술어와 조사 정보를 이용한 논항의 의미역 변환[1)]

서민정○,석미란, 김유섭
한림대학교, 유비쿼터스 컴퓨팅학과
jos3194@naver.com, smr4880@hanmail.net, yskim01@hallym.ac.kr

Semantic Role Transformation of Arguments using Predicate and Josa Information

Min-Jeong Seo○, Mi-Ran Seok, Yu-Seop Kim
Dept. of Ubiquitous Computing, Hallym University

요 약

의미역 결정 (Semantic Role Labeling) 은 문장 내의 술어와 이들의 논항들의 의미 관계를 결정하는 과정을 뜻한다. 의미역 결정을 하기 위해서는 대량의 말뭉치와 다양한 언어 자원이 필요한데, 많은 경우에 PropBank 말뭉치가 사용된다. 한국어 PropBank는 다른 언어에 비해 자료가 적어 그것만을 가지고 의미역 결정을 하기에 적절하지 않다. 또한 한국어 의미 분석을 위해서 지금까지는 세종 말뭉치나 의미역이 활용되어 오기도 하였다. 따라서 한국어 의미역 결정에서는 한국어 PropBank 뿐만 아닌 세종 의미역 표지 부착 말뭉치의 구축 역시 요구되는데 말뭉치 구축 작업이 수동 부착 작업이기 때문에 많은 시간과 비용이 소모된다. 본 논문에서는 이러한 문제점을 해결하기 위해 이미 구축되어 있는 한국어 PropBank 의미역을 세종 의미역으로 자동 변환하는 방법을 제시한다. 자동 변환을 위해서는 먼저 PropBank 의미역의 변환 후보 의미역을 구하여 이들 중에서 가장 적절한 의미역으로 변환한다. 자동 변환을 위해서는 크게 3 가지 특징을 활용하는데, 첫째는 변환 대상 논항의 의미 유사성이고, 둘째는 논항과 의미 관계를 가지고 있는 술어, 그리고 셋째는 논항과 결합되어 있는 조사이다. 이 세 가지 특징을 사용하여 정확한 의미역 변환을 위해 술어, 조사의 의미역 결합 확률 테이블을 구축한다.

주제어: 의미역 변환, 확률 테이블, 논항

1. 서론

의미역 결정(Semantic Role Labeling)이란 문장 내의 술어-논항들의 의미 관계를 결정하는 과정을 뜻한다. 의미역 결정을 하기 위해서는 의미역이 부착되어 있는 대량의 말뭉치가 필요하다. 현재 가장 널리 사용하는 말뭉치는 Proposition Bank(이하 PropBank)[1]이다.

PropBank는 술어-논항 구조를 태그해 놓은 말뭉치를 말한다. 이 말뭉치[2]는 영어와 중국어로는 구축이 되어 있지만, 한국어 PropBank는 기존 영어와 중국어의 말뭉치에 비해 학습자료가 적어 정확도가 적고, 한국어의 특수성이 반영되지 않아 의미역 결정에 그대로 활용하기에는 문제가 있다.[3] 이러한 문제점을 해소하기 위해 한국어에 특화된 세종 계획에서 정의한 의미 표지 부착 말뭉치(이하 세종)를 사용한다. 기존 한국어 의미역 결정 관련 연구들이 세종의 의미 체계를 주로 활용하였기 때문에 세종 의미 표지 부착 말뭉치의 구축 역시 매우 중요하다.

그러나 PropBank 의미역과 세종 의미역이 부착되어 있는 말뭉치를 구축하는 작업은 수동으로 진행되기 때문에 말뭉치 구축에는 많은 시간과 비용이 소모된다. 때문에 새로운 의미 표지 부착 말뭉치를 이미 구축되어 있는 의미 표지 부착 말뭉치를 자동으로 변환하여 구축할 수 있는 방법이 필요하다.

이를 위하여 본 논문에서는 논항에 표지 부착된 한국어 PropBank 의미역을 세종 의미역으로 자동 변환하는 방법을 제시한다. [4]에서는 자동 변환을 위하여 의미역 변환 대상이 되는 논항의 의미 유사도를 계산하였다. 이미 구축된 데이터에서 변환 대상 논항과 의미 상으로 유사한 단어들의 의미역을 찾아 이로 변환하고자 하였다.

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

그러나 이러한 경우에는 논항 그 자체만 고려할 뿐 그 논항의 문맥 정보는 고려하지 못하였다. 따라서 본 연구에서는 논항의 의미 유사도 뿐만 아닌 논항과 의미상으로 밀접한 관계가 있을 것으로 추정되는 논항의 조사 및 술어도 고려한 자동 변환 방법을 제시한다.

2장에서는 Propbank와 세종의 의미역들의 상관관계에 대해 살펴보고, 3장에서는 Propbank에서 세종으로의 의미역을 자동으로 변환하는 방법을 설명한다. 4장에서는 의미역 변환 방법을 이용한 말뭉치 부착 결과와 기존 사용자가 직접 부착한 말뭉치의 유사도를 실험을 통해 확인해 본다. 마지막으로 5장에서는 결론과 향후 연구에 대해 다룬다.

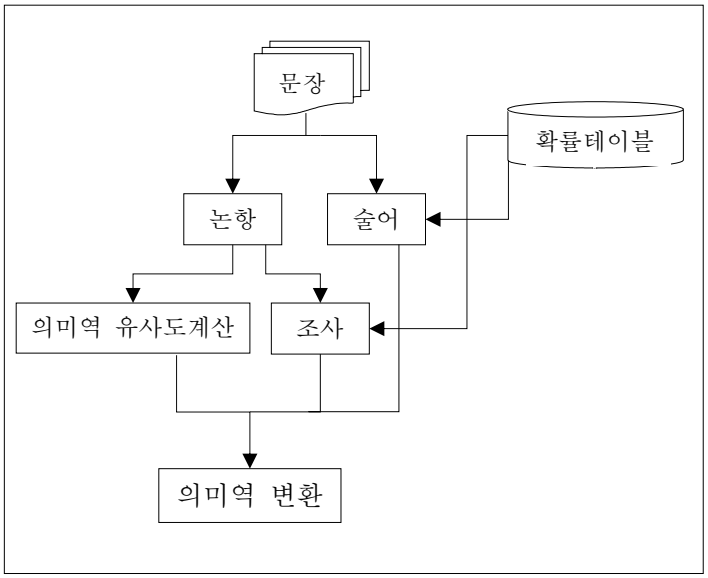
2. Propbank와 세종의 의미역들의 상관관계

[표 1] PropBank 의미역과 세종 후보 의미역

PropBank 의미역	세종 의미역
ARG0(주체자)	AGT(행위주)
ARG1(수동자)	THM(대상)
ARG2(시작점)	FNS(결과상태)
	GOL(도착점)
	LOC(장소)
ARG3(끝점)	GOL(도착점)
M-ADV(부사적 어구)	EFF(영향주)
	CNT(내용)
M-CAU(원인)	EFF(영향주)
M-CND(조건)	EFF(영향주)
	CNT(내용)
M-DIR(방향)	DIR(방향)
M-DIS(문장의 연결)	EFF(영향주)
	CNT(내용)
M-EXT(크기)	CRT(기준치)
	EFF(영향주)
M-INS(도구)	INS(도구)
M-LOC(장소)	LOC(장소)
M-MNR(방법)	INS(도구)
M-PRD(술어의 자격)	INS(도구)
	EFF(영향주)
M-PRP(목적)	PUR(목적)
M-TMP(시간)	LOC(장소)

의미역 결정을 할 때, 장소를 뜻하는 LOC와 같이 PropBank 의미역과 세종 의미역이 1:1 비슷한 확률로 맵핑되는 경우가 있다. 그러나 ARG2(시작점)과 같이 PropBank 의미역과 세종 의미역 사이에는 중의성이 존재하는 경우도 있다. 이러한 PropBank의 의미역은 중의성을 가진 2개 이상의 세종 의미역을 가지게 된다. 따라서 PropBank 의미역과 유사한 변환 후보 의미역을 구하여 유사도를 계산하는 방법을 이용한다. [표 1]는 PropBank 의미역과 유사한 세종 변환 후보 의미역을 나타낸 것이다. ARG0, ARG1, M-CAU와 같이 세종 후보 의미역이 1:1인 경우에는 바로 의미역을 결정하고, ARG2, M-ADV, M-CND와 같이 후보 의미역이 여러 개인 경우, 변환 대상이 되는 PropBank 의미역을 가진 단어와 세종 후보 의미역을 가진 단어들 간의 유사도를 계산하여 의미역을 최종 결정한다.

3. PropBank에서 세종으로의 의미역 자동 변환 방법



[그림 1] 의미역 변환 과정

[그림 1]은 PropBank에서 세종으로 의미역을 자동 변환하는 과정이다. 문장을 논항과 술어의 쌍으로 나눈다. [4]에서 의미역 변환을 할 때, 변환 대상 의미역을 가진 논항과 변환 후보 의미역을 가진 논항끼리의 유사도만 계산하여 의미역을 변환하였다. 그러나 이 방법은 논항 자체의 유사도만 비교하여 의미역 변환의 정확도가 떨어지고, 문맥상의 내용을 고려하지 못했다는 문제점이 있

다. 논항이 주격 또는 목적격으로 쓰인 경우, PropBank 의미역과 세종 의미역이 1:1 맵핑이 되어 의미역 변환이 쉽지만, 부사격 조사 같은 경우에는 조사나 술어에 따라 후보 의미역이 다르게 존재하게 된다. 따라서 변환 대상 의미역과 후보 의미역 사이의 유사도를 계산한 후, 논항의 술어와 조사의 의미역 결합 테이블을 이용하여 정확한 의미역 변환을 한다.

3.1 의미역 유사도 계산

의미역 간의 유사도를 계산하기 위해, 변환 대상이 될 의미역을 가진 단어와 후보 의미역을 가진 단어 간의 유사도를 계산하여 어느 의미역으로 변환될 지를 결정한다. 이를 위해 BOLA(Bank of Laguage resources)[2]의 한국 개념기반 어휘의미망 코어넷[3]을 사용하였다. 코어넷은 총 2,983개의 계층적 개념과 92,448개의 어휘의미가 연결되어있다. 본 논문에서는 CBL1(한국어 명사편)을 이용했다. CBL1을 이용해 조사를 제외한 논항을 의미별 명사 계층구조에 따라 명사를 분류하였다. CBL1은 계층구조로, 개념체계 내에서 위치하는 정보와 상위개념, 단계 정보를 개념번호(숫자)로 표현한다.

변환대상이 될 PropBank 의미역을 P, 변환 후보 세종 의미역 2개를 각각 S1, S2라고 하자. P와 S1의 그룹을 SET1, P와 S2의 그룹을 SET2라고 하자. 또한 P만 가진 것을 p라고 하자.

① p, SET1, SET2로 단어들을 의미별 명사 계층구조에서 찾고, SET1의 단어들을 G1, SET2의 단어들을 G2로 묶는다.

② G1, G2에 속한 단어들의 개념번호(클래스)를 찾고, p의 개념번호도 찾는다.

③ p의 개념번호와 G1의 개념번호를 비교하고, G2 또한 p와 유사도를 비교한다.

④ G1과 G2에서 p와의 거리가 짧은 단어 5개를 뽑고, 5개의 평균값이 작은 쪽으로 PropBank 의미역을 세종 의미역으로 변환한다.

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

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

3.2 술어, 조사 정보를 이용한 결합 확률 테이블

일반 문장은 논항과 술어로 구성된다. 논항이 주격 또는 목적격으로 쓰이는 경우, 주어, 목적어를 나타내는 PropBank의 ARG0, ARG1, 세종의 AGT, THM으로 대부분 1:1로 맵핑된다. 그러나 부사격의 경우, 논항의 조사와 술어에 따라 의미역이 완전히 다르게 부착된다.

학교[용언불가능보통명사]+로[부사격조사]	ARG3(GOL)
학교[용언불가능보통명사]+에[부사격조사]	LOC(LOC)
학교[용언불가능보통명사]+에서[부사격조사]	ARG2(SRC)

[그림 2] 조사와 논항에 따른 의미역 말뭉치

[그림 2]는 논항의 조사와 논항과 같이 쓰이는 술어에 따라 의미역은 완전히 다르게 변환되는 것을 보여준다. 명사 “학교”와 술어 “가다”에 조사 “로”가 붙은 경우, PropBank와 세종에서 도착점을 나타내는 ARG3, GOL로 맵핑된다. “에”가 붙은 경우, 장소를 나타내는 LOC, LOC가 맵핑된다. “에서”가 붙은 경우에는 출발점을 나타내는 ARG2, SRC로 맵핑된다. 이처럼 의미역 결정을 하는데 조사와 술어에 따라 의미역이 완전히 다르게 결정되며, 정확성이 떨어진다. 따라서 본 논문에서는 정확한 의미역 결정을 위해 각 조사와 술어에 대한 정보를 가진 결합 확률 테이블을 구축한다.

테이블의 데이터를 작성하기 위해 본 논문에서는 한국전자통신연구원의 구문 표지 부착 말뭉치를 사용하여, 약 12,000개 문장 데이터의 술어-논항에 대해 의미역을 부착하였다. 의미역을 부착한 후, 확률 테이블은 동사는 4,414개의 테이블로 구축되고, 조사는 241개의 테이블로 구축되었다. 각 테이블은 세종 의미역들이 12,000개의 데이터에서 쓰인 정보를 저장하고 있다.

[표 2] 조사 “에”와 “같이”의 확률관계

조사	세종 의미역	확률
에	LOC	0.43
	GOL	0.26
같이	EFF	0.56
	FNS	0.06
	COM	0.17

[표 2]는 조사 “에”와 “같이”의 정보를 가진 확률 테이블의 부분이다. “에”는 세종 의미역에서 장소를 나타내는 LOC, 도착점을 나타내는 GOL으로 사용되었다. 따라서 조사 “에”는 장소로 많이 쓰였고, 장소에 따라 도착점인 경우에 많이 쓰였다는 것을 알 수 있다. 조사 “같이”는 영향주를 나타내는 EFF, 결과상태를 나타내는 FNS, 동반주를 나타내는 COM으로 쓰였는데, 영향주의 의미로 주로 쓰였다는 것을 알 수 있다.

[표 3] 동사 “근거하다”와 “보다”의 확률관계

술어	세종 의미역	확률
근거하다	THM	0.33
	LOC	0.67
보다	AGT	0.12
	THM	0.51
	LOC	0.14

[표 3]은 동사 “근거하다”와 “보다”의 확률 테이블에서 상위 의미역을 나타낸 것이다. “근거하다”는 세종 의미역에서 장소를 나타내는 LOC를 많이 사용했다. “보다”의 경우, 행위주를 나타내는 AGT, 대상을 나타내는 THM, 장소를 나타내는 LOC로 맵핑되었고, 이 중 대상이라는 의미로 많이 사용했다는 것을 알 수 있다. 따라서 확률테이블은 논항 외, 조사, 술어에 따라 의미역이 다르게 쓰일 수 있고, 문맥 상 논항의 의미역 변환에 필요하다.

4. 실험

```
1     18    때문에[문장접속부사]
2     3          사실주의[용언불가능보통명사]+라는[관형격조사]
3     4          말[용언가능보통명사]+이[주격조사]
4     18    적용되[일반동사]+려면[종속연결어미]
5     7               19_세기[용언불가능보통명사]+의[관형격조사]
6     7               고전적[성상관형사]
7     15              사실주의[용언불가능보통명사]+에서와_같_이[부사격조사]
8     9                    오늘[용언불가능보통명사]+의[관형격조사]
9     10              우리[인칭대명사]+의[관형격조사]
10    11              견지[용언가능보통명사]+에서[부사격조사]
11    15              보[일반동사]+아서도[종속연결어미]
12    14              현실[용언불가능보통명사]+에[부사격조사]
13    14              직접[성상상태부사]
14    15              근거하[일반동사]+었[과거시제선어말어미]+다고[종속연결어미]
15    16         보[일반동사]+ㄹ_수_있[기타보조용언]+는[관형사형전성어미] (동격)
16    17         경우[용언불가능보통명사]+에[부사격조사]
17    18    한하[일반동사]+여서[종속연결어미]
18    0     적용되[일반동사]+어야_하[기타보조용언]+ㄴ다[평서형종결어미]+.[문미기호]
```

[그림 3] 예제의 술어-논항 관계

본 논문에서는 술어-논항과의 관계를 파악하기 위해 한국전자통신 연구원의 구문 표지 부착 말뭉치(etri)를 문장 데이터로 이용한다. 실험을 위해 예제 “때문에 사실주의라는 말이 적용되려면 19세기의 고전적 사실주의에서와 같이 오늘의 우리의 견지에서 보아서도 현실에 직접 근거했다고 볼 수 있는 경우에 한해서 적용되어야 한다.”를 사용한다. 의미역 변환은 3장에서 제시한 방법을 적용하여 실험을 진행하였다.

[그림 3]에서 논항과 술어의 의존관계는 두 개의 인덱스로 표현되는데, 각 줄마다 첫 번째 인덱스는 어절의 순번이고, 두 번째 인덱스는 의존하고 있는 어절의 순번을 가리킨다. 각각의 줄에서 두 번째 인덱스와 첫 번째 인덱스가 같으면 의존관계에 있다고 표현한다. 예를 들면, 논항 “때문에”, “적용되려면”, “한하여”는 술어 “적용되어야 한다.”와 관계가 있다. 이런 논항의 의미역을 변환하기 위해, 논항과 술어의 의존관계를 고려하여 논항과 술어로 분리하면 [표 4]와 같은 결과가 나온다. 논항은 명사 + 조사로 나뉘고, 각 논항에는 PropBank 의미역을 부착하였다.

[표 4] 예문의 논항, 술어 분리 결과

논항	의미역	세종 후보 의미역
말+이	ARG1	THM
견지+에서	LOC	LOC
현실+에	ARG2	FNS, GOL, LOC
사실주의+에서와 같이	ADV	EFF, CNT
경우+에	ARG2	FNS, GOL, LOC

[표 4]와 같이 “말+이”, “견지+에서”는 PropBank 의미역과 세종 의미역이 1:1 맵핑되므로 ARG1(THM), LOC(LOC)로 의미역을 변환한다. 그러나 “현실+에”, “사실주의+에서와 같이”, “경우+에”는 PropBank 의미역과 세종 의미역이 1:N 맵핑되므로 의미역 유사도 계산과 확률 테이블을 이용한 의미역 변환이 필요하다. 각 논항에 해당하는 PropBank 의미역과 세종 의미역 사이의 유사도를 구하고 술어와 조사 결합 확률 테이블을 이용하여 합산하면 [표 5]와 같은 결과가 나온다. [표 5]를 보면, “현실+에”가 ARG2(LOC), “사실주의+에서와 같이”는 ADV(EFF), “경우+에”는 ARG2(LOC)로 가장 높은 수치를 보여 각 세종 의미역으로 PropBank 의미역을 변환한다.

[표 5] 술어와 조사 확률 테이블을 이용한 의미역 계산

논항	PropBank	세종	술어	합계
현실+에	ARG2	GOL	근거하다	0.58
현실+에	ARG2	FNS	근거하다	0.30
현실+에	ARG2	LOC	근거하다	0.71
사실주의+에서와 같이	ADV	EFF	보다	0.53
사실주의+에서와 같이	ADV	CNT	보다	0.50
경우+에	ARG2	GOL	한하다	0.43
경우+에	ARG2	FNS	한하다	0.10
경우+에	ARG2	LOC	한하다	0.84

[표 6] 수동부착 말뭉치와 자동변환 말뭉치의 유사도

논항	수동부착	자동변환	일치
말+이	ARG1(THM)	ARG1(THM)	○
견지+에서	LOC(LOC)	LOC(LOC)	○
현실+에	ARG2(LOC)	ARG2(LOC)	○
사실주의+에서와 같이	ADV(EFF)	ADV(EFF)	○
경우+에	ARG2(CRT)	ARG2(LOC)	X

[표 6]은 기존 예문에 사용자가 직접 부착하여 구축한 말뭉치와 자동 변환 방법을 적용한 말뭉치의 유사도 결과를 나타낸 것이다. 자동 변환 방법을 적용한 말뭉치는 수동부착과 유사한 경우가 4개, 다른 경우가 1개로 나타내는 성능을 보였다.

5. 결론

본 논문에서는 기존 수동 부착 작업의 문제점을 보안하기 위해 기존 PropBank 의미역을 세종 의미역으로 변환하는 방법을 제시한다. 또한 정확한 의미역 결정을 위해 조사와 술어의 결합 확률 테이블을 구축하였다. PropBank 의미역을 세종 의미역으로 바꾸기 위해, 변환 대상 의미역과 후보 의미역 사이의 유사도를 계산하고, 조사와 술어의 테이블을 이용하여 의미역을 최종 결정한다. 자동 변환 방법을 적용한 말뭉치와 기존 수동 부착한 말뭉치를 비교한 결과, 5개의 의미역 중 4개의 올바른 결과가 도출되었다.

향후에는 PropBank 의미역과 세종 의미역을 문장의 논항에 자동 부착하는 방법론을 개발할 것이다. 또한 다양한 명사와 조사, 동사 외에 다양한 문장의 형태소들을 적용하여 보다 정확한 의미역 자동 부착 시스템을 구축할 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] Nianwen Xue, “Annotation Guidelines for the Chinese Proposition Bank”, February 19, 2007

[3] 조정현, 정현기, 김유섭, “한국어 의미 표지 부착 말뭉치 구축을 위한 자동 술어-논항 분석기 개발”, *정보처리학회논문지* B, 제19-B권 제1호, 2012

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