관없이 평균 거리가 0이 되지만 거리 행렬이 대칭이 되지 않기 때문에 이들의 합을 이용하였다. 한편 두 번째 변환은 특징 벡터를 크기가 1인 단위 벡터로 변환하여 기존의 방법을 적용하는 것이다. 기하학적 거리 관점에서 볼 때 두 확률 벡터가 같은 각도를 이루고 있어도 축에 가까울수록 거리가 더 멀다. 그러나 이를 단위 벡터로 변환하면 두 벡터 사이의 각도의 크기 순서와 거리의 순서가 동일하기 때문에 이 변환은 의미를 지닌다. 마지막으로 세 번째 방법은 앞의 두 방법을 같이 이용하는 것이다. 이들 변환을 식으로 정의하면 다음과 같다.

■ **변환 1 : 정규화 시킨 거리의 합을 빈도의 합으로 나눈다.**

$$Distance(w_i, w_j) = \frac{1}{freq(w_i) + freq(w_j)} \times \left(\frac{A(w_i, w_j) - \mu_i}{\sigma_i} + \frac{A(w_i, w_j) - \mu_j}{\sigma_j} \right)$$

■ **변환 2 : $A(w_i, w_j)$의 계산식에 있는 p_{ij}를 p_{ij}'로 대체시켜 단위 벡터에 대해 $A(w_i, w_j)$를 계산한다.**

$$p_{ij}' = \frac{p_{ij}}{|\vec{p_i}|}$$

■ **변환 3 : 변환 2와 같이 단위 벡터로 변환한 다음 변환 1과 같이 정규화 시킨 거리의 합을 빈도의 합으로 나눈다.**

이 세 가지 변환을 적용한 방법 및 cosine coefficient를 가상 단어 군집화에 적용한 결과가 표 4에 나타나 있다. 표 4에서 거리 측정 방법에서 방법 0은 평균점에 대한 불일치의 합을 의미하며, 방법 1~3은 방법0에 대해 변환 1~3을 적용한 경우, 방법 4는 cosine coefficient를 이용하는 경우를 나타낸다.

표 4를 살펴보면 방법1에 의한 유사도 계산 방법이 가장 좋은 성능을 나타내는 것을 알 수 있다. 또한 방법 0과 1, 방법 2와 3의 결과를 각각 비교해 보면 유사도를 정규화 하는 방법이 최소 신장 트리를 이용한 군집화에서 군집 순수도를 향상시

표 4. 거리 측정 방법에 따른 성능

방법	군집화 정확도	군집 순수도
방법 0	91.13%	81.01%
방법 1	96.20%	94.94%
방법 2	86.08%	82.28%
방법 3	94.94%	94.94%
방법 4	81.01%	72.15%

표 5. 방법 1에 의해 잘못 생성된 군집

군집	구성 단어
2	거기0 거기1 거기2 거기3 거기4 거기5 여기0 여기1 여기2
42	아버지0 아버지1 아버지2 아버지3 어머니0 어머니1 어머니2
45	얘기0 얘기1 얘기2 얘기3 얘기4 얘기5 얘기6 이야기1 이야기5 이야기6
51	운동0 운동1 운동2 운동3 운동4 운동5 운동6 활동0 활동1 활동2

키는데 도움이 된다는 것을 알 수 있다. 그러나, 확률 벡터를 단위 벡터로 변환하는 것은 성능 향상을 가져오지 못했는데, 이는 평균점에 대한 불일치의 합이 기하학적 거리와는 차이가 있기 때문으로 생각된다.

한편 표 5는 방법 1에 의해 생성된 군집에서 잘못된 군집으로 판명된 것들을 보여준다. 표 5에 있는 잘못된 군집을 살펴보면 이 실험의 평가 기준과 맞지 않을 뿐이지 잘못된 군집이라 하더라도 대부분 의미가 유사한 단어들로 구성되어 있다.

4. 결론 및 향후 연구

이 논문에서는 평균점에 대한 불일치의 합으로 단어 사이의 거리를 측정하고 최소 신장 트리를 이용하여 군집화 하는 자동 단어 군집화 기법을 제안하였다. 명사의 의미별 군집화 실험 결과 대부분의 군집이 사람의 직관에 크게 벗어나지 않게 구성되었음을 볼 수 있었다. 한편 이 논문에서는 군집화 기법의 성능을 평가하는 방법으로 가상 단어 군집화를 제안하였으며 가상 단어 군집화 결과에 대해 군집화 정확도와 군집 순수도를 측정하였을 때 약 91%의 군집화 정확도와 약 81%의 군집 순수도를 나타내었다. 그리고, 거리 계산 방법을 다양화해서 수행한 실험에서는 평균점에 대한 불일치의 합으로 거리 측정한 다음 이를 정규화 하는 방법이 약 96%의 군집화 정확도와 약 95%의 군집 순수도를 나타내어 가장 좋은 성능을 보였다.

앞으로의 연구에서는 향상된 군집화 알고리즘을 적용할 계획이며 제안된 방법으로 자동 구축된 단어 군집을 여러 자연어 처리 기법에 적용하여 자료 부족 문제를 해결하고 정확도를 향상시키는 작업을 함께 수행할 계획이다.

참고 문헌

[1] Dagan, Ido, Fernando Pereira, and Lillian Lee. Similarity-Based Methods for Word Sense Disambiguation, In *Proceedings of the 35th Annual Meetings of the Association for Computational Linguistics*, pages 56-63, 1997.

[2] Dempster, A.P., N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, Vol. 39, pages 1-38, 1977.

[3] Duda, Richard O. and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[4] Gibbons, Jean Dickinson. Nonparametric Measures of Association, volume 91 of *Quantitative Applications in the Social Sciences*, Sage Publications, Newberry Park, 1993.

[5] Ide, Nancy and Jean Veronis, Introduction to the Special Issue on Word Sense Disambiguation. *Computational Linguistics*, Vol. 24, No. 1, pages 1-40, 1998.

[6] Kaufman, Leonard and Peter J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. *Wiley series in Probability and Mathematical Statistics*. John Wiley and Sons, 1990.

[7] Kullback, Solomon. *Information Theory and Statistics*. John Wiley and Sons, 1959.

[8] Miller, George A. and Walter G. Charles. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, Vol. 6, No. 1, pages 1-28, 1991.

[9] Mizoguchi, R., and O. Kakusho, Hierarchical Clustering Algorithm Based on *k*-nearest Neighbors, In *Proceedings of the IEEE*, Vol. 67, No. 6, pages 930-949, 1979.

[10] Salton, Gerard. Experiments in Automatic Thesaurus Construction for Information Retrieval. In *Proceedings IFIP Congress*, pages 43-49, 1971.

[11] Zahn, C. T. Graph-Theoretical Method for Detecting and Describing Gestalt Clusters. *IEEE Transactions*, C-21, No. 3, pages 269-281, 1971.