규칙의 일반화와 통계 방식을 결합한 한국어 문맥의존 철자오류 교정규칙의 재현율 향상*)

최현수○, 권혁철, 윤애선
부산대학교
{gl2een33, hckwon, asyoon}@pusan.ac.kr

Improving Recall for Context-Sensitive Spelling Correction Rules by Combining Rule-Generalization and Statistical Method

Hyun-soo Choi○, Hyuk-Chul Kwon, Aesun Yoon
Pusan National University

요 약

한국어 맞춤법 검사기는 전자화된 한국어 텍스트에 나타난 오류어를 검색하여 이를 교정할 대치어를 제시하는 시스템이다. 이때 오류어의 유형은 크게 단순 철자오류와 문맥의존 철자오류로 구분할 수 있다. 이 중 문맥의존 철자오류는 어절(word)단위로 봤을 때는 정확하지만, 문맥을 고려하였을 때 오류가 되는 유형으로 교정 난도가 매우 높다. 문맥의존 철자오류의 교정 방법은 크게 규칙을 이용한 방법과 통계 정보에 기반을 둔 방법으로 나뉜다. 이때 규칙을 이용한 방법은 그 특성상 정확도가 매우 높지만, 반대로 재현율이 매우 낮다. 본 논문에서는 본 연구진이 기존에 연구하였던 규칙을 일반화하는 방식에 추가로 조건부확률을 이용한 통계 방식을 결합하여 정확도를 유지하면서 재현율을 향상시키는 방법을 제안한다.

주제어: 한국어 맞춤법 검사기, 문맥의존 철자오류, 선택제약명사 확장, 조건부확률 통계 방식

1. 서론

한국어 맞춤법 검사기(Korean Spelling and Grammar Checker, 이하 KSGC)는 한국어 텍스트 문서에서 오류어(error word)를 검색하여 이를 교정할 대치어(recommended word)를 제시해주는 시스템이다. 이때, 오류어는 그 유형에 따라 검색하거나 교정하는 방식이 다르다. 오류어의 유형은 크게 '단순 철자오류(isolated-term spelling error 또는 non-word spelling error)'와 '문맥의존 철자오류(context-sensitive spelling error 또는 real word spelling error)'로 구분할 수 있다. 단순 철자오류는 "요금 결죄", "감기가 낯다"에서 "*결죄"와 "*낯다"와 같이 어떤 형태소 조합으로도 한 어절을 구성하지 못하는 유형이다. 이는 단순히 형태 분석만으로 쉽게 오류를 검색할 수 있다. 하지만 문맥의존 철자오류는 "요금 결재", "감기가 낳다"에서 "결재"와 "낳다"와 같이 어절(word) 단위로 봤을 때는 정확하지만, 문맥을 고려했을 때 오류가 되는 유형이다. 문맥의존 철자오류는 좌우 문맥의 의미·통사적 관계를 고려해야만 해당 어절의 오류 여부를 판단할 수 있기 때문에 교정난도가 매우 높고, 한국어 맞춤법 검사기의 성능에 큰 영향을 미친다[1]. 문맥의존 철자오류의 교정 방법은 규칙을 이용한 방법과 통계 정보를 기반으로 한 방법으로 나뉜다. 규칙을 이용한 방법은 통계적 방법에 비해 정확도(precision)가 높지만, 재현율(recall)이 매우 낮다.

본 연구진이 개발한 KSGC는 맞춤법을 잘 모르는 일반 사용자들을 위한 범용 시스템을 추구하므로, 잘못된 대치어를 제시하지 않도록 100%에 가까운 정확도를 목표로 한다. 따라서 규칙을 기반으로 한 시스템이며, 재현율이 매우 낮다는 단점이 있다[2]. 하지만 언어교정에 대한 전문적인 지식을 가지고 있는 언어 전문가는 정확도가 크게 떨어지지 않는 선에서 재현율을 높여 오류어를 최대한 많이 검색하기를 원한다. 이때 단순 철자오류는 사전검색이나 형태분석만으로도 오류 교정이 용이하여 정확도와 재현율이 모두 높지만 문맥의존 철자오류는 그 특성상 정확도가 높아지면 재현율은 매우 낮아진다.

본 논문은 기존에 연구하였던 언어학적 규칙에 기반을 둔 통합적 규칙제약 완화를 통한 문맥의존 철자오류 교정 방식[3]에 조건부확률 모델을 이용한 통계 방식을 결합하여 정확도를 95% 정도로 유지하면서 재현율을 높이는 방식을 제안한다. 정확도 95%는 언어 교정에 대한 지식이 있는 실사용자도 그 교정 결과를 신뢰할 수 있는 체감 정확도라고 판단한다.

2. 관련 연구

2.1 문맥의존 철자오류 교정 국내·외 연구 현황

1장에서 언급했듯이, 문맥의존 철자오류 교정 연구는 크게 규칙을 이용한 방법과 통계적 방법으로 구분할 수 있다. 규칙을 이용한 방법은 구문분석 기반 규칙을 이용해 문맥의존 철자오류가 있는 텍스트는 구문 분석이 실패하는 점에 착안한 방법이다[4, 5, 6, 7]. 규칙을 이용한 방법은 구문분석기와 구축한 규칙의 성능에 따라 결과가 달라질 수 있으며, 구문 분석이 실패하였을 때 그 원인을 구분하기 어렵다는 단점이 있다. 통계적 방법으로는 단어와 주위 문맥에 나타난 단어들과의 의미적 연관성(semantic relatedness)을 측정하거나, 일반적으로 n-gram과 같은 통계 모델을 많이 사용한다. 이들 방법에서는 형태나 발음이 비슷한 오류 집합(confusion set)을 만들고 해당 어휘가 오류로 나왔을 때 오류 집합 어휘들의 의미적 연관성이나 조건부 확률,

*) 이 논문은 2014년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2012S1A5A2A03034298)

n-gram 등 확률을 계산한 다음 순위화한다. 하지만 통계적 방법은 어휘의 사용 용례가 충분하지 않으면 자료부족 문제가 발생할 수 있고 정확도 20.7 ~ 50.3%, 재현율 28.1 ~ 76.3% 정도로 성능이 높지 않다[8, 9, 10].

2.2 문맥의존 철자오류 교정 규칙

본 연구진의 KSGC에서 문맥의존 철자오류 교정은 표준국어대사전을 비롯한 한국어 사전과 언어 전문가의 지식을 토대로 삼아 수작업으로 구축한 규칙을 사용한다. [표 1]은 KSGC에서 사용된 발음 유사성에 의한 문맥의존 철자오류 중 하나인 "다리다"를 "달이다"로 교정하는 규칙을 간단하게 나타낸 예시이다.

[표 1] 문맥의존 철자오류 교정 규칙의 예

	N+P 다리(다) ==> N+P 달이(다)
(a1)	Context=B1
(a2)	N=[보약 ¦ 한약 ¦ 약 ¦ 차 ¦ 간장 ¦ 장 ¦ 고기 ¦ 엿 ¦ 국물 ¦ 물 ¦ 멸치젓 ¦ 사골 ¦ 뼈 ¦ ...]
(a3)	P=[을 ¦ 를]
(a4)	Conjugation = 1001+2001
(b1)	Context B1
(b2)	N=[{4} ¦ {5}]
(b3)	P=[을 ¦ 를]
(b4)	Conjugation = 1001+2001
(c1)	Context = B1
(c2)	N=[{4} ¦ {5}]+[-탕 ¦ -국 ¦ -찌개]
(c3)	P=[을 ¦ 를]
(c4)	Conjugation = 1001 + 2001

KSGC의 문맥의존 철자오류 교정 규칙은 각 규칙이 작동하는 핵심 어휘([표 1]에서는 "다리다")를 기준으로 해당 규칙을 적용한다. [표 1]의 규칙은 핵심 어휘 "다리다"를 기준으로 논항 'N+P'(명사+조사)가 왼쪽 첫째 어절에 나타나면(a1) "달이다"로 교정한다. 이때, P는 목적격 조사로 제약하고(a3), N은 논항의 선택제약 명사가 되며(a2), 동사의 활용형 제약(a4)으로 구성된다. 규칙 (b1)~(b4), (c1)~(c4)도 같은 맥락으로 적용되며, 선택제약 명사의 종류가 다르다.

본 연구진의 기존 연구에서는 정확도 감소를 최소로 하고 재현율을 향상시키기 위해, 지금까지 [표 1]의 규칙제약을 완화하는 방식을 연구해 왔다. 한국어 어휘의미망을 이용한 규칙의 논항의 선택제약 명사 N의 확장 방식[11], 조사 제약 조건 P의 완화 방식[12], 그리고 두 방식의 결합과 수의적 삽입 요소를 고려한 통합적 규칙제약 완화 방식[3]을 실험했다. KSGC와 세 가지 규칙 일반화 방식의 성능은 [표 3]에서 볼 수 있는데, 규칙 일반화를 통한 재현율의 향상이 만족스럽지 못했다.

3. 통계적 방식을 이용한 문맥의존 철자오류 교정

본 연구진은 규칙의 핵심 어휘 중에서 발음 유사성에 의한 문맥의존 철자오류 동사를 교정 어휘 쌍으로 선정하여 연구 및 실험을 해왔다. 발음 유사성에 따른 오류는 예를 들어, "한약을 달이다"를 입력할 때 발음 유사성에 의해 "한약을 *다리다*"로 잘못 입력하여 생기는 오류이다. 이때, 규칙기반 방식을 이용한 기존 KSGC에서는 그 특성상 교정 정확도는 매우 높지만 재현율이 매우 낮다. 또한 기존 연구의 교정 규칙의 제약을 완화하거나 규칙을 확장하는 방식을 통해 정확도를 유지하면서 재현율을 향상시키는 결과를 얻었지만 재현율의 향상 결과가 여전히 만족스럽지 않았다.

3.1 조건부확률 모델을 이용한 문맥의존 철자오류 교정

본 논문에서는 기존 연구에서 선정한 발음 유사성에 의한 문맥의존 철자오류 교정 어휘 쌍을 기반으로 문맥 정보와 핵심 어휘 간의 의존관계를 조건부확률 모델을 이용해 확률을 계산하는 통계적 방식을 이용하여 오류를 교정한다. 수식 (1)은 실제 문서에 나타난 단어 s_{ob}와 s_{ob}의 오류에 의해 쓰일 수 있는 어휘 s_i^*를 이용하여 교정 어휘 쌍 중 문맥에 해당하는 어휘를 선택하는 방법을 수식화한 것이다.

$$s' = \underset{s_i^*}{\arg max} \left\{ \prod_{k=1}^{m-1} P(s_i^* | C_L^k) \times \prod_{k=m+1}^{n} P(s_i^* | C_R^k) \right\} P(s_{ob} | s_i^*) \quad (1)$$

s_{ob} : 실제 문서에서 나타난 단어
$s_1, \ldots, s_n$: s_{ob}의 오류에 의해 쓰일 수 있는 어휘
s_{ob}^* : $s_{ob} \rightarrow s_{ob}$로 올바르게 쓰인 경우
s_i^* : 오류에 의해 쓰인 경우 $(i = 1, .., n)$
C_L : 왼쪽 문맥
C_R : 오른쪽 문맥

수식 (1)에서 $\prod_{k=1}^{m-1} P(s_i^* | C_L^k) \times \prod_{k=m+1}^{n} P(s_i^* | C_R^k)$은 좌우 문맥 C_L과 C_R이 나타났을 때 s_i^*의 확률을 나타낸다. 이는 신뢰성 있는 대용량 말뭉치에서 구하는 것이 현실적으로 힘들기 때문에 나이브베이즈의 가정(Naive Assumption)을 도입하다. 만약 전체 n개의 어휘로 이루어진 문장에서 m번째 어휘가 교정 어휘 대상이라면, 왼쪽 문맥 C_L의 개수는 m-1이고 오른쪽 문맥 C_R의 개수는 n-m이다. 이 가정에서 교정 어휘의 좌우 문맥 모두를 보는 것은 큰 의미가 없기 때문에 문맥을 보는 window size 크기를 제한할 것이다.

그리고 수식 (1)에서 s_i^*가 s_{ob}^* 로 쓰였는지 오류인지를 나타내는 $P(s_{ob} | s_i^*)$는 오류가 태깅된 말뭉치가 없으면 구하기 어렵다. 때문에 각각의 오류율인 $error(s_i^*)$을 안다고 가정하고 수식 (2)와 같이 바꿀 수 있다.

$$P(s_{ob} | s_i^*) = \begin{cases} 1 - \sum_{i=1}^{n} error(s_i^*) & s_i^* = s_{ob}^* \\ error(s_i^*) & s_i^* \neq s_{ob}^* \end{cases} \quad (2)$$

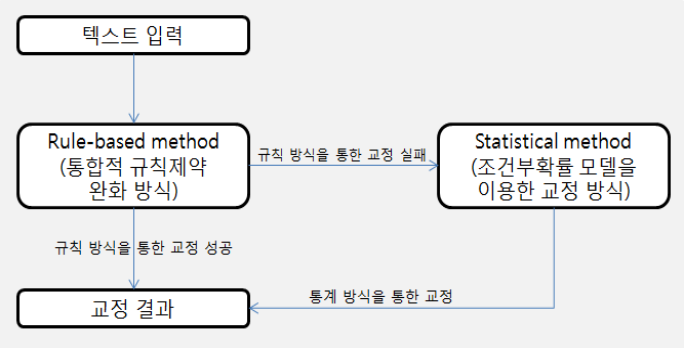
여기서 오류율은 사용자의 목적에 따라 값을 설정할 수 있다. 실제 기존 연구에서 발표된 오타에 의한 오류 발생률은 5%로서[13] 매우 낮으므로 실제 텍스트에서 나타난 단어가 오류일 확률은 매우 낮다.

본 논문에서는 조건부확률 모델과 오류율, window size 간의 관계를 알아보기 위해 오류율을 3~5%, window size를 3~5까지 달리하여 실험했다. 그리고 조건부확률 모델의 확률 계산을 위해 학습 데이터를 태깅하고, 태깅된 문서에 나타난 핵심 어휘 s_{ob}와 s_{ob}의 좌우 문맥 window size 범위 안에서 content word인 명사, 형용사, 동사의 빈도를 측정하여 사전을 구축하였다. 또한, 핵심 어휘 s_{ob}의 좌우 문맥에 동사가 출현하면 동사의 앞(좌), 혹은 뒤(우)의 문맥은 s_{ob}와 관계가 없을 가능성이 높기 때문에 사전 구축에서 제외했다.

3.2 규칙과 통계 방식의 결합

본 논문은 기존 규칙기반의 통합적 규칙제약 완화 방식과 3.1의 조건부확률을 통한 통계방식을 결합하는 방법을 [그림 1]과 같이 제안한다. 정확도를 유지하거나 정확도 감소를 최소화하기 위해 교정 정확도가 높은 규칙기반 통합적

규칙제약 완화 방식을 먼저 적용하고, 재현율 향상을 위해 규칙기반 방식에서 교정하지 못한 텍스트들을 조건부확률 모델을 이용한 통계 교정방식을 통해 교정한다.



[그림 1] 규칙과 통계방식을 결합한 문맥의존 철자오류 교정 방식

통합적 규칙제약 완화는 ① 선택제약 명사 확장 방식과 조사제약 완화방식을 단순 결합한 방식(Simple Integrated Method, SIM), ② 핵심 어휘와 N+P의 거리를 고려하여, P 완화 방식을 동적으로 결합한 방식(Dynamic Integrated Method, DIM), ③ SIM에 부사 삽입과 관형형 구조를 고려하여 규칙을 추가한 통합 방식(Extended Integrated Method, EIM) 총 3가지 방식을 실험하였다. 본 논문에서는 위 세 방식에, 오류율과 window size를 바꾸어 가면서 실험한 통계 방식을 추가하여 SSIM(Statistical & SIM), SDIM(Statistical & DIM), SEIM(Statistical & EIM) 결합 방식을 살펴본다.

4. 실험 및 평가

4.1 실험 환경

본 논문에서 제안하는 방식을 검증하기 위한 문맥의존 철자오류 교정 어휘 쌍은 KSGC가 공개된 웹사이트 '우리말 배움터'에서 일반사용자가 가장 많이 질의하는 문맥의존 철자오류 중에서 발음이 유사한 술어 10개 어휘 쌍으로 선정하였다[2, 14]. [표 2]는 평가 규칙과 평가데이터, 학습 데이터를 정리한 표이다. 성능 평가를 위한 평가 말뭉치는 10쌍(20개)의 규칙별로 정문 100문장(대상어가 포함된 문장), 오류문 100문장(대치어가 포함된 문장), 총 200문장을 구성하였다. 예를 들어 TR3("다리다" -> "달이다")의 규칙의 성능을 평가하기 위해, 대상어 "다리다"가 사용된 정문과 대치어 "달이다"가 사용된 정문에서 "달이다"를 모두 "다리다"로 바꾸어 인위적으로 오류 문장을 만들어 평가 말뭉치를 구성하여 교정의 정확도와 재현율을 구하는 것이다. TR4("달이다" -> "다리다")는 정문과 오류문이 TR3과는 역으로 구성된다. 이때, 평가말뭉치 구성을 위한 원말뭉치로는 '2007 세종 형태분석 말뭉치(1,500만 어절)'를 주로 사용하고, 충분한 정문의 예시가 나타나지 않는 경우, '신문기사 말뭉치(2009-2012)'를 추가했다.

통계 방식의 조건부확률 모델을 위한 학습데이터는 중앙일보에서 수집한 750만 건의 기사(2000년 ~ 2013년)에서 추출한 데이터를 이용하여 [표 2]과 같이 구축하였다.

4.2 실험 결과

[표 3]은 기존 KSGC의 성능과 규칙기반인 통합적 규칙제약 완화의 각 방식별 가장 좋은 성능을 보인 SIM-CWE, DIM-CWE, EIM-CWE의 결과를 나타낸 표이다. P와 R은 오류어 검색과 교정에 대한 Precision(정확도)과 Recall(재현율)이다. 기존 KSGC는 평가 데이터에 대해서 정확도가 모두 100%가 나왔지만 재현율이 평균 29.95%로 매우 낮다. 통합적 규칙제약 완화 방식에서는 평균 정확도를 95.19~96.17%로 95% 선을 유지하면서 EIM-CWE방식에서 평균 재현율이 37.56%까지 향상시켰다. 이는 정확도를 유지하면서 재현율을 향상시켰지만 여전히 재현율의 향상이 만족스럽지 않았다.

[표 2] 평가 데이터 구성

Target Word				바른 문장	오류 문장	대상어 학습 데이터 수
Rule	대상어		대치어			
TR1	낫다	->	낳다	100	100	4,402
TR2	낳다	->	낫다	100	100	4,400
TR3	다리다	->	달이다	100	100	3,447
TR4	달이다	->	다리다	100	100	5,903
TR5	마치다	->	맞히다	100	100	20,000
TR6	맞히다	->	마치다	100	100	20,000
TR7	맞추다	->	맞히다	100	100	20,000
TR8	맞히다	->	맞추다	100	100	20,000
TR9	배다	->	베다	100	100	17,381
TR10	베다	->	배다	100	100	55,241
TR11	안치다	->	앉히다	100	100	485
TR12	앉히다	->	안치다	100	100	7,013
TR13	저리다	->	절이다	100	100	3,199
TR14	절이다	->	저리다	100	100	2,015
TR15	젖히다	->	제치다	100	100	4,183
TR16	제치다	->	젖히다	100	100	57,190
TR17	집다	->	짚다	100	100	20,094
TR18	짚다	->	집다	100	100	22,828
TR19	찢다	->	찧다	100	100	7,698
TR20	찧다	->	찢다	100	100	2,000

[표 3] KSGC와 통합적 규칙제약 완화 방식 각 방식별 결과

Rule	KSGC		SIM-CWE		DIM-CWE		EIM-CWE	
	P	R	P	R	P	R	P	R
TR1	100.0	39.42	96.43	50.94	96.08	47.12	96.67	54.21
TR2	100.0	36.63	97.30	35.64	97.14	34.00	97.62	40.59
TR3	100.0	29.70	86.89	51.46	90.38	46.08	86.96	57.69
TR4	100.0	35.00	100.0	46.53	100.0	43.00	100.0	51.49
TR5	100.0	25.00	96.15	25.00	96.00	24.00	92.59	25.00
TR6	100.0	47.00	89.29	50.00	89.29	50.00	79.37	50.00
TR7	100.0	8.00	100.0	9.00	100.0	8.00	100.0	9.00
TR8	100.0	28.00	90.32	28.00	88.89	23.76	90.32	28.00
TR9	100.0	31.00	100.0	29.70	100.0	26.73	100.0	32.67
TR10	100.0	42.00	100.0	42.00	100.0	42.00	97.73	43.00
TR11	100.0	32.00	94.23	49.00	96.08	49.00	94.23	49.00
TR12	100.0	59.80	100.0	69.61	100.0	68.32	100.0	72.82
TR13	100.0	36.63	100.0	40.59	100.0	39.60	100.0	47.52
TR14	100.0	35.00	100.0	39.00	100.0	33.00	100.0	40.00
TR15	100.0	16.00	87.50	28.00	86.67	26.00	85.29	29.00
TR16	100.0	15.00	96.15	25.00	96.55	28.00	96.67	29.00
TR17	100.0	23.00	100.0	24.00	100.0	24.00	100.0	24.00
TR18	100.0	13.00	85.71	18.00	86.36	19.00	86.36	19.00
TR19	100.0	31.68	100.0	34.31	100.0	29.00	100.0	34.31
TR20	100.0	8.00	100.0	14.85	100.0	14.85	100.0	14.85
평균	100.0	29.59	96.00	35.53	96.17	33.77	**95.19**	**37.56**

[표 4] 오류율과 window size 변화에 따른 조건부확률 모델을 이용한 통계 방식 실험 결과

Rule	오류율 5%						오류율 4%						오류율 3%					
	window size 3		window size 4		window size 5		window size 3		window size 4		window size 5		window size 3		window size 4		window size 5	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
TR1	100.0	50.00	100.0	53.00	100.0	52.00	100.0	47.00	100.0	51.00	100.0	51.00	100.0	41.00	100.0	47.00	100.0	47.00
TR2	97.30	36.00	97.30	36.00	97.30	36.00	97.06	33.00	97.06	33.00	97.14	34.00	96.97	32.00	96.97	32.00	95.83	23.00
TR3	97.50	39.00	95.83	46.00	98.04	50.00	97.37	37.00	95.35	41.00	97.78	44.00	96.88	31.00	97.37	37.00	97.50	39.00
TR4	100.0	30.00	100.0	31.00	100.0	35.00	100.0	28.00	100.0	29.00	100.0	31.00	100.0	23.00	100.0	26.00	100.0	25.00
TR5	94.12	48.00	94.44	51.00	96.08	49.00	93.75	45.00	94.23	49.00	95.83	46.00	93.33	42.00	93.75	45.00	95.74	45.00
TR6	98.21	55.00	96.23	51.00	96.55	56.00	98.04	50.00	97.96	48.00	98.18	54.00	97.96	48.00	100.0	45.00	100.0	47.00
TR7	96.97	32.00	97.06	33.00	97.22	35.00	96.67	29.00	96.43	27.00	96.97	32.00	96.15	25.00	95.65	22.00	96.30	26.00
TR8	97.06	66.00	97.14	68.00	95.83	69.00	98.44	63.00	98.44	63.00	97.01	65.00	98.31	58.00	98.39	61.00	96.88	62.00
TR9	100.0	47.00	100.0	49.00	96.43	54.00	100.0	45.00	100.0	44.00	96.30	52.00	100.0	39.00	100.0	41.00	96.00	48.00
TR10	100.0	50.00	100.0	47.00	100.0	45.00	100.0	49.00	100.0	42.00	100.0	40.00	100.0	42.00	100.0	38.00	100.0	34.00
TR11	85.71	54.00	83.33	55.00	80.82	59.00	86.89	53.00	83.87	52.00	81.43	57.00	87.04	47.00	87.27	48.00	82.81	53.00
TR12	100.0	9.00	100.0	11.00	100.0	9.00	100.0	9.00	100.0	10.00	100.0	8.00	100.0	7.00	100.0	8.00	100.0	5.00
TR13	100.0	55.00	100.0	58.00	100.0	59.00	100.0	55.00	100.0	58.00	100.0	59.00	100.0	53.00	100.0	55.00	100.0	57.00
TR14	100.0	64.00	100.0	64.00	100.0	65.00	100.0	63.00	100.0	64.00	100.0	65.00	100.0	63.00	100.0	64.00	100.0	65.00
TR15	88.16	67.00	85.71	72.00	86.21	75.00	88.57	62.00	85.19	69.00	85.88	73.00	89.86	62.00	86.25	69.00	85.54	71.00
TR16	100.0	37.00	97.14	34.00	96.67	29.00	100.0	35.00	100.0	31.00	100.0	28.00	100.0	30.00	100.0	29.00	100.0	28.00
TR17	100.0	44.00	100.0	43.00	97.73	43.00	100.0	39.00	100.0	40.00	97.56	40.00	100.0	35.00	100.0	36.00	100.0	34.00
TR18	100.0	30.00	96.97	32.00	90.32	28.00	100.0	26.00	100.0	28.00	92.59	25.00	100.0	24.00	100.0	21.00	95.24	20.00
TR19	100.0	51.00	100.0	48.00	100.0	48.00	100.0	46.00	100.0	44.00	100.0	46.00	100.0	40.00	100.0	40.00	100.0	41.00
TR20	98.15	53.00	95.08	58.00	95.24	60.00	98.00	49.00	96.49	55.00	94.92	56.00	97.92	47.00	96.36	53.00	96.49	55.00
평균	97.66	45.85	96.81	47.00	96.22	47.80	97.74	43.15	97.25	43.90	96.58	45.30	97.72	39.45	97.60	40.85	96.92	41.25

[표 4]는 오류율과 window size 변화에 따른 조건부확률 모델을 이용한 문맥의존 철자오류 교정의 결과이다. 오류율을 낮게 볼수록 실제 텍스트에 나타난 대상어가 올바르게 쓰였을 확률이 올라가기 때문에 정확도가 높게 나오고, window size가 커질수록 좌우 문맥에서 볼 수 있는 정보가 많아지기 때문에 재현율이 올라가는 결과를 보였다. 통계 방식만을 이용하여 교정을 하였을 때는 평균 정확도가 96.58~97.74%, 평균 재현율이 39.45~47.80%의 결과를 보였다. 하지만, 기존 KSGC 규칙을 확장할 때 통계 방식만 적용하기에는 규칙 간 편차가 크다는 단점이 있다. 평가 데이터와 학습 데이터의 수가 아직 부족하고, 교정 어휘 대상어와 대치어 간의 학습 데이터 수가 차이가 큰 규칙인 TR11/TR12(“안치다”-“앉히다”)와 TR15/TR16(“젖히다”-“제치다”)은 학습 데이터의 수가 적은 규칙의 정확도가 낮고, 학습 데이터의 수가 많은 규칙의 재현율이 낮게 나오는 결과를 보였다.

[표 5] SSIM 결합 방식 실험 결과

Rule	오류율 5%						오류율 4%						오류율 3%					
	window size 3		window size 4		window size 5		window size 3		window size 4		window size 5		window size 3		window size 4		window size 5	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
TR1	97.10	67.00	97.22	70.00	97.22	70.00	97.06	66.00	97.18	69.00	97.22	70.00	96.83	61.00	97.06	66.00	97.10	67.00
TR2	96.43	54.00	96.43	54.00	96.36	53.00	96.23	51.00	96.36	53.00	96.30	52.00	96.23	51.00	96.30	52.00	95.74	45.00
TR3	88.31	68.00	87.65	71.00	89.02	73.00	88.31	68.00	87.34	69.00	88.61	70.00	87.84	65.00	88.46	69.00	88.61	70.00
TR4	100.0	58.00	100.0	57.00	100.0	59.00	100.0	57.00	100.0	56.00	100.0	57.00	100.0	55.00	100.0	56.00	100.0	56.00
TR5	94.83	55.00	95.08	58.00	96.61	57.00	94.64	53.00	94.92	56.00	96.49	55.00	94.34	50.00	94.74	54.00	96.43	54.00
TR6	91.14	72.00	91.25	73.00	91.36	74.00	91.14	72.00	91.03	71.00	91.14	72.00	91.03	71.00	92.00	69.00	92.21	71.00
TR7	97.30	36.00	97.44	38.00	97.56	40.00	97.06	33.00	96.97	32.00	97.37	37.00	96.88	31.00	96.55	28.00	96.97	32.00
TR8	93.15	68.00	93.42	71.00	92.31	72.00	94.20	65.00	94.37	67.00	93.15	68.00	93.85	61.00	94.29	66.00	92.96	66.00
TR9	100.0	60.00	100.0	59.00	96.97	64.00	100.0	58.00	100.0	56.00	96.92	63.00	100.0	52.00	100.0	53.00	96.77	60.00
TR10	100.0	59.00	100.0	57.00	100.0	57.00	100.0	59.00	100.0	57.00	100.0	55.00	100.0	55.00	100.0	55.00	100.0	53.00
TR11	86.25	69.00	85.54	71.00	82.76	72.00	86.08	68.00	86.08	68.00	83.33	70.00	87.01	67.00	88.31	68.00	85.19	69.00
TR12	100.0	71.00	100.0	71.00	100.0	71.00	100.0	71.00	100.0	71.00	100.0	71.00	100.0	70.00	100.0	70.00	100.0	70.00
TR13	100.0	59.00	100.0	62.00	100.0	63.00	100.0	59.00	100.0	62.00	100.0	63.00	100.0	57.00	100.0	59.00	100.0	61.00
TR14	100.0	69.00	100.0	69.00	100.0	70.00	100.0	68.00	100.0	69.00	100.0	70.00	100.0	68.00	100.0	69.00	100.0	70.00
TR15	86.59	71.00	84.27	75.00	84.62	77.00	87.34	69.00	84.09	74.00	84.62	77.00	87.34	69.00	85.06	74.00	84.44	76.00
TR16	97.92	47.00	97.87	46.00	97.67	42.00	97.92	47.00	97.78	44.00	97.67	42.00	97.78	44.00	97.73	43.00	97.67	42.00
TR17	100.0	52.00	100.0	52.00	98.15	53.00	100.0	50.00	100.0	51.00	98.11	52.00	100.0	47.00	100.0	49.00	100.0	48.00
TR18	93.18	41.00	91.49	43.00	89.13	41.00	92.50	37.00	93.02	40.00	88.37	38.00	92.11	35.00	92.11	35.00	89.47	34.00
TR19	100.0	60.00	100.0	61.00	100.0	60.00	100.0	58.00	100.0	58.00	100.0	59.00	100.0	53.00	100.0	56.00	100.0	56.00
TR20	98.33	59.00	95.38	62.00	95.52	64.00	98.25	56.00	96.77	60.00	95.31	61.00	98.21	55.00	96.67	58.00	96.77	60.00
평균	**96.03**	59.75	95.65	61.00	95.26	**61.60**	96.04	58.25	95.80	59.15	95.23	60.10	95.97	55.85	95.96	57.45	95.52	58.00

[표 6] SDIM 결합 방식 실험 결과

Rule	오류율 5%						오류율 4%						오류율 3%					
	window size 3		window size 4		window size 5		window size 3		window size 4		window size 5		window size 3		window size 4		window size 5	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
TR1	97.01	65.00	97.14	68.00	97.14	68.00	96.97	64.00	97.10	67.00	97.14	68.00	96.72	59.00	96.97	64.00	97.01	65.00
TR2	96.43	54.00	96.43	54.00	96.36	53.00	96.23	51.00	96.36	53.00	96.30	52.00	96.23	51.00	96.30	52.00	95.74	45.00
TR3	91.55	65.00	90.67	68.00	92.11	70.00	91.55	65.00	90.41	66.00	91.78	67.00	91.04	61.00	91.55	65.00	91.78	67.00
TR4	100.0	56.00	100.0	56.00	100.0	58.00	100.0	55.00	100.0	54.00	100.0	55.00	100.0	52.00	100.0	54.00	100.0	53.00
TR5	94.83	55.00	95.08	58.00	96.61	57.00	94.64	53.00	94.92	56.00	96.49	55.00	94.34	50.00	94.74	54.00	96.43	54.00
TR6	91.14	72.00	91.25	73.00	91.36	74.00	91.14	72.00	91.03	71.00	91.14	72.00	91.03	71.00	92.00	69.00	92.21	71.00
TR7	97.30	36.00	97.44	38.00	97.56	40.00	97.06	33.00	96.97	32.00	97.37	37.00	96.88	31.00	96.55	28.00	96.97	32.00
TR8	93.06	67.00	93.24	69.00	92.11	70.00	94.12	64.00	94.20	65.00	92.96	66.00	93.75	60.00	94.12	64.00	92.75	64.00
TR9	100.0	56.00	100.0	56.00	96.83	61.00	100.0	54.00	100.0	52.00	96.72	59.00	100.0	48.00	100.0	49.00	96.55	56.00
TR10	100.0	59.00	100.0	57.00	100.0	57.00	100.0	59.00	100.0	57.00	100.0	55.00	100.0	55.00	100.0	55.00	100.0	53.00
TR11	87.34	69.00	86.59	71.00	83.72	72.00	87.18	68.00	87.18	68.00	84.34	70.00	88.16	67.00	89.47	68.00	86.25	69.00
TR12	100.0	70.00	100.0	70.00	100.0	70.00	100.0	70.00	100.0	70.00	100.0	70.00	100.0	69.00	100.0	69.00	100.0	69.00
TR13	100.0	58.00	100.0	61.00	100.0	62.00	100.0	58.00	100.0	61.00	100.0	62.00	100.0	56.00	100.0	58.00	100.0	60.00
TR14	100.0	68.00	100.0	68.00	100.0	69.00	100.0	67.00	100.0	68.00	100.0	69.00	100.0	67.00	100.0	68.00	100.0	69.00
TR15	86.59	71.00	84.27	75.00	84.62	77.00	87.18	68.00	83.91	73.00	84.44	76.00	87.18	68.00	84.88	73.00	84.27	75.00
TR16	98.04	50.00	98.00	49.00	97.83	45.00	98.04	50.00	97.92	47.00	97.83	45.00	97.92	47.00	97.87	46.00	97.83	45.00
TR17	100.0	52.00	100.0	52.00	98.15	53.00	100.0	50.00	100.0	51.00	98.11	52.00	100.0	47.00	100.0	49.00	100.0	48.00
TR18	93.33	42.00	91.67	44.00	89.36	42.00	92.68	38.00	93.18	41.00	88.64	39.00	92.31	36.00	92.31	36.00	89.74	35.00
TR19	100.0	60.00	100.0	61.00	100.0	60.00	100.0	58.00	100.0	58.00	100.0	59.00	100.0	53.00	100.0	56.00	100.0	56.00
TR20	98.33	59.00	95.38	62.00	95.52	64.00	98.25	56.00	96.77	60.00	95.31	61.00	98.21	55.00	96.67	58.00	96.77	60.00
평균	**96.25**	59.20	95.86	60.50	95.46	**61.10**	96.25	57.65	96.00	58.50	95.43	59.45	96.19	55.15	96.17	56.75	95.72	57.30

[표 7] EIM 방식과 조건부확률 모델을 이용한 통계 방식을 결합한 실험 결과

Rule	오류율 5%						오류율 4%						오류율 3%					
	window size 3		window size 4		window size 5		window size 3		window size 4		window size 5		window size 3		window size 4		window size 5	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
TR1	97.22	70.00	97.33	73.00	97.33	73.00	97.18	69.00	97.30	72.00	97.33	73.00	96.97	64.00	97.18	69.00	97.22	70.00
TR2	96.61	57.00	96.61	57.00	96.55	56.00	96.49	55.00	96.55	56.00	96.49	55.00	96.49	55.00	96.49	55.00	96.15	50.00
TR3	87.80	72.00	87.06	74.00	88.37	76.00	87.80	72.00	86.90	73.00	88.10	74.00	87.50	70.00	87.95	73.00	88.10	74.00
TR4	100.0	60.00	100.0	60.00	100.0	62.00	100.0	59.00	100.0	59.00	100.0	60.00	100.0	58.00	100.0	59.00	100.0	59.00
TR5	94.83	55.00	95.08	58.00	96.61	57.00	94.64	53.00	94.92	56.00	96.49	55.00	94.34	50.00	94.74	54.00	96.43	54.00
TR6	83.72	72.00	83.91	73.00	84.09	74.00	83.72	72.00	83.53	71.00	83.72	72.00	83.53	71.00	84.15	69.00	84.52	71.00
TR7	97.30	36.00	97.44	38.00	97.56	40.00	97.06	33.00	96.97	32.00	97.37	37.00	96.88	31.00	96.55	28.00	96.97	32.00
TR8	93.15	68.00	93.42	71.00	92.31	72.00	94.20	65.00	94.37	67.00	93.15	68.00	93.85	61.00	94.29	66.00	92.96	66.00
TR9	100.0	60.00	100.0	60.00	96.97	64.00	100.0	58.00	100.0	57.00	96.92	63.00	100.0	52.00	100.0	54.00	96.77	60.00
TR10	98.33	59.00	98.28	57.00	98.28	57.00	98.33	59.00	98.28	57.00	98.21	55.00	98.21	55.00	98.21	55.00	98.15	53.00
TR11	86.25	69.00	85.54	71.00	82.76	72.00	86.08	68.00	86.08	68.00	83.33	70.00	87.01	67.00	88.31	68.00	85.19	69.00
TR12	100.0	74.00	100.0	74.00	100.0	74.00	100.0	74.00	100.0	74.00	100.0	74.00	100.0	73.00	100.0	73.00	100.0	73.00
TR13	100.0	62.00	100.0	65.00	100.0	66.00	100.0	62.00	100.0	65.00	100.0	66.00	100.0	62.00	100.0	64.00	100.0	65.00
TR14	100.0	69.00	100.0	69.00	100.0	70.00	100.0	68.00	100.0	69.00	100.0	70.00	100.0	68.00	100.0	69.00	100.0	70.00
TR15	85.54	71.00	84.27	75.00	84.62	77.00	86.25	69.00	84.09	74.00	84.62	77.00	86.25	69.00	85.06	74.00	84.44	76.00
TR16	98.08	51.00	98.04	50.00	97.87	46.00	98.08	51.00	97.96	48.00	97.87	46.00	97.96	48.00	97.92	47.00	97.87	46.00
TR17	100.0	52.00	100.0	52.00	98.15	53.00	100.0	50.00	100.0	51.00	98.11	52.00	100.0	47.00	100.0	49.00	100.0	48.00
TR18	93.33	42.00	91.67	44.00	89.36	42.00	92.68	38.00	93.18	41.00	88.64	39.00	92.31	36.00	92.31	36.00	89.74	35.00
TR19	100.0	60.00	100.0	61.00	100.0	60.00	100.0	58.00	100.0	58.00	100.0	59.00	100.0	53.00	100.0	56.00	100.0	56.00
TR20	98.33	59.00	95.38	62.00	95.52	64.00	98.25	56.00	96.77	60.00	95.31	61.00	98.21	55.00	96.67	58.00	96.77	60.00
평균	95.52	60.90	95.20	62.20	94.82	**62.75**	95.54	59.45	95.35	60.40	94.78	61.30	95.48	57.25	95.49	58.80	95.06	59.35

[표 5~7]은 통합적 규칙제약 완화 방식 각 방식에 조건부확률 모델을 이용한 통계 방식을 결합한 3가지의 결과를 정리한 표이다. 이러한 결합 방식에서는 모두 [표 3]이나 [표 4]에 비해 규칙 간 재현율의 편차가 크게 줄어든 것을 볼 수 있다.

기존 통합적 규칙제약 완화 방식들의 결과들에서는 DIM-CWE 방식이 96.17%로 정확도가 제일 높았고, EIM-CWE 방식이 37.56%로 재현율이 제일 높았다. 통계 방식과 결합한 결과들 또한 비슷하게 나타났다. SSIM 결합 방식에서는 평균 정확도가 오류율 5%, window size 3일 때 96.03%로 가장 높았고 평균 재현율은 오류율 5%, window size 5 일 때 61.60%로 가장 높게 나왔다. SDIM 결합 방식에서는 기존 연구 결과의 양상과 같이 정확도가 가장 높았다. SDIM 또한 오류율 5%, window size 3에서 평균 정확도가 96.25%로 가장 높게 나왔으며 재현율은 오류율 5%, window size 5 일 때 61.10%로 가장 높게 나왔다. SEIM 방식은 3가지 방식 중 평균 정확도가 가장 낮고(오류율 4%, window size 5 일 때 최저 94.78%), 평균 재현율이 오류율 5%, window size 5 일 때 62.75%로 가장 높게 나왔다.

[표 8] 각 결합 방식별 평균 F-measure 결과

	오류율 5%			오류율 4%			오류율 3%		
window size	3	4	5	3	4	5	3	4	5
SSIM	73.66	74.49	74.82	72.52	73.14	73.69	70.61	71.87	72.17
SDIM	73.31	74.18	74.51	72.11	72.70	73.26	70.11	71.38	71.69
SEIM	74.38	75.24	**75.52**	73.29	73.95	74.45	71.58	72.78	73.08

[표 8]은 각 결합 방식의 평균 F-measure 값을 정리한 표이다. 최고 평균 정확도는 SDIM, 최고 평균 재현율은 SEIM 방식에서 나왔고, F-measure 값을 기준으로 봤을 때 오류율 5%, window size 5의 SEIM 방식이 가장 좋은 성능을 보였다. 하지만 SEIM 방식에서도 TR3, TR6, TR11, TR15의 결과는 정확도가 90%이하로 나타났다. TR3(“다리다” -> “달이다”)와 TR6(“맞히다” -> “마치다”)은 통합적 규칙 제약 완화 방식에서 과도한 확장으로 인해 정확도가 낮게 나타났고, TR11(“안치다” -> “앉히다”)와 TR15(“젖히다” -> “제치다”)는 학습 데이터의 크기 차이 때문에 생긴 결과이다.

규칙과 통계를 결합한 방식에서 window size가 커지면서 정확도가 감소하고 재현율이 향상하는 결과를 보였지만, 교정하지 못하는 오류가 여전히 있었다. [표 9]는 window size가 커졌을 때 교정하지 못하는 오류의 대표적인 예이다.

[표 9] window size가 커지면서 생기는 오류 예시

예시 문장	그 애는 자기 병이 다 **낳았다고** 생각했지만 그렇지가 못했어요.									
window size	-1	1	-2	2	-3	3	-4	4	-5	5
문맥		생각하다(v)	병(n)				애(n)			
낳다	0	14	3	0	0	0	88	0	0	0
낫다	0	23	126	0	0	0	0	0	0	0

[표 9]는 TR2(“낳다” -> “낫다”)에서 교정을 하지 못한 예시이다. 4,5행의 숫자는 핵심 어휘와 문맥 단어간의 빈도를 나타낸다. 조건부확률 모델에서 window size를 3까지 보았을 때, 핵심 어휘 “낳다”의 좌우 문맥으로 대명사를 제외한 명사 “병”, 동사 “생각하다”가 있고, “병”으로 인해 “낳다”를 “낫다”로 올바르게 고쳤다. 하지만 window size가 커지면서 더 앞(좌)의 문맥인 명사 “애”로 인해 고치지 못하는 문제가 있었다. 이와 같은 문제를 해결하기 위해서는 더 많은 양의 학습 데이터와 세밀한 통계 기법의 적용이 필요하다.

5. 결론

기존 맞춤법 검사기의 사용자 대부분은 정확도를 중요하게 생각하기 때문에 정확도가 100%에 가깝지만, 재현율은 매우 낮다. 그 문제를 해결하기 위해 규칙기반의 통합적 규칙제약 완화 방식을 통해 정확도를 유지하거나 최소한으로 감소시키면서 재현율을 향상시킴과 동시에 사용자가 정확도와 재현율을 선택할 수 있는 ‘사용자 선택형 KSGC’의 개발을 최종 목표로 연구를 진행하였다. 하지만 규칙 기반의 방식만을 확장하는 것은 그 특성상 재현율을 크게 향상시키기에는 한계가 있었다.

본 논문에서는 기존의 규칙 기반 방식을 이용한 통합적 규칙제약 완화 방식과 조건부확률 모델을 이용한 통계 방식을 결합하는 방법을 제안하였다. 결합 방식 3가지 모두 정확도가 연구 목표인 95%를 유지하면서 재현율을 최고 62.75%까지 향상시킬 수 있었다. 이 결과는 학습데이터와 평가데이터 구성을 서로 다른 범주의 말뭉치에서 구성하여 연구 목표에 도달한 큰 의의가 있다. 하지만 아직 통계 방식을 적용하는 방식의 깊이가 얕고, 학습 데이터와 평가 데이터의 자료부족 문제가 있다.

향후 연구에서는 좀 더 다양하고 정밀한 통계 방식을 적용하고, 단순 결합이 아닌 규칙 방식과 통계 방식의 각 결과에 대한 가중치를 주어 교정하는 연구를 진행할 예정이다.

참고문헌

[1] Christopher, D. Manning, Raghavan Prabhakar, and Schütze Hinrich. "Introduction to information retrieval." An Introduction To Information Retrieval: 151-177. (2008)

[2] “Korean Speller and Grammar Checker 4.5”, <http://speller.cs.pusan.ac.kr>, released in Oct. 2nd 2013.

[3] 최현수, 윤애선, 권혁철. "통합적 방식을 이용한 한국어 문맥의존 철자오류 교정규칙의 재현율 향상", 2014 한국컴퓨터종합학술대회, 한국정보과학회, 2014.06.25~2014.06.27, 부산, pp.580-582

[4] Stephen D. Richardson and Lisa C. Braden-harder. “The experience of developing a large-scale natural language text processing system: CRITIQUE,” Proc. The 2nd Annual Applied Natural Language Conference, pp.195-202. (1988)

[5] Ralph M. Weischedel and Norman K. Sondheimer. “Meta-rules as a basis for processing ill-formed input,” Comutational Linguistics, vol.9, no.3-4, pp.161-177. (1983)

[6] Linda Z. Suri. “Language transfer: A foundation for correcting the written English of ASL signers,” University of Delaware Technical Report TR-91-19. (1991)

[7] Jaime G. Carbonell and Philip J. Hayes. “Recovery strategies for parsing extragrammatical language,” Computational Linguistics, vol.9, no.3-4. (1983)

[8] Hirst, Graeme and Alexander Budanitsky. “Correcting real-word spelling errors by restoring lexical cohesion”, Natural Language Engineering, Vol.11, No.1, pp.87-111. (2005)

[9] Wilcox-O’Hearn, Amber, Graeme Hirst, and Alexander Budanitsky. “Real-word spelling correction with trigrams: A reconsideration of the mays, damerau, and mercer model”, Proceedings of 9th International Conference on Intelligent Text Processing and Computational Linguistics, Vol.4919, pp.605-616. (2008)

[10] Islam, Aminul and Diana Inkpe. “Real-Word Spelling Correction using Google Web 1T 3-grams”, Proceeding of International Conference on Natural Language Processing and Knowledge Engineering, Vol.3, pp.1241-1249. (2009)

[11] 김민호, 최현수, 권혁철 & 윤애선. 한국어 어휘의미망을 이용한 문맥 철자오류 교정규칙의 일반화. 한국정보과학회 학술발표논문집, 653-655. (2013)

[12] 최현수, 윤애선 & 권혁철. 조사제약 조건의 완화에 의한 문맥의존 철자오류 교정의 재현율 향상 방식. 정보과학회 논문지: 소프트웨어 및 응용, 41(3), 249-256. (2014)

[13] 최철, 박세진, 김철중, 권규식 (2000), “쿼르타이 키보드에 기초한 인간공학키보드 설계를 위한 오타율 분석”, 대한인간 공학회 춘계학술대회, pp.142~145.

[14] “우리말 배움터“ [Online]. Available: http://urimal.cs.pusan.ac.kr