인지과학, 제22권 제3호
Korean Journal of Cognitive Science
2011, Vol. 22, No. 3, 271～298.

사전 뜻풀이를 이용한 용언 의미 군집화*

배 영 준[1] 최 호 섭[1] 송 유 화[2] 옥 철 영[3†]

[1]한국과학기술정보연구원
[2]과학기술연합대학원대학교 응용정보과학 전공
[3]울산대학교 컴퓨터정보통신공학과

한국어의 어휘 의미 정보를 명확히 파악하기 위해서는 어휘 의미 체계를 구축해야 한다.

본 논문에서는 어휘 의미 체계 구축의 단계 중 하나인 용언의 의미 군집화를 연구하였다. 용언의 하위범주화 논항(주어 및 목적어)과 선택 제약정보, 부사의 결합정보를 이용한 이전의 연구와는 달리 의미태그가 부착된 사전 뜻풀이의 용언정보를 이용하여 용언의 의미 군집화를 시도하였고, 표제어와 뜻풀이 용언 사이 관계의 종개념과 유개념 관계를 이용하여 계층적 의미 군집화를 시도하였다. 그리고 특정 범주의 일반 샘플을 이용했던 특정 용언의 부류가 아닌, 사전에 존재하는 대부분의 용언들을 대상으로 연구를 진행하였다.

본 논문에서 다의어 수준에서 구분된 총 106,501개의 용언(85,754개의 동사와 20,747개의 형용사)을 대상으로 한국어 용언 의미계층 군집 2,748개를 생성하였다. 이 중 순환정의 군집은 130개가 나타났으며, 중간 계층의 서브군집으로 261개가 나타났다. 군집 내 계층의 최대 깊이는 16단계였다. 그리고 용언 의미 군집 평가를 위해 세종사태의미부류와 비교해 보았다. 그 결과 70.14%의 응집도를 보였다.

주제어 : 의미클러스터링, 용언, 사전 뜻풀이, 뜻풀이 용언

---

* 이 논문은 현대중공업 지원에 의한 울산대학교 전기공학부 일류화 연구비에 의하여 연구되었음.

† 교신저자: 옥철영, 울산대학교 전기공학부, 연구 분야: 자연어 처리
E-mail: okcy@ulsan.ac.kr

서 론

정보검색을 위한 문서 및 자료의 전산화가 빠르게 진행되고 있다. 그 결과로 검색결과의 양은 많아졌지만, 정확한 문서뿐만 아니라 부정확한 문서까지 같이 검색되면서 실제 사용자에게 필요한 정보를 사용자 본인이 다시 선별해야 하는 부차적인 작업이 발생한다. 보다 정확한 정보를 사용자에게 제공하기 위해서는 문서의 형태적 처리, 구문적 처리뿐 아니라 의미적 처리 단계까지 필요하다.

의미기반 정보처리를 위해서 가장 우선 필요한 것이 의미적 언어자원이다. 의미주석이 부착된 말뭉치뿐만 아니라, 명사의 어휘망, 시소러스, 의미망, 온톨로지, 용언의 하위범주 정보, 부사-부사 및 부사-용언의 결합관계, 연어 관계 등의 언어 자원이 의미기반 정보처리에 필요하다. 주로 명사의 어휘망 및 시소러스 및 온톨로지가 가장 많이 활용되고 있지만, 보다 정확한 문장 분석 및 논항 추출을 위해서는 용언의 정보도 무시할 수 없다. 특히 질의응답 시스템에서 자연어 질의의 의미는 하나지만, 자연어 질의의 표현이 여러 가지로 나타날 수 있고, 질문에 나타난 용언도 의미는 비슷하지만 형태가 다른 용언이 사용될 수 있다. 이런 경우 유사한 용언의 군집이 있다면 질문의 요지로의 접근이 쉬워질 것이다.

<예 1> 000선생님께서 **돌아가신** 날짜는?
000가 **죽은** 날짜는 몇 월 며칠인가?
000님이 **사망한** 날은 언제인가?

<예 1>에서 사용된 {돌아가다, 죽다, 사망하다}의 대표 단어는 '죽다'로 볼 수 있고, 이러한 군집이 있다면 모든 질문의 유형을 일일이 기술할 필요 없이 군집된 용언의 대표정보를 이용하여 간단히 문장의 의미를 파악할 수 있다. 이러한 필요성에 따라 그리고 정확한 의미처리를 위해 용언의 군집화 및 계층화를 위한 연구가 진행되었다.

본 논문에서는 사전에 존재하는 대부분의 용언을 대상으로 용언 의미 군집화를 시도한다. 사전 뜻풀이에 의미태그를 부착한 뒤, 그 뜻풀이의 첫 번째 문장에 나타난 용언을 정확히 추출하기 위한 패턴을 분석한다. 분석된 패턴을 바탕으로 뜻풀

이 용언을 추출하여 용언 의미 군집을 시도한 뒤, 계층적 군집화를 시도한다.

관련연구

군집화(clustering)란 주어진 데이터를 의미 있는 그룹으로 분류하는 방법이다. 군집화 기법은 크게 계층적 군집화(hierarchical clustering)과 할당식(partitioning clustering)으로 나눌 수 있고, 구축 방식에 따라 하향식(top-down) 방식과 상향식(bottom-up) 방식으로 나눌 수 있다. 그리고 군집 대상에 따라 텍스트 데이터 군집, 음성 데이터 군집, 영상 데이터 군집 등으로 나눌 수 있다. 그 중 텍스트 데이터 군집에는 문서 군집, 용어 군집 등이 있으며, 문서 군집은 문서 내에 나타난 핵심 용어를 기반 자질로 하여 다른 문서와의 유사도를 파악해 문서간의 군집을 한다. 용어 군집은 문서 또는 문장 내에 같이 자주 나타나는 용어를 묶는 방법과 정제된 사전을 기반으로 정보를 추출해 자질로 묶는 방법 등이 있다.

용언의 군집화의 기준으로 선정된 자질들은 일정 부류의 어휘에 공통적으로 적용되는 것이므로, 그것들을 실제 용언의 분석 과정에서 어떻게 적용하고 처리할 것인지가 중요한 연구 대상이 되었다[1]. 그렇기 때문에 이전의 연구들은 군집화 자질로 볼 수 있는 논항의 구조와 선택제약 정보 및 기타 동사와 명사 쌍(pair) 등의 결합정보를 이용한 규칙 기반[8], 통계적 기법[2, 10], 혼합형 기계학습 기법[3, 4]을 이용하여 군집화를 연구하였다.

그러나 이러한 연구는 다의어가 나타날 경우 구분이 어렵기 때문에 정확한 군집을 하지 못해 모호성이 발생하는데, 의미태그 부착 말뭉치가 있으면 보다 정확한 군집을 할 수 있다. 그리고 용언의 개수가 많아지거나 분야의 범주가 커지면 모호성이 더 증가하기 때문에 기존의 연구들은 용언의 개수를 적절히 한정시키고 특정한 분야의 범주만을 대상으로 한정 하였다. 그리고 학습말뭉치 구축의 많은 노력, 시간, 인력이 들기 때문에 간단히 특정 용언부류만을 대상으로 군집화를 시도하였다.

WordNet이나 의미태그 부착 말뭉치를 이용해 의미기반으로 용언을 군집화하는 연구가 국외에서는 계속 진행 중이다[3, 5, 10]. 그러나 용언의 군집화 관련 국내

연구 중 의미기반으로 용언을 군집화한 연구가 드물다. 35,100개의 '[-하]동사류'을 대상으로 동사 어휘의미망 반자동 구축을 위해 사전정의문의 중심어를 추출하는 연구[11]가 있었고, 영어 워드넷을 참조 모델로 한 한국어 어휘의미망(KorLex)에서는 동사 20,133개, 형용사 20,905개를 대상으로 용언의 어휘망을 구축하는 연구[12]가 있었다.

국외와 달리 의미자질을 포함하는 말뭉치가 국내에는 미비했고, 말뭉치가 있다고 해도 실제 연구용으로 확보가 어려웠기 때문에 용언의 군집화에 대한 연구가 진행되기 어려웠다. 21세기 세종계획으로 만들어진 형태의미분석 말뭉치 1,250만 어절이 있지만 그 중 의미태그 부착 용언은 동형이의어 단계까지만 태그가 부착되어 있고 다의어 단계까지는 태그가 부착되어 있지 않다.

용언 의미 군집화를 위한 기반 작업

사전 뜻풀이 의미 태그 부착

의미 처리 기술을 위한 의미 정보 구축을 위해 다의어 수준의 의미태그가 부착된 표준국어대사전의 뜻풀이를 구축[1]하였다. 다의어 분별을 위해 표준국어대사전의 표제어[2] 중 다의어인 경우는 분리해서 각각 하나의 표제어로 만들었으며, 다의어로 분리된 용언 표제어의 총 개수는 112,024개이다. 이 중 동사는 90,365개이고, 형용사는 21,659개이다. 작업 순서로 우선 표준국어대사전의 뜻풀이를 형태소 분석기를 이용하여 뜻풀이에 품사 태그를 부착하였다. 그리고 의미태그 부착도구를 이

1) 다의어 수준의 의미태그가 부착된 표준국어대사전의 뜻풀이 구축은 2007년부터 시작하여 현재까지 5년 가까이 작업을 진행했으며, 현재 마무리 단계에 있다. 의미태그 부착은 끝났으며 세밀한 오류 유형이나 부분적인 형태소 분석 오류들을 수작업으로 수정하고 있다.

2) 표준국어대사전의 표제어 중 용언의 총 개수는 75,639개이고, 동사는 58,248개이며, 형용사는 17,391개이다. 본 논문에서 사용한 CD 버전 표준국어대사전은 업데이트되기 전 자료라서 현재 국립국어원에서 서비스하고 있는 Web에 있는 표준국어대사전과 개수 차이가 있을 수 있다.

용하여 동형이의어 및 다의어 단어를 대상으로 의미태그를 반자동으로 부착하였으며, 사람을 통한 검증 작업을 거쳤다.

사전 뜻풀이의 용언 추출

용언 의미 군집화를 위해 기반이 되는 용언을 사전 뜻풀이에서 추출한다. 뜻풀이의 첫 문장을 대상으로 하며, 특정 보조용언이 추출될 경우 보조용언 앞의 본용언을 추출하여 용언을 군집화한다.

일반적인 뜻풀이 용언 추출

사전에서 표제어의 주요 정의는 대부분 사전 뜻풀이의 첫 문장에 나타난다. 그래서 첫 문장의 마지막에 나타나는 보조용언을 제외한 용언(이하 뜻풀이 용언)대상으로 집합을 생성한다. 뜻풀이 용언을 추출하기 위해 의미태그 부착 뜻풀이의 마지막 어절부터 잘라서 품사 'VV(동사)', 'VA(형용사)'가 나타난다면 뜻풀이 용언으로 추출한다. <그림 1>은 표제어 '가늘다_000002'[3]의 의미태그 부착 뜻풀이에서

그림 1. '표제어', '뜻풀이 용언' 리스트의 구축 과정

3) 의미태그 번호 6자리: 동형이의어 번호(2자리) / 문형 번호(2자리) / 다의어 번호(2자리)를 뜻한다. 이는 표준국어대사전의 의미 분류체계를 따른 것인데, 다의어를 중복 없이 명확히 구분하기 위해 위 3가지 정보를 조합하여 의미태그 번호를 선정하였다.

마지막의 품사 'VA'의 형용사를 뜻풀이 용언으로 추출하여 매칭 리스트에 추가시키는 과정을 보여주고 있다.

본용언과 보조용언을 이용한 뜻풀이 용언 추출

사전의 뜻풀이에서 단순히 첫 문장의 마지막 보조용언을 제외한 용언만 추출하면 몇몇 용언의 표제어와 뜻풀이 용언의 의미가 이질적이게 되는 문제가 발생한다. 예를 들면 뜻풀이 용언 뒤에 특정 보조용언이 나타나는 경우다. 뜻풀이가 부정(否定)형인 보조용언을 포함하고 있을 경우, <그림 2>와 같이 문장의 뜻풀이 용언을 전처리 없이 추출하면 의미가 상반된 부적절한 용언이 추출된다. <그림 2>의 '추출된 용언'에 나타난 '표제어, 뜻풀이 용언' 쌍 '흐리다_020001, 분명하다_010002'처럼 표제어와 상반된 의미의 뜻풀이 용언이 추출된다. 이럴 경우 전혀 다른 의미의 용언이 같은 집합으로 묶일 수 있다. 그래서 정확한 용언 의미 군집을 생성하기 위해 우선 부정형 보조용언이 나타나는 패턴을 추출하였다.

부정(否定)문은 부정을 나타내는 부사 '아니(안)'나 '못'을 쓰거나 부정의 뜻을 나타내는 용언 {아니다, 아니하다(않다), 못하다, 말다}를 써서 만든다. 그래서 보조용언 {아니다, 아니하다, 못하다, 말다}가 나타나는 뜻풀이의 패턴을 선정하고, 다의어가 구분된 보조용언 패턴을 <표 1>과 같이 정리하였다. 그리고 <표 1>의

그림 2. 부정형 보조용언이 나타나는 표제어의 예

(12)~(15)번과 같이 문법상 부정문을 만드는 보조용언 형태는 아니지만, 의미상 상반된 용언이 추출될 수 있는 형용사 '없다' 패턴도 포함시켰다. 그래서 의미태그 부착 뜻풀이를 분석하여 보다 정확한 뜻풀이 용언 추출에 이용하였다.

표 1. 부정형을 포함한 의미태그 부착 뜻풀이 패턴

no	패턴
1	(용언)+지/EC 못하다_000300/VX+다/EF+./SF
2	(용언)+지/EC 못하다_000401/VX+다/EF+./SF
3	(용언)+지/EC 못하다_000300/VX+게/EC 하다/VX+다/EF+./SF
4	(용언)+지/EC+는/JX 못하다_000401/VX+다/EF+./SF
5	(용언)+지/EC 못하다_000401/VX+ㄴ/ETM 데/NNB+가/JKS 있다/VA+다/EF+./SF
6	(용언)+지/EC 않다_000200/VX+다/EF+./SF
7	(용언)+지/EC 않다_000300/VX+다/EF+./SF
8	(용언)+지/EC 아니하다_000100/VX+ㅁ/ETN+./SF
9	(용언)+지/EC 아니하다_000100/VX+다/EF+./SF
10	(용언)+지/EC 아니하다_000200/VX+다/EF+./SF
11	(용언)+지/EC+도/JX 아니하다_000200/VX+다/EF+./SF
12	(용언)+ㅁ/ETN+이/JKS 없다_010103/VA+다/EF+./SF
13	(용언)+ㅁ/ETN+이/JKS 없다_010102/VA+다/EF+./SF
14	(용언)+는/ETM (명사)/NNG+이/JKS 없다_010301/VA+다/EF+./SF
15	(용언)+ㄴ/ETM (명사)/NNG+이/JKS 없다_010103/VA+다/EF+./SF

부정형 패턴을 적용시켜 추출된 뜻풀이 용언은 의미상 반의어로 대체될 수 있다. 뜻풀이에서 부정형을 가지는 뜻풀이 용언 중 반의어가 있는 뜻풀이 용언은 반의어로 대체하고 그 외의 뜻풀이 용언은 'N_뜻풀이 용언'으로 표시한다. <예 2>은 뜻풀이 용언을 대체할 수 있는 반의어가 없는 예이며, <예 3>는 뜻풀이 용언을 대체할 수 있는 반의어가 있는 예이다.

<예 2> 뜻풀이 용언의 반의어가 없는 예

‘흐리다_020001’의 뜻풀이 : “기억력이나 판단력 따위가 분명하지 아니하다.”

‘흐리다_020001’의 뜻풀이 본용언: 분명하다_010002

[설정]: 흐리다_020001 ≒ **N_**분명하다_010002

<예 3> 뜻풀이 용언의 반의어가 있는 예

‘불합당하다_000000’의 뜻풀이 : “사리에 옳지 아니하다.”

‘불합당하다_000000’의 뜻풀이 본용언: 옳다_010001

[설정]: 불합당하다_000000 ≒ 그르다_010001 ↔ 옳다_010001

용언의 부정형을 ‘N_뜻풀이 용언’의 형식으로 나타내면, 군집될 용언들을 더 이상 확보 할 수 없다. 위의 <예 3>처럼 뜻풀이 용언이 ‘옳다_010001’ 일 때 ‘N_옳다_010001’이 아닌 ‘그르다_010001’로 대체한다면, 현재 용언에서 단절되던 용언 군집을 보다 확장할 수 있다. 그래서 최대한 많은 반의어 확보를 위해 용언의 반의어 쌍 1,884개를 확보한 뒤, 용언의 동의어, 유의어, 비슷한 말을 통해 확장시켰다.

‘~의 사동형’, ‘~의 피동형’ 형식의 뜻풀이 용언 추출

<표 1>의 부정형이 나타나는 뜻풀이와 마찬가지로 <표 2>와 같이 사동사, 피동사 패턴으로 나타나는 뜻풀이 용언은 표제어와 다른 의미를 가진다. 사동(使動)이란 “남으로 하여금 어떤 동작을 하게 하는 동작”을 말하고, 주어와 행위자의 관계에 따라서 주동(主動)과 구분된다. 피동(被動)이란 “어떤 행위나 동작이 주어로 나타내어진 인물이나 사물이 제 힘으로 행하는 것이 아니라, 남의 행동에 의해서 되는 행위”를 말하고, 행위자와 행위의 관계에 따라서 능동(能動)과 구분된다. 사동과 피동을 구분하는 이유는 문장에서 문법상 주어와 목적어가 서로 달라질 수 있기 때문에 이러한 용언들은 서로 다른 의미라 판단하고 사동사일 경우 ‘mk_뜻풀이 용언’ 형식으로 나타내고, 피동사일 경우 ‘psv_뜻풀이 용언’ 형식으로 나타낸다.

표 2. 사동사, 피동사 형태의 용언패턴

no	패턴	예	
		표제어	뜻풀이
1	(용언)+의/JKG 사동사/NNG+./SF	가라앉히다_000100	'가라앉다{1}(1)'의 사동사.
2	(용언)+의/JKG 피동사/NNG+./SF	싸이다_010102	'싸다01{2}(1)'의 피동사.

<표 3>은 지금까지 일반형, 부정형, 사동형, 피동형 사전 뜻풀이에서 추출된 용언을 표제어와 함께 보여준다.

표 3. 사전 뜻풀이에서 추출된 본용언

no	표제어	뜻풀이 용언
1	가볍다_000101	적다_020000
2	가는귀먹다_000000	먹다_010000
3	낚다_000001	잡다_010102
4	날다_000001	움직이다_000001
5	내리우다_000104	mk_내리다_010703
6	흔들리다_000002	psv_흔들다_000002
7	뿌리치다_000001	N_붙잡다_000001

'사전 뜻풀이의 용언 추출'절에서 다루었던 사전 뜻풀이의 일반용언 추출, 부정형 패턴이 나타난 용언의 추출, 사동형, 피동형이 나타난 용언 추출을 위한 시스템 구성도는 <그림 3>과 같다.

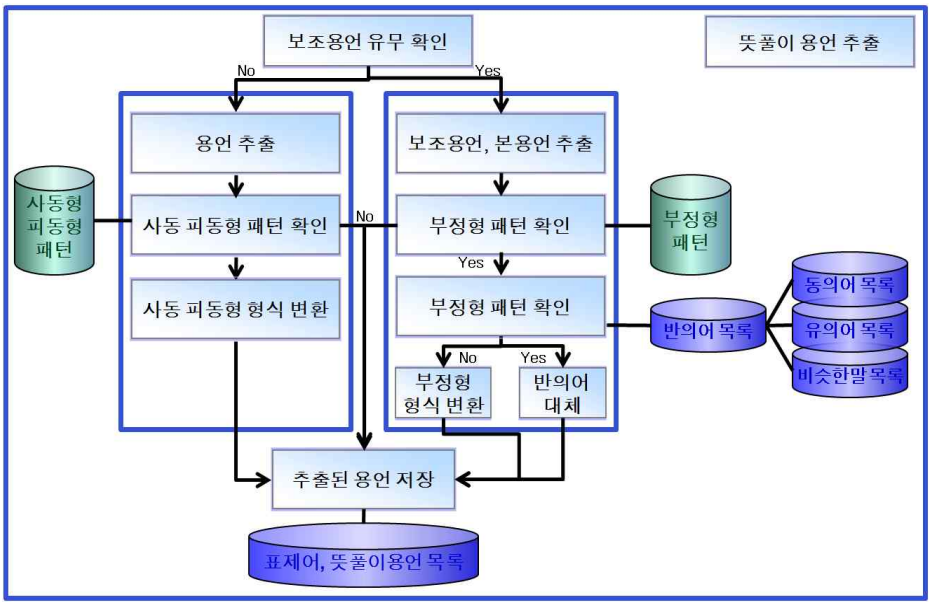

그림 3. 뜻풀이 용언 추출 시스템 구성도

용언 의미 군집화

추출된 용언 정보를 이용하여 집합을 만들고 상위 용언 연결 구조를 만들어 본다. <그림 4>는 본 절에서 실행할 용언 의미 군집을 위한 두 가지 방법을 보여준다. 우선 하나의 용언을 뜻풀이 용언으로 가지는 표제어들을 묶어 용언 집합으로 생성해 보고, 표제어와 뜻풀이 용언을 따라가며 용언의 계층을 생성해본다. 마지막으로 순환되는 용언들을 군집한 뒤 용언들을 연결하고 군집을 형성한다.

용언 집합 생성

앞의 절에서 표제어와 뜻풀이 용언 쌍을 추출하였다. 본 절에서는 이 리스트를 바탕으로 용언의 의미 집합을 만들어 본다. 표제어는 뜻풀이를 통해 정의되고 뜻풀이는 “글의 뜻을 풀이함”이란 정의처럼 표제어의 의미와 비슷하면서 보다 일반

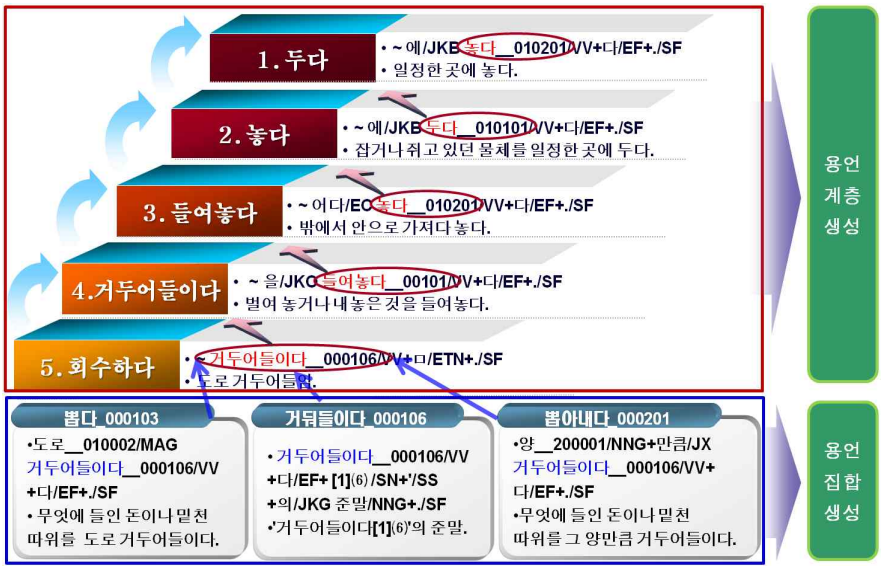

그림 4. 용언 의미 군집화를 위한 방법

적이고 포괄적인 단어를 사용해 표제어를 설명한다. 이러한 사실을 바탕으로 용언의 의미 집합을 생성할 수 있다. 표제어를 종개념으로 보고 뜻풀이 용언을 유개념으로 봤을 때, 같은 유개념 즉, 같은 뜻풀이 용언을 가진 표제어들을 묶을 수 있다. <그림 4>의 '용언 집합 생성' 부분에서 용언 '뽑다_000103', '거둬들이다_000106', '뽑아내다_000201' 용언들은 뜻풀이 용언으로 '거두어들이다_000106'을 가지므로 하나의 용언 집합으로 보고 묶을 수 있다. 용언 의미별로 집합을 정의하고 동일 상위 용언을 사전 뜻풀이에 포함하는 용언들을 추출하면 <표 4>와 같다.

<표 4>에서 '(용언 수)'는 뜻풀이 용언을 '떨어지다'로 가지는 표제어의 개수이다. 그리고 '뜻풀이 용언이 떨어지다인 표제어 리스트' 항목의 용언들 중 다의어는 의미태그가 붙어 있고(예: 흘러내리다_000001), 단의어는 의미태그가 붙어있지 않다(예: 낙하하다). 용언 '① 떨어지다000101'을 뜻풀이 용언으로 가지는 용언은 {곤두박다, 낙마하다, 낙하하다, 뿌리다_000100, 흘러내리다_000001…} 등 541가지가 있고 이 용언들은 하나의 집합으로 묶을 수 있다. 용언 '떨어지다'의 의미 수는 30개이지만 그 중 9개만 나타내었다.

표 4. {떨어지다}를 상위어로 가지는 용언 집합

표제어 (용언 수)	‘떨어지다’의 뜻풀이
	뜻풀이 용언이 ‘떨어지다’인 표제어 리스트
① 떨어지다 000101(541)	위에서 아래로 내려지다.
	곤두박다, 낙마하다, 낙하하다, 뿌리다_000100, 흘러내리다_000001, …
② 떨어지다 000102(2)	어떤 상태나 처지에 빠지다.
	타락하다_000003, 타옥하다
③ 떨어지다 000201(10)	정이 없어지거나 멀어지다.
	멀찍하다, 배리되다_020000, 부즉불리하다_000001, 이심하다_030000, 정떨어지다 …
④ 떨어지다 000204(17)	다른 것보다 수준이 처지거나 못하다.
	손상되다_000004, 쇠락하다, 일합일리하다, 추락하다_020002, 저하되다_000001 …
⑤ 떨어지다 000300(25)	시험, 선거, 선발 따위에 응하여 뽑히지 못하다.
	낙방하다_000001, 낙선하다_00001, 낙제하다_000201, 불합격하다_000001 …
⑥ 떨어지다 000401(7)	함께 하거나 따르지 않고 뒤에 처지다.
	나가떨어지다_000003, 나가떨어지다_000004, 이탈되다, 처지다_010300 …
⑦ 떨어지다 000402(49)	달렸거나 붙었던 것이 갈라지거나 떼어지다.
	낙화하다, 떼다_010101, 문드러지다_000001, 벗겨지다_000001, 분비하다_020000 …
⑧ 떨어지다 000502(24)	일정한 거리를 두고 있다.
	동떨어지다_000001, 띄엄띄엄하다_000001, 멀다_020100, 벽재하다, 유리되다_000001 …
⑨ 떨어지다 000601(24)	값, 기온, 수준, 형세 따위가 낮아지거나 내려가다.
	격락하다, 급락하다, 내려가다_000302, 저락하다_010000, 폭락하다_000001, 하락하다 …

용언의 계층적 연결 구성

용언의 계층화는 명사의 계층화보다 비교적 연구가 적게 이루어졌고, 그만큼 계층적 연결을 시도한 경우가 드물다. 그러나 용언도 서로 비슷한 의미의 용언을 하나의 집합으로 묶을 수 있고, 여러 계층으로 나눌 순 없지만 간단하게 의미의 포괄성 정도에 따라 두세 단계의 계층으로 나눌 수 있다.

용언류들의 의미를 정의할 때 표현 방식이 명사로 끝나는 것이 아니라 용언으로 끝난다. 이러한 용언류들의 정의에 사용된 용언을 이용해 상위 유의어의 연결을 만들 수 있다. 앞의 <표 3>에서 '표제어: 가는귀가먹다_000000'의 용언의 의미가 '뜻풀이 용언: 먹다_010000'의 의미보다 조금 더 구체적이거나 비슷한 의미를 함의한다는 것을 알 수 있다. 이와 같은 정의를 이용해 <표 5>와 같이 용언의 상위 연결 구조를 만들었다.

표 5. {떨어지다}의 상위 용언 연결

표제어	상위 용언 연결
① 떨어지다	mk_옮다_000101>옮기다_000101>내려지다_000101>떨어지다_000101
② 떨어지다	psv_놓다_010201>놓이다_000201>빠지다_020102>떨어지다_000102
③ 떨어지다	있다_010301>*존재하다_000001>있다_010302>쑥스럽다_000000>어색하다_020100>서먹서먹하다_000000>멀다_020301>떨어지다_000201
④ 떨어지다	N_못하다_000201>떨어지다_000204
⑤ 떨어지다	N_뽑히다_000004>떨어지다_000300
⑥ 떨어지다	처지다_010300>*떨어지다_000401
⑦ 떨어지다	떼다_010101>*떨어지다_000402
⑧ 떨어지다	지니다_000101>*가지다_000101>주다_010102>두다_010303>떨어지다_000502
⑨ 떨어지다	내려가다_000302>*떨어지다_000601

상위 용언 추출 과정에서 뜻풀이가 사동형, 피동형, 순환 정의된 경우가 많았으며, 위 표에서 '③ 떨어지다'의 '상위 용언 연결'에서 '있다_010301>*존재하다_000001…' 부분의 *기호는 뜻풀이의 본용언들이 순환정의된 것을 뜻한다. '있다_010301'의 뜻풀이 용언이 '존재하다_000001'이고, '존재하다_000001'의 뜻풀이 용언이 '있다_010301'라는 것이다. 이렇게 순환정의 되는 어휘는 대부분의 의미가 동등하거나 비슷한 경우가 많다. 그래서 이러한 순환정의 되는 용언들을 하나의 묶음으로 묶을 수 있고, 이것들을 하나의 집합으로 볼 수 있다. <그림 5>는 '내다_020302/산출되다_010000/나오다_000403/배출되다_030000/나다_010204'의 5개 용언이 하나의 순환 정의 용언 집합으로 묶인 예이다. '나다_010204'는 뜻풀이용언으로 '배출되다_030000'를 가지며, 순환정의의 마지막 용언 '내다_020302'는 뜻풀이 용언으로 '나다_010204'를 가지므로 5개의 용언이 순환정의 용언 집합으로 묶이게 된다. 5개의 용언 모두 정확히 의미가 일치되는 것은 아니지만 공통적으로 "인물이 배출되다." 라는 의미를 내포하고 있다. 순환정의 용언 집합의 용언들은 <그림 5>의 예와 같이 모두 비슷한 의미를 내포하고 있다.

그림 5. 순환정의 용언 집합의 예

<표 6>은 순환정의 되는 용언들의 집합 중 30개를 정리하였고, 이러한 집합은 총 130개가 추출되었다.

용언의 계층화 과정에서 <그림 6>과 같이 특정 용언이 하나의 하위 용언만 가지는 경우가 나타난다. 이럴 경우 특정 용언과 하위 용언을 하나의 군집으로 묶어

표 6. 상위 유의어 연결 구조에서 나타난 순환정의 용언

no	순환정의 용언 집합	용언 개수
1	내다_020302/산출되다_010000/나오다_000403/배출되다_030000/나다_010204	5
2	구분하다_030000/갈라놓다_000002/구별하다/가르다_000100/나누다_000101	5
3	일치하다/합치하다_000001/맞다_010302/들어맞다	4
4	굽히다_000001/구부리다_010000/휘다_000001/굽다_020000	4
5	덜하다/작다_010001/약하다_010101	3
6	씌우다_010001/덮다_000101/쓰다_020101	3
7	아득하다_000202/막연하다_000001/막막하다_020002	3
8	말하다_000104/나무라다_000001/꾸짖다	3
9	도망가다/달아나다_000102/도망치다	3
10	죄송하다/송구하다_010000/황송하다	3
11	중요하다_020000/긴요하다/요긴하다	3
12	지도하다/인도하다_020001/이끌다_000003	3
13	맡다_010101/담당하다	2
14	분명하다_010001/뚜렷하다	2
15	만들다_000106/결성하다	2
16	만들다_000102/편찬하다	2
17	겁다_000101/어렵다_000101	2
18	검토하다/따지다_010203	2
19	맑다_010001/깨끗하다_000002	2
20	매만지다_000001/손질하다_000001	2
21	무성하다_000001/우거지다_000100	2
22	묵다_020000/머무르다_000001	2
23	뭉치다_000200/단결하다_010000	2
24	미끈하다_000103/훤칠하다_000001	2
25	더럽다_000001/지저분하다_000002	2
26	반복하다_010000/되풀이하다_010000	2
27	휘어지다/구부러지다	2
28	갑갑하다_000005/답답하다_000102	2
29	밝다_000101/환하다_010101	2
30	털털하다_020001/소탈하다	2

주면 불필요하게 계층이 깊어지는 문제를 해결할 수 있다. 단일 연결된 용언들이 단말 노드에 나타나는 경우, 의미가 보다 세분화되는 경우와 <예 4>와 같이 '~의 잘못', '~의 방언', '~의 북한어'처럼 의미는 같으나 다른 표현으로 쓰인 경우로 구분된다. 중간 노드에 나타나는 경우는 <표 7>과 같이 의미적으로 유사성이 높은 용언들이 묶이기 때문에, 단말 노드에 나타나는 경우와 구분하여 단일 연결된 용언들이 단말 노드가 아닌 중간 노드에 나타날 때만 하나의 군집으로 묶어주었다. <그림 6>은 '사정없다_000000', '매몰차다_000001', '얀재없다_000000'이 단일 연결된 용언들이고, 이중 단말노드인 '얀재없다_000000'을 제외한 중간노드 '사정없다_000000'과 '매몰차다_000001'이 서로 묶인 것을 볼 수 있다. 두 개의 용언의 뜻풀이를 보면 '남의 사정을 헤아려 돌봄이 없이 매몰차다.', '인정이나 싹싹한 맛이 없고 아주 쌀쌀맞다.'로 의미가 비슷하기 때문에 하나의 군집으로 묶을 수 있다.

그림 6. 단일 연결된 용언 중간 계층 병합

단일 연결된 용언이 나타난 군집의 총 개수는 4,779개이며, 단말노드에 나타난 개수가 4,518개, 중간노드에 나타난 개수가 261개였다. 따라서 중간 계층에서 총 261개의 군집이 묶였다. 단일 연결된 용언은 <표 7>의 9번과 같이 최대 연결 개수는 3개였다.

<예 4> 단말노드로 사용되는 단일 연결된 용언들의 <표제어, 뜻풀이> 예

- 얇따랗다 : '얄따랗다'**의 잘못**.
- 골라다 : '고르다01'**의 방언**(경남).
- 앞차림하다 : '앞가림하다'**의 북한어**.

표 7. 단일 연결된 용언이 중간 노드에 나타나는 경우 중 10가지 예

번호	단일 용언 계층 (하위 용언 개수)
1	…>**헝클어지다_000002(1)>어수선하다_000101**(7)>지저분하다_000001(10)
2	…>움직이다_000001(869)>뒤끓다_000002(12)>**꼬이다_010001(1)>틀어지다_000101**(4)…
3	…>하다_010303(27)>가지다_000102(90)>**획득하다_000000(1)>얻다_010301**(39)…
4	…>정하다_030100(61)>**지목하다_000000(1)>꼽다_010002(1)>세다_020001**(11)…
5	…>있다_010302(1934)>**불명예스럽다_000000(1)>욕되다_000000**(13)…
6	…>나아가다_000001(150)>가다_010301(9)>**평행하다_000001(1)>나란하다_000002**(2)…
7	…>있다_010302(1934)>**소통되다_000002(1)>통하다_000601**(24)>맥락관통하다_000000(0)
8	…>분해되다_000103(2)>**부패하다_000002(1)>상하다_020102**(1)>조상하다_040000(0)
9	psv_담다_010002(1)>**담기다_010002(1)>스미다_000200(1)>깃들다_000002(1)**>피어리다_000000(0)
10	psv_열다_020101(1)>열리다_020101(50)>**벌리다_010103(1)>뜨다_050001**(77)…

용언 의미 군집화 결과 및 평가

용언 의미 군집화를 위해 다의어 단위로 구분된 용언 106,501개가 사용되었으며, 이중 동사는 85,754개, 형용사는 20,747개가 의미 군집에 사용되었다. 군집된 용언 중 최상위용언의 총 개수는 2,915개이고, 최상위 용언 중 부정형은 781개, 사동형

은 777개, 피동형은 1,060개, 일반용언은 297개이다. 일반용언 중 순환정의된 용언 의미 군집은 130개이며 군집에 포함된 용언은 285개이고 12개의 용언은 순환되지 않고 계층을 형성하는 도중 단절된 용언으로 볼 수 있다. 단절된 원인을 분석해 보면, <그림 7>처럼 뜻풀이에 용언이 전혀 사용되지 않아 뜻풀이 용언을 추출할 수 없을 경우 계층을 생성하는 과정에서 단절된 현상이 발생하였다. 이 문제는 뜻풀이를 업데이트된 웹의 표준국어대사전 뜻풀이로 대체하거나 새로운 분류항을 만들거나 뜻풀이의 마지막 명사 연결 규칙 등을 만들어 해결할 수 있을 것이다.

그림 7. 뜻풀이에 용언이 없어 뜻풀이 용언을 추출할 수 없는 예

<표 8>은 용언을 상위 연결 구조로 만들었을 때, 최상위에 놓인 용언 중 하위 용언의 개수가 많은 순으로 50개 추출한 것이다. 가장 많은 용언을 하위어로 포함한 최상위 용언은 '이루어지다_000200'으로 16,349개의 하위어를 가진다.

최상위 용언의 하위 용언을 분석해 보면 부정형을 최상위 용언으로 가지는 하위 용언의 개수는 10,797개이며 사동형을 최상위 용언으로 가지는 하위 용언의 개수는 22,632개이고 피동형을 최상위 용언으로 가지는 하위 용언의 개수는 10,583개이다.

가장 계층의 깊이가 큰 군집은 16단계의 계층 구조를 이루고 있다. 실제 군집 결과 18단계[4]가 최고 단계였지만, 이는 일부 오류에 의한 것으로 가장 긴 군집 계

4) 오류에 의한 18단계 계층구조: 있다_010301>*존재하다_000001>있다_010302>새롭다_000001>대신하다_000000>바꾸다_000101>이동하다_000001>가다_010101>들어가다_010101>들여보내다_000101>먹다_020101>**맛보다_010101**>누리다_010000>주다_010101>드리다_010001>바치다_010101>내다_020207>납부하다_000000>반납하다_010000

위 계층은 의미적 연결이 잘못된 예이다. 그 이유로 두 가지를 생각해 볼 수 있다. 첫 번째는 [용언+연결어미+용언+연결어미+용언+종결어미] 형식의 뜻풀이를 가지며, 마지막

표 8. 최상위 용언 50개

최상위 용언	하위 용언개수	최상위 용언	하위 용언 개수
이루어지다__000200	16,349	품다__010200	457
당하다__010501	12,128	일치하다__000000	425
mk_옮다__000101	8,741	떨어지다__000402	395
있다__010301	5,109	mk_맡다__010101	385
mk_드러나다__000102	4,688	mk_없다__010201	385
psv_보다__010101	3,551	psv_듣다__010101	365
mk_나다__010403	3,299	치르다__000201	342
mk_나다__010506	3,143	똑바르다__000001	339
지니다__000101	2,119	N_놓다__010101	338
psv_놓다__010201	1,902	N_떨어지다__000402	334
차리다__000003	1,773	N_미치다__020200	331
결정하다__010000	1,585	psv_모으다__000200	331
N_다르다__010001	1,539	N_일어나다__000207	329
지나다__000302	1,511	mk_죽다__010001	317
N_존재하다__000001	1,390	mk_끝나다__000001	293
존재하다__000001	1,148	덜하다__000000	261
N_나타나다__000300	749	돕다__000103	245
mk_들다__010201	643	mk_쓰다__020101	240
떼다__010101	596	매만지다__000001	231
정하다__030201	536	N_성립되다	231
N_오래다	521	N_깨끗하다__000002	226
일어나다__000202	520	비비다__000101	225
N_갖추다__000001	490	약하다__010101	222
향하다__000001	477	거처하다__000000	217
이용하다__010001	476	psv_열다__020101	217

용언이 확장된 의미를 가진 용언인 경우(용언: 누리다_010000). 두 번째는 의미태깅 오류가 있는 경우 (용언: 맛보다_010101). 실제 의미인 맛보다_010102가 태깅된 경우 “치르다_000201>*겪다_000001>맛보다_010102”로 용언군집이 끝난다.

이러한 경우들을 해결하기 위한 방법으로 가장 의미적 포괄성이 높은 용언을 선정해 최상위 용언으로 고정한 뒤 군집화를 시도하거나, 용언의 군집 진행 도중 상위 용언과 하위용언과의 의미자질 비교 등을 통해 상위 계층으로의 연결을 중단하는 등의 방법으로 해결할 수 있을 것이다. 이는 다음 연구로 넘기기로 한다.

층은 16단계였다. <예 5>는 16단계의 계층 구조를 가진 용언 중 최하위어 '표준화하다_000002'의 계층 구조를 보여준다. 계층 구조를 따라 상위용언으로 연결하다 보면 의미적으로 점점 멀어지거나 이질적인 현상이 나타나게 된다. 특히 10단계 이상의 용언 연결 구조에서 특히 이러한 현상이 나타났다.

<예 5> '표준화하다_000002'의 계층구조

mk_나다_010506>내다_020306>짓다_010110>내리다_010402>판단하다_000001>생각하다_000101>살펴보다_000003>알아보다_000101>살피다_010002>보다_010104>바라보다_000001>붓다_020004>집중하다_020002>모으다_000105>통일하다_000102>표준화하다_000002

군집화 방식이 달라서 정확한 비교는 어렵지만, 목적어-동사 관계를 상대 엔트로피를 이용하여 군집화한 결과[4]와 간단히 비교해 보았다. 이 연구는 다의어 단계까지의 의미태그 부착 말뭉치가 없이 군집화를 시도하였다. 그래서 실제 사용한 말뭉치가 다르지만, 본 논문의 의미태그 부착 말뭉치를 사용한 군집화와 의미태그 부착 말뭉치를 사용하지 않은 군집화 연구의 결과 차이를 살펴볼 수 있다. 앞의

표 9. 상대 엔트로피를 이용한 9개의 군집결과

군집 번호	상위 유의어 연결
1	완화하다 철폐하다
2	해독하다 암호화하다
3	팔다 판매하다 거듭하다 되풀이하다
4	극대화하다 향상시키다 약화시키다 회복시키다
5	거론하다 언급하다
6	처치하다 물리치다
7	떠넘기다 회피하다
8	머금다 삼키다 글썽이다
9	끄덕이다 숙이다 갸우뚱하다

연구에서는 <표 9>와 같이 용언들이 9개의 군집으로 묶였다. 이는 목적어를 대상으로 동사를 묶었기 때문에 서로 관련은 있으나 '해독하다, 암호화하다'와 같이 상반된 의미의 동사끼리 묶이는 결과가 나타났다.

'해독하다'는 다의어이며 <그림 8>과 같이 사전에는 3가지 의미의 표제어들이 있고 본 논문의 결과로 군집된 용언들이 포함되어 있다. '해독되다_020002'는 {풀다_000104, 해제하다_010002, 해결하다, 처리하다_000001}와 군집되었고 '암호화하다'는 {바꾸다_0000101, 엇바꾸다_000001, 변환하다_020001, 변형하다_000002, 부호화하다}와 묶여 다른 군집 결과를 보여준다. 이와 같이 의미태그 부착 말뭉치를 이용한 결과가 의미적으로 보다 세밀한 군집을 만드는 것을 볼 수 있다. 의미태그 부착 말뭉치 없이 용언을 군집했을 때 결과가 <그림 8>과 같이 군집되면 이보다 더 좋은 방법이 없겠지만, 현재 군집화 기법으로는 자질과 말뭉치 부족 등으로 인해 이러한 결과를 도출하기 어렵다. 본 논문에서 추출한 군집의 정확성을 검증하고 군집을 세밀화 한다면 다른 용언 의미 군집 연구의 결과와 비교세트로 추천될 수 있을 것이다.

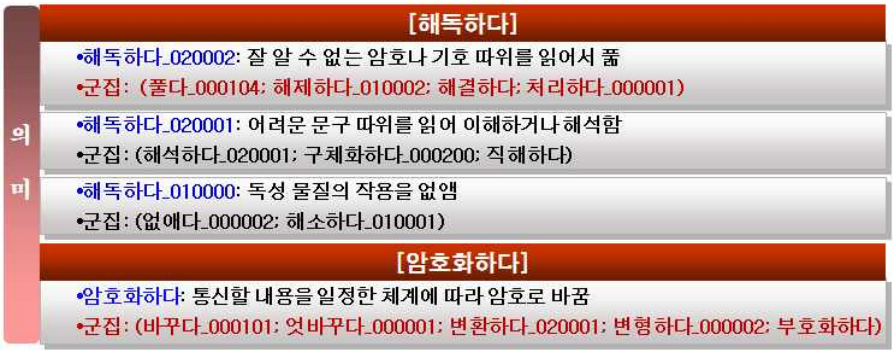

그림 8. '해독하다'와 '암호화하다'의 뜻풀이와 군집 결과

용언의 군집화의 평가를 위해서는 일관된 기준을 설정한 뒤 전문가가 판단하는 방법이 있고, 아니면 이미 정확히 만들어진 군집들과 비교해 보는 방법이 있다. 전자의 방법은 상당한 시일과 인력을 필요로 한다. 후자의 방법을 이용하려면 아주 잘 정의된 용언 의미 군집들이 필요한데, 세종사태의미부류 및 KorLex 등과 서로

비교하여 군집의 적절성 의부를 판단할 수 있다. 본 논문에서는 세종사태의미부류의 용언과 군집화된 용언을 비교하여 용언 의미 군집을 평가해 보았다.

세종의미부류체계는 한국어 어휘들이 표상하는 의미 영역을 엄격하고 정밀하게 분할하고, 각 의미영역에 대해 이를 공유하는 어휘들과 해당 의미 영역의 정의에 근거가 되는 어휘들을 함께 묶어 놓은 위계적 어휘·의미 분류 체계이다. 세종의미부류체계의 의미 영역은 크게 <구체물>, <집단>, <장소>, <추상적 대상>, <사태>의 대부류로 나눈 다음, 각각에 대해 점진적인 의미 분할을 시도하여 <사건>, <상태변화>, <정적사태>, <행위>, <현상>등과 같이 보다 세밀한 의미 영역을 갖는 부류들로 나누어 놓았다. 본 논문에서는 용언 군집의 평가를 위해 <사태> 부류의 <사건>, <상태변화>, <행위>, <현상> 등의 부류에 포함된 용언들과 의미 군집된 용언들을 비교 분석해 보았다.

사태부류와 용언 의미 군집의 비교를 위해 표준국어대사전의 다의어 구분된 용언들과 사태부류의 용언들을 매핑한 자료를 활용하였다. 세종사태의미부류(이하 사태부류)의 용언과 군집된 용언의 매핑된 용언(다의어 단위로 구분)의 개수는 19,096개이며, 2,748개 용언 의미 군집 중 매핑된 용언 의미 군집 개수는 1,297개이다. 이 중 711개 군집은 매핑된 용언 개수가 하나이기 때문에 제거하고, 586개의 군집만 평가에 사용하였다. 매핑된 용언이 하나인 군집은 군집의 정확성 여부를 판단할 수 있는 비교 가능한 용언이 없기 때문에 제거하고, 최소한 2개 이상의 용언이 매핑된 군집을 대상으로 실험을 진행하였다.

용언 군집의 응집도를 확인하기 위해 용언 의미 군집별로 개별 용언으로 쪼갠 후, 각각 용언에 매핑된 사태부류를 수집하였다. 군집이 사태부류 중 어떤 영역에 속해야 하는지 사람이 판단하기 전까지 정확한 답은 구할 수 없지만, 하나의 군집의 용언들이 특정 사태부류 중 하나로 집중적으로 모인다면 용언 의미 군집이 잘 묶였다고 볼 수 있다. 그래서 군집별로 용언들이 속해있는 사태부류를 모아 통계를 낸 뒤, 그 중 가장 많은 용언들이 속한 사태부류를 중심점으로 선택하고, 중심점을 기준으로 상위의 사태부류에 대해서는 0.5의 가중치를 주고 동위와 하위의 사태부류에 대해서는 0.75의 가중치를 주었다. 사태부류에서 상위 계층으로 가면 의미가 포괄적으로 변하기 때문에 의미가 동등하거나 보다 세밀해지는 동위·하위의 사태부류 보다 낮은 가중치를 주었다. 그리고 계층이 3단계보다 위인 사태부류

는 대부분의 부류들을 하위 또는 동위 부류로 포함하기 때문에 군집 응집도를 측정하기에는 부적합하다. 그래서 동위 사태부류 중 계층이 3단계보다 아래인 부류만 가중치 대상으로 제한하였다.

<그림 9>는 최상위 용언이 '비비다_000101|문지르다'인 용언 의미 군집(이하 용언 의미 군집 '비비다_000101|문지르다')과 매핑된 사태부류를 보여준다. 24개의 매핑된 용언 중 15개의 용언이 '① 사태>행위>물리적행위>단독행위>결과행위'에 위치하고, 가장 많은 개수의 용언을 가지므로 의미부류 '결과행위'가 중심점이 된다. '② …결과행위>개선행위'는 '결과행위'의 하위부류이므로 '다리다_010000, 다리미질하다, 닦다_010001' 3개의 용언은 0.75의 가중치를 가진다. '③ …>몸짓및동작'은 '결과행위'와 동위 의미부류이므로 '칫솔질하다' 1개의 용언도 0.75의 가중치를 가진다. 이와 같이 아래의 수식대로 계산해보면 용언 의미 군집 '비비다_000101|문지르다'에 대한 응집도는 87.50%로 나타난다. 수식에서 Coh(T)는 하나의 군집에 대한 응집도이며, T는 용언, K는 군집에 포함된 용언 개수이다. 실험결과 용언 의미 군집의 평균 응집도는 70.14%로 나타났다.

$$Coh(T) = \sum_i \frac{P(T_i)}{K} \quad (1)$$

$$P(T_i) = \begin{cases} Tu: T_i \times 0.5 & \text{중심점보다 상위 부류인 경우} \\ Tl: T_i \times 0.75 & \text{중심점과 동위 또는 하위 부류인 경우} \end{cases}$$

<그림 9>에서 용언군집의 24번째 '갈다_020104'가 사태부류 '⑥ …>생리행위'의 용언으로 포함되어있고, 4번째 '갈아붙이다'는 사태부류 '① …>결과행위'의 용언으로 포함되어있다. '갈다_020104'의 뜻풀이는 "윗니와 아랫니를 마주 대고 문지르다."이고, '갈아붙이다'의 뜻풀이는 "분함을 억제하지 못할 때나 결심을 굳게 할 때, 독한 마음으로 이를 바짝 갈다."이다. 뜻풀이를 확인해보면 '갈아붙이다'는 "이를 마주 대고 문지르다"는 의미보다는 "독한 마음을 먹고 벼르다."라는 의미가 있다. 그러나 용언 군집에서는 '…문지르다>**갈다_020104>갈아붙이다**'처럼 같은 군집의 하위 계층으로 '갈아붙이다'가 포함되어 있는 것을 볼 수 있다. 이는 실제 사전의 '갈다' 뜻풀이 중 '벼르다'는 의미의 뜻풀이가 없고, 뜻풀이 중 가장 유사한 뜻풀이로 의미 태그 부착을 하였기 때문에 위와 같은 결과를 보인 것으로 판단된다. 만약 관용구 사전이 있다면 이와 같은 경우를 올바로 처리할 수 있을 것이다.

그림 10. 용언군집과 사태부류의 비교

<그림 10>은 '치다_020002', '굴리다_000108', '경치다_010000'의 용언이 사태부류와 용언군집에 포함된 위치를 보여준다. 용언의 뜻풀이를 보면 '치다_020002'가 '경치다_010000'와 '굴리다_000108'보다 의미가 더 포괄적인 것을 알 수 있다. 그래서 '굴리다_000108'과 '경치다_010000'은 '치다_020002'의 하위 계층으로 둘 수 있다. 이와 같이 용언 의미 군집은 사태부류보다 세밀하게 군집되었다는 것을 알 수 있다.

결 론

본 논문에서는 다의어 기반으로 용언을 의미 군집화 하였기 때문에, 질의응답에서 질의를 인식하는 기존의 방법으로는 처리하기 까다로운 요소들을 용언 의미 군집정보를 이용해 간단히 처리할 수 있다. 앞의 서론에서 살펴보았던 {죽다, 돌아가다, 사망하다}에 대한 군집화 및 계층을 살펴보면 <그림 11>과 같다.

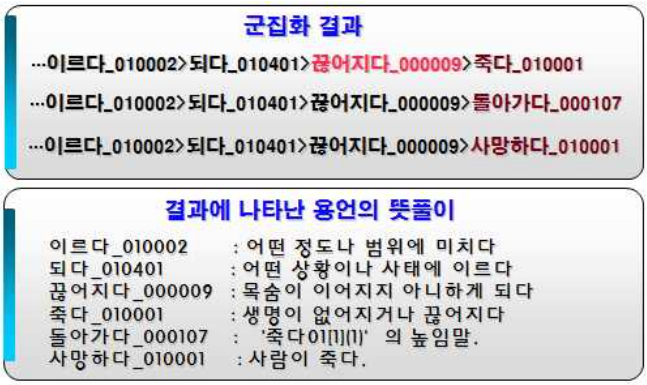

그림 11. '죽다_010001/돌아가다_000107/사망하다_010001'의 군집화 결과

위와 같이 서론의 <예 1>에서는 제시하지 못했던 "000의 목숨이 **끊어진** 날짜는?"과 같은 유형의 질의문도 용언 의미 군집정보를 통해 부가적으로 '죽다'라는 요지로 통합해 처리할 수 있다. 정보처리 담당자가 미처 생각하지 못했던 유형이나 의미 정보가 미흡하여 처리하지 못했던 유형들을 용언 의미 군집정보를 이용하면 보다 쉽게 추가적인 데이터 구축 없이 질의문을 인식할 수 있다.

본 논문에서는 사전에 존재하는 대부분의 용언을 대상으로 군집화하고 계층을 만들어 보았다. 지금까지의 방법론과는 달리 의미태그가 부착된 사전의 뜻풀이를 이용하여 bottom-up방식으로 용언의 집합을 만들고, 유의어 집합에서 자주 나타나는 용언의 계층구조를 분석하여 간단한 용언 계층구조를 만들었다. 그리고 세종사태의미부류와 비교를 통해 용언 의미 군집의 평가를 시도하였다. 그 결과 70.14%의 응집도를 보였다.

가장 계층의 깊이가 큰 군집은 16단계의 계층 구조를 이루고 있는데 최하위 용언에서 최상위 용언으로 연결하는 과정 중 의미적 결합이 약해지거나 자칫 의미적으로 이질성이 나타날 수도 있기 때문에, 제시한 용언 의미 군집을 보다 세부적인 용언 의미 군집을 만들 필요가 있다. 세부적인 용언 의미 군집을 만들기 위해 뜻풀이 용언 근처의 부사 정보와 논항정보 및 제약정보 등의 결합적 정보를 이용하여 세부적인 하위 군집화를 시키거나 다른 군집으로 분류하는 연구 및, 가장 의미적 포괄성이 높은 용언을 선정해 최상위 용언으로 고정한 뒤 군집화를 시키는 연구 등을 시도할 것이다.

참고문헌

[1] 남경완 (2008). **국어 용언의 의미 분석.** 태학사.

[2] Chagnaa Altangerel, 이창범, 옥철영 (2006). 독립성분분석과 Box-Cox 변환을 이용한 동사 개념 클러스터링. **제 18회 한글 및 한국어 정보처리 학술대회**, pp. 164-170.

[3] Schulte im Walde, S. (2000). Clustering verbs seman-tically according to their alternation behaviour, Proceedings of the 18th International Conference on Computational Linguistics (COLING2000), pp.747-753.

[4] 박성배, 장병탁, 김영택 (2000). Self-Organizing Map을 이용한 한국어 동사 클러스터링. **한국정보과학회 가을 학술발표논문집**(Ⅱ), pp.183-184.

[5] Lin Sun and Anna Korhonen. (2009). Improving Verb Clustering with Automatically Acquired Selectional Preferences, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp.638-647.

[6] 김혜경, 윤애선 (2006). 단어클러스터링을 이용한 동사 어휘의미망의 활용 및 평가. **한국인지과학회 춘계학술대회**, pp.1-7.

[7] 김미훈, 남상협, 이용훈, 이종혁 (2007). EM 알고리즘을 이용한 전문용어 온톨로지 클래스간 관계 정의를 위한 동사 클러스터링. **제19회 한글 및 한국어 정보처리 학술대회,** pp.233-240.

[8] 김종복, 임경섭, 이주원. (2006). 전산학적 응용을 위한 한국어 용언 위계 구축. **한국사전학회,** 제8호, pp.145-162.

[9] Jinying Chen and Martha Palmer. (2004). Chinese Verb Sense Discrimination Using an EM Clustering Model with Rich Linguistic Features, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp.295-302.

[10] Andreas Vlachos, Zoubin Ghahramani, and Anna Korhonen, (2008), Dirichlet process mixture models for verb clustering, In Proceedings of the ICML workshop on Prior Knowledge for Text and Language.

[11] 김혜경, 윤애선. (2006). 동사 어휘의미망의 반자동 구축을 위한 사전정의문의 중심어 추출. 한국언어정보학회, 10-1, pp.47-69

[12] 윤애선, 황순희, 이은령, 권혁철. (2009). 한국어 어휘의미망 “KorLex 1.5”의 구축. 한국정보과학회, 36-1, pp.92-108

1 차원고접수 : 2011. 5. 20
2 차원고접수 : 2011. 7. 22
최종게재승인 : 2011. 8. 11

(*Abstract*)

Semantic Clustering of Predicates using Word Definition in Dictionary

Young-Jun Bae[1] Ho-Seop Choe[1] Yoo-Hwa Song[2] Cheol-Young Ock[3]

[1)]Korean Institute of Science and Technology Information

[2)]University of Science and Technology

[3)]Dept. of Computer Engineering and Information Technology, University of Ulsan

The lexical semantic system should be built to grasp lexical semantic information more clearly.

In this paper, we studied a semantic clustering of predicates that is one of the steps in building the lexical semantic system. Unlike previous studies that used argument of subcategorization(subject and object), selectional restrictions and interaction information of adverb, we used sense tagged definition in dictionary for the semantic clustering of predicate, and also attempted hierarchical clustering of predicate using the relationship between the generic concept and the specific concept. Most of the predicates in the dictionary were used for clustering.

Total of 106,501 predicates(85,754 verbs, 20,747 adjectives) were used for the test. We got results of clustering which is 2,748 clusters of predicate and 130 recursive definition clusters and 261 sub-clusters. The maximum depth of cluster was 16 depth. We compared results of clustering with the Sejong semantic classes for evaluation. The results showed 70.14% of the cohesion.

Key words : semantic clustering, predicate, dictionary's definition, predicate in definition