문자 인식기의 특성과 말뭉치의 통계 정보를 이용한 문자 인식 결과의 후처리

손훈석, 최성필, 권혁철
부산대학교 전자계산학과

The Postprocessing of a Korean OCR using the Output of the Word Recognition and the Statistical Information from a Corpus

Hoon-Seok Son, Sung-Pil Choi, Hyuk-Chul Kwon
Department of Computer Science, Pusan National University

요약

한국어 문자 인식 후처리는 인식기가 제공하는 후보 음절을 바탕으로 후처리를 하였다. 이 논문은 문자 인식기가 제공하는 후보 음절 대신에 인식기의 인식 결과를 분석하여 인식기의 오인식 통계 정보에 따라 인식 결과 음절의 후보 음절을 생성한다. 여기서 생성된 후보 어절을 각 음절의 확률 값을 이용하여 확률이 가장 높은 어절을 선택한다. 이때 한국어 대용량 말뭉치에서 추출한 어절의 통계정보를 이용하여 그 어절의 확률 값을 구한다.

이 기법의 장점은 후보 음절의 조합으로 생성된 어절의 확률 값과 그 어절의 말뭉치상의 확률 값을 이용한 결과 말뭉치에 포함된 미등록어 정보에 따라 형태소 분석이 되지 않는 미등록어 처리가 가능하다. 또한 후보 어절 중 형태소 분석이 성공하는 어절이 두 개 이상 있을 경우 실제 거의 쓰이지는 않지만 단지 음절의 확률 값이 높아 우선으로 선택되는 경우를 방지하였다. 실험은 약 1,000page 분량의 실험을 통해 오인식 결과를 수집하고, 4000만 원시 말뭉치에서 구한 어절의 통계정보를 이용하였다. 그 결과 문자 인식기의 98.05%의 어절 인식률을 후처리 결과 99.52%로 향상시켰다.

1. 서론

현대 사회의 정보화 및 컴퓨터 산업의 급속한 발달로 인하여 정보 처리 자동화에 대한 요구가 높아지고 있다. 매년 방대한 양의 정보가 생산되고 있고, 기존 문서를 적절히 처리하기 위해서는 문자 인식 시스템이 필요하다. 문자 인식 시스템의 인식 결과는 입력 문서의 왜곡, 문자의 유사성, 인식 기술의 한계와 문서 자체의 오류 등으로 오류가 있을 수 있다. 따라서 인식률을 향상시키기 위해 문자 인식 후처리 시스템을 이용하여 인식 오류를 교정해야 한다. 이 논문은 한국어 문자 인식 시스템을 위한 후처리기 인식률을 개선하는 방법을 제시한다.

기존 문자 인식 후처리 기법은 n-gram 사전을 이용한 방법과 형태소 분석기를 이용한 방법으로 나뉜다. 이 중 n-gram 사전을 이용한 방법은 영어권에서 널리 사용된다. 그러나 한국어, 일본어, 터키어와 같이 어형의 변형 꼴이 많은 언어에서 믿을 만한 n-gram 사전을 구하는 것이 어렵다. 특히, 한국어는 명사 간 띄어쓰기에 대한 맞춤법 규정이 약하므로, n-gram을 적용하기가 어렵다. 이러한 문제점을 극복하고자 형태소 분석기를 이용한 후처리 기법이 사용된다.

형태소 분석기를 이용한 문자 인식 후처리 시스템에서 처리 속도를 빠르게 하고 인식률을 높이기 위해 다양한 기법을 사용한다. 후처리 속도를 개선하기 위해서 ①음절 bi-gram, ②viable-prefix, ③거리평가함수가 사용되고, 인식률을 향상시키기 위해 ①형태소 분석 정보, ②음절 혼동 행렬이 사용된다. 그러나 기존 문자 인식 시스템은 사용 빈도가 낮은 어절이 형태소 분석됨으로써 인식 결과로 잘못 제시되는 문제와 미등록어를 다른 어절로 대치하여 제시하는 문제가 있다.

형태소 분석을 이용한 문자 인식 후처리 시스템에서 발생하는 문제점을 해결하기 위해 말뭉치에 기반한 기법과 형태소 분석기에 기반한 기법을 결합한 혼합 기법을 사용한다[1].

지금까지의 후처리 시스템은 인식기에서 제공하는 인식 결과에 대한 후보 음절을 통해 후처리를 하였다. 인식기가 인식 결과와 함께 후보 음절을 제공하지 않으면 후처리를 할 수 없다. 이 논문에서 구현한 후처리기는 인식기의 인식 결과를 분석하여 후보 음절을 만들고, 각 후보 음절이 인식기가 인식하고자 했던 음절이었을 확률 값을 구하여 후보 음절에 부여한다. 이렇게 구한

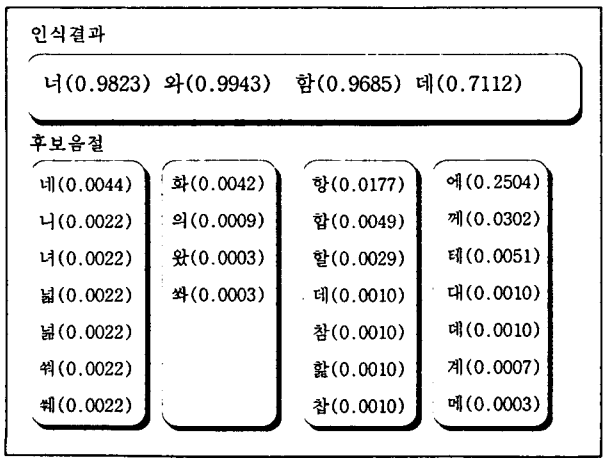
후보 음절에서 후보 어절을 생성하고 각 음절의 확률 값으로 각 후보 어절의 확률 값을 구한다[8].

후보 어절 중 형태소 분석되는 어절이 두 개 이상 있을 때 후처리 시스템은 음절 확률 값을 통한 후보 어절의 확률 값과 말뭉치 통계 정보를 통한 어절 확률 값을 고려하여 인식 어절을 선택한다. 따라서 상대적으로 빈도가 낮은 어절이 제시되는 경우가 감소하고, 대용량의 말뭉치를 기반으로 정보를 추출함으로써 미등록어를 다른 말로 잘못 교정하는 경우를 감소시킨다.

2. 후보 음절 생성

문자 인식 시스템은 오류가 있을 수 있다. 그래서 후처리 시스템은 인식기의 특성에 따라 유사한 문자 간의 혼동에 따른 후보 음절들을 이용하여 인식기의 오류를 교정해야 한다.

기존의 후처리는 문자 인식기에서 각 음절 당 인식 신뢰도에 따라 10개의 후보 음절을 후처리 시스템에 제공해 준다[8]. 그리고 후보 음절과 함께 후보 음절 각각에 인식 신뢰도를 넘겨 준다. 이렇게 인식기가 제공해 주는 후보 음절은 그 인식기의 특성에 맞는 적당한 후보 음절이다. 그러나 후보 음절을 제공하지 않는 인식기의 인식 결과는 기존의 방법으로는 후처리 할 수가 없다. 이 논문에서 적용하는 후처리 기법은 후보 음절을 제공하지 않는 인식기의 결과를 후처리 시스템에서 후보 음절을 생성하여 후처리한다. 후보 음절을 생성하는 방법은 많은 문서의 이미지를 인식기로 인식해 인식기가 인식하기 위한 원문과 인식 결과를 비교 분석한다. 인식기가 오인식한 음절과 원문의 바른 음절의 쌍을 모아서 그 인식기의 인식 결과에 대한 후보 음절을 얻는다. 문자 인식기의 인식 오류는 모양이 비슷한 문자 간에 주로 발생하므로 가능하다. 하지만 이 오류는 인식기에 따라 다르다. 그래서 해당 인식기의 인식 결과를 후처리 하려면 그 인식기의 결과를 바탕으로 후보 음절을 생성해야 한다. 후처리 시스템에서는 이렇게 생성된 후보 음절에 가중치를 주고 그 가중치에 따라 후처리를 한다. 후보 음절의 가중치는 후보 음절을 생성하는데 사용되었던 문서를 모집단으로 두고 베이즈의 법칙(Bayes' rule)을 사용하였다. 인식기가 문서를 인식하는 실험이 모두 수행되고 그 결과를 통해 생성된 오인식 음절 쌍을 통해 생긴 후보 음절 각각이 원문이었을 사후 확률(posterior probability)을 이용한다. [식 1]을 이용하여 인식 결과 음절 R을 보고 후보 음절 S_i 각각이 인식기가 인식하기 전 원문이었을 확률 값으로 후보 음절의 가중치를 정한다.



[그림 1] 확률 값에 따라 정렬된 후보 음절

$P_A(Si)$	원문에서 음절 Si가 나타날 확률
$P_B(R)$	인식 결과에서 음절 R이 나타날 확률
$\Pr(Si \mid R)$	인식 결과가 R일 때 원문이 Si이었을 확률
$\Pr(R \mid Si)$	원문에 Si가 나왔을 때 인식기가 R로 인식할 확률

$$\Pr(Si \mid R) = \frac{\Pr(R \mid Si) P_A(Si)}{P_B(R)}$$

($1 \le i \le$ 인식결과 R의 후보 음절 개수)

[식 1] 후보 음절의 확률

[그림 1]을 보면 원문은 '너와 함께'였는데 인식기가 '너와 함데'로 인식하였다. 후처리 시스템은 인식 결과인 '너와 함데'를 보고 각 음절 '너', '와', '함', '데'에 따른 후보 음절과 그 후보 음절 각각이 원문이었을 확률 값을 구하여 후처리를 시작한다.

3. 시스템의 구성

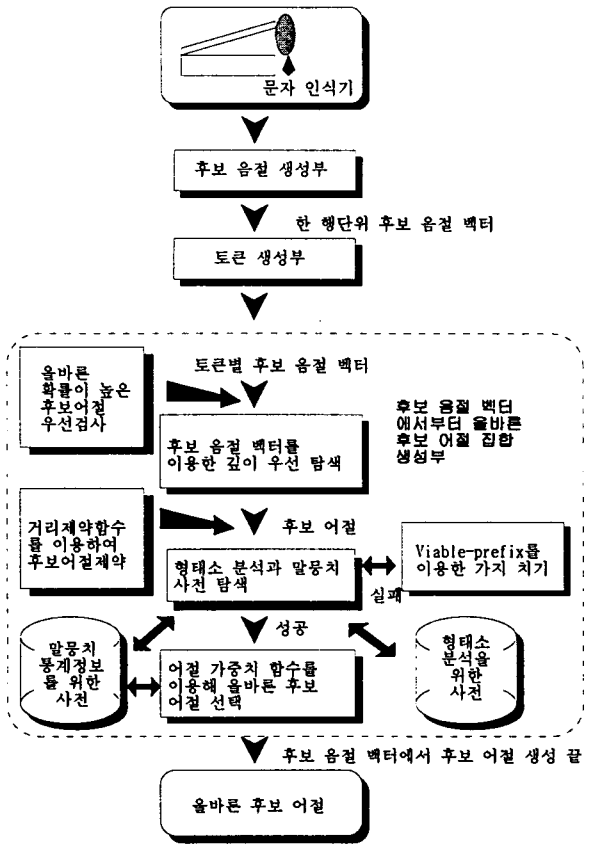
3.1 후처리 시스템의 구성

[그림 2]는 이 논문에서 구현한 후처리 시스템의 전체 구성도이다.

3.2 말뭉치 분석 정보

문자 인식 후처리기의 핵심 요소인 형태소 분석기에 대해서는 많은 연구가 진행되었다. 지난 20여 년 간 연구한 한국어 형태소 분석기는 그 성능이 실용화 수준에 도달했다. 그러나 올바른 어절이 형태소 분석에 실패하거나 틀린 어절이 분석에 성공하는 경우가 아직 발생하고 있으며 이 부분은 한국어 처리의 중요한 연구 주제 중 하나이다.

이 논문은 대용량 말뭉치(약 4000만 어절)를 분석하여 추출한 어절의 통계 정보를 이용하여 형태소 분석기의 문제점을 개선한다. [표 1]는 대용량 한국어 말뭉치에서 빈도 순으로 어절의 수에 따라 말뭉치 전체에서



[그림 2] 후처리 시스템의 전체 구성도

어절을 포함하는 정도를 보여준다.

빈도 순 상위 어절 개수	전체 말뭉치에서 차지하는 어절 개수	전체 말뭉치에서 차지하는 비율
10만	35,623,490	81.08%
20만	37,868,003	86.19%
30만	38,990,096	88.75%
40만	39,706,753	90.38%
50만	40,237,416	91.59%

[표 1] 말뭉치에서 어절 수에 따른 포함 정도

[표 1]에 따르면 4,000만 어절 말뭉치는 총 2,752,416가지의 어절로 구성되어 있으며 이 중에서 빈도순으로 50만 어절이 4,000만개의 어절로 이루어진 한국어 말뭉치에서 91.59%를 차지한다. 이 값으로 보아 말뭉치에서 빈도 순으로 상위 50만 어절이 일반 문서에서 어절의 90%이상을 차지한다.

그리고 말뭉치에는 오류 어절이 포함되어 있다. 문자 인식 후처리기는 올바른 어절 뿐만 아니라 사투리, 띄어쓰기 등 일반 문서에서 자주 틀리는 오류도 분석해서 제시할 수 있어야 한다. 그러므로 자주 틀리는 오류를 처리하기 위해 말뭉치의 오류 어절을 가공하여 말뭉치 사전에 추가하였다. [표 2]는 말뭉치 사전에 추가한 오류 어절의 예를 나타낸다. [표 2]에 있는 어절과 같이 말뭉치 사전에 추가한 오류 어절은 주로 자주 쓰이는 띄어쓰기 오류와 사투리이다.

이 논문에서 구현한 문자 인식 후처리 시스템은 빈도순으로 상위 50만 말뭉치 사전을 사용한다. 이 사전에 올바른 어절과 함께 [표 2]에서 보는 바와 같이 오류 어절도 포함하는 사전을 구성하였다. 상위 50만 어절에 포함되어 있는 사투리, 띄어쓰기 등의 오류는 일반 문서에서 자주 발생할 수 있는 어절이므로 이런 어절을 후처리 할 때 형태소 분석만을 사용하여 바른 인식결과를 틀리게 교정하는 경우를 방지할 수 있다.

말뭉치에 존재하는 오류 어절	올바른 어절
여러가지	여러 가지
두번째	두 번째
씌어진	쓰인, 씐
그리구	그리고
몇가지	몇 가지
센치미터	센티미터
알맞는	알맞은

[표 2] 말뭉치 사전에 포함된 오류 어절 예

3.2 어절 가중치 함수를 이용한 기법

이 논문에서 만든 후보 음절은 그 음절이 인식기가 인식하고자 하였던 원문이었을 확률 값의 순서로 정렬되어 있다. 그리고 해당 음절에 대해서 첫 번째 후보 음절과 두 번째 후보 음절이 올바른 인식 결과 음절일 가능성이 높다.

형태소 분석기의 분석 정보만을 이용하는 후처리 시스템은 사용 빈도가 낮은 어절이 형태소 분석됨으로써 인식 결과로 잘못 제시되는 문제와 미등록어를 다른 어절로 대치하여 제시하는 문제가 있다. 이러한 인식률에 관한 문제점을 해결하기 위해 부산대학교 문자 인식 후처리 시스템은 말뭉치를 이용한 형태소 분석 기법을 문자 인식 후처리 시스템에 적용하였다[1]. 이 기법은 후보 음절 벡터로부터 신뢰도가 1,2순위의 후보 음절들로써 후보 어절을 생성하고 이들 각각을 말뭉치의 빈도 가중치를 이용하여 검증하는 기법이다. 이 기법은 '말뭉치 사전에 존재하는 어절이 두 개 이상일 때 거의 쓰이지 않는 어절이 후보 음절의 확률 값이 높아서 후처리 최종 결과 어절로 제시된다' 는 문제점이 있다.

이러한 문제 점을 해결하기 위하여 후보 어절의 가중치 함수를 이용한 기법을 사용한다. 이 기법은 각 후보 음절의 확률 값을 통해 후보 어절이 원문이었을 확률과 말뭉치에서 구한 어절의 일반 문서에서 나타날 확률을

이용한 기법이다. [식 2]는 이 논문에서 사용한 가중치 신뢰 함수를 나타내고 있다. 먼저 후보 어절을 구성하는 각 후보 음절이 인식기가 인식하고자 했던 음절일 확률을 곱한다. 이렇게 구한 값이 그 후보 음절의 음절 확률 값으로 구한 후보 어절의 확률 값이다. 그리고 말뭉치의 통계 정보를 이용하여 후보 어절이 일반 문서에 나타날 확률 값을 다시 곱한다.

$$\begin{matrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ c_{31} & c_{32} & \cdots & c_{3n} \\ \vdots & \vdots & \cdots & \vdots \end{matrix}$$: 후보 음절

$\Pr(c_{ij} \mid R)$	인식 결과가 R일 때, 순위가 i 등이고, 어절에서 j 번째 후보 음절이 원문이었을 확률
$S=(c_1, c_2, \cdots, c_n)$	후보 어절
n	어절 길이
$CPr(S)$	어절 S가 말뭉치에 있을 확률
$W(S)$	후보 어절의 가중치

용어 설명

$$W(S) = \prod_{j=1}^{n} \Pr(c_i \mid R_j) \times CPr(c_1 c_2 \cdots c_n)$$

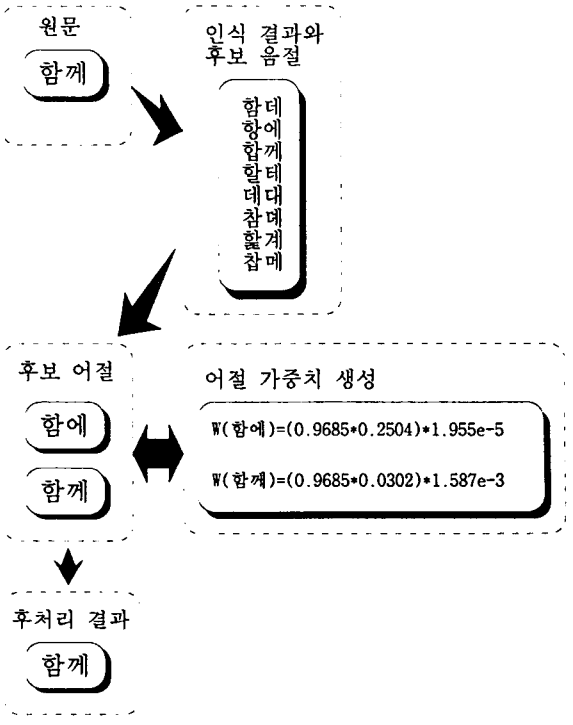
[식 2] 후보 어절의 가중치 함수

[그림 3]은 후보 음절 중 말뭉치에 존재하는 어절이 두 개 이상 있을 때 [식 2]를 사용하여 후처리 결과를 결정하는 과정을 설명하고 있다. 올바른 어절이 '함께'인데 인식결과가 '함데'일 경우 결과 '데'에 대한 후보 음절은 음절 확률 값에 의해 '에', '께', '테'의 순서로 생성된다. [식 2]를 사용하여 음절 확률 값을 통한 어절의 확률 값과 말뭉치의 통계 정보를 이용한 어절의 확률 값을 곱해서 각 후보 음절의 가중치를 구한다. 높은 가중치의 어절을 최종 후처리 결과로 결정하면 음절 확률 값은 높지만 거의 사용되지 않는 '함에' 대신에 어절 확률 값이 높은 '함께'가 후처리 결과로 선택된다. 다른 예를 들어 보면 원문이 '겪어야'이고 인식 결과가 '겸어야'로 나왔을 때, 형태소 분석이 되고 말뭉치에 존재하는 후보 어절은 '경어야'와 '겪어야' 순으로 두 개가 남는다. 이 경우도 역시 인식 결과 음절 '겸'에 대해 '경(0.1550)'이 '겪(0.0310)' 보다 확률 값이 높다. 그러나 말뭉치에서 어절 확률 값이 '겪어야'가 월등히 높아 '겪어야'가 후처리 결과로 결정된다.

4. 실험 및 결과

4.1 실험 방법

이 실험에서 사용한 인식기는 주노시스템에서 개발한



[그림 3] 어절 가중치 함수를 이용하여 후처리 결과 결정

'스피드리더'라는 문자 인식기이다. 인식기의 인식 결과를 얻기 위해 약 1000page 가량의 다양한 문서를 스캐너로 읽어 들여 이미지를 구하고 이를 인식기로 인식시켜 문서의 원본과 인식 결과를 비교하였다. 비교 결과 9,381개의 오인식 음절 쌍을 구했고, 이 오인식 음절 쌍에서 완성형 코드 내의 문자 1,890개와 완성형 코드를 벗어나는 문자 1,192개에 대한 후보 음절을 만들었다. 그리고 후보 음절에 대한 가중치는 인식 결과를 얻기 위해 사용한 1,000page를 전체 모집단으로 두고 [식 1]을 이용해서, 인식 결과를 보고 후보 음절 각각이 인식기가 인식하고자 하였던 음절이었을 확률 값으로 구했다. 또한, 말뭉치 상의 어절 확률 값을 얻기 위해 원시 말뭉치 약 4,000만 어절을 분석하여 빈도 순으로 상위 50만 어절을 사용하였다. 실험 데이터는 인식 결과를 얻기 위해 사용한 문서와 독립적인 22,481개의 어절로 구성된 교과서와 논설문 데이터를 사용하였다.

4.2 실험 결과

이 논문은 후처리기의 인식률 개선 성능 평가를 위해서 다음의 세 가지 방법을 이용하여 결과에 대한 비교, 분석을 한다.

- 후처리를 하지 않은 경우
- 형태소 분석기 만을 사용하는 경우
- 형태소 분석과 어절의 말뭉치 빈도를 결합한 경우
- 어절 가중치 함수를 사용한 경우

후처리를 하지 않은 경우는 인식기의 결과가 올바른지 아닌지를 검사한다. 형태소 분석기와 말뭉치 빈도를 결합한 방법은 후보 음절 순위 1,2위로 생성되는 후보 어절을 형태소 분석기와 대용량 한국어 말뭉치(4,000만 어절)에서 추출한 상위 빈도 50만 어절 사전을 통해서 검증하였다. 어절 가중치 함수를 이용한 방법은 후보 어절 중에 상위 빈도 50만 어절 사전에 존재하는 어절이 두 개 이상 있을 경우는 후보 어절의 후보 음절의 조합으로 생성된 확률 값과 후보 어절이 말뭉치에 나타날 확률 값을 이용한 방법이다.

[표 3]은 말뭉치 통계정보를 사용한 후처리기를 인식기에 반영하였을 때 전체 인식률을 후처리기를 사용하지 않은 인식기의 인식률과 형태소 분석과 말뭉치 빈도를 사용한 후처리의 인식률에 비교한 결과이다.

* 후처리 대상 어절 : 22,481개

	인식기	형태소 분석기를 주로 사용	형태소 분석과 말뭉치 사용	어절 가중치 함수를 사용
개수	22,043개	22,318개	22,359개	22,372개
인식률	98.05%	99.27%	99.46%	99.52%

[표 3] 처리 결과 어절 인식률

[표 3]을 보면 인식기의 어절 인식률 98.05%를 형태소 분석기를 주로 사용한 후처리로 99.27%, 형태소 분석과 말뭉치를 사용하여 99.46%로 개선 시켰고, 여기에 어절 가중치 함수를 사용한 후처리를 통해 99.52%까지 향상시켰다. 이 실험 결과를 통해 인식기에서 후보 음절을 제공하지 않고 인식기의 인식 결과만을 가지고 후처리에서 후보 음절을 생성했을 때 인식기의 오인식률 1.95%를 형태소 분석기를 주로 사용한 후처리를 하여 0.73%로 줄여 오인식률을 62.56% 감소시켰고, 형태소 분석과 말뭉치를 결합한 방법을 사용하여 0.54%로 줄여 오인식률을 72.15% 감소시켰다. 여기에 어절 가중치 함수를 추가로 사용하여 오인식률을 0.48%로 줄여 오인식률을 75.11% 감소시켰다.

5. 결론

이 논문에서는 문자 인식기가 인식 과정에서 발생하는 오류를 교정하는 문자 인식 후처리 시스템을 계발하였다. 기존의 후처리 시스템은 문자 인식기에서 후보 음절을 제공하여 후처리를 하였다. 이 논문에서 제안한 시스템은 문자 인식기가 제공하는 후보 음절 대신에 인식기가 인식한 결과를 분석하여 후보 음절을 생성하였다. 후보 음절을 생성하는 방법은 인식기의 인식 결과인 각 음절의 후보 음절이 인식기가 인식하고자 했던 원래 음절이었을 확률 값을 통해 후보 음절의 우선 순위를 주었다. 그 결과 문자 인식기에서 후보 음절을 제공하였을 때와 비슷한 결과를 얻었다.

이 논문에서 개발한 시스템은 인식기의 인식 결과를 분석하여 후보 음절을 제공하지 않는 인식기의 인식 결과도 후처리를 할 수 있다. 그리고 기존의 형태소 분석기와 말뭉치를 사용했을 때의 문제점을 해결하기 위해서 후보 음절의 확률을 통한 후보 어절의 확률 값과 대용량 말뭉치의 통계 정보를 통한 어절의 확률 값을 이용하여 어절 가중치 함수를 사용하여 해결하였다.

인식기가 후보 음절과 후보 음절에 대한 신뢰 값을 제공한다는 것은 그 인식기가 문자를 인식할 때의 상황을 고려하여 후처리 시스템으로 넘겨 주는 것이다. 이 논문에서 제시한 방법은 인식기가 후보 음절을 제공해 주지 않고 인식 결과만을 분석하여 후처리 시스템에서 후보 음절을 생성하므로 각 인식 결과 음절에 대해 후보 음절과 후보 음절에 대한 확률 값을 생성할 때 그 값이 고정된다. 그래서 비록 확률 값을 사용하여 인식기의 결과를 향상시키기는 했지만 [그림 3]에서와 같이 원문이 '함에'이고 인식 결과가 '함데'일 때 후처리 시스템은 '함께'로 교정한다는 한계가 있다. 이러한 한계를 극복하는 것은 향후 연구 과제로 남긴다.

후처리 시스템이 후처리에 실패한 경우의 원인을 보면 다음의 두 종류가 대부분을 차지했다.

- 인식기가 인식에 실패하여 특수기호가 결과로 나오거나 한글을 영어나 숫자로 인식한 경우
- 인식 결과에 대한 후보 음절 중에 맞는 음절이 없을 경우

맞는 말인 후보 음절이 없는 경우를 대비하기 위해서는 후보 음절을 생성하기 위한 문서를 더욱 다양하고 많은 양으로 늘여야 한다.

그러나 후보 음절을 생성하기 위해 인식기의 인식 결과를 완벽하게 분석한다는 것은 무한대의 실험 데이터가 필요하다. 이런 문제점을 해결하기 위해서는 음소 단위의 오인식 결과를 분석해서 음절 혼동 행렬을 생성하여 후보 음절을 보완해야 한다.

[참고 문헌]

[1] 김민정, 규칙과 말뭉치를 이용한 한국어 형태소 분석과 중의성 제거, 부산대학교 전자계산학과 이학박사 학위논문, 1997.

[2] 이원일, 홍남희, 이종혁, 이근배, Binary N-gram과 형태소 분석기를 이용한 한국어 철자 교정기, 한국정보과학회 93년 학술발표논문집, pp.813-816, 1993.

[3] 유진희, 이종혁, 이근배, 형태소 분석과 언어 평가를 통한 문자 인식 후처리기, 한국 인공지능 연구회 94년 춘계 학술발표논문집, 1994.

[4] 이병훈, 윤준태, 송만석, 말뭉치를 기반으로 한 한국어 철자 교정기의 구현, 제5회 한글 및 한국어 정보처리 학술대회 학술발표논문집, pp.285-294, 1993.

[5] 박진우, 이일병, 통계적 방법에 의한 후처리, 제6회 한글 및 한국어 정보처리 학술대회 학술발표논문집, pp.518-526, 1994.

[6] 이성환, 김은순, 주소 및 성명에서의 한글 인식을 위한 효율적인 오인식 교정 알고리즘, 정보과학회 논문지, Vol. 20, No. 5, May, 1993.

[7] 부산대학교 정보통신연구소, 한글 철자 검사기/교정기 이식 및 글자 인식을 위한 후처리 기술에 관한 연구 2차 과제, 최종 연구/개발 보고서, 삼성전자, 1995.

[8] 황호정, 한글 문자 인식을 위한 후처리기의 개발, 부산대학교 전자계산학과 이학석사 학위 논문, 1995.

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

[10] 심철민, 김민정, 이영식, 권혁철, 단어 간 지배 관계 및 연관 관계를 이용한 한국어 교열 시스템, 제5회 한글 및 한국어 정보처리 학술발표논문집, pp.303-316, 1993.

[11] 민병우, 이성환, 김홍기, 문자 인식을 위한 후처리 기법의 사례 연구, 제1회 문자 인식 워크샵 발표 논문집, pp.91-104, 1993.

[12] 고왕경, 기초확률론, 경문사.

[13] Chul-Min Sim, Min-Jung Kim, Hyuk-Chul Kwon, Automatic Revision of Korean Texts by Collocation Words, Proceedings of ICCPOL, pp.280-284, 1994.

[14] Tsuyoshi Kitani, An OCR Post-Processing Method for Hanwritten Japanese Documents, NLPRS 91, 38-45, Nov. Singapore, 1991.

[15] Katsumi Marukawa, Masashi Koga, Yoshihiro Shima, A Post-Processing Method for Handwritten Kanji Name Recognition Using Furigana Information, ICDAR 93 2st Int. Conf. on Document Analysis and Recognition, pp.218-221, 1993.