담화에서의 어휘지도를 이용한 한국어 무형대용어 탐지 및 해결 말뭉치 생성

윤 호[O1], 남궁영[1], 박혁로[2], 김재훈[1]
한국해양대학교[1], 전남대학교[2]
4168615@naver.com, young_ng@kmou.ac.kr, hyukro@jnu.ac.kr, jhoon@kmou.ac.kr

Building a Korean Zero-Anaphora Detection and Resolution Corpus in Korean Discourse Using UWordMap

Ho Yoon[O1], Young Namgoong[1], Hyuk-Ro Park[2], Jae-Hoon Kim[1]
Korea Maritime and Ocean University[1], Chonnam National University[2]

요 약

담화에서 의미를 전달하는 데 문제가 없을 경우에는 문장성분을 생략하여 표현한다. 생략된 문장성분을 무형대용어(zero anaphora)라고 한다. 무형대용어를 복원하기 위해서는 무형대용어 탐지와 무형대용어 해결이 필요하다. 무형대용어 탐지란 문장 내에서 생략된 필수성분을 찾는 것이고, 무형대용어 해결이란 무형대용어에 알맞은 문장성분을 찾아내는 것이다. 본 논문에서는 담화에서의 무형대용어 탐지 및 해결을 위한 말뭉치 생성 방법을 제안한다. 먼저 기존의 세종 구어 말뭉치에서 어휘지도를 이용하여 무형대용어를 복원한다. 이를 위해 본 논문에서는 동형이의어 부착과 어휘지도를 이용해서 무형대용어를 복원하고 복원된 무형대용어에 대한 오류를 수정하고 그 선행어(antecedent)를 수동으로 결정함으로써 무형대용어 해결 말뭉치를 생성한다. 총 58,896 문장에서 126,720개의 무형대용어를 복원하였으며, 약 90%의 정확률을 보였다. 앞으로 심층학습 등의 방법을 활용하여 성능을 개선할 계획이다.

주제어: 무형대용어, 담화, 동형이의어 부착, 어휘지도

1. 서론

일반적으로 화자는 의미전달에 큰 문제가 없으면 문장성분을 생략한 채로 청자에게 발화한다. 예를 들어 그림 1에서 (1)로 표시된 곳에는 서술어 '하다'의 주어인 '너는'이 생략되었고 (2)의 위치에서는 서술어 '보다'의 주어인 '나는'이 생략되었다.

P1 : (1) 지하철에서 뭐 해?
P2 : (2) 신문을 봐

그림 1. 무형대용어의 예

이처럼 의미 전달에 문제가 없는 생략된 문장성분을 무형대용어(zero anaphora)라고 한다[1-3]. 한국어 담화에서는 맥락이나 상황에 따라서 문장성분의 생략이 빈번하므로 무형대용어가 자주 발생된다[1].

무형대용어의 탐지(detection)와 해결(resolution)은 자연언어를 이해하는 데 매우 중요한 과정이다. 특히 질의응답이나 대화시스템 등에서 정확한 정보의 이해와 전달을 위해 필요한 중요과정이다. 무형대용어 탐지란 문장 내에서 생략된 필수성분을 찾는 것이고, 무형대용어 해결이란 무형대용어에 알맞은 문장성분을 찾아내는 것이다. 한국어에서는 언어학적으로는 무형대용어에 대한 연구가 다소 진행되었으나[1], 한국어처리 분야에서는 무형대용어에 대한 연구가 매우 초보적인 단계이다[4-7]. 또한 무형대용어에 관한 말뭉치는 ETRI의 엑소브레인 한국어 언어분석 학습데이터[1]가 있으나 현재는 제한된 범위 내에서만 공개되고 있는 실정이다.

본 논문에서는 무형대용어 탐지 및 해결을 위한 말뭉치 생성 방법을 제안한다. 먼저 세종 구어 말뭉치[8]에서 무형대용어를 복원한다. 이를 위해 본 논문에서는 동형이의어 부착과 어휘지도를 이용해서 무형대용어를 복원하고 복원된 무형대용어에 대한 오류를 수정하고 그 선행어(antecedent)를 결정함으로써 무형대용어 해결 말뭉치를 생성한다. 무형대용어 복원 알고리즘은 현재 약 90%의 정확률을 보이며 앞으로 심층학습 등의 방법을 활용하여 성능을 개선할 계획이다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 소개하고, 3장에서는 어휘지도를 이용하여 기존의 세종 구어 말뭉치를 무형대용어 복원 말뭉치로 변환하는 방법을 기술한다. 4장에서는 변환된 말뭉치에 대한 분석을 수행한다. 마지막으로 5장에서는 결론 및 향후 연구 방향을 기술한다.

2. 관련 연구

무형대용어 탐지에 대한 연구는 미흡한 실정이다[4-7]. 무형대용어에 대한 연구는 주로 무형대용어 해결에 대해 초점이 맞춰져서 진행되었다. 무형대용어 해결과 비슷한 한국어 백과사전 문서에 생략된 표제어 복원

1) https://itec.etri.re.kr/itec/sub02/sub02_01_1.do?t_id=1210-2017-00482

에 대해서 격틀과 최대엔트로피(maximum entropy) 모델을 사용한 방법[4]이 연구되었고, Structual SVM 모델을 이용한 방법[5]을 통해 생략된 필수성분 복원 성능을 향상시켰다. 최근에는 무형대용어 해결의 연구에 심층학습 모델을 이용하고 있으며, 심층학습 모델 중 하나인 합성곱 신경망(Convolutional Neural Network, CNN)을 사용한 연구가 진행되었다[6]. 또한 단어표상과 1개의 은닉층을 이용한 신경망모델을 사용한 연구도 진행되었다[7].

3. 무형대용어 복원 및 말뭉치 생성

이 장에서는 무형대용어 복원을 방법을 기술한다. 본 논문에서는 세종 구어 말뭉치[8]를 대상으로 수행한다. 세종 구어 말뭉치는 기본적으로 형태소부착말뭉치이다. 그림 2는 세종 구어 말뭉치 예문이고, 그림 3은 복원된 말뭉치의 예문이다.

P1 : 지하철에서 뭐 해?
P2 : 신문을 보다가 버스로 갈아타.
P1 : 거기서 사람 많이 내리더라.
P2 : 학생들이 많이 내려.
P1 : 버스로 갈아탄 다음에 어디로 가?

그림 2. 기존 세종 구어 말뭉치

P1 : 지하철에서 뭐 [Noun가] 해?
P2 : 신문을 [Noun가] 보다가 버스로 [Noun가] 갈아타.
P1 : 거기서 사람 많이 내리더라.
P2 : 학생들이 많이 내려.
P1 : 버스로 [Noun가] 갈아탄 다음에 어디로 [Noun가] 가?

그림 3. 무형대용어 복원 말뭉치

그림 3에서의 'Noun가'는 생략된 주격이 복원되었음을 의미한다. 이와 같은 방법으로 필수격을 모두 복원한다.

그림 4는 세종 구어 말뭉치로부터 무형대용어의 복원 과정을 그림으로 표현한 것이며, 자세한 과정은 다음과 같다.

3.1 세종 구어 말뭉치 전처리

세종 구어 말뭉치는 다양한 구어 말뭉치를 수집하였는데 이 중 설교, 독백, 강연 유형의 말뭉치에는 주로 화자가 한 명만 등장한다. 따라서 이러한 말뭉치는 제외하고 일상대화, 전화, 토론 말뭉치를 중심으로 수집하여 재구축하였다. 이렇게 수집된 구어 말뭉치에 대한 정보는 표 1과 같다.

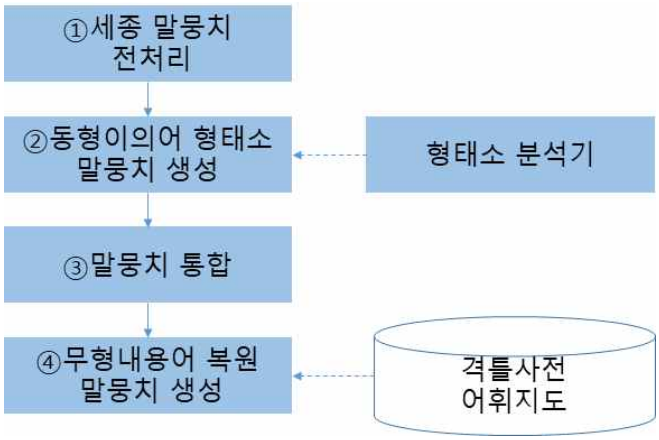

그림 4. 말뭉치 변환 과정

표 1. 세종 구어 말뭉치 정보

말뭉치	전체말뭉치	재구축된 말뭉치
문서수	155	98
발화수	73,723	65,125
문장수	187,185	124,886

또한 세종 구어 말뭉치에서 분석된 형태소의 결과에 어미가 모음만으로 표현된 것이 있어서 이를 정제하였다. 예를 들어 '좋아서'는 '좋/VA+아서/EC'로 분석되는 반면 '모아서'는 '모으/VV+ㅏ서/EC'로 분석된다. 이를 모두 '아서/EC'로 통일하였다.

3.2 동형이의어 형태소 말뭉치 생성

세종 구어 말뭉치의 원문을 Utagger 형태소 분석기[9]를 사용하여 모든 용언에 어깨번호를 부착한다. 즉 이를 통해 동형이의어 관계인 용언을 구분한다. 예를 들어 '버스로 갈아타'라는 문장을 분석하게 되면 '갈아타/VV__000000'와 같이 분석된다. 이때 '000000'은 어깨번호를 나타낸다. 이 과정에서 구묶음[10]을 수행하여 보조용언상당어구에 대해서는 무형대용어를 복원하지 않도록 한다. 예를 들어 "먹을 수 있다"와 같은 어구에서 "있다"에 대해서는 무형대용어를 복원하지 않는다.

3.3 말뭉치 통합

세종 구어 말뭉치는 형태소 분석 정확률이 높지만, 어깨번호가 부착되어 있지 않다. 반면에 3.2절에서 구축된 동형이의어 형태소 말뭉치는 형태소 분석 정확률이 낮을 수 있으나, 어깨번호가 부착되어 있다. 두 말뭉치의 형태소 분석 결과가 같다면 어깨번호가 부착된 말뭉치로 통합하고, 형태소 분석이 다르다면 세종 구어 말뭉치의 결과를 따르는 방법으로 통합한다. 통합과정에서 어깨번호가 필요할 경우에는 수동으로 추가하였다.

3.4 무형대용어 복원 말뭉치 생성

3.3절에서 통합된 말뭉치를 바탕으로 UPropBank 격틀사전[2]과 UWordMap 어휘지도[3]를 이용하여 무형대용어 복원 말뭉치를 생성한다. 그림 5는 무형대용어 복원 알고

리즘을 표현한 것이다. 그림 5의 알고리즘을 적용하여 각 용언의 필수성분을 확인하고 생략된 성분이 발견되면 그 성분을 복원하여 말뭉치를 생성한다. 그림 5에서 말하는 격틀 후보의 경우, 하나의 용언에 대해서 여러 가지의 격틀 후보가 존재한다. 예를 들어 그림 3에 쓰인 '갈아타다'와 같은 용언의 뜻은 '타고 가던 것에서 내려 다른 것으로 바꾸어 타다'라는 의미의 용언이다. 하지만 '갈아타다'의 격틀은 '{X:행동주 Y:착점-으로/로},{X:행동주 Y:대상-을/를}'로 2개가 존재한다. 그림 5에서 ⑴조사 정보, ⑵조사가 생략된 명사 정보는 다음과 같다.

⑴ 조사 정보

용언 앞에 존재하는 명사 중 조사가 붙어 있다면, 조사정보를 이용하여 격틀에서 해당 성분을 제거한다. 예를 들어 '버스를 갈아타다'와 같은 경우, 조사정보를 확인하여 '를'을 확인하고 격틀 후보 중에서 '를'이 들어가 있는 격틀 '{Y:대상-을/를}'을 제거한다. 위의 예를 따랐을 때, 격틀 후보의 결과는 '{X:행동주 Y:착점-으로/로},{X:행동주}'가 된다.

⑵ 조사가 생략된 명사 정보

용언 앞에 존재하는 명사 중 조사가 붙지 않은 명사들에 대해 어휘지도를 이용하여 격틀에서 제거한다. 예를 들어 '사과 먹어' 와 같은 경우 격틀 후보는 '{X:행동주 Y:대상-을/를}'이다. 명사 '사과' 에 조사 '를' 이 생략되었지만, 어휘지도를 이용하여 용언 '먹다' 와의 관계가 목적어이므로 격틀 후보에서 '{Y:대상-을/를}' 을 삭제한다. 따라서 이때 격틀 후보의 결과는 '{X:행동주}' 가 된다.

이후 격틀 후보 중 격틀 개수가 가장 적은 격틀을 반환하게 된다. 격틀 후보의 개수가 같을 경우 가장 먼저 위치해있는 격틀을 사용하여 우선 복원하고 수동으로 수정하는 과정에서 이를 교정할 것이다.

현재는 수동으로 이들을 교정하고 있으며 교정 과정에서 무형대용어 복원 오류뿐 아니라 선행어(antecedent)를 결정함으로써 무형대용어 탐지 및 해결 말뭉치를 구축할 것이다.

4. 무형대용어 복원 말뭉치의 분석

표 1의 재구축된 말뭉치에 대해서 3장의 변환 과정을 적용하였다. 총 124,886문장 중에서 58,896문장에 대해서 무형대용어 복원 말뭉치가 생성되었으며, 담화의 경우 용언이 쓰이지 않고 간단한 대답만으로 이루어지는 문장도 많아서 위와 같은 결과가 나왔다. 58,896문장에 대해서 생성된 무형대용어는 126,720개로 각 문장 당 2.15개의 무형대용어가 복원되었다.

실험적으로 100문장을 검사하여 무형대용어가 잘못 복원된 경우를 집계하였다. 100문장에서 243개의 무형대용어가 복원되었으며 이 중 잘못 복원된 무형대용어는 24

```
def To_Zero_Anaphora_Corpus(corpus, uprop, uword):
    # 어절을 형태소, 품사 단위로 분리
    morph, tag = split_eojeol(corpus)

    # UpropBank를 이용하여 격틀 후보 추출
    candidate_list = uprop(morph)

    # 후보 중 조사 정보, 조사가 생략된 명사 정보를
      이용하여 격틀 제거
    candidate_list = Delete_josa(morph, candidate_list)
    candidate_list = Delete_noun(morph,uword,
                     candidate_list)

    # 격틀 후보 중 격틀이 가장 적은 격틀을 반환
    index = find_min_index(candidate_list)
    return candidate_list[index]
```

그림 5. 무형대용어 복원 탐지 말뭉치 변환 알고리즘

개였다. 약 90.2%의 무형대용어가 올바르게 복원되었다. 잘못 복원된 무형대용어에 대해 살펴보면 표 2와 같고 세 가지 유형의 오류가 존재하였다.

표 2. 유형 별 오류 개수

유형	오류 개수
(1) 조사가 생략된 명사 정보	17
(2) 동형이의어 분석	4
(3) 대명사	3
합계	24

(1) 조사가 생략된 명사 정보에 관련된 오류

담화에서 등장하는 많은 명사들이 어휘지도에 등장하지 않아서 생기는 오류이다. 예를 들어 '육호선 탔어'와 같은 문장에서 어휘지도에 조사가 생략된 명사 '육호선'과 용언 '타다'의 결과가 존재하지 않아서 필수격이 추가되는 오류이다.

(2) 동형이의어 분석에 관련된 오류

담화에 등장하는 용언의 동형이의어 분석이 틀릴 경우 등장하는 오류이다. 예를 들어 "사람이 없잖아"라는 문장을 동형이의어 분석을 하였을 때, '없다__010101'로 분석되어 격틀이 '{X:대상}'이어야 했지만, 실제로 분석결과는 '없다__010201'로 나타나서 격틀이 '{X:대상 Z:처소-에}'로 분석되어 필수격이 추가되는 오류이다.

(3) 대명사와 관련된 오류

담화에 등장하는 대명사와 용언의 어휘지도 결과가 존재하지 않아서 생기는 오류이다. 예를 들어 '나는 집에 가' 라는 문장에서 '나'라는 대명사와 '가다'라는 용언의 어휘지도 결과가 존재하지 않아서 필수격이 추가되는 오류이다.

2) ftp://203.250.77.242/UPropBank.zip

3) http://nlplab.ulsan.ac.kr/doku.php?id=uwordmap

5. 결론 및 향후 연구

본 논문에서는 무형대용어 탐지 및 해결을 위한 말뭉치 생성 방법을 제안하였다 먼저 세종 구어 말뭉치에서 무형대용어를 복원했다. 이를 위해 본 논문에서는 동형이의어 부착과 어휘지도를 이용해서 무형대용어를 복원하고 복원된 무형대용어에 대한 오류를 수정하고 그 선행어를 수동으로 결정함으로써 무형대용어 해결 말뭉치를 생성했다. 또한 세종 구어 말뭉치를 무형대용어 복원 말뭉치로 변환하는 알고리즘을 기술하였다. 세종 구어 말뭉치를 동형이의어 형태소 말뭉치와 비교하여 어깨번호를 추가하고 어휘지도와 격틀사전을 이용하여 생략된 약 126,720개의 무형대용어를 복원하였다. 오류 분석을 통해 약 90%의 무형대용어가 올바르게 복원되었고 어휘지도에 결과가 존재하지 않을 때 주격이 추가되는 오류가 주로 등장하였다. 현재는 수동으로 이들을 교정하고 있으며 교정 과정에서 무형대용어 복원 오류뿐 아니라 선행어를 결정함으로써 무형대용어 탐지 및 해결 말뭉치를 구축할 것이다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017M3C4A7068187, 한국어 정보처리 원천 기술 연구 개발).

참고문헌

[1] N.-R. Han, Korean Zero Pronoun Analysis and Resolution, Ph.D Dissertation, Department of Linguistics, University of Pennsylvania, 2006.

[2] S. Nariyama, “Grammar for ellipsis resolution in Japanese”, Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 135-145 2002.

[3] 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, pp. 1-22, 2007.

[4] S. Lim, C. Lee and M. Jang, “Restoring an Elided Entry Word in a Sentence for Encyclopedia QA System”, Proceedings of 2nd International Joint Conference on Natural Language Processing, pp. 215-219, 2005.

[5] 황민국, 김영태, 나동열, 임수종, 김현기, “Structural SVM을 이용한 백과사전 문서 내 생략 문장성분 복원”, 지능정보연구, 제21권, 제2호, pp. 131-150, 2015.

[6] 김영태, 백지엽, 김민형, 나동열, 임수종, “Convolutional Neural Network 기반 무형대용어 해결 기법”, 한국지능정보시스템학회 학술대회논문집, pp. 22-23, 2018.

[7] 김영태, 백지엽, 김민형, 나동열, 임수종, “신경망 모델을 이용한 무형대용어 해결 기법”, 한국정보과학회 학술발표논문집, pp. 611-613, 2018.

[8] 김흥규, 강범모, 홍정하, "21세기 세종계획 현대국어 기초말뭉치: 성과와 전망", 한국정보과학회 언어공학연구회 학술발표 논문집, pp. 311-316, 2007.

[9] 신준철, 옥철영, “기분석 부분 어절 사전을 활용한 한국어 형태소 분석기”, 정보과학회논문지 : 소프트웨어 및 응용, 제39권, 제5호, pp. 415-424, 2012.

[10] 남궁영, 김창현, 천민아, 박호민, 윤호, 최민석, 김재균, 김재훈, 한국어 말덩이 정의와 구묶음: Bi-LSTM/CRFs 모델을 이용하여, 정보과학회논문지, vol. 47, no. 6, pp. 587~595, 2020.