한글 문자 인식에서의 오인식 문자 교정을 위한 단어 학습과 오류 형태에 관한 연구

이 병 희[†] · 김 태 균[††]

요 약

본 논문에서는 문자 인식 과정을 거치고 난 후에 발생하게 되는 오인식된 문자들을 언어적 지식을 이용하여 교정하는 문자 인식 후처리에 관하여 논한다. 문자 인식의 오인식 교정시스템의 경우 후보 단어가 많을 때 많은 후보 단어중에서 가장 적당한 단어를 후보 단어로 올려주기 위해서는 여러 가지 정보가 필요하다. 본 논문에서는 이러한 정보로 이용할 수 있는 것으로 단어들의 특성과, 문자 인식시에 발생하는 오인식 형태, 단어 학습에 관하여 논한다. 이를 위한 실험으로 15만여의 단어가 수록된 국어 사전을 입력하고 초중고 국어교과서에 나타난 단어들의 사용빈도수를 조사하여 국어 사전에 등록된 단어 중에서 10.7%정도가 실제 초중고 국어교과서에 사용되고 있다는 것을 알 수 있었다. 또한 실제 문자 인식 시스템들을 가지고 여러 문서를 입력하고 인식하여 오인식이 자주 일어나는 글자들의 형태를 분류하여 보았다. 그리고 한국어처리 관련 서적이나 논문을 처리하고자 한국어에 관련된 책의 찾아보기에 나타난 단어를 학습시켜 후보 단어들의 과다로 인하여 정확한 단어를 예측하기 힘들던 문제를 해결하고자 하였다.

A Study on Word Learning and Error Type for Character Correction in Hangul Character Recognition

Byeong-Hee Lee[†] · Tae-Kyun Kim[††]

ABSTRACT

In order to perform high accuracy recognition of text recognition systems, the recognized text must be processed through a post-processing stage using contextual information. We present a system that combines multiple knowledge sources to post-process the output of an optical character recognition(OCR) system. The multiple knowledge sources include characteristics of word, wrongly recognized types of Hangul characters, and Hangul word learning. In this paper, the wrongly recognized characters which are made by OCR systems are collected and analyzed. We input a Korean dictionary with approximately 150,000 words, and Korean language texts of Korean elementary/middle/high school. We found that only 10.7% words in Korean language texts of Korean elementary/middle/high school were used in a Korean dictionary. And we classified error types of Korean character recognition with OCR systems. For Hangul word learning, we utilized indexes of texts. With these multiple knowledge sources, we could predict a proper word in large candidate words.

1. 서 론

대량의 문서를 신속하게 입력하기 위해서는 문서 자동 입력 장치 개발과 이들 장치를 통해서 들어온 영상을 인식하는 문자 인식 기술이 반드시 필요하다. 이러한 요구에 부응하여 문자 인식에 대한 연구가 활성화되어 국내에서도 20여 년전부터 지금까지 꾸준

† 정 회 원:충남대학교 컴퓨터공학과 박사과정
†† 정 회 원:충남대학교 컴퓨터공학과 교수
논문접수:1995년 9월 23일, 심사완료:1996년 1월 12일

히 우리 글인 한글을 중심으로 한글 문자 인식 연구가 있어 왔다.

최근에 들어서는 이러한 연구들의 결실로 우리 기술로 만들어 낸 문자 인식 시스템들이 나오고 있는 실정이다. 하지만 인식된 문서에서 발생하게 되는 오인식을 언어적 지식을 이용하여 후처리를 행하는 연구는 아직까지도 미흡하다고 하겠다[1].

한국어 처리를 위해서는 형태소 분석, 통사 분석, 의미 분석, 화용 분석 등의 처리를 수행하는데, 형태소 분석기를 이용하여 실용화가 이루어진 것이 철자 교정기, 또는 맞춤법 검사기이다. 문자 인식 과정중에 발생하게 되는 오인식을 교정하기 위해서도 형태소 분석이 필요하다.

물론 언어적 지식을 이용하여 잘못된 글자들을 교정하려는 철자 교정기의 연구는 10여 년전부터 있어 왔으나[2] 이러한 연구는 문자 인식의 특성들을 이용하지 않는다는 측면에서 철자 교정기와는 다른 면을 가지게 된다.

또한, 문자 인식 후처리를 위해서는 형태소 분석을 행하여 처리하는 구조적 접근 방식과 문서내의 통계적 정보를 이용하는 통계적 접근 방식이 있을 수 있다. 하지만 구조적 방법은 자연어 처리에서의 근본적 문제인 중의성(ambiguity)과 관련되어 현재로는 쉽게 해결될 문제가 아니다. 반면 통계적 접근 방식은 철자 교정기나 오인식 후처리기에서의 단어 검색에 드는 비용을 줄일 수 있는 방법이긴 하나 교정률이 떨어진다는 단점이 있다[3].

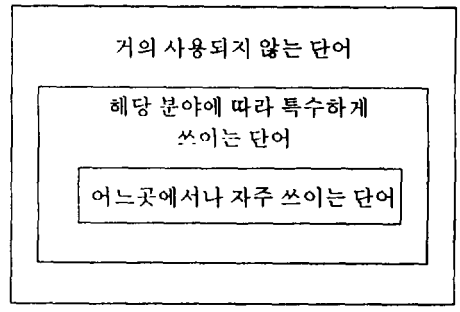
구조적 후처리 방식과 통계적 후처리 방식은 각각의 장단점을 가지고 있으므로 이 둘을 복합한 후처리 방식들도 제안되었다. 복합적 방식의 후처리는 통계적 방식의 빠른 처리속도와 교정 후보 문자열 생성 능력의 장점을 가지며, 구조적 방식의 높은 교정률의 장점을 취하려는 의도에서 한글에서도 적용된 바 있다[4, 5].

본 논문에서는 단어의 사용 빈도수를 이용하여 구조적 방법의 근본적 문제인 중의성을 어느 정도 해결하고자 한다. 또한 대부분의 문서가 어떤 주제를 가지고 쓰여졌기 때문에 해당 분야에서 자주 사용되는 단어들에 대한 정보를 이용하고자 비교적 간단히 얻을 수 있는 서적의 뒤에 있는 찾아보기에 나타난 단어들을 학습하는 방법을 이용한다.

본 논문의 구성은 다음과 같다. 2장에서는 문서에서 등장하는 단어들의 성질에 대해서, 3장에서는 문자 인식시에 발생하는 오인식 형태에 대하여 기술하고, 4장에서는 단어 학습과 교정을, 5장에서는 결론 및 향후 연구 과제에 대하여 논한다.

2. 문서에서 등장하는 단어들의 성질

본 논문에서는 (그림 1)과 같은 단어 계층구조[6]의 성질을 이용한다. 단어들이 사용되는 성질을 보면 어느 곳에서나 자주 쓰이는 단어(예, 우리, 그리고)가 있는데 이들은 단어의 수도 적고 정적인 성질을 가지며, 다음으로 해당분야에 따라 특수하게 나타나는 단어가 있는데 이들은 단어의 수도 많고 해당 분야에 따라 출현이 있고 없는 동적인 성질을 갖는다. 마지막으로 거의 사용되지 않는 단어들은 어느 곳에서나 자주 쓰이는 단어와 해당분야에 따라 특수하게 나타나는 단어들을 제외한 사전에 나타나는 대부분의 단어들이다.



(그림 1) 단어 계층구조
(Fig. 1) The hierarchy of words

이러한 성질을 이용할 경우, 국어사전에 등록된 수많은 단어 중에서 실제로 쓰이는 단어는 많지 않으며 또한 단어를 어떤 주제로 학습할 경우 여러 후보 단어 중에서 정확한 후보 단어를 제시할 수 있어 처리 효율이나 교정률의 향상을 가져올 수 있다.

해당분야에 관련된 단어들만을 가진 단어 사전이라든가 실험실용의 조그만 단어 사전을 가지고 실험을 했을 때는 교정률이 좋았지만, 크고 단어가 많이

수록된 사전을 이용하면 교정률이 현저히 떨어지는 경우가 종종 있어 왔다. 수록된 단어가 많아질수록 오히려 교정의 정확률은 떨어질 수 있다. 그렇다고 조그만 단어사전만을 이용하면 일반성이 없고 단어 사전을 확장시 문제가 된다.

이런 문제를 해결하기 위해서 본 논문에서는 단어가 많이 수록된 단어 사전을 자주 사용되는 단어는 가중치나 순위를 높이고, 또 해당분야의 특수한 단어들은 어렵지 않게 학습시킬 수 있는 서적의 뒤에 나오는 찾아보기로 단어 학습을 하는 방법을 이용하고자 한다.

이렇게 하면, 단어들이 어느 곳에서나 나타나는 일반적인 성질과 해당 분야의 특수한 단어들의 성질도 함께 이용할 수 있으며 오히려 많은 단어들 때문에 발생하던 교정의 정확성 감소를 해결할 수 있다.

철자교정 시스템과 마찬가지로 가장 훌륭한 문자 오인식 교정시스템이라고 하면, 오류가 있는 모든 어절만을 찾아내어 그 오류에 대한 최소 개수의 후보를 제시하고 그 후보 중에는 맞는 어절이 반드시 들어 있어야 한다[7]. 그러나 실제로는 오류를 찾아내지 못하거나, 맞는 어절을 틀리는 어절로 해석하거나, 후보 어절을 만들어 내지 못하거나, 후보 어절을 과다하게 생성하는 등의 여러 문제점이 드러나고 있다. 물론 사전에 등록되지 않은 복합어나 고유명사 등으로 인해 처리의 한계가 있는 것도 사실이다.

본 논문에서는 문자 인식에서 발생하는 오류 형태를 알아보고 국어사전에 등록된 대부분의 단어들이 실생활에서는 거의 사용되지 않고 소수만이 사용된다는 성질을 이용하여 과다하게 생성되는 후보 단어를 줄이고자 한다.

3. 문자 인식시에 발생하는 오인식 형태

오류를 넓게 분류하면, 맞춤법오류, 구문오류 및 의미오류 등으로 나눌 수 있다. 이 중에서 구문 오류나 의미 오류는 형태소 분석만 가지고서는 처리하기 곤란한 오류들이므로 보통의 철자교정 시스템에서는 이들을 처리하지 않고, 형태소 분석으로 어느 정도 처리할 수 있는 맞춤법 오류를 교정하는 것이 보통이다. 맞춤법 오류는 어절에 발생한 오류로 철자 오류, 문장부호 오류, 띄어쓰기 오류 등이 있는데, 문장부호 오류는 드물게 나타나며, 띄어쓰기 오류는 상용 문서편집기에서 쉽게 처리되고 있다. 따라서 철자 오류가 맞춤법 오류중에서 대부분을 차지하고 있다.

인쇄물에서 발생하는 오류의 종류로는 띄어 쓰기, 맞춤법 오류 및 어미와 조사의 오용이 전체오류의 약 60%를 차지한다고 한다[8]. 또한 지금까지의 철자오류 교정시스템들은 문서편집기를 사용하여 문서를 작성할 때 범하게 되는 오류에 관한 연구가 많은 부분을 차지하고 있었다.

그러나 문서편집기에서는 문서의 내용을 자판을 눌러서 직접 입력하게 되므로 자판을 잘못 누르는 것이 대부분인 경우의 오류를 교정하고자 하는 연구를 해 왔으며, 이런 자판을 잘못 누르는 오류는 인쇄물을 인식하는 오프라인 문자 인식에서는 거의 찾아 볼 수 없는 오류 형태이므로 기존의 문서편집기를 위한 철자 교정을 위한 알고리즘들은 문자 인식을 위한 오류 교정 시스템으로는 그대로의 적용이 거의 불가능하다.

그리하여 본 논문에서는 실제 문자 인식 시스템에서 인식된 여러 결과를 이용하여 분석하여 본 결과, 문자 인식시스템에서 발생하는 오류의 대부분은 문자 인식 시스템이 글자의 모양을 보고 인식을 하게 되므로, 주로 발생하는 오류는 글자간의 유사성 때문이었다. 또한 문서편집기에서 단어나 어절의 첫글자는 오류가 거의 발생하지 않는다는 것도 문자 인식 시스템에서는 적용할 수 없었으며 문자 인식에서의 오인식은 철자 오류와 관계가 깊었다.

또한, 문자 인식 시스템을 위한 오인식 교정을 위해서는 무엇보다도 문자 인식 시스템들이 어떻게 문자 인식을 행하는지에 관한 연구가 선행되어져야 한다. 그리하여 문자 인식 시스템에서 자주 오인식되는 글자들에 대한 예들을 오인식 교정 시스템에서 이용할 수 있도록 오인식 문자 혼동 테이블을 작성하였다. 〈표 1〉은 오인식된 글자들의 형태이다.

이 표는 신문, 잡지, 간행물 등을 영상 입력장치인 HP ScanJet 4c 스캐너상에서 300dpi로 입력 받아 상용 소프트웨어인 한국인식 기술의 HiART 글눈과 합산컴퓨터의 아르미로 인식한 다음 오인식되는 문자들을 조사한 것이다.

(그림 2)는 오인식의 유형을 비율별로 나타낸 것이다. 〈표 1〉에서 보듯이 본 논문에서는 오인식의 형태

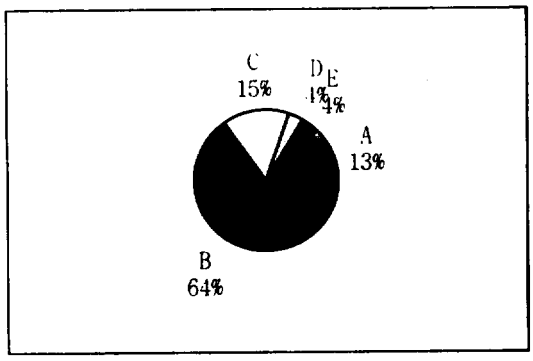
〈표 1〉 오인식된 글자들의 형태
〈Table 1〉 The error types of wrongly recognized characters

오인식 형태	오인식된 글자들의 예
글자형태무관(A)	엄/떠, 합/라, 희/료, 엽/빱, 쓰/새, 며/시, 고/싯, 면/싶 …
1자소 오류(B)	갈/갈, 격/경, 필/필, 필/펕, 될/뉠, 늣/늦, 답/담, 직/저, 쩍/쩌 …
2자소 오류(C)	팬/펭, 병/냥, 넬/뉘, 핵/락, 맞/방, 영/싱, 현/엿, 펜/펭 …
1획 첨가(D)	능/눙, 등/듕, 림/림, 풀/풀 …
1획 탈락(E)	층/중, 측/즉, 화/희, 호/흐 …

를 글자 형태에 무관한 오인식, 한 자소의 대치로 인한 오류, 2자소의 대치로 인한 오류, 한 획의 첨가로 인한 오류, 1획의 탈락으로 인한 오류로 분류하였다.

문자 인식시 인식 알고리즘의 잘못으로 인해서 발생하는 글자 형태에 무관하게 인식하는 오인식 A가 13%정도 이였으며, 이 오인식은 인식 알고리즘을 잘 학습시키거나 훈련을 하면 줄어들을 수 있는 오류이다.

한 자소가 오인식되는 오류 B는 64%이였다. 여기서 주목해서 볼 문자 예는 직/저, 쩍/쩌와 같은 경우이다. 이들은 실제로는 두 자소가 틀린 것이나 문자 인식의 측면에서는 저와 쩌의 "ㅓ"의 위치를 분별하지 못해서 이루어진 오류로 이는 간단히 교정이 가능하기 때문에 B의 오류 형태에 포함시켰다. 문자 인식 시스템의 인식률이 99%가 넘는 인식 시스템들은 보통 한 자소만 틀리는 이러한 종류의 오류가 많았다.



(그림 2) 비율별 오인식의 형태
(Fig 2) The percentages of wrongly recognized characters

두 자소가 오인식된 오류 C는 15%이였다. 한글의 초성, 중성, 종성을 모두 갖춘 글자에서 많이 발생했으며 이 오류는 한 자소가 오인식되는 오류 B보다 처리가 어려운 오류이다.

한 획이 첨가 되거나 탈락하는 오류 D, E는 인식하려고 하는 문서에 잡음(noise)이 있을 때 발생하는 오류로 각각 4%이였다. 이러한 한 획의 첨가라든가 탈락으로 인하여 발생하는 오류는 영상 입력 장치 하드웨어인 스캐너(scanner)의 해상도가 좋아짐에 따라 감소될 수 있는 오류이다.

본 연구에서 오인식되는 유형을 분석해 본 결과는 다음과 같다.

1. 글자간 모양의 유사성으로 오인식되는 것이 가장 많다. 문자 인식이 입력되어 들어온 영상인 글자의 모양을 보고 인식이 이루어지기 때문에 글자의 유사성으로 인한 오인식이 대부분이다. 이렇게 글자간의 유사성으로 오인식된 오류는 문자 인식 기술만으로는 이러한 오인식을 줄이는데 한계가 있으며 이런 오류는 언어적 지식을 이용하는 문자 인식 후처리가 반드시 필요하다.
2. 획수가 많은 글자는 오인식이 많다. 즉 획수가 많아 복잡한 글자는 오인식이 많다. 이것은 영상 입력 장치인 영상 스캐너의 해상도가 높아지면 어느 정도 해결될 문제이다. 문자 인식 후처리에서는 획수가 많은 글자들은 사용빈도가 그리 높지 않은 것이 대부분이므로 한글의 음절별 출현 빈도수를 이용하면 해결될 수 있다.
3. 문자 인식 알고리즘의 잘못으로 인하여 글자 형태와는 거의 관련성이 없이 엉뚱한 형태로 오인식 되는 글자들도 있다. 이것은 인식 알고리즘을 잘 고치거나 문자 학습을 시키면 될 것이다. 인식부에서 해결할 수 있는 이러한 오류는 인식 알고리즘을 잘 정비하여 오류가 발생하지 않게 하여야 한다. 가능한한 오류의 가능성은 문자 인식 후처리전에 이루어 지는 것이 좋다. 왜냐하면 문자 인식 후처리에서 교정을 하려면 형태소 분석과 단어의 검색을 많이 해야 되므로 그만큼 교정 비용이 많이 든다.
4. 스캐너의 해상도가 좋아짐에 따라 획의 첨가나 탈락으로 인한 오류는 줄어들 것이다. 과거에 비

해 요즘의 스캐너들은 그 해상도가 매우 발전하고 있다. 이러한 하드웨어의 발달은 인식전의 전처리나 특징추출 단계에서 문제가 되던 입력 영상의 불완전 문제를 해결하는데 도움이 된다.

5. 문서편집기에서의 맞춤법 검사기나 철자 교정기에서 이루어지던 단어의 첫 글자가 옳다는 가정을 하기가 어렵다. 왜냐하면 자판으로 입력할 때는 사용자가 첫 글자를 잘못 치는 경우가 거의 없으나 인식시스템의 경우는 첫 글자나 두번째 글자나 오인식이 일어날 확률이 같다.

위와 같은 오류의 유형을 보면 문서편집기에서의 맞춤법 검사기나 철자 교정기에서 쓰이던 교정 방법을 그대로 적용하는 것이 거의 불가능하다는 것을 알 수 있다.

4. 단어 학습과 교정

본 장에서는 한글 문자 인식 후처리를 이용하여 얼마만큼 인식률 향상을 가져 올 수 있는지를 확인하기 위하여 단어 학습과 교정에 관하여 알아보도록 한다.

4.1 단어 학습

단어 학습과 단어 교정을 위하여 가장 먼저 국어사전[9]에 등록된 단어들을 입력하였다. 전체 15만여개에 달하는 단어를 고어는 입력시키지 않고, 동음이의어는 한 단어로 입력하여 보니, 총 7만8천여 단어가 되었다.

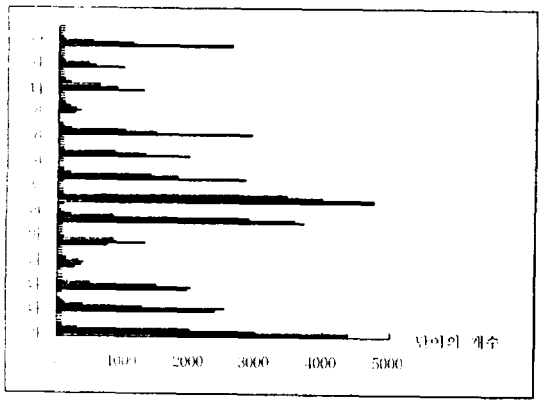


그림 3) 한글의 단어별 개수
(Fig 3) The number of Hangul characters by word

(그림 3)은 국어사전[9]를 입력하였을 때 가나다 14개에 대해 각 글자를 또다시 글자수에 따라 9개로 분류하였을 때의 단어별 개수이다.

이에 따르면 국어의 단어는 2자와 3자짜리 글자가 가장 많다. 국어의 단어가 2자짜리가 많다는 것은 오인식 교정의 측면에서 교정을 어렵게 하는 요인이 된다. 왜냐하면 두 글자중 하나가 틀렸다고 하면, 가장 많은 2자짜리 단어들이 후보단어가 될 수 있으므로 실제로는 교정을 불가능하게 만들기 때문이다. 본 연구에서 조사된 국어의 평균 단어 길이는 2.96자이였다. 물론 이 수치는 어절별 글자수가 될 때에는 2.96자가 좀 넘을 것으로 예상된다.

그러나 국어사전을 보고 입력한 7만8천여 단어가 일상 문서에서 전부 다 사용되는 것은 아니다. 실제로 이들 단어중 얼마가 보통의 문서나 책에서 이용되는 지를 알아보기 위하여 초중고 국어 교과서[10]에 나타난 단어들을 입력하고 조사해 보았다. 국어 사전을 입력하여 얻은 총 7만8천여 단어중에서 초중고 국어 교과서에서는 8천4백여 단어가 이용되고 있어 실제로 우리 국어 사전에서 이용되는 단어의 10.7%정도가 초중고 국어 교과서에서 이용되고 있었다.

물론 초중고 국어 교과서에 나타난 단어들이 우리가 일상생활에서 쓰고 있는 단어들을 대표한다고 볼 수는 없으나 이 결과를 보면 우리가 사용하는 단어는 실제로 국어사전에 나타난 많은 단어들을 모두 사용하는 것이 아니라 10%가 조금 넘는 단어가 쓰이고 있다는 것을 알 수 있었다. 이렇게 단어의 소수가 대부분의 문서에서 집중적으로 등장한다는 단어 사용의 지역성(locality)를 이용하여 원하는 단어가 다수의 후보 단어들 중에 포함되어 있을 때 원하는 단어의 순위를 높일 수 있었다.

또한 해당 분야에 따라 특수하게 쓰이는 단어들을 쉽게 학습할 수 있는 방법으로 보통 서적의 뒷부분에 있는 찾아보기에 나오는 단어들을 학습시켜 보았다. 서적의 뒷부분에 있는 찾아보기는 그 서적의 내용을 대표할 수 있는 단어들을 선정해 놓은 것으로 (그림 1) 단어 계층 구조에서 "해당 분야에 따라 특수하게 쓰이는 단어들"이다. 이렇게 서적의 찾아보기를 이용하면 학습시키는데 큰 어려움없이 해당 분야에 따라 특수하게 쓰이는 단어들을 학습시킬 수 있다.

해당 분야에 따라 특수하게 쓰이는 단어들을 이용하여 단어 교정을 하는데 도움을 주기 위하여 본 논문에서는 국어 문법서인 [11]을 가지고 실험해 보았다. 국어 문법서에 등장하는 단어들은 국어 문법에 관한 내용을 대표할 수 있는 단어들이 많이 들어가 있다. [11]의 찾아보기를 입력하고 가중치를 별도로 주어 단어의 교정 과정에서 이용할 수 있도록 하였다.

4.2 단어의 교정

지금까지 만들어진 사전과 문자 혼동 테이블, 단어 학습을 이용하여 한글 문자 오인식 교정을 하기 위하여 본 논문에서는 입력 문서를 스캐너로 받아 문자인식 시스템으로 인식한 결과를 교정하여 보았다. (그림 4)는 [11]의 한 부분을 문자인식 시스템으로 인식한 결과이며 밑줄 친 부분은 오인식된 문자를 가리킨다.

단어 교정을 위해서는 먼저 인식된 어절에 대해 형태소 분석을 하여 맞는 어절인지 틀린 어절인지를 판별하고 만일 틀렸다고 하면 교정을 위해 맞는 어절을 제시하여야 한다. 하지만 국어의 경우는 형태소 분석이 까다롭고 특히, 용언의 경우 활용과 불규칙이 많아 처리하기가 쉽지 않다. 오인식 교정을 위한 국어의 형태소 분석을 위해 본 논문에서는 다음과 같은 가정과 처리 과정을 하였다.

오프라인 문자 인식 시스템을 통해 나온 오인식된 결과를 수정하고자 하는 오인식 교정은 구문 정보, 의미 정보까지를 이용하고자 하는 자연어 처리를 위한 기존의 형태소 분석과 같은 자세한 분석은 필요치 않다. 그리하여 본 논문에서는 360여개의 조사와 어미들로 구성된 기능어 사전을 이용하여 어절을 분리하고 국어 사전을 입력한 단어들과 단어의 유사성과 문자 혼동 테이블을 이용하여 후보 단어들을 생성하여 보았다. 오인식 교정을 위한 국어 형태소 분석에 관한 연구는 앞으로 계속 연구하여야 할 부분이며 아직은 본 논문에서 미약한 부분이기도 하다.

위 (그림 4)를 가지고 교정을 시도한 결과는 〈표 2〉와 같다. 〈표 2〉에서 A는 학습을 안 했을 때의 후보수 즉 국어사전의 단어를 단순히 입력한 것을 가지고 실험하였을 때의 후보수, B는 초중고 국어 교과서로 학습하였을 때의 후보수, C는 B에다가 찾아보기로 학습까지 했을 때의 후보수, ()의 O, ×는 원하는 교정 단어의 포함 여부, 그리고 순위는 원하는 교정 단어의 순위이다.

〈표 2〉에서 "조금 ㅏ석"은 "조금씩"을 오인식한 것으로 이 단어는 단어의 유사성을 이용하여 후보 단어들을 각 사전에서 검색하면 어렵지 않게 고칠 수 있었다. 원래 단어가 "자리"인 "카리"는 문자 혼동 테이블의 정보도 이용하고 각 국어 사전에서 단어를 검색해 보아도 2자짜리 후보 단어가 너무 많아 교정에 실패한 예이다. 이 "카리"의 예에서 보듯이 학습한 자료가 불완전한 자료일 경우 오히려 후보 단어들 중에서 원하는 단어를 제외시키는 경우도 있었다. "기능"이 원래 단어인 "기?"은 찾아보기에서 이 단어가 이 문서에서 쓰인다는 정보를 이용하여 후보 단어들이 많았음에도 불구하고 교정에 성공한예이다.

국어의 단어 중에서 2자짜리 단어가 많아 두 글자 중에 한 글자만 틀려도 교정할 수 없는 문제를 해결하기 위해서는 문자 인식 시스템의 오인식 성질을 이용할 수 있는 견고한(robust) 오인식 문자 혼동 테이블이나, 기능어나 불용어(stopword), 문서내 사용빈도들을 잘 고려하여 교정한다면 교정도 가능하리라

명사, 대명사, 수사는 의미는 조금ㅏ석 다르지마는 모두 무엇의 자리를 채울 수 있기 때문에 기능상으로는 큰 차이가 없다. 목적이나 서술이의 자리에 나타나기도 하지마는 주어의 카리에 나타나서 문장의 주체, 곧 몸의 노릇을 하는 것 이 중요한 기?이라 할 수 한다.

(그림 4) 문서 인식 시스템으로 인식된 텍스트의 예
(Fig 4) A recognized text example by a OCR system

〈표 2〉 오인식 교정시의 후보 단어들의 수
〈Table 2〉 The number of candidate words

오인식 글자	A	B	C
조금 ㅏ석	2(O), 2위	2(O), 1위	2(O), 1위
카리	184(O), 120위	16(×), 120위	1(×), 120위
기?	212(O), 33위	18(×), 49위	19(O), 1위

생각된다[12, 13, 14].

본 논문에서는 단어학습을 고려하면 기존의 방법으로는 후보 문자 집합이 많을 때 올바른 정정이 사실상 불가능하던 것을 문서의 내용을 어느 정도 알 수 있도록 학습을 하면 이러한 후보 집합을 줄이는데 훌륭한 정보로서 이용될 수 있다는 것을 보였다. 본 방법은 문서의 내용을 대표하고 있는 단어들을 가지고 후처리가 이루어지므로 오인식 교정률도 높일 수 있으며 해당 분야별 사전과 함께 후처리 될 시에는 더욱 우수한 성능을 보이리라 생각된다.

5. 결론 및 향후 연구 과제

지금까지 문자 인식 시스템에 의해 오인식된 문자들을 교정하기 위한 한글 문자 인식 후처리에 대하여 알아 보았다. 국어사전에 등록된 수많은 단어가 실제로는 거의 사용되지 않는 단어들이 대부분이고 실제로 초중고 국어 교과서에 나타나는 단어들은 국어사전에 등록된 단어들 중의 10. 7%정도가 쓰이고 있었으며, 이러한 성질을 이용하여 교정 후보 단어들이 너무 많고 그 중에서 원하는 단어가 100위 후보안에도 안 들어 오던 문제를 어느 정도 해결할 수 있었다.

그렇지만 초중고 국어 교과서가 일상 단어들을 대표하지 못하기 때문에 즉, 학습된 자료가 불완전할 때는 오히려 원하는 단어를 후보 단어들 중에서 제외시키 버리는 문제도 있어서 이를 위한 해결책도 앞으로 해결하여야 할 문제로 남는다.

그리고 문자 인식시에 오인식이 일어나는 글자들에 대하여 알아 보았으며 오인식 유형을 분류하고 분석하여 보았다. 오인식 교정을 위해서는 인식시스템과 오인식 유형을 잘 분석해 보아야 한다.

또한 문서의 내용에 관한 지식을 이용하기 위하여 서적의 뒷면에 나오는 찾아보기의 단어들을 간단히 학습시켜 보았다. 그 해당 분야에 따라 특수하게 등장하는 단어들을 알 수 있으면 후보 단어들 중에서 원하는 단어의 순위가 높아져 교정의 정확률을 높일 수 있었다.

향후 연구 과제로는 오인식 교정을 위한 강력한 국어의 형태소 분석 알고리즘이 필요하며, 오인식 교정률 향상을 위해서 쉽게 구문정보나 의미 정보를 자동적으로 추출하여 후처리에 이용할 수 있는 방법에 관한 연구가 필요하다.

참 고 문 헌

[1] 이성환, 문자 인식:이론과 실제, 홍릉과학출판사, 서울, 1993.

[2] 채영숙, 김재원, 권혁철, "도움말 기능을 가진 문서 철자 검색/교정 시스템," 한국정보과학회 가을 학술발표논문집, Vol. 17, No. 2, pp. 815-818, 1990.

[3] 박진우, 이일병, "통계적 방법에 의한 후처리," 제 6회 한글 및 한국어 정보처리 학술발표 논문집, pp. 518-526, 1994.

[4] 이원일, 홍남희, 이종혁, 이근배, "Binary N-gram과 형태소 분석기를 이용한 한국어 철자 교정기," 한국정보과학회 봄 학술발표논문집, pp. 813-816, 1993.

[5] 황호정, 도정인, 권혁철, "한글 문자 인식을 위한 후처리기의 개발과 속도 개선," 제 2회 문자 인식 워크샵, pp. 180-188, 1994.

[6] J. L. Peterson, "Computer Programs for Detecting and Correcting Spelling Errors," Communication of ACM, Vol. 23, No. 12, pp. 676-686, 1980.

[7] 임한규, 김웅모, "철자오류의 통계자료에 근거한 철자오류 교정시스템," 한국정보처리학회 논문지, 제2권 제6호, pp. 839-846, 1995.

[8] 미승우, 새 맞춤법과 교정의 실제, 어문각, 1994.

[9] 엣센스 국어사전, 민중서림, 1995.

[10] 원영섭, 초·중·고 국어 교과서에 나타난 띄어쓰기·맞춤법 용례, 세창출판사, 1993.

[11] 남기심, 고영근, 표준 국어문법론, 탑출판사, 1993.

[12] Hisao Niwa, Kazuhiro Kayashima and Yasuharu Shimeki, "キ-ワ-ド抽出と最適文節選擇による文

字認識後處理," 日本電子情報通信學會論文誌, D-II Vol. J76-D-II, No. 5, pp. 940-948, 1993.
[13] 이병희, 김태균, "키워드 추출법을 이용한 한글 문자 인식 후처리," 한국정보과학회 가을 학술발표논문집 A, 제21권, 제2호, pp. 411-414, 1994.
[14] 이병희, 이인동, 김태균, "한국어 오인식 수정을 위한 자기 구조화 휴리스틱스에 기반한 사전 구성," 한국정보과학회 가을 학술발표논문집, 제20권, 제2호, pp. 1155-1158, 1993.

이 병 희

1992년 충남대학교 컴퓨터공학과 졸업(학사)
1994년 충남대학교 대학원 컴퓨터공학과(공학석사)
1994년~현재 충남대학교 컴퓨터공학과 박사과정 재학중
1995년~현재 충남대, 충북대 컴퓨터공학과 시간강사
관심분야: 패턴인식, 문자인식, 자연어처리

김 태 균

1971년 서울대학교 공업교육학과(학사)
1980년 일본동경공업대학 대학원 물리정보학과(공학석사)
1984년 일본동경공업대학 대학원 물리정보학과(공학박사)
1974년~현재 충남대학교 컴퓨터공학과 교수
관심분야: 패턴인식, 영상처리, 멀티미디어