동사 어휘의미망 평가를 위한 단어클러스터링 시스템의 활용 방안*

The Method of Using the Automatic Word Clustering System for the Evaluation of Verbal Lexical-Semantic Network

김 혜 경(Hae-Gyung Kim)**
윤 애 선(Ae-Sun Yoon)***

목 차

초 록

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 68,455,856어절 규모이며 ,특성벡터와 벡터공간모델을 이용하여 시스템A를 완성하였다. 시스템B는 구축된'[-하]동사류' 3,656개의 어휘의미를 포함하는 동사 어휘의미망을 활용하여 확장된 것으로 확장대상정보를 선택하여 특성벡터를 재구성한다 .대상이 되는 실험 데이터는 '다국어 어휘의미망-코어넷'으로 클러스터링 결과 나타난 어휘의 세 번째 층위까지의 노드 동일성 여부로 정확률을 검수하였다 .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.

ABSTRACT

For the recent several years, there has been much interest in lexical semantic network. However, it seems to be very difficult to evaluate the effectiveness and correctness of it and invent the methods for applying it into various problem domains. In order to offer the fundamental ideas about how to evaluate and utilize lexical semantic networks, we developed two automatic word clustering systems, which are called system A and system B respectively. 68,455,856 words were used to learn both systems. We compared the clustering results of system A to thoseof system B which is extended by the lexical-semantic network. The system B is extended by reconstructing the feature vectors which are used the elements of the lexical-semantic network of 3,656 '-ha' verbs. The target data is the 'multilingual Word Net-CoreNet'.When we compared the accuracy of the system A and system B, we found that system B showed the accuracy of 46.6% which is better than that of system A, 45.3%.

키워드: 어휘의미망, 시소러스, 단어클러스터링, 특성벡터, '-하'동사류
Lexical-semantic Network, Thesaurus, Automatic Word Clustering, Feature Vector, '-ha' verb

---

* 본 연구는 한국학술진흥재단의 지원(과제번호 :KRF-2005-005-J0602,과제명 :주변의 경계지형에 표출된 언어현상 재해석)'을 받아 이루어졌음을 밝힌다.
** 부산대학교 인지과학협동과정(haegyungk@gmail.com)
*** 부산대학교 불어불문학과, 인지과학협동과정 교수(asyoon@pusan.ac.kr)
논문접수일자 2006년 8월 14일
게재확정일자 2006년 9월 15일

1. 서 론

최근 들어 어휘의미망(Lexical-Semantic Network) 및 시소러스(thesaurus)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)'(Fellbaum 1998)과 유럽에서 이 워드넷의 1.5 버전을 모형으로 유럽 8개 국어를 대상으로 구축한 다국어 어휘의미망인 '유로워드넷(Euro WordNet)'(Vossen 2005) 등이 있다. 아시아에서는 일본과 중국에서 구축한 어휘의미망이 대표적이다. '하우넷(HowNet)'(Dong and Dong 2006)은 중국어와 영어를 대상으로 하여 보편적인 의미체계를 구성하고자 한 개념망이며, 일본의 NTT사(Nippon Telegraph Telephone Corporation: 일본 전신 전화사)의 '어휘대계'(Ikehara and others 1997)는 기계번역 시스템의 번역사전 중에서 일본어 의미사전에 관한 부분으로, 단어의미속성체계, 단어의미사전, 구문의미사전으로 구성되어 있다. 이러한 어휘의미망은 명사뿐만이 아니라 동사나 형용사, 부사 등의 여타 다른 품사를 이용해서 개념체계(conceptual system)를 구축하고 그 결과물을 발표하고 있다. 국내에서도 한국어와 일본어, 중국어의 다국어에 기반하여 '전문용어언어공학연구센터(KORTERM, 이하 '코텀'으로 표기함.)'에서 구축한 '다국어 어휘의미망(한국과학기술원 전문용어언어공학연구센터 2005)인 '코어넷(CoreNet)'이 있다. 또한 워드넷 2.0 버전을 바탕으로 한국어에 맞게 수정하고 보완되어 구축 중인 부산대의 'KorLex'와 한국어사전에서 상위어 개념을 자동으로 추출하고 이를 이용하여 의미 계층 구조를 만들어낸 울산대의 'U-WIN(UOU-Word Intelligent Network)'이 있다(옥철영, 2005; 최호섭, 옥철영 2002).

이러한 어휘의미망은 외국과 마찬가지로 국내에서도 한국어를 중심으로 꾸준히 구축되고 공개되어 가고 있는 것은 사실이나, 아직 만들어진 어휘의미망에 대한 실질적인 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 개별적인 개념에 대한 명칭이라든지, 개념명의 위치, 혹은 그것의 상하 관계에 대해 전문가 집단이 재편성되어 끊임없이 서로 비교하여(cross-checking) 옳고 그름에 대한 결론을 도출해 내는 방식이다. 다른 하나는 만들어진 어휘의미망을 별도의 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. 그 예로 다음 (1)과 같은 문장을 살펴보자.

(1) 농촌 아이들은 누구나 제 몫의 일을 하면서 학교에 다닌다.

이때 예문 (1)에서 쓰인 '학교'를 건물의 일종으로 취급하여야 할 것인가 아니면 일종의 교육기관으로 봐야 할 것인가 아니면 두 가지 의미를 모두 포함할 수 있는 제 3의 개념명으로 봐야 할 것인가. 이것은 언어학의 애매성(ambiguity)과도 관련된 문제로 어휘의미망을 구축할 때 사전편집가들의 끊임없는 고민으로 남는다. 개념명 자체에 대한 평가는 구축 시에 발생했던 이러한 언어학적 관점의 문제가 평가 시에도 여전히 남는다는 것이다.

이러한 문제를 극복하기 위해 최근에는 유사도 측정과 같은 객관적인 방식의 연구가 진행되기도 한다(김준수 2004: 20, 25, 28). 유사도 측정에 대한 방식은 크게 에지(Edge) 기반 측정 방법과 노드(Node) 기반 측정 방법, 의미 기반 측정 방법으로 나뉜다. 에지 기반 측정 방법은 밀도(부모 노드에서 자식노드로 미치는 링크의 총수), 계층구조에서 노드의 깊이, 링크 타입, 길이 등의 특성이 고려된다. 노드 기반 측정 방법은 대용량 코퍼스 내의 개념의 빈도를 세어 많은 빈도량이 할당된 개념은 특정 주제에 대해 매우 세부적인 정보로, 적은 빈도량은 더 일반적이고 덜 세부적인 개념으로 보는 방법이다. 의미 기반 측정 방법은 관련된 단어 의미 간의 비교나 공기(共起) 통계에 기반을 둔 유사성 측정법이다.

하지만 이러한 어휘의미망 자체 평가 방식도 미국의 워드넷과 같이 최하위노드까지 섬세하게(fine-grained) 동의어, 반의어 등의 어휘의미관계를 분류한 어휘의미망에는 적합하나, 그렇지 못한 어휘의미망에는 부적합하다. 예를 들어 일본의 '어휘대계'라든지 한국어의 '코어넷'과 같이 2,700여 개의 상위노드(upper node) 이하 개념명의 어휘의미 그룹을 지니는 어휘의미망에서는 최하위노드(terminal node)에 속하는 각 어휘의미에 대해 유사도를 측정하는 것이 불가능하다.

이 같은 경우에는 후자의 시스템에의 활용 방식을 선택하는 것이 유리한데, 이 방식은 평가 대상 어휘의미망이 시스템에 적용된 이후의 시스템 향상의 결과치를 적용되기 이전의 결과치와 비교하여 보여주는 방식이다. 따라서 어휘의미망의 유효성에 대해 가장 명백하며(clear) 객관적으로 입증할 수 있는 방식이라 하겠다. 또한 어떠한 어휘의미망의 유형이든지 활용가능하다는 장점을 지닌다.

본 논문에서는 후자의 방식을 선택하여 구축된 동사 어휘의미망을 시스템에 활용하여 활용 전후의 성능을 비교하고 이를 통한 어휘의미망에 대한 활용 및 평가 방안을 제시하고자 한다. 시스템에 활용하여 평가하고자 하는 어휘의미망은 김혜경, 윤애선 (2006)을 통해 발표된 '[-하]동사류' 어휘의미망 3,656개이며, 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.

2. 단어클러스터링 시스템 A

단어클러스터링이란 용도에 따라 비슷한 특성을 갖는 단어를 같은 클래스로 병합하는 것을 말한다(신중호, 박혁로, 이기호 1993). 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 이용하여 정의하고, 이 특성벡터로 벡터공간모델(Vector Space Model)을 이용하여 단어 유사도를 측정하여 단어클러스터링을 수행한다.

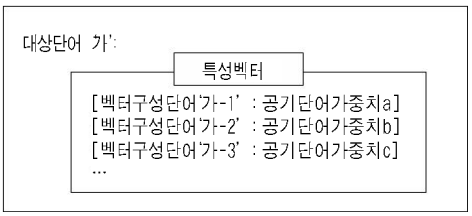
본 논문에서 사용된 학습 데이터는 1998년 5월부터 2000년 4월까지의 한국일보 기사 말뭉치로 정치, 경제, 사회, 국제, 문화 등의 전반적인 내용을 싣고 있으며, 총 68,455,856 어절 규모이다.

2.1절에서는 공기정보를 이용한 특성벡터의 구성 방법에 대해 설명하고, 2.2절에서는 2.1 절에서 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 단어 간의 여러 가지 정보를 활용하여 클러스터링을 수행할 수 있는 단어를 말한다.

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



〈그림 1〉 대상단어와 특성벡터의 구조도

각 공기 단어의 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 $Term_Weight_{ij}$는 단어 i의 공기 단어 j에 대한 공기 단어 가중치가 된다. f_{ij}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, 단어 간의 유사도를 판정하여 유사도가 높은 단어 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.

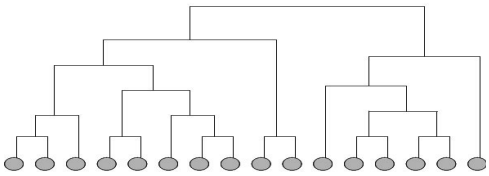
i = {w1i, w2i, w3i, … , wni}
j = {w1j, w2j, w3j, … , wnj}

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

위 식에서 wni는 i라는 특정벡터의 n번째 단어에 부여된 가중치를 의미하고 wnj는 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. 다음의 〈그림 2〉는 계층적 클러스터링 기법의 도식화하여 그림으로 나타낸 구조도이다.



〈그림 2〉 계층적 클러스터의 구조도
(이경순 2001: 26)

본 논문에서는 계층적 클러스터링의 한 종류인 계층적 결합 클러스터링을 사용하였다. 단어클러스터링을 위한 알고리즘은 아래 〈표 1〉과 같다.

클러스터 간 유사도 측정은 그룹 평균 링크(group average link) 방법을 이용하였다. 즉, 두 클러스터의 각 구성요소 사이의 유사도의 평균을 구해서, 가장 큰 값을 갖는 두 클러스터를 하나로 묶어 나가는 방법이다.

〈표 1〉 단어클러스터링 알고리즘

```
클러스터링 대상 단어를 각각 개별 클러스터로 정의
while (TRUE)
begin
    현재 클러스터 중에서 가장 높은 유사도를 나타내는 클러스터 C1, C2를 선택;
    if (유사도 최대값이 임계치보다 낮다)
        while문 종료;
    C2의 모든 단어를 C1에 추가;
    C2를 비움;
end;
```

단계 1	클러스터링 대상단어를 각각 개별 클러스터로 정의한다.
단계 2	가장 높은 유사도를 나타내는 두 클러스터를 선택하여, 하나의 클러스터로 합친다.
단계 3	새로 생성된 클러스터와 다른 클러스터사이의 유사도를 계산한다.
단계 4	유사도 최대값이 지정된 임계치(thresholds)에 도달할 때까지 반복한다.

이 때 클러스터 간 유사도는 다음의 〈표 2〉와 같이 계산한다.

〈표 2〉 클러스터 간 유사도 측정

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

- 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

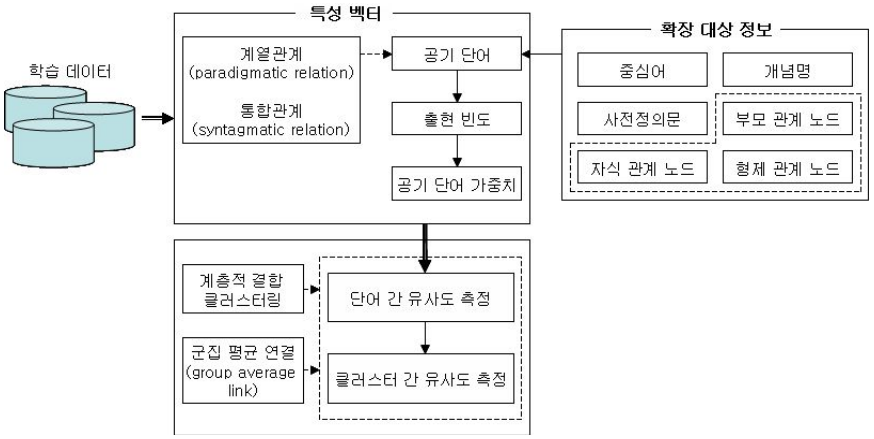
C_i에 존재하는 모든 단어와 C_j에 존재하는 모든 단어 사이의 개별 단어간 유사도를 계산한 다음, 이를 그 경우의 수만큼 나누어서 클러스터 간 유사도를 계산한다.

3. 확장된 단어클러스터링 시스템 B

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

아래의 〈그림 3〉은 단어클러스터링 과정에서 시스템B로의 확장 과정을 보여주는 그림이다.

3.1절에서는 본 논문에서 제시하는 동사 어휘의미망에 대해 간략하게 소개한다. 다음으로 3.2절에서는 시스템B의 확장을 위해 사용될 동사 어휘의미망의 확장 대상 정보에 대해 소개하고, 3.3 절에서는 동사 어휘의미망의 정보 중에서 확장 대상 정보로 시스템 확장을 시키는 방법과 그 결과 나온 시스템B에 대해 기술할 것이다.



〈그림 3〉 시스템B의 단어클러스터링 과정

3.1 동사 어휘의미망

어휘의미망은 자연언어처리 과정에서 발생할 수 있는 어휘 간의 개념으로 인한 문제를 효율적으로 극복하기 위해 구상된 체계이다. 개념체계는 특정분야, 혹은 일반분야에서 지구상에 존재하는 언어로 표현될 수 있는 다양한 개념에 명칭을 붙이고 그 개념명의 서로 유기적인 관계에 대한 설명을 나무구조 등으로 설명하는 체계이다.

본 논문의 시스템B에서의 확장을 위해 활용되는 동사 어휘의미망은 '[-하]동사류'의 어휘의미망 3,656개 항목이다.

'[-하]동사류'란 서술성명사(predicative noun)와 기능동사(support verb)로 이루어진 동사쌍이다. 이 때, 기능동사로 쓰인 '-하다'는 동작성이나 상태성의 의미를 지니는 선행 요소와는 달리, 한 문장을 이루기 위해 사용되는 형태, 통사적인 도구이다. 따라서 시제나 인칭(person), 수(number), 그리고 상(aspect) 등을 나타내는 문법적 표지 역할만을 담당한다. 예를 들어 다음의 (2)에서의 명사 '운동'과 '-하다' 간의 관계가 된다.

(2) 그 남자는 방 안에서 운동한다.

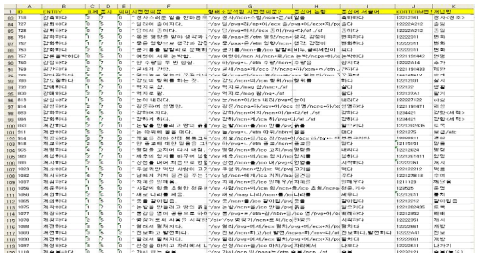
위의 (2)에서 '운동'은 동작성을 지닌 서술성명사이며 '-하다'는 서술어로서 형식적이며 문법적인 기능을 하는 기능동사이다.

3,656개 항목은 〈우리말큰사전〉에 등재된 '[-하]동사류' 35,100개 중 선별하여 최종 결정된 데이터 리스트이다. 결정 방식은 세 가지로 요약될 수 있다. 첫째, 의성어와 의태어는 개념체계의 특정 개념명 몇 군데에 집중되어 분포되므로 제외한다. 둘째, 구축하게 될 동사 어휘의미망은 현대 한국어 화자가 실질적으로 사용가능한 어휘의미망을 구현하고자 기본어휘로 국한한다. 셋째, 기구축된 어휘의미망에서 서술성명사를 통해 자동으로 개념명을 부여할 수 있는 경우는 제외한다.

최종 결정된 데이터 리스트는 중심어를 추출하고, 추출된 중심어를 기반으로 개념명과 개념번호를 부여하는 데 이용된다.[1]

3.2 확장 대상 정보

데이터는 다음의 〈그림 4〉에서 보는 바와 같이 각 어휘에 대해 총 6개의 정보로 구성되어 있다.



〈그림 4〉 '[-하]동사류'의 어휘의미망의 예

〈그림 4〉 중에서 F~K까지가 시스템B를 위한 확장대상정보로 활용된다. 〈그림 4〉의 C, D와 E열의 번호는 〈우리말큰사전〉의 표제어와 품사, 의미에 따라 구분[2]되어 번호가 부여된

1) 중심어 추출과 개념명 부여 방식에 대해서는 3.2절의 확장 대상 정보를 소개하면서 자세히 기술한다.
2) 동음이의어 구분에 대한 일련번호는 '표제어 번호'로 표기하고, 다의어 구분에 대한 일련번호는 '의미번호'로 표기한다.

형태적 제약 정보를 이용한 중심어 추출의 가장 기본이 되는 제약정보는, 〈그림 5〉의 ①과 ②로써, 사전정의문의 역방향으로 형태소를 분석하는 과정에서 제일 처음 출현하는 서술어 - '일반동사(pvg)'나 '서술성명사+동사파생접미사 쌍(ncpa+xsv)[5]' - 혹은 명사(ncn)가 그 사전정의문의 중심어가 된다는 것이다.

예를 들어, 〈표 3〉에서 보는 '망각하다/1/타동사/0'의 '사전정의문2'를 형태소 분석하고, 역방향 분석 과정을 거치게 된다. 이 과정에서 제일 처음 출현하는 서술어는 '잊다'이며, 이 '잊다'가 바로 '망각하다/1/타동사/0'의 사전정의문의 중심어가 된다.

표제어에 개념을 부여할 때 사용되는 개념명은 코텀의 '코어넷'[6]에서 구현된 2,937개의 개념명을 지닌 개념체계를 따른다. 우선 개념명

〈표 3〉 기본적인 제약 정보

표제어	사전정의문1[7]	사전정의문2	형태소 분석 결과
망각하다/1/타동사/0	잊어버리다①	죄다 잊다.	죄다/mag 잊/pvg+ 다/ef+./sf

후보를 선정하고 테스트셋(test-sets)을 통한 검증의 단계를 거쳐서 어휘의미에 부여하게 될 최종 개념명과 개념번호를 결정한다.

구축할 대상 어휘를 선택하고 중심어를 검색하고 중심어에 사용된 용어를 서술어 중심으로 '코어넷'의 기구축된 한국어 어휘목록에서 같은 용어를 검색한다. 만일 동일한 용어가 '코어넷' 한국어 어휘목록에 있고 그 의미가 중의적이지 않다면 개념명 후보로 선택하도록 한다. 다음의 〈표 4〉에서의 중심어 '걷다'와 같은 예이다.

'독보하다/1/자동사/0'의 중심어는 '걷다'이며, 2,937개의 개념명 중에서 한국어 어휘목록의 '동일한 용어 검색'을 통해 개념명 후보로 '12212141[보행〈발동작〉]'[8]과 '122221221[사라짐]'이 선택되었다. 다음으로 〈표 5〉의 테스

3) 형태소분석에 사용된 태그는 한국과학기술원에서 개발된 태그셋인 '카이스트 태그'이다. '카이스트 태그'에 대한 자세한 것은 박석문(2000)을 참조할 것.

4) 중심어란 동사 사전정의문에서 표제어의 의미를 가장 잘 표현하고 있는 논항을 포함하거나 그렇지 않은 하나의 어휘로 이루어진 단어를 일컫는다. 예를 들어 다음과 같이 〈우리말큰사전〉에 등재된 '간택하다'라는 동사 중 하나의 의미를 살펴보자.
'간택하다
사전정의문: 여럿 가운데서 특별히 가리다'
위의 사전정의문에서 '여럿 가운데서 (특별히)'는 양태소이며, '가리다'는 표제어 '간택하다'와 동의관계에 있으면서도 의미전달이 좀 더 용이한 기본어휘에 속하는 단어(general word)인 '풀이말'이다. 따라서 '간택하다'에서는 '가리다'가 중심어가 된다(김혜경, 최기선, 윤애선 2005).

5) 카이스트 형태소 분석 방식에서는 서술성의 유무에 따라 '일반동사(pvg)'의 위치에 '서술성명사+동사파생접미사 쌍(ncpa+xsv)'이 쓰인 경우가 있다. 4.3.1절의 제약 정보에서 쓰이는 '일반동사(pvg)'의 위치에는 항상 '서술성명사+동사파생접미사 쌍(ncpa+xsv)'도 같이 쓰일 수 있다.

6) '코어넷'의 개념명은 초기에 일본어 '어휘대계'의 명사의미체계 부분의 2,710개의 개념명과 노드 규칙을 한국어로 번역하는 것부터 시작되었다. 이후 한국어에 맞게 개념명을 적절하게 수정하기도 하고 추가하여 보완하기도 하여, 최종 2,937개의 개념명을 지닌 한국어 어휘의미망을 만들었다.

7) '사전정의문1'과 '사전정의문2'의 구분은 사전정의문이 명사로의 링크만을 담고 있는 경우, 명사로의 링크정보만으로 구성된 사전정의문을 '사전정의문1', 링크된 명사로 사전정의문을 다시 찾아 최종 문장 형태로 연결된 사전정의

트셋을 거치게 된다. 사용된 테스트셋은 함의(entailment)[9] 관계에 대한 질의와 응답으로 구성되어 있다.

다음 〈표 5〉에서 보는 바와 같이 '보행〈발동작〉'은 '발맞추다/1/자동사/0', '행진하다/2/자동사/0', '도보하다/1/자동사/0' 등과 더불어 '걷다'의 개념명이 된다.

동사 어휘의미망을 통해 구축된 중심어와 개념명은 단어로 제공되므로, 확장 대상 정보로 각 어휘가 지닌 정보를 그대로 사용하면 된다.

사전 정의문은 형태소 분석되어 품사 태깅된 형태로 존재하므로, 명사, 동사, 형용사 등의 실질형태소를 중심으로 추출하여 이를 확장 요소에 적용한다. 예를 들어 '정가하다'는 '값을 매기다'라는 사전정의문을 지니며, 여기서 실질형태소인 '값'과 '매기다'가 '정가하다'의 사전정의문이 지니는 확장요소에 포함된다.

〈표 4〉 '독보하다/1/자동사/0'의 중심어로 추출된 '걷다'

표제어	사전 정의문1	사전 정의문2	형태소 분석 결과	중심어
독보하다/1/자동사/0	독보2	혼자서 걸음.	혼자/ncn+서/jca 걷/pvg+음/xsn+./sf	걷다

〈표 5〉 중심어 '걷다'의 테스트셋

〈(누구 혹은 무엇이) (***걷다***)〉가 〈(누구 혹은 무엇이) [보행〈발동작〉] 하는 것이다〉는 '참'이고,
〈(누구 혹은 무엇이) [보행〈발동작〉]하는 것〉이 〈(누구 혹은 무엇이) (***걷다***)〉가 '반드시 참인 것은 아니다'이라면,

"걷다"는 개념명으로 [보행〈발동작〉]을 지닌다

부모 관계 노드, 자식 관계 노드, 혹은 형제 관계 노드는 개념번호로 검색하여 정보를 사용할 수 있다. 번호 부여 방식은 1,2,3,… 의 순서대로 번호를 부여하되, 하위 노드에 대해서는 1층위(level)씩 자리를 추가하여 번호를 부여하고 있다. 만약 1이라는 개념번호의 1층위의 하위 노드에 개념명이 두 개가 있다면 그 개념번호는 11과 12이며, 다시 11에 대해 1층위의 하위 노드에 세 개의 개념명을 지닌다면 111과 112, 113의 개념번호를 부여하는 방식이다.

3.3 시스템B의 확장 방법

문을 '사전정의문2'라 명명한다. 즉, '발전하다/1/자동사/0'는 '사전정의문1'에서 '→발전]'로 구성되어 있으며, 최종 '사전정의문2'의 정보를 통해 '더 잘 되거나 나아지거나 활발해지거나 하는 일.'이라는 정의문을 얻게 되었다.

8) '코어넷'과 번호를 공유하므로 '코어넷'과 같은 번호 부여 방식이다.

9) 워드넷에서도 함의를 이용하여 하위 관계를 정의한다. "Troponymy is a particular kind of entailment. ...Every troponym V_1 of a more general V_2 also entail V_2."

3.1에서 제시된 정보에 따라 시스템 B에서는 시스템A가 지니고 있던 특성벡터를 구성하는 단어에 대해 아래의 〈표 6〉과 같은 확장을 위한 정보를 추가한다. 〈표 6〉의 첫 번째 열은 3.1에서 기술한 '[-하]동사류'가 지니고 있는 정보 중에서 확장을 위해 사용되는 확장 대상 정보이며, 두 번째 열은 이러한 확장 대상 정보가 시스템B에서 사용된 총 개수이다.

위 〈표 6〉에서 관련어는 중심어와 개념명을 통틀어 일컫는다. 확장 시 사용되는 '[-하] 동사류'의 어휘의미망은 반드시 중심어를 지니며 그 중심어에는 개념명이 부여되어 있다. 그 개수에 있어 2,417개와 2,420개로 차이를 보이는 것은 적용된 목록은 동일하나, 하나의 어휘에 대해 두 개 이상의 중심어나 개념명을 지닌 경우이다. 다음 〈표 7〉의 '가감하다'와 같은 표제어이다.

〈표 7〉의 '가감하다/1/타동사/0'는 '더하다'와 '덜다'로 중심어의 서술어 형태가 둘 이상으로 나타난다. 서로 다른 의미의 서술어를 지닐 때는 그에 따른 서로 다른 둘 이상의 개념명을 부여하게 된다. 따라서 '가감하다/1/타동사/0'는 '더하다'와 '덜다'에 대해 '증가'와 '감소'의 개념명을 부여받았다.

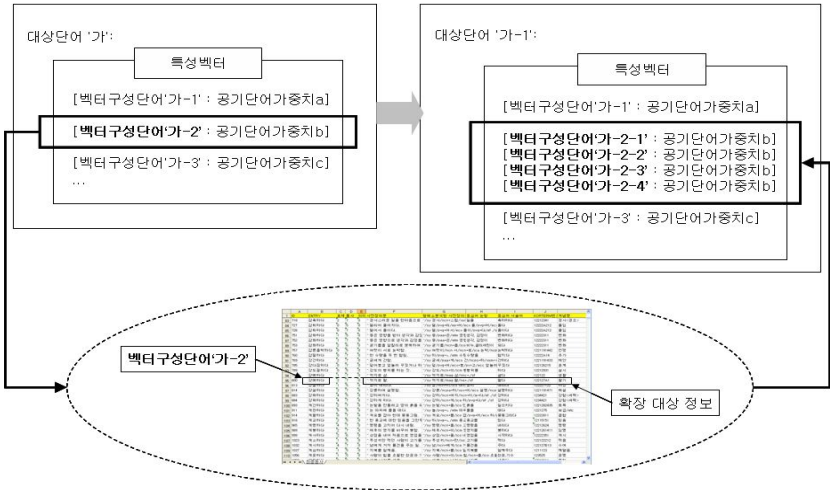
확장 방법은 〈그림 6〉과 같이 먼저 시스템A의 특성벡터의 구성요소인 벡터 구성 단어, 즉 각 공기 단어를 순차적으로 검색하면서 만약 '[-하]동사류'가 발견된다면, 이에 해당하는 확장 대상 정보를 기존의 정보에 추가한다.

〈표 6〉 확장 대상 정보의 사용된 갯수

확장 대상 정보		사용된 개수
'[-하]동사'		1,463개 어휘(2,269개 어휘의미)
사전정의문		2,269개
관련어	중심어	2,417개
	개념명	2,420개
부모 관계 노드		1개
자식 관계 노드		10개
형제 관계 노드		1,672개

〈표 7〉 둘 이상의 중심어

표제어	사전정의문	형태소 분석 결과	중심어	개념명
가감하다/1/타동사/0	더하거나 덜다.	더하/pvg+거나/ecc 덜/pvg+다/ef+./sf	더하다, 덜다	증가, 감소



〈그림 6〉 시스템B의 확장 방법

확장 대상 정보로 추출되어 추가되는 단어에 대한 공기 정보인 빈도는 확장 대상이 된 초기의 단어가 지니는 빈도로 정보를 부여한다. 즉, '[가게]'에 대한 특성벡터로 '[정가하다:387]', '[물건:298]', '[점원:219]'…가 존재하고, '정가하다'가 '[-하]동사류'에 존재한다면 '정가하다'에 대한 확장대상정보는 모두 공기 빈도가 387을 지니게 된다.

추가된 정보를 포함하여 다시 특성벡터를 재구성하고 단어클러스터링의 과정을 반복하여 확장된 시스템B를 이룬다.

〈표 8〉은 시스템A와 시스템B에서 유사한 단어클러스터링의 모양이 시스템의 향상으로 인해 어떻게 변화되어 나타나는 지를 보여주는

〈표 8〉 단어클러스터링이 향상된 일례

시스템 A	시스템 B
가공 (12111342) *원료 (113222)*	가공(12111342) 재고(12313) 생산량(12374) 소진(1223111122)
재고 (12313)	
생산량 (12374)	
소진 (1223111122)	
거듭나다(12232116) *도약 (1132231)*	거듭나다(12232116) *도모 (122117E1)* 끌어들이다(12222682) 충실 (1234111)
끌어들이다 (12222682)	
충실 (1234111)	
결정(122117932) 확정(122117B2) *열다(1221282435) 다음주(1239123)*	결정(122117932) 확정(122117B2) 일정(123927) 정하다(122117531) 늦다(1239253)
따르다(123382) 한편(11111411111) 방침(11322442) 일정(123927) 정하다(122117531) *재개 (12222351)*	
내달(1239122) 예정(122117E1) 이달(1239231) 늦다 (1239253) *미루다 (12211782)*	

일례이다.

〈표 8〉의 세 번째 클러스터에서 제시되는 예처럼 '결정', '확정' 등의 단어가 '열다', '다음주'와 같은 단어와 군집화된 시스템A에 비해 시스템B에서는 '결정', '확정', '정하다', '늦다 가 함께 군집화되어 시스템이 향상되었음을 알 수 있다.

4. 비교 실험 결과 및 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.

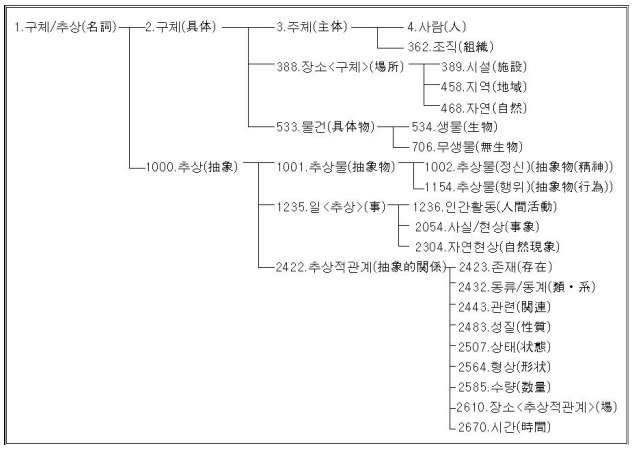
비교 방식은 단어클러스터링된 어휘 그룹이 어휘의미망에서 하나의 상위노드를 지니는 지를 검증하는 것으로 결과물을 분석하고자 한다.

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

4.1 비교 방식

시스템A로 '코어넷'에 사용된 단어를 클러스터링한 결과 총 1,471개의 클러스터가, 시스템B로는 총 1,751 개의 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 '코어넷'의 한국어 부분은 총 12층위의 노드로 이루어져 있으며 총 2,937개의 개념명을 지니고 있다. 어휘는 이 개념명 각각에 유의어 관계로 그룹화되어 있다. '코어넷'의 상위 4층위까지의 노드에 대해 나무구조로 나타낸 것은 다음의 〈그림 7〉과 같다.



〈그림 7〉 '코어넷'의 상위 노드

'코어넷'의 개념번호에서 첫 번째 층위가 '구체/추상', 두 번째 층위가 '구체', '추상', 세 번째 층위가 '구체'에 대해 '주체', '장소', '구체물', '추상'에 대해 '추상물', '일', '추상적 관계'이다. '코어넷'의 개념체계는 2,937개의 개념명이 적게는 5층위에서 많게는 12층위까지를 이루고 있다. 5층위까지를 이루는 개념체계의 부분 중 한 예는 개념명 '관계'로 '관계', '관련', '연고' 등과 같은 단어가 분포되어 있다. 12층위의 개념명은 '목축'과 같은 개념명으로 '낙농', '목양', '양돈' 등과 같은 단어가 분포되어 있다. '코어넷'의 가장 큰 특징 중의 하나는 최종 노드(terminal node)의 규모가 유의어 구분의 수준이기 때문에 단어클러스터링 시스템의 결과물과 비교가 용이하다는 것이다. 실제로 개념명 '관계'내에는 '관계', '관련', '연고' 뿐만이 아니라 그 하위 개념으로 분류 가능한 '인과관계', '이해관계', '인간관계' 등을 포함하여 80여 개의 단어가 하나의 개념명으로 묶여 있다.

비교는 〈그림 8〉에서 보는 바와 같은 순서로 진행된다.

클러스터의 정확도는 각 단어마다 '코어넷'에서의 개념번호를 서로 비교하되, 상위 3층위까지 같은 번호를 지니는지의 여부로 점수를 부여한다. 시스템으로 클러스터링한 결과물에 대해서 상위 3층위까지의 개념번호의 동일 여부를 비교해 보면, 〈표 9〉에서 보는 바와 같이 '내년', '연말'은 같은 클러스터링으로, '내놓다'는 다른 클러스터링으로 분류하는 정도의 기준이다.

〈표 9〉 클러스터링 분류 정도

내년, 연말, *내놓다*
산정, 산출
작다, 크기
능력, 뛰어나다
늘리다, 줄이다, 축소

점수는 정확률을 위해 각 클러스터뿐만 아니라 클러스터 내의 단어 각각에 대해 개별적으로 배점을 하여 평균 점수를 구하였다. 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 3층위까지의 개념번호가 동일하지만 나머지 한 단어에서는 개념번호

1. '코어넷'의 단어를 개념명을 없앤 어휘 수준으로 나열한다.
2. 단어클러스터링의 과정을 거쳐 각각 시스템A와 시스템B의 결과물을 산출한다.
3. 산출된 결과물을 기존의 '코어넷'이 지닌 개념명 혹은 개념번호와 비교하여 얼마나 유사한가를 서로 비교한다.

〈그림 8〉 시스템A와 시스템B의 비교 과정

가 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 시스템A 에서는 4,115 개의 단어로 1,471개의 클러스터가 형성되고 6,361 개의 단어가 클러스터되지 못하고 남았다. 시스템B에서는 5,789개의 단어로 1,751 개의 클러스터가 형성되고 5,739개의 단어가 클러스터에 실패했다. A, B 두 시스템의 클러스터된 단어의 개수와 클러스터 개수를 정리하면 〈표 10〉에서 보는 바와 같다.

〈표 10〉 시스템A와 시스템B의 클러스터링 결과

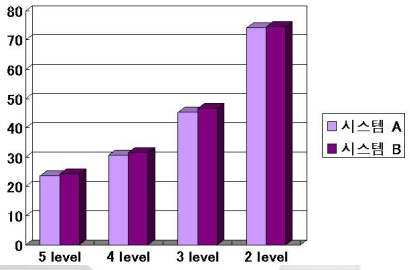
	시스템A	시스템B
클러스터 성공 단어 개수/ 클러스터 실패 단어 개수	4,115 / 6,361	5,789 / 5,739
클러스터 갯수	1,471	1,751

다음으로 〈표 11〉은 〈표 10〉에서의 클러스터된 집단인 시스템A의 1,471개와 시스템B의 1,751개의 클러스터로 4.1절에서 소개된 비교 방식을 적용한 결과의 최종 정확률이다.

〈표 11〉 시스템A와 시스템B의 클러스터 정확률

	시스템 A	시스템 B
클러스터의 정확률	45.3%	46.6%

실험은 4.1절에서 소개된 바와 같이 상위3층위까지의 개념번호의 동일성 여부로 이루어졌다. 실험의 오류를 없애기 위해 상위2층위부터 5층위까지의 정확률을 모두 검증하였다. 다음의 〈그림 9〉에서 알 수 있듯이 2~5까지의 결과가 상위 3층위와 비교해 볼 때, 그 결과 변화에는 큰 변화가 없음을 알 수 있다.



〈그림 9〉 2~5층위까지의 정확률 비교

5. 결론 및 향후

본 논문에서는 단어클러스터링 시스템 개발을 통하여, 어휘의미망에 의해 확장되기 전후의 클러스터링을 수행하여 데이터를 서로 비교하여 어휘의미망의 평가 및 활용 방안을 제시하였다. 클러스터링된 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. 본 연구에서는 동사, 그 중에서 '[-하]동사류'에 대한 어휘의미

망을 시스템에 활용시킨 결과를 보여주었다. 향후 연구는 동사 어휘의미망 뿐만이 아니라 폭넓은 분야의 어휘의미망을 활용하여 좀 더 다양한 시스템에 체계적인 평가를 통해 시스템의 향상은 물론, 연구되고 있는 많은 어휘의미망에 의미 있는 평가 방안을 확대시켜 나가야 할 것이다.

참 고 문 헌

기민호. 2001.『단어클러스터링 기반 정보처리 도구 개발 기술』. 정보통신부 우수신기술 지정·지원 사업 최종 보고서.

김준수. 2004.『의미정보와 시소러스를 이용한 한국어 어휘 중의성 해소 모델』. 울산대학교 컴퓨터정보통신공학과 박사학위논문.

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

김혜경, 윤애선. 2006. 동사 어휘의미망의 반자동 구축을 위한 사전정의문의 중심어 추출.『언어와 정보』. 10(1).

박석문. 2000.『코퍼스 품사 태깅 매뉴얼』 한국과학기술원.

신중호, 박혁로, 이기호. 1993. 단어의 유사성 척도와 클러스터링 알고리즘.『한국 인지과학회 논문지』. 9(2).

옥철영. 2005. 한국어 Wordnet 구축: 명사를 중심으로.『한국언어정보학회 2005 정기학술대회 발표 논문집』, 1-15.

이경순. 2001.『정보검색에서 벡터공간 검색과 클러스터 분석을 통한 문서 순위 결정 모델』. 한국과학기술원 전자전산학과 박사학위논문.

조현양, 최성필. 2004. 계층적 결합형 문서 클러스터링 시스템과 복합명사 색인방법과의 연관관계 연구.『한국문헌정보학회지』. 38(4): 179-192.

최준호. 2004.『의미적 멀티미디어 정보검색을 위한 개념간 유사도 측정 방법』. 조선대학교 전자계산학과 박사학위논문.

최호섭, 옥철영. 2002. 한국어 의미망 구축과 활용.『한국어학』. 17: 301-329.

한국과학기술원 전문용어언어공학연구센터. 2005.『다국어 어휘의미망』. KAIST PRESS.

한글학회. 1991.『우리말큰사전』. 어문각.

Baeza-Yates, Ricardo, and Berthier. Ribeiro-Neto. 1999. *Modern Information Retrieval*, ACM Press.

Dong, Zhendong, and Quiang. Dong. 2006. *How-Net and the Computation of Meaning*. World Scientific Publishing.

Fellbaum, Christiane. 1998. *Wordnet: An Electronic Lexical Database*. MIT Press.

Ikehara, Satoru. et al. 1997. *The Semantic System, volume 1 of Goi-Taikei -- A Japanese Lexicon*. Iwanami Shoten.

Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, 2000, *DATA MINING Methods for Knowledge Discovery*, Kluwer Academic Publishers.

Vossen, Piek. 2005. EuroWordNet General Document.