한국어 생략어복원 가이드라인

류지희†, 임준호, 임수종, 김현기

한국전자통신연구원 언어지능연구그룹
{chrisjihee, joonho.lim, isj, hkk}@etri.re.kr

Korean Zero Anaphora Resolution Guidelines

Jihee Ryu†, Joon-Ho Lim, Soojong Lim, Hyunki Kim
Electronics and Telecommunications Research Institute

요 약

말과 글에서 유추가 가능한 정보에 대해서는 사람들이 일반적으로 생략해서 표현하는 경우를 볼 수 있다. 사람들은 생략된 정보를 문맥적으로 유추하여 이해하는 것이 어렵지 않지만, 컴퓨터의 경우 생략된 정보를 고려하지 못해 주어진 정보를 완전하게 이해하지 못하는 문제를 낳게 된다. 우리는 이러한 문제를 생략어복원을 통해 해결할 수 있다고 여기면서 본 논문을 통해 한국어 생략어복원에 대해 정의하고 기술 개발에 필요한 말뭉치 구축 시의 생략어복원 대상 및 태깅 사례를 포함하는 가이드라인을 제안한다. 또한 본 가이드라인에 의한 말뭉치 구축 및 기술 개발을 통해서 엑소브레인과 같은 한국어 질의응답 시스템의 품질 향상에 기여하는 것이 본 연구의 궁극적인 목적이다.

주제어: 자연어처리, 한국어 생략어복원, 생략어복원 태깅 가이드라인, 엑소브레인

1. 서론

우리의 일상 언어 사용에서 경제성의 원리가 작용되어 청자가 알고 있는 것이나 충분히 유추가 가능한 정보는 축약하거나 생략하여 표현하는 경우가 있다. 축약되었거나 생략된 표현은 대용어(anaphora: 조응어 또는 조응대용어)로 나타날 수 있고, 컴퓨터가 이것을 명확하기 인식하기 위하여 대용어 해결(anaphora resolution)이라는 자연어처리 문제로 정의하여 다루고 있다[1]. 생략어복원(zero anaphora resolution)은 어떠한 동사 표현 어구나 명사 표현 어구에서 일부 문장 성분이 미리 나타나 유추가 가능하거나 암묵적으로 알고 있기에 문장 내에서 생략된 해당 성분을 찾아 복원해주는 문제이다. 본 논문에서는 생략된 문장 성분을 생략어(zero anaphora: 생략된 대용어 또는 무형대용어)라 하고, 생략된 문장 성분이 종속되는 대상을 지배소(head)라 하고, 생략어가 복원되어야할 원래 표현을 선행어(antecedent)라고 한다.

생략어복원은 상호참조해결와 달리 선행어를 대신하여 사용된 대용어가 대명사나 약어 등의 형태로 나타나는 것이 아니라 아예 생략되었다는 것이 차이점이라고 할 수 있다. 대용어가 생략되어 있기 때문에 주어진 문장을 읽다가 특정 동사 표현 어구나 명사 표현 어구 내에서 생략어가 존재함을 먼저 알아내야 한다. 그 뒤, 해당 생략어에 대한 선행어를 결정하는 과정에서 문서 내에 나타난 표현 이외에도 암묵적이기에 문서 내에 존재하지 않는 표현까지 고려해야하는 특수성이 있다.

이러한 생략어를 복원시킨 결과는 텍스트 상에서 이전에 언급되었거나 암묵적으로 표현된 정보를 찾아주어 텍스트의 의미를 보다 명확하게 이해하게 해주고, 담화나 문서 내에서 언급하는 대상에 대한 정보를 일관성 있게 유지하게 해준다. 따라서 생략어복원 문제의 해결은 문서에서 등장하는 개체와 그에 대한 정보를 이해하는데 상당히 중요한 역할을 하며 정보 검색, 정보 추출, 질의응답, 문서 요약, 기계 번역 등에서 유용하게 사용될 수 있다.

생략어복원을 이해하기 위해서 [2]에서 들었던 예시를 통해 명시적인 대용어와 암묵적인 대용어가 나타나는 경우를 살펴볼 수 있다.

(ㄱ) **철수는**† 학교에 갔다. 가는 도중 **그는**¶ 영희를 만났다.
(ㄴ) **철수는**† 학교에 갔다. 가는 도중 영희를 만났다.

(ㄱ)에서 대용어 “그는”의 선행어는 이전 문장의 “철수는”이다. 대명사로 표현된 대용어 “그는”의 선행어를 찾아내는 대용어 해결 문제는 상호참조해결로 해결이 가능하다. 반면, (ㄴ)에서는 대용어가 생략되어 있고, 이것은 생략어복원으로 해결이 가능하게 된다. 두번째 문장에서 “만나다”는 동사에 대한 주어가 생략되어 있어 생략어가 있음을 먼저 인지한 뒤, 해당 생략어에 대한 선행어는 앞에서 등장했던 “철수는”임을 판단할 수 있어야 한다. 즉, “가는 도중 철수는 영희를 만났다.”로 생략어를 복원시키는 것이 정보를 보다 명확하고 구체적으로 드러낼 수 있다는 사실을 인지해야 한다. 이러한 복원 결과로부터 지식을 생성할 때, (ㄷ)과 같이 표현할 수 있는 관계 정보를 알 수 있게 된다.

(ㄷ) [**철수**† - 만나다 - 영희]

생략어복원 문제 중에서도 위키피디아와 같은 백과사전 본문에서는 표제어를 암묵적으로 알고 있다고 판단하여 문장 내에서 대부분 표현하지 않는 측면이 있다.

(ㄹ) **지미 카터는**† 조지아 주 섬터 카운티 플레인스 마을에서 태어났다. 조지아 공과대학교를 졸업하였다. 그 후 해군에 들어가 전함·원자력·잠수함의 승무원으로 일하였다.

(ㅁ) 지미 카터는† 조지아 주 섬터 카운티 플레인스 마을에서 태어났다. 지미 카터는¶ 조지아 공과대학교를 졸업하였다. 그 후 해군에 들어가 전함·원자력·잠수함의 승무원으로 지미 카터는¶ 일하였다.

(ㄹ)은 "지미 카터"에 대한 위키피디아 본문 설명의 일부분이다. 이 텍스트의 첫번째 문장에서 나타난 선행어 "지미 카터는"으로 복원시킬 수 있는 대용어들이 첫 번째 문장 이후에 모두 생략되어 있음을 알 수 있다. 생략어가 복원된 결과의 예로 (ㅁ)과 같은 결과를 들 수 있고, 복원되는 생략어의 위치는 한국어의 어순 특성상 자유로울 수 있음을 알 수 있다. 생략어복원 문제는 필요에 따라 백과사전 본문 내에서 백과사전의 표제어로 생략된 대용어를 복원시키는 표제어 복원 문제로 축소시킬 수 있다.[2-4]

한국어 뿐만 아니라, 중국어[5-9]와 일본어[10-14]에 대해서도 이러한 생략어복원 문제를 해결하기 위한 방법들이 각각 제안되어 왔다. 방법론 면에서도 규칙과 구문적 패턴을 활용하는 방법부터 전통적인 기계학습 방법에서 최근에는 딥러닝을 활용하는 방법까지 다양하게 시도되고 있다.

이러한 생략어복원을 해결하는 기술을 한국어 질의응답 시스템인 엑소브레인에 활용하기 위해서 본 논문은 생략어복원 대상과 태깅 사례에 대한 가이드라인을 제안하고자 한다. 본 논문의 구성은 다음과 같다. 2장에서는 생략어복원의 대상을 소개하고, 3장에서는 태깅 결과물 포맷과 함께 태깅 사례들을 소개한다. 그리고 4장에서는 말뭉치 구축 도구를 소개한다. 마지막 5장에서는 결론을 맺도록 한다.

2. 생략어복원 대상

생략된 모든 정보를 복원하는 데에는 한계가 있으므로, 본 논문에서는 각 생략어복원 주요 개념에 대한 후보를 다음과 같이 정한다.

- 지배소 후보
  - 동사 표현 어구
  - 주어 및 목적어 역할을 하는 명사 표현 어구
- 생략어 후보
  - 동사 표현 어구에서 생략된 주어
  - 동사 표현 어구에서 생략된 목적어 및 필수 부사어
  - 주어 및 목적어 표현 어구에서 생략된 관형어
- 선행어 후보
  - 해당 문서의 표제어
  - 필자가 염두에 두고 있는 포커스
  - 상호참조해결에 의해 탐지된 멘션 및 개체
  - 암묵적으로 또는 상식적으로 알고 있거나 아직 알려지지 않은 개념

지배소 후보에 대해서는 2.1절에서, 생략어 후보에 대해서는 2.2절에서, 선행어 후보에 대해서는 2.3절에서 각각 설명한다.

2.1. 지배소 후보

2.1.1. 동사 표현 어구

본 논문의 생략어복원 대상이 되는 생략어에 대한 지배소 후보로서 먼저 동사 표현 어구를 생각해볼 수 있다. 동사 표현 어구는 일반적으로 담화나 문서 내에서 개체와 개체 또는 개체와 값 간의 사건, 행동 및 상태와 같은 주요한 정보를 서술하는 형태이다. 이러한 동사 표현 어구에서 생략 표현된 정보가 있어 이를 일차적인 지배소 후보로 생각할 수 있다. 동사 표현 어구를 인식하기 위해 의존 구문분석 결과 및 의미역 부착 결과를 참고할 수 있으며, 구문 태그로 VP[용언]나 VNP[긍정지시사구]를 가지는 것이 이에 해당한다고 할 수 있다. 이에 대한 예시는 다음과 같다.

케냐 공화국 또는 케냐는 동아프리카의 공화국이다*[VNP]. 인도양에 면해*[VP] 있으며 북동쪽으로 소말리아, 북쪽으로 에티오피아와 남수단, 서쪽으로 우간다, 남쪽으로 탄자니아와 국경을 맞닿고*[VP] 있다.

2.1.2. 주어 및 목적어 역할을 하는 명사 표현 어구

다음으로 주어 및 목적어 역할을 하는 명사 표현 어구를 생각해볼 수 있다. 명사 표현 어구는 일반적으로 담화나 문서 내에서 어떠한 구체적 개체나 추상적 개념의 명칭을 나타내는 형태이다. 이러한 명사 표현 어구에서도 생략된 정보가 있어 이를 이차적인 지배소 후보로 생각할 수 있다. 그 중에서도 주어 및 목적어 역할을 하는 명사 표현 어구는 문장에서 주요한 성분이므로 생략 표현된 정보까지 이해하는 것이 중요할 수 있다. 주어 및 목적어 역할을 하는 명사 표현 어구를 인식하기 위해 구문 태그로 NP[체언]를 가지면서 기능 태그로 SBJ[주어]와 OBJ[목적어]를 가지는 것이 이에 해당한다고 할 수 있다. 이에 대한 예시는 다음과 같다.

케냐 공화국 또는 케냐는*[NP_SBJ] 동아프리카의 공화국이다. 인도양에 면해 있으며 북동쪽으로 소말리아, 북쪽으로 에티오피아와 남수단, 서쪽으로 우간다, 남쪽으로 탄자니아와 국경을*[NP_OBJ] 맞닿고 있다. 수도는*[NP_SBJ] 나이로비이며 공용어는*[NP_SBJ] 영어와 스와힐리어이다.

2.2. 생략어 후보

한가지 전제할 것은 생략어 태깅은 생략된 성분을 파악하는 것이 목표이지 자연스러운 완성된 문장을 만드는 것이 목표가 아니라는 것이다. 따라서 생략어가 어느 지배소에 속하는지, 선행어는 무엇인지까지만 태깅해주면 된다.

2.2.1. 동사 표현 어구에서 생략된 주어

생략어 후보 종류 중에 가장 많은 비율을 차지하는 것은 바로 생략된 주어이다. 일반적으로 동사는 최소 1개의 주어가 필수적으로 필요하지만 앞에서 등장한 개체를 이미 알고 있을 것이라는 가정 하에 생략하여 표현하는 경우가 많다. 이에 대한 예시는 다음과 같다.

케냐 공화국 또는 케냐는 동아프리카의 공화국이다. [?는]¶ 인도양에 면해*[VP] 있으며 북동쪽으로 소말리아, 북쪽으로 에티오피아와 남수단, 서쪽으로 우간다, 남쪽으로 탄자니아와 [?는]¶ 국경을 맞닿고*[VP] 있다.

2.2.2. 동사 표현 어구에서 생략된 목적어 및 필수 부사어

생략어 후보 종류 중에 다음으로 많은 비율을 차지하는 것은 바로 생략된 목적어 및 필수 부사어이다. 일반적으로 타동사는 필요에 따라 1개 또는 그 이상의 목적어 및 필수 부사어가 필요하지만 생략하여 표현하는 경우가 있다. 이에 대한 예시는 다음과 같다.

> 월스트리트 저널은 다우존스가 발행하는 조건으로서 세계 10대 신문 중 하나이며, 세계적으로 가장 영향력이 큰 경제지이다. 1889년 다우존스사의 다우가 기업과 금융 관계를 전문적으로 보도하고자 [?를]¶ 창간했다*[VP].

2.2.4. 주어 및 목적어 표현 어구에서 생략된 관형어

적은 비율이기는 하지만 정보 추출 관점에서 추가로 다룰 필요가 있는 것은 바로 생략된 관형어이다. 주어 및 목적어를 표현하는 데 있어서 앞에서 등장한 개체가 관형어가 됨에도 불구하고 이미 알고 있는 정보가 될 때는 이를 생략한 채 표현하는 경우가 있다. 이에 대한 예시는 다음과 같다.

> 케냐 공화국 또는 케냐는 동아프리카의 공화국이다. 인도양에 면해 있으며 북동쪽으로 소말리아, 북쪽으로 에티오피아와 남수단, 서쪽으로 우간다, 남쪽으로 탄자니아와 국경을 맞닿고 있다. [?의]¶ 수도는*[NP_SBJ] 나이로비이며 [?의]¶ 공용어는*[NP_SBJ] 영어와 스와힐리어이다.

2.3. 선행어 후보

본 절에서 설명하는 선행어 후보는 표제어이면서 멘션이 되는 복합적인 형태도 가능한데, 여기에서의 구분은 선행어를 어느 한가지 종류로 규정하려는 것이 아니고 선행어의 후보로 살펴야할 대상을 단계적으로 제공함으로써 태깅 작업자가 놓치지 않고 선행어를 찾도록 가이드하기 위함이다. 선행어 후보에 대한 전제사항은 상호참조해결에 의해 탐지된 멘션 및 개체에 대해서는 멘션의 중심어 중 생략어와 가까운 위치의 것을 선행어로 결정하는 것을 원칙으로 한다는 것과, 포커스에 대한 고려는 새롭게 고려되는 의미 자질이므로 포커스로 여길 수 있는 것을 선행어로 결정할 수는 있지만 포커스에 대한 태깅은 추후에 다시 논의하여 태깅 가이드를 구체화한 뒤 진행하도록 한다는 것이다. 본 가이드라인에서는 후보의 발생 개수가 적고 명시적인 것에서부터 범위가 넓거나 암묵적인 것으로 순서를 정하여 소개한다.

2.3.1. 해당 문서의 표제어

생략어에 대한 선행어 후보로서 먼저 해당 문서의 표제어를 생각해볼 수 있다. 해당 문서의 표제어는 텍스트를 이해하는 데 있어서 가장 먼저 접하는 정보이면서 해당 문서 전체를 한 마디로 설명할 수 있는 개념이기 때문에 필자는 독자가 이를 이미 알고 있다고 생각할 수 있다. 특히 백과사전 종류의 텍스트에서는 대부분의 문장이 표제어를 구체적으로 설명하기 위한 문장들이 많으며, 컴퓨터가 문서 내의 문장들을 분석할 때는 생략된 표제어를 복원시켜야 백과사전에서 설명하는 내용을 일관적으로 이해할 수 있게 된다. 이러한 선행어가 사용된 예는 다음과 같다.

> (표제어 : 케냐†)

> 케냐 공화국 또는 케냐는 동아프리카의 공화국이다. [케냐†는]¶ 인도양에 면해*[VP] 있으며 북동쪽으로 소말리아, 북쪽으로 에티오피아와 남수단, 서쪽으로 우간다, 남쪽으로 탄자니아와 [케냐†는]¶ 국경을 맞닿고*[VP] 있다.

2.3.2. 필자가 염두에 두고 있는 포커스

생략어에 대한 선행어 후보로서 다음으로 필자가 염두에 두고 있는 포커스를 생각해볼 수 있다. 일반적인 말과 글에는 각 단락 또는 문장 단위에서 필자가 생각하고 있는 주된 포커스(focus)가 명시적으로 또는 묵시적으로 존재하는 경우가 있다. 포커스는 단위 텍스트 내에서 존재하지 않거나, 새롭게 등장하거나, 그대로 유지되거나, 기존의 포커스 중 하나로 이동할 수 있다. 상호참조해결이나 생략어복원과 같은 대용어 해결 문제에서 현재 등장한 멘션에 대해 참조되는 선행 멘션이나 생략어에 대한 선행어는 이러한 포커스인 경우가 있다. 따라서 생략어복원 문제에서 선행어를 더 정확하게 선택하는데 있어서, 현재의 포커스가 무엇인지 고려하여 생략어에 대한 선행어를 결정할 필요가 있다. 특히 뉴스 분야의 텍스트에서는 사건의 전말을 설명하기 위하여 전체 사건 내에서 부분적으로 포커스를 차츰 옮겨가면서 전체 사건의 내용을 전반적으로 다루는 경우가 종종 있다. 이러한 선행어가 사용된 예는 다음과 같다.

> 문재인 대통령은 21일 '경제 사령탑'인 경제부총리 겸 기획재정부 장관 후보자에 김동연(60) 아주대 총장, 외교부 장관 후보자에 여성인 강경화(62) 유엔 사무총장 정책특보를 각각 내정했다. … 김동연 부총리 후보자†는 충북 음성 출신으로 '고졸신화의 인간승리 드라마'로 불린다. [김동연 부총리 후보자†는]¶ 덕수상고 졸업 뒤 은행에 취직해*[VP] 직장생활을 하며*[VP] 행정고시와 입법고시에 동시 합격한*[VP_MOD] 입지전적의 인물로 평가 받는다*[VP]. …

2.3.3. 상호참조해결에 의해 탐지된 멘션 및 개체

생략어에 대한 선행어 후보로서 일반적으로는 상호참조해결에 의해 탐지된 멘션 및 개체를 생각해볼 수 있다. 상호참조해결에 대한 보다 상세한 내용은 한국어 상호참조해결 가이드라인 및 관련 연구 결과[15]를 참고할 수 있다. 상호참조해결(coreference resolution)은 임의의 개체(entity)에 대하여 다른 표현으로 사용되는 단어들을 찾아 서로 같은 개체로 연결해주는 자연어처리 문제이다.[16] 멘션(mention)은 상호참조해결의 대상이 되는 모든 명사구를 의미한다. 멘션에서 해당 구의 실질적인 의미를 나타내는 단어를 중심어라하며, 멘션은 중심어를 중심으로 이를 수식하는 수식어까지도 포함한다. 개체(entity)는 동일한 멘션의 집합으로써 상호참조해결의 결과이다. 선행 멘션(antecedent)과 현재 등장한 멘션 간의 참조를 해결하면 하나의 개체로 포함된다. 이러한 선행어가 사용된 예는 다음과 같다.

> 비텐베르크 대학교의 요한 스타우피츠 교수†는 루터가 성서에 대해 진지하게 공부하면 평안을 찾을 것이라고 생각하였다. 그래서 [요한 스타우피츠 교수†는]¶ 그를 성서학 교수사제로 임명하였는데*[VP], 스타우피츠 교수의 결정은 루터가 신앙적인 고민을 해결하는데 도움이 되었다.

2.3.4. 암묵적으로 또는 상식적으로 알고 있거나 아직 알려지지 않은 개념

생략어에 대한 선행어 후보로서 마지막으로 암묵적으로 또는 상식적으로 알고 있거나 아직 알려지지 않은 개념을 생각해볼 수 있다. 필자는 이러한 개념들은 굳이 명시적으로 표현하지 않아도 될 것이라고 생각할 수 있다. 게다가 만약 아직 알려지지 않은 개념이라면 표현하기 어려울 수 있다. 이런 개념에 대해 사람은 내용을 이해하면서 자연스럽게 생각해낼 수 있어도 컴퓨터는 외부 지식의 도움을 받지 않고서는 알기 어려운 정보일 수 있다. 특히 이런 종류의 선행어는 해당 문서 어디에도 발견되지 않을 수 있어서 대상을 정확히 찾는데 더 어려움이 있다. 이러한 선행어가 사용된 예는 다음과 같다.

케냐 공화국 또는 케냐는 동아프리카의 공화국이다. … 동아프리카에서 [어느 발굴가†에 의해]¶ 발견된✱[VP_MOD] 화석에 따르면 조상이 2백만 년 전 이 지역에서 살았다고 한다.

3. 생략어복원 태깅 사례

생략된 정보 중에는 사람조차 구체적으로 알기 어려운 정보가 존재하기도 하며, 알더라도 정보로서의 가치가 적은 것도 존재한다. 본 생략어복원 태깅의 궁극적 목적은 주어진 텍스트에서 정보로서의 가치가 있는 생략된 정보를 복원시킴으로써, 사람이 보기에도 의미가 명확해지고 그것을 해석하는 컴퓨터가 중요한 정보들을 놓치지 않고 잘 파악하여 엑소브레인을 포함하여 앞서 제시한 응용 등에서 도움을 받기 위함이다.

실제 태깅 결과에 대한 이해를 위하여 본 장에서 다루는 예시들은 생략어복원 태깅 결과물 포맷으로 어떻게 표현되는지를 함께 살펴본다. 3.1절에서 태깅 결과물 포맷을 소개하고 이전 장에서 정의한 생략어복원 대상인 생략어와 선행어를 중심으로 태깅 사례들을 3.2절과 3.3절에서 소개한다. 포커스에 대한 태깅은 추후에 정한다.

3.1. 태깅 결과물 포맷

생략어복원 태깅의 결과물은 엑소브레인 언어분석 말뭉치[17-18]와 같은 Json 포맷을 따르고 있다. 이 중에서 생략어복원에 해당하는 부분은 다음과 같다.

```
"ZA" : [
        {
          "id" : integer, // 0부터 시작함
          "type" : { "s" | "o" | "a" },
          "head_wid" : integer, // 0부터 시작함
          "ant_text": "string",
          "ant_sid" : integer, // 0부터 시작함
          "ant_wid" : integer, // 0부터 시작함
          "ant_is_title" : { 0 | 1 } // 0=No, 1=Yes
        }, …
       ]
```

3.2. 생략어 태깅

생략어 태깅의 전형적인 순서는 먼저 문장 내에서 지배소 후보들을 먼저 찾은 뒤, 각 지배소 후보에서 생략어 후보가 포함되어 있는지를 찾아보는 것이다. 구체적인 생략어 태깅 사례들을 살펴보면 다음과 같다.

가) 동사 표현 어구에서 필수 성분 중 생략된 것이 있다면 해당 성분을 생략어로 태깅한다.

[?는]¶ 인도양에 면해✱[VP] 있으며 북동쪽으로 소말리아, 북쪽으로 에티오피아와 남수단, 서쪽으로 우간다, 남쪽으로 탄자니아와 [?는]¶ 국경을 맞닿고[VP] 있다.

```
"ZA" : [
        {
          "id" : 0,
          "type" : "s", // 주어
          "head_wid" : 1, // 2번째 어절(면해)
          ...
        }, …
       ]
```

나) 동사 표현 어구에서 필수 성분 중 생략된 것이 여러 개가 있다면 각 성분을 모두 생략어로 태깅한다.

[?가]¶ [?를]¶ 출시한지✱[VP] 6개월이 지나 가격이 많이 떨어진 상태다.

```
"ZA" : [
        {
          "id" : 0,
          "type" : "s", // 주어
          "head_wid" : 0, // 1번째 어절(출시한지)
          ...
        },
        {
          "id" : 1,
          "type" : "o", // 목적어
          "head_wid" : 0, // 1번째 어절(출시한지)
          ...
        }
       ]
```

다) 여러 절로 구성된 문장에서 개별적인 절 단위에서 생략되어 있는 문장 성분이 있다면 생략어로 태깅한다.

대사의 오용과 남용을 강하게 성토한 그는 1517년 95개 논제를 게시함으로써✱[VP_AJT] 도미니코회 수사이자 대사령 설교 담당자인 요한 테첼에 [?는]¶ 맞섰다✱[VP].

```
"ZA" : [
        {
          "id" : 0,
          "type" : "s", // 주어
          "head_wid" : 17, // 18번째 어절(맞섰다)
          ...
        }
       ]
```

라) 주어 및 목적어 표현 어구에서 생략한 정보성이 있는 관형어를 생략어로 태깅한다.

[?의]¶ 수도는✱[NP_SBJ] 나이로비이며 [?의]¶ 공용어는✱[NP_SBJ] 영어와 스와힐리어이다.

```
"ZA" : [
        {
          "id" : 0,
          "type" : "a", // 관형어
          "head_wid" : 0, // 1번째 어절(수도는)
          ...
        }, …
       ]
```

3.3. 선행어 태깅

선행어 태깅은 앞서 제시한 선행어 후보를 태깅 작업

자가 정한 순서에 따라 후보들을 검토한 뒤, 해당 선행어를 복원하였을 때 필자가 의도했던 의미에 따라 명확해진 정보가 주어지는지 확인하는 과정을 거친다. 구체적인 선행어 태깅 사례들을 살펴보면 다음과 같다.

가) 해당 문서 표제어 또는 제목 내의 표현 일부가 선행어가 될 수 있다면 선행어로 태깅한다.

(표제어 : 케냐[T])
케냐 공화국 또는 케냐는 동아프리카의 공화국이다. [케냐[†]는]¶ 인도양에 면해[✱][VP] 있으며 북동쪽으로 소말리아, 북쪽으로 에티오피아와 남수단, 서쪽으로 우간다, 남쪽으로 탄자니아와 [케냐[†]는]¶ 국경을 맞닿고[✱][VP] 있다.

```
"ZA" : [
        {
          "id" : 0,
          "type" : "s", // 주어
          "head_wid" : 1, // 2번째 어절(면해)
          "ant_text": "케냐",
          "ant_sid" : -1, // 미 존재 문장
          "ant_wid" : -1, // 미 존재 어절
          "ant_is_title" : 1 // 표제어
        }, …
       ]
```

나) 해당 단락 및 문장 내에 포커스가 선행어가 될 수 있다면 선행어로 태깅한다.

문재인 대통령은 21일 '경제 사령탑'인 경제부총리 겸 기획재정부 장관 후보자에 김동연(60) 아주대 총장, 외교부 장관 후보자에 여성인 강경화(62) 유엔 사무총장 정책특보를 각각 내정했다. … 김동연 부총리 후보자[†]는 충북 음성 출신으로 '고졸신화의 인간승리 드라마'로 불린다. [김동연 부총리 후보자[†]는]¶ 덕수상고 졸업 뒤 은행에 취직해[✱][VP] 직장생활을 하며[✱][VP] 행정고시와 입법고시에 동시 합격한[✱][VP_MOD] 입지전적의 인물로 평가받는다[✱][VP].

```
"ZA" : [
        {
          "id" : 0,
          "type" : "s", // 주어
          "head_wid" : 4, // 5번째 어절(취직해)
          "ant_text": "후보자",
          "ant_sid" : 6, // 7번째 문장
          "ant_wid" : 2, // 3번째 어절(후보자는)
          "ant_is_title" : 0 // 표제어 아님
        }, …
       ]
```

다) 일반적으로 생략어 앞에서 나타난 멘션들 중 선행어를 찾아 태깅한다.

… 비텐베르크 대학교의 요한 스타우피츠 교수[T]는 루터가 성서에 대해 진지하게 공부하면 평안을 찾을 것이라고 생각하였다. 그래서 [요한 스타우피츠 교수[†]는]¶ 그를 성서학 교수사제로 임명하였는데[✱][VP], 스타우피츠 교수의 결정은 루터가 신앙적인 고민을 해결하는데 도움이 되었다.

```
"ZA" : [
        {
          "id" : 0,
          "type" : "s", // 주어
          "head_wid" : 4, // 5번째 어절(임명하였는데)
          "ant_text": "교수",
          "ant_sid" : 26, // 26번째 문장
          "ant_wid" : 4, // 5번째 어절(교수는)
          "ant_is_title" : 0 // 표제어 아님
        }
       ]
```

라) 중요한 개체를 나중에 말하는 화법 등에서는 생략어 뒤에서 나타난 멘션들 중에서도 선행어를 찾아 태깅한다.

제2차 세계 대전 당시 [이 영국 수학자[T]는]¶ 독일군 암호를 풀어[✱][VP] 전쟁을 승리로 이끌었으며[✱][VP] 컴퓨터의 원조인 자동 기계 이론을 개척했다[✱][VP]. 이 영국 수학자[†]는 과연 누구일까?

```
"ZA" : [
        {
          "id" : 0,
          "type" : "s", // 주어
          "head_wid" : 6, // 7번째 단어(풀어)
          "ant_text": "수학자",
          "ant_sid" : 1, // 2번째 문장
          "ant_wid" : 2, // 3번째 단어(수학자는)
          "ant_is_title" : 0 // 표제어 아님
        }, …
       ]
```

마) 암묵적으로 또는 상식적으로 알고 있거나 아직 알려지지 않은 개념이 선행어로 될 수 있다면 태깅한다.

케냐 공화국 또는 케냐는 동아프리카의 공화국이다. … 동아프리카에서 [어느 발굴가[†]에 의해]¶ 발견된[✱][VP_MOD] 화석에 따르면 조상이 2백만 년 전 이 지역에서 살았다고 한다.

```
"ZA" : [
        {
          "id" : 0,
          "type" : "s", // 주어
          "head_wid" : 1, // 2번째 어절(발견된)
          "ant_text": "[Unknown]", // 알 수 없음
          "ant_sid" : -1, // 미 존재 문장
          "ant_wid" : -1, // 미 존재 어절
          "ant_is_title" : 0 // 표제어 아님
        }
       ]
```

4. 생략어복원 말뭉치 구축 도구

엑소브레인 언어분석 말뭉치는 형태소분석, 어휘의미분석, 개체명인식, 구문분석, 의미역인식, 상호참조해결, 생략어복원에 대한 언어분석 정답을 제공한다. 현재까지 공개된 엑소브레인 언어분석 말뭉치는 언어분석 기술 개발을 위한 학습용으로는 그 양이 많지는 않으나, 동일 문장에 대해서 형태소분석부터 어휘의미분석, 개체명인식, 구문분석, 의미역인식, 상호참조해결, 생략어복원까지의 언어분석 정답을 포함하고 있기 때문에, 세부 언어분석 기술 뿐만 아니라 전체 언어분석 파이프라인을 평가하기 위한 용도로도 활용이 가능하다.[17-18]

엑소브레인 언어분석 말뭉치에 생략어복원을 태깅하기 위하여 반자동 태깅 도구를 가이드라인 수립과 함께 구축하였다. 엑소브레인 언어분석 결과를 고려하면서 생략어복원에 대한 태깅을 진행한다. 다만, 엑소브레인 언어분석이나 원문 상의 오류에 대해서는 작업자가 감안하고 작업해야 하며, 알려지지 않은 선행어에 대한 직접적인 유추는 하지 않고 [Unknown]으로 선행어 텍스트를 고정하여 태깅한다. 제공되는 주요한 기능은 다음과 같다.

- 작업 목록 및 작업 진행 상황 확인
- 작업 환경 설정 및 태깅 결과 시각화

- 각 문장 내 단어의 의존 관계 정보 확인
- 문서 내 존재하는 선행어 태깅
- 문장 내 존재하는 생략어 태깅

기존에는 작업자가 컴퓨터에서 툴을 이용하여 태깅한 것을 모아서 검토자에게 전달하여 일괄적으로 검토하는 방식이었다. 보다 효율적인 태깅 작업을 위하여 새롭게 구축한 방식은 서버에서 제공하는 태깅 대상 문서에 대해서 권한이 있는 작업자들과 검토자가 실시간으로 작업하고 함께 검토할 수 있는 환경이며, 태깅 작업자가 웹 브라우저에서 작업한 내용이 Json 포맷으로 서버에 실시간으로 저장되도록 설계하였다.

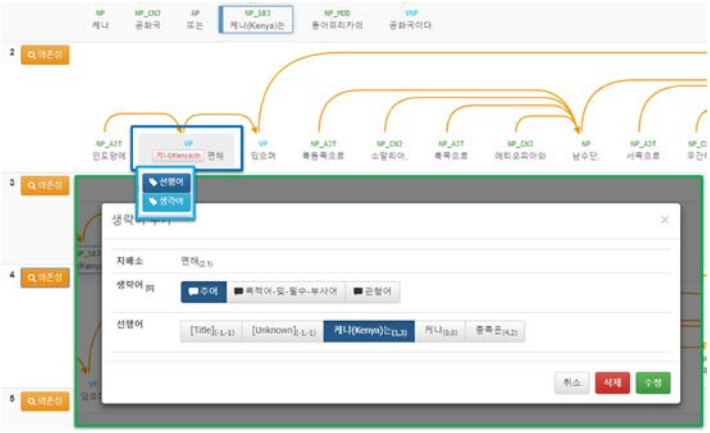

그림 1. 생략어복원 태깅 도구의 편집 화면

5. 결론

본 논문에서는 언어 사용에서 발생하는 생략된 정보를 명확하게 밝히기 위한 생략어복원 문제에 대해 한국어 생략어복원에 대해 정의하고 가이드라인을 제안하였다. 본 가이드라인을 통해서 한국어 생략어복원 말뭉치를 구축하는데 있어서 고려되는 지배소, 생략어, 선행어에 대한 대상을 정하였고, 태깅 포맷과 함께 실제 태깅 사례들을 살펴보았다. 우리는 본 가이드라인을 이용하여 종전에 개발된 한국어 생략어복원 시스템을 개선해 나가고 있으며, 새롭게 고려되는 개념과 방법론을 통하여 궁극적으로 엑소브레인 시스템의 품질이 보다 향상되기를 기대한다. 다만, 정보 추출 및 질의 응답 관점에서 도움이 되는 생략된 정보들을 찾는 데에 초점을 두었기 때문에 모든 생략된 정보들을 찾는 데에는 한계점이 존재한다. 향후, 본 연구에서 제안한 가이드라인을 통한 말뭉치 구축과정에서 발생할 수 있는 불분명한 기준 등에 대해서는 추가적인 개선이 필요할 것이며, 기술 개선과 함께 포커스에 대한 구체적인 사항도 단계적으로 가이드라인에 포함시킬 것이다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 일환으로 수행하였음. [2013-0-00131, (엑소브레인-1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발]

참고문헌

[1] G. Hirst, *Anaphora in Natural Language Understanding*. Springer Verlag, Germany, 1981.
[2] 황민국, 김영태, 나동열, 임수종, "무형대용어 해결 기술을 이용한 백과사전 표제어 복원," 제26회 한글 및 한국어 정보처리 학술대회 논문집, pp. 65–69, 2014.
[3] 황민국, 김영태, 나동열, 임수종, 김현기, "Structural SVM을 이용한 백과사전 문서 내 생략 문장성분 복원," 지능정보연구, vol. 21, no. 2, pp. 131–150, Jun. 2015.
[4] 임수종, 이창기, 장명길, "백과사전 질의응답을 위한 생략된 표제어 복원에 관한 연구," 한국정보과학회 학술발표논문집, vol. 32, no. 2, pp. 541–543, Nov. 2005.
[5] C. Yeh and Y. Chen, "Zero Anaphora Resolution in Chinese with Shallow Parsing," *Journal of Chinese Language and Computing*, vol. 17, no. 1, pp. 41–56, 2007.
[6] D. S. Wu and T. Liang, "Zero anaphora resolution by case-based reasoning and pattern conceptualization," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7544–7551, 2009.
[7] F. Kong and G. Zhou, "A Tree Kernel-based Unified Framework for Chinese Zero Anaphora Resolution," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 882–891, 2010.
[8] C. Chen and V. Ng, "Chinese Zero Pronoun Resolution with Deep Neural Networks," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 778–788, 2016.
[9] Y. Qingyu, Z. Weinan, Z. Yu, and L. Ting, "A Deep Neural Network for Chinese Zero Pronoun Resolution," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3322–3328, 2017.
[10] R. Iida, K. Inui, and Y. Matsumoto, "Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 625–632, 2006.
[11] R. Iida, K. Inui, and Y. Matsumoto, "Zero-Anaphora Resolution by Learning Rich Syntactic Pattern Features," *ACM Transactions on Asian Language Information Processing*, vol. 6, no. 4, 2007.
[12] R. Sasano, D. Kawahara, and S. Kurohashi, "A Fully-Lexicalized Probabilistic Model for Japanese Zero Anaphora Resolution," in *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 769–776, 2008.
[13] K. Imamura, K. Saito, and T. Izumi, "Discriminative Approach to Predicate-Argument Structure Analysis with Zero-Anaphora Resolution," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 85–88, 2009.
[14] R. Iida and M. Poesio, "A Cross-Lingual ILP Solution to Zero Anaphora Resolution," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 804–813, 2011.
[15] C. Park, K.-H. Choi, C. Lee, and S. Lim, "Korean Coreference Resolution with Guided Mention Pair Model Using the Deep Learning," *ETRI Journal*, vol. 38, no. 6, pp. 1207–1217, Dec. 2016.
[16] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky, "Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules," *Computational Linguistics*, vol. 39, no. 4, pp. 885–916, 2013.
[17] 임준호, 배용진, 김현기, 김윤정, 이규철, "의존 구문분석을 위한 한국어 의존관계 가이드라인 및 엑소브레인 언어분석 말뭉치," 제27회 한글 및 한국어 정보처리 학술대회 논문집, pp. 234–239, 2015.
[18] 임수종, 권민정, 김준수, 김현기, "ExoBrain을 위한 한국어 의미역

가이드라인 및 말뭉치 구축,” 제27회 한글 및 한국어 정보처리
학술대회 논문집, pp. 250–254, 2015.