격틀사전을 활용한 존댓말 교정 프로그램

장윤정, 임리나, 이재성
충북대학교 소프트웨어학과

e-mail : yjjang@chungbuk.ac.kr

Honorific Correction Program Using Case Frame Dictionary

Yunjeong Jang, Lina Lim, Jae Sung Lee
Dept. of Computer Science, Chungbuk National University

요 약

한국어 맞춤법 검사기는 현재 여러 곳에서 개발하여 서비스를 제공하고 있으며 다양한 목적으로 사용되고 있다. 하지만 이들은 맞춤법이나 띄어쓰기를 검사할 뿐, 존댓말과 같은 높임표현이 잘못 사용되었는지는 검사할 수 없다. 최근에는 맞춤법 검사를 위해 통계적 방식과 딥러닝을 사용하고 있지만, 본 연구에서는 규칙 기반을 활용하여 사용자가 텍스트를 입력하면 잘못된 높임표현을 탐지하여 그에 대한 오류 정보를 제공하고, 올바른 표현으로 교정하는 프로그램을 개발하였다. 구문 분석기를 사용하여 주어-서술어 구조를 파악하고, 형태소 분석기를 활용하여 높임표현을 탐지한 후 격틀사전의 정보를 이용하여 존댓말 오류를 탐지한다. 본 연구는 격식이 있는 문서를 검사할 때나, 한국어 높임말에 익숙하지 않은 외국인들에게 도움이 될 수 있을 것이다.

1. 서론

대화를 하거나 글을 작성할 때, 올바른 맞춤법을 사용하는 것은 중요하게 여겨지며 맞춤법 오류가 있는지 확인하기 위해 맞춤법 검사기를 활용한다[1,2,3,4]. 하지만 기존의 맞춤법 검사기들은 철자오류나 띄어쓰기를 검사할 뿐, 존댓말과 같은 높임표현이 잘못 사용되었는지에 대해서는 검사하지 않는다. 국립국어원 홈페이지[5]에서 국어지식 상담 건수는 높임법을 검색한 결과 1,130건, 높임말 426건, 높임표현 871건, 존댓말 111건(2019년 9월 7일 기준)으로 올바른 높임표현 사용에 대해 궁금해 하는 사람들이 많음을 알 수 있다. 본 논문에서는 실생활에 잘못 쓰이고 있는 높임말에 대해 올바른 표현을 알 수 있도록 사용자가 텍스트를 입력하면 잘못된 높임표현을 탐지하여 오류 정보를 제공하고, 올바른 표현을 추천하여 교정해주는 프로그램을 제공한다. 이는 격식을 갖추어야 할 문서를 검사할 때나, 한국어 높임말에 익숙하지 않은 외국인들에게 도움이 될 수 있다.

본 연구는 존칭 선어말 어미에 대한 주체 높임법에 대해서만 탐지하고 교정하며, 단문에서만 적용된다. 또한 보편적으로 인정되는 높임표현만을 적용하였으며, 국립국어원 국어지식 상담 답변을 참고하였다. 철자 오류의 교정 방법에는 딥러닝이나 통계 정보를 사용하는 방법이 보편적이지만[6], 본 연구에서는 규칙 기반을 활용한다. 오류 결정 규칙은 검사 단어와 같이 사용될 수 없는 어휘, 품사, 범주화된 사전정보 등으로 구성된다[7].

2. 제안 시스템

2.1 활용 요소

본 프로그램은 높임말 탐지 및 오류 판단, 문장 교체를 위해 다음과 같은 세 가지 요소를 활용한다.

(1) 의존 구문 분석 (Dependency parsing)

문장의 구조를 파악하기 위하여 Google Cloud Natural Language API를 사용한다[8]. 문장이 입력되면 품사 단위로 품사 종류, 각 어절 간의 의존 구조, 의존 관계 레이블 레벨을 추출한다. 입력된 문장에서 서술어는 <root>로 태깅되고, <root>의 주어는 <nsubj>로 태깅된다. 이러한 <root>와 <nsubj>간의 관계를 파악하여 해당 용언이 올바르게 사용되었는지 판단한다.

(2) 형태소 분석

한국어 정보처리를 위한 파이썬 패키지 KoNLPy에서 제공하는 꼬꼬마(KKMA) 형태소 분석기와 코모란(Komoran) 형태소 분석기를 사용하여 형태소 분석을 한다[9]. 형태소 분석이란, 문장이 입력되면 형태소 단위로 문장을 나눈 후, 해당하는 형태소가 문장 내에서 어떤 범주에 있는 지를 태깅하는 것을 말한다. 예를 들면, '손님, 주문하신 커피 나왔습니다.'라는 문장이 입력되었을 때, 꼬꼬마 형태소 분석기 기준으로 '손님'과 '커피' 모두 일반명사이므로 (NNG)로 태그되고, '나오(다)'는 동사이므로 (VV)가 태그된다. 또한 어떤 형태소 분석기를 사용하는지

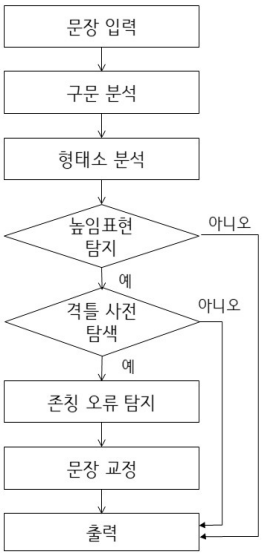
에 따라 형태소 태그가 다르므로 Korean POS tags comparison chart[10]을 참고 하였다.

(3) 격틀 사전

격틀에 대한 정보를 모든 서술어(혹은 서술어 역할을 할 수 있는 서술성 명사)에 대해서 모아 놓은 사전을 격틀 사전이라 한다[11]. 본 프로그램은 국립국어원 21세기 세종계획 최종 평가물로 배포된 격틀사전[12]를 사용하였다. 격틀사전 중에 명사 사전, 서술어 사전, 범주 사전을 사용하였다. 명사 사전은 해당 명사가 어떤 범주에 속하는지 정의 되어 있고, 서술어 사전은 해당 서술어에 어떤 주어가 올 수 있는 지 정의되어 있다. 범주 사전은 하위 범주들에 대한 상위 범주가 정의되어 있다. 격틀 사전은 MongoDB에 저장하여 사용한다.

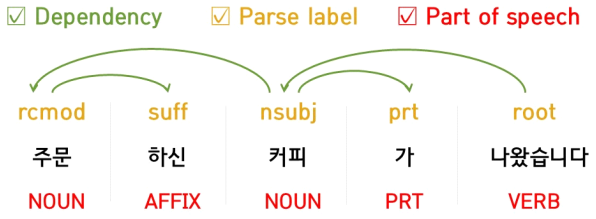
2.2 존댓말 교정 알고리즘

본 프로그램에서는 잘못 쓰이는 높임말을 교정하기 위해 (그림 1)의 과정을 거친다.



(그림 1) 시스템 동작을 나타내는 흐름도

과정은 크게 구문 분석, 형태소 분석, 높임표현 탐지, 격틀사전 탐색, 존댓말 오류 탐지, 문장 교정으로 이루어진다. 일반적으로는 형태소 분석 후 구문 분석을 하지만 본 프로그램에서는 구문 분석기 API[8]을 사용하기 때문에, 구문 분석기로 주어와 용언을 탐지 후, 형태소 분석을 하여 다음 과정을 진행한다.



(그림 2) 구문 분석 결과와 의존 구조 그래프

구문 분석 과정에서는 입력된 문장의 주어와 서술어를 찾고, 관계를 파악한다. <root>로 태그 된 품사는 해당 문장의 서술어이고, <root>와 의존성을 가지는 <nsubj>는 해당 서술어와 관련된 주어이다. 예를 들어, '주문하신 커피가 나왔습니다.'라는 문장에서 '나왔습니다'는 서술어이고, '커피'는 주어이다(그림 2).

형태소 분석 단계에서는 꼬꼬마 형태소 분석기를 사용하여 처리하며, 구문 분석 결과 중 필요한 부분을 다시 형태소 분석하여 높임표현 탐지나 원형 복원에 활용할 수 있도록 한다.

높임표현 탐지 단계에서는 분석된 형태소 태그들을 활용하여 높임표현의 유무를 탐지한다. 입력된 문장에 존칭 선어말 어미 태그(EPH 혹은 EP)가 존재하면 존댓말로 판단한다. 존댓말로 판단되면 상태 플래그를 1로 변환하고 존댓말이 아니라고 판단되면 상태 플래그를 -1로 변환 후 용언 원형 복원단계로 들어간다. 용언 원형 복원은 격틀 사전 검색을 위해 필요하여 이를 통해 존댓말 오류를 탐지할 수 있으므로 매우 중요하다. 자세한 용언 원형 복원 방법은 2.3에서 기술한다.

존댓말오류를 탐지하기 위해, 용언 원형 복원의 결과로 원형이 된 용언을 격틀 사전에서 검색 후 해당 용언이 사용 될 수 있는 상위 범주를 파악한다. '커피가 나오셨습니다.'에서 서술어는 '나오다'이고 주어는 '커피'이다. <nsubj>로 태깅된 주어 '커피'는 명사 사전을 통해 범주를 파악한다. '커피'의 범주는 [음식, 음료]이다. 범주 사전에 정의 된 '음식'의 상위 범주는 [구체인공물, 구체물, 온]으로 인간이 포함되어 있지 않고, '음료'의 상위 범주도 [음식, 구체인공물, 구체물, 온]으로 인간이 포함되어 있지 않으므로, 높임표현이 불가한 주어이다. 서술어인 '나오다'는 서술어 사전에 ['전기전자기구', '기체', '화면', '말', '인간', '가스', '상품', '뉴스', '행위']등이 주어로 올 수 있다고 정의되어 있다. '나오다'는 '인간'이 주어로 올 수 있으므로 높임표현이 가능한 용언(서술어)이지만, 주어인 '커피'는 인간의 범주에 들어있지 않으므로 높임표현을 쓸 수 없다. 따라서 해당 문장은 존댓말 오류 문장으로 판단한다.

마지막 단계인 문장 교정은 올바르지 못한 문장이라고 판단된 문장을 수정하고 잘못된 부분을 화면에 표시하는 과정이다. 문장 교체는 용언 원형 복원기 알고리즘과 유사하게 작동된다. 하지만 용언 원형 복원기는 시제 어미, 존칭 선어말 어미 등의 모든 어미를 제거하고 남은 어간으로 원형을 복원하는데 반해, 문장 교체기는 존칭 선어말 어미만을 제거하여 높임표현에서 낮춤표현으로 교체한다.

2.3 용언 원형 복원

용언 원형 복원은 높임표현이라고 판단되는 용언을 격틀 사전에서 검색할 수 있는 형태인 용원의 원형으로 복원하는 것을 말한다. 하지만 꼬꼬마 형태소 분석기만을 사용하였을 때 두 가지 문제점이 나타나는 것을 발견했다. 첫 번째 문제점은 존칭 선어말 어미 태그가 붙어야 하는

형태소를 동사 자체로 판단하는 경우가 있는 것이다. 예를 들면, '가다'의 높임표현인 '가시다'는 '가+시+다'로 태그 되어야 하지만, '가시+다'로 태그 되는 문제이다. 두 번째 문제점은 용언의 어간을 잘못 판단하여 격틀사전에 없는 용언을 생성하는 것이다.

이 두 가지의 문제점을 해결하기 위해 꼬꼬마 형태소 분석 결과로 존댓말이 아니라고 판단되는 문장(상태 플래그가 -1이었던 문장)을 코모란 형태소 분석기를 통해 재검사 한다. 재검사를 하는 경우는 원형으로 복원된 용언에 '시'라는 글자가 존재하는 경우와 격틀 사전에 없는 용언으로 복원 되는 경우이다. 코모란 형태소 분석기는 전반적으로 꼬꼬마 형태소 분석기보다 성능이 좋지는 않았지만 꼬꼬마 형태소 분석기에서 잘못 판단하는 문장을 올바르게 판단하기 때문에 이차적으로 사용하게 되었다. 해당 성능평가는 3.2에서 자세하게 다루었다. 만약 코모란 형태소 분석기를 사용 한 후에도 존댓말로 판단되지 않으면 문장에 오류가 없는 것으로 판단하고 '올바른 표현입니다.'를 화면에 출력한다.

다음 규칙1, 2, 3은 용언 원형 복원 시 어간의 형태소 규칙을 나타낸 것이다. 규칙 1, 2, 3에 따라 어간을 찾아 원형으로 복원한다. 해당 규칙에서 V는 VV(동사), VA(형용사)와 같은 용언들을 모두 포함한 것을, N은 NNG(일반명사), NNP(고유명사) 등의 모든 명사들을 포함한 것을 말한다. 꼬꼬마 형태소 분석기를 사용한 경우와 코모란 형태소 분석기를 사용한 경우의 알고리즘이 거의 유사하기 때문에 본 논문에서는 꼬꼬마 형태소 분석기를 사용하여 용언 원형 복원을 했을 경우만 나타내었다.

규칙 1) 용언(V)+'다' 경우

규칙 1은 용언의 어간이 동사나 형용사 단일인 형태의 경우를 나타낸다. 규칙 1의 경우에는 <표 1>과 같이 VV나 VA가 태깅된 형태소를 찾고 해당하는 형태소에 '다'를 더하여 용언의 원형으로 복원한다.

<표 1> 용언 규칙 예시

* 동사 (VV + '다')
▶ 커피가 나오셨습니다. -> 나오다. [('커피', 'NNG'), ('가', 'JKS'), **('나오', 'VV')**, ('시', 'EPH'), ('었', 'EPT'), ('습니다', 'EFN'), ('.', 'SF')]
* 형용사 (VA + '다')
▶ 아버지께서 예쁘십니다. -> 예쁘다 [('어머니', 'NNG'), ('께서', 'JKS'), **('예쁘', 'VA')**, ('시', 'EPH'), ('ㅂ니다', 'EFN'), ('.', 'SF')]

규칙 2) 명사(N) +동사파생 접미사(XSV) + '다' 경우

용언의 어간에 파생명사가 오는 경우가 있다. 규칙 2는 그러한 경우를 정의했다. <표 2>는 규칙 2의 예시이다. 단일 명사나 복합명사일 경우 N태그와 XSV태그를 탐지하여 이를 용언의 원형으로 복원한다. '-화하다'의 경우에는 N과 XSN(명사 파생 접미사)와 XSV를 함께 용언의 원형으로 복원한다.

<표 2> 명사 규칙 예시

* 단일 명사 (N+XSV+'다')
▶ 삼촌은 그 일을 마감하실 것이다. -> 마감하다 [- **('마감', 'NNG'), ('하', 'XSV')**, ('시', 'EPH'), ('ㄹ', 'ETD'), ('것', 'NNB'), ('이', 'VCP'), ('ㅂ니다', 'EFN'), ('.', 'SF')]
* 복합 명사 (N+N+XSV+'다')
▶ 아버지께서 프로그램을 원상복구하십니다. -> 원상복구하다 [- **('원상', 'NNG'), ('복구', 'NNG'), ('하', 'XSV')**, ('시', 'EPH'), ('ㅂ니다', 'EFN'), ('.', 'SF')]
* -화하다 (N+'XSN'+'XSV'+'다')
▶ 아버지께서 프로그램을 문서화하셨습니다. -> 문서화하다 [- **('문서', 'NNG'), ('화', 'XSN'), ('하', 'XSV')**, ('시', 'EP'), ('었', 'EP'), ('습니다', 'EF'), ('.', 'SF')]

규칙 3) 기타 경우

규칙 3은 규칙1, 2의 경우에 포함되지 않는 기타 경우를 나타낸다. <표 3>은 규칙3에 대한 예시이다.

<표 3> 기타 예시

* N+VV+ECE+VV +'다'
▶ 삼촌은 그 일에 발 벗고 나섰습니다. -> 발벗고나서다 [- **('발', 'NNG'), ('벗', 'VV'), ('고', 'ECE'), ('나서', 'VV')**, ('었', 'EPT'), ('다', 'EFN'), ('.', 'SF')]

3. 실험 및 성능평가

3.1 실험

실험 데이터는 격틀 사전에 나오는 서술어에 대하여 각각 높임표현의 활용 예제 450문장을 만들어 사용하였다. <표 4>는 사용자로부터 입력된 문장이 문장 교체기를 통해 어떻게 출력될 것인지를 보여주는 예제이다. 사용자가 ①번 '손님, 주문하신 커피가 나오셨습니다.' 라는 문장을 입력하면 커피에게 높임표현을 사용한 것이므로 잘못된 문장이 된다. 그러므로 '나오셨습니다'는 '나왔습니다'로 교정된다. 이와 마찬가지로 ②번도 사물을 높이는 표현을 사용했으므로 '예쁘시네요'는 '예쁘네요'로 바뀐다. 반면에 ③번 '어머니께서 오십니다.'는 올바른 문장이므로 '교정할 문장이 없습니다.'라고 출력된다.

<표 4> 문장 교체기 적용 예제

※ 입력문장 -> 출력문장
① 손님, 주문하신 커피가 나오셨습니다. -> 손님, 주문하신 커피가 나왔습니다.
② 신발이 예쁘시네요. -> 신발이 예쁘네요.
③ 어머니께서 오십니다. -> '교정할 문장이 없습니다.'

3.2 성능 평가

테스트 문장은 격틀사전에 있는 서술어에 대해 네이버 국어사전[13]의 뜻을 참고하여 만들었다. 평가 데이터는 총 1680문장으로, 올바른 문장(존댓말 오류가 없는 문장) 1260문장, 존댓말 오류가 포함된 문장 420문장으로 구성하였다. 평가 측정 방법은 아래의 정확률(Precision) 식 (1)과 재현율(recall) 식 (2)를 사용하였다.

$$\text{정확률} = \frac{\text{올바르게 존댓말 오류로 찾은 문장 수}}{\text{존댓말 오류로 찾은 문장 전체 수}} \quad (1)$$

$$\text{재현율} = \frac{\text{올바르게 존댓말 오류로 찾은 문장 수}}{\text{존칭 오류가 있는 전체 문장 수}} \quad (2)$$

<표 5> 존댓말 오류 평가 결과

	정확률	재현율	F1-Score
꼬꼬마	83.2%	89.5%	86.2%
코모란	76.7%	87.9%	81.9%
꼬꼬마 + 코모란	89.1%	93.1%	91.1%

<표 5>은 꼬꼬마 형태소 분석기와 코모란 형태소 분석기를 각각 사용하였을 때와 꼬꼬마 형태소 분석기와 코모란 형태소 분석기를 함께 사용하였을 때 존댓말 오류를 탐지한 결과이다. 꼬꼬마 형태소 분석기를 사용하여 오류 탐지를 했을 경우 정확률은 83.2%, 재현율은 89.5%이며, F1-Score를 계산한 결과 86.2%이다. 코모란 형태소 분석기를 사용하여 오류 탐지를 했을 경우 정확률은 76.7%, 재현율은 87.9%이며, F1-Score를 계산한 결과는 81.9%이다.

꼬꼬마 형태소 분석기와 코모란 형태소 분석기를 함께 사용하여 오류 탐지를 했을 경우 잘못된 문장에서는 총 420개 중 391개의 문장에서 존댓말 오류를 탐지하여 93.1%의 성공률을 보였으며, 올바른 문장의 경우 총 1260개 중 48개의 문장을 잘못 탐지하여 3.8% 잘못된 탐지를 하였다. 결과적으로 정확률은 89.1%, 재현율은 93.1%이다. F1-Score를 계산한 결과는 91.1%이며, 이는 형태소 분석기를 단독으로 사용한 것보다 높은 성능을 보였다.

4. 결론 및 향후 연구

본 논문에서는 문장에 잘못 쓰인 존댓말을 교정하는 방법을 제안하여, 사용자들이 잘못된 높임표현을 알고 그에 대해 올바르게 교정 할 수 있도록 하는 것에 의의를 둔다. 형태소 분석을 활용하여 높임표현을 탐지하고, 의존 구문 분석을 활용하여 주어와 서술어를 추출한다. 추출한 주어와 서술어를 격틀 사전을 통해 관계를 파악하여 문장 내 오류를 탐지한다. 탐지된 오류는 형태소 분석 결과를 이용하여 옳은 문장으로 교체한다. 이 프로그램은 높임표현에 익숙하지 않은 사람들을 대상으로 올바른 주체 높임법을 사용할 수 있도록 도울 수 있다.

향후에는 문장교체기에서 사용하는 기존의 규칙 기반 기법에서 딥러닝을 활용한 기법으로 교체하는 것을 계획하고 있다.

감사의 글

"이 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 사업의 연구결과로 수행되었음" (2019-0-01183)"

참고문헌

[1] 네이버맞춤법검사기, 2018.09, https://www.naver.com/
[2] 부산대 한국어 맞춤법/문법 검사기, 2019.07.08., https://speller.cs.pusan.ac.kr/
[3] 다음 맞춤법 검사기, 2018.09, https://dic.daum.net/grammar_checker.do
[4] saramin 글자수세기/맞춤법검사기, 2019.04, http://www.saramin.co.kr/zf_user/tools/character-counter
[5] 국립국어원, https://www.korean.go.kr/
[6] 김민호, 최성기, 권혁철, "어절 N-gram을 이용한 문맥 의존 철자 오류 교정", 한국정보과학회 학술발표논문집, 580-582, 2014
[7] 소길자, 권혁철, "선택 제약 명사의 의미 범주 정보를 이용한 용언의 문맥 의존 오류 검사 및 교정", 한국정보통신학회논문지/18(1), 2014, 25-31, 한국정보통신학회
[8] Google Cloud Natural Language API, https://cloud.google.com/natural-language/
[9] 박은정, 조성준, "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지", 제 26회 한글 및 한국어 정보처리 학술대회 논문집, 2014
[10] Korean POS tags comparison chart, https://docs.google.com/spreadsheets/d/1OGAjUvalBuX-oZvZ_-9tEfYD2gQe7hTGsgUpiiBSXI8/edit#gid=0
[11] 신명철, "한국어 부사격의 의미역 결정을 위한 격틀 사전과 기계학습법의 통합", 포항공과대학교, 2006
[12] 국립국어원, 격틀사전-21세기 세종계획 최종성과물(수정판), 2011
[13] 네이버 국어사전 Beta, 2018.10, https://ko.dict.naver.com/#/main