DOI: 10.3745/KIPSTB.2009.16B.5.427

교정사전과 신문기사 말뭉치를 이용한 한국어 철자 오류 교정 모델

이 세 희[†] · 김 학 수[††]

요 약

인터넷 및 모바일 환경의 빠른 발전과 함께 신조어나 줄임말과 같은 철자 오류들을 포함하는 텍스트들이 활발히 통용되고 있다. 이러한 철자 오류들은 텍스트의 가독성을 떨어뜨림으로써 자연어처리 응용들을 개발하는데 걸림돌이 된다. 이러한 문제를 해결하기 위해서 본 논문에서는 철자오류 교정사전과 신문기사 말뭉치를 이용한 철자 오류 교정 모델을 제안한다. 제안 모델은 구하기 쉬운 신문기사 말뭉치를 학습 말뭉치로 사용하기 때문에 데이터 구축비용이 크지 않다는 장점이 있다. 또한 교정사전 기반의 단순 매칭 방법을 사용하기 때문에 띄어쓰기 교정 시스템이나 형태소 분석기와 같은 별도의 외부 모듈이 필요 없다는 장점이 있다. 신문기사 말뭉치와 실제 휴대폰에서 수집한 문자 메시지 말뭉치를 이용한 실험 결과, 제안 모델은 다양한 평가 척도에서 비교적 높은 성능(오교정률 7.3%, F1-척도 97.3%, 위양성율 1.1%)을 보였다.

키워드 : 철자 오류 교정 모델, 신문기사 말뭉치, 철자 오류 교정 사전

A Spelling Error Correction Model in Korean Using a Correction Dictionary and a Newspaper Corpus

Sehee Lee[†] · Harksoo Kim[††]

ABSTRACT

With the rapid evolution of the Internet and mobile environments, text including spelling errors such as newly-coined words and abbreviated words are widely used. These spelling errors make it difficult to develop NLP (natural language processing) applications because they decrease the readability of texts. To resolve this problem, we propose a spelling error correction model using a spelling error correction dictionary and a newspaper corpus. The proposed model has the advantage that the cost of data construction are not high because it uses a newspaper corpus, which we can easily obtain, as a training corpus. In addition, the proposed model has an advantage that additional external modules such as a morphological analyzer and a word-spacing error correction system are not required because it uses a simple string matching method based on a correction dictionary. In the experiments with a newspaper corpus and a short message corpus collected from real mobile phones, the proposed model has been shown good performances (a miss-correction rate of 7.3%, a F1-measure of 97.3%, and a false positive rate of 1.1%) in the various evaluation measures.

Keywords : Spelling Error Correction Model, Newspaper Corpus, Spelling Error Correction Dictionary

1. 서 론

인터넷, 메신저 또는 휴대폰 등을 사용하여 문자를 주고받을 때는 자신의 의도를 빠르게 표현해서 전달해야 하기 때문에 철자 오류가 발생하는 경우가 빈번하다[1]. 이런 철자 오류들은 타이핑(typing) 실수 또는 맞춤법 지식의 부재 때문에 발생하는 기본적 철자 오류와 일부러 맞춤법에 어긋나게 적는 의도적 철자 오류로 나눌 수 있다. 의도적 철자 오류는 구어체적 특성을 가지는 오류가 많으며 신조어, 비속어, 줄임말, 약어, 은어, 맞춤법 파괴 등이 섞여 발생한다. 또한 의도적 철자 오류는 유행 및 시기에 따라 변화하며 새로 생성되고 사라지는 특징을 가진다. 이와 같은 철자 오류를 포함한 문장은 형태소 분석을 기반으로 한 정보 추출 또는 정보 검색 등의 상위 자연어 처리 응용의 성능을 떨어뜨린다. 따라서 이런 철자 오류의 특성에 맞는 철자 오류 교정 시스템이 필요하다. 본 논문에서는 교정 사전과 신문기

※ 이 논문은 2008년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(KRF-2008-313-D00907).
† 준 회 원 : 강원대학교 컴퓨터정보통신공학전공 석사과정
†† 정 회 원 : 강원대학교 컴퓨터정보통신공학전공 교수
논문접수 : 2009년 6월 9일
수 정 일 : 1차 2009년 7월 14일, 2차 2009년 7월 20일
심사완료 : 2009년 7월 28일

사 말뭉치를 이용하여 통계적으로 철자 오류를 교정하는 시스템을 제안한다. 제안 시스템은 실용성을 높이기 위해서 철자 오류 문자열의 추가와 삭제가 쉬운 교정 사전을 기반으로 한다. 또한 효율적인 통계 학습을 위해서 철자 오류가 비교적 적고 수집이 용이한 신문기사를 학습 말뭉치로 사용한다. 사용자로부터 문장이 입력되면 제안 시스템은 단순한 패턴 매칭(pattern matching) 방법을 이용하여 교정 사전에 등록된 문자열을 철자 오류 후보 문자열로 분류한다. 그리고 신문기사를 이용하여 학습된 문맥정보를 기반으로 철자 오류 후보 문자열이 진짜 오류 문자열인지를 판단한다.

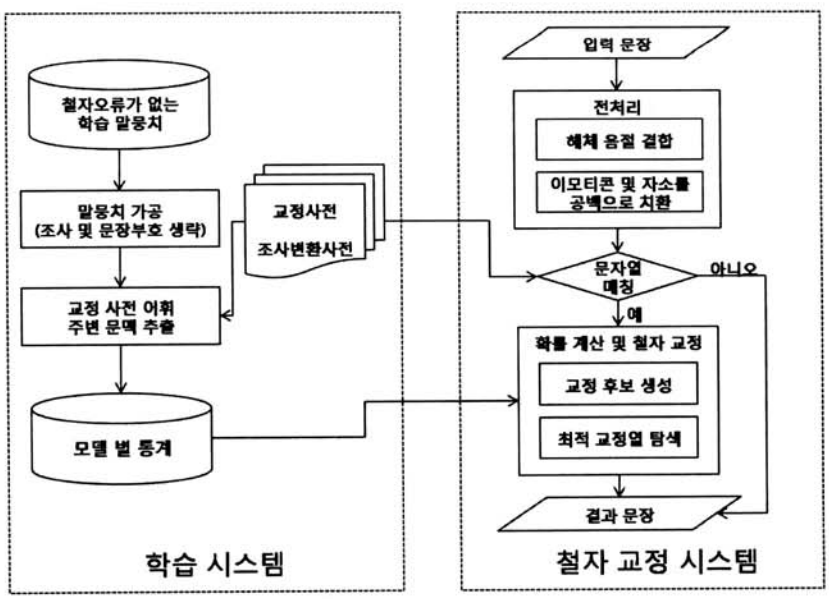
본 논문의 구성은 다음과 같다. 2장에서 철자 오류 교정에 관한 연구들을 살펴보고, 3장에서는 제안하는 철자 오류 교정 모델을 자세히 설명한다. 4장에서 실험 데이터 및 평가척도를 설명하고, 다양한 실험 결과를 보인다. 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

철자 오류에 관한 기존 연구로는 일반 사전을 사용한 방법[2-4], 형태소 분석결과를 이용한 방법[5], 자소 단위 철자 오류 교정 방법[6], 교정사전을 사용한 방법[7] 등이 있었다. 일반 사전을 사용한 철자 교정은 주로 영어권 국가들이 사용하는 방법으로 모든 입력문장의 각 단어들을 일반 사전에서 검색한 후, 일반 사전에 존재하지 않는 단어는 철자 오류라고 가정한다. 철자 오류로 가정된 문자열은 편집거리(edit-distance)[4], 메타폰(metaphone) 알고리즘[4] 등을 사용하여 해당 단어와 사전의 단어 중 거리가 가까운 교정 후보를 생성한다. 그리고 생성된 여러 교정 후보들 중 정답에 가장 가까운 단어를 선택하여 교정하는 방법이다. 이러한 일반 사전 기반의 방법은 모든 철자 오류 후보 단어들을 미리 구축해야 하는 단점이 있으며, 용언의 활용이 많은 한국어의 경우에는 어휘의 기본형을 찾는데 따르는 비용이 클 뿐만 아니라 잘못된 기본형을 찾을 가능성이 높다는 단점이 있다. 형태소 분석 결과를 이용한 방법은 형태소분석기에서 분석이 실패한 어절은 철자 오류가 포함된 오류어절로 판단한다. 오류어절에 대해 교정규칙정보와 말뭉치통계정보를 적용하여 철자 오류를 교정한다. 형태소 분석기를 사용하는 방법은 형태소 분석기의 성능에 의존적이며 철자 오류 교정 문제에 형태소 분석 문제가 추가되어 복잡도가 증가하는 단점이 있다. 자소 단위 철자 오류 교정 방법은 자소 단위의 변환확률을 이용한다. 자소 변환확률은 오류가 포함된 말뭉치와 그 말뭉치를 수작업으로 교정한 말뭉치를 자소단위로 일대일 대응시켜 계산한다. 자소 단위 철자 교정 방법은 오류가 포함된 말뭉치를 수작업으로 교정하여 병렬말뭉치를 생성해야 하는 단점이 있다. 교정사전을 사용한 방법은 단계별 사전과 형태소 분석 결과를 이용하기 때문에 위에서 언급한 단점들을 복합적으로 포함한다. 이러한 기존 철자오류 교정 모델들이 가지고 있는 문제점들을 해결하기 위해서 본 논문에서는 신문기사나 교과서와 같이 철자 오류가 거의 없는 말뭉치만을 이용하여 통계적으로 철자 오류를 교정하는 방법을 제안한다.

3. 철자 오류 교정 모델

(그림 1)에서 보는 것과 같이 전체 시스템은 학습 시스템과 철자 교정 시스템으로 나뉜다. 학습 시스템은 학습 말뭉치에서 교정사전에 있는 각 어휘에 대한 문맥을 추출한다. 철자 교정 시스템은 전처리 과정을 거친 입력 문장을 오류



(그림 1) 제안된 철자 오류 교정 시스템의 구조도

문자열과 부분 매칭(matching)을 시도한다. 매칭된 오류문자열이 발견되면 그에 대한 교정 후보를 생성한다. 그리고 학습 시스템에서 추출한 문맥 정보를 사용하여 현재 문맥에 가장 적합한 교정 후보를 선택함으로써 철자 오류를 교정한다.

3.1 교정 후보 생성

사용자가 문장을 입력하면 제안 시스템은 제일 먼저 입력문장에 대한 전처리를 수행한다. 제안 시스템은 전처리를 통해서 해체된 음절을 결합하고 자질로써 의미 없는 문자열이 문맥으로 사용되지 않도록 한다. 입력문장의 전처리는 2단계로 수행된다. 1단계로 해체된 음절을 결합하고, 2단계로 이모티콘과 모음이 생략된 문자열을 공백으로 대체한다. 해체된 음절의 예는 'ㅇㅏㅃㅏ', 'ㄱㅡㄹㅐ' 등이 있으며 규칙을 이용하여 '아빠', '그래'와 같이 교정한다[8]. 2단계에서는 '-_-', '^^;;'와 같은 이모티콘이나 'ㅋㅋㅋ', 'ㅎㅎ'와 같은 특수 문자열은 제안 시스템의 처리 대상이 아니기 때문에 공백으로 치환한다. 이러한 종류의 문자열을 단순히 제거하지 않고 공백으로 치환하는 이유는 문장 경계 또는 어절 경계 정보로 사용될 수 있기 때문이다. 공백으로 치환된 문자열들도 전달하고자 하는 의미가 있을 수 있기 때문에 철자 오류 교정 과정이 완전히 끝난 후 다시 복원한다.

전처리 과정이 끝나면 제안 시스템은 <표 1>과 같은 교정사전의 각 엔트리(entry)를 입력문장과 단순 매칭하는 방법을 이용하여 교정 후보 집합을 생성한다.

교정 후보 집합은 매칭된 오류문자열과 그것에 대한 교정문자열들의 모임을 말한다. 교정 후보를 생성할 때 오류문자열의 중의성이 없다면 교정문자열이 집합에 추가되고, 중의성이 있는 단어라면 교정사전에 수록된 오류문자열과 교정문자열이 집합에 추가된다. 오류문자열이 중의성을 가지고 있는지 여부는 대용량의 철자 오류가 없는 말뭉치에서 오류문자열이 출현했는지를 보고 판단한다. 철자 오류가 없는 말뭉치에서 한 번도 출현하지 않은 오류문자열은 올바른 표현이 아닐 가능성이 크기 때문에 본 논문에서는 이러한 문자열은 중의성이 없다고 가정한다. 반대로 중의성이 있는 오류문자열은 문맥에 따라 철자 오류가 아닐 가능성도 있기 때문에 오류문자열과 교정문자열 모두를 교정 후보 집합에 추가하는 것이다. 예를 들어 교정사전에 '남친->남자친구'가 있고, 학습 말뭉치에서 '남친'이 한 번도 출현하지 않았다면 시스템은 입력문장에서 중의성이 없는 오류문자열 '남친'에 대한 교정 후보로 '남자친구'만 생성한다. <표 1>의 교정사전에서 'ㄹ텨->ㄹ래'에서 'ㄹ'은 '텨'와 '래' 앞 음절의 종성을 의미하며, '지성ㅎ->죄송ㅎ'에서 'ㅎ'은 '지성'과 '죄송' 뒤 음절의 초성을 의미한다. '@'는 조사 접속 문제의 가능성이 있는 교정 후보를 나타낸다. 조사 접속 문제는 오류문자열과 교정문자열의 종성 정보가 다른 경우 발생한다. 예를 들어 '엉아->형'의 교정은 교정 후 문자열이 명사이며 '엉아'의 마지막 음절 '아'는 종성이 없지만 '형'은 종성이 존재하므로 조사 접속 문제가 발생한다. 반대로 '남친->남자친구'의 교정은 마지막 음절에 종성이 있는 '남친'을 마지막 음절에 종성이 없는 '남자친구'로 교정하기 때문에 조사 접속 문제가 발생한다. 조사 접속 문제로 문장이 어색해지는 예는 <표 2>의 1과 같다. 조사 접속 문제로 문장이 어색해 지는 것을 해결하기 위해 교정문자열 뒤에 오는 문자열이 조사와 매칭될 경우에 조사 변환 사전을 사용해 무조건 변환하면 <표 2>의 2번과 같은 문제가 발생한다. 이러한 조사 접속 문제를 해결하기 위해서는 문자열의 교정후보를 생성할 때 바로 뒤에 나타나는 문자열이 <표 3>과 같은 조사 변환 사전에 매칭되면 조사에 대한 교정후보도 생성한다.

교정 후보 집합 열의 생성 과정을 예를 들어 설명하면 (그림 2)와 같다.

(그림 2)에서는 교정사전에 '욕시->역시', '남친->남자친

〈표 1〉 교정사전의 예

오류문자열	교정문자열
햇으	했어
ㄹ텨	ㄹ래
께여	게요
ㅁ니다	ㅂ니다
지성ㅎ	죄송ㅎ
남친	@남자친구

〈표 2〉 조사 접속 문제의 예

번호	교정 전 문장	교정 후 문장
1	내 남친이 전화했어.	내 남자친구이 전화했어.
2	내 남친이순신이 전화했어.	내 남자친구가순신이 전화했어.

〈표 3〉 조사 변환 사전의 예

종성 있음	종성 없음
이	가
과	와
관	완
을	를

원본문장	뉴	욕	시	에	서	b	남	친	을	b	봐	ㅆ	ㅇ	ㅜ	면	b	좋	겠	지	만	ㅋ	ㅋ	욕	시	b	무	린	가	?
전처리문장	뉴	욕	시	에	서	b	남	친	을	b	봤		우		면	b	좋	겠	지	만	b		욕	시	b	무	린	가	?
교정 후보 집합 열	뉴	욕시 역시		에	서	b	남자친구		을 를	b	봤		우면 으면			b	좋	겠	지	만	b		욕시 역시		b	무	린	가	?

(그림 2) 교정 후보 집합 열의 생성 예

구', '우면->으면'이, 조사 변환 사전에는 '을<->를'이 등록되어 있다고 가정한다. 먼저 전처리를 통해 '봐ㅆ'이 '봤'으로, 'ㅇㅜ'이 '우'로 교정되며, 'ㅋㅋ'이 공백(b)으로 치환된다. 그리고 교정사전 매칭을 통해 '욕시'와 '우면'에 대한 교정 후보 '욕시, 역시'와 '우면, 으면'이 생성되고, '남친'에 대한 교정 후보 '남자친구'가 생성된다. '남친'이 교정 후보에 포함되지 않은 이유는 학습 말뭉치에서 출현하지 않아서 중의성이 없다고 판단되었기 때문이다. 다음으로 조사 변환 사전과의 매칭을 통해 '남친' 뒤의 조사 '을'에 대한 교정 후보 '을'과 '를'이 생성된다.

3.2 철자 오류 교정

교정 후보 집합 열이 생성되면 철자 오류 교정은 식 (1)과 같이 다수의 교정 후보 중에서 문맥에 맞는 최적의 교정 문자열을 찾는 문제로 해석될 수 있다.

$$M(S_{1...n}) = \arg\max_{C_{1...n}} P(C_{1...n}|S_{1...n}) \quad (1)$$

식 (1)에서 S_i는 i번째 교정 후보 집합을 나타내며, C_i는 S_i에 포함된 하나의 교정 후보이다. n은 생성된 교정 후보 집합의 수이다. 식 (1)을 예를 들어 도식적으로 표현하면 (그림 3)과 같다.

식 (1)을 만족하는 $C_{i...n}$를 구하기 위해 $P(C_{1...n}|S_{1...n})$을 베이즈 정리를 이용해 풀면 식 (2)와 같다[11].

$$\begin{aligned} M(S_{1...n}) &= \arg\max_{C_{1...n}} P(C_{1...n}|S_{1...n}) \quad (2)\\ &= \arg\max_{C_{1...n}} \frac{P(S_{1...n}|C_{1...n})P(C_{1...n})}{P(S_{1...n})} \\ &\approx \arg\max_{C_{1...n}} P(S_{1...n}|C_{1...n})P(C_{1...n}) \\ &\approx \arg\max_{C_{1...n}} \prod_{i=1}^{n} P(S_i|C_i)P(C_i) \end{aligned}$$

식 (2)에서 $P(C_i)$는 교정 후보인 C_i^{error}또는 $C_i^{correct}$가 나타날 확률로써 본 논문에서는 학습 말뭉치에서 후보 문자열이 출현한 수를 기초로 하여 식 (3)과 같이 계산한다.

$$P(C_i) \approx \frac{freq(C_i)}{freq(C_i^{error}) + freq(C_i^{correct})} \quad (3)$$

식 (3)에서 $freq(C_i^{error})$와 $freq(C_i^{correct})$는 i번째 교정 후보 집합에 포함된 오류문자열과 교정문자열이 학습 말뭉치에서 출현한 횟수를 각각 의미한다. 예를 들어서, (그림 2)의 2번째 교정 후보 집합에서 '욕시'가 나타날 확률은 'P(욕시)=freq(욕시)/(freq(욕시)+freq(역시))'이며, '역시'가 나타날 확률은 'P(역시)=freq(역시)/(freq(욕시)+freq(역시))'이다.

식 (2)의 $P(S_i|C_i)$는 교정 후보 집합에서 특정 교정 후보가 선택될 확률로써 본 논문에서는 식 (4)와 같이 주변 문맥을 이용하여 계산한다.

$$P(S_i|C_i) \approx \prod_{k=1}^{m} P(f_i^k|C_i) \quad (4)$$

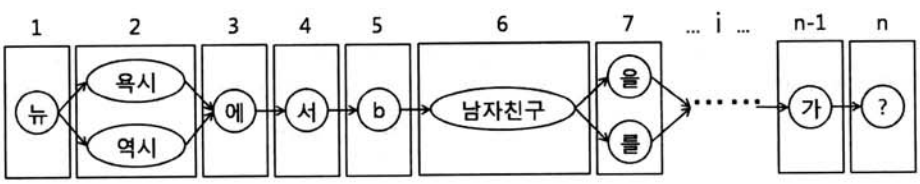
식 (4)에서 f_i^k는 C_i의 문맥으로 사용되는 k번째 자질이며, m은 사용된 자질의 수이다. 예를 들어서, 문맥으로 사용하는 자질이 교정 후보 좌우 1음절 이라면 m은 2가 되며, (그림 2)의 2번째 교정 후보 집합에서 '욕시'가 선택될 확률은 'P(S_2|욕시) = freq(뉴욕시)/freq(욕시)*freq(욕시에)/freq(욕시)'이다.

학습 시스템은 학습 말뭉치에서 교정사전의 문자열이 출현한 수와 식 (4)에 사용되는 자질이 출현한 수를 저장한다. 본 논문에서는 식 (4)에 사용된 자질에 따라 다양한 모델을 만들었으며, 각 모델이 사용하는 자질은 4.3절에서 자세히 설명한다. 확률 계산이 끝나면 비터비(Viterbi) 알고리즘[9]을 사용하여 최적의 교정 후보열을 결정한다.

4. 실험 결과 및 분석

4.1 실험 데이터

학습 말뭉치로는 1,705,935문장(18,529,683어절)의 조선일보 신문기사(이하 신문기사 말뭉치)를 사용하였다. 평가 말뭉치로는 100여명의 대학생 휴대폰에서 실제로 수집한 1,038개의 문자메시지(이하 휴대폰 평가 말뭉치)를 사용하였다. 실험의 신뢰성 확보를 위해서 신문기사 말뭉치에 10배 교차검증(10-fold cross validation)을 수행하여 올바른 문자열을 철자 오류로 잘못 판단하는 비율을 측정하였다. 휴대폰 평가 말뭉치는 통신어와 신조어를 포함하는 실제 철자 오류 문장에 대한 제안 모델의 성능을 평가하기 위해 구축한 것이다. 휴대폰 평가 말뭉치를 이용한 실험은 신문기사 말뭉치 전체를 학습하여 구축한 시스템을 이용하였다. 휴대폰



(그림 3) 교정 후보 집합 열의 예

〈표 4〉 휴대폰 평가 말뭉치의 구성

	오류문자열과 매칭된 문자열 수	실제 오류문자열의 수	부분 매칭된 옳은 문자열의 수
원본 문장	839	797(95.0%)	42(5.0%)
공백 제거 문장	839	797(95.0%)	42(5.0%)
띄어쓰기 교정 문장	828	787(95.0%)	41(5.0%)

평가 말뭉치는 <표 4>와 같이 3가지 형태(원본 문장, 공백을 모두 제거한 문장, 원본에서 띄어쓰기만 교정한 문장)로 가공하여 실험을 진행하였다. 원본 휴대폰 평가 말뭉치는 약 25Kbyte크기에 1,038개의 문자메시지로 구성되며 1,835개의 어절로 이루어졌다.

<표 4>에서 보는 것과 같이 교정사전의 오류문자열과 매칭된 문자열은 실제 오류문자열이거나 옳은 문자열이 부분 매칭된 경우로 나뉜다. 교정사전은 평가 말뭉치 수집과는 무관하게 별도로 구축하였으며, 통신언어 어휘집[10]에서 중의성이 많은 1음절 어휘와 적용 가능성이 매우 낮은 어휘들을 제거하는 방법으로 구축하였다. 또한 인터넷 검색과 대학생들에게 설문을 통해 자주 틀리는 어휘들을 평가 말뭉치 수집 이전에 수작업으로 추가하였다. 결론적으로 평가 말뭉치와는 완전히 독립적으로 1,990개의 오류문자열과 교정문자열의 쌍을 포함하는 교정사전을 구축하였다.

4.2 평가 방법

제안 모델을 평가하기 위해서 오교정률(miss-correction rate)과 이진 분류기 성능 측정 방법을 사용하였다. 신문기사 말뭉치는 일반적으로 철자 오류의 비율이 낮은 편에 속한다. 그러므로 교정사전에 등록된 오류 문자열이 신문기사 말뭉치에서 나타났다면 옳은 문자열이 부분 매칭된 것일 가능성이 높다. 이러한 가정을 뒷받침하기 위해서 신문기사 말뭉치와 매칭된 문자열의 철자오류 여부를 판단하는 실험을 수행했으며, 그 결과 5%만이 실제 철자 오류 문자열이었다. 본 논문에서는 상기한 가정에 기초하여 제안 시스템이 옳은 문자열을 틀리게 교정하는 비율인 오교정률을 계산한다. 오교정률은 신문기사 말뭉치에 식 (5)를 사용하여 10배 교차 검증으로 계산한다. 신문기사 말뭉치에 철자 오류가 5% 포함되어 있지만 식 (5)의 분모와 분자에 오류의 영향이 모두 미치기 때문에 실제 오교정률과 계산된 오교정률 사이에 심각할 정도의 큰 차이는 없을 것으로 생각된다.

$$\text{오교정률} = \frac{\text{제안시스템이교정한문자열의수}}{\text{테스트데이터와교정사전내오류문자열이매칭된수}} \quad (5)$$

이진 분류기 성능 측정 방법은 <표 5>를 기반으로 식 (6)~(9)를 계산하는 것이다. 본 논문에서는 휴대폰 평가 말뭉치에 제안 시스템을 적용시켜 <표 5>를 얻고, 이를 기반으로 식 (6)~(9)와 같은 다양한 척도를 이용하여 제안 모델의 성능 측정을 하였다.

〈표 5〉 이진 분류기 성능 측정 방법

시스템 \ 정답	철자 오류	정상
철자 오류	true positive	false positive
정상	false negative	true negative

$$\text{정확률} = \frac{true\,positive}{true\,positive + false\,positive} \quad (6)$$

$$\text{재현율} = \frac{true\,positive}{true\,positive + false\,negative} \quad (7)$$

$$F1-\text{척도}(F1-measure) = \frac{2*\text{정확률}*\text{재현율}}{\text{정확률}+\text{재현율}} \quad (8)$$

$$\text{위양성률}(false-positive\ rate) = \frac{false\,positive}{true\ positive + false\,positive} \quad (9)$$

4.3 실험 모델

실험은 (그림 4)에서 보는 것과 같이 제안 시스템이 사용하는 문맥 자질에 따라 35개의 서로 다른 모델을 구축하여 진행하였다. 문맥 자질로는 공백정보와 좌우 음절 정보를 사용하였다. (그림 4)에서 '±n'은 교정 후보의 좌(-), 우(+) n번째 음절을 의미한다. 두 셀(cell)이 합쳐진 표시는 음절 바이그램(bigram) 자질을 의미한다. 'O'는 해당 자질을 사용한다는 의미이며, 'X'는 해당 자질을 사용하지 않는다는 의미

모델명	좌측 음절 자질			우측 음절 자질			모델명	좌측 음절 자질			우측 음절 자질		
	-2	-1	공백	공백	+1	+2		-2	-1	공백	공백	+1	+2
A	X	O	O	O	O	X	A"	X	O	X	O	O	X
B	O	O	O	O	O	X	B"	O	O	X	O	O	X
C	X	O	O	O	O	O	C"	X	O	X	O	O	O
D	O	O	O	O	O	O	D"	O	O	X	O	O	O
E	O		O	O	O	X	E"	O		X	O	O	X
F	X	O	O	O	O		F"	X	O	X	O	O	
G	O		O	O	O		G"	O		X	O	O	
A'	X	O	O	X	O	X	H	X	O	X	X	O	X
B'	O	O	O	X	O	X	I	O	O	X	X	O	X
C'	X	O	O	X	O	O	J	X	O	X	X	O	O
D'	O	O	O	X	O	O	K	O	O	X	X	O	O
E'	O		O	X	O	X	L	O		X	X	O	X
F'	X	O	O	X	O		M	X	O	X	X	O	
G'	O		O	X	O		N	O		X	X	O	
							O	X	O	△	△	O	X
							P	O	O	△	△	O	X
							Q	X	O	△	△	O	O
							R	O	O	△	△	O	O
							S	O		△	△	O	X
							T	X	O	△	△	O	
							U	O		△	△	O	

(그림 4) 각 모델별 자질 설명

이다. '△'는 '±n' 음절을 계산할 때 공백을 포함시킨다는 의미이다. 모델 O에서 U를 제외한 나머지 28개 모델은 교정 후보 좌우에 공백이 있을 경우, 공백의 유무 정보만 사용하고 공백을 무시한 상태에서 좌우 음절을 추출한다.

(그림 4)의 모델들과는 별도로 기저모델(baseline model)을 만들어서 최저 기본 성능을 계산하고 제안 모델들과 비교하였다. 기저모델은 중의성이 있는 오류문자열을 철자 오류 판단하는데 학습 말뭉치 내에서 출현 빈도를 사용한다. 즉, 학습 말뭉치에서 출현 빈도가 높은 어휘를 교정 대상으로 선택한다. 예를 들어 교정사전에 '남친->남자친구'가 있고, 학습 말뭉치에서 '남친'이 20번, '남자친구'가 100번 출현했다면 '남자친구'가 '남친'보다 더 많이 출현했으므로 '남친'이라는 어휘는 '남자친구'로 무조건 변환하는 것이다.

4.4 실험 결과 및 분석

<표 6>은 신문기사 말뭉치와 휴대폰 평가 말뭉치(원본)에 대한 실험 결과의 일부를 보여준다. 나머지 모델들은 성능 차이가 크게 없어서 생략하였다. 오교정률은 신문기사 말뭉치에 대한 실험 결과이고, 이진 분류기의 성능 측정 방법을 사용한 결과는 휴대폰 평가 말뭉치를 사용한 실험 결과이다.

<표 6>에서 X는 기저모델을 의미한다. 기저모델은 교정사전에 매칭된 문자열 111,699개 중 74,882개를 수정함으로써 67.0%라는 매우 높은 오교정률을 보였다. 오교정률이 가장 낮은 모델은 교정 후보 문자열 좌우 음절 바이그램에 양쪽 공백정보를 모두 사용한 '모델 G'였다. 단문메시지에 대한 실험 결과에 따르면 정확률과 위양성률은 '모델C"가 가장 좋았고, 재현율은 '모델K'가 가장 좋았으며, F1-척도는 '모델 J'가 가장 좋았다.

문자메시지를 포함해서 근래에 많이 통용되는 통신어들은 일반적으로 띄어쓰기 오류를 많이 포함하고 있기 때문에 철자 오류 교정 모델도 띄어쓰기 오류에 강건해야한다. 띄어쓰기 오류에 대한 강건성 실험을 위해서 <표 7>에서 보는 것과 같이 음절자질 추출 방법이 서로 다른 모델 중 가장 성능이 좋은 '모델 J' 와 '모델 Q'를 대상으로 실험을 수행하였다.

<표 7>의 표준편차를 살펴보면 '모델 J'가 '모델 Q'보다 띄어쓰기 오류에 강건하다는 것을 알 수 있었다. 그 이유는 음절자질 추출 방법에서 공백은 추출하지 않고 공백 이후에 나타나는 음절을 추출하였기 때문에 띄어쓰기 오류에 영향을 덜 받는 것으로 생각된다. '모델 Q'가 띄어쓰기를 교정한 데이터에서 성능이 더 높은 것도 음절자질 추출 방법에서 공백을 포함하였기 때문으로 보인다.

〈표 6〉 모델별 성능 비교

모델	오교정률 (%)	정확률 (%)	재현율 (%)	F1-measure(%)	위양성율 (%)
X	67.0	95.7	92.9	94.3	4.2
A	4.6	98.7	92.1	95.3	1.2
B	5.3	98.0	92.7	95.2	1.9
C	5.2	98.8	93.7	96.2	1.1
D	5.5	98.0	94.3	96.1	1.9
E	4.8	98.3	90.8	94.4	1.6
F	4.1	98.5	90.4	94.3	1.5
G	3.6	98.0	88.5	93.0	1.9
A'	6.1	98.8	93.6	96.1	1.1
B'	6.8	98.2	93.9	96.0	1.7
C'	6.8	98.9	94.6	96.7	1.0
D'	7.1	98.1	95.1	96.6	1.8
I	7.2	98.0	96.2	97.1	1.9
J	7.3	98.8	95.9	97.3	1.1
K	7.5	98.2	96.4	97.2	1.7

제안 모델의 성능을 비교하기 위해서 강승식(2008)[7]과 성능을 비교하였다. 강승식(2008)의 재현율은 옳게 교정한 문자열의 수를 실험 말뭉치에 존재하는 전체 오류 문자열의 수로 나누어 계산하였다. 본 논문의 재현율은 철자 오류 평가 대상이 되는 문자열만을 대상으로 하기 때문에 강승식(2008)의 재현율과는 차이가 있다. 정확한 비교를 위해 강승식(2008)의 방법으로 본 논문에서 가장 좋은 성능을 보인 '모델 J'의 성능을 계산하면 재현율이 75.7%였고 정확률이 98.8%였다. 강승식(2008)에 발표된 재현율은 57.0%였고, 정확률은 91.3%이었다. 수치상으로는 제안 모델이 강승식(2008)보다 좋은 성능을 보이지만 철자 오류 교정 방법, 실험 말뭉치 등이 서로 다르기 때문에 단순히 비교하기는 어려울 것으로 보인다. 그러나 실용적인 관점에서 비교해 보면 제안 모델이 교정사전에 존재하지 않는 어휘는 철자 오류로 인식하지 못한다는 단점이 있지만 강승식(2008)보다 다음과 같은 장점이 있을 것으로 생각된다. 첫째, 새로운 어휘의 추가 및 삭제가 쉽다. 둘째, 교정사전의 오교정률을 학습 데이터를 통해 측정할 수 있으므로 부작용(side effect)이 큰 어휘를 사전에 인지하고 대응할 수 있다. 셋째, 형태소분석기와 띄어쓰기 교정시스템과 같은 별도의 외부 모듈을 사용할 필요가 없다.

학습 말뭉치에 한 번도 출현하지 않은 어휘는 중의성이

〈표 7〉 띄어쓰기 오류 강건성 비교

모델	F1-척도 (%)			표준편차
	원본	공백 제거	띄어쓰기 교정	
J	97.3	97.3	97.1	0.12
Q	96.9	96.8	97.6	0.44

〈표 8〉 중의성 없는 교정어휘에 대한 성능 평가

	교정사전 매칭 수	무조건 변환 문자열 수	맞은 수	틀린 수
원본 문장	839	326	326	0
공백 제거	839	326	326	0
띄어쓰기 교정	828	321	321	0

〈표 9〉 오교정 유형

오류 유형	단순 오류	인명	지명	그 밖의 고유명사	띄어쓰기 오류	실험 말뭉치 오류
비율 (%)	51.0	16.0	14.0	9.5	4.5	5.0

없는 철자 오류라는 가정을 확인하기 위해서 휴대폰 평가 말뭉치에서 중의성 없이 무조건 교정하는 문자열의 비율을 <표 8>과 같이 살펴보았다. 학습 말뭉치를 통해 중의성이 없다고 판단되어 무조건 교정하는 문자열은 단문메시지 데이터에 포함된 철자 오류를 100% 옳게 교정하였다. 교정사전 전체에서 중의성 없는 교정어휘 쌍은 1,489개였으며, 이것은 전체 교정사전의 74.8%에 해당되는 숫자였다.

철자 오류 교정 시스템의 성능에 심각한 영향을 끼칠 수 있는 오교정 유형을 분석한 결과는 <표 9>과 같다. <표 9>에서 보는 것과 같이 전체 오류 중에 인명, 지명 등의 고유명사에 의해 발생하는 오류는 39.5%를 차지했으며, 띄어쓰기 오류 때문에 발생하는 오류는 4.5%였다. 신문기사 말뭉치에 오류문자열이 포함된 경우도 있었는데 그 비율은 5.0%였다. 인명 오류의 예를 살펴보면 "그런 도다가 발탁한 인물이 이케다."는 문장에서 '이케->이렇게'로 교정이 이루어진 경우가 있었다. 지명의 경우에는 "부산시 사하구 안락동"이라는 문장에서 '하구->하고'로 교정된 경우가 있었다. 잘못된 띄어쓰기 때문에 발생한 오류의 예는 "그지역 청소년에게" 라는 문장에서 '그지->그렇지'로 교정된 경우가 있었다.

5. 결 론

본 논문에서는 교정사전과 철자 오류가 없는 말뭉치만을 이용하여 자동으로 철자 오류를 교정하는 방법을 제안하였다. 교정 후보 문자열 주변 자질을 사용하는 방법에 따라 여러 모델을 만들고, 신문기사 말뭉치와 휴대폰 평가 말뭉치를 사용하여 다양한 성능 측정 방법으로 제안시스템을 평가하였다. 실험 결과에 따르면 교정문자열 앞 1음절과 뒤 2음절 각각을 문맥 자질로 사용하는 '모델 J'가 가장 좋은 성능을 보였으며, F1-척도는 97.3%였고 오교정률은 7.3%였다.

참 고 문 헌

[1] 조동욱, 이현경, "인터넷 상에서 쓰이는 통신 언어에 대한 분석 및 문제점 해결 방안", 한국콘텐츠학회/한국통신학회 2003 추계 종합학술대회 논문집, 제1권 제2호, pp.79-83, 2003.

[2] 노형종, 차정원, 이근배, "띄어쓰기 및 철자 오류 동시교정을 위한 통계적 모델", 정보과학회논문지: 소프트웨어 및 응용, 제34권 제2호, pp.131-139, 2007.

[3] Stéphanie Jacquemont, Francois Jacquenet, Marc Sebban "Correct your text with Google", 2007 IEEE/WIC/ACM International Conference on Web Intelligence, pp.170-176, 2007.

[4] Johannes Schaback "Multi-Level Feature Extraction for Spelling Correction", IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data, pp.78-86, 2007.

[5] Eric Brill, Robert C. Moore, "An Improved Error Model for Noisy Channel Spelling Correction", In Proc. of the 38th Annual Meeting of the ACL, pp.286-293, 2000.

[6] 윤근수, 권혁철, "교정률 최적화를 위한 한국어 철자교정기의 모듈 배열", 정보과학회논문지: 소프트웨어 및 응용, 제32권 제5호, pp.366-377, 2005.

[7] 강승식, 장두성, "SMS 변형된 문자열의 자동 오류 교정 시스템", 정보과학회논문지: 소프트웨어 및 응용, 제35권 제6호, pp.386-391, 2008.

[8] 이주호, 김학수, "2단계 규칙을 이용한 해체된 한글 음절의 결합", 인지과학, 제19권 제3호, pp.283-295, 2008.

[9] H. L. Lou, "Implementing the Viterbi Algorithm, Fundamental and real-time issues for processor designers", IEEE Signal Processing Magazine, pp.42-52, 1995.

[10] 조오현, 김경용, 박동근, "통신언어의 실태와 개선 방안", 통신언어 어휘집, 문화관광부, 2001.

[11] 김현준, 정재은, 조근식, "가중치가 부여된 베이지안 분류자를 이용한 스팸 메일 필터링 시스템", 정보과학회논문지: 소프트웨어 및 응용, 제31권 제8호, pp.1092-1100, 2004.

이 세 희

e-mail : nlpshlee@kangwon.ac.kr

2008년 강원대학교 컴퓨터학부(공학사)

2008년~현 재 강원대학교 컴퓨터정보통신공학전공 석사과정

관심분야 : 철자 오류 교정, 기계학습, 정보검색

김 학 수

e-mail : nlpdrkim@kangwon.ac.kr

1996년 건국대학교 전자계산학과(공학사)

1998년 서강대학교 컴퓨터학과(공학석사)

2003년 서강대학교 컴퓨터학과(공학박사)

2004년~2005년 CIIR in UMass, Amherst (박사후연구원)

2005년~2006년 한국전자통신연구원(선임연구원)

2006년~현 재 강원대학교 컴퓨터정보통신공학전공 교수

관심분야 : 자연어처리, 대화시스템. 정보검색, 질의응답시스템