Universal conceptual cognitive annotation(UCCA) 주석 체계의 한국어 적용 연구

오태환[0], 한지윤[1], 최현수[1], 박석원[1], 김한샘[†]
연세대학교 국어국문학과[0], 연세대학교 언어정보학협동과정[1†]
{ghksl0604, clinamen35, choehyonsu, pswon27, khss}@.yonsei.ac.kr

A Study on UCCA for Korean Semantic Analysis

Tae-Hwan Oh[0], Ji-Yoon Han[1], Hyon-Su Choe[1], Seok-Won Park[1], Han-Saem Kim[†]
Department of Korean Language and Literature, Yonsei University, Seoul, South Korea[0]
Institution of Language and Information Studies, Yonsei University, Seoul, South Korea[1†]

요 약

본 논문은 Universal conceptual cognitive annotation(보편 개념 인지 주석, 이하 UCCA)를 한국어에 적용하는 방안에 대해 제시하였다. 우선 기존의 한국어 의미 분석 체계들의 장단점을 살펴본 뒤, UCCA가 가지고 있는 상대적인 장점들을 소개하였다. UCCA는 모든 언어에 대하여 일관적인 기술을 하려는 Meaning representation framework의 하나로, 보편언어적인 의미 분석 체계를 가지고 있다. 본고는 주석 단위와 문법적 요소의 관점에서 한국어의 특성을 반영하여 UCCA를 한국어에 적용하는 방안을 검토하였다.

주제어: Universal conceptual cognitive annotation, UCCA, 한국어 의미 분석, 의미 분석

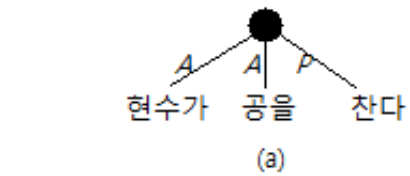
1. 서론

1997년부터 2007년에 걸쳐 한국어 언어 자원을 구축한 21세기 세종계획을 통해 형태 분석 말뭉치와 구문 분석 말뭉치, 형태-의미 분석 말뭉치가 완성되었다. 그러나 21세기 세종계획의 형태-의미 분석 말뭉치는 각 단어에 동형어 번호를 붙인 수준에 머물렀다[1]. 이후 한국전자통신연구원은 인공지능 프로젝트인 엑소브레인 개발을 위하여 구축한 의미역 분석 말뭉치를 공개하였고, 울산대학교에서도 의미역을 주석한 UCorpus-DP/SR 및 UPropBank를 공개하였다[2].

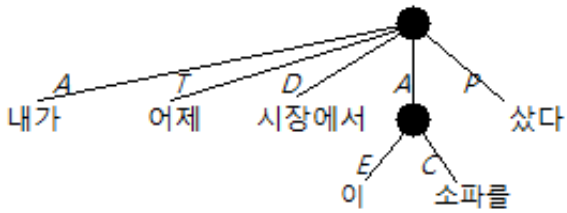
한국전자통신연구원의 의미역 분석 말뭉치는 Korean Proposition Bank(KPB)를 기반으로 하여 서술어와 서술어의 필수역, 부가역을 중심으로 문장을 분석한 것이다[3]. 기구축된 한국어 서술어 사전을 이용하여 의미를 분석하기 때문에 한국어에 특화된 분석 결과를 내놓을 수 있고, 주석표지가 16개로 비교적 복잡하지 않다는 장점이 있다. 그러나 반대로 KPB에 등재되지 않은 서술어에 대해서는 뚜렷한 의미 분석이 어렵다는 점, 한국어에 특화된 의미 분석이기 때문에 보편언어적인 분석이나 다른 언어와의 비교 연구가 어렵다는 한계점을 가지고 있다. Choe et al.(2019)에서는 문장의 논리적 의미를 그래프 구조로 표상하는 Abstract Meaning Representation(이하 AMR)을 한국어에 적용하고자 하였다[4][5]. 그러나 이 방법론 또한 KPB를 이용하여 의미 분석을 하고 있기 때문에 KPB에 등재되지 않은 서술어가 나타날 경우 이를 바르게 처리하기가 어렵다. 또한 100개 이상의 주석 표지를 가지고 있어 주석의 난도가 높다.

본고에서는 한국어 의미 분석을 위하여, AMR과 함께 Meaning representation framework의 한 유형이며 보편적인 의미 분석 체계로 새로 주목받고 있는 Universal conceptual cognitive annotation(보편 개념 인지 주석, 이하 UCCA)를 소개하고, 이를 한국어에 적용시켜 볼 것이다.

2. UCCA



(a)



(b)

그림 1 : UCCA 주석 예시

위 그림1과 같이 UCCA는 문장을 장면(Scene)으로 분석한다[6]. 여기서 장면이란 행동 또는 상태에 대한 지칭으로, 그것이 일어난 시간, 장소 등의 정보를 가질 수

있다. 간단히 말하여, 한 가지의 행동이나 상태가 이뤄진 문장은 곧 하나의 장면이라고 할 수 있다. 따라서 '철수가 밥을 먹었다.' 는 한 개의 장면을, '철수가 밥을 먹고 잠을 잤다.' 는 두 개의 장면을 가지고 있다.

UCCA는 문장의 표면형을 주석의 대상으로 삼으며 KPB와 같은 기구축된 언어 사전이 필요하지 않다. 또한 주석 표지가 13개로 간소하며 문장 표면에 나타나는 구문 구조를 바탕으로 의미를 주석하기 때문에 상대적으로 주석 난도가 낮은 편이다.

2.1 주석 대원칙

UCCA 주석 체계에서 의미 분석을 위한 여섯 가지의 대원칙은 아래와 같다[7].

- UCCA 주석 작업은 주로 한 단락 또는 여러 단락 이상의 여러 문장을 대상으로 한다. 전체 문장을 분석하기에 앞서 전체 문서를 읽어 맥락을 이해하여야 한다.
- UCCA 주석 작업은 문장을 일정한 단위로 나누며, 각 단위들의 관계, 관계의 참여자 또는 참여자와의 관계를 주석된다.
- 구두점을 제외한 모든 토큰은 단위에 포함된다.
- 단위는 다른 하위 단위를 포함하여 계층 구조를 이룰 수 있다.
- 각 단위에는 범주가 할당되는데, 이는 단위가 참여하는 관계 속에서의 역할을 나타낸다. 할당된 범주는 단위를 이루는 단어의 의미를 나타내는 것은 아니다.(즉, 단어의 의미가 아닌 문장 내에서의 의미 역할에 따라 범주가 할당된다.)
- UCCA 주석에서는 중의성을 따로 표시하지 않는다. 작업자의 판단에 따라서 가장 가능성이 높은 주석을 한다.

2.2 주석 단위

UCCA 주석에서 각 단위들은 크게 장면(Scenes), 비장면 단위(Non-scene Units)와 장면 간 관계(Inter-Scene relations)로 나눌 수 있다.

장면(Scene categories)		
P	Process	진행
S	State	상태
A	Participant	참여자
D	Adverbial	부가 표현
T	Time	시간

표 1 UCCA 장면

각 장면은 장면의 유형을 결정하는 하나의 주 관계를 가진다. 만약 장면이 시간에 따라 변하는 것이라면 진행(P), 시간에 관계없이 고정적인 것이라면 상태(S)가 할당된다.

각 장면은 임의의 수의 참여자(A)와 시간(T), 부가 표현(D)을 가질 수 있다.

비장면 단위(Non-scene Units)		
C	Center	중심
E	Elaborator	상술 표현
N	Connectors	연결어
Q	Quantifier	양화사

표 2 UCCA 비장면 단위

장면이 아닌 단위들도 더 세부적으로 분석될 수 있는데, 이를 비장면 단위라고 한다. 비장면 단위 내에서 다른 단위의 수식을 받는 단위를 중심(C)이라고 한다. 비장면 단위들은 무조건 한 개 이상의 중심을 가지고 있어야 한다. 중심에 다른 정보를 더하거나 수식해주는 단위를 상술 표현(E)이라고 하며, 둘 이상의 단위를 연결시켜주는 단위는 연결어(N), 수량을 나타내는 단위는 양화사(Q)라고 한다.

장면 간 관계(Inter-Scene relations)		
E-Scene	Elaborator Scene	상술 장면
A-Scene	Participant Scene	참여 장면
C-Scene	Center Scene	중심 장면
H-Scene	Parallel Scene	병렬 장면
L	Linker	연계어
G	Ground	배경

표 3 UCCA 장면 간 관계

하나의 장면은 다른 단위의 하위 성분이 될 수도 있다. 상술 장면(E-Scene)은 비장면 단위인 중심(C)을 수식하게 된다. 참여 장면(A-Scene)은 상위 장면의 참여자가 된다. 중심 장면(C-Scene)은 단위가 장면을 일으키지만 장면이 포함된 구 자체는 장면으로 분석될 수 없는 특정한 경우에만 한정된다. 장면 간 관계를 만족하는 장면 중 상술 장면, 참여 장면, 중심 장면이 아닌 장면은 곧 병렬 장면(H-Scene)이다. 병렬 장면은 연계어(L)를 가질 수도 있으며, 그러지 않을 수도 있다. 배경(G)는 어떤 발화 사건(장면)을 외부에서 판단하는, 다른 장면과 직접적인 관계를 가지고 있지 않은 단위에 할당된다.

기타(Others)		
R	Relator	관계어
F	Function	기능어
UNA	Unanalisable	주석불가

표 4 UCCA 기타 단위

그 외 단위로는 관계어(R), 기능어(F)가 있다. 관계어와 기능어는 장면과 비장면 단위의 하위 요소로 모두 나타날 수 있다. 관계어는 여러 개의 성분들을 이어주는 것인데, 병렬 장면을 이어주는 연계어(L)나 비슷한 유형이나 역할을 가진 중심을 이어주는 연결어(N)을 제외한 다른 성분들을 말한다. 기능어는 다른 관계나 참여자를 도입하지 않으며, 더 큰 단위의 일부로서만 해석될 수 있는 것들이다. 또는 시제나 초점과 같이 UCCA의 주석

체계에서 다루어지지 않는 의미들을 나타내기도 한다. 주석불가(UNA)는 UCCA 주석 체계에 따라서 주석할 수 없는 단위에 할당된다.

3. UCCA의 주석의 실제

3.1. 장면 단위 주석

모든 장면 단위들은 무조건 주관계인 진행(P)이나 상태(S)에 포함되어야 한다. 나머지 장면 단위인 참여자(A), 부가 표현(D), 시간(T)은 수의적으로 등장할 수 있다.

• 진행(P)
한국어 문장의 동사를 진행(P)로 분석한다.
예) 철수가A 간다P.

• 상태(S)
한국어 문장의 형용사를 상태(S)로 분석한다.
예) 철수가A 좋다S.

• 참여자(A)
하나의 장면은 여럿의 참여자를 가질 수 있다. 참여자에는 주관계의 중요한 객체들이 해당되며 장소까지 포함된다. 또한 구체 명사와 추상 명사 모두 참여자에 해당된다.
예) 철수가A 집으로A 간다P.

• 부가 표현(D)
부가 표현은 새로운 장면을 도입하지는 않으며, 기존 장면이나 장면의 진행(P), 상태(S)를 수식한다. 한국어에서는 '아마, 빨리, 힘들게' 등의 부사어가 이에 해당한다.
예) 철수가A 빠르게D 집으로A 간다P.

• 시간(T)
장면이 발생한 시간을 확정하는 단위는 시간(T)로 주석된다. 시간에는 빈도와 기간도 포함된다. 한국어의 시간 명사구가 이에 포함된다.
예) 철수가A 어제T 빠르게D 집으로A 갔다P.

3.2. 비장면 단위 주석

비장면 단위는 새로운 장면의 도입은 없지만 내부적으로 더 분석할 수 있는 단위이다. 비장면 단위에는 필수적으로 한 개 이상의 중심(C)이 필요하다.

• 중심(C)
중심은 상술 표현(E)이나 양화사(Q)의 수식을 받거나 연결어(N)으로 이어진다. 만약 이어지는 단위들 중에서 중심과 상술 표현을 구분하기 어려울 경우에는 둘 모두에게 중심을 할당한다.
예) 그의E 손C / 해저C 500미터C

• 상술 표현(E)
상술 표현은 중심에 몇 가지 정보를 더한다.
예) 영국의E 여왕C / 초콜릿E 쿠키C

• 연결어(N)
연결어는 비슷한 유형이나 역할을 가지는 두 개 이상의 중심을 연결한다.
예) 철수C 그리고N 영희C

• 양화사(Q)
양화사는 중심의 수량이나 규모를 나타내는 표현들에 할당된다. 한국어의 수관형사가 이에 해당된다.
예) 세Q 사람C / 사과C [삼Q 키로C]Q

3.4. 장면 간 관계 주석

어떤 장면은 다른 장면의 하위 요소가 될 수도 있다. 장면 간 관계는 대괄호([])로 묶어서 나타낸다.

• 상술 장면(E-Scene)
한국어에서는 체언을 수식하는 관형절이 상술 장면이 될 수 있다.
예) [[[내E 공책을C]A 찢은P]E 강아지는C]A 갈색이다S.

• 참여 장면(A-Scene)
상위 장면의 참여자가 되는 명사절 속의 장면이 참여 장면이 된다.
예) 나는A [너와A 마주하기가P]A 싫다S.

• 중심 장면(C-Scene)
중심 장면은 최상위 장면의 참여자가 되는 명사절 속에 다시 관형절의 수식을 받는 명사절이 내포되는 등의 특수한 상황에 나타날 수 있다.
예) [저기에A [[나와A 친한P]E [[키가A 큰P]E 사람이C]C]A 보인다P].

• 병렬 장면(H-Scene)
병렬 장면은 두 개의 절이 대등하게 접속할 때 나타난다. 병렬 장면은 연계어(L)와 함께 나타날 때도 있고, 그렇지 않을 때도 있다.
예) [그는A 춤을A 추면서P]H [노래를A 불렀다P]H.

• 연계어(L)
연계어는 병렬 장면을 이어주는 단위이다. 한국어에서는 접속부사나 '만약, 혹시' 등의 일부 문장부사가 이에 속한다.
예) 만약L [너가A 간다면P]H [나도A 갈거야P]H.

• 배경(G)
배경은 발화 장면에서 조금 떨어진 지점에서 장면에 대한 평가를 내리는 절에 해당한다.
예) 놀랍게도G, 우리는A 정시에T 도착했다P.

UCCA의 나머지 두 주석 표지인 관계어(R)와 기능어(F)

는 한국어에 적용할 수 없다. 영어에서의 관계어는 주로 'of, in' 등의 전치사로, 한국어에서는 대부분 조사가 그 역할을 분담하고 있다. 본고에서는 한국어 UCCA 분석 단위를 어절로 설정하고 있기 때문에 어절 구성 요소인 조사에 이러한 단위를 적용할 수 없다. 또한 기능어는 영어의 'is, am'과 같은 계사, 'to'나 'let' 등에 할당된다. 관계어와 마찬가지로 한국어에서는 이러한 단위가 단독 어절로 나타나지 않아 별도의 주석 표지를 할당할 수 없다.

3.4. 한국어 적용의 쟁점

한국어에 UCCA의 주석 체계를 그대로 적용하기 어려운 부분에 대해서는 다음과 같이 기준을 정하였다. 비록 UCCA가 보편성을 전제로 하는 의미 분석 체계이기는 하지만, 한국어의 개별적인 특성을 고려하지 않고는 올바른 적용을 할 수 없을 것이다.

- 주석 단위

한국어는 한 어절 안에 여러 개의 형태소나 단어가 포함되어 있을 수 있다. 따라서 한국어는 분석의 단위를 형태소, 어절의 두 가지로 설정할 수 있으며 둘 중 무엇을 분석 단위로 선택하는가에 따라서 전혀 다른 분석 체계를 만들 수 있다.

본고에서는 UCCA 주석 체계가 비교적 형태 분석과는 연관성이 떨어진다는 점, 형태소를 단위로 설정할 경우 문장의 표면형을 그대로 살리는 분석이 어려워진다는 점 등을 이유로 어절을 주석 단위로 삼는다. 또한 형태소를 분석 단위로 삼을 경우 조사나 지정사가 모두 분리되어 각각의 표지를 할당받게 되므로 주석의 복잡성과 난도가 올라가게 된다.

- 지정사 '-이-'

한국어의 지정사와 유사한 성격을 지닌 영어의 Be동사는 체언이 아닌 형용사와도 함께 나타나며, 특히 UCCA 주석 체계에서는 체언과 함께하는 Be동사는 상태(S)로, 형용사와 함께하는 Be동사는 기능어(F)로 달리 분석하는 모습을 보인다. 한국어의 지정사는 전통 문법에서 주로 용언의 일종으로 취급받으며, 체언과 결합하여 문장 내에서 서술어의 기능을 한다. 본고에서는 이미 어절을 분석 단위로 삼았기 때문에 체언과 지정사의 결합을 하나의 단위로 취급할 것이며, 전통 문법의 견해를 따라 이를 상태(S)로 주석하기로 한다.

- 보조용언

UCCA 주석 체계에서는 'want, have, shall, may' 등의 영어 조동사와 'not'을 부가적인 의미를 더하는 부가 표현(D)으로 주석하고 있으므로 한국어의 보조용언 '싶다, 하다, 않다' 등도 부가 표현으로 주석하기로 한다.

4. 결론

본고에서는 한국어 문장의 개념 표상을 위하여 지금까지 적용되지 않았던 UCCA라는 주석 체계를 소개하였다. UCCA는 문장의 표면을 그대로 보이며 주석 표지의 개수가 적어 주석이 용이하다. 또한 보편언어적 의미 분석 체계이기 때문에 다른 언어와의 비교 연구 등에 유리한 점을 가지고 있다. 한국어에 UCCA 주석 체계를 본격적으로 적용하기 위해서는 '버리다, 있다, 되다' 등 문법적 기능을 가지고 있는 보조용언과 '-ㄹ 수 있-'과 같은 우언적 구성 등 세부적인 사항에 대한 가이드라인 등을 갖추어야 한다는 과제가 남아 있다. UCCA를 비롯한 다양한 의미 주석 체계에 대한 연구가 한국어 문장의 의미 처리에 도움이 되기를 기대한다.

참고문헌

[1] 국립국어원, "국어 기초 자료 구축", 21세기 세종계획 국어 기초 자료 구축 분과 보고서, 2005.

[2] 김완수, 옥철영, "한국어 격틀 사전과 의미역 빈도 정보를 사용한 한국어 의미역 결정", 한국정보과학회 학술발표논문집, pages 651-653, 2015.

[3] 임수종, 권민정, 김준수, 김현기, "ExoBrain을 위한 한국어 의미역 가이드라인 및 말뭉치 구축", 제27회 한글 및 한국어 정보처리 학술대회 논문집, 2015.

[4] Hyon-Su Choe, Ji-Yoon Han, Hye-Jin Park, and Han-Saem Kim, "Copula and Case-Stacking Annotations for Korean AMR", In Proceedings of the First International Workshop on Designing Meaning Representations (pp. 128-135), 2019.

[5] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schnetder, N., "Abstract Meaning Representation (AMR) 1.2.6 Specification", 2019.

[6] Omri Abend and Ari Rappopor, "Universal conceptual cognitive annotation (ucca)", Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013.

[7] Omri Abend and Ari Rappoport, "UCCA's Foundational Layer: Annotation Guidelines v2", 2018.