論文98-35C-2-8

어절 정보를 이용한 한국어 문자 인식 후처리 기법

(A Postprocessing Method for Korean Optical Character Recognition using Eojeol Information)

李 英 和 *, 金 桂 成 *, 金 永 勳 **, 李 相 祚 *

(Young-Hwa Lee, Kye-Sung Kim, Young-Hun Kim, and Sang-Jo Lee)

요 약

본 논문에서는 한국어 문장에서 얻은 언어적, 어절 정보를 이용하여 문서 인식기의 오류를 검출하고 교정하고자 한다. 먼저, 한국어 문장을 어절 단위로 분석한 결과, 어절 내의 구성 성분을 16가지로 분류하였다. 이 성분을 어절 구성 상태 전이도로 나타낸 다음 어절 구성 성분간 좌우 접속 정보를 얻었다. 이 정보는 오인식 형태소의 품사를 유추하여 오인식 어절의 가능한 형태소 분석 수를 제한하여 생성될 후보 문자의 수를 줄였다. 그리고, 어절간 인접 정보를 조사하여 휴리스틱 정보로 이용하여 교정 속도를 향상시켰다. 교정에서는 어휘에 대한 역순 사전을 재구성해 두고 오인식 문자의 위치에 상관없이 교정 후보를 사전에서 탐색할 수 있도록 하였다. 이 결과 기존의 후처리기를 포함한 문서 인식기의 성능을 형태소 분석, 오인식 검출, 후보 문자 수 측면에서 개선하였고, 인식률 97.03%를 98.02%로 향상시켰다.

Abstract

In this paper, we will to check and to correct mis_recognized word using Eojeol information. First, we divided into 16 classes that constituents in a Eojeol after we analyzed Korean statement into Eojeol units. Eojeol-Constituent state diagram constructed these constituents, find the Left-Right Connectivity Information. As analogized the speech of connectivity information, reduced the number of cadidate words and restricted case of morphological analysis for mis-recognition Eojeol. Then, we improved correction speed using heuristic information as the adjacency information for Eojeol each other. In the correction phrase, construct Reverse-Order Word Dictionary. Using this, we can trace word dictionary regardless of mis-recongnition word position. Its results show that improvment of recognition rate from 97.03% to 98.02% and check rate, reduction of chadidate words and morphlogical analysis cases.

I. 서 론

컴퓨터를 이용한 정보처리가 일반화되면서 많은 분야에서 문서 인식 시스템의 사용을 필요로 하고 있다. 그러나, 입력 문서 자체의 오류, 인식 기술의 한계성과 문자간의 구조적 유사성등으로 인한 오인식으로 완벽한 문서 인식은 불가능한 현실이다. 오인식 문자에 대해서 문서 인식기 자체에서도 후처리를 하고 있지만 시스템의 실용화를 위해 문서 인식 결과에 대한 교정은 보다 많은 언어적 정보를 이용한 별도의 후처기가 필요하다. 따라서, 본 논문에서는 한국어 문장을 어절 단위로 분석하여 얻은 어절 정보를 이용하여 오인식 문자를 교정하고자 한다.

* 正會員, 慶北大學校 컴퓨터工學科
(Dept. of Computer Engineering, Kyungpook National Universirty)
** 正會員, 安東 專門大 電子計算科
(Dept. of Computer Engineering, Andong College)
接受日字:1997年8月29日, 수정완료일:1998年2月5日

형태소 분석과 문자/자소 대치에 의한 후처리 기법[1]은 한국어의 언어적 정보를 이용한다는 데 의미가 있지만, 형태소 분석에 실패한 어절을 구조적 유사 문자/자소 대치로 교정할 경우 후보 문자의 수가 상당히 많으며, 이 중에는 어절 구성에 맞지 않는 문자도 많이 포함되어 있다. 또한, 생성된 대량의 후보가 어절 구성 성분으로서 적합한지를 판정하기위해 다시 형태소 분석을 하거나 사전을 검색해야 하는 번거로움이 있다.

본 논문에서는 한국어 문장을 어절 단위로 분석한 결과, 어절 내의 구성 성분을 16가지로 분류하여 어절 구성 상태 전이도에 따라 어절 구성 성분간 좌우 접속 정보를 이용함으로써 오인식 어절의 가능한 형태소 분석 수를 제한하여 생성될 후보 문자의 수를 줄였다. 그리고, 인식기가 범하는 오류 중에서 오인식 어절에 인접한 어절들을 조사하여 휴리스틱 정보로 이용하여 교정 속도를 향상시켰으며, 어휘에 대한 역순 사전을 재구성해 둠으로써 오인식 문자의 위치에 상관없이 교정 후보를 사전에서 탐색할 수 있도록 하였다.

본 논문에서 제시한 후처리 시스템은 오인식 문자의 검출과 교정단계로 나누어 구현하였다. 오인식 문자 검출기는 양방향 최장 일치에 의한 형태소 분석기[2]를 개선하여 오인식 어절을 검출하고, 어절내의 접속 정보로 오인식 문자를 포함한 형태소의 품사를 유추한다. 이것은 생성될 교정 후보 문자의 수를 줄이고 언어적 정보를 이용함으로써 의미없는 후보 문자의 생성을 막는다. 다음으로 교정 단계에서는 검출 결과와 역순 사전, 유사 문자의 구조 정보[1] 그리고 어절 간 인접 정보로 후보 문자를 생성한다. 후보 문자의 신뢰도는 구조적 유사도와 자소 및 문자 발생 확률에 의해 평가한다.

논문의 구성은 1장 서론에 이어 2장에서는 후처리에 이용한 어절 정보를 설명하고 3장에서는 전체 시스템의 구성을 단계별로 소개하였다. 4장에서는 실험 및 결과를 보이고 끝으로 5장에서 결론을 다루었다.

Ⅱ. 어절 정보

본 논문에서 이용한 어절 정보는 크게 어절내 정보와 어절간 정보 두 가지로 분류된다. 먼저 어절내의 정보를 살펴보면, 어절내의 구성 성분간 접속 정보와 격조사가 포함된 어절의 종성 유무에 따른 조사 유형 정보가 있다.

한국어 어절의 구성 상태 전이도는 사용 목적이나 사용자에 따라 조금씩 차이를 보인다[3] [4] [5]. 본 논문에서는 오인식 문자 검출에 사용한 양방향 최장 일치 형태소 분석기에서 분석되는 형태소를 기준으로 어절 상태 전이도를 구성하였다. 형태소 분석과 사전에 없는 미등록어(오인식 어절 포함)의 품사 유추가 용이하도록 구성 성분을 16가지로 나누었으며, 표 1에서 품사별 접속 번호를 보이고 있다.

표 1. 품사별 번호
Table 1. A Part of Speech Number.

번호	품사	번호	품사
1	체언	9	선어말어미
2	특수조사(도,만,조차, ...)	10	어미
3	접미사	11	명사형어미(ㅁ,음,기)
4	격조사	12	관형사형어미(ㄴ,ㄹ, ...)
5	-하-/-되-/-시키-	13	부사형 어미(이,히)
6	-이-	14	부사
7	용언	15	관형사
8	보조용언	16	감탄사

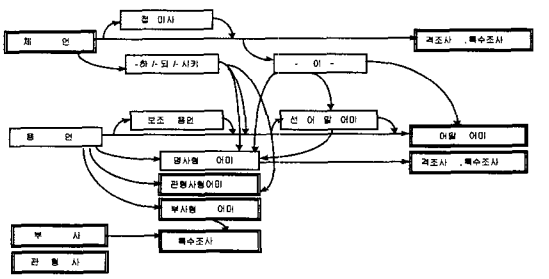

그림 1. 한국어 어절의 상태 전이도
Fig. 1. A State Diagram of Korean Eojeol.

그림 1은 본 논문에서 제시하는 한국어 어절의 구성 상태 전이도이다. 그림 1을 살펴보면, 오인식 형태소의 품사를 보다 정확하게 유추하기 위해 조사를 특수조사와 격조사로, 전성어미를 명사형, 관형사형, 부사형 전성 어미로 세분하였다. 또한 명사에 "-하다/-되다/-시키다"가 결합된 어절은 사용빈도가 높으므로 순수한 체언과 구분하였고 보조용언도 테이블을 구성하여 용언과 분리시켰다. [3]에서는 특수조사가 수식언에만 연결되어 있어, 이 경우 "책상까지"가 "책상까지"로 오인식되었다면 "책상"을 수식언(관형사,부사)

으로만 추정하므로 올바른 후보를 생성하지 못한다. 하지만 본 논문에서 제안한 어절 구성 상태 전이도는 체언으로 추정하여 적절한 교정 후보를 생성한다. 그림 1에서 겹네모는 어절 구성이 종료됨을 의미한다.

상태 전이도는 형태소간 좌우 접속 정보표로 재구성하여 어절내의 형태소 각각이 좌우에 가질 수 있는 형태소를 표시하여 둠으로써 오인식 문자와 미등록어로 인해 형태소 분석이 실패한 경우에도 좌우에 접속된 다른 형태소에 의해 형태소의 품사를 유추할 수 있다. 또한, 오인식 문자의 위치가 어절의 앞부분에 있더라도 인접한 형태소의 접속 정보를 이용하여 품사를 유추할 수 있다. 예를 들어, "정리한"이 "평리한"으로 오인식된 경우에 형태소 분석기의 출력은 다음과 같다.

(1) 평리(?) + -하 + ㄴ
(2) 평리하(?) + ㄴ
(3) 평리한(?)

(1)에서 "평리"는 "-하"의 좌접속 정보에 의해 체언 중에서 "-하"와 결합이 가능한 체언으로 품사가 유추된다. (2)의 경우에는 어미 "ㄴ"의 좌접속 정보로 인해 "평리하"가 용언으로, (3)은 독립어로 사용이 가능한 체언, 부사, 관형사로 유추된다.

어절내의 또 다른 정보는 앞 형태소의 종성 유무에 따라 사용이 달라지는 조사와 어미의 형태이다. 조사와 어미 "은/는", "이/가","을/를", "아/야", "과/와", "으로/로", "았/었/였" 등이 예이며, "였"은 "하다"동사 뒤에서만 사용되는 형태론적 조건의 "았/었"에 대한 이형태이다. 형태소 분석 결과 이러한 조사와 어미가 나타난 어절에서는 좌 접속 형태소의 끝 음절에 종성이 있는지 없는지를 조사하여 교정에 이용한다.

한국어 문장에서 한 문장 내에 있는 어절들은 상호간 의미적 관계를 가지고 있다. 이 관계를 어절의 나열 순서 측면에서 보면, 간단한 형태로 특정한 두 어절이 인접하여 사용되는 경우와 그렇지 않은 경우가 있다. 거의 고정된 순서로 인접하여 사용되는 어절들 중에서 인식기가 오류를 발생하는 예를 조사하여 오인식 어절을 교정할 때 인접된 어절 정보를 이용함으로써 보다 빠르고 정확하게 교정할 수 있다. 이 외에도 인식기가 자주 범하는 오류 문자를 교정 문자와 같이 기억해 두면 교정이 쉬워진다. 표 2는 어절간 인접 정보의 일부를 예로 나타내고 있다.

표 2. 어절간 인접 정보
Table 2. The Adjacency Information of Eojeol.

오인식 형태	교정 형태
ㅂ(습)녀다(?)	ㅂ(습)니다
ㅁ시다(?)	ㅂ시다
ㄹ녀다(?)	ㅂ니다
-할 + 대(?)	때
훌륭한, 좋은 + 성괴(?)	성과
명사+-해(?)+게	하
:	:
:	:

Ⅲ. 전체 시스템의 구성

본 논문에서 구현한 시스템은 크게 오인식 어절 검출기와 오인식 어절 교정기로 구성되었다. 그림 2는 전체 시스템의 구성을 보여주고 있다.

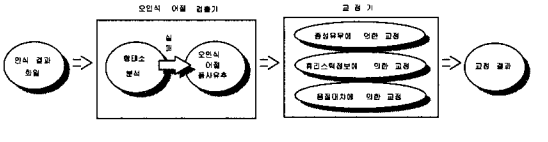

그림 2. 전체 시스템의 구성
Fig. 2. A Configuration of a Total System.

1. 오인식 어절 검출기

오인식 어절 검출기는 먼저 양방향 최장 일치에 의한 형태소 분석을 행한다. 양방향 최장 일치법은 입력 어절에 대해 조사어미 사전을 검색하여 좌방향 일치를 행하여 어절의 최장 조사나 어미를 구한 후, 구해진 최장 조사나 어미에 대한 겹침 정보를 이용해 우방향 최장일치의 끝을 계산에 의해 구하는 방법이다. 이 방법은 사전 탐색 회수가 적으므로 형태소 분석에 소요되는 시간을 줄일 수 있다. [2]의 양방향 최장 일치 형태소 분석기는 한국어 분석을 위해서 구현되었으나 본 시스템에서는 오인식 어절 검출기의 목적에 맞도록 재구성하였다. 그림 3은 형태소 분석과 어절 정보를 이용한 오인식 어절 검출기의 흐름을 나타내고 있다. 양방향 최장 일치를 행한 분석 결과에 대해 종성 유무와 좌우 접속 정보를 검사하여 어절의 오인식 여부를 판단한다. 오인식 문자로 인해 완전한 형태소 분석이 불가능한 어절은 나머지 부분으로 유추할 수 있는 가능한 분석을 모두 출력한다. 출력된 분석 결과는 어절

내 좌우접속 정보을 이용하여 오인식 문자를 포함한 형태소의 품사를 유추하여 교정 후보 생성에 필요한 정보를 그림 4와 같이 출력한다.

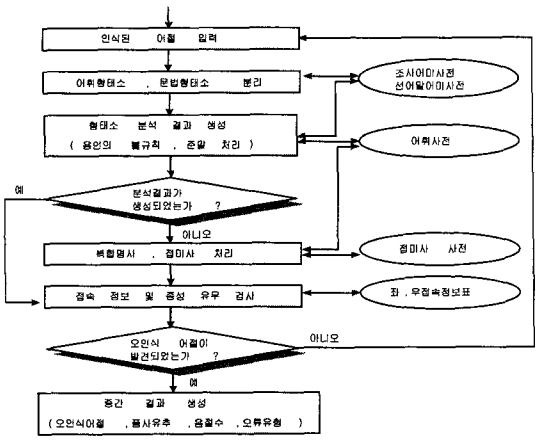

그림 3. 오인식 어절 검출기의 전체 구성도
Fig. 3. A Configuration of Misrecognized Word Checker.

<오인식 형태소,{유추된 품사 정보},음절수,오류 유형>

그림 4. 검출된 오인식 어절의 출력 정보
Fig. 4. A Output Information of Checked Misrecognized Word.

그림 4에서 음절수 정보는 후보 문자 대치 후에 사전을 탐색할 때 이용되고, 오류 유형 정보는 종성 여부, 어절간 인접 정보, 문자 대치, 자소 대치 등의 교정 방법을 선택하는 데 이용된다.

지금까지 설명한 검출기의 오인식 어절 출력 결과를 "정의하고"가 "정익하고"로 오인식된 경우를 예로 들어 살펴보자. 좌방향으로 "고"가 분리되고 우방향으로 "정"이 분리된다. 검출기는 분리 가능한 모든 정보를 생성하여 "정익하고"의 경우 생성 결과는 아래와 같이 모두 5가지이다.

① 정익하(?)/고 : <정익하, {용언}, 3, 음절대치>
② 정익하(?)/고 : <정익하, {체언}, 3, 음절대치>
③ 정익하고(?) : <정익하고, {체언}, 4, 음절대치>
④ 정익(?)/하/고 : <정익, {체언(하다)}, 2, 음절대치>
⑤ 정/익하(?)/고 : <익하, {체언,접미사}, 2, 음절대치>

①,②는 "고"가 조사,어미로 사용되므로 생성된 후보이다. ③은 "고"의 oN(체언 겹침정보)에 의해 생성하였다. ④에서 추정된 품사유형 "체언(하다)"는 체언이면서 '-하다' 가 붙을 수 있는 명사를 의미한다. 즉, 오인식된 "정익"은 '하다'성 명사이고 2음절이며 음절대치 교정을 한다. ⑤는 양방향 최장일치로 조사된 "정"의 우접속정보와 "고"의 좌접속 정보를 이용해 미분석 부분 "익하"를 접미사 혹은 체언으로 추정하였다.

2. 오인식 어절 교정기

교정에 앞서, 오인식 어절 및 문자의 형태 조사 결과를 보면 한 어절에서는 한 문자가 잘못 인식된 경우가 89%로 가장 많고 두 문자가 잘못 인식된 경우는 8%이다. 두 문자 이상의 오인식은 대부분이 4음절이상의 비교적 긴 단어와 긴 외국어에서 발생되었다. 한 문자의 경우를 다시 분석해 보면, 한 자소가 잘못 인식된 경우가 71%이고, 두 자소가 오인식 경우는 13%, 글자 형태와 무관하게 발생된 오인식이 12%였다. 한국어 문장을 구성하는 어절의 길이가 거의 2,3음절임을 감안할 때, 이 조사에서 4음절을 넘지 않는 오인식 어절에는 특수한 경우를 제외하고 한 문자가 잘못 인식되었음을 알 수 있다[3].

교정에 사용되는 사전은 형태소 분석에 이용한 사전 외에 단어의 역순 형태로 저장된 역순사전을 구성하였다. 어절의 대부분을 차지하는 2,3음절의 단어에 대한 역순사전은 고정시켜두고 그 이상의 음절로 구성된 단어에 대한 역순사전은 오인식이 일어났을 때 동적으로 생성되고 소멸될 수 있도록 하였다. 역순사전은 오인식 문자의 위치에 상관없이 교정 후보를 찾을 수 있어서 편리하다. 오인식 어절에서 첫음절이 오인식된 경우[1]는 사전의 탐색이 엉뚱한 방향으로 흐를 수 있으나, 본 시스템에서는 역방향 탐색을 가능하게 함으로써 이러한 문제를 해결할 수 있었다.

오인식 문자 교정기의 전처리 단계로 먼저 오인식 어절에 대한 가능한 형태소 분석의 여러 경우 중에서 실제 음절의 쓰임새가 사용 불가능하거나 발생 확률이 희박한 경우는 고려하지 않는다. 이것은 [11]의 음절 쓰임새의 분류와 본 논문에서 조사한 문자 발생 확률을 이용한다. 문자 발생 확률은 앞 뒤 문자와 연관하여 사용된 문자의 발생 빈도수를 구하였다.

위의 과정을 통해 검출기의 출력 결과중에서 어절 구성이 가능한 경우에 대하여 문자 교정을 시도한다. 어절 길이 정보를 참조하여 길이에 적절한 역순 사전

에서 후보 어절을 생성한 다음 문자 단위 교정과 어절 간 인접 정보에 의한 교정 그리고 자소 대치에 의해 후보 문자를 선택한다.

"조정된"이 "조징된"으로 오인식된 경우의 교정 과정을 예로 들어 보면, 먼저 검출기의 출력은 네가지로 나온다.

조징된-> ① <조징,{체언(되다)},2,3>
② <조징된,{체언,부사,관형사,감탄사},3,3>
③ <조징되,{대명사},3,3>
④ <조징되,{용언},3,3>

①은 2음절 체언 중에서 "조_"로 시작하는 단어들을 탐색, 2음절 역순 사전에서 "징_"으로 시작하는 단어들을 탐색하여 체언 중에서 "되다"와 결합 가능한 단어만 출력하여 {조정, 조직, 상징, 응징}을 후보 문자로 생성한다. ②,③,④의 경우는 해당되는 후보 문자가 없다. 구해진 후보 문자들 중에서 구조적 유사성을 고려하여 {조정,조직}을 후보로 선택한다.

Ⅳ. 실험 결과

이 장에서는 시스템의 실험 결과를 분석하기 위해 후처리기가 없는 문자 인식기, 기존 후처리기[1] 그리고 개선된 후처리기를 포함한 문서 인식기를 이용하여 처리 결과를 비교해 보인다. 비교 내용은 검출기의 성능과 생성된 후보 문자의 수, 전체 시스템의 인식률이다. 인식 자료는 신문 사설 5편, 초등학교 교과서 10쪽, 소설 10쪽을 사용하였고 표 3은 이를 나타내고 있다.

표 3. 실험 자료의 음절과 어절 개수
Table 3. Number of Character and Eojeol for Example.

구 분	신문 사설	초등 교과서	소 설
음절 수	3095	2241	5098
어절 수	931	794	1826
한 어절 평균음절 수	3.32	2.82	2.79

우선, 후처리기를 사용하지 않은 문자 인식기의 인식 결과는 표 4과 같다. 이 인식 결과를 후처리기에 의해 교정한 결과는 표 5에서 보여준다.

표 4. 후처리기가 없는 문자 인식기의 인식 결과
Table 4. The Processed Result of Recognizor without Post-processor.

		인식 기		
		성공	실패	인식률
신문 사설	음절수	2904	191	93.84
	어절수	816	115	87.61
초등 교과서	음절수	2190	51	97.74
	어절수	760	34	95.78
소설	음절수	4895	203	96.03
	어절수	1656	170	90.74

표 5. 후처리기에 의한 처리 결과
Table 5. The Processed Result.

		기존 후처리기					개선된 후처리기				
		검출기		교정기		전체 인식률	검출기		교정기		전체 인식률
		성공	실패	성공	실패		성공	실패	성공	실패	
신문 사설	음절수	100	91	85	15	96.57	117	74	105	12	97.21
	어절수	67	48	58	9	93.85	74	41	67	7	94.89
초등 교과서	음절수	16	35	4	12	97.92	30	21	26	4	98.87
	어절수	15	19	10	5	96.94	22	12	15	7	97.65
소설	음절수	84	119	57	27	97.13	114	89	100	14	97.98
	어절수	107	63	91	16	95.69	123	47	110	13	96.72

표 5에서 개선된 후처리기가 있는 인식기의 음절 인식률이 기존 시스템의 인식률 97.03%보다 향상된 98.02%임을 알 수 있다. 또한, 어절 정보의 이용으로 인해 어절 교정률이 높아져서 전체 인식률 96.35%가 97.22%로 되었다. 그림 5는 형태소 분석 회수를 비교한 것으로 오인식 어절의 평균 형태소 분석 회수가 기존의 방법보다 줄어들었음을 알 수 있다.

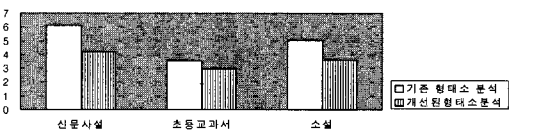

그림 5. 오인식 어절의 평균 형태소 분석 회수
Fig. 5. The Average of Morphological Analyzer for Misrecognition Eojeol.

그림 6은 어절 정보를 이용한 본 시스템이 기존 시스템에 비하여 적은 수의 후보 문자를 생성함으로써 후처리기의 속도 개선을 보여주고 있으며, 그림 7은 후처리기가 없는 인식기와 기존의 후처리기, 개선된 후처리기를 이용한 문서 인식률을 나타내고 있다.

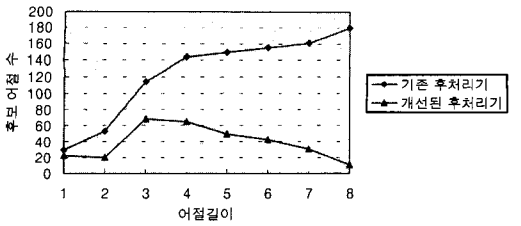

그림 6. 평균 후보 어절 수
Fig. 6. The Number of Candidate Eojeol.

역순 사전과 어절 정보의 이용으로 인해 후보 어절의 수가 많이 줄어 들었지만, 사전 의존도가 높아짐으로 미등록어의 교정에 대한 연구가 계속되어야 한다.

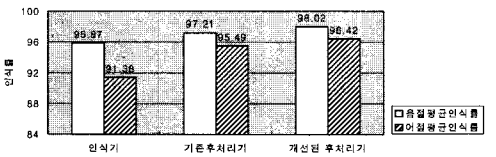

그림 7. 처리 결과 인식률
Fig. 7. The Correction Rate.

V. 결 론

본 논문에서는 기존의 후처리기보다 처리 속도와 인식률을 개선한 기법을 제시하였다.

우선 사전 탐색 회수가 적은 양방향 일치에 의한 형태소 분석을 이용하여 오인식 어절을 검출하였다. 검출된 오인식 어절은 어절 구성 성분에 대한 접속 정보를 이용하여 오인식 문자로 인하여 분석이 불가능한 형태소의 품사를 유추함과 더불어 어절 정보의 이용으로 교정 후보 문자의 생성 개수를 줄일 수 있었다.

이 결과 기존의 후처리기를 포함한 문서 인식기의 성능 97.03%를 98.02%로 향상시켰다. 앞으로 미등록어의 오인식 처리와 구문적 의미를 반영해야 올바른 교정이 되는 오인식 어절의 검출 등을 연구해야 하겠다.

참 고 문 헌

[1] 이영화, 김계성, 김영훈, 이상조, "문자인식 후처리를 위한 형태소 분석기와 문자 교정기의 구현", 전자공학회논문지-C 제34권 제5호, pp. 82-92, 1997

[2] 최재혁, "양방향 최장일치에 의한 형태소 분석기의 구현", 경북대학교 전자공학과 박사학위논문, 1993

[3] 이병희, 김태균, "오프라인 한글 문자 인식을 위한 효율적인 오인식 단어 교정 방법", 한글정보처리학회 논문지, 제3권 제6호, pp. 1598-1605, 1996

[4] 강승식, "한국어 형태소 분석을 위한 단어 유형 분류와 자료구조", 제8회 한글 및 한국어 정보처리 학술대회, pp. 241-245, 1996

[5] 강승식, "한국어 형태소 분석기 HAM의 형태소 분석 및 철자검사 기능", 제8회 한글 및 한국어 정보처리 학술 대회, pp. 246-254, 1996

[6] 이병훈, "말뭉치 분석을 기반으로 한 한국어 철자 교정기 구현", 연세대학교 전산과학과 석사학위논문, 1994

[7] 김민정, 권혁철, "언어적, 경험적 제약을 이용한 한국어 문자 인식 후처리 기법", 정보과학회 논문지(B) 제24권 제1호, pp. 25-31, 1997

[8] Thomas-N.Turba, "Checking for Spelling Typographical Errors in Computer Based Text", SIGPLAN Notices, pp. 298-312, 1981.

[9] Tsuyoshi Kitani, "An OCR Post-processing Method for Hanwritten Japanese Documents", NLPRS '91, pp. 38-45, Nov. Singapore, 1991.

[10] R.L.Kashyap, B.J.Oommen, "Spelling Correction Using Probabilistic Methods", Pattern Recognition Letters, pp. 147-154, 1984.

[11] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 컴퓨터공학과 박사학위 논문, 1993

저 자 소 개

李 英 和(正會員) 第 34卷 C編 第 5號 參照

金 桂 成(正會員) 第 34卷 C編 第 5號 參照

金 永 勳(正會員) 第 34卷 C編 第 5號 參照

李 相 祚(正會員) 第 34卷 C編 第 5號 參照