한국어 격틀 사전과 용언의 하위 범주 정보를 사용한 한국어 의미역 결정[1)]

김완수○, 옥철영
울산대학교, 한국어처리연구실
kimwansu@outlook.com, okcy@ulsan.ac.kr

Korean Semantic Role Labeling using Case Frame and Subcategory of Predicate

Kim Wansu○, Ock CheolYoung
Korean Language Processing Lab. University of Ulsan, Korea

요 약

의미역 결정은 문장의 서술어와 그 서술어에 속하는 논항들 사이의 의미관계를 결정하는 문제이다. 본 논문에서는 UPropBank 격틀 사전과 UWordMap의 용언의 하위 범주 정보를 이용하여 의미역을 부착하였다. 실험 결과 80.125%의 정확률로 의미역을 부착하는 성능을 보였다.

주제어: 의미역(Semantic Role), 격틀사전, PropBank, 하위 범주, UWordMap

1. 서론

의미역 결정은 문장의 서술어와 그 서술에 속하는 논항들 사이의 의미관계를 결정하는 문제로, 문장의 서술어와 논항들 사이의 '주어', '목적어'와 같은 문법 관계를 '행동주', '경험주', '대상'등과 같은 의미 관계로 사상(mapping)하는 문제로 볼 수 있으며, 일반적으로 구문 분석을 수행한 후에 의미역 결정이 수행된다. 의미역 결정은 기계 번역, 정보 추출, 질의응답과 같은 다양한 자연 언어 처리 응용에서 성능 향상을 위해 사용될 수 있다[1,2].

이전 논문에서 사용한 방법은 격틀 사전과 용언에 올 수 있는 격조사별 의미역의 통계를 사용하여 의미역을 부착하는 방법이다[3]. 이 방법은 용언의 격틀 전체를 고려하지 않고 한 어절의 격조사와 용언만을 가지고 의미역을 결정하기 때문에 다양한 격틀을 가진 용언에 대해서 의미역을 부착할 때 정확한 의미역을 부착하기 어려운 문제가 있다.

본 논문에서는 국립국어원 표준국어대사전을 기반으로 구축한 UPropBank 격틀 사전[4]과 용언이 가질 수 있는 하위 범주 정보[5]를 사용하여 용언에 대해 한국어 의미역 부착을 시도하였다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 UWordMap의 하위 범주 정보에 대해 기술하고, 4장에서는 한국어 격틀 사전과 용언의 하위 범주 정보를 사용하여 의미역을 결정하는 방법에 대해 기술하고, 마지막으로 5장에서 결론과 향후 연구에 대해 기술한다.

2. 관련 연구

의미역을 결정하는 방법은 격틀 사전에 기반한 방법과 말뭉치에 기반한 방법으로 나눌 수 있다. 격틀 사전을 이용하는 방법은 격틀 사전에 기술된 격틀에 따라 서술어-논항 관계와 격틀 사이의 유사도 계산 등을 통해서 의미역이 결정되기 때문에 높은 정확률을 보이지만, 격틀 사전 구축이 어렵고 격틀에 기술되지 않은 문장 형태에는 적용하지 못하는 문제가 있다.

말뭉치를 이용하는 방법은 말뭉치에 의미역을 태깅하여 의미역 결정 학습 데이터를 생성한 후, 기계 학습을 적용하여 의미역을 결정하는 방법이다. 말뭉치를 이용하는 방법은 격틀 사전에 기반한 방법보다 적용률이 높다는 장점이 있지만, 의미역 부착 말뭉치의 구축이 어렵다는 단점이 있다. 영어권에서는 PropBank 등의 의미역 부착 말뭉치를 이용하는 방법이 많이 연구되어 있지만, 한국어의 경우에는 의미역이 태깅된 말뭉치를 구축하기 어려워서 격틀 사전을 이용하는 방법이 주로

1) 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (R0101-15-0176)

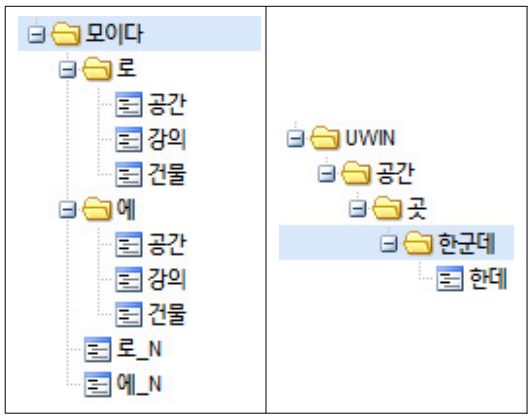
연구되어 왔고[1], 최근에는 기계 학습 기법을 이용한 한국어 의미 부착 방법[2]이 연구되고 있다.

이전 논문에서 사용한 방법은 동형이의어, 다의어, 의존관계 분석된 문장의 각 어절에서 의존하는 용언에 대해 격틀 사전을 사용하여 의미역 후보를 좁히고 나서 후보가 1개인 경우 바로 결정하고, 2개 이상인 경우에는 기존에 구축한 격조사와 용언별 의미역 통계를 사용하여 의미역을 결정한다[3].

본 논문에서는 국립국어원 표준국어대사전을 기반으로 구축한 격틀 사전과 용언의 하위 범주 정보를 사용하여 서술어의 필수 논항의 의미역을 결정하는 방법을 제안한다.

3. UWordMap 하위 범주 정보

UWordMap[6]은 표준국어대사전에 있는 용언의 하위 범주 정보, 단어의 상하위어 관계 정보가 구축된 데이터베이스이다. 용언의 하위 범주 정보에 있는 단어는 용언의 필수 논항에 함께 사용할 수 있는 모든 단어 중 가장 공통된 단어('로'의 하위에 있는 단어)와 그 중 예외인 단어('로_N'의 하위에 있는 단어)의 정보가 있다[그림 1]. 단어의 상하위어 관계 정보는 단어의 뜻풀이, 용례 등의 어휘 정보를 활용하여 계층적으로 분류한 것이다[그림 2].



[그림 1] 용언(모이다 010201)의 하위범주 정보(왼쪽)
[그림 2] 단어(한군데)의 상하위어 관계 정보(오른쪽)

본 논문에서는 격틀 사전을 사용하여 의미역을 정확하게 결정할 수 없는 필수 논항에 대해 용언의 하위 범주 정보를 사용하여 각 격조사별로 용언과 함께 쓰일 수 있는 단어를 확인, 상하위어 관계 정보를 사용하여 문장에서 주어진 단어가 용언과 함께 쓰일 수 있는 단어에 속하는 것인지를 확인하는 방법으로 의미역을 결정한다.

4. 격틀 사전과 하위 범주를 사용한 의미역 부착

UTagger-SR은 의미역 부착 말뭉치를 구축하는 도구이다[5]. 입력한 문장을 UTagger-DP[7]를 사용하여 형태의미 분석 및 의존관계 분석 후, 문장에 있는 서술어를 추출하면 의미역 부착 작업자가 각 어절-서술어에 UPropBank 격틀 사전을 사용하여 의미역 후보를 한정하여 한 가지로 확정된 의미역은 미리 부착하고, 확정되지 않은 후보 중에서 알맞은 의미역을 선택하는 방법으로 의미역 부착 말뭉치를 구축하는 프로그램이다.

본 논문에서는 UTagger-SR에 용언의 하위 범주를 이용하여 적합한 의미역을 선택하는 기능을 추가하여 의미역 자동 부착을 시도하였다. 문장에서 한 어절에 의미역을 부착하는 과정은 아래와 같다.

문장: 처음 듣는 얘기에 모두들 신기해서 눈이 한군데로 모여 왔다.

의미역부착 | 형태소.의존관계

순서	의존	어절	2 듣_01	5 신기하_04	8 모이_01
1	2	처음/NNG			
2	3	듣__01/VV+는/ETM			
3	5	얘기/NNG+에/JKB			
4	5	모두__01/NNG+들__09/XSN			
5	8	신기하__04/VA+어서/EC			
6	8	눈__01/NNG+이/JKS			THM
7	8	한군데/NNG+로/JKB			DIR
8	9	모이__01/VV+어/EC			
9	9	오__01/VX+았/EP+다/EF+./SF			

[그림 3] 의미역 자동 부착 예문

4.1 의존 관계를 사용하여 의미역을 부착할 용언 선택

UTagger-DP로 분석된 의존관계를 따라서 현재 어절의 지배소인 용언을 찾는다. 의존하는 어절에 용언이 없는 경우 용언이 발견될 때까지 의존하는 어절 번호를 계속 따라가서 찾는다.

- 예문에서 6번 어절인 '눈이'는 8번 어절을 의존하고, 8번 어절에 있는 '모여'는 용언인 '모이다'가 있으므로 '모이다' 용언을 선택한다.
- 7번 어절도 마찬가지로 8번 어절을 의존하고, 8번 어절에 용언이 있으므로 '모이다' 용언을 선택한다.

모이다 010101	[동사]	{X:대상}	모으다[1]㈀'의 피동사. >>한데 합치다.	재료가 다 모이면 한데 섞어라. 이 산은 조그마한 돌들이 모여서 된 돌산이다. 잡다한 일들이 너무 많이 모여서 이제는 처리하기 힘든 실정이다.
모이다 010102	[동사]	{X:대상}	모으다[1]㈁'의 피동사. >>특별한 물건을 구하여 갖추어 가지다.	그는 우표가 하나 둘씩 모일 때마다 그렇게 기뻐할 수 없었다. 어느 정도 귀중한 골동품들이 모이면 전시회를 열 작정이다.
모이다 010103	[동사]	{X:대상}	모으다[1]㈂'의 피동사. >>돈이나 재물을 써 버리지 않고 쌓아 두다.	돈이 좀 모였어? 이웃 돕기 성금이 많이 모였을까? 돈이 좀 모이면 들어와서 살겠어요.
모이다 010104	[동사]	{X:대상}	모으다[1]㈄'의 피동사. >>정신, 의견 따위를 한곳에 집중하다.	젊은이들의 창의적인 의견이 어느 정도 모이면 그중에서 가장 실용적인 것을 골라 사업화할 계획입니다.
모이다 010105	[동사]	{X:대상}	모으다[1]㈅'의 피동사. >>힘, 노력 따위를 한곳에 집중하다.	여러분의 작은 정성이 모여 큰 힘이 됩니다. 우리들의 힘이 지금보다 더 크게 모일 때까지는 자중합시다.
모이다 010201	[동사]	{X:대상 Z:처소-에} {X:대상 Z:방향-으로/로}	모으다[2]'의 피동사. >>여러 사람을 한곳에 오게 하거나 한 단체에 들게 하다.	오랜만에 온 식구가 한 장소에 모였다. 모인 김에 어디 놀러나 갈까? 우리는 내일 약속 장소로 모이기로 하고 헤어졌다. 그 많은 사람들이 한 시간 안에 한곳으로 모이기는 쉽지 않을 것이다.
모이다 010202	[동사]	{X:대상 Z:착점-에} {X:대상 Z:착점-으로/로}	모으다[1]㈆'의 피동사. >>다른 이들의 관심이나 흥미를 끌다.	팬들의 관심은 온통 그 선수의 은퇴 여부에 모여 있었다. 사람들의 시선은 새로 출시된 자동차로 모여 있었다.

[그림 4] UPropBank 격틀 사전

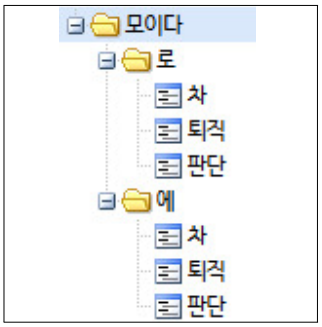
4.2 격틀 사전을 사용한 의미역 선택

4.1에서 의미역을 부착할 용언이 결정되었으면 해당 용언에 의존하는 어절의 격조사를 확인하여 격틀 사전에서 일치하는 격틀을 찾는다. 이 때 문장에 주격이 있는지 확인하여 주격이 없는 경우에는 격틀에서 주격 부분을 제외하고 비교한다. 찾은 격틀에서 의미역을 확인하여 한 어절에 대해 의미역 후보가 1개로 좁혀지면 의미역을 부착한다.

- 예문에서 '모이다' 용언을 의존하는 어절 중 필수 논항은 6번 어절에 있는 '이'와 7번 어절에 있는 '로'이다.
- [그림 4]의 격틀 사전에서 '모이다' 용언의 문형에 따르면 예문의 문형과 일치하는 것은 '모이다 010201'과 '모이다 010202'이다.
- 이 중에서 주격인 '눈이'에 대한 의미역은 양쪽 모두 '대상'으로 같으므로, 6번 어절 '눈이' 대한 의미역은 '대상'이다.
- 반면 '한군데로'에 대한 의미역은 '방향'과 '착점'으로 나뉘므로 격틀 사전의 정보만으로는 의미역을 정할 수 없다.

4.3 UWordMap 어휘 하위범주를 사용한 의미역 선택

4.2에서 찾은 의미역 후보가 2개 이상일 때는 현재 어절의 격조사와 의미역이 일치하는 격틀 중에서 UWordMap 어휘의 하위범주를 사용하여 의미역을 선택한다. 용언의 동형이의어 중 남아 있는 격틀 후보와 일치하는 격틀을 가진 단어의 하위 범주를 검색하여 의미역을 부착할 위치에 있는 어절에 있는 단어가 해당 용언에서 사용할 수 있는 단어인지 확인한다. 이때 용언의 하위 범주에는 사용할 수 있는 가장 공통적인 단어가 들어가므로 단어의 상위어를 검색하여 하위 범주에 사용할 수 있는 단어가 나오는지 확인한다. 의미역을 부착할 위치에 있는 어절에 붙일 수 있는 의미역 후보가 1개로 좁혀지면 의미역을 부착한다.



[그림 5] 모이다 010202의 하위 범주 정보

- 4.2의 과정에서 결정하지 못한 '한군데로'에 대한 의미역을 용언의 하위범주를 사용하여 결정하도록 격조사 '로'에 대해서 '모이다 010201'과 '모이다 010202'의 하위범주를 검색한다.
- 하위 범주 검색한 결과 '모이다 010201'의 '로'와 함께 쓰일 수 있는 단어는 '공간', '강의', '건물'이다.
- '모이다 010202'의 '로'와 함께 쓰일 수 있는 단어는 '차', '퇴직', '판단'이다.
- 7번 어절의 '한군데로'의 '한군데'는 용언의 하위범주에 없지만, [그림 2]와 같이 '한군데'의 상위어를 따라가면 '모이다 010201'의 '로'와 함께 쓸 수 있는 '공간'이 나오고, '모이다 010202'와 함께 쓸 수 있는 단어는 나오지 않는다.
- 따라서 하위 범주 정보가 있는 '모이다 010201'의 격틀인 {X:대상 Z:방향-으로/로}가 적용되어 '한군데로'의 의미역은 '방향'으로 결정된다.

4.4 격조사-용언 별 의미역 통계를 사용한 의미역 선택

4.2와 4.3의 방법을 사용해도 의미역 구분이 되지 않는 경우에는 남아 있는 의미역 후보 중에서 격조사-용언 별 의미역 통계를 조사하여 어절의 격조사-용언을 기준으로 가장 많이 나온 의미역을 부착한다[3].

5. 결론

본 논문에서 격틀 사전과 용언의 하위 범주 정보를 이용하여 28,788개의 한국어 문장에서 3,880가지의 용언에 대해 의미역을 부착하는 실험을 하였다. 정확도는 문장의 필수 논항에 부착된 의미역과 정답 파일에 있는 의미역의 일치 여부로 평가한다. 아래의 3가지 방법으로 의미역을 부착해본 후 정확도를 평가한 결과는 [표 1]과 같다.

① 격틀 사전, 격조사-서술어 통계 사용
이전 논문[3]에서 사용했던 방법으로, 격틀 사전에서 격조사와 의존하는 용언을 바탕으로 적합한 용언을 찾을 때, 격틀 사전에 있는 격틀 전체를 보지 않고 격조사-의미역 쌍에서 한 격조사에 대한 의미역이 1개인 경우에는 바로 부착하고, 2개 이상인 경우에는 격조사-서술어 별 의미역 통계를 사용하여 후보 중 가장 많이 나온 의미역을 부착한다.

② 문형 일치 격틀 사전, 격조사-서술어 통계 사용
본 논문의 4.1~4.4의 의미역 부착 과정 중 하위범주 정보(4.3)를 사용하지 않고 의미역을 부착한다.

③ 문형 일치 격틀 사전, 용언의 하위범주, 격조사-서술어 통계 사용
본 논문의 4.1~4.4의 의미역 부착 과정을 모두 사용하여 의미역을 부착한다.

[표 1] 의미역 부착 방법에 따른 정확률

실험 방법	정확하게 부착된 의미역의 수	부착된 의미역의 수	정확률
①	27,557	36,806	74.87%
②	23,057	28,502	80.896%
③	23,930	29,866	80.125%

본 논문에서 격틀 사전과 어휘 하위 범주를 사용하여 한국어 문장에 의미역을 부착하는 실험을 하였다. 실험 결과 격틀 사전을 사용하여 의미역을 결정할 때 용언의 격틀과 문장의 문형이 일치하는 의미역을 부착할 경우 80.896%의 정확률로 이전의 실험 방법으로 의미역을 부착하였을 때보다 더욱 정확하게 의미역을 부착하였다. 또한, 용언의 하위 범주를 사용하여 의미역을 부착하면 격틀 사전만으로는 의미역을 구분하기 어려운 경우에도 용언과 함께 쓸 수 있는 단어인지를 구분하여 의미역을 더욱 정확하게 부착할 수 있을 것으로 예상하였으나 용언의 하위 범주 정보가 완전하게 구축되지 않아 80.125%의 정확률로 용언의 하위 범주를 사용하지 않았을 때 보다 다소 떨어짐을 보였다.

향후 연구로 한국어 의미역 결정에 필요한 자질을 더 찾아내어 더욱 세밀한 규칙을 적용하여 의미역 부착의 정확도를 향상시키는 방법을 연구할 예정이다.

참고문헌

[1] 김병수, 이용훈, 나승훈, 김준기, 이종혁, “부트스트래핑 알고리즘을 이용한 한국어 격조사의 의미역 결정”, 한국컴퓨터종합학술대회, vol.33, no.1, pp.4-6, 2006.
[2] 이창기, 임수종, 김현기, “Structural SVM 기반의 한국어 의미역 결정”, 정보과학회논문지, vol.42, no.2, pp.220-226, 2015.
[3] 김완수, 옥철영, “한국어 격틀사전과 의미역 빈도 정보를 사용한 한국어 의미역 결정”, 2015 컴퓨터종합학술대회 논문집, pp.651-653, 2015.6.
[4] 김윤정, 김완수, 옥철영, “전산언어학에서의 한국어 필수논항의 의미역 상정과 제고”, 한국언어정보학회 <언어와 정보>, vol.18, no.2, pp.169-200, 2014.
[5] 옥철영, “어휘의미 관계 데이터베이스 확장”, 국립국어원 발간자료, 2010.11.
[6] 배영준, 옥철영, “한국어 어휘지도(UWordMap)와 API 소개”, 한국정보과학회언어공학연구회 제 26회 한글 및 한국어 정보처리 학술대회, pp.27-31, 2014.
[7] 김홍순, 옥철영, “동형이의어 분별에 의한 한국어 의존관계 분석”, 정보처리학회논문지 소프트웨어 및 데이터공학, 제3권 제6호, pp.219-230, 2014.