오프라인 한글 문자 인식을 위한 효율적인 오인식 단어 교정 방법

이 병 희† · 김 태 균†

요 약

문자 인식 과정을 거치고 난 후에 발생하게 되는 오인식된 문자들을 언어적 지식을 이용하여 교정하는 문자 인식 후처리 과정이 반드시 필요하다. 본 논문에서는 한국어의 형식 측면에서 품사를 재분류하고 사전을 구성하며 한글 어절의 상태 전이도를 구성하고 형태소 분석을 위해 Head-tail구분법을 적용해 단어를 분리하였다. 또한 본 논문에서는 효율적인 단어분리와 교정을 위해 여러 문서와 책들로부터 세롭게 조사의 결합형으로 900여개를, 규칙 어미의 활용형으로 800여개를 수집하였다. 그리고 불규칙 용언의 활용형을 위해 국어학에 나오는 9개의 불규칙을 조사하여 활용형을 구축하였고 자동적 교체와 불구동사의 활용형도 사전에 등록하여 어절을 분석하는데 이용하였다. 어느 인식시스템을 가지고 문서를 인식한 결과 93.7%의 인식률을 보인 것을 본 단어 교정 방법을 적용한 결과 97%로 인식률을 향상시킬 수 있었다.

An Efficient Correction Method for Misrecognized Words in Off-line Hangul Character Recognition

Byeong-Hee Lee† · Tae-Kyun Kim†

ABSTRACT

In order to achieve high accuracy of off-line character recognition(OCR) systems, the recognized text must be processed through a post-processing stage using contextual information. In this paper, we reclassify Korean word classes in terms of OCR word correction. And we collect combinations of Korean particles(approximately 900) and Korean verbal forms(around 800). We aggregate 9 Korean irregular verbal phrases defined from a Korean linguistic point of view. Using these Korean word information and a Head-tail method, we can correct misrecognized words. A Korean character recognizer demonstrates 93.7% correct character recognition without a post-processing stage. The entire recognition rate of our system with a post-processing stage exceeds 97% correct character recognition.

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 활성화되면서 국내에서도 20여 년전부터 지금까지 꾸준히 우리 글인 한글을 중심으로 한글 문자 인식 연구가 있어 왔다.

최근에 들어서는 이러한 연구들의 결실로 우리 기술로 만들어 낸 문자 인식 시스템들이 나오고 있는

† 정 회 원:충남대학교 컴퓨터공학과
논문접수:1996년 5월 16일, 심사완료:1996년 8월 8일

실정이다. 하지만 인식된 문서에서 발생하게 되는 오인식을 언어적 지식을 이용하여 후처리를 행하는 연구는 아직까지도 미흡하다고 하겠다[1].

문자 인식 시스템에서 출력되는 결과는 오인식된 문자를 포함하고 있으며, 1)문자 인식 시스템이 오인식한 경우와, 2)사용자의 실수인 경우, 두가지로 분류할 수 있다[2, 3]. 1)의 오인식은 문자 인식 시스템이 불완전하거나, 영상 입력 장치인 스캐너(scanner)에 의해 발생하는 경우이다. 2)의 오류는 사용자가 부주의로 문자를 잘못 썼거나, 맞춤법 실수에 의해 발생하는 오류이다. 따라서, 인간이 문장을 이해해 나가듯이 문맥적 정보를 이용하여 오류를 교정하는 오인식 교정 단계가 반드시 필요하다.

그러나 영상 입력 장치인 스캐너를 통해 들어온 영상을 문자 인식하는 오프라인 문자 인식 시스템의 경우에는 입력되는 문서가 대부분 사람에 의해 오류 교정이 이루어진 것들로서 1)의 오류가 대부분이다. 그리하여 지금까지 컴퓨터의 자판(keyboard)을 통해 입력된 글자들을 교정하기 위해 개발된, 오류의 형태가 2)에 해당하는 맞춤법 검사기 알고리즘[4, 5]은 문자 인식 시스템을 위한 오인식 교정 시스템에 적용이 어렵다. 또한 지금까지 국내에서 주종을 이루어 왔던 기계 번역이나 자연어 이해 시스템과 같은 자연어 처리를 위한 형태소 분석기[6]는 맞는 문장(correct sentence)을 가지고 구문이나 의미 정보까지 분석하는 것이 목적이어서 오인식된 문장(incorrect sentence)을 교정하기 위한 문자 인식 후처리와는 목적부터가 다르다.

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 반드시 필요하다.

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[7].

구조적 후처리 방식과 통계적 후처리 방식은 각각의 장단점을 가지고 있으므로 이 들을 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[8, 9].

본 논문에서는 한글 문자인식에서 발생하는 오인식된 문자들을 교정하기 위해서 한국어를 형식 측면에서 품사를 재분류하고 사전을 구성하며 Head-tail 구분법을 이용하여 형태소를 분석하는 방법에 관하여 기술하고자 한다. 또한 본 논문에서는 효율적인 단어분리와 교정을 위해 조사의 결합형과 규칙 어미의 활용형, 불규칙 용언의 활용형을 수집하여 이를 적용하는 방법을 제시한다. 본 논문의 구성은 다음과 같다. 2장에서는 품사 분류에 대해서, 3장에서는 사전의 구성에 대하여, 4장에서는 오인식 단어 교정을 위한 형태소 분석에 대하여, 5장에서는 실험 및 결과에 대하여, 마지막 6장에서는 결론 및 향후 연구 과제에 대하여 기술한다.

2. 품사 분류

본 논문에서는 오인식 문자들을 교정하기 위한 형태소 분석 방법을 개발하고자 언어학에서의 품사 분류 방법과 조어법(word-formation), 그리고 전산학의 기계번역, 자연언어 이해 시스템, 자동색인에서 적용하던 형태소 분석 방법을 한글 문자 인식 시스템의 오인식 성질에 바탕을 두고 오인식된 문자들을 교정하기 위해 의미, 기능, 형식에 따라 품사 분류를 하는 기존의 방법을 형식에 따라 분류하고 단어를 분리하는 방법에 관하여 논한다.

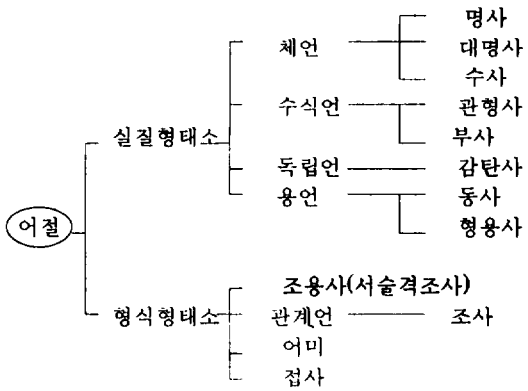
품사 분류의 기준으로는 문장 속에서 단어가 가지는 의미(의미론적 기준)와 기능(통사론적 기준) 및 형식(형태론적 기준)에 따른다[10, 11]. 본 논문에서의 품사 분류 기준은 크게 실질 형태소와 형식 형태소의 범주에 따라 문법상 나타나는 형태소의 역할 차이에 의해 품사를 분류한다. 한 어절내에서 형태소들을 분리해내려면 형태소의 구조를 분석하여야 한다. 예를

들어, '철수가 동화를 읽었다'를 형태소 분석하면 다음과 같다.

(1) 철수, 동화, 읽

(2) 가, 를, 었, 다

(1)은 구체적인 대상이나, 동작, 상태와 같은 어휘적 의미를 표시하므로 실질 형태소라고 부르며, (2)는 실질 형태소에 붙어 주로 말과 말 사이의 관계나 기능을 형식적으로 표시하므로 형식형태소라 일컫는다. 이런 형식 측면에 의해 품사 구분을 하면 (그림 1)과 같다.



(그림 1) 형식 측면에서의 품사 분류
(Fig. 1) The classification of part-of-speech by form

3. 사전의 구성

오인식 단어 교정을 위한 사전은 (그림 1)의 형식 측면에서의 품사 분류에 따라 실질 형태소 사전과 형식 형태소 사전으로 구성한다. 실질 형태소 사전에는 체언, 수식언, 독립언, 용언으로 분류하여 입력하였다. 형식형태소에는 조용사, 조사, 어미를 사전에 수록하고 접사에서 접두사는 체언에 교착된 형태로 하고, 접미사는 형식 형태소 사전에 등록한다. 사전의 전체 구성은 (그림 2)와 같다.

실질 형태소 사전: 체언, 수식언, 독립언, 용언
형식 형태소 사전: 서술격조사, 조사, 어미, 접미사

(그림 2) 한국어 사전의 구성
(Fig. 2) The construction of Korean word dictionary

조용사는 명사, 의성어, 의태어등에 붙어서 용언화시키는 역할을 한다. '하다'는 순수 우리말이나 한자에 붙어서 용언이 되는 경우가 국어의 경우에 많다. '하다'를 독립적인 형태소로 분류하면 '활용하다, 설명하다, 분류하다'에서 '하다'라는 공통 요소를 없앨 수 있으므로 사전의 크기를 감소시킬 수 있다. 그리하여 본 논문에서는 체언과 '하다, 되다, 스럽다'를 분리하여 저장하고 교착정보를 등록하여 접속 여부를 나타낸다. 서술격 조사 '이다'에 대해서도 체언과 분리하여 따로 저장한다.

관계언인 조사는 문장 안에서의 기능과 의미에 따라 격조사, 접속 조사, 보조사로 나뉜다. 또한 조사는 앞 체언의 받침의 유무 즉, 음운론적 조건에 의하여 교체되는 이형태(異形態)가 있다. 그리고 조사는 조사끼리 서로 결합될 수 있다. 지금까지 조사에 대한 연구가 있었으나 조사 상호간에 배열 순서와 그 배열에 따른 규칙을 추출하는데 어려움이 많았다. 그 주된 원인은 조사 기능의 상관관계를 중심으로 한, 자연어 처리 방법이 활발하게 논의되지 못했기 때문이다. 따라서 기존의 분류 방법과 달리 배열 순위와 기능에 따라 조사를 재분류하는 방법이 필요하다[12]. 그리하여 본 논문에서는 배열 순위와 기능에 따라 조사들의 결합된 형태들을 조사해 보았다. 복합조사는 형태는 적지만 복합조사를 검사하는 별도의 처리를 수행하여야 하기 때문에 본 논문에서는 서적이나 글에서 등장하는 복합조사를 조사하여 모두 사전에 수록하였다.

실질 형태소 사전에는 단어, 품사정보, 조용사 교착정보 및 용언의 불규칙정보가 포함되도록 구성한다. 형식 형태소 사전에는 단어, 결합조건(품사의 음운규칙에 따른 실질 형태소에 대한 결합 조건을 기술함)이 포함되도록 구성하였다.

4. 오인식 단어 교정을 위한 단어 분석

4.1 언어학과 자연언어 처리에서의 어절 분석

형태소간의 접속 형태를 파악하기 위해서는 한글 어절에 대한 구조분석이 요구된다. 한글 어절을 분석하는데는 언어학의 경우, 각 품사의 굴곡현상을 다루는 굴절형태론(inflectional morphology)과 단어와 그

파생형의 관계를 다루는 파생형태론(derivational morphology)으로 구분하며 조어법(word-formation)으로 파생법(derivation)과 합성법(compounding)으로 나누고 있다[13]. 언어학에서의 조어법을 이용하면 단어의 생성(generation)면에서는 우수할지 모르나 생성을 제한하는 규칙을 만들기가 쉽지 않으며 과잉생성의 문제를 유발할 수 있어 본 오인식 문자 교정을 위한 단어 분석 방법으로는 그대로의 적용이 힘들다.

지금까지 전산학[6, 12, 15]에서 제시된 여러 가지 한국어 형태소 분석 방법 중에서, 많이 알려진 것들로는 기계번역(machine translation)과 정보검색을 위한 자동색인(automatic indexing)등에서 이용하기 위한 자연언어 처리를 위한 형태소 분석과 워드프로세서의 철자교정 시스템을 위한 형태소 분석을 위하여 한글 어절에 대한 구조분석 연구가 진행되어 왔다.

하지만 형태소 분석 시스템이 어떠한 일을 수행해야 하고, 어떤 기능을 갖추어야 하는지에 따라 한 어절을 어디까지 분석하는가가 달라진다. 형태소 분석기를 어떤 방법으로 설계하고 구현해야 할 것인지는 무엇을 위해 형태소 분석을 하는지에 따라 달라지게 된다. 그리하여 기계번역을 위해서는 번역의 등가성(translation equivalence)을 만족시키기 위해 분석의 깊이가 단순한 구문 분석 수준에서 의미 분석 수준에 이르기까지 다양하다.

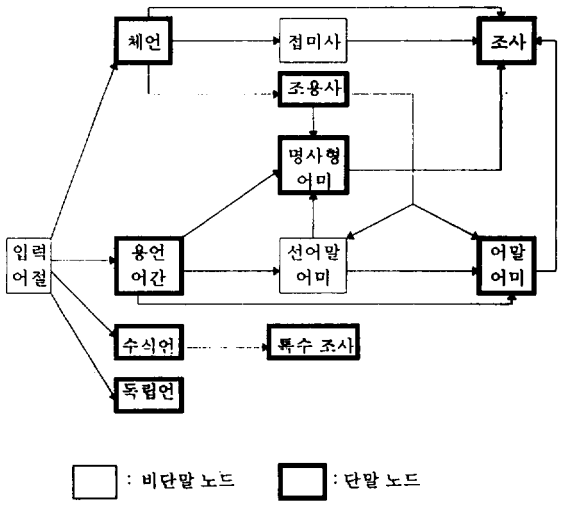
예를 들어 영한 번역 시스템의 경우는 영어의 전치사는 국어의 조사에 영향을 많이 끼치므로 구문이나 의미 분석까지 이루어져야 하며, 영어의 'He is a student'는 한국어의 '그는 학생이다'로 될 수 있겠지만 발화 상황을 고려한다면 한국어에서는 존비법이 잘 발달되어 있기 때문에, 청자가 화자보다 높은 사람이라면 '그는 학생입니다' 또는 '그분은 학생이십니다'로 되어야 한다.

또한 자동 색인에서는 형태소 분석 결과로부터 주제를 나타내는 단어나 구를 식별해 내며 구문 분석 수준에서 이루어지는 자동 색인에서는 문장의 구문 구조를 분석한 다음 전치사구나 명사구 등을 이루는 단어군을 찾아내고, 이 가운데 빈번하게 나타나는 단일어 또는 복합어를 색인어로 취하므로 한국어의 경우 용언의 활용에 크게 노력을 기울이지 않아도 된다.

4.2 오인식 문자 교정을 위한 한국어 어절 분석

본 논문에서는 한국어 어절 분석을 위해 다음과 같은 가정을 한다. [14]의 경우에는 기계번역이나 자연언어 이해 시스템을 위해 자세히 선어말어미를 분석하고자 하였으나, 오프라인 인쇄체 문서를 인식하고자 할 경우 인식하고자 하는 문서나 책에 나오는 단어나 어절은 회화체 형식은 거의 나오지 않고 대부분이 평서형이며 극 존칭 높임·공손 선어말 어미와 같은 어미의 형성은 처리 효율성을 위하여 제외한다. 예를 들어, 일상 문서에서는 거의 등장하지 않는 '잡으시었겠사오이다'나 '잡으시었겠사오더니라'를 처리하지는 않기로 한다.

본 논문에서 오인식 교정을 위해 구성한 어절 분석 상태 전이도는 (그림 3)과 같다.



(그림 3) 한글 어절의 상태 전이도
(Fig. 3) A state diagram of Hangul Eojeol

4.3 오인식 문자 교정을 위한 형태소 분석론

한국어의 형태소 분석론으로는 최장일치법(longest match stategy), Head-Tail 구분법, 접속 정보를 이용한 방법(tabular parsing method), 음절 단위 분석법, 사전을 중심으로 하는 방법 등 여러 가지 기법들이 제시되고 있다[15].

본 논문에서 이용하고자 하는 방법은 Head-tail 구분법[15]이다. Head-tail 구분법은 한국어 형태소 분석과 전반적인 형태론적 현상을 처리하는 방법에 대하

여 기술하고 있다. 단어를 구성하고 있는 형태소를 분리하기 위한 방법으로는 형태론적 변형을 중심으로 하여 단어를 변형되지 않는 부분(head)과 활용 현상이 일어나는 부분(tail)으로 구분한다. 이 방법에서는 먼저 분절 가능한 tail을 모두 찾아서 테이블을 구성한다. 이 테이블에 저장된 tail 정보로부터 head를 추정하는데 head가 발견되면 head와 tail간의 결합 관계를 확인하고 head에 대한 사전을 검색하는 과정으로 결과를 생성한다.

Head-tail 구분법은 불규칙 활용이나 음운 현상과 같은 형태론적 변형 문제를 해결하기 위해서 불규칙 용언을 처리 유형에 따라 세분화하여 유형을 분류하고 그들 사이의 결합 관계를 접속정보표로 구성한다. 이 방법에서 사용한 접속정보와 접속정보표는 단지 형태론적 변형을 처리하기 위하여 불규칙 용언과 어미에 대해서만 기술한 것이다. Head-tail 구분법은 단어를 head 부분과 tail 부분으로 구분하여 접속정보표를 이용하여 형태론적 변형을 처리하는 top-down 방식이다.

4.4 불규칙 활용과 유형

'체언+조사'로 구성된 어절이나 '규칙용언+어미'로 구성된 어절은 각 형태소를 구분하는 경계가 명확하지만, 불규칙 활용 어절인 경우에는 그 경계가 명확하지 않다. 그 이유는 불규칙 활용 어절에서 어간과 어미가 결합할 때 어간 혹은 어미의 일부가 변형되기 때문이다. 어간의 일부가 변형된 경우에는 변형된 어간을 어휘형태소 사전에 수록하거나 그 원형을 추정함으로써 형태소 분석을 할 수 있으나, 어미의 모양이 변형된 경우에는 주어진 형식형태소 사전만으로 변형된 어미를 분리해 낼 수 없기 때문에 형식형태소 사전에 변형된 어미를 추가하거나 아니면 다른 방법으로 처리하여야 한다. 그런데 어미의 모양이 변형되는 경우는 불규칙 용언이 '아/어'로 시작되는 어미와 결합할 때에만 발생하므로 불규칙 용언의 처리 문제는 '아/어'의 변이체를 어떻게 처리하느냐에 따라 실질형태소 사전의 내용과 문법형태소 사전의 내용뿐만 아니라 불규칙 활용 어절의 처리 방법도 달라질 수 있다[16].

국어학에서 분류하는 불규칙 활용에는 다음과 같은 것들이 있다.

1) ㄷ, ㅂ, ㅅ, 우, 르, 여, 러, 거라, 너라 불규칙 동사
2) ㅂ, ㅅ, 르, 여, 러, ㅎ불규칙 형용사

이들 불규칙 용언들은 개수가 그리 많지 않고 활용 형태도 특이하므로 불규칙 용언과 활용 형태를 일상 문서에 나오는 것들을 모두 수집하여 사전이나 테이블에 등록한다.

5. 실험 및 결과

5.1 실질형태소와 형식형태소의 분리

형태론적 변형들은 모두 형태소의 경계에서 일어난다는 공통점이 있고 형태소 경계에서 일어나는 형태론적 변형을 처리할 때는 형태소 분리 문제와 형태론적 변형의 복합되는 현상을 동시에 고려하여야 한다. 특히, 용언의 불규칙 활용에서 어간과 어미가 모두 변하는 경우 원형을 복원하는 문제는 형태소 분석 알고리즘을 더욱 복잡하게 하는 요인이다.

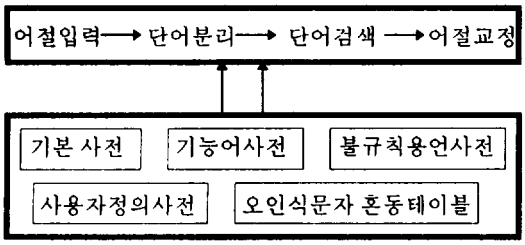
형태론적 변형을 처리하는 형태소 분석 알고리즘은 규칙으로 처리하는 방법, 사전을 기반으로 하여 처리하는 방법으로 분류할 수 있다. 대부분의 방법론들은 규칙을 기반으로 하는 방식을 취하고 있으나, 불규칙 활용이 일어난 어간이나 형태소 분석에 필요한 여러 가지 정보를 최대한 사전에 기술하는 사전을 중심으로 처리하는 방법과 빈도수가 높은 단어들의 분석결과를 기분석 사전에 수록함으로써 효율을 높이고 있다.

불규칙 활용이 일어나지 않는 어절은 형식형태소 사전에 의하여 바로 어간과 어미를 분리할 수 있고, 어간의 모양만 바뀐 경우에도 어미를 먼저 분리하면 되므로 어간과 어미를 쉽게 분리할 수 있다. 그러나 어미의 모양이 바뀐 경우에는 어미의 표층 형태가 형식형태소 사전에 존재하지 않으므로 원형을 복원한 후에 어미를 분리하여야 한다. 어간을 먼저 분리해 내는 방법이 있으나 어간이 수록된 실질형태소 사전은 크기가 매우 클 뿐만 아니라 어간과 어미가 하나의 음절로 통합되어 나타난 경우도 있으므로 비효율적이다. 더구나 어간과 어미가 모두 바뀐 경우에는 사전의 검색만으로 처리할 수 없다. 형태변이가 일어난 어간을 사전에 수록함으로써 처리할 수도 있으나, 그렇게 하면 하나의 어간에 대해 두 개 이상의 표제어가 사전에 존재하는 경우가 발생하여 각각의 표제

어에 대해 결합이 가능한 어미를 제약조건으로 기술해 주어야 하는 어려움이 있다. 따라서 형태소 사전에는 어간과 어미의 원형만을 수록하고 형태변이가 일어난 어간이나 어미는 표층 형태로부터 형태변이가 일어났을 가능성이 있는 어미를 추출해 낸 후에 어미 혹은 어간의 원형을 추론하는 방법으로 원형을 복원할 수 있다.

5.2 전체 시스템 구성

본 논문에서 오인식 교정을 위해 구성한 전체 시스템의 구성은 (그림 4)과 같다. 여기에서 기본 사전은 7만 8천여개의 단어가 수록된 초중고 교과서에서 등장하는 단어들로 학습된 사전[17, 18, 19]을 이용하였다.



(그림 4) 전체 시스템 구성도
(Fig. 4) An overall system

기능어 사전은 실질형태소와 형식형태소, 어간과 어미를 구별하기 위해서 조사가 연결된 형태와 규칙용언의 어미가 활용되는 형태를 수집하고 조사해 보았다. 기능어 사전에는 조사가 결합된 형태와 규칙 활용형의 어미가 1자 짜리 단어 96개, 2자 짜리 444개, 3자 짜리 645개, 4자 짜리 376개, 5자 짜리 112개, 6자 짜리 40개, 7자 짜리 1개로 모두 1,714개의 기능어들이 수록되어 있다. 8자 이상의 기능어들도 있을 수는 으나 문서나 서적에는 거의 등장하지 않고 있었다. 1,714개의 기능어 중에서 조사의 결합형은 900여개, 규칙 활용형의 어미 활용은 800여개이다. 규칙 활용형의 어미 활용은 현재까지 800여개를 수집하여 수록했으며 대략 이들 800여개로 어미 활용의 경우는 처리에 불편이 없었다. 어미 활용의 경우 조합 가능한 하지만 실제로는 어느 문서를 보아도 나타나지 않는 예, 예를 들어 '잡으시었겠사오더니라'와 같은 경우의 '으시었겠사오더니라'는 사용 빈도가 없어 사전에 등록하지는 않았다.

불규칙 용언 사전을 구축하기 위해서 본 논문에서는 [11, 16, 20]에서 보인 9가지의 불규칙 활용과 자동적 교체가 일어나는 현상, 불구동사의 활용을 수집하여 수록하였다. 현재까지 수집된 불규칙 용언의 활용형은 500개로 자주 쓰이는 용언은 거의 포함되어 있다. 이렇게 구축된 불규칙 용언과 불규칙 용언 활용형 500여개는 활용형을 이용하여 단어를 분리하고 단어검색을 하여 원형을 복원하여 어절을 교정할 수 있도록 하였다.

오인식 문자 혼동 테이블은 인식 시스템이 오인식을 발생하는 문자를 조사하여 테이블로 구성한 것이며 사용자 정의 사전은 고유명사나 특수명사등을 수록하기 위한 사전이다.

5.3 오인식 교정의 예

(그림 5)는 본 논문에서 실험을 위해 인식시스템으로 인식한 문서의 예이다.

(그림 4)에서의 사전과 Head-tail 구분법을 이용하여 단어를 분리하고 단어의 문자열 유사도(similarity)를 계산하여 후보 어절을 제시하도록 하였다. (그림 5)의 예를 가지고 어절을 교정하는 처리과정과 처리내용은 〈표 1〉에 나타내었다.

(그림 5)의 경우 인식 시스템을 가지고 인식하였을 때, 132문자 중에서 124문자가 올바로 인식이 되어 93.9%의 인식률을 보였으며, 이렇게 인식된 문서를 가지고 단어 교정을 하였을 때는 4자를 교정할 수 있어 인식률을 97%로 올릴 수 있었다.

5.4 본 시스템의 한계

본 시스템은 복합 명사나 미등록어의 처리는 할 수 없다. 기본 사전에 수많은 국어대사전에 나오는 단어들을 모두 등록하면 오히려 틀린 단어를 맞다고 하므로 오히려 정확률이 떨어질 수 있다. 또한 용언의 활용형이나 조사의 결합형이 지금까지 조사한 여러 활용형과 조사의 결합형을 벗어나는 경우도 분석에 실패할 수 있다.

(그림 5)에서 '국어학자'를 '국어화자'로 오인식한 경우, 현재 사전을 이용하면 '국어화자'라는 말이 특수하여 '국어학자'라고 올바로 제시해 주지만 만일 '국어화자'라는 말이 국어학에서는 많이 쓰이므로 이

이에 따라서 장난감 수준의 시스? 을 가지고 한국어 정보 처리가 끝난 것처럼 주장하는 국어화자가 있는가 하면, 전산학자들이 만든 시스? 을 무조건적으로 비판만 하는 화자들도 있다. 이 이면에는 전산학적 접근에 대한 불산과 자신의 화문에 대? 배타성, 그리고 그 반? 로, 컴퓨터의 능력에 대한 과신 등이 있음을 알 수 있다.

(그림 5) 인식시스템으로 인식한 문서의 예
(Fig. 5) A text example recognized by an OCR system

〈표 1〉 어절 교정의 예
〈Table 1〉 An example of Eojeol correction

처리 과정	처리 내용	처리 과정	처리 내용
시스?은	검사시작	반?로	검사시작
시스?＋은	기능어 사전을 이용하여 단어분리	반?＋로 반＋?로	기능어 사전을 이용하여 단어분리
시스?	기본 사전에서 문자열 유사도를 이용하여 후보생성	반＋?로 반?＋로	'기로', '대로', '므로', '에로', '으로'의 후보. 기본 사전에서 문자열 유사도를 이용하여 후보생성
시스?	'시스템', '시스터'의 후보 문자열 생성	반＋?로 반?＋로	'기로', '대로', '므로', '에로', '으로'의 후보. '반만', '반쯤', '반대' 등 198개의 후보 단어
시스템은(O) 시스터은(X)	접속정보를 이용하여 단어결합	반＋?로 반?＋로	접속정보를 이용하여 단어결합 '반대로', '반기로', '반에로', '반으로' '반대로', '반두로'등←61개의 후보
시스템은	올바른 후보 제시	'반대로', '반두로', '반기로', '반에로', '반으로' 등 63개의 후보	올바른 후보 제시

'국어화자'를 기본사전에 등록할 경우 맞는 말이 된다.

또한 희화체나 시에서와 같은 문체들이 나올 경우, 그리고 신문이나 보도매체에서와 같이 신조어가 많이 나올 경우도 올바른 후보 어절을 제시할 수 없게 된다.

6. 결론 및 향후 연구 과제

지금까지 본 논문에서는 오프라인 인쇄체 한글 문서 인식에서 발생하는 오인식 문자들을 교정하기 위해 단어 분석을 하는 방법에 관하여 제시하였다. 오

인식 문자 교정을 위해 품사 분류 기준으로 형식에 의하여 구별하고 사전을 구성하였으며, 형태소 분석론으로는 Head-tail구분법을 이용하였다. 처리가 까다로운 조사의 결합형과 용언의 활용형은 일상 생활에서 많이 쓰이는 형태들을 조사하여 이들을 이용하는 방법을 적용하였고 불규칙 활용의 경우도 모두 조사하여 교정에 이용하는 방법을 적용하였다.

향후 연구 과제로는 후보 어절을 생성할 때 여러가지 정보를 이용한 최적의 순위(ranking) 결정 방법, 복합 명사와 미등록어 처리, 신조어 처리, 자연언어 처리의 근본 문제인 중의성(ambiguity)해결에 관한 연구가 필요하다.

참 고 문 헌

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

[2] 홍승우, 이종현, 오상헌, "한글 어절 맞춤법 오류 검출을 위한 형태소 분석기," 한국정보과학회 가을 학술발표논문집, Vol. 20, No. 2, pp. 1143-1146, 1993.

[3] 이종연, 오상헌, "N-GRAM 한글 사전을 이용한 오인식 단어의 교정 알고리즘," 제5회 한글 및 한국어 정보처리 학술발표 논문집, pp. 271-283, 1993.

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

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

[6] 김재훈, 서정연, 자연언어 처리를 위한 한국어 품사 태그, 한국과학기술원, 인공지능연구센터, CAIR-TR-94-55, 1994.

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

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

[9] 황호정, 도정인, 권혁철, "한글 문자 인식을 위한 후처리기의 개발과 속도 개선," 제2회 문자 인식 워크샵, pp. 180-188, 1994.

[10] 조규빈, 하이라이트 고교문법, 지학사,1992.

[11] 남기심, 고영근, 표준 국어문법론, 탑출판사, 1993.

[12] 이 영 주, "자동 색인을 위한 한국어 형태소 분석 알고리즘," 한글 및 한국어 정보처리 학술 발표 논문집, pp. 240-246, 1989.

[13] 박영수, 김동석 외, 언어학개론, 형설출판사, 1993.

[14] 강승식, 김영택, "한국어 형태소 분석기에서 선어말어미의 분석 모형," 한국정보과학회 논문지, 제18권, 제5호, pp. 505-513, 1991.

[15] 강 승 식, "한국어의 형태론적 특성과 형태소 분석 기법," 한국정보과학회지, 제12권, 제8호, pp. 47-59, 1994.

[16] 강승식, 김영택, "한국어 형태소 분석기에서 불규칙 용언의 분석 모형," 한국정보과학회 논문지, 제19권, 제2호, pp. 151-164, 1992.

[17] 엣센스 국어사전, 민중서림, 1995.

[18] 원영섭, 초·중·고 국어 교과서에 나타난 띄어쓰기맞춤법 용례, 세창출판사, 1993.

[19] 이병희, 김태균, "한글 문자 인식에서 오인식 교정을 위한 오류 형태와 단어 학습에 관한 연구," 한국정보과학회 봄 학술발표논문집, 제23권, 제1호, pp. 301-304, 1996.

[20] 이철수, 국어형태학, 인하대출판부, 1994.

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