레벨 기반의 유사도 계산을 이용한 PropBank 의미역과 Sejong 의미역 간의 자동 변환[1)]

윤영신○, 석미란, 김유섭
한림대학교 유비쿼터스 컴퓨팅학과
youngshin@hallym.ac.kr, smr4880@hanmail.net, yskim01@hallym.ac.kr

Automatic Transformation of Semantic Roles between PropBank and Sejong using Similarity Estimation based on Tree Level

Young-Shin Youn○, Mi-Ran Seok, Yu-Seop Kim
Dept. of Ubiquitous Computing, Hallym University

요 약

의미 표지 부착 작업은 구문 표지 부착된 문장의 술어-논항 구조를 파악하여 논항에 적절한 의미역을 부착하는 과정이다. 이 작업을 통하여 생성되는 의미 표지 부착 말뭉치는 의미역 결정에 있어서 절대적으로 필요한 자원이 된다. 의미 표지 부착 말뭉치로는 세계적으로 PropBank가 널리 활용되고 있는데 이를 한국어에 적용시키기 위해서는 PropBank 의미역과 Sejong 의미역 간의 자동 변환이 필요하다. 이전에 제안되었던 이종 의미역 간의 자동변환 방법에서는 명사 계층의 구조 정보를 반영하지 않았다는 문제점이 있었다. 본 논문에서는 이러한 문제점을 보강하기 위하여 명사 계층구조를 반영하여 한국어 PropBank 의미역을 Sejong 의미역으로 자동 변환하는 방법을 제안한다. 제안하는 방법은 PropBank와 Sejong의 맵핑 관계 중에서 1:N으로 맵핑되는 PropBank 의미역을 기준으로 명사 계층구조에서 변환 대상 의미역을 가지고 있는 단어와 변환 후보 의미역을 가진 단어들의 개념번호를 뽑아 두 단어 간의 거리를 측정한다. 그리고 레벨 당 가중치를 주어 유사도 계산을 하여 유사도가 적은 값으로 의미역을 자동 변환한다. 본 논문에서 제안하는 방법은 0.8의 성능을 보인다.

주제어: 이종 의미역, 자동 변환, 유사도 계산, 계층구조, 가중치

1. 서론

의미역 결정은 어휘 중의성 해소와, 자연어 처리의 의미 분석에 있어서 매우 중요한 요소로, 문장 내의 술어-결정을 위해서는 통계학에 기반한 방법론들이 사용되는데, 이를 위해서는 술어-논항 구조를 파악할 수 있는 의미역이 표지 부착된 말뭉치가 필요하다.[2]

의미 표지 부착 작업은 수동 구축작업으로 시간과 비용이 과다하게 필요하기 때문에 새로운 의미역 체계를 갖추는 의미 표지 부착 말뭉치를 구축하기 위해서는 하나의 표지 부착 말뭉치를 다른 표지 부착 말뭉치로 자동 변환할 필요가 있다.

의미역 결정에 있어서는 세계적으로 Propositional Bank(이하 PropBank)[3]가 많이 활용 되는데, 기존에 세종 의미역 체계로 구축된 말뭉치를 PropBank 의미역 체계로 자동 변환하게 된다면 큰 비용을 들이지 않고도 한국어 PropBank 말뭉치를 구축할 수 있을 것이고 이를 통해서 PropBank 기반의 다양한 의미역 결정 방법론을 쉽게 한국어에 적용시킬 수 있을 것이다.

이에 [4]에서는 논항에 표지 부착된 세종 의미역을 PropBank 의미역으로 자동 변환하는 방법론을 제시하였다. 하지만, [4]의 연구는 매우 특정한 의미역에 대해서만 변환을 시도함으로써 그 가능성만을 확인하려고 했고, Sejong 의미역에서 PropBank 의미역으로의 자동 변환만을 고려하였을 뿐, 그 반대의 경우에 대해서는 실제 변환 결과에 대한 결과를 제시하지 못하였다. 그리고 두 개념간의 유사도를 계산할 때, 계층구조의 특수성을 반영하지 않았다는 문제점이 있었다.

본 논문에서는 이러한 문제점을 보완하기 위해서 PropBank 의미역과 Sejong 의미역의 관계를 재분석하여 다양한 의미역에 대해서 대응 관계를 분석하였고, 그 결과를 토대로 논항에 부착된 PropBank 의미역을 Sejong 의미역으로 자동으로 변환시켜 보았다. 자동 변환을 위해서는 변환 대상이 되는 의미역을 가지고 있는 단어와 변환 후보 의미역을 가지고 있는 단어 간의 유사도를 계산하여 어느 의미역으로 자동 변환될지를 결정한다. 유사도 계산은 두 명사 어휘의 계층구조를 반영하여 각각의 계층구조의 경로의 길이를 레벨 당 가중치를 주어 유사도를 계산하였다.

2. 의미역 간의 관계 재분석

[표 1]는 현재까지 구축된 한국어 PropBank[5] 10,263개 문장 중에서 PropBank에서 사용되는 의미역이 Sejong

1) 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2010-0010612)

의미역으로 맵핑되는 결과를 재분석한 결과를 보여준다. [5]에서 구축된 의미 표지 부착 말뭉치는 PropBank 의미역 뿐만 아니라 Sejong 의미역도 함께 부착되어 있다. [표 1]의 첫 번째 열은 PropBank 의미역을, 두 번째 열은 Sejong 의미역을, 그리고 마지막 열은 맵핑되는 세종 의미역의 빈도수와 그 비중을 보여준다.

[표 1] PropBank 의미역과 Sejong의미역 관계 재분석

PropBank	Sejong	개수(%)
ARG0	AGT	6172(76%)
ARG1	THM	19912(90%)
ARG2	FNS	1119(26%)
	GOL	1067(24%)
	LOC	826(19%)
ARG3	GOL	544(60%)
ADV	EFF	551(69%)
	CNT	161(20%)
CAU	EFF	487(81%)
CND	EFF	36(32%)
	CNT	30(27%)
DIR	DIR	195(80%)
DIS	EFF	120(43%)
	CNT	82(29%)
EXT	CRT	330(52%)
	EFF	199(31%)
INS	INS	794(97%)
LOC	LOC	3274(86%)
MNR	INS	1470(75%)
PRD	INS	159(50%)
	EFF	95(30%)
M-PRP	PUR	146(91%)
M-TMP	LOC	1107(78%)

[표 1]은 현재까지 구축된 전체 ETRI 문항 수 10,000개 문장의 태깅 결과 각각의 PropBank 의미역에 맵핑되는 Sejong 의미역들의 빈도수와 그 비중을 상위 2-3개로 간추려 놓은 것이다. PropBank와 Sejong 의미역의 관계를 재분석하는 이유는 마지막으로 분석한 ETRI문항 수 3,759개의 데이터보다 더 많은 데이터로 자세한 분석이 필요했기 때문이다.

[표 1]의 결과, 동사의 주어를 나타내는 ARG0는 주격을 나타내는 AGT와 주로 맵핑된다. 목적어를 나타내는 ARG1은 목적격을 나타내는 THM과, 장소를 나타내는 LOC은 장소와 시공간을 나타내는 LOC와 1:1로 맵핑되는 것을 볼 수 있다. 이처럼 1:1로 맵핑되는 것이 아닌 1:N으로 맵핑되는 것은 중의성이 높기 때문에 주로 맵핑되는 여러 개의 의미역 중에서 하나를 선택하여 자동변환 해야 한다. [표 2]는 1:N으로 맵핑되는 PropBank 의미역과 Sejong 의미역을 [표 1]에서 따로 분리하여 표시한 것이다.

PropBank의 ADV 의미역은 부사적어구로 전치사 + 명사의 형태로 동사를 꾸며주는 것을 의미하는데, 이것은 Sejong 의미역의 영향주를 의미하는 EFF와, 내용, 발화/사유/인지 행위 등의 내용을 의미하는 CNT로 주로 맵핑되는 것을 볼 수 있다. 본 논문에서는 PropBank 의미역 ADV와 주로 맵핑되는 Sejong 의미역 EFF, CNT을 실험 데이터로 사용하였다.

[표 2] 1:N로 맵핑되는 PropBank의미역

PropBank	Sejong	개수(%)
ARG2	FNS	1119(26%)
	GOL	1067(24%)
	LOC	826(19%)
ADV	EFF	551(69%)
	CNT	161(20%)
CND	EFF	36(32%)
	CNT	30(27%)
DIS	EFF	120(43%)
	CNT	82(29%)
EXT	CRT	330(53%)
	EFF	199(31%)
PRD	INS	159(50%)
	EFF	95(30%)

3. 유사도 계산

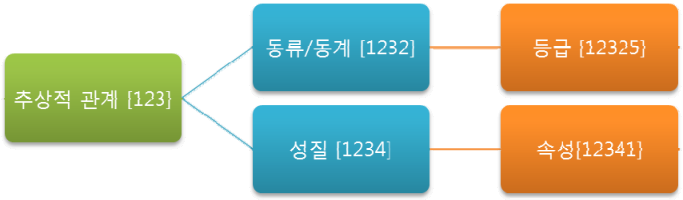
유사도 계산을 위해 BOLA(Bank of language resources)[2]의 한국어 개념 기반 어휘의미망인 코어넷[3] 중 CBL1(한국어 명사편)을 활용하여 의미별 명사 계층구조에 따라 명사를 분류한 후, 두 명사 클래스 간의 경로의 길이를 레벨 당 가중치를 두어 계산하였다.

[그림 1]은 CBL1의 계층구조로 각 숫자는 개념체계 내에서의 위치에 관한 정보, 즉 상위개념과 단계정보를 제공한다. 개념번호의 자릿수는 단계를 알려주며, 마지막 자릿수를 제거한 번호가 상위개념에 해당한다. [그림 1]에서 등급을 나타내는 12325의 상위 개념으로는 동류/동계를 나타내는 1232가, 더 상위인 123은 추상적 관계를 나타낸다.

본 논문에서는 트리 형식으로 된 명사 클래스간의 레벨 당 가중치를 두어 유사도를 계산했다. 계층구조를 반영하지 않고 유사도 계산을 했을 경우, 1234와 123사이의 에지, 123과 1232사이의 에지, 1232와 12325사이의 에지를 계산하여 4라는 값을 가질 수 있다. 클래스의 숫자가 1의 자리로 갈수록 더 작은 클래스를 나타내므로, 두 명사가 좀 더 유사하다는 것을 알 수 있다. 이를 이용하여 트리 구조로 된 명사 클래스의 레벨 당 유사도값을 더 적게 주는 방식을 사용하였다. 가장 큰 클래스가

2) http://semanticweb.kaist.ac.kr/org/bora/index.html
3) http://semanticweb.kaist.ac.kr/org/bora/CoreNet_Project/index.html

다를 경우 1을, 하위 클래스로 갈수록 유사도 값을 1/2로 줄여가며 계산을 하였다. 따라서 레벨3과 레벨4 각각의 가중치 0.125와 0.0625를 합쳐 0.188이란 유사도를 가진다.



[그림 1]한국어 명사편 어휘의미망

4. 의미 표지 부착 체계에서의 표지 자동변환 방법

① PropBank 의미역 ADV와 Sejong 의미역 EFF, CNT로 태깅된 결과 값에 포함된 단어들을 계층적 개념 체계에 적용하여 각 단어의 개념번호를 찾아낸다.

② ADV로 태깅된 의미역을 Sejong (CNT or EFF)로 자동변환하기 위해, 각각 Sejong의미역으로 태깅된 개념번호 중 랜덤으로 150개를 뽑아 5개의 Group을 생성한다.

③ Group1에 포함되어 있는 단어들을 testdata로 하여, Group2,3,4,5에 포함되어있는 단어들 간의 유사도를 계산한다. 마찬가지로 나머지 Group들도 각각 다른 Group들의 단어들과 유사도를 계산한다.

④ 각 Group들의 유사도 계산 결과, testdata와의 경로의 길이가 짧은 단어 5개를 뽑는다. 5개의 유사도 계산 결과 중 맵핑되는 Sejong 의미역의 수가 더 많은 쪽으로 Sejong의미역을 맵핑한다.

②에서 개념번호 중 랜덤으로 150개를 뽑는 이유는 PropBank 의미역에서 Sejong 의미역으로의 관계분석에서 빈도수의 차이가 많이 날 경우, 한쪽으로만 편향될 가능성이 있기 때문에, 똑같은 개수로 실험하기 위해서 이다.

5. 실험

본 논문에서는 PropBank 의미역은 ADV로, Sejong 의미역은 EFF와 CNT로 태깅 된 721개의 데이터에서 랜덤으로 추출한 명사 클래스 150개를 5개의 Group로 나눠 생성하여 사용하였다. 실험은 testdata가 들어있는 Group의 명사 클래스와 나머지 Group안의 각각의 클래스와 유사도 계산을 할 때, 명사 계층구조를 반영하여 트리로 된 계층구조의 레벨 당 가중치를 주어 계산한다.

testdata가 들어있는 Group을 G1, testdata를 제외한 나머지 Group들을 G2라고 하자. [표 3]에서 첫 번째 열은 G1의 명사 계층구조 클래스이고, 두 번째 열은 G2의 명사 계층 구조 클래스이며, 세 번째 열은 G1과 G2의 유사도를 계산하여 유사도가 높은 5개의 결과를 오름차순으로 나타낸 것이다. 그리고 네 번째 열은 G2의 계층 구조가 어느 세종 의미역으로 가는지 자동 태깅 결과를 출력한 것이다.

[표 3] 유사도 계산 결과 상위 5개 출력

G1개념번호	G2 개념번호	유사도	자동 태깅
111122	111121	0.031	EFF
111122	11111	0.094	CNT
111122	11113212	0.118	EFF
111122	11111211	0.118	EFF
111122	11113211	0.118	EFF

유사도 결과 상위 5개에서 G2의 계층구조가 어떤 세종 의미역으로 가는지 그 수를 세서 다섯개 결과 중 많은 쪽으로 Sejong의미역을 맵핑하고 수동 태깅 결과와 비교를 하여 성능평가를 하였다. [표 3]에서는 '이혼' 이라는 단어인 G1개념번호 111122에 대하여 EFF로 맵핑된 수가 더 많으므로 PropBank가 ADV인 경우, 그 의미역이 가진 명사의 개념번호가 111122일 때, EFF로 자동 태깅한다. [표 4]는 위와 같은 과정으로 자동으로 태깅된 결과와 수동으로 태깅한 결과를 비교한 것이다.

[표 4] 자동 태깅과 수동 태깅의 비교

개념번호	자동 태깅	수동 태깅
12325	CNT	CNT
111122	EFF	EFF
1221191453	EFF	EFF
12373	CNT	EFF
122118222	CNT	CNT
11226	EFF	CNT
121142	CNT	CNT
122129123	EFF	EFF
11224	EFF	EFF
123721	CNT	CNT

[표 3]에서 다루었던 '이혼' 이라는 단어를 제외한 다른 단어 10개에 대하여 유사도 계산을 하고 자동 태깅과 수동 태깅의 비교를 한 결과, 결과가 같은 경우가 8개이고, 결과가 다르게 나온 경우는 2개였다.

위와 같은 방식을 사용하여 실험 데이터 총 300개의 유사도 계산을 실행한 결과, 수동 태깅 결과와의 비교에서 238개 즉, 0.8의 성능평가를 보였다.

6. 결론

본 논문에서는 PropBank 의미역 표지 부착된 말뭉치를 자동으로 Sejong 의미역으로 변환하기 위해 계층구조를

반영한 유사도 계산을 이용하여 이를 자동 변환에 활용하는 방법론을 제안하였다. 실험 결과 5개 Group에서 임의로 뽑은 10개의 명사에 대해 이 방법론을 적용한 결과 8번의 올바른 결과가 도출되었다.

향후에는 PropBank의 ADV 의미역뿐만 아니라 1:N으로 맵핑된 ARG2, CND, DIS, EXT, PRD의미역에 대하여 Sejong 의미역으로 자동 변환 할 수 있는 방법론을 개발할 것이며, 반대로 Sejong의미역에서 PropBank의미역으로 자동 변환 할 방법론도 개발 할 것이다.

참고문헌

[1] 강신재, 박정혜, “한국어 전산처리에서 규칙과 확률을 이용한 구문관계에 따른 의미역 결정”, 한국산업정보 학회 논문지, 제8권 제1호, 2003

[2] 이민지, 이윤정, ‘한국어 의미 표지 부착 말뭉치’, 제 24회 한글 및 한국어 정보처리 학술대회 논문지, pp.99-103, 2012

[3] M. Palmer, D. Gildea, and Paul Kingsbury, “The Proposition Bank: An Annotated Corpus of Semantic Rules,” *Computational Linguistics,* 31(1), 71-106, 2005.

[4] 석미란, 윤영신, 김유섭, ‘개념 계층구조 상의 유사도를 이용한 이종 의미역의 자동변환’, 한국정보과학회 2014 한국 컴퓨터종합학술대회 논문집, pp.1773-1775, 2014

[5] 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