외국인 학습자를 위한 문맥 기반 실시간 국어 문장 교정

박영근○, 최재성, 김재민, 이성동, 이현아
금오공과대학교, 컴퓨터 소프트웨어공학과
owo2323@naver.com, bload4712@nate.com, richjam9284@naver.com
xkdnrl@naver.com, halee@kumoh.ac.kr

Context Based Real-time Korean Writing Correcting for Foriengers

[Young-Keun Park○, Jae-Sung Choi, Jae-Min Kim, Seong-Dong Lee, Hyun-Ah Lee]
Dept. of Computer Sofrware Engineering, Kumoh National Institute of Technology

요 약

외국인 유학생과 국내 체류 외국인을 포함하여 한국어를 학습하고자 하는 외국인이 지속적으로 증가함에 따라, 외국인 한국어 학습자의 교육에 대한 관심도 높아지고 있다. 기존 맞춤법 검사기는 한국어를 충분히 이해할 수 있는 한국인의 사용에 중점을 두고 있어, 외국인 한국어 학습자가 사용하기에는 다소 부적절하다. 본 논문에서는 한국어의 문맥 특성과 외국인의 작문 특성을 반영한 한국어 교정 방식을 제안한다. 제안하는 시스템에서는 말뭉치에서 추출한 어절 바이그램에 대한 음절 역색인을 구성하여 추천 표현을 빠르게 제시할 수 있으며, 키보드 후킹에 기반한 사용자인터페이스를 제공하여 사용자 편의를 높인다.

주제어: 한국어 교정, 유사도 계산, 어절 바이그램, 음절 역색인

1. 서론

외국인 유학생이 10만 명, 국내 체류 외국인이 200만 명에 이르는 등 국내 체류 외국인이 늘어남에 따라 한국어 학습에 관심을 가지는 외국인의 숫자가 크게 증가하고 있다. 교육부의 2016년 교육기본통계에 따르면 외국인 대학생이 총 10만 4262명, 재학 대학이 236개 대학에 이르는데 반해, 언어 능력(영어 또는 한국어) 기준을 통과한 학생이 절반을 넘는 곳은 24.6% 밖에 되지 않았다[1]. 또한 서울시 교육청의 '다문화학생 교육지원 기본계획' 에 따르면 다문화 학생 1만 1642명에 비해 다문화 교사는 87명에 불과하여, 부족한 교사 수로 인해 한국어 교육이 제대로 이루어지지 못 하고 있다[2]. 한국어 학습자들의 한국어 구사 능력 향상을 도울 수 있는 시스템이 다방면으로 필요한 시점이다.

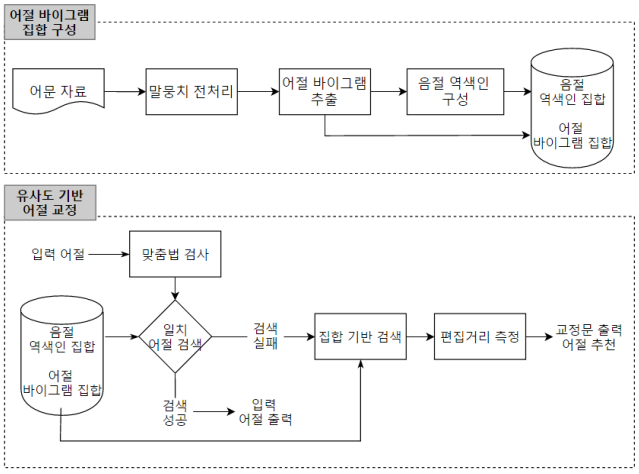
외국인의 한국어 작문 오류는 접속사 및 조사 오류, 문장 성분 누락, 부적절한 어절간 호응, 음운 차이로 인한 오류 등을 주로 포함하여, 한국인들의 일반적인 작문 오류와는 다른 특성을 나타낸다[3,4]. 기존의 한국어 작문을 돕기 위한 맞춤법 검사 시스템들[5~7]은 명백히 호응 관계가 잘못된 문장도 단일 어절 내에서의 맞춤법이 정확하다면 잡아주지 못 하는 경우가 많아, 외국인의 작문 교정이 적합하지 않은 측면이 있다.

본 논문에서는 한국어 학습자 대상의 한국어 문장 교정 시스템을 제안한다. 시스템에서는 미리 구축한 어절 바이그램 집합에서 사용자 입력과 가장 유사한 표현을 검색하여 입력 문장에 대한 추천 교정을 제시한다. 시스템에서는 어떤 작문 환경에서도 추천 교정을 신속히 얻을 수 있도록 키보드 후킹 기반 백그라운드 프로세스로 교정 시스템을 구축하는 동시에, 어절 바이그램과 음절 단위 역색인 구조, 빠른 속도의 유사도 계산 방식을 채용하여 실시간으로 교정을 제시하고자 한다.

2. 본론

2.1 시스템 구조

아래 [그림 1]은 외국인 대상 한국어 문장 교정 시스템의 구조도이다. 시스템은 어절 바이그램 집합 구성과 유사도 기반 어절 교정의 두 부분으로 구성된다. 어절 바이그램 집합 구성에서는 잘 작문된 한국어 말뭉치로부터 어절 바이그램 집합을 추출하고, 유사 어절의 빠른 검색을 위해 역색인을 음절 단위로 구성한다. 유사도 기반 어절 교정에서는 추출된 바이그램 집합 중에서 입력된 어절과 가장 유사한 어절을 추천하기 위해, 집합 기반 검색과 개선된 편집거리를 사용한다. 시스템은 효율적인 역색인을 사용하여 짧은 시간 안에 교정 문장을 추천하며, 어절 바이그램과 음소 기반 편집거리를 통한 문장 교정으로 외국인의 작문 특성을 반영한 교정이 가능하다. 아래에서는 각 과정을 상세히 설명한다.



[그림 1] 시스템 구조도

2.2 어절 바이그램 집합 구성

제안하는 시스템에서는 추천 문장을 제시하기 위해 말

뭉치의 어절 바이그램을 이용한다. 어절 바이그램 집합 구성은 말뭉치 전처리와 어절 바이그램 추출, 음절 역색인 구성의 단계로 구성된다.

어절 바이그램에서 숫자와 고유명사는 불필요하게 추출 자료의 크기를 증대시킬 수 있다. 시스템에서는 숫자와 고유명사를 미리 정의된 특수문자로 대체하여 자료 크기의 문제와 자료 희소성의 문제를 동시에 해결하고자 한다. 예를 들어 "2016년도 중국발 황사가 기승을 부리고 있다."는 전처리를 거쳐 "#년도 @발 황사가 기승을 부리고 있다."로 변환시킨다.

전처리된 말뭉치로부터 어절 바이그램을 추출한다. 예를 들어 "#년도 @발 황사가 기승을 부리고 있다."에서는 '#년도 @발', '@발 황사가', ..., '부리고 있다'의 5개의 어절 바이그램이 추출된다. 제안하는 시스템에서는 국립국어원의 언어정보나눔터에 공개된 말뭉치[8]와 일부 뉴스에서 추출한 2203개의 말뭉치에서 바이그램을 추출한다.

추출된 바이그램 집합에 대한 유사도 검색의 속도 향상을 위해서, 추출된 각 바이그램에 어절 ID를 부여하고, 어절 바이그램의 각 음절을 기준으로 한 역색인을 구성한다. 아래는 추출된 역색인의 예를 보인다. 그림의 왼쪽은 어절 바이그램과 각각의 ID를, 오른쪽은 추출된 역색인을 보인다. 예를 들어 ID i인 "밥을 짓는다"의 각 음절 '밥', '을', '짓' 등은 어절의 첫음절인 '밥'의 역색인 파일에 포함된다. 이러한 구조는 유사도 기반 어절 교정에서 빠른 탐색을 지원한다.

ID	어절 바이그램
i	밥을 짓는다
j	밥을 주기에
k	밥을 먹는다
l	밥 짓기에

⇒

파일명	음절	음절 역색인
밥.txt	을	i, j, k, ...
	짓	i, l, ...
	는	i, k,...
	다	i, k, ...
	주	j, ...
	기	j, l, ...
	에	j, l, ...
	먹	k, ...
	기	j, l, ...
	...	

[그림 2] 어절 바이그램 집합 구성 예시

2.3 유사도 기반 어절 교정

유사도 기반 어절 교정에서는 맞춤법 검사와 유사 어절 탐색으로 추천 교정을 제시한다. 맞춤법 교정에서는 다음 맞춤법 검사 API[7]를 이용한다. 유사 어절 탐색에서는 추출된 어절 바이그램과 음절 역색인을 이용한 어절 바이그램 유사도 계산을 사용하여, 입력된 어절 바이그램과 유사한 말뭉치의 어절 바이그램 후보들을 찾는다. 마지막으로 얻어진 후보들에 대해 개선된 편집거리를 적용하여 입력 어절 바이그램에 가장 적합한 추천 교정을 제시한다.

대부분의 한국어 맞춤법 검사는 각 어절의 맞춤법을 독립적으로 평가한다. 따라서, '밥을 짓는다'를 '밥을 찢는다'로 표현하면 교정하지 못하는 경우가 많다. 본 시스템은 말뭉치에서 추출한 어절 바이그램을 사용하므로 문맥 정보에 기반하여 추천 교정을 제시할 수 있다. 또한 전체 문장에 대하여 교정을 제시하기 위해서는 시스템 속도가 느려질 수 있으므로, 제안하는 시스템에서는 두 어절 단위로 추천 교정을 제시하여 사용자 편의를 높이고자 한다.

시스템에서는 사용자가 작성한 표현과 가장 유사한 말뭉치의 표현을 찾기 위해 [9]의 유사도 계산 방식을 사용한다. [9]에서는 유사 표현의 빠른 탐색을 위해 자료를 블록으로 구성한다. 아래 [그림 3]에서는 '밥을 찢는다'에 대한 블록 구성 예시를 보인다. 첫음절 '밥'을 사용하여 음절 역색인 파일을 결정하고, 다음 음절부터 2음절 단위로 블록을 추출하여 '을찢', '는다'의 블록을 얻는다. 추출된 블록에 대해서 [그림 2]의 음절 역색인에서 유사 어절을 탐색한다. '을찢'은 말뭉치에서 동시에 나타는 어절 바이그램이 없기 때문에 후보가 없고, '는다'에 대해서는 i와 k를 유사 어절 후보로 얻을 수 있다. 후보가 존재하는 블록으로 타 음절들을 검색한 결과, i와 k가 후보로 도출되어 시스템에서는 이를 후보로 제시한다. 이처럼 2.2 에서 제시한 음절 역색인을 이용하면 유사 어절 바이그램을 빠르게 탐색할 수 있다.

사용자 입력	음절	블록 구성			
밥을 찢는다	밥	을찢		는다	
		을	찢	는	다
		i,j,k	-	i,k	i, k
		-		i, k	
		i			

최종 후보 어절

밥을 짓는다, 밥을 먹는다

[그림 3] 유사도 측정 예시

문장 교정을 위하여 시스템에서는 최종적으로 편집거리를 사용하여 유일한 추천 교정을 결정한다. 음소 단위의 편집거리를 사용하면서 개선된 편집거리 방식을 적용하여 외국인의 작문 특성을 반영하고자 한다.

외국인의 작문에서는 '찢다'와 '짓다', '얘를'과 '예를', [7]에서 교정하지 못하는 '외냐하면'과 같은 이중모음과 격음, 경음에서 작문 상 혼동하는 경우를 자주 발견할 수 있다[4]. 시스템에서는 편집거리의 삽입/삭제/수정 중 수정에서 혼돈되기 쉬운 이중모음과, 격음, 경음에 대해 낮은 편집거리를 부여하여 외국인의 작문 특성을 반영한다.

[표 1]은 외국인의 작문 특성을 반영한 개선된 편집거리 계산을 위한 음소별 코드 표를 보인다. 혼돈하기 쉬운 모음과 자음이 가까운 코드를 가지도록 구성되어 있다. 예를 들어 '왜냐하면'과 '외냐하면'에서 모음 '새'와 '시'의 차이는 외국인이 자주 실수하는 이중모음에 대한 실수이다. 본 시스템에서 모음 새'는 11,

‘ㅚ’는 12의 코드를 가지며, 두 음소의 코드의 차이를 코드 개수로 나눈 |11-12|/21 ≈ 0.048을 두 어절의 음소별 편집거리로 부여한다. 또한 ‘밥을 찢는다’와 ‘밥을 짓는다’의 편집거리에서는, ㅉ과 ㅈ은 1/19, ㅈ과 ㅅ이 2/26 의 편집거리를 가지므로, 두 음소 사이의 편집거리는 두 값을 합한 0.13이 된다.

시스템서는 후보로 얻어진 유사 어절 바이그램 각각에 대해서 편집거리를 계산하여 편집거리가 가장 작은 값을 추천 교정으로 제시한다. 추천 교정의 편집거리가 입력 어절 길이의 절반을 초과할 경우에는 추천 교정을 제시하지 않고, 사용자 입력이 의심스럽다는 경고만을 제시한다.

[표 1] 개선된 편집거리 계산을 위한 음소별 코드 체계

초성 (19)

0	ㄱ	7	ㄹ	14	ㅅ
1	ㅋ	8	ㅇ	15	ㅆ
2	ㄲ	9	ㅎ	16	ㅈ
3	ㄴ	10	ㅁ	17	ㅉ
4	ㄷ	11	ㅂ	18	ㅊ
5	ㅌ	12	ㅍ		
6	ㄸ	13	ㅃ		

중성(21)

0	ㅏ	7	ㅖ	14	ㅠ
1	ㅐ	8	ㅗ	15	ㅝ
2	ㅑ	9	ㅘ	16	ㅞ
3	ㅒ	10	ㅙ	17	ㅟ
4	ㅓ	11	ㅚ	18	ㅡ
5	ㅔ	12	ㅛ	19	ㅢ
6	ㅕ	13	ㅜ	20	ㅣ

종성 (26)

0	ㄱ	5	ㄵ	10	ㅆ	15	ㄿ	20	ㅎ	25	ㅇ
1	ㅋ	6	ㄶ	11	ㄷ	16	ㅀ	21	ㅁ	26	ㅎ
2	ㄲ	7	ㅈ	12	ㅌ	17	ㄺ	22	ㅂ		
3	ㄳ	8	ㅊ	13	ㄹ	18	ㄻ	23	ㅍ		
4	ㄴ	9	ㅅ	14	ㄻ	19	ㄼ	24	ㅄ		

2.4 키보드 후킹을 이용한 실시간 교정

본 시스템은 최초 작동 이후 사용자의 별도 요청 없이, 키보드 후킹을 통해 입력에서 2어절 입력 때, 새로운 어절을 인식할 때마다 분석 작업을 거쳐 출력한다.

아래는 시스템의 사용자 인터페이스를 보인다. 사용자가 입력한 문장에서 잘못된 표현은 적색으로 표현하여 추천 교정을 제시한다. 기본 화면인 [그림 4]에서 틀린 표현과 교정 어절을 확인할 수 있다. 확대 버튼을 눌러 볼 수 있는 상세 화면인 [그림 5]에서는 추천 어절 후보들을 볼 수 있다.

[그림 4] 교정 예시

[그림 5] 상세 교정 예시

3. 결론

본 논문에서는 외국인 학습자를 위한 문맥 기반 실시간 교정 시스템을 제안하였다. 형태소 분석, 어절 간의 문맥을 고려하지 않은 맞춤법 검사로 교정할 수 없었던 외국인 작문 특성을 반영한 어절 바이그램과 개선된 편집거리를 제안하였다. 향후 연구로는 기존 연구에서 속도 및 적중률 문제로 적용할 수 없었던 문맥 범위의 확장, 외국인과 일반인 모두에게서 발생할 수 있는 보편적 작문 오류의 고려, 모바일 환경을 포함한 다양한 입력 환경에서의 교정 연구를 진행하고 있다.

참고문헌

[1] 2016년 교육기본통계, 교육부, 2016. 8.
[2] 2016학년도 다문화학생 교육 지원 기본 계획, 서울특별시 교육청, 2016, 2.
[3] 서수백. “외국인 유학생의 한국어 쓰기 오류 분석”, 우리말글, 62, 127-157, 2014.
[4] 김은경, “외국인을 위한 한국어 음운 지도 방안 연구”, 충남대학교 석사학위논문, 2006.
[5] http://speller.cs.pusan.ac.kr/, 한국어 맞춤법/문법 검사기, ㈜나라인포테크.
[6] http://www.naver.com, 네이버 맞춤법 검사기.
[7] https://developers.daum.net/services/apis/grammar-checker/intro, 다음 맞춤법 검사기.
[8] https://ithub.korean.go.kr/user/total/database/corpusManager.do, 국립국어원 언어정보나눔터.
[9] 고은비, 이종우, 이재원, “효율적인 집합-기반 POI 데이터 검색 알고리즘”, 정보과학회 컴퓨팅의 실제 논문지, 19(5), 242-251. 2013.