평균점에 대한 불일치의 합을 이용한 자동 단어 군집화†

이호, 서희철, 임재창
고려대학교 컴퓨터학과
서울시 성북구 안암동 5가 1번지
우: 136-701
leeho@nlp.korea.ac.kr, hceso@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 가장 공간에 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 상위출현 비도를 갖는 약 1200개의 명사들 을 의미에 따라 군집화 하는 것이며 두 번째 실험은 이 논문에서 제시한 자동 군집화 방법의 성능을 객관적으로 평가하기 위한 것으 로 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 실험에서 는 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 중 하나는 자료 부족(data sparseness)이다. 자료 부족 문제는 학습시키기 위함 단어들의 분포적 특성을 이용하여 자동으로 단어를 군집화하는 기법으로 제안된 것이다. 이 기법은 코퍼스에서 가장 상위출현 비도를 갖는 약 1200개의 명사들을 의미에 따라 군집화 하는 것이다. 두 번째 실험은 이 논문에서 제시한 자동 군집화 방법의 성능을 객관적으로 평가하기 위한 것이며 가상 단어(pseudo word)에 대한 군집화 결과를 나타내었다. 실험 결과 이 방법은 가상 단어에 대해 약 91%의 군집화 정확도와와 약 81%의 군집 순수도를 나타내었다. 한편 두 번째 실험에서 는 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.

† 이 논문은 1998년도 과학계제 핵심기술 연구 과제 "다의의 단어 의미 중복성 해결에 관한 연구" 지원에 의한 결과임

한 대표적인 방법으로는 평균화(smoothing) 기법, 부류 기반 모형(class based model), 유사도 기반 모형(similarity based model) 등이 있다[5]. 이들 방법 중 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 이용하는 비도에 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 연구되고 있다. 일반적으로 자동 단어 군집화 기법에서는 각 단어에 대해 특성(feature)을 추출하고 이 특성에 의해 고유한 분포를 생성하여 단어 사이의 유사도(similarity) 또는 거리(distance)를 측정하여 유사도가 높은 단어들끼리 군집(cluster)을 형성한다. 이러한 군집화 방법은 단어들의 특성 벡터(feature vector)를 이용하여 유사성을 측정한다. 이 논문에서는 제안하는 군집화 기법에서는 유사한 단어들끼리 자동으로 묶어주는 것이다. 이 기법은 코퍼스에서 가장 상위출현 비도를 갖는 약 1200개의 명사들을 의미에 따라 군집화 하는 것이다. 두 번째 실험은 이 논문에서 제시한 자동 군집화 방법의 성능을 객관적으로 평가하기 위한 것이며 가상 단어(pseudo word)에 대한 군집화 결과를 나타내었다. 실험 결과 이 방법은 가상 단어에 대해 약 91%의 군집화 정확도와와 약 81%의 군집 순수도를 나타내었다. 한편 두 번째 실험에서 는 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.
2. 자동 단어 균일화

이 절에서는 본 논문에서 이용하는 특정 벡터를 추출하는 방법과 평균값에 대한 불일치의 합을 이용하여 단어의 유사도를 측정하는 방법 및 유사도를 이용하여 자동으로 단어를 균일화 하는 방법에 대해 소개한다.

2.1 특정 벡터

단어 사이의 유사도를 계산하기 위해서는 각 단어에 대한 특정 벡터를 추출하는 작업이 필요하다. 이때 사용되는 벡터들은 균일화 하는 의도에 따라 달라진다. 예를 들어 단어를 풍에 따라 균일화 하는 경우에는 단어의 풍사적 특성을 잘 반영할 수 있는 특징들이 사용되어야 하며 단어를 의미에 따라 균일화 하는 경우에는 의미에 관련된 특성을 이용하여야 한다. 이 논문에서는 의미 및 유사성을 이용하여 균일화하는 것을 목표로 한다.

논문 [8]에서는 인간이 두 단어가 의미적으로 유사한 판단할 때 그 단어들이 사용된 문장의 유사성을 살펴본다는 것을 증명하였다. 따라서 문장에 나타나는 단어들은 특정으로 이용하면 단어의 의미에 따라 균일화 할 수 있다. 특히 명사와 그 명사를 지배하는 동사 사이에는 밀접한 관계가 있기 때문에 명사를 균일화 할 때에는 동사가, 동사에 균일화 할 때에는 명사가 좋은 특성이 될 수 있다. 이러한 성질을 이용하여 이 논문에서는 명사에 대한 균일화를 할 때 문장에 나타나는 모든 단어를 대신 명사로 부착된 조사의 격과 그 명사를 지배하는 동사만을 이용하여 만들어진 각 동사 성을 특징으로 사용한다.

각 명사에 대해 포프스트로부터 모든 <격, 동사> 쌍에 대한 출현 빈도를 구해내기 위해서는 간단한 구문 분석기가 필요하지만 본 논문에서는 구문 분석 정보를 이용하지 않고 단문 분리가 가능한 문장만 단문으로 분리하여 다음 여거에 <격, 동사> 정보를 추출하여 균일화하는 작업을 한다. 격을 결정하는 과정에서는 주격, 목적격, 보조 격가 사용되었을 경우에는 격 정보를 그대로 사용하지만 부사격 격의 경우에는 조사의 형태에 따라 세부적인 격을 정확히 판별할 수 있는 것이나 이는 사용하고 나머지 부사격 조사나 보조자는 고려 대상에서 제외한다.

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

2.2 유사도 측정 방법


문자의 집합 \(W = \{w_1, w_2, \ldots, w_n\}\) 와 각 문자 \(w_i\)에 대해 \(<격, 동사>\) 쌍의 출현 확률 벡터 \(p_i = (p_{i1}, p_{i2}, \ldots, p_{im})\)가 주어졌을 때 각 방법의 유사도 계산식은 다음과 같다. \(n\)은 명사의 종류 수, \(m\)은 격의 종류 수, \(p_{ij}\)는 명사 \(w_i\)가 나타난 문장에서 \(j\)번째 <격, 동사> 쌍 \(v_j\)가 나타나는 확률을 의미한다.

- \(L_1\) norm

\[L_1(w, w_i) = \sum_{k=1}^{m} |p_{ik} - p_{ik}|\]

- \(L_2\) norm

\[L_2(w, w_i) = \sqrt{\sum_{k=1}^{m} (p_{ik} - p_{ik})^2}\]

- cosine coefficient

\[\cos (w, w_i) = \frac{\sum_{k=1}^{m} p_{ik} p_{ik}}{||p_i|| ||p||}\]

- \(\tau\) coefficient

\[\tau(w, w_i) = \frac{2}{m(m-1)} \sum_{k=0}^{m} \sum_{k=1}^{m} \frac{p_{ik} - p_{ik} \cdot p_{ik} - p_{ik}}{||p_i|| ||p||}\]

- KL divergence

\[D(w_i|w) = \sum_{k=1}^{m} p_{ik} \log \frac{p_{ik}}{p_{ik}}\]

- total divergence to the average

\[A(w_i|w) = D\left(\frac{w_i + w}{2}\right) + D\left(\frac{w_i + w}{2}\right)\]

\[= \sum_{k=1}^{m} \left[p_{ik} \log \frac{2p_{ik}}{p_{ik} + p_{ik}} + p_{ik} \log \frac{2p_{ik}}{p_{ik} + p_{ik}}\right]\]

\[= 2 \log 2 - H(w_i) - H(w_i) + H(w_i + w_i)\]

\[= [0, 2 \log 2]\]

이 논문에서는 이들 여러 가지 방법 중에서 평균 유사도 함수를 계산하는데 사용한다.
경에 대한 불필요함을 할당하여 단어 사이의 거리를 계산한다. KL divergence와 비교할 때 이 방법은 몇 가지 장점을 가지고 있다. 첫째로, KL divergence를 이용하여 구한 거리는 0~∞ 사이의 범위를 가지지만 평균점에 대한 불필요한 카바스의 합에 의해 구한 거리는 0~2log2 사이의 고정
된 범위의 값만을 가진다. 둘째로, KL divergence
에서는 \( p_k \)의 값이 0일 경우 값을 계산할 수 없기 때문에 평균화가 부가적으로 필요하다. 그러나
평균점에 대한 불필요한 합을 이용할 때에는 \( p_k \)
이나 \( p_n \)가 0일 경우에도 아무런 문제가 발생하지
않는다.

2.3 최소 신장 트리를 이용한 군집화

단어 사이의 유사도나 거리가 주어졌을 때 단어
t들을 군집화 하는 알고리즘에는 간접적 군집화
(aggregate clustering)[10], k-means 알고리즘
[3], expectation maximization(EM) 알고리즘[2], k-nearest neighbor 알고리즘[9], 최소 신장

이 논문에서는 이를 방법 중에서 최소 신장 트리
을 이용한 군집화 방법을 이용한다. 최소 신장 트
리를 이용한 군집화 알고리즘에서는 각 단어들을
전치 연 결 그래프 상에서의 정점(node)으로, 단어
사이의 거리를 두 정점을 이루는 가중치(weight)로
간주한다. 이 방법에서는 군집
화 대상 단어들과 특정 벡터로부터 추출된 거리
를 이용하여 최소 신장 트리를 생성한 다음 각
군집이 가장 높은(즉, 거리가 가장 먼) k-1 개의
간선을 제거하여 k개의 군집을 생성한다. 최소
신장 트리를 이용한 군집화 방법은 유사도 해제
완료 없이 전층에 계산 복잡도가 낮은 경점
을 가지고 있지만 하나의 거리만의 군집이며, 단어
만으로 만든 거리는 군집을 생성할 수 있는
단위를 가지고 있다. 본 논문에서는 이러한 문제점을 완
복시키기 위해 군집의 최대 크기를 제한하였다.

3. 실험 및 평가

이 장에서는 2 장에서 소개한 유사도 계산 방법
과 군집화 알고리즘에 의해 두 가지의 군집화 실험을 수행한다. 첫 번째 실험은 군집화 대상 명사
가 나타난 문장에서 \( \text{여행}, \text{동사} \) > 능동사가 나타난 조
전복 횡단을 이용하여 명사를 군집화 하는 것이며
두 번째 실험은 단어의 단위에 대한 군집화를
해제한 군집화 기법의 성능을 평가하는 것
이다.

3.1 실험 1: 명사의 의미별 군집화

이 실험에서는 명사를 의미에 따라 자동적으로
군집화 한다. 일반적으로 의미가 유사한 명사들은
동사들과의 관계가 서로 유사하다. 그리고, 한국어
에서는 명사에 부착된 조사의 경향을 이용하여
명사와 동사 사이의 관계를 파악할 수 있기 때문에
\( \text{여행}, \text{동사} \) 등은 명사를 의미별로 군집화 하는
작업에서 특별히 유용하다.

<table>
<thead>
<tr>
<th>군집</th>
<th>단어</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>가계 감독 교실 택 도서관, 부언 술집 내시 식당 화장실</td>
</tr>
<tr>
<td>4</td>
<td>가까이 근처 동네 둘러 둘러, 사찰 입구 주변 주택 지역</td>
</tr>
<tr>
<td>6</td>
<td>가량 매주 사례 상자 술안주 사례 주민 모임</td>
</tr>
<tr>
<td>8</td>
<td>가스 기분 난방 물건, 자동 세탁 불수</td>
</tr>
<tr>
<td>10</td>
<td>가습기 짐가 높은 마음 속 머릿속</td>
</tr>
<tr>
<td>12</td>
<td>가을 겨울 밤 방학 봉</td>
</tr>
<tr>
<td>14</td>
<td>가족 앞 새끼 식구 하우</td>
</tr>
<tr>
<td>16</td>
<td>야도 자녀 자식 주민 환자</td>
</tr>
<tr>
<td>18</td>
<td>외 교재 참조 일대 행</td>
</tr>
<tr>
<td>20</td>
<td>장부 관리 문단 품목</td>
</tr>
</tbody>
</table>

이 실험에서는 약 800만 어절 크기의 코퍼스에서
출현 번도가 10회 이상인 1264 개의 명사를 추출한 다음 21절에서 설명한 방법으로 30회
어석어의 번도만 가지는 9524 개의 \( \text{여행}, \text{동사} \) > 등
를 추출하여 구성한 특정 벡터를 이용하여 군
집화를 수행하였다. 이때 생성된 전체 군집의 최
대 크기는 10 단어로, 최대 15개는 1.7log2로 제
한하였다. 그 결과 235 개의 군집이 생성되었다.

표 1은 이 실험을 통해 얻어진 군집화 결과의
립니다. 표 1을 살펴보면 대부분의 군집이 사람
의 의관에 일치되는 것을 알 수 있다. 군집화가
잘못된 경우에는 주로 명사, 부사성 명사, 물질
명사 등은 대부분 같은 부류의 명사들과 군집을
이름을 이루고 있음을 알 수 있다. 표 1의 실험 결과에서
군집 6에 "사례"나 "예", "사각"이 포함된 것은 군집
6에 포함된 명사들이 "들다"는 동사의 목적어로
사용되는 특성이 강한 점이 크게 작용한 결과이
다. 이들 "들다"의 의미는 목적어에 따라 여러
가지로 나타낼 수 있지만 이 실험에 사용된 코
퍼스에서는 의미 정보가 포함되어 있지 않기 때
문에 이와 같은 오류가 발생한 것이며 이 오류는
의미 정보가 제공되면 해결될 수 있다.

표 2는 비교적 균일한 개수를 나타낸다. 표 2를
보면 비교가 1 단어 혹은 10 단어간의 크기가 많음
을 볼 수 있는데 이는 최소 신장 트리를 이용한
방법의 특성 때문인 것으로 추정된다. 하지만 군
집화 되지 않은 단어의 수가 104개로 전체 단어
의 10% 미만으로 적을 뿐 아니라 이들은 중심
(centrifol)까지의 거리가 가장 가까운 군집에 재
합당 시킬 수도 있기 때문에 큰 문제가 되지는
않는다.

한편, 최소 신장 트리를 이용한 군집화 기법에서
표 2. 크기별 군집의 개수

<table>
<thead>
<tr>
<th>군집 크기</th>
<th>군집 개수</th>
<th>군집 크기 × 군집 개수</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>101</td>
<td>1010</td>
</tr>
</tbody>
</table>

는 단어들이 하나의 거대 군집으로 집중되는 경우가 종종 발생하는데 이는 단어별 평균 유사도가 불균등하게 분포되어 있을 때 특히 잘 나타난다. 그림 1은 평균점에 따른 분리하는 방법으로 거리가 측정하였을 때 단어의 변화에 따른 그 단어가 다른 단어 사이의 평균 거리 값을 나타낸다. 그림 1에서 보듯이 단어의 변동가 높음수록 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 단어의 구조를 가져왔는데 이와 같은 현상은 그림 1에서 보여지듯이 거리 측정 방법의 특성에 그 원인이 있음은

3.2 설명 2: 가상 단어의 군집화

군집화 기법의 성능을 평가할 때 군집화 결과 중에서 어떤 단어나 군집이 정확하고 어떤 것이 잘못되었는지를 직접 조사하기는 어렵기 때문에 군집화 기법의 성능 평가는 주로 간결적인 방법으로 이루어진다. 군집화 기법의 성능 평가에 일반적으로 많이 사용되는 방법은 군집화에 의해 정의가 얼마나 손실되었는지를 살펴보는 것이다. 다시 말해, 단어 단위로 측정된 정밀도와 군집으로 변환되어 측정된 정밀도의 차이가 적응수록 더 좋은 성능을 가진 군집화 기법으로 평가할 수 있다. 하지만 이런 평가 방법은 두 개 이상의 군집화 기법 사이의 상대적인 비교만 가능할 뿐만 아니라 정보의 손실이 적다고 해서 비가정한 군집화 방법이라도 보장하지 못한다. 이러한 이유 때문에 이 실험에서는 가상으로 생성한 단어를 군집화 함으로써 간결적인 평가를 수행하고자 한다. 이 실험에서 사용되는 가상 단어는 다음과 같이 생성된다. 먼저 크리스마스에서 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, '계획1', '계획2', '계획3'이라는 가상 단어를 생성시킨다. 가상 단어를 생성하면 크리스마스에서 원래 단어가 출현한 적이 기록에 대해 그 단어의 가상 단어들 중에서 임의로 선택한 하나로 대체한다. 이 작업을 수행하고 나면 크리스마스부터 가상 단어에 대한 특성 벡터를 추출할 수 있다. 따라서, 이후 과정에서는 실험 1에서의 방법과 동일하게 가상 단어에 대해 군집화를 수행하고 나서 동일한 단어로부터 생성된 가상 단어들이 같은 군집에 포함되는지를 조사하면 군집화의 성능을 알아볼 수 있다. 여기서 출현 빈도가 1,000~3,000 범위에 비례적 고빈도에 속하는 단어를 선택한 이유는 출현 빈도가 너무 낮은 단어일 경우 여러 개의 가상 단어로 나누었을 때 특정 벡터가 밝혀지지 않는 군집화 결과에 부작용하기 때문이다. 이 연구에서는 가상 단어의 군집화 결과에 대해 군집화 정확도와 군집 순수도의 측면에서 평가를 수행한다. 군집화 정확도는 전체 단어에 대해 정확히 군집화된 단어의 비율이며 군집 순수도는 전체 군집에 대해 올바른 군집의 비율이다. 여기서 정확한 군집화된 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 지표는 수식으로 표현하면 다음과 같다.

군집화 정확도 = \( \text{정확한 군집화된 단어의 수} \times 100\% \)

군집 순수도 = \( \text{올바른 군집의 수} \times 100\% \)

출현 빈도가 1,000~3,000 범위에 속하는 79개의 단어를 44개의 가상 단어로 나누는 다음 전체 가상 단어를 79개의 군집으로 만들어 군집화의 최대 단어의 수를 10 단어로 제한한 결과와 91.33%의 군집화 정확도와 81.01%의 군집 순수도를 나타내었다.

그림 1. 단어의 변화와 평균 거리의 관계

- 422 -
표 3. 잘못 생성된 군집

<table>
<thead>
<tr>
<th>군집</th>
<th>구성 단어</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>거기0, 거기1, 거기2, 거기3, 거기4, 거기5</td>
</tr>
<tr>
<td></td>
<td>여기0, 여기1, 여기2</td>
</tr>
<tr>
<td>4</td>
<td>결과0, 결과1, 결과2</td>
</tr>
<tr>
<td></td>
<td>현상0, 현상1, 현상2, 현상3, 현상4, 현상5</td>
</tr>
<tr>
<td>6</td>
<td>계획0, 계획1, 계획2</td>
</tr>
<tr>
<td></td>
<td>내용0, 내용1, 내용2, 내용3, 내용4, 내용5</td>
</tr>
<tr>
<td>7</td>
<td>고개0, 고개1, 고개2, 고개3, 고개4, 고개5</td>
</tr>
<tr>
<td></td>
<td>고개6, 머리0, 머리1</td>
</tr>
<tr>
<td>8</td>
<td>과정0, 과정1</td>
</tr>
<tr>
<td></td>
<td>상황0, 상황1, 상황2, 상황3, 상황4, 상황5</td>
</tr>
<tr>
<td>9</td>
<td>관계0, 관계1, 관계2, 관계3, 관계4, 관계5</td>
</tr>
<tr>
<td></td>
<td>의미0, 의미1, 의미2, 의미3, 의미4, 의미5</td>
</tr>
<tr>
<td>11</td>
<td>기능0, 기능1, 기능2, 기능3, 기능4, 기능5</td>
</tr>
<tr>
<td></td>
<td>무어0, 무어1, 무어2, 무어3, 무어4, 무어5</td>
</tr>
<tr>
<td>23</td>
<td>방식0, 방식1, 방식2, 방식3, 방식4</td>
</tr>
<tr>
<td></td>
<td>방향0, 방향1, 방향2, 방향3, 방향4</td>
</tr>
<tr>
<td>28</td>
<td>사회0, 사회1, 사회2</td>
</tr>
<tr>
<td></td>
<td>세계0, 세계1</td>
</tr>
<tr>
<td>30</td>
<td>생활0, 생활1, 생활2, 생활3, 생활4, 생활5</td>
</tr>
<tr>
<td></td>
<td>역할0, 역할1, 역할2, 역할3</td>
</tr>
<tr>
<td>31</td>
<td>서울0, 서울1, 학교0, 학교1, 학교2, 학교3</td>
</tr>
<tr>
<td></td>
<td>학교4, 학교5, 학교6</td>
</tr>
<tr>
<td>36</td>
<td>아버지0, 아버지1, 아버지2, 아버지3</td>
</tr>
<tr>
<td></td>
<td>아호, 아호0, 아호1, 아호2, 아호3, 아호4</td>
</tr>
<tr>
<td>37</td>
<td>예기0, 예기1, 예기2, 예기3, 예기4, 예기5</td>
</tr>
<tr>
<td></td>
<td>예기6, 예기7, 예기8, 예기9</td>
</tr>
<tr>
<td>41</td>
<td>운동0, 운동1, 운동2, 운동3, 운동4, 운동5</td>
</tr>
<tr>
<td></td>
<td>운동6, 운동7, 운동8, 운동9</td>
</tr>
<tr>
<td>47</td>
<td>이유0, 이유1, 이유2</td>
</tr>
<tr>
<td></td>
<td>필요0, 필요1, 필요2, 필요3, 필요4</td>
</tr>
</tbody>
</table>

단어들의 군집이 두 개 혹은 그 이상으로 분리되며, 구분화 가이드가 바로 잘 나와 있는 경우로 분류한 군집화 정확도는 실제의 길이의 점화도는 실험 결과보다 더 높은 것으로 나타났다.

표 4. 거리 측정 방법에 따른 성능

<table>
<thead>
<tr>
<th>방법</th>
<th>군집화 정확도</th>
<th>군집 순수도</th>
</tr>
</thead>
<tbody>
<tr>
<td>방법 0</td>
<td>91.13%</td>
<td>81.01%</td>
</tr>
<tr>
<td>방법 1</td>
<td>96.20%</td>
<td>94.94%</td>
</tr>
<tr>
<td>방법 2</td>
<td>86.08%</td>
<td>82.28%</td>
</tr>
<tr>
<td>방법 3</td>
<td>94.94%</td>
<td>94.94%</td>
</tr>
<tr>
<td>방법 4</td>
<td>81.01%</td>
<td>72.15%</td>
</tr>
</tbody>
</table>

규 분포로 변환하면 단어의 빈도에 상관없이 평균 거리가 0이 되지만, 거리 행렬이 대칭이 되지 않기 때문에 이들의 합을 이용하였다. 한편 두 변 할 방법은 특정 텍스트의 크기 1인 단위 텍스트를 변환하여 기존의 방법을 적용하는 것이다. 기하학 적 거리 관점에서 보면 두 홀름 벡터가 같은 각 도를 이루고 있어도 최우거절 벡터가 더 많다. 그러나 이들 단위 벡터로 변환하면 두 벡터 사이의 각도의 전위 순서와 거리 순서가 일치하기 때문에 이 변환은 허용을 지니 있다. 마지막으로 제 2번째 방법은 앞의 두 방법을 같이 이용하는 것이다. 이들 변환을 총으로 정의하면 다음과 같다.

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

이 세 가지 변환을 적용한 방법 및 cosine coefficient를 구한 단어군집화에 적용한 결과가 표 4에 나타난다.

이 세 가지 변환을 적용한 방법 및 cosine coefficient를 구한 단어군집화에 적용한 결과가 표 4에 나타난다. 다시 변환 0은 평균정치에 대한 몰입지의 경계를 의미하며, 방법 1 3은 변환 1 3을 적용한 경우, 방법 4는 cosine coefficient를 이용하는 경우를 나타낸다.

표 5. 범위 1에 의해 잘못 생성된 군집

<table>
<thead>
<tr>
<th>군집</th>
<th>구성 단어</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>거기0, 거기1, 거기2, 거기3, 거기4, 거기5</td>
</tr>
<tr>
<td></td>
<td>여기0, 여기1, 여기2</td>
</tr>
<tr>
<td>42</td>
<td>아버지0, 아버지1, 아버지2, 아버지3</td>
</tr>
<tr>
<td></td>
<td>아호0, 아호1, 아호2, 아호3</td>
</tr>
<tr>
<td>45</td>
<td>예기0, 예기1, 예기2, 예기3, 예기4, 예기5</td>
</tr>
<tr>
<td></td>
<td>예기6, 예기7, 예기8, 예기9</td>
</tr>
<tr>
<td>51</td>
<td>운동0, 운동1, 운동2, 운동3, 운동4, 운동5</td>
</tr>
<tr>
<td></td>
<td>운동6, 운동7, 운동8, 운동9</td>
</tr>
</tbody>
</table>

- 423 -
(제 10회 한국 및 한국어 정보처리 학술대회)

키는데 도움이 된다는 것을 알 수 있다. 그러나, 확률 백터를 단위 백터로 변환하는 것은 성능 향상을 가져오지 못했다. 이는 평균값에 대한 불일치의 함이 기하학적 거리와 차이가 있기 때문으로 생각된다.

한편 표 5는 방법 1에 의해 생성된 군집에서 잘 맞춘 군집으로 판명된 것을 보여준다. 표 5에 있는 잘못된 군집을 살펴보면 이 실험의 평가 기준과 맞지 않게 잘못된 군집이 비리리도 대부분 의미가 유사한 단어들로 구성되어 있음으로.

4. 결론 및 향후 연구

이 논문에서는 평균값에 대한 불일치의 함으로 단어 사이의 거리를 측정하고 최소 신경 트리를 이용하여 군집화 하는 자동 단어 군집화 방법을 제안하였다. 명사의 의미별 군집화 실험 결과 대부분의 군집이 사람의 직관에 크게 받아들이지 않게 구성되어 있음을 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 평균값에 대한 불일치의 함으로 거리 측정한 다음 이 온 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