논항의 의미 정보를 이용한 동사의 유사도 추정[1)]

이채훈○, 석미란, 김유섭
한림대학교, 유비쿼터스 컴퓨팅학과
chaehoon@naver.com, smr4880@hanmail.net, yskim01@hallym.ac.kr

Similarity Estimation between Verbs Using Semantic Information of their Argument

Chae-Hun Lee○, Mi-Ran Seok, Yu-Seop Kim
Dept. of Ubiquitous Computing, Hallym University

요 약

한국어의 경우 동사와 형용사는 문장에서의 역할이 명사와는 다르며, 동사의 의미는 동반하는 논항의 의미적, 통사적 특성에 따라 분화되므로 근본적으로 논항과 함께 고려되어야 한다. 논항이라 함은 명제를 표시하는 방법 중 하나로 관계와 논항으로 표시하는 방법이 있는데, 여기서 관계는 문장의 동사, 형용사 또는 다른 관계항에 해당하며, 논항은 특정시간, 장소, 사람, 대상을 지칭하는 것으로서 흔히 명사에 해당한다. 본 논문에서는 동사간의 의미 유사도를 추정하기 위하여, 수동으로 구축한 의미역 표지부착 말뭉치인 한국어 PropBank의 의미역인 ARG1에 해당하는 명사들을 동사의 주요 논항으로 보았다. 그리고 이들 주요 논항간의 의미 거리를 '코어넷 한국어 명사편'에서 계산하여 동사별로 이를 합산함으로써 이 계산한 값을 동사의 유사도로 추정하였다. 또한 본 연구에서 제안된 방식과 '코어넷 한국어 동사편'에서 동사간의 거리를 계산한 값 사이의 상관계수를 구하여 보았다.

주제어: 논항, 계층정보, 동사, 유사도

1. 서 론

개념간의 의미적인 유사도 및 관계도(Semantic Similarity/ Realatedness)를 구하는 연구는 고전적인 연구에서는 데이터베이스 통합이나 시스템 통합, 그리고 현대의 연구에 있어서는 태그 및 키워드 추출, 연관 단어 추천 등에 걸쳐 다양한 분야에 활용되어 온 연구이다[1].

국내에서는 의미망, 의미 표지 부착된 말뭉치 등 의미적 언어자원의 부족으로 인해, 영어 어휘망인 WordNet에서의 개념간 유사도 측정 방법을 활용하는 연구[2]가 진행될 뿐, 한국어 의미망을 바탕으로 한 개념간의 유사성 측정 방법이나 이를 활용하는 방법에 대한 연구가 미흡하다[3].

한국어의 경우 동사와 형용사는 문장에서의 역할이 명사와는 다르며, 동사의 의미는 동반하는 논항의 의미적, 통사적 특성에 따라 분화되므로 동사의 의미는 근본적으로 논항과 함께 고려되어야 한다[4].

개념간의 유사도 측정 방법은 크게 세 가지로, 링크 기반 방법[2)], 정보량 기반 방법[3)] 주석 기반 방법[4)][3]이 있다. 본 논문에서는 링크 기반 방법을 사용하여 동사의 주요 논항들의 유사도를 계산하고, 이 결과를 합산함으로써 동사간의 의미 유사도를 추정한다. 논항들의 성질을 반영하기 위해 수동으로 만든 의미역 표지 부착 말뭉치를 활용한다. 의미역 표지부착 말뭉치로는 가장 대표적인 말뭉치 중 하나인 Proposition Bank(이하 PropBank)[5]의 한국어 버전[6]을, 의미 체계로는 BOLA(언어자원은행)[5)]의 한국어 개념 기반 어휘의미망인 코어넷[6)]을 활용한다.

2장에서는 예비실험을, 3장에서는 동사의 의미 유사도 추정방법에 대하여, 4장에서 실험방법을 설명한 뒤, 마지막으로 5장에서는 결론과 향후 연구에 대하여 기술한 뒤 논문을 마무리한다.

1) 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2010-0010612)
2) 링크 기반 방법 : 의미망에서 두 개념의 최단 경로의 수, 노드의 깊이, 관계의 종류 등의 정보를 이용하는 방법
3) 정보량 기반 방법 : 대용량의 말뭉치에서의 개념의 발생빈도를 확률로 계산하는 방법
4) 주석 기반방법 : 관련 단어들의 공기정보를 활용한 방법
5) http://semanticweb.kaist.ac.kr/org/bora/index.html
6) http://semanticweb.kaist.ac.kr/org/bora/CoreNet_Project/index.html

2. 예비 연구

2.1 PropBank

‘밥을 집에서 먹었다’ 라는 문장에서, 술어 ‘먹다’에 대하여 어절 ‘밥’은 먹은 대상으로 ‘ARG1’ 의미역 표지가 부착되고, ‘집’은 먹은 행위를 한 위치이므로 ‘LOC’가 부착된다. ‘밥’은 먹은 대상이므로 논항에 해당하고, ‘집’은 먹은 장소이므로 역시 논항이다. [표 1]에 있는 의미역 모두 논항에 해당 될 수 있지만, ‘집’은 술어 ‘먹다’에 필수적인 요소가 아니다.

따라서 본 연구에서는 ‘ARG1’을 술어와 가장 밀접한 필수 논항으로 보고, ‘ARG1’에 해당하는 명사들을 ‘코어넷 한국어 명사편’에서 찾아 거리 계산한 값을 합산하여 동사간의 유사도를 추정하였다.

[표 1] 의미역 사례

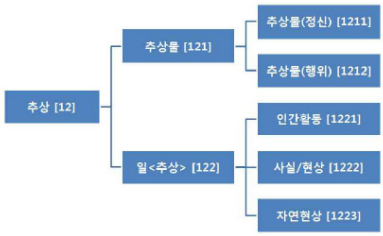
의미역	정의	의미역	정의
ARG1	피동작주	TMP	시간
LOC	장소	EXT	범위
DIR	방향	INS	도구
MNR	방법		

2.2 계층적 개념 체계

본 논문에서는 유사도를 추정하기 위한 의미 체계로 코어넷을 사용하였다. 코어넷은 총 2,938개의 계층적 개념과 92,448개 어휘의미가 연결되어 있다. 여기서는 코어넷중 CBL1(코어넷 한국어편)을 이용하였다.

[그림 1]은 CBL1의 계층구조로 각 숫자(클래스 번호)는 개념체계 내에서의 위치에 관한 정보, 즉 상위개념과 단계정보를 제공한다. [그림 1]에서 자연현상을 나타내는 1223은 3단계이며 상위개념으로 일<추상>을 나타내는 122가 있고, 더 상위엔 12는 추상을 나타낸다.

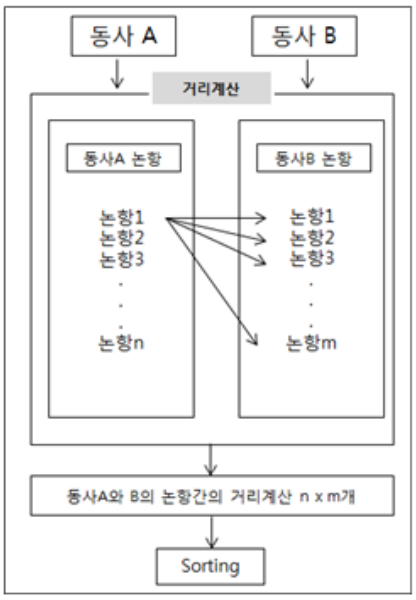
각 클래스간의 경로의 길이를 측정함으로써 유사도 추정을 하였는데, 인간활동과 자연현상의 유사도 추정을 위해 경로의 길이를 측정하면, 1221과 122사이의 에지, 122와 1223사이의 에지로 총 경로의 길이가 2라는 것을 알 수 있다[7].



[그림 1] 한국어 명사편 어휘의미망

3. 동사의 의미 유사도 추정

동사의 의미 유사도 추정은 약 9200개의 수동으로 만든 의미역 표지부착 말뭉치를 이용하였다. 말뭉치에서 술어역할을 하는 동사의 ‘ARG1’에 해당하는 명사들을 ‘코어넷 한국어 명사편’에서 클래스 번호들을 찾아, 그 거리를 계산하여 거리가 가까운 순서로 정렬한 뒤, 거리가 가까운 상위 30%의 평균을 동사의 의미 유사도라고 가정하였다. 상위 30%를 설정한 이유는 코어넷에 하나의 단어가 복수 개의 클래스에 나타나는 어의 중의성 문제가 발생한다. 이러한 문제 때문에 단어가 포함되는 모든 클래스간의 거리의 평균값을 계산하게 되면, 거의 모든 단어간의 거리가 차별화 되지 않는 경향이 있다. 따라서 본 연구에서는 근접한 상위 30%만을 계산대상으로 함으로써 단어간의 거리를 보다 차별화 하고자 하였다. [그림 2]는 동사간의 유사도 추정 과정을 그림으로 간단히 표현한 것이다.



[그림 2] 동사간의 유사도 추정 과정

예를 들어, [그림 3]은 “다듬다” 라는 동사의 ‘ARG1’에 해당하는 명사 ‘채소’, ‘문장’, ‘숙주’, ‘자신’을 ‘코어넷 한국어 명사편’에서 클래스 번호를 찾은 것이다. ‘채소’를 ‘코어넷 한국어 명사편’에서 찾으면 ‘작물’ 클래스에서 2번, ‘야채’ 클래스에서 2번 나온다. ‘문장’은 ‘윗사람’, ‘장막’, ‘마크’, ‘문장’, ‘책’에서 각각 1번씩 나오며, ‘자신’은 4가지 클래스에서 각각 한 번씩 나온다. 동사 ‘다듬다’는 총 16개의 클래스 번호를 가지게 되는데,

이 번호들을 다른 동사와의 유사도 추정을 하는데 모두 이용한다. [그림 4]을 보면 '가두다'에 해당하는 명사 클래스번호는 3개가 있다. '다듬다'와 '가두다' 두 동사의 의미 유사도를 추정하면 '다듬다'의 클래스번호 16개와, '가두다'의 클래스번호 3개를 1:1로 비교하여, 총 48개의 클래스간의 거리를 얻게 되는데, 이를 거리가 짧은 순으로 정렬하여 상위 30%에 해당하는 클래스간의 거리 14개의 평균을 '다듬다'와 '가두다'의 동사의 의미 유사도로 한다.

```
채소 작물 [11312121, 677]
채소 작물 [11312121, 677]
채소 야채 [11322512, 841]
채소 야채 [11322512, 841]
문장 윗사람 [111114221, 142]
문장 장막 [113226132, 885]
문장 마크 [12113253, 1102]
문장 문장(전체) [1211411, 1111]
문장 책(내용) [121145, 1119]
숙주 옷감 [11322412, 816]
숙주 야채 [11322512, 841]
자신 일인칭 단수 [11111111, 8]
자신 자신 [11111151, 32]
자신 확신 [12211722, 1399]
자신 신구/완급 [123925, 2710]
```

[그림 3] "다듬다"의 ARG1에 해당하는 명사 클래스 번호 모음

```
관계자 대리 [11113331, 342]
시민 주민 [11111531, 158]
시민 민중 [111115422, 168]
```

[그림 4] "가두다"의 ARG1에 해당하는 명사 클래스 번호 모음

4. 실험

3장에서 제안한 방식으로 약 230개의 동사를 이용하여, 29588개의 동사쌍을 만들었다. 같은 동사들을 '코어넷 한국어 동사편'에서 찾아 동사간의 거리를 계산하였다. [표 2]는 두 방법의 평균과 표준편차이다.

[표 2] 두 방법의 평균과 표준편차

	논항간의 거리를 계산하여 추정한 동사의 유사도	'코어넷 한국어 동사편'에서의 동사간 거리
평균	7.914762	6.865008
표준편차	1.039635	1.274787

실험은 임의의 295개의 동사쌍을 이용해 두 방법 사이의 상관계수를 구해보았다. [표 3]은 동사쌍을 두 가지 방법으로 유사도를 추정한 값의 일부이다. 두 방법의 상관계수를 구함으로써 논항의 유사도와 동사의 유사도와의 관계를 측정해 볼 수 있다. 본 실험에서 295개의 동사쌍의 상관계수는 '0.482578'로 비교적 강한 상관관계를 가진다.

[표 3] 두 방법의 유사도 비교 값 일부

	논항간의 거리를 계산하여 추정한 동사의 유사도	'코어넷 한국어 동사편'에서의 동사간 거리
걸다_걸치다	5.921146953	5.285714286
따다_뜯다	7	4.833333333
걸다_채우다	7.215189873	5.80952381
뿌리다_팔다	6.355172414	5.444444444
깎다_깎이다	4.888888889	1

5. 결론 및 향후 연구

본 논문에서는 동사간의 의미 유사도를 추정하기 위해 논항과 함께 유사도를 추정하는 방법을 제안하였다.

향후 연구에서는 여러 가지 조건을 사용하여 더 좋은 성능을 낼 수 있도록 연구할 것이며, 역으로 동사간의 거리를 계산하여 논항의 유사도를 구하는 실험을 통해 두 결과를 비교분석하여 유사도를 추정 할 계획이다.

참고문헌

[1] 최영석, 박진수, "의미간의 유사도 연구의 패러다임 변화의 필요성 - 인지 의미론적 관점에서의 고찰", 지능정보연구 제19권 1호, 111-123, 2013

[2] 조미영, 최준호, 김판구, "개념 기반 이미지 검색 시스템을 위한 WordNet 적용 방안", 정보과학회 가을학술발표논문집, 2002

[3] 임지희, 배영준, 최호섭, 옥철영, "U-WIN을 이용한 의미 유사도 측정과 활용", 한국컴퓨터종합학술대회 논문집 제 34권 제 1호(C),189-193, 2007.

[4] 이주호, 배희숙, 김은혜, 김혜경, 최기선, "명사 워드넷과 단일어 사전을 이용한 한국어 동사 워드넷 구축", 제 14회 한글 및 한국어 정보처리 학술대회, 92-97, 2002

[5] M. Palmer, D. Gildea, and Paul Kingsbury, "The Proposition Bank: An Annotated Corpus of Semantic Rules", Computational linguistics, 31(1), 71-106, 2005

[6] Young-Bum Kim, Heemoon Chae, Benjamin Snyder and Yu-Seop Kim*, "Training a Korean SRL System with Rich Morphological Features", The 52nd Annual Meeting of the Association for Computational Linguistics, 637 - 642, 2014

[7] 석미란, 윤영신, 김유섭, “개념 계층구조 상의 유사도를 이용한 이종 의미역의 자동 변환”, 한국정보과학회 2014 한국컴퓨터 종합 학술대회 논문집,1773-1775, 2014