격틀 구조에 기반한 유사 동사 추출

조정현○ 정현기 김유섭
한림대학교 컴퓨터공학과
{showcjh, mayapple, yskim01}@hallym.ac.kr

Similar Verb Words Extraction based on their Case Frame Structure

Junghyun Cho○ Hyunki Jung Yu-Seop Kim
Dept. of Computer Engineering, Hallym University

요 약

한국어 Propbank를 구축하기 위해서는 유사 동사를 군집화하고 군집에 포함되는 동사들의 구문 및 의미 특성을 모아놓은 정보가 필요하다. 본 연구에서는 이러한 군집화의 초기 단계로써 개별 동사들의 격틀 구조에 기반하여 동사간의 유사도를 추정하여 유사 동사를 추출하고자 하였다. 본 연구는 개별 동사의 격틀 정보를 추출하기 위하여 세종 계획의 용언 사전과 KAIST 언어자원의 동사 격틀 사전을 활용하였다. 또한 격틀을 세분화하여 보다 상세한 격틀 정보를 생성하기 위하여 격틀이 가지고 있는 논항의 특성을 활용하였다. 동사의 유사도를 측정하기 위하여 개별 동사들은 벡터로 표현하였고, 벡터의 원소는 해당 동사가 다른 동사와 세분화된 격틀을 공유하는 정도로 하였다. 실험에서는 두 용언 사전에서 개별적으로 위의 과정을 진행하여 각 동사와 유사한 동사들을 추출하였다.

주제어: 한국어 Propbank, 유사 동사 추출, 동사 군집화, 용언 격틀 정보

1. 서 론

PropBank[1]은 동사의 술어-논항 (Predicate-Argument) 구조를 태그해 놓은 말뭉치로써 영어의 경우 의미 역 결정 (Semantic Role Labeling)에 단독 또는 복합적으로 활용되고 있다[2, 3, 4, 5, 6, 7]. 한국어의 경우 이러한 말뭉치가 아직 구축되지 않아 의미 역 결정과 같은 문장 단위 의미 분석 관련 연구에 큰 어려움을 겪고 있다[8, 9].

본 연구팀에서는 이러한 어려움을 해소하고자 한국어 Propbank를 구축하고자 하는데, 이의 전 단계로서 기존의 구문 태그된 말뭉치에 술어-논항 정보를 초벌로 자동 태그해주는 자동 술어-논항 분석기[10]를 구현하고자 한다. 이 분석기의 구현을 위해서는 VerbNet[11]과 같은 동사 사전이 필요한데, 본 연구에서는 한국어 VerbNet의 기초 작업의 하나로써 격틀에 기반한 동사 유사도를 측정하여 개별 동사의 유사 동사를 추출하고자 하였다.

본 연구에서는 기존의 격틀 정보를 추출하기 위하여 세종 계획[12]의 용언 사전과 KAIST 언어자원[13]의 동사 격틀 사전을 활용하였다. 본 실험을 두 자원을 통합하지 않고 병렬적으로 사용하였는데, 이는 두 자원의 본 연구와의 적합도를 가늠하기 위함이다. 사전에서 추출된 격틀 정보는 그 자체만으로는 지나치게 광범위하여 직접 활용될 수 없기 때문에 격틀이 가지고 있는 논항의 특성을 활용하여 격틀 정보를 세분화 하였다. 세종 계획과 KAIST 언어자원은 모두 체언의 의미 체계를 가지고 있어, 논항 특성 분석에 이를 활용하였다. 개별 동사들은 유사도 측정을 위하여 벡터로 표현되었는데, 벡터의 원소를 계산하기 위하여 해당 동사가 다른 동사와 세분화된 격틀을 얼마나 공유하는 가를 계산하였다. 또한 유사도는 코사인 값을 계산하여 추정하였다. 실험에서는 개별 동사와 유사한 것으로 나타난 동사의 예를 보여주는데, 이를 세종 계획 및 KAIST 용언 사전을 서로 비교하며 보여준다.

2장에서는 본 연구에서 활용된 언어 자원인 세종 계획의 용언 사전과 KAIST 언어자원의 동사 격틀 사전에 대하여 상세히 기술하였다. 3장에서는 두 자원에서 추출된 격틀 정보 및 이들 정보를 세분화하여 실제 벡터를 구성하여 유사도를 추정하는 과정을 보여준다. 4장에서는 두 자원으로 추출된 유사 동사들을 서로 비교하며 이에 대하여 논하였고, 마지막으로 5장에서는 본 연구를 요약하고 향후 연구에 대하여 기술하였다.

2. 언어 자원

본 연구에서 사용한 언어자원은 21세기 세종 계획의 용언 사전과 KAIST 언어자원의 동사 격틀 사전이다. 세종 계획 용언 사전은 총 15,174개의 동사와 1,269개 유형의 격틀로 이루어져 있고, KAIST 언어자원의 동사 격틀 사전은 2,731개의 동사와 304개 유형의 격틀로 되어있다.

언어자원 파일의 형태는 세종 계획 용언 사전의 각 동사들은 XML 파일로 되어있고 KAIST의 동사격틀 사전의 각 동사들은 TEXT 파일로 되어있다.

그림 1은 세종 계획 용언 사전에서 '가꾸다' 라는 동사의 XML 파일 내용의 주요 부분을 보여주고 있다. 각각 세부적으로 살펴보면 <sense n="01">은 표제항이 갖는 각각의 의미를 구별하여 의미 정보 및 통사 정보를 제공하고 있는 센스 구획을 나타낸다. <sem_class> ... </sem_class>는 의미 부류를 나타내고 <trans> ... </trans>는 영어 대역어를 나타낸다. 그리고 가장 중요한 문형 및 논항에 관한 정보를 제시하는 문형 구성 구획인 <frame_grp type="FTR"> ... </frame_grp> 이 있다. 여기서 "FTR"은 일반타동사를 의미한다. 이 안에 있는 <frame>X=N0-이 Z=N2-에 Y=N1-을 V</frame>은 문형(격틀) 정보를 보여준다. 안에 있는 <subsense>는 하위 센스별로 각 논항의 선택제약을 제시한다. 이 <subsense>는 선택 제약이 차이나는 경우에 여러 개가 존재 할 수 있다. 그 안에는 용언의 의미역 및 선택 제약을 나타내는 <sel_rst arg="X" tht="AGT">인간 </sel_rst>, <sel_rst arg="Y" tht="THM">식물 </sel_rst>, <sel_rst arg="Z" tht="LOC">지상장소 </sel_rst> 가 있다. 첫 번째 "X"에는 "AGT"(행위주) 의미역과 의미부류인 인간이 들어가고, 두 번째 "Y"에는 "THM"(대상) 의미역과 의미부류인 식물이 들어가고, 세 번째 "Z"에는 "LOC"(장소) 의미역과 의미부류인 지상장소가 들어간다. 마지막으로 <eg> ... </eg>는 위의 문형과 선택 제약을 반영하는 전형적인 용례를 나타낸다. <sense n="02"> 부분도 <sense n="01"> 부분과 마찬가지로 나타내고 있다.

그림 2 는 KAIST 언어자원 동사 격틀 사전에서 '가꾸다' 라는 동사의 TEXT 파일 내용을 보여주고 있다. KAIST 언어자원의 동사 격틀 사전은 세종 계획의 용언 사전과는 형태가 조금 다르다. '1. 키우다/손질하다'는 '가꾸다'의 의미 분류 중 하나이다. 아래 2번과 3번도 마찬가지로 '가꾸다'의 의미 분류이다. 세종 용언 사전과 비교하면 <sense n="01"> 부분과 같다고 할 수 있다. '1. 키우다/손질하다' 아래의 '(1) N0이 N1을 v' 는 격틀 구조를 나타낸다. 이것은 세종 용언 사전의 <frame> 과 같다고 할 수 있다. 이 격틀 구조는 하나의 의미 분류에서 다른 여러 개가 나타날 수 있다. 그 아래에는 각 N0, N1에 들어갈 수 있는 개념과 세부 명사를 나타내고 있다. 첫 번째를 보면 'N0:[5 인간] 사람 N1:[672 식물(개체)] 화초' 에서는 [ ] 안의 5, 672는 개념번호로 KAIST 언어자원에서 구축한 개념체계의 일련번호 이며, 인간과 식물(개체)은 가각의 해당 개념 번호의 개념명칭 이다. 사람과 화초는 각각의 개념에 해당하는 명사이다. 그 아래에는 다른 개념 또는 다른 세부 명사에 해당되는 형태를 나타내고 있다.

```
- <sense n="01">
  - <sem_grp>
      <sem_class>지속적활동</sem_class>
      <trans>cultivate</trans>
      <trans>grow</trans>
    </sem_grp>
  - <frame_grp type="FTR">
      <frame>X=N0-이 Z=N2-에 Y=N1-을 V</frame>
    - <subsense>
        <sel_rst arg="X" tht="AGT">인간</sel_rst>
        <sel_rst arg="Y" tht="THM">식물</sel_rst>
        <sel_rst arg="Z" tht="LOC">지상장소</sel_rst>
        <eg>할머니께서는 앞 마당에 오이와 호박을 정성껏 가꾸시어 매년 집으로 보내 주신다.</eg>
        <eg>수진이는 멋진 집을 짓고, 집앞 정원에 꽃을 가꾸고 사는 것이 소원이다.</eg>
      </subsense>
    </frame_grp>
  </sense>
- <sense n="02">
  - <sem_grp>
      <sem_class>지속적활동</sem_class>
      <trans>decorate</trans>
    </sem_grp>
  - <frame_grp type="FTR">
      <frame>X=N0-이 Y=N1-을 V</frame>
    - <subsense>
        <sel_rst arg="X" tht="AGT">인간</sel_rst>
        <sel_rst arg="Y" tht="THM">(방|집|건물내부)</sel_rst>
        <eg>호진이는 집을 가꾸는 일에 여념이 없었다.</eg>
      </subsense>
    </frame_grp>
  </sense>
```

그림 1 세종계획 용언 사전 XML 형태

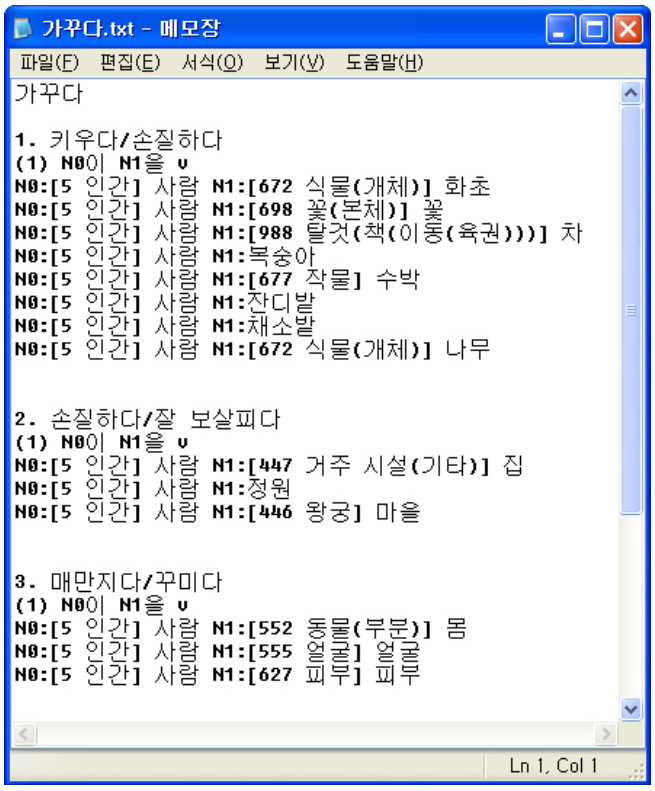

그림 2 KAIST 동사 격틀 사전 TEXT 형태

위와 같은 형식으로 된 두 개의 언어자원의 격틀 정보를 이용해 동사들을 나누게 될 것이다.

3. 동사 분류

3.1 격틀별, 세부 유형별 동사 분류

이 두 개의 언어자원을 가지고 동사들을 분류 하기위해 먼저 각 언어자원의 격틀별로 동사를 나누고 다시 세부 유형별로 동사를 나눈다.

각 언어 자원에서 크게 격틀 별로 동사를 분류하게 되면 아래 표 1에서 보는 바와 같은 격틀들로 분류된다. 이 표 1은 각 언어 자원 동사 격틀의 일부 예를 보여주는 것이다. 같은 격틀별로 동사들을 분류하게 되면 세종 용언 사전의 격틀 수는 총1,269개 이고, KAIST 동사 격틀 사전의 격틀 수는 총304개가 된다.

표 1 각 언어 자원의 격틀의 예

KAIST 격틀	세종 격틀
N0과 N1이 v	X=N0-이 Y=N1-에 V
N0에 v	X=N0-이 Y=N1-에서 V
N0은 N1을 v	X=N0-이 Z=N2-에 Y=N1-을 V
N0이 N1(으로) v	X=N0-이 Y=N1-을 Z=N2-로 V
N0이 N1과 N2(에) v	X=N0-이 Y=N1-을 V
N0이 N1과 v	X=N0-이 V
N0이 N1에 N2을 v	X=N0-이 Y=N1-로 V
...	...

이렇게 나눈 격틀을 세부 유형별로 다시 나눈다. 세종 용언 사전 격틀의 세부 유형별 분류 방법은 각 용언에서 논항의 의미역과 선택 제약을 통해 나눈다. 그림 1 XML 파일의 <sel_rst arg="X" tht="AGT">인간</sel_rst> 에서 논항 X의 의미역은 'AGT(행위주)'가 되고 선택 제약은 '인간'이다. 의미역은 표 2와 같이 15가지로 구분되고 선택 제약은 세종 명사 의미 부류 체계에 따라 제공한다. 그림 3에서와 같이 세종 의미 부류 체계는 '구체물, 집단, 장소, 추상적 대상, 사태' 5가지의 최상위 개념을 기준으로 총 645개의 개념들이 있다. 본 연구에서는 상위 2단계 개념 69개와 최상위 개념 5개를 통해 동사를 분류하였다. 예를 들면 '인간'의 경우는 세종 명사 부류 체계에서 '구체물/구체자연물/생물/인간' 이다. 그러므로 인간은 2단계 개념인 '구체자연물' 클래스로 분류된다. 그리고 선택 제약이 최상위 개념은 경우는 독립적으로 분류한다.

표 2 세종 의미역 목록

Attribute	Value
AGT	행위주
EXP	경험주
MAG	심리행위주
COM	동반주
THM	대상
LOC	장소
DIR	방향
GOL	도착점
FNS	결과상태
SRC	출발점
INS	도구
EFF	영향주
CRT	기준치
PUR	목적
CNT	내용

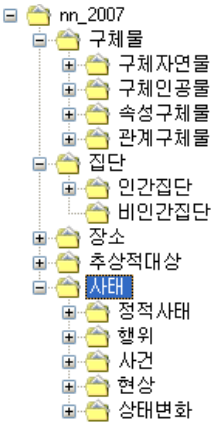

그림 3 세종 명사의미 부류 체계

KAIST 동사 격틀 사전의 세부 유형별 분류 방법은 예를 들어 2장 그림 2에서 본 'N0:[5 인간] 사람 N1:[672 식물(개체)] 화초' 와 같은 세부 유형으로 나누는 것이다. 하지만 이렇게 나누게 되면 격틀이 너무 많은 수로 세분화 되고 하나의 세부 유형에는 하나 또는 두 개의 동사만 포함될 뿐이다. 따라서 이렇게 나누는 것은 동사를 분류 하는데 의미가 없을 것이다. 그래서 세부 유형을 더 큰 단위로 묶을 필요가 있다. 한 가지 방법으로 세부 유형에서 개념들을 그것들의 관계에 따라 큰 단위로 묶는 방법이 있을 수 있다. 이것이 가능한 이유는 KAIST 언어자원 중에 코어넷 한국어 명사편이 있기 때문이다. 이것은 개념체계 내에서의 위치에 관한 정보, 즉 개념들의 상하위 개념과 단계 정보를 나타내 주고 있다. 코어

넷 명사편에서 하나의 예를 보면 '인간 [11111, 5]' 과 같은 형식으로 되어 있다. 여기서 인간은 개념명, [ ] 안의 뒤에 숫자 5는 인간의 개념 번호이다. 앞의 숫자 11111은 인간이라는 개념의 상위 개념과 단계를 표현한 것이다. 자릿수 5는 단계를 뜻하며 11111의 마지막 자릿수 1을 빼면 인간의 상위 단계인 1111(사람)이 나온다. 마지막 자릿수를 다시 빼면 1111(사람)의 상위 단계 개념이 나오게 된다. 이런 형식으로 개념들 간의 상하위 개념과 단계를 알 수 있다.

KAIST 언어자원의 개념은 상하위 11단계로 총 2,938개의 개념들이 있다. 본 연구에서 KAIST 언어자원의 개념을 묶는 단위는 상위 4단계까지의 개념들을 기준으로 묶었다. 상위 4단계의 개념은 21개이고 상위 1, 2, 3단계의 개념은 8개이다.

표 3은 상위 4단계에 있는 개념들 21개와 그 개념의 하위에 포함되는 개념들의 개념 번호를 나타낸 개념 클래스 표이다. 이 개념 클래스에 따라서 격틀 내의 유형들에서 오른쪽의 포함되는 개념들이 나오면 왼쪽의 상위 4단계 개념으로 바꾸어 분류한다. 예를 들어 위에서 봤던 'N0:[5 인간] 사람 N1:[672 식물(개체)] 화초'를 표 2에 따라서 바꾸게 되면 'N0:[4 사람] N1:[533 생물]' 이 되는 것이다.

표 3 상위 4단계 개념 클래스

개념 클래스	
상위 4단계 개념	포함되는 개념
사람 [1111, 4]	4 - 361
조직 [1112, 362]	362 - 387
시설 [1121, 389]	389 - 457
지역 [1122, 458]	458 - 467
자연 [1123, 468]	468 - 532
생물 [1131, 533]	534 - 705
무생물 [1132, 534]	706 - 999
추상물(정신) [1211, 1002]	1002 - 1153
추상물(행위) [1212, 1154]	1154 - 1235
인간활동 [1221, 1236]	1236 - 2053
사실/현상 [1222, 2054]	2054 - 2303
자연현상 [1223, 2304]	2304 - 2421
존재 [1231, 2423]	2423 - 2431
동류/동계 [1232, 2432]	2432 - 2442
관련 [1233, 2443]	2443 - 2482
성질 [1234, 2483]	2483 - 2506
상태 [1235, 2507]	2507 - 2563
형상 [1236, 2564]	2564 - 2584
수량 [1237, 2585]	2585 - 2609
장소(추상적관계) [1238, 2610]	2610 - 2669
시간 [1239, 2670]	2670 -

표 4는 상위 1, 2, 3단계의 개념들을 보여준다. 이 개념들을 포함하는 개념은 없다. 따라서 상위 1, 2, 3단계의 8개의 개념들은 따로 개념 클래스에 포함되지 않고 독립적인 개념으로 표시한다.

표 4 상위 1, 2, 3단계 개념

구체/추상 [1, 1]
구체 [11, 2]
주체 [111, 3]
장소<구체> [112, 388]
물건 [113, 533]
추상 [12, 1000]
추상물 [121, 1001]
추상적관계 [123, 2422]

3.2 벡터를 이용한 동사 유사도 측정

이렇게 격틀별, 세부 유형별로 동사들을 분류해 놓고 각 동사가 가지고 있는 세부 유형을 다른 동사들의 세부 유형과 비교하기 위해서는 각각의 동사들이 가지고 있는 세부 유형들을 알아야 한다. 위에서 나눈 세부 유형별 동사를 반대로 동사별 세부유형으로 바꾸어 비교를 하게 된다. 아래 그림 4가 동사 '가꾸다'의 동사별 세부유형을 정리한 파일의 예이다.

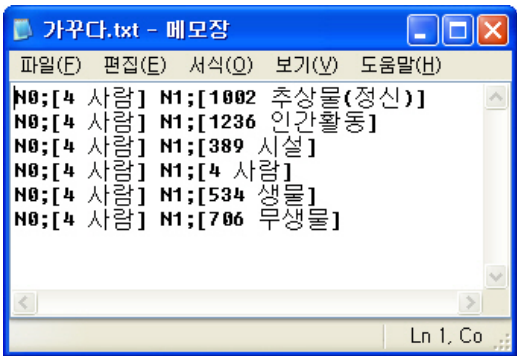

그림 4 동사별 세부 유형

동사들의 세부 유형 비교는 세종 용언 사전의 동사들과 KAIST 동사 격틀 사전의 동사들을 따로 비교한다. 먼저 하나의 동사는 다른 모든 동사와 세부 유형을 비교한다. 두 동사의 세부 유형 비교 방법은 아래 식 (1)을 이용한다. 그러면 최대 1 최소 0의 값을 가지게 된다.

$$\text{세부유형비교식} = \frac{\text{겹치는세부유형수}}{\text{둘중유형의개수가작은수}} \quad (1)$$

위의 식 (1)을 통해 나온 결과가 벡터의 원소가 된다. 이렇게 차례로 각 동사가 다른 모든 동사들과 세부 유형 비교를 마치게 되면 하나의 동사마다 동사 개수 만큼의 크기, 즉 세종 동사들은 15,174개의 벡터, KAIST 동사들은 2,731개의 벡터가 만들어진다. 아래 그림 5는 동사 '가꾸다'의 벡터의 예를 보여 준다.

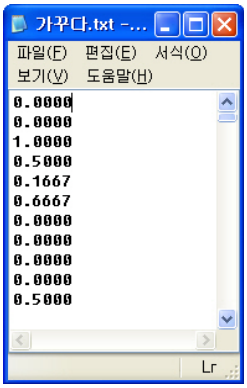

그림 5 동사별 벡터

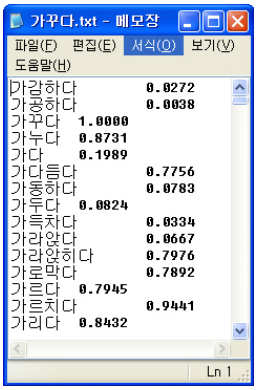

그림 6 유사도

이렇게 각각의 동사마다 벡터가 만들어지면 벡터를 이용해 유사도를 구한다. 유사도는 아래 수식 (2)와 같이 계산한다.

$$\text{유사도}(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\|\vec{y}\|} = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}x_i^2} \times \sqrt{\sum_{i=1}^{n}y_i^2}} \qquad (2)$$

유사도 역시 하나의 동사는 모든 동사들과의 유사도를 구한다. 그러면 위의 그림 6 '가꾸다' 동사의 유사도 예처럼 하나의 동사는 다른 모든 동사들과의 유사도를 가지고 있게 된다.

4. 실험 및 결과

세종 용언 사전과 KAIST 동사 격틀을 3장에서 계산된 각 동사별 유사도를 통해 동사들을 비교한다. 실험을 위하여 기초 어휘 빈도 조사 결과[14]에서 빈도가 높은 동사 10개와 상대적으로 빈도가 낮은 동사 10개를 추출하였고, 세종 용언 사전과 KAIST 격틀 사전에서 유사 단어를 각각 5개씩 뽑아서 어떤 동사들이 나오고 있는지 서로 비교하였다. 표 5는 세종 용언 사전 동사와 KAIST 격틀 사전 동사에서 고빈도 동사의 유사어를 뽑아 비교한 표이며 ( )안의 숫자는 유사도 이다.

표 5 고빈도 동사의 유사어

동사	만나다	들어오다	마시다	넣다	맞다
세종	사귀다 (0.976) 상봉하다 (0.957) 회견하다 (0.957) 대면하다 (0.951) 접하다 (0.877)	가입하다 (0.627) 입학하다 (0.621) 입단하다 (0.621) 입대하다 (0.621) 취학하다 (0.621)	썰다 (1.000) 먹고떨어지다 (1.000) 둘러매다 (1.000) 쓸어내리다(1.000) 토막내다 (1.000)	맡기다 (0.629) 보태다 (0.623) 등재하다 (0.589) 선적하다 (0.566) 투입하다 (0.560)	증폭하다 (0.899) 배역하다 (0.874) 세력화시키다 (0.860) 추첨하다 (0.859) 구제하다 (0.849)
KAIST	기다리다 (0.870) 쳐다보다 (0.840) 해치다 (0.837) 모시다 (0.816) 따라다니다(0.793)	부임하다 (0.725) 입학하다 (0.722) 편입하다 (0.612) 합격하다 (0.604) 입소하다 (0.563)	쐬다 (0.946) 개봉하다 (0.910) 조리하다 (0.910) 시식하다 (0.910) 흡입하다 (0.910)	건네다 (0.349) 걸치다 (0.198) 켜다 (0.150) 뿌리다 (0.148) 기울이다 (0.131)	피하다 (0.843) 까다 (0.841) 패다 (0.827) 부리다 (0.811) 깨물다 (0.807)

표 6은 저빈도 동사들과 높은 유사도를 나타낸 동사들을 비교한 것이며 ( )안의 숫자는 유사도 이다.

표 6 저빈도 동사의 유사어

동사	가득차다	계획하다	개발하다	대응하다	덤비다
세종	뒤덮이다 (0.903) 바글대다 (0.903) 우글거리다(0.903) 재배치되다(0.868) 위치되다 (0.865)	알아채다 (0.980) 당연시하다(0.958) 갈구하다 (0.929) 감질내다 (0.922) 검토하다 (0.922)	히트시키다(0.974) 운용하다 (0.964) 표준화시키다 (0.964) 공증하다 (0.964) 실용화하다(0.964)	혼용되다 (0.845) 결합되다 (0.841) 상호작용하다 (0.814) 얽히고설키다 (0.805) 평행하다 (0.767)	종종거리다(1.000) 종종대다 (1.000) 깐죽깐죽하다 (1.000) 삿대질하다(1.000) 으르렁대다(0.988)
KAIST	휩싸이다 (0.662) 물들다 (0.658) 넘치다 (0.622) 놓이다 (0.615) 중복하다 (0.599)	주최하다 (1.000) 기획하다 (1.000) 개최하다 (1.000) 강행하다 (1.000) 결행하다 (1.000)	관측하다 (0.925) 개량하다 (0.925) 발명하다 (0.889) 수입하다 (0.881) 게시하다 (0.875)	보복하다 (0.953) 대처하다 (0.953) 간섭하다 (0.905) 대전하다 (0.903) 경합하다 (0.903)	응전하다 (0.917) 달려들다 (0.911) 경주하다 (0.881) 달라붙다 (0.862) 어울리다 (0.840)

위 결과를 보면 고빈도 동사들에 비하여 저빈도 동사들이 직관적으로 더 유사한 동사들을 추출하는 것을 알 수 있다. 고빈도 동사의 경우 격틀의 수가 매우 많고 다의어인 경우가 많아 유사성이 집중되지 못하여 이러한 현상이 벌어지는 것으로 생각된다. 반면에 저빈도 동사의 경우에는 비교적 의미가 명확하고 격틀 사전 구축에 활용된 말뭉치에서도 다양한 형태로 나타나지 않았기 때문에 비교적 단순한 격틀 구조를 가진 것이 이런 현상의 원인이라 생각된다.

5. 결론 및 향후 연구방향

본 연구는 한국어 PropBank 구축의 가장 기초적인 단계로써 유사한 동사를 추출하고자 하였다. 이를 위하여 세종 계획 및 KAIST의 동사 격틀 사전을 활용하였고, 이들 언어 자원에서 제공되는 의미 체계를 활용하여 격틀의 논항을 분류하였으며, 이로써 격틀을 보다 세분화할 수 있었다. 또한 동사를 타 동사와 격틀을 공유하는 정도로 벡터를 구성하여 유사도를 추정할 수 있었다.

이 연구에서의 유사도는 전적으로 격틀(세부 유형)에 기반한 것이다. 즉 구문 정보(격틀의 세부 유형)에 지나치게 의존하는 측면이 강하다. 비록 PropBank가 격틀에 기반한다고는 하지만 동사의 의미적인 측면 역시 매우 중요하기 때문에 부족함이 있다. 따라서 향후 연구로는 격틀 정보 뿐만 아니라 동사 의미 체계도 함께 활용하여 동사의 유사도를 추정해야 한다. 또한 한국어 VerbNet의 구축을 위해서는 이러한 유사도를 기반으로 결국 군집화를 해야 할 것이다.

참고문헌

[1] Palmer, M., P. Kingsbury, and D. Gildea, "The Proposition Bank: An Annotated Corpus of Semantic Roles," *Computational Linguistics*, **31**(*1*), pp.71-106, 2005.

[2] Xue, N., and M. Palmer, "Automatic Semantic Role Labeling for Chinese Verbs," *Procs. of International Joint Conference on Artificial Intelligence*, 2005.

[3] Kingsbury, P., B. Snyder, N. Xue, and M. Palmer, "PropBank as a Bootstrap for Richer Annotation Schemes," *Procs. of sixth Workshop on Interlinguas*, Machine Translation Summit IX, 2003.

[4] Johansson, R., and P. Nugues, "Dependency-based Syntactic-Semantic Analysis with PropBank and NomBank," Procs. of CoNLL-2008, 2008.

[5] Giuglea, A., and A. Moschitti, "Knowledge Discovering using FrameNet, VerbNet and PropBank," *Workshop on Ontology and Knowledge Discovery at ECML-04*, 2004.

[6] Hovy, E., M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "OntoNotes: The 90% Solution," *Procs of the Human Language Technology Conference of the NAACL*, 2006.

[7] Giuglea, A., and A. Moschitti, "Semantic Role Labeling via FrameNet, VerbNet and PropBank," *Annual Meeting of Association for Computational Linguistics,* 2006.

[8] 김병수, 이용순, 나승훈, 김병기, 이종혁, "부트스트래핑 알고리즘을 이용한 한국어 격조사의 의미역 결정," *한국정보과학회 2006 한국컴퓨터종합학술대회 논문집(B)*, pp.4-6, 2006.

[9] 김병수, 이용훈, 이종혁, "비지도 학습을 기반으로 한 한국어 부사격의 의미역 결정," *정보과학회논문지:소프트웨어및응용*, 제34권 제2호, pp.112-122, 2007.

[10] Palmer, M., J. Rosenzweig, and S. Cotton, "Automatic Predicate Argument Analysis of the Penn Treebank," Procs. of HLT 2001, First International Conference on Human Language Technology Research, 2001.

[11] Schuler, K. K., "VerbNet: A broad-coverage, comprehensive verb lexicon", Dissertations of University of Pennsylvania, 2005.

[12] http://www.sejong.or.kr

[13] 최기선(2001), KAIST 언어자원 2001년도판, 과학기술부 핵심 소프트웨어 과제 결과물 2000 (http://kibs.kaist.ac.kr)

[14] 서상규, 남윤진, 진기호, "한국어 세계화 추진을 위한 기반 구축 사업 1차년도 보고서," 문화관광부 한국어 세계화 추진 위원회, 1998.