단어클러스터링을 이용한 동사 어휘의미망의 활용 및 평가

김 혜 경, 윤 애 선
부산대학교 인지과학협동과정
haegyungk@gmail.com / asyoon@pusan.ac.kr

The Application and Evaluation of Verbal Lexical-Semantic Network Using Automatic Word Clustering

Hae-Gyung KIM / Aesun YOON
Pusan National University
haegyungk@gmail.com / asyoon@puan.ac.kr

요 약

최근 수년간 한국어를 위한 어휘의미망에 대한 관심은 꾸준히 높아지고 있지만, 그 결과물을 어떻게 평가하고 활용할 것인가에 대한 방안은 이루어지지 않고 있다. 본 논문에서는 단어클러스터링 시스템 개발을 통하여, 어휘의미망에 의해 확장되기 전후의 클러스터링을 수행하여 데이터를 서로 비교하였다. 단어클러스터링 시스템 개발을 위해 사용된 학습 데이터는 신문 말뭉치 기사로 총 68,455,856 어절 규모이며, 특성벡터와 벡터공간모델을 이용하여 시스템A를 완성하였다. 시스템B는 구축된 '[-하]동사류' 3,656개의 어휘의미를 포함하는 동사어휘의미망을 포함하여 확장된 것으로 확장대상정보를 선택하여 특성벡터를 재구성한다. 대상이 되는 실험 데이터는 '다국어 어휘의미망-코어넷' 으로 클러스터링 결과 나타난 어휘들의 세 번째 층위까지의 노드 동일성 여부로 정확률 검수를 하였다. 같은 환경에서 시스템A와 시스템B를 비교한 결과 단어클러스터링의 정확률이 45.3%에서 46.6%로의 향상을 보였다. 향후 연구는 어휘의미망을 활용하여 좀 더 다양한 시스템에 체계적이고 폭넓은 평가를 통해 전산시스템의 향상은 물론, 연구되고 있는 많은 어휘의미망에 의미 있는 평가 방안을 확대시켜 나가야 할 것이다.

1. 서론

최근 들어 어휘의미망(Lexical-Semantic Network)에 대한 관심이 높아지고 있다. 외국에서는 물론, 국내에서도 다양한 방법으로 이에 대한 연구가 진행되고 있으며, 다년간의 연구를 바탕으로 한 시스템 개발이 활발히 이루어지고 있다. 대표적인 것으로는 미국의 프린스턴(Princeton) 대학에서 영어를 대상으로 구축한 '워드넷(Wordnet)' [1]이 있으며, 유럽에서 이 워드넷의 1.5 버전을 모형으로 유럽 8개 국어를 대상으로 구축된 다국어 어휘의미망인 '유로워드넷(EuroWordNet)' [2]이 있다. 아시아에서는 중국의 '하

1) 1980년대 초부터 미국 프린스턴 대학(Princeton University)의 인지심리학자인 밀러(Georges A. Miller)가 주축이 되어 영어를 대상으로 구축하기 시작한 어휘의미망으로, 현재 2.1버전이 웹서비스(http://wordnet.princeton.edu/)되고 있다.

2) 유럽공동체(European Community)의 언어 정보화 계획의 일환으로 1996년부터 1999년까지 진행되었으며, 프린스턴

우넷(HowNet)' [3]과 일본 NTT사(Nippon Telegraph Telephone Corporation: 일본 전신 전화사)의 '어휘대계' [4]를 들 수 있다. 국내에서도 한국어와 일본어, 중국어의 다국어에 기반하여 KAIST KORTERM에서 구축한 '다국어 어휘의미망 - 코어넷' 과 워드넷 2.0 버젼을 바탕으로 한국어에 맞게 수정하고 보완되어 구축중인 부산대의 'KorLex' , 한국어사전에서 상위어 개념을 자동으로 추출하고 이를 이용하여 의미 계층 구조를 만들어낸 울산대의 'U-WIN(UOU-Word Intelligent Network)' 등이 있다.

. 이러한 어휘의미망들이 외국과 마찬가지로 국내에서도 한국어를 대상으로 꾸준히 구축되고 공개되어 가고 있는 것은 사실이나 아직 만들어진 어휘의미망에 대한 활용 및 평가 방식에 대한 연구는 전무한 것이 사실이다.

어떤 하나의 어휘의미망이 올바른 나무구조(tree-structure)를 이루고 있는지를 평가하는 방법은 크게 두 가지로 나누어 볼 수 있다. 하나는 어휘의미망 자체에 대한 평가, 즉 어휘의미망 내부의 각각의 개념명이라든지, 개념명의 위치라든지, 그것의 상하 관계에 대해 전문가 집단이 재편성되어 끊임없이 서로 비교하여(cross-checking) 옳고 그름에 대한 결론을 내는 방식이다. 다른 하나는 만들어진 어휘의미망을 또 다른 자연언어처리 시스템에 활용하여 활용된 자연언어처리 시스템의 성능을 향상시킨 실험 결과에 따라 어휘의미망에 대한 가치를 평가하는 방식이다.

전자의 방식은 개념명 자체에 대한 평가이므로 객관성 부여에 많은 어려움이 있다. 개념명이라는 것의 성질 자체가 전문가들의 의해서도 그 하나하나의 분류 기준의 잣대를 명확히 정의내리기 어려운 것이 사실이다. 최근에는 개념명에 대한 평가 기준으로 유사도 측정과 같은 방식의 연구가 진행되기도 한다.[2] 하지만, 이 방식도 미국의 워드넷과 같이 최하위노드까지 섬세하게(fine-grained) 동의어, 반의어 등의 어휘의미관계를 분류한 어휘의미망에는 적합하나, 그 외의 특징을 나타내는 어휘의미망에는 부적합하다. 예를 들어 일본의 '어휘대계' 라든지 한국어의 '코어넷' 과 같이 상위노드(upper node) 이하 개념명의 어휘의미 그룹을 지니는 어휘의미망에서는 최하위노드(terminal node)에 속하는 각 어휘의미에 대해 유사도를 측정하는 것이 불가능하다.

다음으로, 후자의 평가 방식인 어휘의미망을 시스템에 활용하는 방식은 어휘의미망이 시스템에 적용된 이후의 시스템의 향상의 결과치를 실험을 통해 검증하여 보여주는 방식이므로 어휘의미망의 유효성에 대해 가장 명백하며(clear) 객관적으로 입증할 수 있는 방식이라 하겠다. 또한 어떠한 어휘의미망의 유형이든지 활용가능하다는 장점을 지닌다.

본 논문에서도 후자의 방식을 선택하여 구축된 동사 어휘의미망을 시스템에 활용하여 활용 전후의 성능 비교를 통한 어휘의미망에 대한 활용 및 평가 방안을 제시하고자 한다. 시스템에 활용하여 평가하고자 하는 어휘의미망은 [3]을 통해 발표된 '[-하]동사류' 어휘의미망 3,635개이며, 어휘의미망이 활용되는 시스템은 단어클러스터링 시스템이다. 2장에서는 본 논문에서 사용될 단어클러스터링 시스템에 대해 소개하고 3장에서는 '[-하]동사류' 어휘의미망을 활용하여 개선시킨 확장된 단어클러스터링 시스템에 대해 설명하고자 한다. 다음으로 4장에서는 단어클러스터링 시스템과 확장된 단어클러스터링 시스템의 성능을 비교하기 위한 데이터 실험을 소개하고 실험 결과를 비교 분석하는 과정에 대해 기술할 것이다. 마지막 5장에서는 결론 및 향후 연구에 대해 언급하겠다.

2. 단어클러스터링 시스템 A

단어클러스터링이란 용도에 따라 비슷한 특성을 갖는 단어들을 같은 클래스로 병합하는 것을 말한다.[5] 본 논문에서 사용되는 단어클러스터링의 방법으로 대용량 코퍼스를 이용하여 대상단어와 주변단어의 유사성을 기준으로 같은 집단으로 클러스터링 하는 방법을 제안한다. 즉, 단어클러스터링에 기준이 되는 특성벡터(feature vector)를 대용량 코퍼스에서 의미 공기 정보를

대학의 WordNet 1.5를 모형으로 유럽의 8개국어로의 다국어 어휘의미망으로 구축하였다.

3) 중국 과학 아카데미(Chinese Academy of Sciences)의 동진동(董振東;Zhendong DONG) 교수를 중심으로 1988년 처음 발표되기 시작한 중국어와 영어로 된 어휘의미망이다.

4) 1986년부터 1997년까지 10여 년에 걸쳐 일본 전신 전화사에서 커뮤니케이션과학기술연구소의 이케하라 사토루(Ikehara S.) 교수를 중심으로 일영 기계 번역 시스템 ALT-J/E(Automatic Language Translator-Japanese to English)의 일환으로 구축한 것이다.

이용하여 정의하고, 이 특성벡터로 벡터공간모델(Vector Space Model)[5]을 이용하여 단어 유사도를 측정하여 단어클러스터링을 수행한다.

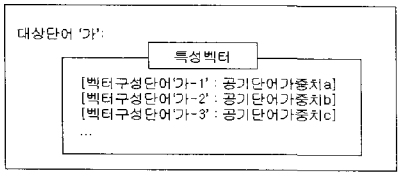
본 논문에서 사용된 학습 데이터는 1998년 5월부터 2000년 4월까지의 한국일보 기사 말뭉치로 정치, 경제, 사회, 국제, 문화 등의 전반적인 내용을 싣고 있으며, 총 68,455,856 어절 규모이다.

2.1절에서는 공기정보를 이용한 특성벡터의 구성 방법에 대해 설명하고, 2.2절에서는 2.1절에서 구성된 특성벡터를 이용하여 단어 간 유사도를 측정하는 방식과 단어클러스터링을 실행하기 위한 알고리즘에 대해 기술한다.

2.1 특성벡터의 구성 방법과 유사도 측정

단어클러스터링 시스템에서 특성벡터란 말뭉치를 분석하여 얻은 단어들 간의 여러 가지 정보를 활용하여 클러스터링을 수행할 수 있는 단어들을 말한다.

특성벡터는 말뭉치에서 공기 단어들을 추출하고 이를 단일화하여 구성한다. 즉, <그림 1>과 같이 코퍼스에서 유사도를 위해 추출한 대상단어가 있다면 그 대상단어에 대한 특성벡터는 다시 색인어인 벡터구성단어와 코퍼스에서의 공기단어 가중치(co-occurrence term weight)로 이루어진다.



<그림 1> 대상단어와 특성벡터의 구조도

각 공기 단어들의 공기단어가중치는 공기단어와 그 공기 단어의 코퍼스내에서의 출현빈도를 통해 다음과 같은 수식으로 계산할 수 있다.

$$Term\_Weight_{i,j} = f_{i,j} \times \log \frac{N}{n_i}$$

위의 식에서 $Term_Weight_{i,j}$는 단어 i의 공기 단어 j에 대한 공기 단어 가중치가 된다. $f_{i,j}$는 단어 i와 단어 j가 함께 출현한 빈도이며, n_i는 단어 i가 전체 특성벡터에 출현한 빈도, N은 전체 특성벡터 개수가 된다.

최종 추출된 특성벡터의 개수는 총 461,153개이다.

다음으로, 단어들 간의 유사도를 판정하여 유사도가 높은 단어들 간의 클러스터링을 해 나가는데, 이 때 유사도를 판정하는 방식으로 벡터공간모델의 수식을 이용한다. 두 단어 사이의 유사도는 벡터 공간에서 두 벡터 사이의 각도에 대한 코사인 값이다. 다시 말해서 두 벡터의 내적과 같다.

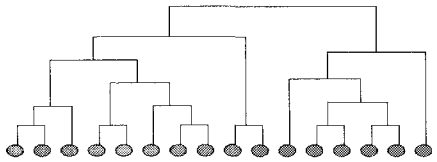
- $i = \{W_{1i}, W_{2i}, W_{3i}, \ldots, W_{ni}\}$
- $j = \{W_{1j}, W_{2j}, W_{3j}, \ldots, W_{nj}\}$

$$Sim_{i,j} = \frac{\sum_{t=1}^{n} Term\_Weight_{ti} \times Term\_Weight_{tj}}{\sqrt{\sum_{t=1}^{n} Term\_Weight_{ti}} \times \sqrt{\sum_{t=1}^{n} Term\_Weight_{tj}}}$$

위 식에서 W_{mi}는 i라는 특정벡터의 m번째 단어에 부여된 가중치를 의미하고 W_{nj}는 j라는 질의벡터의 n번째 단어의 가중치를 의미한다. $Term_Weight_{i,j}$는 단어 i의 공기 단어 j에 대한 공기 단어 가중치이며, n은 특성벡터에 출현한 모든 단어의 단일화된 개수를 뜻한다.

2.2 클러스터링 수행

클러스터링의 방법에는 크게 계층적 클러스터링과 비계층적 클러스터링이 있는데, 계층적 클러스터링의 방법이 좀 더 효율적이고 생산적인 방법으로 일반적으로 이용되고 있다[6]. 다음의 <그림 2>는 계층적 클러스터링 기법의 도식화하여 그림으로 나타낸 구조도이다.



<그림 2> 계층적 클러스터의 구조도[6]

본 논문에서는 계층적 결합 클러스터링의 한 종류인 계층적 결합 클러스터링을 사용하였다.[6] 단어 클러스터링을 위한 알고리즘은 아래 <표 1>과 같다.

클러스터링 대상 단어를 각각 개별 클러스터로 정의
while (TRUE)
begin
　현재 클러스터 중에서 가장 높은 유사도를 나타내는

5) 코넬대학의 Gerald Salton 교수가 만든 모델로서 사용자 질의와 코드를 벡터로 표현하고 이 두 벡터 사이의 유사도를 이용하여 벡터를 계산해 내는 모델이다.

```
    클러스터 C1, C2를 선택;
    if (유사도 최대값이 임계치보다 낮다)
        while문 종료;
    C2의 모든 단어를 C1에 추가;
    C2를 비움;
end;
```

<표 1> 단어클러스터링 알고리즘

클러스터 간 유사도 측정은 그룹 평균 링크(group average link) 방법[7]을 이용하였다. 즉, 두 클러스터 사이의 각 구성요소들 사이의 유사도들의 평균을 구해서, 가장 큰 값을 갖는 두 클러스터를 하나로 묶어 나가는 방법이다.

이 때 클러스터 간 유사도는 다음의 <표 2>와 같이 계산한다.

$$Sim\ (C_i, C_j) = \frac{\sum_{a=1}^{N_i} \sum_{b=1}^{N_j} Sim\ (C_{ia}, C_{jb})}{N_i \times N_j}$$

- C_i, C_j : 유사도 측정 대상 클러스터
- C_{ia} : C_i에 속하는 단어
- C_{jb} : C_j에 속하는 단어
- N_i : C_i에 속하는 단어 개수
- N_j : C_j에 속하는 단어 개수

<표 2> 클러스터 간 유사도 측정

C_i에 존재하는 모든 단어와 C_j에 존재하는 모든 단어 사이의 개별 단어간 유사도를 계산한 다음, 이를 그 경우의 수만큼 나누어서 클러스터 간 유사도를 계산한다.

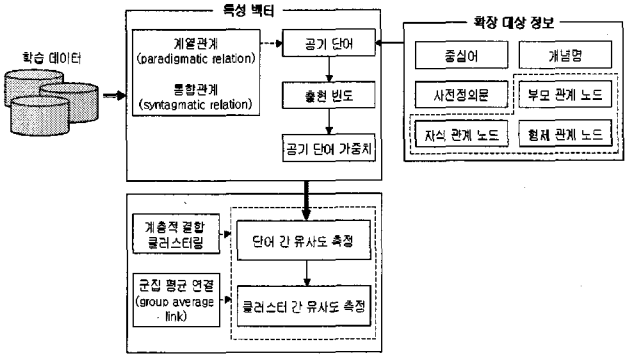
6) 계층적 클러스터링은 문서의 계층적 관계에 따라 군집화(grouping)를 시키는 방법으로써 또다시 하향식(top-down) 방법과 상향식(bottom-up) 방법으로 나눌 수 있다. 하향식 방법은 전체 단어집합을 하나의 클러스터로 보고 상이한 단어 혹은 클러스터를 분리하는 과정을 수행하는 클러스터링이다. 그리고 상향식 방법은 결합 클러스터링(agglomerative clustering)이라고도 하는데 문서 집합의 모든 문서 각각을 하나의 클러스터로 보고 유사한 클러스터를 결합함으로써 군집화를 수행하는 방법이다.[7]

7) 계층적 클러스터링의 방법에는 가장 가까운 두 개의 클러스터를 선택하는 기준에 따라 단순 링크(single link), 완전링크(complete link), 그룹 평균 링크(group average link), 워드(Ward) 기법 등이 있다. 이 중 그룹 평균 링크 방법은 클러스터 내의 모든 구성 요소가 클러스터 사이의 유사도에 영향을 주기 때문에, 느슨하게 생성되는 단순 링크 기법의 클러스터와 밀접하게 생성되는 완전 링크 기법의 클러스터의 중간 형태의 구조를 생성하게 된다. 따라서, 클러스터링 방법의 비교 연구에서 좋은 평가를 받는다.[6]

3. 확장된 단어클러스터링 시스템 B

시스템B는 시스템A의 특성벡터를 이루는 단어들 중에서 동사 어휘의미망이 포함하고 있는 단어들과 일치하는 단어가 있다면 벡터 정보에 해당 단어의 정보를 추가하여 확장한 시스템이다. 즉, 특성벡터에 포함된 단어들을 순차적으로 검색하면서 만약 확장할 동사어휘의미망에 포함된 '[-하]동사류' 의 어휘가 발견된다면, 발견된 어휘에 해당하는 확장 대상 정보들을 기존의 정보에 추가하여 특성벡터를 변화시킨다. 변화된 특성벡터로 단어클러스터링을 수행하며 그 결과 시스템B를 완성시킨다.

아래의 <그림 3>은 단어클러스터링 과정에서 시스템B로의 확장 과정을 대략적으로 보여주는 그림이다.



<그림 3> 시스템B의 단어클러스터링 과정

3.1절에서는 본 논문에서 제시하는 동사 어휘의미망에 대해 간략하게 소개하면서 확장 대상이 되는 정보에 대해 기술할 것이다. 다음으로 3.2절에서는 확장 대상 정보로 시스템 확장을 시키는 방법과 그 결과 나온 시스템B에 대해 설명할 것이다.

3.1 동사 어휘의미망

시스템B에서의 확장을 위해 활용되는 동사 어휘의미망은 '[-하]동사류' 의 어휘의미망 3,656개 항목이다. 데이터는 아래의 <그림 4>에서 보는 바와 같이 각 어휘에 대해 총 6개의 정보로 구성되어 있다.

<그림 4> '[-하]동사류'의 어휘의미망의 예

<그림 4>의 '[-하]동사류' 의 어휘의미망은 C, D와 E열의 번호는 한글학회 우리말큰사전의 표제어와 품사, 의미에 따라 구분되어 번호가 부

여된 것이며, F열과 G열은 사전정의문과 그 사전정의문을 형태소분석한 것으로 형태소분석에 사용된 태그는 카이스트 태그[8]이다. H열과 I열의 중심어는 동사가 지닌 F열과 G열의 사전정의문을 분석하여 부여한 것이다.[9]

J열과 K열의 개념명과 개념번호는 F열과 G열의 중심어에 부합하는 어휘의미망의 개념명과 그에 대한 개념번호를 부여한 것으로 KORTERM에서 개발한 '코어넷' 과 그 명칭과 번호를 공유한다.

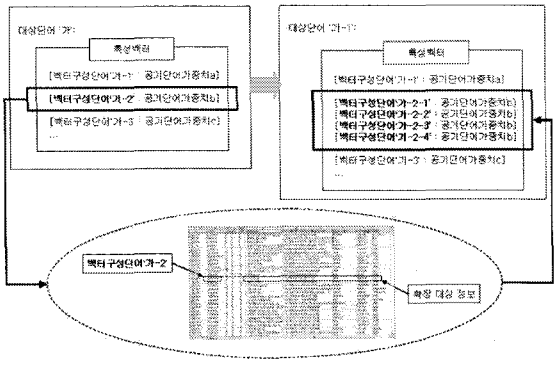
3.2 시스템B의 확장 방법

3.1에서 제시된 정보에 따라 시스템B에서는 시스템A가 지니고 있던 특성벡터를 구성하는 단어에 대해 아래의 <표 3>과 같은 확장을 위한 정보들을 추가한다. <표 3>의 첫 번째 열은 3.1에서 기술한 '[-하]동사류' 가 지니고 있는 정보들 중에서 확장을 위해 사용되는 확장 대상 정보들이며, 두 번째 열은 이러한 확장 대상 정보들이 시스템B에서 사용된 총 개수이다.

확장 대상 정보		사용된 갯수
'[-하]동사류'		1,463개 어휘(2,269개 어휘의미)
관련어	중심어	2,417개
	개념명	2,420개
사전정의문		2,269개
부모 관계 노드		19개
자식 관계 노드		75개
형제 관계 노드		1,672개

<표 3> 확장 대상 정보

확장 방법은 <그림 5>와 같이 먼저 시스템A의 특성벡터의 구성요소인 벡터 구성 단어, 즉 각 공기 단어들을 순차적으로 검색하면서 만약 '[-하]동사류' 가 발견된다면, 이에 해당하는 확장 대상 정보들을 기존의 정보에 추가한다.



<그림 5> 시스템B의 확장 방법

추가된 정보를 포함하여 다시 특성벡터를 재구성하고 이로써 2의 (그림 1)에서 기술된 단어클러스터링의 과정을 반복하여 확장된 시스템B를 이룬다.

4. 성능 비교 실험 및 결과 분석

실험은 2에서 소개된 시스템A와 3에서 소개된 시스템B의 두 버전을 같은 데이터로 단어클러스터링을 수행하여 변화된 결과를 비교 분석한다. 본 논문에서는 실험 대상이 되는 데이터로 2005년도에 KORTERM에서 출판한 '코어넷' 에 사용된 총 23,972개의 어휘(59,698개 어휘의미)를 대상으로 한다.

단어클러스터링된 어휘 그룹이 어휘의미망에서 하나의 상위노드를 지니는 지를 검증하는 것으로 결과물을 분석하고자 한다.

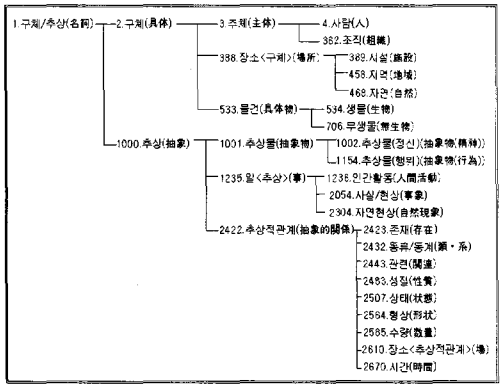
실험은 노드 정보를 완전히 없앤 어휘의미망에 사용된 단어를 사용하여 시스템A와 시스템B 각각으로 클러스터링하고 그 결과의 정확률을 어휘의미망의 노드 정보로 비교하여 판정한다.

4.1 비교 방식

시스템A로 코어넷을 클러스터링 한 결과 총 1,471개의 클러스터가, 시스템B로는 총 1,751개의 클러스터 결과물이 나왔다. 클러스터 내의 단어는 최소 2개가 하나의 클러스터를 이루는 것부터 많게는 두 시스템에서 공히 17개가 하나의 클러스터를 이루는 것까지 다양한 모습을 나타내었다.

검증을 위해 사용된 어휘의미망 코어넷은 총 12층위의 노드로 이루어져 있으며 총 2,937개의 개념명을 지니고 있다. 어휘는 이 개념명 각각에 유의어 관계로 그룹화 되어 있다. 코어넷의 상위 4층위까지의 노드에 대해 나무구조로 나타낸 그림은 다음의 <그림 6>과 같다.

8) '카이스트 태그' 는 한국과학기술원에서 개발된 태그셋으로 자세한 것은 박석문(2000)을 참조할 것.
9) 중심어 부여 방식에 대한 자세한 것은 [3]을 참조.



<그림 6> 코어넷의 상위 노드

클러스터의 정확도는 각 단어마다 코어넷에서의 개념번호를 서로 비교하였다. 코어넷 개념번호가 상위 세 번째 층위까지 같은 번호를 지니는지의 여부로 점수를 부여한다.

점수는 정확률을 위해 각 클러스터뿐만 아니라 클러스터 내의 단어 각각에 대해 개별적으로 배점을 하였다. 즉 클러스터 내의 각 단어는 클러스터내의 단어 수만큼 '1/n' 의 점수를 지니며, 각 클러스터가 가질 수 있는 총점은 100점이다. 예를 들어 3개의 단어가 하나의 클러스터를 이루며 나머지 다른 2개의 단어에서의 상위 세 번째 층위까지의 개념번호가 동일하지만 나머지 한 단어에서는 개념번호가 다르다면, 대상 클러스터는 총 66.7%의 점수를 지닌다.

4.2 실험 및 결과 분석

클러스터링에 사용된 단어 수는 각각 시스템A에서는 10,476개, 시스템B에서는 11,528개이다. 이 중 시스템A에서는 4,115개의 단어로 1,471개의 클러스터가 형성되고 6,361개의 단어가 클러스터되지 못하고 남았다. 시스템B에서는 5,789개의 단어로 1,751개의 클러스터가 형성되고 5,739개의 단어가 클러스터에 실패했다.

다음으로 <표 4>는 클러스터된 집단인 시스템A의 1,471개와 시스템B의 1,751개의 클러스터로 4.1에서 소개된 비교방식을 적용한 결과의 최종 정확률이다.

	시스템A	시스템B
클러스터의 정확률	45.3%	46.6%

<표 4> 시스템A와 시스템B의 클러스터 정확률

5. 결론 및 향후 연구

본 논문에서는 단어클러스터링 시스템 개발을 통하여, 어휘의미망에 의해 확장되기 전후의 클러스터링을 수행하여 데이터를 서로 비교하여 어휘의미망의 평가 및 활용 방안을 제시하였다. 클러스터링 된 그룹의 정확률을 기존의 구축된 어휘의미망의 노드와 비교함으로써 작업자의 주관적인 판단이 아닌 객관적인 데이터로 설명할 수 있었다. 어휘의미망을 적용하기 이전의 시스템에서는 45.3%의 정확률을 어휘의미망을 적용하고 난 후에는 46.6%의 단어클러스터링 시스템의 향상을 보였다. 어휘의미망의 자연언어처리 시스템에서의 활용에 관한 연구는 아직 시작 단계에 있다. 향후 연구는 어휘의미망을 활용하여 좀 더 다양한 시스템에 체계적이고 폭넓은 평가를 통해 전산시스템의 향상은 물론, 연구되고 있는 많은 어휘의미망에 의미 있는 평가 방안을 확대시켜 나가야 할 것이다.

참고 문헌

[1] 기민호, 2001, 단어클러스터링 기반 정보처리 도구 개발 기술, 정보통신부 우수신시술 지정·지원 사업 최종 보고서.

[2] 김준수, 2004, 의미정보와 시소러스를 이용한 한국어 어휘 중의성 해소 모델, 울산대학교 컴퓨터정보통신공학과 박사학위논문.

[3] 김혜경, 최기선, 윤애선, 2005, '[-하]동사류' 어휘의미망 구축을 위한 사전 정의문 분석, 한국사전학회 제 7회 학술대회 발표논문집, p.153-169.

[4] 박석문, 2000, 코퍼스 품사 태깅 매뉴얼, 한국과학기술원.

[5] 신중호, 박혁로, 이기호, 1993, 단어의 유사성 척도와 클러스터링 알고리즘, 한국 인지과학회 논문지 제 9권 제 2호.

[6] 이경순, 2001, 정보검색에서 벡터공간 검색과 클러스터 분석을 통한 문서 순위 결정 모델, 한국과학기술원 전자전산학과 박사학위논문.

[7] 조현양, 최성필, 2004, 계층적 결합형 문서 클러스터링 시스템과 복합명사 색인방법과의 연관관계 연구, 한국문헌정보학회지 제 38권 제 4호, p.179-192.

[8] 최준호, 2004, 의미적 멀티미디어 정보검색을 위한 개념간 유사도 측정 방법, 조선대학교 전자계산학과 박사학위논문.

[9] 한국과학기술원 전문용어언어공학연구센터, 2005, 다국어 어휘의미망, KAIST PRESS.

[10] Baeza-Yates, R., & Ribeiro-Neto, B., 1999,

Modern Information Retrieval, ACM Press.
[11] Fellbaum, C., 1998, Wordnet: An Electronic Lexical Database, MIT Press.
[12] Ikehara, S. et al., 1997, The Semantic System, volume 1 of Gio-Taikei -- A Japanese Lexicon. Iwanami Shoten.
[13] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, 2000, DATA MINING Methods for Knowledge Discovery, Kluwer Academic Publishers.
[14] Y.Zhao & G.Karypis, 2005, Hierarchical Clustering Algorithms for Document Datasets, Data Mining and Knowledge Discovery, 10, p.141-168.