동적 프로그래밍을 이용한 OCR에서의 띄어쓰기 교정

박호민†○, 김창현‡, 노경목†, 천민아†, 김재훈†
†한국해양대학교, 컴퓨터정보공학과
‡한국전자통신연구원
homin2006@hanmail.net, chkim@etri.re.kr, kmq7542@gmail.com, minah0218@kmou.ac.kr, jhoon@kmou.ac.kr

Using Dynamic Programming for Word Segmentation in OCR

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

요 약

광학 문자 인식(OCR)을 통해 문서의 글자를 인식할 때 띄어쓰기 오류가 발생한다. 본 논문에서는 이를 해결하기 위해 OCR의 후처리 과정으로 동적 프로그래밍을 이용한 분절(Segmentation) 방식의 띄어쓰기 오류 교정 시스템을 제안한다. 제안하는 시스템의 띄어쓰기 오류 교정 과정은 다음과 같다. 첫째, 띄어쓰기 오류가 있다고 분류된 어절 내의 공백을 모두 제거한다. 둘째, 공백이 제거된 문자열을 동적 프로그래밍을 이용한 분절로 입력 문자열에 대하여 가능한 모든 띄어쓰기 후보들을 찾는다. 셋째, 뉴스 기사 말뭉치와 그 말뭉치에 기반을 둔 띄어쓰기 확률 모델을 참조하여 각 후보의 띄어쓰기 확률을 계산한다. 마지막으로 띄어쓰기 후보들 중 확률이 가장 높은 후보를 교정 결과로 제시한다. 본 논문에서 제안하는 시스템을 이용하여 OCR의 띄어쓰기 오류를 해결할 수 있었다. 향후 띄어쓰기 오류 교정에 필요한 언어 규칙 등을 시스템에 추가한 띄어쓰기 교정시스템을 통하여 OCR의 최종적인 인식률을 향상에 대해 연구할 예정이다.

주제어: 광학 문자 인식, 띄어쓰기 오류 교정, 분절, 동적 프로그래밍

1. 서론

현대 사회의 정보화 및 컴퓨터 산업의 급속한 발달로 사회분야 여러 곳에 정보 처리 자동화의 바람이 불고 있다. 그에 따라 새로이 생산되는 많은 양의 정보와 기존에 존재했던 문서들을 적절히 처리하기 위한 문자 인식 시스템의 필요성이 대두하고 있다[1]. 대표적인 문자 인식 방법으로 광학 문자 인식(OCR)이 있다. 광학 문자 인식은 사람이 적거나 기계가 인쇄한 아날로그 데이터의 영상을 통해 컴퓨터가 이해할 수 있는 디지털 데이터로 변환하는 것을 말한다[2]. 이 방법은 영상이 선명하지 않은 경우, 기계가 인식하기 힘든 악필, 글자가 잘못 인쇄된 서류에서 문자 인식을 할 때 오류가 발생한다[3]. 발생하는 오류의 종류로는 띄어쓰기 오류와 철자 오류가 있다. 본 논문에서는 띄어쓰기 오류에 대해서만 다룬다. OCR에서의 띄어쓰기 오류는 자동 줄 바꿈(soft return), 강제 줄 바꿈(hard return)에서 주로 발생한다. 이는 아날로그 데이터에서 디지털 데이터로의 변환 과정에서 생기는 잡음 때문이다.

이와 같은 OCR에서의 띄어쓰기 오류를 교정하기 위해 본 논문에서는 동적 프로그래밍을 이용한 확률 기반 방식 띄어쓰기 오류 교정 시스템을 제안한다. 제안하는 시스템은 OCR의 인식 과정에서 오류를 교정하는 것이 아니라, 인식 후의 후처리 과정에서 얻은 단서를 바탕으로 띄어쓰기 오류를 교정한다.

본 논문의 구성은 다음과 같다. 2장에서 제안하는 동적 프로그래밍을 이용한 띄어쓰기 오류 교정 방식에 대하여 소개하고, 3장에서는 결론 및 향후 연구에 대해 기술한다.

2. 동적 프로그래밍을 이용한 띄어쓰기 오류 교정

2.1 동적 프로그래밍

동적 프로그래밍은 주로 수학과 컴퓨터 공학에서 복잡한 문제를 단순한 여러 작은 문제로 분할하여 해결하는 방법이다[4]. 작은 문제들의 계산 결과를 저장해 놓은 뒤, 이후에 같은 문제를 풀게 되면 저장된 결과를 사용하기 때문에 계산 횟수를 줄일 수 있다.

본 논문에서 제안하는 띄어쓰기 오류 교정 시스템에서는, 띄어쓰기 오류를 교정하기 위해 뉴스 기사 말뭉치 및 그에 따른 띄어쓰기 확률 모델을 참조하여 모든 띄어쓰기 후보들의 확률을 계산한다. 이 과정에서 수행되는 동일한 띄어쓰기 후보에 대한 확률 계산 결과를 해시 테이블에 저장하여 계산 횟수를 줄인다.

2.2 띄어쓰기 오류 교정 시스템

일반적으로 OCR을 통해 인식된 문자열은 철자 오류와 띄어쓰기 오류를 포함하고 있다. OCR에서 문자 인식과 동시에 오류를 교정하는 것은 하드웨어적으로 처리해야 하므로 비용과 시간이 많이 필요하므로, 대부분 OCR 인식 후에 그림 1과 같은 후처리 단계를 추가하여 소프트웨어적으로 오류를 해결한다. 후처리 단계는 오류 검출, 철자 교정, 띄어쓰기 교정으로 이루어져 있다.

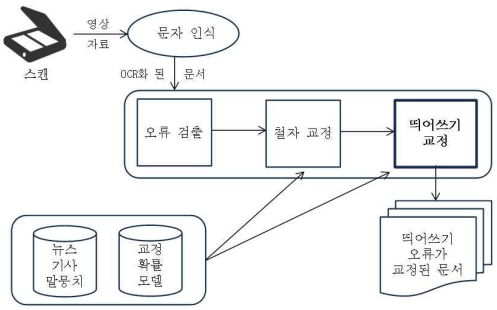

그림 1. OCR에서의 문자 인식 후처리 시스템

오류 검출 단계에서 분류기가 인식 결과를 분석하여 표 1과 같이 문자열이 포함하고 있는 오류의 종류에 따라 인식된 문자열의 각 어절에 해당하는 오류 태그를 부착한다. E는 인식된 문자열의 어절에 오류가 존재하지 않음, S는 띄어쓰기 오류, T는 철자 오류, ST는 띄어쓰기 오류와 철자 오류를 모두 포함하고 있다는 의미이다.

표 1. 오류의 종류

종류	설명
E	오류 없음
S	띄어쓰기 오류
T	철자 오류
ST	띄어쓰기 오류 + 철자 오류

오류 검출에 따라 후처리 과정으로 철자 교정과 띄어쓰기 교정을 진행하게 된다. 철자 교정을 띄어쓰기 교정보다 먼저 하는 이유는 말뭉치와 확률 모델을 띄어쓰기 오류 교정에 이용하기 위해선 오탈자가 없어야하기 때문이다.

2.3 동적 프로그래밍을 이용한 띄어쓰기 오류 교정

본 논문에서 제안하는 띄어쓰기 오류 교정 시스템은 2.2절에서 언급한 바와 같이 OCR 후처리의 가장 마지막 단계에 해당한다. 철자에 오류가 있다면 말뭉치와 확률 모델에 해당 음절에 대한 확률이 존재하지 않는 경우가 발생하여 확률 계산에 치명적인 오류를 포함하기 때문이다. 본 논문에서는 띄어쓰기 오류에 대해서만 다루므로 철자 오류에 대한 교정은 완료되었다고 가정한다. 예를 들어, “나는 학교를 즐겁게 다닌다”라는 문장을 OCR로 인식하여 얻은 데이터가 “나는 학 교를 즐겁게 다닌다”라서 후처리 과정을 통해 띄어쓰기 오류를 교정해야 할 경우, 그림 2와 같이 오류 검출 분류기를 통해 각 어절에 부착된 오류 태그를 단서로 사용한다.

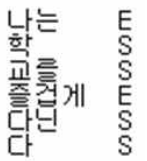

그림 2. 어절에 따른 오류 검출

그림 2에 따르면 ‘학’, ‘교를’, ‘다’, ‘닌다’라고 인식된 어절이 띄어쓰기 오류를 포함하고 있다. 이 오류를 교정하기 위해 그림 3처럼 인식된 문자열의 공백을 모두 제거한다. 그 후, 음절 단위 분절을 통해 가능한 모든 띄어쓰기 후보를 찾는다.

나는학교를즐겁게다닌다

그림 3. 모든 공백이 제거된 문장

동적 프로그래밍을 이용한 음절 기반의 띄어쓰기 과정은 그림 4와 같다. 짙은 색으로 표시된 부분은 해당 단계에서 어절로 확정된 부분이다. 흰색 배경의 문자열은 띄어쓰기를 진행해야하는 부분으로, 음절 단위로 하나씩 분리하여, 뉴스 기사 말뭉치와 교정 확률 모델을 참고하여 띄어쓰기를 할 위치와 그에 따른 띄어쓰기 확률을 계산한다. 띄어쓰기 확률은 말뭉치에 해당 문자열이 존재하는지, 띄어쓰기 확률 모델에 해당 어절 후보가 정답일 확률이 얼마인지를 계산한다. 이 계산 과정에서 앞서 설명한 동적 프로그래밍 기법이 사용되며, 반복적인 부분들의 계산 양을 줄여주어 전체적인 교정 속도가 빨라진다.

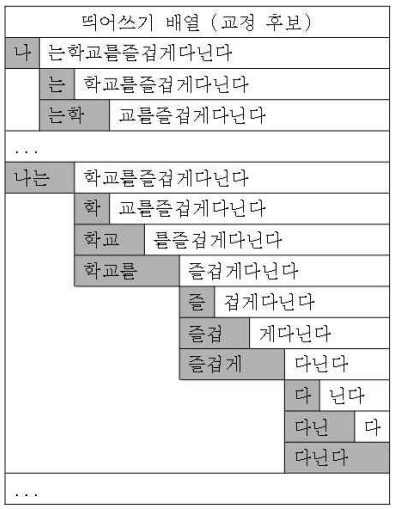

그림 4. 최적의 교정 후보 탐색

가능한 모든 어절의 조합을 후보로 찾아낸 후, 각 후보의 전체 확률을 계산한다. 전체 확률은 해당 후보의 각 어절에 대한 확률들을 곱하여 구할 수 있다. 이 전체 확률이 가장 높은 후보가 최종 띄어쓰기 교정 결과가 된다. 최종 결과 화면은 그림 5와 같다. 교정 결과는 줄 단위로 제시되며, 3개의 항목을 포함한다. 각 항목은 탭(tab) 단위로 구분된다. 첫 번째 항목은 OCR에서 인식된 어절이고, 두 번째 항목은 띄어쓰기 교정 결과를 태그 형식으로 나타낸 것이다. <w>는 어절의 시작을 의미하며 </w>는 어절의 끝을 나타낸다. <w alt=“교정 어절”> “교정 전 어절”</w> 는 “교정 전 어절”을 “교정 어절”로 바꿔야 한다는 의미이다. 마지막 항목은 오류 검출 단계에서 분류기가 해당 어절에 부착한 오류 태그이다.

```
나는    <w>나는</w>                    E
학      <w alt="학교를">학 교를</w>    S
교를    <w/>                           S
즐겁게  <w>즐겁게</w>                  E
다닌    <w alt="다닌다">다닌 다</w>    S
다      <w/>                           S
```

그림 5. 띄어쓰기 오류가 교정되어 출력된 문장

설명한 것을 바탕으로 그림 5의 최종 교정 결과를 살펴보면 띄어쓰기 오류가 있던 '학', '교를' 이라는 어절이 합쳐져서 '학교를' 이라는 어절로 교정되고, '다닌', '다' 라는 어절이 합쳐져 '다닌다' 라는 어절로 교정됐다.

3. 결론

본 논문에서는 광학 문자 인식(OCR)을 통한 문자 인식 시 발생하는 띄어쓰기 오류를 해결하기 위해 동적 프로그래밍을 이용한 분절 방식의 띄어쓰기 후처리 시스템을 제안하였다. 일반적인 띄어쓰기 오류는 자동 줄 바꿈, 강제 줄 바꿈에 의해 발생하므로 문자열 내부의 공백을 전부 없앤 뒤, 음절의 분절을 수행하여, 말뭉치와 띄어쓰기 확률 모델에 근거한 오류 교정을 실시하였다. 단순 띄어쓰기 오류 문장에서의 교정은 잘 진행되었으나, 실제 언어 체계와 달라지거나 구어체가 입력으로 들어오면 사용하는 말뭉치와 띄어쓰기 확률 모델의 특성 상 교정이 어려운 점이 있었다.

향후에는 실제 언어 체계와 다르거나 구어체가 입력으로 들어와도 띄어쓰기 교정을 할 수 있는 규칙을 추가하고, 말뭉치와 띄어쓰기 확률 모델을 확장하여 띄어쓰기 교정 시스템의 성능을 향상시킬 계획이다.

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] 손훈석, 최성필, 권혁철, 문자 인식기의 특성과 말뭉치의 통계 정보를 이용한 문자 인식 결과의 후처리, 제9회 한글 및 한국어 정보처리 학술대회, 1997
[2] 국립특수교육원, 특수교육학용어사전, 하우, 2009.
[3] 전남열, OCR로 문자 인식된 자료의 확률적 띄어쓰기 후처리 시스템, 전남대학교 대학원, 석사 학위 논문, pp.10-11, 2001.
[4] Neapolitan, R. and Naimipour, K., Foundation of Algorithms(4th Edition), Jones & Bartlett Publishers, pp.91-92, 2009.