論文99-36C-7-6

오인식 형태소 추정에 의한 한국어 문자 인식 후처리 기법

(A Postprocessing Method of Korean Character Recognition by Mis-recognized Morphology Presumption)

金 永 勳 *, 李 英 和 **, 李 相 祚 **

(Young Hun Kim, Young Hwa Lee, and Sang Jo Lee)

요 약

본 논문에서는 형태소 분석을 이용한 후처리에서 속도 개선을 위해 사전 탐색 횟수를 줄이는 새로운 방법을 제안한다. 본 논문에서 제안하는 방법은, 오인식 어절 검출을 위한 형태소 분석 과정에서 분석되는 일부의 형태소 정보를 최대한 이용하여 오인식 어절의 형태소 구성을 추정한 후, 형태소 단위의 교정을 한다. 형태소 단위의 교정은 어절보다 길이가 짧으므로 최악의 경우라도 생성되는 후보의 수가 어절 단위의 교정보다 적다, 특히, 생성된 후보가 형태소 단위이므로 사전 탐색만으로 올바른 후보를 선택할 수 있으므로 형태소 분석으로 인한 사전 탐색 횟수를 줄일 수 있다. 본 논문에서 제안한 형태소 정보를 이용한 후처리는 기존의 어절 단위 후처리에 비해 생성된 후보의 형태소 분석이 필요 없다. 생성된 후보가 형태소이므로 사전 탐색에 의해 올바른 후보를 선택할 수 있었다. 이로 인해 사전 탐색 횟수는 어절 단위 후처리와 비교하였을 때 60%나 감소되었으며 후처리 결과 문자 인식기의 음절 인식률이 94%에서 97%로 향상되었다.

Abstract

We proposed the new method of postprocessing which not only reduces the frequency of dictionary access using morphological analysis but improve the recognition rate of character recognizer. In this paper, after estimating morphological construction of mis-recognized word using the part of speech that is analyzed, correct presumed mis-recognized morphology. The postprocessing using a morphology unit reduce candidate because of short than word and frequency of dictionary access because there is no need to morphological analysis for candidate. To select right candidate is only necessary to dictionary access.

The proposed results show that reduced the frequency of dictionary access to 60% than postprocessing method using a word unit and recognition rate improved from 94% to 97%.

Ⅰ. 서 론

문자 인식 시스템에서 후처리란 문자 인식 결과의 중의성을 해결하는 과정으로서, 오인식된 문자를 찾아 바르게 교정하거나 언어 처리에 의해 여러 개의 후보 문자 중에서 가장 적합한 하나의 후보를 선택함으로써 인식기의 인식률을 높이는 과정이다[3,6].

문자 인식 시스템이 개발된 이래로 지금까지 시스템의 인식률을 높이기 위해 연구되어 온 후처리 기법은 영어권을 중심으로 문맥적 지식의 표현 방법에 따라 크게 세 가지 유형으로 나뉠 수 있다. 첫째는 문맥적 지식의 확률적 표현에 기초한 상향식 방법, 둘째는 문맥적 지식의 구조적 표현에 기초한 하향식 방법 그리

* 安東科學大 情報處理學科
(Dept. of Information Processing, Andong Science College)
** 正會員, 慶北大學校 컴퓨터工學科
(Dept. of Computer Engineering, Kyungpook National Universirty)
接受日字:1998年11月19日, 수정완료일:1999年6月22日

고 셋째는 상향식과 하향식을 결합한 복합적 방법이다. 하지만 한국어는 영어에 비해 어순이 자유롭고 문장 성분의 생략이 많으며 용언의 활용과 조사, 어미의 다양한 변화로 어절의 형태가 복잡하다. 이러한 이유로 한국어의 후처리는 영어와는 다른 접근 방법을 취해야 한다[6]. 1980년대 중반부터 연구되어 온 한국어 문자 인식 후처리 기법은 사전과 확률치를 이용한 방법과 형태소 분석과 언어적 정보를 이용한 방법으로 나누어진다. 전자에는 사전을 이용하여 오류 어절을 분석하고 오류 음절을 추측한 후 사전 내의 올바른 어절로 교정하는 방법[7], 말뭉치를 기반으로 철자를 교정하는 방법[4], 통계적 방법에 의한 후처리[1]가 있는다. 이 방법에 의해 생성된 후보는 한국어 맞춤법에 어긋난 형태일 수 있다. 어절 단위의 후처리에서는 생성된 후보의 검증을 위해 형태소 분석이 필요하다[5].

생성된 후보의 검증을 위해 형태소 분석을 이용하는 후자에는 후보 어절에 대한 형태소 분석을 행하고 형태소간 접속 확률인 언어 평가를 이용하여 올바른 후보들 중 하나의 어절을 결정하는 방법[3]과 후보 음절 벡터에서 형태소 분석과 di-gram, viable-prefix를 이용하여 속도를 개선한 방법[9] 등이 있다. 형태소 분석을 이용한 후처리는 선택된 후보가 한국어 맞춤법에 타당하므로 교정률이 비교적 높으나 생성된 후보 어절의 맞춤법 검사를 위해 모든 후보를 다시 형태소 분석해야 한다. 한 어절을 형태소 분석하기 위해 필요한 사전 탐색 횟수를 보면 적게는 평균 3회[8]이며 많게는 10.6회이다. 문자 인식기의 인식 결과로 후보 음절 벡터를 입력받은 후처리기에서 한 어절의 음절 수가 n이고 각 음절의 후보 음절 개수가 m이라면 최악의 경우 생성되는 후보 어절의 수는 m^n 개이다. 즉, 사전 탐색 횟수는 $3\times m^n$ 에서 $10\times m^n$ 회가 필요하다.

이러한 문제점을 해결하기 위해서는 형태소 분석을 필요로 하는 생성 후보의 개수를 줄여야 한다. 그러나, 생성되는 후보의 단위가 어절인 경우엔 생성된 많은 후보 어절을 모두 형태소 분석하여 올바른 후보를 선택해야 하므로 사전 탐색 횟수를 줄이는데는 한계가 있다.

본 논문에서는 형태소 분석을 이용한 후처리에서 속도 개선을 위해 사전 탐색 횟수를 줄이는 새로운 방법을 제안한다. 본 논문에서 제안하는 방법은, 오인식 어절 검출을 위한 형태소 분석 과정에서 분석되는 일부의 형태소 정보를 최대한 이용하여 오인식 어절의 형태소 구성을 추정한 후, 형태소 단위의 교정을 한다. 형태소 단위의 교정은 어절보다 길이가 짧으므로 최악의 경우라도 생성되는 후보의 수가 어절 단위의 교정보다 적다, 특히, 생성된 후보가 형태소 단위이므로 사전 탐색만으로 올바른 후보를 선택할 수 있으므로 형태소 분석으로 인한 사전 탐색 횟수를 줄일 수 있다. 본 논문에서 사용한 말뭉치는 초등학교 교과서, 동아일보 사설 모음, 공학 논문지, 소설 등의 내용 일부로서 약 40만 어절이다.

Ⅱ. 형태소 정보를 이용한 후처리

본 논문에서 제안한 후처리 방법은 문자 인식 결과를 먼저 형태소 분석하여, 형태소 분석에 실패한 어절 즉 오인식 어절을 검출한다. 검출된 오인식 어절은 분석된 일부 형태소의 형태소간 접속 정보를 이용하여 오인식 어절의 가능한 형태소 구성을 추정한다. 오인식 어절에서 추정된 형태소 구성 정보는 형태소 단위의 후보를 생성하고, 후보 중 올바른 후보는 사전 탐색에 의해 선택되므로 형태소 분석이 필요 없다.

1. 오인식 어절 검출을 위한 형태소 분석

형태소 분석기의 기능은 응용 분야와 사용 목적에 따라 달라진다. 예를 들어, 한국어 구문 분석의 전 단계로 쓰기 위한 형태소 분석기는 어절을 구성하고 있는 모든 형태소에 대하여 가능한 많은 분석 결과를 출력해야 되며, 자동 인덱싱(automatic indexing)이나 정보 검색 시스템에 사용될 형태소 분석기는 명사구에 대한 분석 결과를 출력하면 된다.

본 논문의 오인식 어절 검출을 위한 형태소 분석기는 철자 검사를 목적으로 하므로 모호성이 내포된 단어나 어절에 대한 모든 분석이 반드시 필요하지는 않다. 예를 들어, "나는", "감기는" 등의 어절이 전체 문장에서 어떤 의미로 분석되어야 옳은가를 고려하여 "감(동사)+기+는", "감기(명사)+는" 그리고 "감기(동사)+는"의 모든 분석 결과를 출력하기보다는 한 어절을 구성하고 있는 형태소들 사이에서 그들의 결합이 어색하지 않고 합당한가를 검사한다. 분석에 성공하는 경우가 있으면 철자 오류 판정을 피한다. 또한, 한국어 분석에 사용되는 기존의 형태소 분석기는 올바른 어절이 입력된 경우에만 분석에 성공하여 그 결과를 출력

하고 그렇지 않은 어절에 대해서는 분석 과정을 출력하지 않고 실패라는 의미의 메시지만 출력한다. 본 논문의 형태소 분석기는 철자 오류가 있는 어절이 입력되어 완전한 분석에 실패하더라도 이미 분석된 부분의 형태소 정보를 출력하도록 한다. 이것은 인식 결과에서 오인식 어절을 검출하기 위한 형태소 분석 과정에서 오인식 형태소를 만나기 이전에 했던 분석 정보를 이용하여 나머지 형태소의 품사를 추정하는데 이용된다.

본 논문의 오인식 어절 검출을 위한 형태소 분석기는 후보 음절 벡터를 가진 인식 결과에서 어절 단위로 입력받아서 철자 오류가 있어서 형태소 분석에 실패한 어절은 오인식 어절로 간주하여 교정을 위한 다음 단계로 출력된다. 형태소 분석의 성공은 입력 어절을 구성하고 있는 형태소들이 한국어 맞춤법에 맞도록 결합되어 있음을 의미한다.

$W,\ l\langle n_i q_i\rangle, r\langle n_j q_j\rangle,\ lp\ , rp$

W : 오인식 어절

$l\langle n_i p_i\rangle$:우방향 탐색 결과

$r\langle n_j q_j\rangle$:좌방향 탐색 결과

$n_i,\ n_j$: 분석된 음절 수

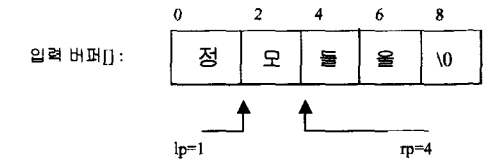
$q_i,\ q_j$: 분석된 형태소의 품사 정보

$lp\ , rp$: 우방향,좌방향탐색위치값

그림 1. 형태소 분석기에서 오인식 어절의 출력 형태
Fig. 1. Morphology Analyzer Output of Mis_recognized Word.

형태소 분석은 먼저 왼쪽에서 오른쪽 방향(우방향)으로 최장 일치법에 따라 어휘를 탐색한다. 우방향 탐색에서 실질 형태소가 분석되며 만약 입력 어절의 남은 부분이 있으면 오른쪽에서 왼쪽 방향(좌방향)으로 분석하면서 조사나 어미를 분석한다. 우방향 탐색의 끝 위치(lp)와 좌방향 탐색의 끝 위치(rp)가 일치하는 경우는 분석이 완료된다. 그러나, 양방향 탐색 위치가 일치하지 않는 경우 즉, $lp<rp-1$ 이면 남은 어절에 대해서 선어말 어미, 접미사, 하/되/이 의 분석을 계속한다. 모든 처리 루틴을 거친 뒤에도 입력 어절이 완전히 분석되지 않으면 오인식 어절로 판단하여 이미 분석된 일부 형태소 정보와 함께 오인식 형태소를 추정하는 다음의 처리 과정으로 출력된다. 오인식 어절 검출을 위한 형태소 분석기는 그림 1의 출력 형태와 같이, 형태소 분석에 실패한 오인식 어절 전체와 부분 어절에 대한 형태소 분석 결과 그리고 우방향, 좌방향의 위치 값을 함께 출력한다.

부분 형태소 분석 결과는 분석된 형태소의 음절수와 사전에 등재된 형태소의 품사 정보이다. 예를 들어, 인식을 원한 어절 "정보들을"의 인식 결과 "정모들을"를 형태소 분석하면 출력 결과는 그림 2와 같다.



-> 출력 : 정모들을,l<1체언>, r<1접미사,1조사>, 1,4

그림 2. 오인식 어절의 형태소 분석 결과 예
Fig. 2. Example for Morphology Analyzer Output of Mis_recognized Word.

위의 정보들은 다음 절에서 설명할 오인식 어절의 형태를 추정하는 함수의 입력이 된다.

2. 오인식 어절의 형태 추정

이 절에서는 형태소 분석기에서 검출된 오인식 어절의 부분 형태소 정보를 이용하여 미분석 형태소의 품사와 오인식 어절의 형태를 추정하는 과정을 설명하고자 한다.

표 1. 형태소의 기호
Table 1. Symbol of Morphogy.

형태소	기호	형태소	기호
체언	n	어미	e
용언	a	조사	j
부사	v	선어말 어미	s
관형사	d	하/되(활용)	f
감탄사	h	함/됨	m
단일어	o	이(활용)	g
접미사	b	임	y

체언은 사전의 품사 정보가 명사, 대명사, 수사, 접속사인 어휘이며, 용언은 품사 정보가 동사, 형용사인 어휘이다. 본 논문에서는 체언, 부사, 관형사, 감탄사 그리고 용언(활용) 등이 단독으로 어절을 구성하고 있는 경우 이들을 단일어로 묶어서 같은 형태로 처리한

다. 이것은 하나의 형태소로 어절이 구성된 경우에는 교정 후보의 품사를 정확하게 추정하기 어려우므로 단일어로 추정한다. 표 1은 본 논문에서 사용한 형태소의 기호를 나타낸다.

본 논문에서는 오인식 어절의 형태를 추정하기 위해 어절의 길이별로 가능한 형태소의 구성을 조사하여 테이블을 구성하였다. 구성된 테이블의 정보와 오인식 어절에서 분석된 일부 형태소 정보를 이용하여 분석되지 못한 오인식 형태소의 품사를 추정한다. 표2는 2음절, 3음절 어절의 형태 정보 테이블의 예이다.

표 2. 어절의 형태 정보 테이블
Table 2. Information Table of Word Type Presumption.

음절수	형태소 개수	어절 구성 형태	가중치	음절수	형태소 개수	어절 구성 형태	기중치
2	1	2o	47.58	3	2	2ale	8.91
2	2	1n1j	29.40	3	2	2n1b	0.72
2	2	1ale	20.89	3	2	2n1f	3.59
2	2	1n1f	1.04	3	2	2n1j	46.71
2	2	1n1b	0.09	3	2	2v1f	0.01
2	2	1v1f	0.04	3	2	2v1j	0.01
2	2	1v1j	0.06	3	3	1a1s1e	9.85
3	1	3o	16.48	3	3	1n1b1j	1.58
3	2	1a2e	3.16	3	3	1n1f1e	0.54
3	2	1n2b	0.00	3	3	1n1y1j	0.09
3	2	1n2j	6.65	3	3	1n1m1j	0.05
3	2	1v2j	0.03	3	3	1v1f1e	1.62

<어절의 음절수>는 오인식 어절의 전체 음절수를 나타낸다. 어절의 음절수에 따라 구성되는 형태소가 다르므로 어절의 길이별로 구분하였다. 오인식 어절의 음절수와 같은 음절수에 대한 어절 형태만 참조한다.

$$n_1\,p_1\;\; n_2\,p_2\;\; n_3\,p_3\;\; \ldots\ldots n_n\,p_n$$

$$0 \le n_1, n_2, n_3, \ldots, n_n \le |\text{어절}|$$

$$n_1 + n_2 + n_3 + \ldots + n_n = |\text{어절}|$$

$p_1, p_2, p_3, \ldots, p_n$: 형태소 품사 정보

그림 3. 어절 구성 형태
Fig. 3. Structure of Word Type.

<형태소 개수>는 어절을 구성하고 있는 형태소가 몇 개인지를 나타낸다. 이 정보는 오인식 어절에서 분석된 일부 형태소가 있을 경우에 구성된 형태소의 개수를 참조하여 가능한 형태를 추정하기 위해 필요하다. <어절의 형태>는 음절수에 따라 다르게 구성되어 있으며 어절의 형태 26가지 중에서 어절의 음절수를 고려하여 가능한 모든 형태들로 구성하였다. 어절의 형태는 그림 3과 같이 구성된다.

<가중치>는 어절의 출현 빈도를 이용하여 구한 값이다. 말뭉치에서 맞춤법이 올바른 어절을 음절 길이별로 나누어 각 길이별 어절 수를 구하였다. 나누어진 어절을 형태소 분석한 후에 어절 형태들의 출현 빈도를 미리 구해둔 해당 길이별 어절 개수로 나누어 구하였다. 가중치에 사용된 값은 식 1에 의해 구한다.

$$\text{가중치 } (\text{어절형태}_i) = \frac{\text{어절형태}_i\text{의 출현횟수}}{\text{같은 음절수를 가진 어절의 개수}} \times 100 \quad (1)$$

이제, 어절의 형태소 정보 테이블을 이용하여 형태소 분석기에서 검출된 오인식 어절의 형태를 추정하는 과정을 설명하고자 한다.

형태소 분석에 실패한 오인식 어절 중에서 2음절 이상인 어절에 대해서 해당 어절의 가능한 형태소 구성을 추정한다. 1음절 어절 경우는 이 추정 과정을 거치지 않고 교정한다. 어절의 형태소 정보 테이블과 오인식 어절에서 분석된 일부 형태소 정보를 이용하여 분석되지 못한 형태소의 품사를 추정한다. left(형태소 정보)와 right(형태소 정보)는 각각 분석된 실질 형태소와 형식 형태소 정보가 있음을 의미한다. 즉 형태소 분석 결과에서 어느 한 정보만 출력된 경우에는 어절 형태 정보 테이블에서 출력된 형태소의 품사를 포함하는 어절의 형태를 찾아서 추정된 형태 집합{}에 추가한다. 출력된 형태소 분석 결과에 left(형태소 정보)와 right(형태소 정보)가 모두 포함된 경우는 분석된 정보를 최대한 이용하여 중간에 분석되지 못한 일부 형태소를 추정한다. 가능한 어절 형태 [ ]는 분석된 형태소의 음절 길이와 품사가 같고 어절의 구성 형태소 개수는 하나 더 많은 어절 형태를 가장 먼저 찾아서 저장한다. 다음으로 분석된 형태소 정보를 포함할 수 있는 어절의 형태를 찾아 저장한다. 분석된 형태소를 포함한다는 것은 형태소의 품사는 일치하고 음절 길이는 더 긴 어절 형태를 의미한다. 그런데, 분석된 양쪽 형태소가 결합 불가능한 경우가 있다. left(형태소 정보)와 right(형태소 정보)가 체언과 어미 혹은 용언과

조사인 경우로서, 이러한 어절은 두 가지로 분석된다.

첫째는 중간에 분석되지 못한 형태소가 이 둘 형태소와 각각 결합할 수 있는 형태이다. 즉, 한 어절에 체언과 어미가 같이 있으려면 체언 뒤에 "하/되/이(활용)"이 필요하며, 용언과 조사의 경우는 용언 뒤에 "명사형 어미"가 있어야 한다.

다음으로는 양쪽에서 분석된 형태소 중 하나는 잘못 분석되었다는 가정을 할 수 있다. 어휘 사전의 단어 중에서 첫 음절이 같으면서 품사가 다른 경우 혹은 조사/어미 중에서 끝 음절이 중복된 경우이다. 따라서, 어느 한 쪽의 분석된 정보를 무시하고 추정 형태를 찾는다. 이렇게 찾아진 어절 형태는 가중치의 내림차순으로 정렬되어 출력된다.

분석된 형태소 정보가 하나도 없을 경우에는 단일어 형태 정보를 우선으로 선택하고 가중치 순서대로 가능한 모든 형태를 포함한다. 추정된 형태 집합{}의 우선순위는 정보 최대 일치, 정보 최대 포함, 출현 빈도의 순서로 결정된다.

3음절 이상 오인식 어절에서 형태소 분석 결과가 "l<a>, r<j>"이거나 "l<n>, r<e>"인 경우는 다음과 같이 추정된다.

1) 형태소 분석 결과 : l<xa>, r<x′j>

가능한 분석1-1. 체언이나 부사의 앞 음절이 용언과 중복 : <xa>가 <yn>이나 <yv>로 바뀌어야 한다.(x<y)

가능한 분석1-2. 어미의 일부가 조사의 음절과 중복 : <x′j>가 <y′e>로 바뀌어야 한다. (x′<y′)

2) 형태소 분석 결과 : {(l<xn>, r<x′e>), (l<xv>, r<x′e>)}

가능한 분석2-1. 용언의 앞 음절이 체언이나 부사와 중복 : <xn>나 <xv>를 <ya>로 바뀌어야 한다.(x<y)

가능한 분석2-2. 조사의 일부가 어미의 음절과 중복 : <x′e>가 <y′j>로 바뀌어야 한다. (x′<y′)

가능한 분석2-3. 체언이나 부사가 어미와 같은 어절에 쓰일 수 있도록 매개체 추정 : 미분석된 형태소는 {<하/되/이(활용)>}를 포함한다.

예를 들어, "매달리어"가 "매단리어"의 형태로 오인식되면 형태소 분석 결과는 l<1n>, r<1e>이다. 용언 "매달리다"의 첫음절이 체언 "매"와 중복되기 때문이다. 이러한 형태는 가능한 분석2-2, 2-3에 의해 어절의 형태를 추정한다. 체언에 인접하여 사용되는 "하/되/이(활용)"중에서 "한, 할, 해, 인, 일, 되"는 체언의 음절과 중복되므로 체언으로 추정하는 형태도 출력한다.

입력 어절에서 모든 형태소를 분석한 후 분석된 인접 형태소가 결합 가능한 형태소인지 검사하는 과정에서 실패할 경우가 있다. 예를 들어, "지장하게"를 형태소 분석하면 "지장(2n), 하(1f), 게(1e) "로 모두 분석되었으나 체언 "지장"은 "하"와 결합이 불가능하다. 이러한 경우는 인접이 불가능한 두 형태소를 번갈아 추정 형태로 정한다. 즉, 2n이 오인식 형태소일 가능성과 1f가 오인식일 가능성을 우선 고려한다.

그림 4는 검출된 오인식 어절에 대한 어절 형태 추정 결과이다. 출력 결과는 하나 이상일 수 있으며 가능한 어절 형태 [ ] 에 저장된 순서대로 정렬된다.

$$\{W,\ [\ n_x a_i \langle n_y a_j \rangle n_z a_k]_l,\ f_l\ \}$$

W : 오인식어절

$n_x a_i \langle n_y a_j \rangle n_z a_k$: 추정된 어절 형태 $_l$의 구성

f_l : 어절형태 $_l$의 가중치

$\langle n_y a_j \rangle$: 추정된 오인식 형태소의 정보

그림 4. 오인식 어절의 형태소 정보 추정 결과
Fig. 4. Morphology Presumption Output about Mis-recognized Word.

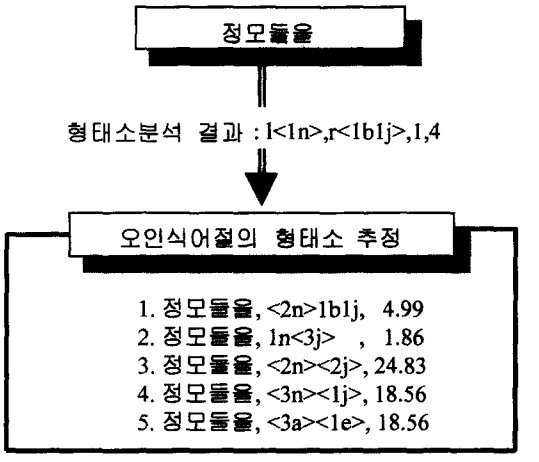


그림 5. 오인식 어절의 형태소 추정 결과 예
Fig. 5. Example of Morphology Presumption Output about Mis-recognized Word.

예를 들어, 오인식 어절 "정모들을" 의 어절 형태 추정 결과는 그림 5이다. <>안의 정보가 추정된 오인

식 형태소이다.

3. 교정 후보의 생성과 선택

이 절에서는 오인식 어절의 형태 추정 결과를 이용하여 교정 후보를 생성하는 방법과 최종 교정 후보를 선택하는 방법을 설명한다.

후보 음절 벡터에서는 후보 음절이 인식 신뢰도에 의해 순서적으로 정렬되어 있다. 그러므로 해당 음절의 첫 번째 후보 음절과 두 번째 후보 음절로 이루어진 후보 중에 올바른 형태소가 있을 확률이 높다. 따라서, 본 논문에서는 인식 신뢰도가 높은 음절을 이용하여 올바른 후보를 우선 생성한다.

인식 신뢰도가 높은 음절에 의해 올바른 후보를 선택하지 못한 경우 후보 음절 벡터를 깊이 우선 탐색하며 후보를 생성한다. 이 과정에서는 먼저 선택된 올바른 후보의 거리값을 기준으로 다음 후보 생성을 제한한다. 생성된 후보의 거리값이 선택된 후보의 거리값보다 큰 경우에는 생성을 억제하여 사전 탐색 횟수를 줄인다.

형태소 정보를 이용한 후보 생성 억제는 추정된 어절의 형태에서 형태소 길이가 작은 후보를 먼저 생성하여 사전 탐색에 실패하면 그 추정 형태는 더 이상 후보를 생성하지 않도록 한다. 특히 추정된 형태소 정보에 "하/되/이(활용), 접미사" 정보를 우선 고려하여 후보를 생성한다. 이 형태소들은 길이가 1이므로 후보 음절들을 순서대로 사전 탐색하면 된다. 예를 들어, "알아내는"을 교정할 경우에 생성되는 후보의 수를 비교해 보면 그림 6과 같다.

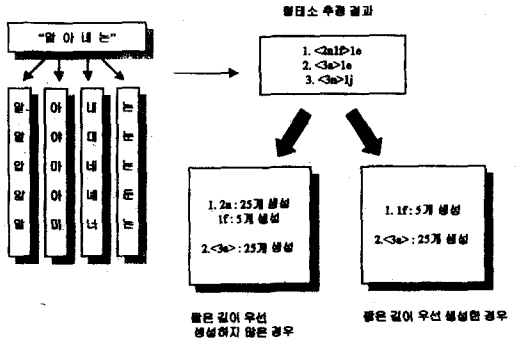


그림 6. 음절 길이가 짧은 형태소 후보 우선 생성의 예

Fig. 6. Example of Priority Rule about Create Candidate-Word.

다음은 사전 탐색에 실패한 형태소 정보를 기억했다가 같은 형태소를 가진 다음 추정 형태에 대한 후보 생성을 미리 억제할 수 있다. 예를 들어, 오인식 어절의 추정 형태가 "<2n2j>, <2n1b1j>, <2n1f1e>, <3n1j>.." 등인 경우에 처음 2n2j에서 2n이 실패한 경우, 다음의 추정 형태 2n1b1j, 2n1f1e은 실패 형태소 2n을 공통으로 가지고 있으므로 여전히 실패한다. 이 경우에 모두 후보를 생성하기보다는 실패 형태소 정보를 이용하여 다음 추정 형태에서 같은 형태소를 가진 어절은 후보 생성에서 미리 제거하여 불필요한 사전 탐색을 방지한다.

본 논문의 후처리기에서는 추정된 어절의 형태 정보를 이용하여 교정 후보를 생성하고 선택한다. 후보 생성은 추정된 어절 형태 중에서 분석된 정보를 가장 많이 가졌거나 해당 어절의 형태가 말뭉치에서 출현한 빈도가 많은 순으로 처리된다. 추정 형태에서 가장 먼저 선택되는 후보 즉 후보의 거리값이 가장 작은 후보가 최종 교정 어절이 된다.

문자 인식기에서 후보 음절 벡터를 입력으로 받아 후처리를 할 경우, 후보 음절 벡터에 인식을 원한 음절이 없으면 최종 후보를 선택하지 못한다. 문자 인식기는 인식 알고리즘의 특성으로 인해 임의의 음절이 특정한 음절로 자주 오인식될 수 있으며 이를 위해 인식 결과를 분석하는 과정에서 대치음절 정보를 조사하여 교정에 이용한다. 후보 음절 벡터에서 인식 신뢰도가 가장 좋은 음절에 대한 대치음절로 후보를 생성하여 사전을 탐색한다. 대치음절 정보는 문자 인식기에 의존되므로 문자 인식기가 바뀌면 수정하는 것이 효과적일 것이다. 그림 7은 대치음절 정보를 이용하여 교정 후보를 선택하는 예이다.

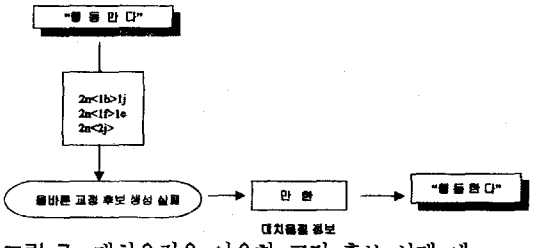


그림 7. 대치음절을 이용한 교정 후보 선택 예

Fig. 7. Example of Select Candidate-Word by Substitution Character.

Ⅲ. 실험 및 결과 분석

이 장에서는 본 논문에서 제시한 후처리기의 성능을 평가하기 위해 말뭉치의 일부 문서에 대한 인식 결과

를 실험하고 후처리 결과를 분석한다.

본 논문에서 사용한 사전의 종류는 240,300개 단어의 어휘 사전, 987개 단어의 조사/어미 사전, 76개 단어의 선어말 사전 그리고 74개 단어의 접미사 사전이다. 어절 형태 추정의 말뭉치는 동아 일보 사설 모음, 초등학교 읽기, 쓰기, 바른 생활, 즐거운 생활, 사회, 국어 교과서, 논문지, 소설, 통신 자료에서 얻은 문서들이다. 본 후처리기는 팬티엄-150 MHz PC에서 C-언어로 구현하였다. 실험에 사용된 자료의 종류와 어절 수는 표 3과 같다.

표 3. 실험 자료의 종류와 어절개수
Table 3. A Kind and Words of Postprocessing Data.

구분	신문 사설	초등 교과서	소설	공학 논문	전체
전체 어절 개수	1862	1588	1652	1336	6438
1음절 어절 개수	164	111	132	134	546
2음절 어절 개수	438	447	446	307	1638
3음절 어절 개수	586	579	516	385	2066
4음절 어절 개수	384	238	298	273	1193
5음절 어절 개수	138	127	99	94	458
6음절 어절 개수	93	52	83	80	308
7음절이상 어절 개수	54	34	78	63	229

본 논문에서 제시한 후처리 방법의 성능을 비교하기 위해서 같은 인식 결과로 몇 가지 실험을 한다. 표 4와 그림 8은 본 논문에서 제안한 오인식 어절에서의 형태소 분석 추정 정보와 인식 신뢰도가 높은 후보를 우선 생성, 거리값 제한에 의한 후보 생성 등으로 인한 속도 개선 효과를 나타낸다.

실험 0 : 추정된 형태소 정보 모두 이용
실험 A: 형태소 추정 정보(60%이용)
실험 B: 형태소 추정 정보 + 인식 신뢰도 높은 후보 우선 생성
실험 C: 형태소 추정 정보 + 인식 신뢰도 높은 후보 우선 생성 + 거리값 제한에 의한 후보 생성
실험 D: 형태소 추정 정보 + 인식 신뢰도 높은 후보 우선 생성 + 거리값 제한에 의한 후보 생성+ 형태소 정보를 이용한 후보 생성

실험 0은 오인식 어절의 가능한 어절 형태를 모두 고려한 경우인데, 후보 음절 벡터에 올바른보가 없을 때 생성된 후보의 수가 증가된다.

표 4. 실험별 생성 후보 개수 비교
Table 4. Comparison for Created Candidate-Words.

	1음절	2음절	3음절	4음절	5음절	6음절	7음절이상
실험 0	3.10	29.08	52.56	298.31	294.93	1305.01	3402.77
실험 A	3.10	15.34	10.00	109.70	105.30	743.20	1703.80
실험 B	3.10	15.00	7.70	57.04	56.79	543.02	645.50
실험 C	3.10	14.00	7.50	53.20	51.50	398.10	629.60
실험 D	3.10	13.40	6.90	39.21	34.10	153.30	302.92

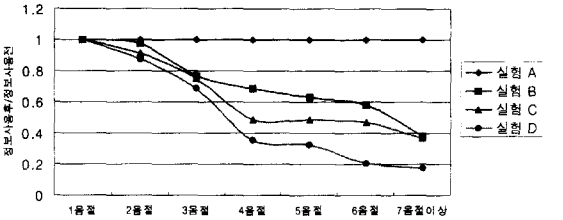


그림 8. 제안한 정보 사용에 따른 생성 후보 수의 비교
Fig. 8. Comparison Graph for Created Candidate-Words.

표 4와 그림 8에서 2음절 어절보다 3음절 어절에서 사전 탐색 횟수가 적게 나타난 것은 2음절 어절의 형태소 추정 결과는 실질 형태소로만 구성된 2o 형태의 오인식 어절이 많은 반면에 3음절 어절에서는 출현 빈도에 의한 우선 순위와 동일한 형태 즉,2n1j 혹은 2a1e의 오인식 어절이 불필요한 후보가 생성되지 않았다. 정보의 사용에 따른 효과는 추정 형태소 정보 중에서 특정 형태소 정보를 먼저 교정하거나 실패 형태소 정보를 이용한 경우가 가장 높은 효과를 보였고, 인식 신뢰도가 높은 음절을 이용하여 후보를 우선 생성한 경우도 후보의 수를 줄였다.

표 4의 실험 결과를 기존의 어절 단위 후처리와 비교해 보면 표 5로 나타난다. 사전 탐색 횟수로 비교하면 어절 단위 후처리에서는 (사전 탐색 횟수=생성후보 개수*형태소 분석에 필요한 사전 탐색 횟수)이지만, 본 논문에서 제안한 형태소 단위 후처리는 (사전탐색 횟수=생성 후보 개수)이다.

표 5에서 어절 단위 후처리는 어절의 길이가 늘어날수록 생성된 후보의 개수와 그에 따른 사전 탐색 횟수의 증가율이 형태소 단위의 후처리보다 크다는 것을 알 수 있다. 어절 단위의 후처리와 비교하였을 때 사

전 탐색 횟수가 평균 60% 줄었다.

표 5. 교정 단위(형태소/어절)에 따른 사전 탐색 횟수 비교
Table 5. Comparison for Frequency of Dictionary Access.

교정 단위	1음절	2음절	3음절	4음절	5음절	6음절	7음절
형태소	3.10	13.40	6.90	39.21	34.10	153.30	302.92
어절 (후보수)	3.07 (3.07)	17.90 (13.30)	35.60 (10.12)	74.16 (13.01)	178.88 (23.85)	312.24 (39.03)	1679.10 (167.91)

실험 결과에서 후처리에 의한 교정 결과는 표 6과 같다. 실험에서 사용한 문서에 대한 인식기 인식률은 어절 인식률이 60%이고 음절 인식률이 83%였다. 후처리 결과 어절 인식률은 93.01%로 음절 인식률은 96.97%로 향상된 표 6과 같은 결과를 보인다. 표 6에서 1음절 어절이 다소 낮은 교정률을 가지는데 이것은 오인식된 형태나 철자 오류가 없어서 검출되지 못하였거나 후보 음절 중에서 인식을 원하지 않은 상위 신뢰도 음절이 사전 탐색에 성공하여 최종 후보로 선택되었기 때문이다. 그리고 논문을 제외한 문서에서 7음절 이상의 어절에서도 낮은 교정률을 보이는데 이것은 7음절 이상 어절에 대한 어절 형태 정보가 논문에 많이 영향을 받았기 때문에 다양한 형태의 긴 어절이 쓰인 소설에서 교정률이 떨어진다.

표 6. 교정 결과
Table 6. Result of Postprocessing.

구분	신문사설		초등교과서		소설		공학논문	
	전체 어절수	바른 인식	전체 어절수	바른 인식	전체 어절수	바른 인식	전체 어절수	바른 인식
1음절	164	141	111	98	132	118	134	124
2음절	438	430	447	441	446	441	307	303
3음절	586	579	579	568	516	507	385	379
4음절	384	366	238	224	298	285	273	262
5음절	138	125	127	117	99	82	94	83
6음절	93	90	52	48	83	79	80	73
7음절이상	54	51	34	28	78	74	63	60

실험 결과에서 오인식 어절이 후처리에 실패한 원인은 원인1) 철자 오류가 없어 형태소 분석에 성공한 오인식 어절, 원인2) 선택기준치에 의한 출력될 추정 형태 결정에서 제외된 어절, 원인 3) 최종 후보 선택에서 제외된 경우, 원인 4) 후보 음절 벡터에 인식을 원한 음절이 없는 경우이다. 실험 결과 후처리에 실패한 어절에서 각 원인의 비중을 보면 그림 9와 같다.

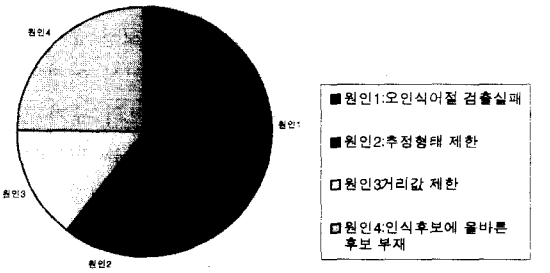


그림 9. 후처리 실패 어절 중 원인별 차지 비율
Fig. 9. A Proportion of 4-failure cause.

비교적 많은 비율을 차지하는 원인 1)과 원인 3)의 해결은 오인식 어절의 바른 교정이 형태소 분석만으로는 한계가 있음을 보여준다. 보다 의미적으로 올바른 어절의 교정이 되기 위해서 문맥적 지식과 의미 분석의 연구가 필요하다. 또한 음절수가 5이상인 어절에서 미리 분석된 형태소 정보가 하나 이하일 경우 생성되는 후보 개수를 줄이기 위해 양방향 일치에 의한 형태소 분석에 외에도 어절의 임의의 위치에서 특정 형태소를 분리해 내는 연구가 계속되어야 한다.

Ⅳ. 결 론

본 논문에서는 한국어 문자 인식 후처리에서 필요한 사전 탐색 횟수를 감소시키고 문자 인식기의 인식률을 높이는 방법으로 형태소 정보를 이용한 새로운 후처리 기법을 제안하였다. 형태소 정보를 이용한 후처리 기법은 형태소 분석과 사전을 이용한 방법으로 먼저 입력된 인식 결과를 형태소 분석하여 분석에 실패한 어절을 오인식 어절로 검출하였다. 오인식 어절 검출을 위해 형태소 분석을 하는 과정에 얻어진 일부 형태소 정보를 이용하여 오인식 어절의 형태소 구성을 추정한 후 추정된 형태소를 사전 탐색만으로 교정하였다.

한국어의 어절은 인접 가능한 형태소의 결합으로 구성된다. 한 어절에 포함된 임의의 형태소는 인접한 나머지 형태소를 제한한다. 이러한 한국어 어절의 맞춤법에서 오인식 어절의 분석된 일부 형태소 정보를 이용하여 가능한 어절의 형태를 추정하였다. 어절의 형태를 추정하기 위해 사전 품사의 정보를 기반으로 어

절의 모든 형태를 음절 길이별로 구분하여 어절의 형태 정보 테이블을 구성하였다. 이미 분석된 오인식 어절의 길이에 따른 가능한 형태소 구성 정보에 우선 순위를 부여하기 위해 말뭉치에서 같은 형태의 어절에 대한 출현 빈도를 구하였다. 오인식 어절에 대한 추정된 형태소 정보가 하나 이상인 경우에 교정 순서는 출현 빈도가 높은 순서대로 처리된다. 오인식 어절의 가능한 추정 형태가 많을수록 생성될 후보의 수가 급증하므로 이를 제한하기 위해 분석된 정보를 최대한 많이 가지는 형태를 우선 선택한다. 선택된 어절 형태의 출현빈도가 선택 기준치에 미달되면 분석된 정보를 최대한 포함하는 형태를 계속 찾아서 선택된 어절의 형태별 출현 빈도합이 선택기준치가 되도록 가능한 형태를 선택한다. 이렇게 추정된 오인식 형태소 정보를 이용하여 후보 음절 벡터에서 교정 후보를 생성한다. 불필요한 후보 생성으로 인한 사전 탐색 횟수를 줄이기 위해 인식 신뢰도가 높은 음절을 이용하여 후보를 우선 생성하여 사전 탐색에 의해 선택한다. 앞의 단계에서 교정 후보 생성에 실패하면 거리값을 이용하여 선택된 후보의 거리값보다 작은 거리값의 후보만 사전 탐색하여 올바른 후보를 찾는다. 그리고, 이미 사전 탐색에 실패한 형태소 정보를 기억해 두었다가 그 형태소를 포함하는 추정 형태는 후보 생성에서 제외하였다. 끝으로 선택된 올바른 후보가 없는 경우 대치 음절 정보를 이용하여 인식기가 자주 범하는 오인식 음절로 대치하면서 후보를 선택한다.

본 논문에서 제안한 형태소 정보를 이용한 후처리는 기존의 어절 단위 후처리에 비해 생성된 후보의 형태소 분석이 필요 없다. 생성된 후보가 형태소이므로 사전 탐색에 의해 올바른 후보를 선택할 수 있었다. 이로 인해 사전 탐색 횟수는 어절 단위 후처리와 비교하였을 때 60%나 감소되었다. 후처리 결과 문자 인식기의 음절 인식률이 94%에서 97%로 향상되었다. 문자 인식기의 성능이 좋아지면서 한 어절에 포함된 오인식 문자의 수도 줄어들고 있다. 오인식된 문자의 위치를 보다 정확하게 찾아내어 오인식 문자를 포함한 최소 단위의 교정은 후처리 시스템의 속도를 개선할 수 있었다. 하지만, 형태소 분석에 의한 오인식 검출로 인해 철자 오류가 없는 오인식 어절과 최종 교정 후보가 문맥적으로 의미에 맞지 않을 경우 기각하지 않고 그대로 출력하는 문제점이 있다. 앞으로 이러한 문제를 해결하기 위해 문맥적 지식의 사용과 의미 분석에 의한 제약 정보를 조사하여 해결해야 하겠다.

참 고 논 문

[1] 부산대학교 정보통신 연구소, "한글 철자 검사기/교정기 이식 및 글자 인식을 위한 후처리에 관한 연구," *제2차 과제, 최종 연구/개발 보고서, 삼성전자*, 1995

[2] 심철민, "어절 간 연관 관계와 오류 유형 추정 규칙에 기반한 한국어 철자 교정기," *부산대학교 석사 학위 논문*, 1995

[3] 윤면기, "자연어 처리를 위한 형태소 분석(I)," *인하대학교 석사 학위 논문*, 1984

[4] 이병희, 김태균, "한글 문자 인식에서의 오인식 문자 교정을 위한 단어 학습과 오류 형태에 관한 연구," *한국 정보처리학회 논문지*, 제3권, 제5호, pp.1273~1280, 1996

[5] 이상조, "한·영 기계 번역을 위한 중심어 기반 구 구조 변환 사전," *서울대학교 박사 학위 논문*, 1994

[6] 이영식, "사전 근사 탐색과 Heuristics를 이용한 한국어 철자 오류 교정 시스템 구현," *부산대학교 석사 학위 논문*, 1994

[7] 이영화, 김계성, 김영훈, 이상조, "문자 인식 후처리를 위한 형태소 분석기와 문자 교정기의 구현," *대한 전자 공학회 논문지*, 제34권, C편, 제5호, pp.82~92, 1997

[8] 최재혁, "양방향 최장 일치법에 의한 한국어 형태소 분석기의 구현," *경북대학교 박사 학위 논문*, 1993

[9] 허윤영, "언어적 지식과 경험적 제약을 결합한 문자 인식 결과의 후처리," *부산대학교 석사 학위 논문*, 1996

[10] 황호정, 도정인, 권혁철, "한글 문자 인식을 위한 후처리기의 개발과 속도 개선," *제2회 문자 인식 워크샵 발표 논문집*, pp.180~189, 1994

[11] K. Abend, "Compound decision procedures for unknown distributions and for dependent state in nature," *L. N. Kanai (Ed.), Pattern Recognition, Thompson Book Co.*, pp.207~247, 1968.

[12] W. W. Bledsoe and J. Browning, *Pattern Recognition*, New York: J. Wiley, pp.301~316, 1966.

[13] R. S. Boyer and J. S. Moore, "A fast string searching algorithm," *Comm. of*

ACM, Vol. 20, No. 10, pp.762~772, 1977.

[14] E. Charniak, "Statistical Language Learning," *MIT Press, Cambridge*, MA, 1993.

[15] K. W. Church and M. Hill, "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," *Second Conference on Applied Natural Language Processing*, pp.136~143, 1988.

저 자 소 개

金 永 勳(正會員) 第 34卷 C編 第 5號 參照
현재 안동과학대학 정보처리학과

李 相 祚(正會員) 第 34卷 C編 第 5號 參照
현재 경북대학교 컴퓨터공학과

李 英 和(正會員) 第 34卷 C編 第 5號 參照