오프라인 인쇄체 문장부호, 일본 문자, 영문자, 한자 인식에서의 오인식 문자 교정에 관한 연구

이 병 희† · 김 태 균†

요 약

최근 상용 오프라인 문자 인식시스템들이 계속 발표되고 있다. 본 논문에서는 적은 메모리와 빠른 시간대에 검색이 가능한 자기조직화 데이타 구조를 가진 단어 사전을 구축하고 검색하는 알고리즘을 제시하며 오프라인 문자 인식 시스템을 이용하여 오인식 교정의 측면에서 문장부호, 일본문자, 영문자, 한자를 인식한 후에 나온 오인식된 문자들을 수집하여 오인식 형태를 재분류하였다. 영문자에 대해서는 영문자의 오인식 형태와 오인식의 예들을 조사하고 오인식이 자주 일어나는 글자들에 대해 오인식 혼동 테이블을 작성하였으며 25,145개의 영어 단어가 입력된 자기조직화된 영어 단어 사전을 가지고 교정을 행하여 0.5%의 인식률 향상을 가져왔다. 한자에 대해서도 영문자와 마찬가지로 오인식 형태를 조사하고 혼동 테이블을 작성하였으며 34,593개의 단어가 입력된 자기조직화된 한자 단어 사전을 이용하여 교정을 행하여 인식률을 6.1% 향상시켰다.

A Study on the Character Correction of the Wrongly Recognized Sentence Marks, Japanese, English, and Chinese Character in the Off-line Printed Character Recognition

Byeong-Hee Lee† · Tae-Kyun Kim†

ABSTRACT

In the recent years number of commercial off-line character recognition systems have been appeared in the Korean market. This paper describes a "self-organizing" data structure for representing a large dictionary which can be searched in real time and uses a practical amount of memory, and presents a study on the character correction for off-line printed sentence marks, Japanese, English, and Chinese character recognition. Self-organizing algorithm can be recommended as particularly appropriate when we have reasons to suspect that the accessing probabilities for individual words will change with time and theme. The wrongly recognized characters generated by OCR systems are collected and analyzed. Error types of English characters are reclassified and 0.5% errors are corrected using an English character confusion table with a self-organizing dictionary containing 25,145 English words. And also error types of Chinese characters are classified and 6.1% errors are corrected using a Chinese character confusion table with a self-organizing dictionary carrying 34,593 Chinese words.

† 정 회 원:충남대학교 컴퓨터공학과
논문접수:1996년 2월 8일, 심사완료:1996년 9월 16일

1. 서 론

대량의 문서를 신속하게 입력하기 위해서는 문서 자동 입력 장치 개발과 이들 장치를 통해서 들어온 영상을 인식하는 문자 인식 기술이 필요하다. 이러한 요구에 부응하여 문자 인식에 대한 연구가 활성화되면서 국내에서도 20여 년 전부터 지금까지 꾸준히 우리 글인 한글을 중심으로 한글 문자 인식 연구와 함께 영문자와 한자 인식에 관한 연구도 병행되어져 오고 있다[1].

문자 인식 시스템은 크게 전자펜과 같은 입력 장치를 통하여 글자를 쓰는 동시에 인식을 행하는 온라인(on-line) 문자 인식과 영상 스캐너(scanner)를 통하여 들어온 입력 영상을 인식하는 오프라인(off-line) 문자 인식으로 나눌 수 있는데 90년대를 접어들면서 국내에서는 주로 온라인 인식에 노력이 집중되어 온라인 영문자, 한글, 한자 인식 시스템과 오인식된 문자를 교정하는 시스템이 나오게 되었다.

최근에는 오프라인 문자 인식 시스템들이 상용화고 있다. 하지만 오프라인 인식 시스템을 통하여 인식된 문서에서 발생하게 되는 오인식을 언어적 지식을 이용하여 후처리를 행하는 연구는 국내에서 한글[2, 3]이나 영문자에 대해서는 몇몇 시스템이 나와 있지만 아직도 미흡한 실정이다. 특히 국내에서 오프라인 한자 인식에 관한 연구는 제한적이며 오프라인 한자 인식 시스템에서 발생하는 오인식을 교정하고자 하는 연구는 거의 발표된 예가 없다. 본 논문에서는 오프라인 문자 인식에서 발생하게 되는 오인식에 관하여 연구하고자 하며 이하 문자 혹은 문서 인식 시스템이라 하면 오프라인 문자 혹은 문서 인식 시스템을 일컫기로 한다.

오프라인 문자 인식 시스템을 구성하는 인식 알고리즘을 보면 인식 알고리즘이 글자 모양의 형태를 보고 인식하기 때문에 문자 모양이 다른 영문자, 한글, 한자의 인식 방법이 각각 다른 것이 대부분이다. 오프라인 문자 인식에 관한 연구와 시스템들은 지금까지 대부분 문장부호나 특수기호를 포함한 영문자 전용 인식과 한글 인식 전용, 한자 인식 전용등과 같이 어느 나라 문자 세트만을 인식하는 쪽으로 연구가 진행되어 왔다.

하지만 국내의 경우, 한 문서내에 문장부호, 특수기호, 영문자, 한글, 한자가 동시에 사용되고 있으므로 문장부호, 특수기호, 영문자, 한글, 한자 문자들이 한 문서내에 혼합되어 나타날 때 이들을 적당히 알아서 인식할 수 있는 인식 방법에 관한 연구와, 각각의 문자들이 혼용된 문서의 인식에서 발생하게 되는 오인식을 교정해 주는 문자 인식 후처리의 연구도 필요하다.

그리하여 최근에는 한 문서내에 여러 가지 문자가 동시에 사용될 때, 이들 여러 문자들을 동시에 인식하고자 하는 연구와 시스템들에 관한 연구가 이루어져 왔으며 몇몇 시스템들은 여러 문자들이 한 문서내에 혼용된 경우에도 이들을 인식할 수 있는 강건한(robust)한 인식 시스템들이 나오고 있다. 여러 가지 문자를 인식할 경우에는 인식 대상 문자들이 많아져 지금까지 각각의 나라 문자에 대해 각각 다른 인식 알고리즘을 적용하던 것을 통합적인 하나의 인식 알고리즘으로 만들기도 어렵고, 또한 인식하고자 하는 문자가 많아짐에 따라 인식 시간도 많이 걸리고 인식률도 떨어지게 된다.

오프라인 문자 인식의 오인식 교정에 관한 연구 측면에서도 지금까지는 여러 나라 문자가 동시에 나타날 때 이들을 교정하는 연구는 거의 없는 실정이다. 이에 본 논문에서는 오인식되는 문자들의 성질인 통계적 정보를 얻고 이 정보를 교정단계에서 이용하고자 문장 부호, 영문자, 일본문자, 한자의 문자들을 인식하고 인식할 때 발생하는 오인식을 고찰하여 교정 시스템에서 중요한 정보로 쓰일 문자 혼동 테이블(confusion table)을 구성하고자 한다.

물론 여러 나라 문자들이 동시에 나타날 때의 오인식 글자들의 통계적 정보를 얻으려고 하면 좋겠지만 이렇게 하기 위해서는 국내에서 쓰이고 있는 최소한 9,000여 글자에 달하는 2바이트 완성형 한글 한자 및 특수 문자에 나오는 오인식 정보를 얻어야 한다. 하지만 아직 이 9,000여 글자에 달하는 글자를 동시에 인식할 수 있는 시스템은 극소수의 문자 인식 시스템에서 이루어 진다고는 하나 본 연구에서 확인한 바로는 그런 시스템은 아직 없었다. 그리고 인식된 문자 코드를 가지고 문맥적 정보를 이용하여 오인식을 교정하는 문자 인식 후처리 측면에서도 어느 나라 문자가 다른 나라 문자로 잘못 인식하는 경우는 체크하기가 어렵지 않기 때문에 본 논문에서는 각 나라 문자에 대해서만 오인식을 비교하기로 한다.

본 논문에서는 적은 메모리와 빠른 시간내에 검색이 가능한 자기조직화 데이타 구조를 가진 단어 사전을 구축하고 검색하는 알고리즘을 제시하며 오프라인 문자 인식 시스템을 이용하여 문장부호, 일본문자, 영문자, 한자를 인식한 후에 나온 오인식된 문자들을 수집하여 오인식 형태와 원인을 비교·분석하고 교정을 행하는 방법에 대해서 제시한다.

2. 이론적 배경과 사전 검색 방법

2.1 이론적 배경

오프라인 문자 인식 시스템에서 오인식된 문자들을 교정하기 위해서는 인식 알고리즘의 특성, 즉 어느 글자가 어느 글자로 잘못 인식되는 경우의 수가 많다는 등의 통계적 자료와 단어의 출현 가능성 여부 등의 여러 가지 정보를 이용하여야 한다. 사람은 단어내 문자의 연결관계, 단어간의 연결관계, 문장의 구조, 문서의 주제 등의 다양한 문맥적 지식을 이용하여 비교적 정확하게 문서 인식을 수행한다. 그렇지만 컴퓨터가 사람이 사용하는 모든 문맥적 지식으로 의미를 이해하고 교정하는 것은 구현상의 어려움과 많은 계산량으로 인하여 그 효율성이 떨어지게 된다. 그리하여 일반적으로 자동 문서인식 시스템에서는 오인식되는 문자들의 통계적 정보와 단어내 문자의 연결관계에 관한 정보를 가지고 오인식 교정을 하게 된다[4].

문자 인식 시스템에서의 오인식 교정 알고리즘은 인식된 단어를 야기시킬 수 있는 단어 중에서 확률이 가장 높은 것을 입력 단어로 결정한다. 문자 인식 시스템에서 입력된 입력단어를 $Z=Z_1Z_2, ..., Z_n$이라 하고, 그 입력단어에 대한 인식단어를 $X=X_1X_2, ..., X_n$이라 하면, 인식단어가 X일때 입력단어가 Z일 확률 $P(Z|X)$는 Bayes의 정리에 의하여 다음과 같이 표현된다.

$$P(Z|X)=\frac{P(X|Z)\cdot P(Z)}{P(X)}$$

위의 식에서 $P(X|Z)$는 단어 Z가 입력되어 X로 인식될 단어간 혼동확률(confusion probability)을 나타내며 문자 인식 시스템의 특성을 반영한다. $P(Z)$와 $P(X)$는 각각 X와 Z의 사전확률(priori probability)을 나타낸다. 이와 같이 인식단어로 X가 주어졌을 때 모든 입력 가능한 단어중에서 위 식에서의 $P(Z|X)$를 가장 크게하는 단어 Z를 구하는 것이 본 오인식 교정 알고리즘의 목적이다.

2.2 사전 검색

자기조직화(self-organizing) 방법의 기본 개념은 자주 사용되는 단어를 단어사전의 앞 부분에 가져다 놓는 것이다. 이렇게 하면 수많은 단어들 중에서 흔히 사용되는 단어들이 앞으로 나오게 되고 사용되지 않는 단어들은 사전의 뒤에 위치하게 된다. 이 방법은 보통 글에서 나오는 단어는 수많은 단어들 중에서 한정되어 있다는 성질을 이용하자는 것으로 이렇게 하면 단어의 출현 확률을 이용할 수 있어 교정률이 좋아지며 단어사전에 나오는 모든 단어를 모두 검색하지 않아도 되며 글이나 문서에 따라 다르게 나오는 단어 출현 정보를 이용하기 쉬워진다. 이러한 것을 Zipf의 법칙이라 하며 실제적으로 구현하는 방법으로는 Move-to-Front 방법, Count를 이용하는 방법, Transpose 방법등이 있다[5, 6, 7].

보통 글에서 나오는 단어들의 출현 분포는 Zipf의 법칙에 따른다. Zipf의 법칙이란 텍스트에 있는 단어들을 가장 자주 나오는 단어를 앞에다 두는 순서로 정렬하였을 때 출현 빈도(frequency)와 출현 순위(rank)를 곱한 값이 상수가 된다는 것이다. 다시 말하면,

$$f_i \approx i f_i$$

여기서 f_i는 i번째 등장하는 단어의 출현 빈도이다. 이를 이용하여 Zipf의 확률 분포를 수식으로 표현하면,

$$p_i=\frac{1}{iH_n} \qquad (\text{단 } 1\le i\le n)$$

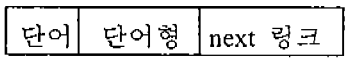
여기에서 H_n은 조화수열이며, 1차 모멘트와 분산은 다음과 같다.

$$\mu_1'=\frac{n}{H_n}$$

$$\mu_2'=\frac{n(n+1)}{2H_n}$$

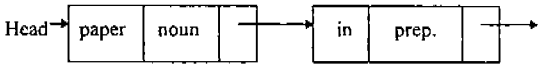
$$\sigma^2=\frac{n}{H_n}\left(\frac{n+1}{2}-\frac{n}{H_n}\right)$$

본 논문에서는 실제 단어 사전을 위한 데이타 구조와 알고리즘으로 Move-to-Front 방법을 이용하였다 [8]. 본 논문에서는 C언어를 이용하여 Move-to-Front 방법의 리스트(list) 데이타 구조를 정의하고 알고리즘을 구성하였다. 정의한 리스트 구조는 (그림 1)과 같다.



(그림 1) Move-to-Front 알고리즘을 위한 리스트 데이타 구조
(Fig. 1) The list data structure for the Move-to-Front algorithm

(그림 1)에 보인 리스트 구조를 이용하여 단어와 단어형을 저장하고 next 링크를 연결한 리스트 구현은 (그림 2)와 같다.



(그림 2) Move-to-Front 알고리즘을 위한 리스트 구현
(Fig. 2) The implementation for the Move-to-Front algorithm

이렇게 구성된 연결된 리스트 구조를 검색하여 해당단어를 앞으로 위치하게 하는 알고리즘을 C언어로 구현한 코드는 (그림 3)과 같다.

```
/* linked list 를 검색하여 해당단어를 앞으로 위치하게 한다. */
NODE *search(char key[])
{
        NODE *p, *q;

        if(head==NULL) return NULL;
        if(strcmp(key,head->word)==0) return head;
        p=head;
        while(p->next!=NULL) { /*  다음 단어가 끝이 아닐때*/
                if(strcmp(p->next->word,key)==0) /* key 발견*/
                {
                        q=head;
                        head=p->next;
                        p->next=p->next->next;
                        head->next=q;
                        return head;
                }
                p=p->next;
        }
        return NULL;
}
```

(그림 3) C언어로 작성된 Move-to-Front 알고리즘
(Fig. 3) Code description of the Move-to-Front algorithm in C

3. 문장부호, 숫자, 영문자의 오인식

영문자 인식은 영어 문화권을 중심으로 많은 연구가 있어 왔으며 문자의 개수가 많지 않고 문자의 모양이 그리 복잡하지 않아 상용 제품들도 이미 여럿이 나와 있는 실정이다.

문장 부호들을 인식할 경우, 문서내에 문장 부호가 영문자나 한글과 함께 나오게 되면 문장 부호는 영문자나 한글의 문자에 비해 '.'와 ','는 크기가 작아 잡영(noise)으로 인식되거나 영문자나 한글에 붙어서 영문자나 한글의 문자에 간섭 현상을 발생시켜 다른 문자를 오인식되게 하는 경우도 많았다.

문장에서는 특수문자들은 거의 나오지 않고 주로 느낌표, 물음표, 큰 따옴표, 작은 따옴표, ':'와 ';', '.'와 ','등이 자주 나타난다. 문자 인식 시스템의 경우 ':'와 ';', '.'와 ','사이에 각각 오인식이 자주 발생하며 ','의 상하 위치를 잘못 파악하여 ','와 ''간에 오인식이 발생했다.

영문자 'l'과 숫자 '1'간에도 오인식이 자주 발생했으며 영문자의 소문자 'o', 대문자 'O'와 숫자 '0'간에도 오인식 자주 발생하였으며 이들 문자는 글자 모양이 유사하여 인식 알고리즘만으로는 구별이 거의 불가능하여, 글자의 전후를 보아 교정하는 문자 인식 후처리가 필요하다.

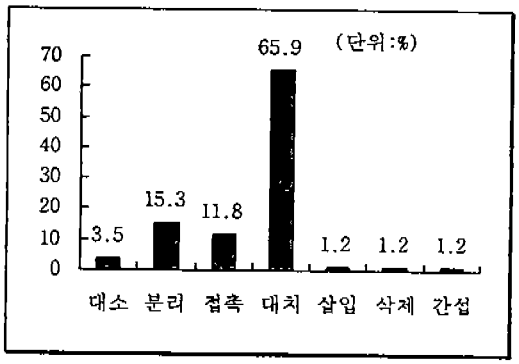
영문자는 모든 글자를 따로 떼어놓고 인식하였을 경우 한글이나 한자의 경우보다 인식률이 상당히 좋았지만 실제 문서에서와 같이 글자들이 모여 단어를 이룰 경우 인식률이 감소되는 경우가 많았다. 이는 글자간에 서로 붙거나 간섭 현상으로 인해 문자 분리가 어려워 인식률이 떨어지는 현상이다. 이를 통해 영문자의 경우는 문자 인식 전처리 과정에서 문자 분리 과정이 중요함을 알 수 있었다.

본 논문에서는 오인식의 유형과 예들을 본 연구실에서 나온 문자 인식 시스템과 상용 문자 인식 시스템들을 가지고 조사하여 보았으며 이들을 이용하여 오인식 혼동 테이블을 작성하였다. 〈표 1〉은 오프라인 문자 인식 시스템에 의한 영문자의 오인식 형태와 예를 나타낸다. 〈표 1〉에서 대소문자 오류는 대소문자를 구별하지 못해서 발생하는 오류이며, 분리와 접촉 오류는 문자 인식의 전처리시에 글자를 제대로 분리하지 못해서 발생하는 오류이며, 간섭 오류는 〈표 1〉의 예와 같이 ru가 따로 있을 때는 제대로 인식하던 것을 문서상에서 같이 붙어 있으면 서로 상대에게 간섭 현상을 보임으로써 두 글자 모두가 오인식 되는 형태이다.

〈표 1〉 오프라인 문자 인식 시스템에 의한 영문자 오인식의 형태와 예

〈Table 1〉 The types and examples of english character errors generated by off-line character recognition systems

오인식 형태	오인식의 예
대소문자 오류	c/C, k/K, o/O, p/P, s/S, u/U
분리 오류	m/rn, m/nr, o/0, n/rr, U/ll, W/VV
접촉 오류	rn/m, ll/H, if/K, ei/d, ck/d
대치 오류	a/o, e/s, f/t, i/j, l/I, n/m, y/V
삽입 오류	t
삭제 오류	i
간섭 오류	ru/ni



(그림 4) 오프라인 문자 인식 시스템에 의한 영문자 오인식의 비율

(Fig. 4) The percentages of english character errors generated by off-line character recognition systems

(그림 4)는 〈표 1〉에 나타난 오류 형태의 각각의 비율을 나타낸 것이다.

이렇게 볼 때 오프라인 영문자 오인식은 다음과 같은 특성들을 지니고 있다.

1. 한글[9]이나 한자에 비해 영문자는 크기가 작고 모양이 복잡하지 않다.
2. 영문자는 글자의 개수가 적은데 반해 글자체는 매우 다양하다.
3. 글자체가 한글이나 한자는 정사각형 형태를 따르는데 반해 영문자는 영문자의 가로의 길이가 세로의 길이의 보통 1/2밖에 되지 않는 직사각형 형태를 따른다.
4. 글자의 개수가 한글이나 한자에 비해 훨씬 적은 대신 대소문자간 글자의 유사성 때문이거나 혹은 영문자는 총 글자 개수중 한 글자가 차지하는 비율이 크므로 한 글자만 틀려도 인식률이 상당히 감소된다.
5. 영문자는 글자의 모양이 복잡하지는 않지만 장식선이나 글자체의 다양성으로 인해 인식이 어렵다.
6. 영문자는 문자 인식의 전처리 과정에서 글자와 글자간에 서로 붙어 있거나 큰 글자 밑으로 조그만 글자가 들어가 있는 경우가 있어 문자 분리가 어렵고, 특히 비스듬히 쓰여진 이탤릭체는 여러 가지 형태의 글자체가 혼용되어져 있는 경우, 문자 분리가 힘들어 인식률이 떨어 진다.

이와 같은 영문자 오인식 특성들은 문서편집기를 위한 철자 교정 알고리즘과 다른 점들이다.

4. 일본문자와 한자의 오인식

오프라인 인쇄체 일본 문자 인식은 본국인 일본에서 활발하며 오인식을 교정하는 연구[10]도 발표되고 있다. 요즘은 국내에서도 일본어가 많이 쓰이고 있으며 이에 따라 문자 인식에서도 일본 문자들도 인식되는 시스템들이 나오고 있다. 일본 문자로는 히라가나(ひらがな)와 가다가나(カタカナ) 그리고 한자(Kanji)가 쓰이고 있으며 일본 한자(Kanji)는 우리 나라에서 쓰이고 있는 한자와 다른 글자들이 상당수이다.

일본 문자에서 히라가나와 가다가나는 글자 모양이 복잡하지 않아 오인식이 그리 많지는 않았지만, は/ば/ぱ와 같은 청음(清音, せいおん)/탁음(濁音, だくおん)/반탁음(半濁音, はんだくおん)간에 탁음이나

반탁음의 소실, 탁음과 반탁음간의 오인식등이 있었으며, や와 ゃ에서와 같은 요음(拗音, ようおん) 오인식이 있었다.

또한 '力'와 'カ', '匕'와 'ヒ', '了'와 'ア' 같은 한자와 가다가나 문자간에도 글자 모양이 유사하여 오인식이 있었다.

한자는 한자 문화권인 중국, 대만, 일본[11, 12, 13, 14], 그리고 국내에서도 하드웨어적으로 구성한 [15]의 연구가 있으며 온라인 한자 인식[16]에 관한 연구는 오프라인 한자 인식에 비하여 활발한 편이며 오프라인 한자 인식을 위한 특징 추출(feature extraction)을 위한 연구[17]가 있다. 하지만 아직 오프라인 한자 인식에 관한 연구는 국내에서 부족한 부분이며 실제 생활에서 쓰일 수 있는 시스템이 나오기 위해서는 더욱더 연구가 필요하다.

한자는 중국에서는 6,763자, 대만에서는 5,401자의 한자가 사용되고 있으며, 일본에서는 산업표준(JIS)으로 6,349자의 한자(Kanji)가 있으나 약 2만자 정도의 한자가 사용되고 있다. 국내에서 한자는 KS C 5601 완성형코드로 4,888자가 쓰이고 있으며, 일상적으로는 교육용 기초 한자 1,800자가 주로 사용되며, 학술적으로 필요한 한자는 적어도 1만에서 5만을 넘는다. 한자 사전에서 한자를 분류하고 배열하는 방법은 우선 부수를 따라 나누고 그 다음에 획수를 따라 나열하여 놓으며 부수는 모두 214개이고 획수는 최대 64개까지가 있다[18].

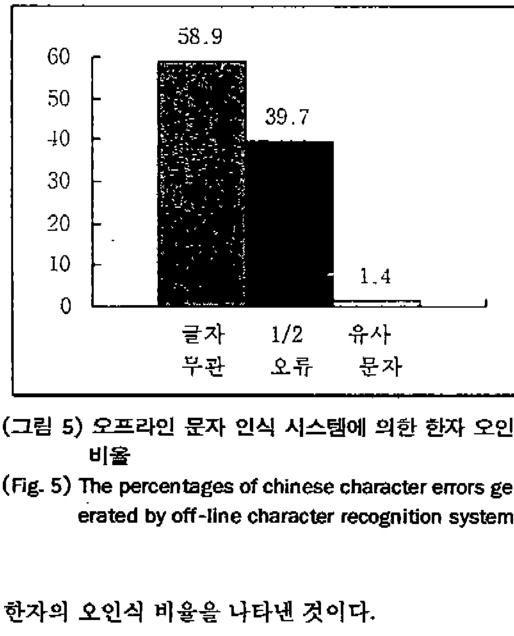
한자는 문자 인식을 하는데 있어서 주로 직선과 사선 성분으로 구성되어 있으나 획수가 많고 문자의 개수가 방대하여 인식에 많은 어려움이 있다. 한자는 글자의 수가 많아 학습에 시간이 많이 걸리고 획이 많은 글자일수록 오인식되는 경우가 많았다. 본 논문에서는 국내에서 쓰이고 있는 상용 한자 4,888자를 오프라인 한자 인식 시스템을 이용하여 인식하였을 때 발생하는 오인식을 유형별로 분류하고 〈표 2〉와 같이 한자 오인식 테이블을 구성하였다. 〈표 2〉에서 글자 모양에 무관한 오인식은 글자간에 모양이 유사한 점이 없는데도 잘못 오인식되는 형태이며 글자의 1/2오인식은 한 글자 모양의 1/2이하가 오인식된 형태이며, 유사 문자 오류는 문자가 유사하여 오인식되는 형태이다.

(그림 5)는 〈표 2〉의 오인식 형태에 따른 오프라인 한자의 오인식 비율을 나타낸 것이다.

〈표 2〉 오프라인 문자 인식 시스템에 의한 한자 오인식 형태와 예

〈Table 2〉 The types and examples of Chinese character errors generated by off-line character recognition systems

오인식 형태	오인식의 예
글자모양에 무관한 오인식	哥/懇, 穀/察, 珏/韆, 桿/抑, 僩/嗔, 截/職
글자의 1/2 오인식	訶/訴, 肝/肋, 蝎/蜈, 鉀/錦, 講/謙, 醵/醮
유사 문자로 인한 오인식	季/秊, 共/井, 訃/計, 跌/趺, 唄/俎, 皓/晧



(그림 5) 오프라인 문자 인식 시스템에 의한 한자 오인식 비율

(Fig. 5) The percentages of chinese character errors generated by off-line character recognition systems

이렇게 볼 때 오프라인 한자 오인식은 다음과 같은 특성들을 지니고 있다.

1. 한글이나 영문자에 비해 한자는 글자의 크기가 커서 전처리나 특징추출에 시간이 많이 걸리지만, 글자체는 그리 많지 않으며, 획수는 많아 글자 모양이 복잡하다.
2. 문자 인식의 전처리 과정에서 발생하는 문자 분리 실패로 인한 오인식은 한글이나 영문자보다는 적다.
3. 한자는 글자의 개수가 많아 유사문자가 많을 것 같지만 비율로 보면 영문자나 한글에 비해 적었다.
4. 한자는 영문자나 한글에 비해 총글자의 개수 중에서 자주 쓰이는 글자가 한정되어 있다.
5. 한자는 획수가 많음으로 인해 영문자나 한글보다는 높은 해상도의 영상 입력 장치로 입력받아야 한다.

5. 영문자, 한글, 일본문자, 한자 오인식의 비교

본 장에서는 앞서 언급한 영문자, 한글, 일본문자, 한자 오인식에 대해서 각 글자들이 각기 독특한 특성들을 가지고 있고 오류형태가 달라 비교가 어렵기는 하지만 오프라인 인쇄체 문자 인식이라는 측면에서 비교·분석하도록 한다. 한글에 대한 오인식은 [9]에서 나온 결과를 가지고 비교하도록 한다.

우선, 영문자는 오인식 형태를 대소문자 오류, 분리 오류, 접촉 오류, 대치 오류, 삽입 오류, 삭제 오류, 간섭 오류로 나누었으며, 한글은 글자형태무관, 1자소 오류, 2자소 오류, 1획 첨가, 1획 탈락으로, 일본 문자는 글자 형태에 무관한 오인식, 탁음과 반탁음간의 오인식, 요음의 오인식, 한자와 가나문자간 오인식으로, 한자는 글자모양에 무관한 오인식, 글자의 1/2오인식, 유사 문자로 인한 오인식으로 나누었다.

영문자의 오인식은 문자 인식 전처리 과정에서 문자 분리의 오류로 인해 오인식을 발생시키는 분리 오류, 접촉 오류, 간섭 오류가 발생하였으며, 영문자는 특히 문자 분리가 문자 인식을 하기 위해서는 매우 중요하였다. 영문자는 또한 영문자의 특수성이라 볼 수 있는 대소문자 오류가 있었으며 대치 오류가 65.9%로 가장 많았다.

한글은 다른 나라 문자들에 비해 초성, 중성, 종성이 조합하여 쓰인다는 특성이 있었으며 초성, 중성, 종성의 조합을 문자 모양에 따라 구분한 한글 6형식에 글자들이 포함되므로 한 자소만 오인식하는 1자소의 오류, 두 자소를 오인식하는 2자소 오류가 있었다. 그리고 1획 정도의 첨가와 탈락 오류가 있었으며 글자형태 무관 오류가 64%이었다.

일본 문자는 일본 문자의 특성이라 볼 수 있는 탁음과 반탁음, 요음의 오인식이 있었으며, 한자와 가다가나 문자간에도 오인식이 있었다.

한자는 문자수가 방대함과 국내의 경우는 아직 일본이나 대만, 중국에 비해 문자 인식 기술이 약간 떨어지고 있었으며 글자 모양에 무관한 오인식이 58.9%나 되었다. 한자는 또한 둘 이상의 글자가 합하여 또 다른 한 문자를 만드는 회의문자와 형성문자가 많아 한 문자의 1/2이하가 오인식되는 특성들이 있었다.

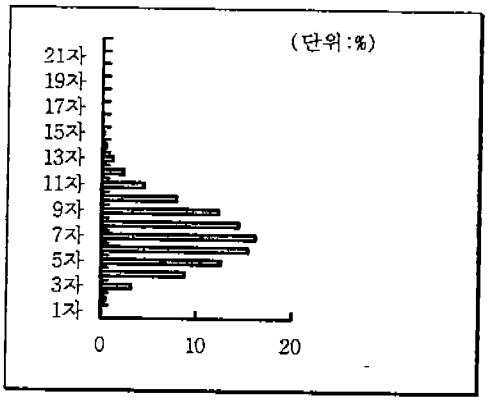
영문자, 한글, 일본문자, 한자에 대해서 공통적으로 글자가 대치되는 오류가 가장 많았으며 각 나라 문자들은 위에서 언급한 바와 같은 특성들을 각각 지니고 있었다.

6. 오인식 교정

오프라인 문자 인식 시스템을 가지고 오인식된 결과를 교정하기 위해서 본 논문에서는 신문, 잡지, 간행물 등을 영상 입력장치인 HP ScanJet 4c 스캐너상에서 300dpi로 입력 받아 상용 소프트웨어인 한국인식 기술의 HiART 글눈과 합산컴퓨터의 아르미로 인식한 다음 오인식된 문자들 교정하여 보았다.

6.1 영문자 오인식 교정

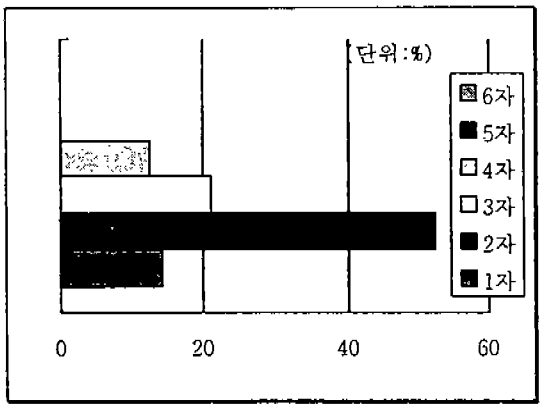
영문자 오인식 교정을 위한 사전으로 영어 단어를 글자 개수별로 나누어 총 25,145단어를 입력하였다. 입력한 영어 단어의 평균 글자수는 7.22이었다. 영어 단어의 개수별 점유 비율은 (그림 6)과 같다.



(그림 6) 영문자의 단어별 개수 점유율
(Fig. 6) The percentages of english words by length

또한 본 연구에서는 국내에서 자주 쓰이는 상용 한자 4,888자 및 한자 단어를 입력하여 총 34,593개의 단어 사전을 구축하였다. 입력한 한자 단어의 평균 글자수는 2. 32자 이었으며 단어 개수별 점유 비율은 (그림 7)과 같다[19].

(그림 8)는 영문자를 오프라인 문자 인식 시스템을 가지고 인식하였을 때의 데이타이며 밑줄 그은 부분은 오인식이 일어난 단어이다. 영문자의 경우는 문자



(그림 7) 한자의 단어별 개수의 점유 비율
(Fig. 7) The percentages of chinese words by length

의 수가 적어 오인식이 발생하는 패턴을 쉽게 분석할 수 있으며 한국어와 같은 복잡한 형태소 분석이 필요하지 않아 교정이 비교적 쉽다. (그림 6)에서 보듯 단어의 분포가 어느 한 곳에 집중되지 않고 널리 분포되어 있어 한 글자가 틀렸다고 해도 어렵지 않게 정확한 단어를 찾을 수 있었다.

(그림 8)에서 밑줄 친 오인식된 단어들 "ln"과 "lt"을 교정하는 과정은 다음과 같다.

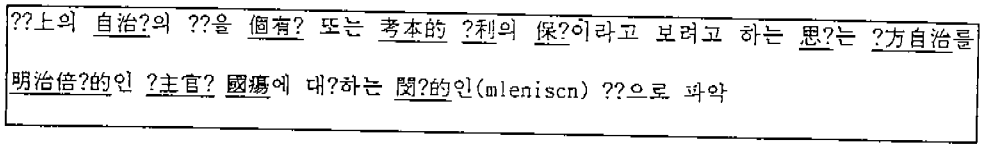
1. 복수형, 소유격, 3인칭 단수, 과거/과거분사/현재분사, 비교급/최상급 등의 접미사(suffix)를 분리한다.
2. 대문자를 소문자로 변환한 후, 문자열 유사도(string similarity)를 이용하여 단어 사전을 참조하여 후보 문자열을 생성한다. 후보 문자열을 생성할 때, 중복된 단어들은 제거한다.
3. 첫 글자의 대문자여부와 접속관계를 이용하여 분리된 단어를 결합하여 후보문자열을 제시한다. 이때 영문자 오인식의 통계적 정보로 구성된 영문자 혼동 테이블의 정보를 함께 이용하여 후보열을 제시하게 된다.

"ln"의 경우 기존의 글자열 매칭을 통하여 교정하고자 하면, "an", "en", "in", "on", "la", "lo"와 같은 후보 문자열이 생성될 수 있으나 'l'은 'I'와 오인식이 자주 발생한다는 영문자 오인식 혼동 테이블 정보를 이용하면 정확한 후보는 "In"이라는 것을 알 수 있다. "lt"도 기존의 단어사전 정보를 이용할 경우, "at", "et", "Ft", "it", "Mt", "St", "la", "lo"등의 후보 문자열을 제시할 수 있으나, 'l'은 'I'와 오인식이 자주 발생한다는 영문자 오인식 혼동 테이블 정보를 이용하면 정확한 후보는 "It"이라는 것을 알 수 있다.

(그림 8)과 같은 문서를 인식하였을 때 인식률이 99.1%되는 것을 본 교정을 거쳤을 때 99.6%가 되어 0.5%의 인식률 향상을 가져올 수 있었다.

ln this paper the Maximum Block Method is suggested for the Feature Extraction of Strokes of off-line Chinese characters. The Maximum Block Method is a technique which enlarges the block from the first found pixel that extracts the skeleton and features of the input characters. lt was proven that the suggested algorithm is suitable for the Chinese characters, because of the characteristics of the language. Chinese characters are composed of the complex strokes of straight and oblique lines and also comprised of cross points

(그림 8) 오프라인 문자 인식 시스템으로 인식한 영문자 문서의 예
(Fig. 8) An example of an english document recognized by an off-line character recognition system



(그림 9) 오프라인 문자 인식 시스템으로 인식한 한자 문서의 예

(Fig. 9) An example of a chinese document recognized by an off-line character recognition system

6.2 한자의 오인식 교정

(그림 9)는 오프라인 문자 인식 시스템을 가지고 한자가 포함된 문서를 인식하였을 때의 데이타이다. 한자의 경우도 한글과 같은 복잡한 형태소 분석이 필요하지 않다. 오프라인 한자 인식에서는 문자 분리 과정에서 오분리가 그리 발생하지 않는 특성이 있으므로 글자 개수별로 단어들을 사전에 등록하였다. 그리하여 한자 교정에서 글자열 매칭(string matching)을 이용하여 교정 후보 단어들을 생성할 수 있게 하였다. 본 논문에서는 각 글자 개수별로 34,593 단어가 입력된 한자 단어 사전을 이용하여 글자열 매칭을 이용하여 교정을 해 보았다.

〈표 3〉는 (그림 9)의 데이타를 글자 개수별 한자 단어 사전을 가지고 글자열을 매칭함으로써 교정하였을 때의 오인식 단어와 이 오인식 단어의 후보 단어를 나타낸다. 한자를 교정하고자 할 때 "?利"나 "保?", "國瘍"의 경우에서 보듯이 2자짜리 단어의 경우는 교정 후보가 너무 많이 생성되며 교정하고자 하는 단어의 순위도 낮아 교정에 어려움이 많았다. "自治?"와 "?方自治"와 같은 3자짜리나 4자짜리의 단어의 경우는 교정이 쉬웠다. "明治倍?的"의 경우는 단어가 자주 쓰이지 않는 단어여서 사전에 등록되지 않아 교정에 실패한 경우이며, "?主官?"와 "関?的"의 경우는 단어 중에서 틀린 글자의 개수가 많아 교정을 할 수 없는 경우이다.

〈표 3〉 오인식 단어와 교정 후보 단어

〈Table 3〉 The wrongly recognized words and their candidate words

오인식단어	교정 후보 단어
自治?	自治國, 自治權, 自治領, 自治的, 自治會
個有?	固有法, 固有性, 固有質
考本的	根本的, 基本的
?利	高利, 公利, 功利등 32개의 후보 단어
保?	保健, 保身등 29개의 후보 단어
思?	思量, 思想등 12개의 후보 단어
?方自治	地方自治
明治倍?的	후보 단어 없음
?主官?	후보 단어 없음
國瘍	國歌, 火傷등 91개의 후보 단어
関?的	293개의 후보 단어

(그림 9)와 같은 문서를 인식하였을 때 인식률이 30.3%되는 것을 본 교정을 거쳤을 때 10위 안에 올바른 단어가 들어가는 경우를 조사하여 보니 45.4%가 되어 15.1%의 인식률 향상을 가져올 수 있었고 1위 후보로 교정된다고 하였을 때는 6.1%의 인식률을 향상시킬 수 있었다.

7. 결론 및 향후 연구 과제

지금까지 본 논문에서는 적은 메모리와 빠른 시간 내에 검색이 가능한 자기조직화 데이타 구조를 가진 단어 사전을 구축하고 검색하는 알고리즘을 제시하였으며 문장부호, 일본문자, 영문자, 한자 문자 인식 후처리에 대하여 알아 보았다. 문장 부호들을 인식할 경우에 특수문자들은 거의 나오지 않고 주로 느낌표, 물음표, 큰 따옴표, 작은 따옴표, ':'와 ';', '.'와 ','등이 자주 나타났으며, 문자 인식 시스템의 경우 ':'와 ';', '.'와 ','사이에 각각 오인식이 자주 발생하며 ','의 상하 위치를 잘못 파악하여 ','와 '''간에 오인식이 자주 발생했다.

·오프라인 영문자 인식의 경우는 문자들의 수가 적기는 하지만 글자체가 다양하고 작아 문자 인식의 전처리 과정에서의 문자 분리에 실패하는 경우가 많았

다. 하지만 영문자 인식 후처리에서는 영어 단어들의 특성이 비교적 간단하고 처리가 쉬워 교정에 큰 어려움은 없었다.

오프라인 일본 문자중에서 히라가나와 가다가나에서는 청음/탁음/반탁음에서 탁음이나 반탁음의 소실, 탁음과 반탁음간에 오인식이 자주 발생했으며, 문자의 크기를 잘못 판단하여 생기는 요음의 오인식도 자주 발생했다. 또한 한자와 가나문자간에 글자 모양이 유사하여 발생하는 오인식도 있었다.

본 논문에서는 오프라인 한자의 오인식 특성에 대한 기초 연구로 지금까지 국내에서 제한된 개수의 문자 수준에서만 연구가 행하여 오던 것을 상용한자 4,888자에 대해 한자의 오인식 오류 유형과 예들을 고찰하였다. 한자의 경우는 국내의 경우 아직 한글이나 영문자의 인식률에 상당하는 인식 시스템이 거의 없고 글자 자체도 복잡하고 글자수도 많아 인식도 어렵고 한자 단어의 측면에서도 2자나 3자짜리 단어가 대부분이어서 교정에도 어려움이 많았다.

앞으로 한자의 오인식 교정률 향상을 위해서는 한자의 인식률 향상도 이루어짐과 동시에 오인식 특성에 관한 연구들도 이루어져야 할 것이다. 또한 한자의 2자짜리 단어가 많음으로 인해 발생하는 교정의 어려움도 해결하여야 할 문제로 남는다.

마지막으로 국내의 경우는 한 문서내에 문장부호, 특수부호, 영문자, 한글, 한자들이 동시에 쓰이고 있다는 점을 감안하여 이들 문자들이 한 문서내에 동시에 나타날 때 발생하는 오인식에 관한 종합적인 오인식 교정을 위한 연구가 필요하리라 본다.

참 고 문 헌

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

[2] 박진규, "한글 문서 인식 시스템의 오인식 수정에 관한 연구," 한국과학기술원 석사학위 논문, 1988.

[3] 홍남희, 이원일, 이종혁, 이근배, "어절 정보와 문자열 정보를 이용한 문자 인식에서의 오인식 수정 기법에 관한 연구," 제1회 문자인식 워크샵 논문집, pp. 109-113, 1993.

[4] 민병우, 이성환, 김홍기, "문자 인식을 위한 후처리 기법의 사례 연구," 제 1회 문자 인식 워크샵 발표 논문집, 1993.

[5] Bentley, and C.C. McGeoch, "Amortized Analyses of Self-Organizing Sequential Search Heuristics," Communications of ACM, Vol.28, No.4, pp.404-411, 1985.

[6] C.J. Wellis, L.J. Evett, P.E. Whitby, and R.J. Whitrow, "Fast Dictionary Look-Up for Contextual Word Recognition," Pattern Recognition, Vol.23, No.5, pp.501-508, 1990.

[7] G.H. Gonnet, and R.Baeza-Yates, Handbook of Algorithms and Data Structures in Pascal and C, Addison-Wesley, 1991.

[8] 이병희, 이인동, 김태균, "한국어 오인식 수정을 위한 자기 구조화 휴리스틱스에 기반한 사전 구성," 한국정보과학회 가을 학술발표논문집, 제20권, 제2호, pp.1155-1158, 1993.

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

[10] 高尾哲康, 西野文人, "日本語文書リーダ後處理の實現と評價," 일본정보처리학회논문지, Vol.30, No.11, 1989.

[11] Jun S. Huang and Ma-Lung Chung, "Separating Similar Complex Chinese Characters By Walsh Transform," Pattern Recognition, Vol.20, No.4, pp.425-428, 1987.

[12] Hsi-Jian Lee and Bin Chen, "Recognition of Handwritten Chinese Characters Via Short Line Segments," Pattern Recognition, Vol.25, No.5, pp.543-552, 1992.

[13] S. Hsu, K. Takahashi, S. Ozawa and H. Fujita, "印刷漢字の順序情報をったスト-ク抽出法," 일본 전자통신학회논문지, Vol.J65-D, No.2, 1982.

[14] Kahan, T. Pavilidis, and H.S. Baird, "On the Recognition of Printed characters of any font and size," IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.PAMI-9, pp.274-288, March, 1994.

[15] 여진경, 이우일, 정호선, IDMLP를 이용한 한자 인식에 관한 연구, 대한전자공학회 논문지, 제28

권 B편, 제10호, 1991.

[16] 김기철, 이성환, "이완 정합을 이용한 획순에 무관하게 흘려 쓴 한자의 온라인 인식," 한국정보과학회논문지, 제22권, 제3호, pp.423-430, 1995.

[17] 김의정, 김태균, "오프라인 한자 특징을 위한 최대 블록화 방법에 관한 연구," 한국정보처리회 봄 학술발표논문집, 제22권, 제1호, pp.225-228, 1995.

[18] 김정수, 한자 자료의 전산화를 위해서 새김의 표준화를 제안함, 1989년도 한글 및 한국어정보처리 학술발표논문집, pp.107-113, 1994.

[19] 이병희, 김태균, "오프라인 영/한자 문자 인식을 위한 오인식 형태와 교정에 관한 연구," 한국정보처리학회 봄 학술발표논문집, 제23권, 제1호, pp.568-571, 1996.

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