나이브 베이즈 분류기와 혼동 행렬을 이용한 OCR에서의 철자 교정

노경목[†○], 김창현[‡], 천민아[†] 김재훈[†]
[†]한국해양대학교 컴퓨터공학과
[‡]한국전자통신연구원
kmq7542@gmail.com, chkim@etri.re.kr, minah0218@kmou.ac.kr, jhoon@kmou.ac.kr

Using Naïve Bayes Classifier and Confusion Matrix Spelling Correction in OCR

Kyung-Mok Noh[†○], Chang-Hyun Kim[‡], Min-Ah Cheon[†], Jae-Hoon Kim[†]
[†]Department of Computer Engineering, Korea Maritime and Ocean University
[‡]Electronics and Telecommunications Research Institute

요 약

OCR(Optical Character Recognition)의 오류를 줄이기 위해 본 논문에서는 교정 어휘 쌍의 혼동 행렬(confusion matrix)과 나이브 베이즈 분류기(naïve Bayes classifier)를 이용한 철자 교정 시스템을 제안한다. 본 시스템에서는 철자 오류 중 한글에 대한 철자 오류만을 교정하였다. 실험에 사용된 말뭉치는 한국어 원시 말뭉치와 OCR 출력 말뭉치, OCR 정답 말뭉치이다. 한국어 원시 말뭉치로부터 자소 단위의 언어 모델(language model)과 교정 후보 검색을 위한 접두사 말뭉치를 구축했고, OCR 출력 말뭉치와 OCR 정답 말뭉치로부터 교정 어휘 쌍을 추출하고, 자소 단위로 분해하여 혼동 행렬을 만들고, 이를 이용하여 오류 모델(error model)을 구축했다. 접두사 말뭉치를 이용해서 교정 후보를 찾고 나이브 베이즈 분류기를 통해 확률이 높은 교정 후보 n개를 제시하였다. 후보 n개 내에 정답 어절이 있다면 교정을 성공하였다고 판단했고, 그 결과 약 97.73%의 인식률을 가지는 OCR에서, 3개의 교정 후보를 제시하였을 때, 약 0.28% 향상된 98.01%의 인식률을 보였다. 이는 한글에 대한 오류를 교정했을 때이며, 향후 특수 문자와 숫자 등을 복합적으로 처리하여 교정을 시도한다면 더 나은 결과를 보여줄 것이라 기대한다.

주제어: 광학 문자 인식, 혼동 행렬, 나이브 베이즈 분류기, 철자 교정

1. 서론

OCR(Optical Character Recognition)은 이미지를 문자로 인식하는 기술이다. 필기체, 명함, 차량번호판, 문서 인식 등 널리 이용되고 있으며, 문서 인식은 정보의 전산화 측면에서 노동력과 비용을 줄이기 위해 꼭 필요한 기술이다. 현재 OCR은 높은 인식률을 보여주지만 여전히 오류가 존재한다. 이러한 오류를 줄이기 위한 여러 가지 후처리 기법이 연구되었다[1,2].

본 논문에서는 OCR의 인식률을 향상시키기 위한 철자 교정 시스템을 제안한다. 제안된 교정 알고리즘은 [3]의 철자 교정 알고리즘을 기반으로 구성하였으며, 특수 문자, 알파벳, 한자 등의 철자 오류는 제외하고 한글 오류에 대해서만 교정하였다.

제안된 철자 교정 시스템에 사용된 말뭉치는 약 530만 개의 어절로 구성된 한국어 원시 말뭉치와 약 7만 6천개의 어절로 구성된 OCR 정답 말뭉치, 오류가 있는 약 8만 1천 개의 어절로 구성된 OCR 출력 말뭉치를 이용하여 접두사 말뭉치와 혼동 행렬(confusion matrix)을 구축하고 나이브 베이즈 분류기(naïve Bayes classifier)를 통해 오류 어절에 대한 교정 후보를 제시하였다[3].

본 논문의 구성은 다음과 같다. 2장에서는 철자 교정 시스템의 구성과 교정 방법을 설명한다. 3장에서는 실험 방법과 성능을 평가하고, 4장에서는 결론을 맺는다.

2. 제안된 철자 교정 시스템의 구성

제안된 철자 교정 시스템은 크게 학습 단계와 교정 단계로 나누어지며, 시스템의 구성도는 그림 1과 같다.

2.1 학습 단계

학습 단계에서는 약 530만 개의 어절로 이루어진 한국어 원시 말뭉치에서 어절 단위로 한글을 추출한 뒤, 초

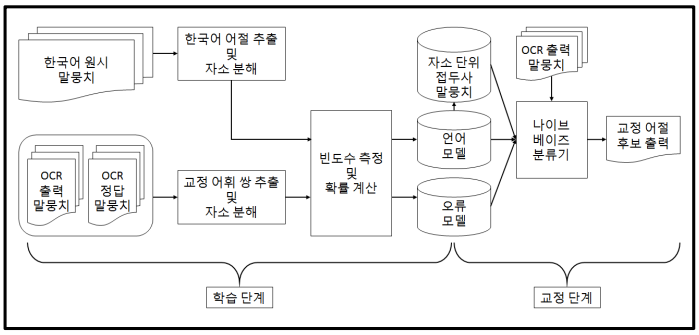

그림 1. 철자 교정 시스템의 구성도

성, 중성, 종성으로 분해하고, 분해된 어절의 빈도수를 측정한 후 확률을 계산하여 언어 모델(language model)을 구축하였다. 구축된 언어 모델을 기반으로 접두사 말뭉치를 구축하였는데 접두사 말뭉치는 교정단계에서 교정 후보를 검색하기 위해 사용된다. 한국어 원시 말뭉치에서 어절의 종류는 약 88만 개가 있으며, 이를 통해서 약 318만 개의 접두사가 있는 말뭉치를 구축하였다.

약 8만 1천 개의 어절로 이루어진 OCR 출력 말뭉치는 띄어쓰기 오류와 철자 오류, 띄어쓰기와 철자의 복합 오류가 포함된 말뭉치이며, 약 7만 6천 개의 어절로 이루어진 OCR 정답 말뭉치는 OCR 출력 말뭉치를 사용자가 수동으로 오류를 교정한 말뭉치이다. 두 말뭉치를 비교하여 교정 어휘 쌍을 추출한 뒤, 자소 분해하여 혼동 행렬을 구축하였다. 표 1는 혼동 행렬의 일부이다.

표 1. 자소 단위의 혼동 행렬 일부

오류 자소¦정답 자소	빈도수
ㅜ¦ㅡ	36
ㅛ¦ㅛ.	31
ㄹ¦ㅌ	19
ㅗ,¦ㅗ	7
ㅓ¦ㅣ	4
ㅟ¦ㅜ	3

예를 들어, “붙들어”가 “불들어”로 오인식되었을 경우, 이를 자소 단위로 분해해서 혼동 행렬을 만들면 “ㄹ¦ㅌ”이 된다. “ㅛ¦ㅛ.”는 “ㅛ.”가 “ㅛ”로 오인식된 것으로 “.”가 삽입된 것이고, “ㅗ,¦ㅗ”는 “ㅗ”가 “ㅗ,”로 오인식된 것으로 “,”가 삭제된 것이지만, 일부는 사용자가 수동으로 정답을 교정하면서 실수한 것도 존재했다.

이러한 혼동 행렬을 이용하여 언어 모델과 같은 방법으로 오류 모델을 구축하였다.

2.2 교정 단계

교정 단계에서는 교정 후보를 찾기 위해 오류 어절을 어두와 어미로 분할한다. 어두와 어미 사이에서 발생하는 자소의 삽입, 삭제, 대치에 대해서 편집거리 1에 해당하는 모든 교정 후보 중에서, 접두사 말뭉치에 존재하는 어절만을 교정 후보로 고려한다. 이를 재귀적으로 반복하여 편집거리 2까지의 교정 후보를 찾고, 교정 후보에 대한 교정 어휘 쌍의 확률과 교정 후보 단어의 확률을 나이브 베이즈 분류기를 통해서 확률이 가장 높은 후보를 출력한다.

또한, 오류 어절은 특수 문자, 숫자, 알파벳 등이 존재하는데 철자 교정은 한글만 하였으며, 한글이 아닌 문자는 교정 전에 제외하였다가 교정 후 다시 포함하여 최종적으로 정답 어절과 일치한 어절만을 교정 성공으로 판단했다. 예를 들어, 정답 어절 “웬일이냐?”가 “웬일이나?”라고 오인식되었을 경우 “?”를 제외한 “웬일이나”를 교정하여 “웬일이냐”를 교정 후보로 찾고, 여기에 “?”를 다시 포함하여 “웬일이냐?”를 최종 교정 후보로 출력하였다.

3. 실험 방법과 결과

제안된 철자 교정 시스템의 성능을 평가하기 위해서 기존의 한국어 원시 말뭉치를 확장하였다. OCR 정답 말뭉치에서 80%를 한국어 원시 말뭉치에 추가하여 언어 모델과 접두사 말뭉치를 구축했다. 그리고 말뭉치 확장에 사용된 OCR 정답 말뭉치 80%와 OCR 출력 말뭉치 80%를 이용하여 혼동행렬을 만들고, 오류 모델을 구축했다. 성능 평가에는 OCR 출력 말뭉치의 나머지 20%를 사용하였다. 철자 교정 오류만을 평가하기 위해서 OCR 출력 말뭉치의 20%에서 띄어쓰기 오류만 수동으로 교정했다.

철자 오류만 있는 OCR 출력 말뭉치의 20%는 약 1만 5천 개의 어절을 가지며, 인식률은 약 97.73%였다. 철자 오류 어절은 345개가 있었다. 345개의 오류 어절에서 확률이 높은 n개의 교정 후보군을 제시하고, 정답 어절이 후보군에 속할 경우, 교정을 성공하였다고 판단할 때, 교정 후의 OCR의 인식률은 그림 2와 같다. 그림 2를 보면 교정 후보를 많이 출력하지 않고 3~4개만 출력해도 교정 후보를 잘 찾아내는 것을 알 수 있다.

교정 후에도 정답 어절을 찾지 못한 어절들에 대해서 오류의 유형을 파악해 본 결과, 표 2와 같다. 표 2는 교정 후보 수를 3으로 했을 때의 결과이다.

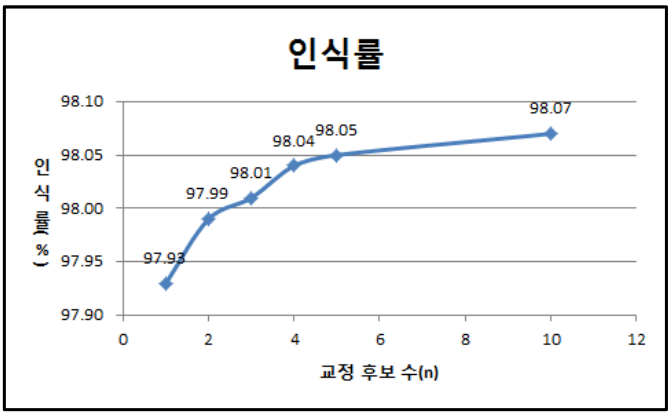

그림 2. 인식률 그래프

표 2. 교정 후보 3의 실패 유형

오류 유형		수
①삽입, 삭제 오류		5
대치 오류	②한글 외 문자만 틀린 경우	207
	③자소만 바뀐 경우	42
	④문자가 변형된 경우	49
총 오류 수		303

표 2를 살펴보면, ①문자가 삽입되거나 삭제된 경우는 많지 않았는데, 스캔 이미지에 잡음(noise)이 있거나 OCR이 어절이나 문자를 미인식한 경우였다. ②한글은 일치했으나 특수 문자, 숫자, 알파벳 등의 이유로 교정 실패한 경우는 상당히 많았다. 예를 들면 반각 기호 콤마 “,”가 전각 기호 콤마 “，”로 인식 되거나, “?”문자가 “T”로 인식된 경우이다. ③자소만 바뀐 경우는 “손뼉이라도”가 “손떡이라도”, “올라가마”가 “을라가마”등으로 인식된 경우이고, 이는 원시 한국어 말뭉치에는 없으나 문서 내에서만 존재하는 개체명, 대명사, 구어체 등이 있었다. 교정 후보를 찾았으나 혼동 행렬을 이용한 오류 모델에서 확률값이 낮아 후보 순위가 낮은 어절들도 있었다. ④문자가 변형된 경우는 문자가 한글이 아닌 다른 문자로 변형된 경우이다. 예를 들면, “아들”이 “of들”, “아낀다고”가 “0}낀다고”, “싫다고!”가 “싫다괴”등으로 인식된 경우이다.

4. 결론

OCR에서 발생한 오류 어절에 대해서 띄어쓰기 오류를 제외하고 철자 오류만을 교정하는 시스템을 구현하였다. 그 중에서도 알파벳과 숫자, 특수 문자를 제외한 한글에 대해서만 교정을 시도하였으며, 3개의 교정 후보 수를 기준으로 약 0.28%의 성능 향상을 보였다. 이는 한글만 처리하였을 때의 결과이며, 오류 유형에 따라 원시 말뭉치의 확장, 문서 내에서만 발생하는 어절, 특수 문자와 숫자 등을 복합적으로 처리한다면 좀 더 나은 성능이 나오리라 기대한다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업[R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발]의 일환으로 수행하였고 OCR 말뭉치를 제공해주신 국립중앙도서관에 감사드립니다.

참고문헌

[1] 이영화, 김계성, 김영훈, 이상조, “문자 인식 후처리를 위한 오인식 단어 검출 및 교정기의 구현”, 한국정보과학회 학술발표논문집, 24(1), pp. 517-520, 1997.

[2] 손훈석, 최성필, 권혁철, “문자 인식기의 특성과 말뭉치의 통계 정보를 이용한 문자 인식 결과의 후처리”, 제9회 한글 및 한국어 정보처리 학술대회 발표논문집, pp. 188-193, 1997.

[3] P. Norvig, Natural Language Corpus Data: Beautiful Data, pp. 219-242, 2009.