한국어의 음절 결합 특성 및 통사적 어휘 특성을 이용한 문자인식 후처리 시스템

황영숙, 박봉래, 임해창
고려대학교 컴퓨터학과 자연어처리 연구실

Post-processing for Korean OCR Using Cohesive Feature between Syllables and Syntactic Lexical Feature

Young-Sook Hwang, Bong-Rae Park, Hae-Chang Rim
NLP Lab. Dept. of Computer Science & Engineering Korea Univ.

지금까지의 한글 문자인식 후처리 연구분야에서 미등록어와 비문맥적 오류 문제는 아직까지 잘 해결하지 못하고 있는 문제이다. 본 논문에서는 단어로서 가능한지를 결정하는 기준으로 확률적 음절 결합 정보를 사용하여 형태소 분석 기법만을 사용했을 때 발생할 수 있는 미등록어 문제를 해결하고, 통사적 기능의 어말 어휘를 고려한 문맥 결합 정보를 이용함으로써 다수의 후보 어절 가운데에서 최적의 후보 어절을 선택하는 방법을 제안한다. 제안된 시스템은 인식기에서 내보낸 후보 음절과 학습된 혼동 음절을 조합하여 하나 이상의 후보 어절을 생성하는 모듈과 통계적 언어 정보를 이용하여 최적의 후보 어절을 선정하는 모듈로 구성되었다. 실험은 1000만 원시 코퍼스에서 추출한 음절 결합 정보와 17만 태깅된 코퍼스에서 추출한 어절 결합 정보를 사용하였으며, 실제 인식 결과에 적용한 결과 문자 단위에서는 94.1%의 인식률을 97.4%로, 어절 단위에서는 87.6%를 96.6%로 향상시켰다. 교정률과 오교정률은 각각 문자 단위에서 56%와 0.6%, 어절 단위에서 83.9%와 1.66%를 보였으며, 전체 실험 어절의 3.4%를 차지한 미등록어 중 87.5%를 올바로 인식하는 한편, 전체 오류의 20.3%인 비문맥 오류에 대해서 91.6%를 올바로 교정하는 후처리 성능을 보였다.

I. 서론

정보를 자동생산, 관리하고자 하는 현대 정보 사회의 요구와 함께 문자인식 시스템에 대한 연구가 활발히 진행되고 있어 인식 시스템에 대한 기대가 점점 높아지고 있다. 그러나 현재의 인식 성능은 사용자들이 만족할 수 있을 만한 수준에는 이르지 못하고 있다. 이러한 문자인식의 한계를 극복하고 인식 성능을 효과적으로 향상시키기 위해서는 인식 결과를 분석하여 적절한 처리 방법을 선택하는 것이 필요하다.

한국어의 단어 및 문장 구조상에서 문자인식의 오류 유형을 살펴보면 크게 비단어 오류(non-word error)와 비문맥 오류(real-word error)로 나누어 볼 수 있다. 비단어 오류는 다음의 예 1)에서 보는 바와 같이 인식결과로 출력된 문자열이 단어로서 인정되지 못하는 오류이다. 비문맥 오류는 출력된 문자열이 그 자체만을 살펴볼 때는 단어로서 인정되지만, 문맥적으로 살펴볼 때는 올바르게 사용되지 않은 오류를 말한다. 예 2)를 살펴보면, 하나의 문장을 구성하기 위해 나열된 문자열들 중 '기출이', '이르기', '꽃하고' 라는 문자열들은 그 자체는 어절로서 인정될 수 있지만 문맥상으로는 잘못 사용되고 있음을 알 수 있다.

예 1) 기줄이, 아직모, 틸용화, 이르치, 꽂하고, ...

예 2) 기출이 아직도 실용화 단계에 이르기 꽃하고.....

이러한 인식 오류들은 대부분 문서 영상내에 존재하는 잡음이나 문자간 접촉 혹은 유사 문자간의 혼동으로 인해 발생하며, 문자단위의 처리에 의존하는 현재의 문자인식 방법만으로는 해결하기 어려운 문제이다. 그러므로 비단어 오류나 비문맥 오류를 교정하기 위해서는 인식기의 유사문자에 대한 혼동 정보와 사용 언어의 특성 정보를 이용하여 후처리를 수행하는 것이 효과적이라 할 수 있다.

지금까지 진행되어 온 문자인식 후처리 관련 연구를 살펴보면 단어 구조를 중심으로 비단어 오류를 교정하는 방법론과 문장 구조로 확장하여 비문맥적 오류를 탐색, 교정하기 위한 방법론들이 제시되고 있다. 국외의 경우, 단어내에서의 전이 확률과 혼동 확률을 이용한 Viterbi 알고리즘, 문자열 정보를 이용한 N-gram 알고리즘등이 비단어 오류의 검출 및 교정을 위해 제시되었고, 품사 태깅이나 공기 패턴 정보를 이용한 방법들이 비문맥 오류를 탐색하거나 교정하기 위해 제시되었다 [May91,Tong96].

한글 문자인식 후처리 연구 분야에서도 한국어의 특성을 고려한 여러가지 방법론이 제안되어 왔다[김민정97, 민병우91, 박진우94, 이종연93, 유진희95, 홍남희93, 황호정94]. [이종연93, 홍남희93]에서는 비단어 오류를 검출하기 위해 형태소 분석을 하고, 형태소 분석 결과

검출된 오류를 교정하기 위해 자소 단위의 n-gram 사전을 이용하는 방법을 제시하였는데, 비단어 오류는 교정해도 비문맥적 오류는 교정하지 못한다는 문제가 제기되고 있다. [유진희95]는 비문맥적 오류를 교정하기 위해 형태소 분석과 언어평가 함수를 이용하는 방법을 제안하였는데, 여기서는 인식기의 후보 문자와 유사 문자들의 조합으로 후보 어절을 생성하고 형태소 분석과 어절단위의 N-gram 품사 태깅, 그리고 공기 패턴 정보를 이용하여 비단어 오류와 비문맥 오류를 교정하는 방법을 제시하고 있다. 또한 한국어 어절의 Viable-Prefix 정보와 형태소 분석, 그리고 공기 패턴 정보를 이용하여 인식률과 처리 속도를 개선하고자 하는 방법이 [김민정97, 황호정94]에서 제시되기도 하였다.

그러나 지금까지의 연구들은 대부분 형태소 분석을 중심으로 한 후처리 방법으로써 형태소 분석만을 수행했을 때 발생할 수 있는 미등록어 문제를 전혀 고려하지 못하고 있다. 즉, 형태소 분석으로 비단어 오류를 검출하는 방법은 형태소 분석 실패 어절들을 올바른 후보 어절에서 배제시키기 때문에 올바른 인식 결과가 미등록어가 되는 경우에는 교정할 수 없다는 문제가 제기된다. 또한 품사 태깅과 공기 패턴 정보를 이용하는 경우 후보 음절과 혼동 음절의 조합으로 생성된 후보 어절들 사이에 동품사 중의성 문제가 발생할 수 있고, 결과적으로 공기 패턴 정보에 대한 의존도가 심해질 수 있다. 이는 필요한 공기 패턴 정보를 자동으로 추출하기가 어렵다는 문제와 공기 패턴 정보의 저장, 탐색과 관련하여 시스템 효율성이 문제로 대두되게 된다.

본 논문에서는 형태소 분석 기법만을 사용했을 때 발생할 수 있는 미등록어 문제를 음절 결합 확률 정보와 형태소 품사 결합 확률 정보를 함께 이용하여 해결하는 방법을 제안하다. 또한 통사적 기능의 어말 어휘를 고려한 어절 태깅 방법을 사용함으로써 동품사 중의성 문제를 어느 정도 약화시키고, 공기정보를 사용하지 않으면서 비문맥 오류를 교정하는 방법을 제안한다.

II. 미등록어와 비문맥 오류를 고려한 후처리

문자인식 시스템의 성능 향상을 도모하기 위해서는 인식기의 특성을 고려하는 한편, 인식기의 유사 문자간 혼동에 따른 비단어 오류와 비문맥 오류를 복합적으로 다루어야 한다. 인식기의 특성은 유사 문자 간의 혼동 정보로 나타날 수 있는데, 학습의 정도에 따라 유사 문자들에 대한 혼동 정도가 다르게 나타날 수 있다. 이를 인식기에서는 후보 문자들과 함께 신뢰도로써 제시하며, 후보 문자들 사이에서도 나타나지 않는 혼동 문자들에 대해서는 인식 결과들에 대한 별도의 재학습 과정을 거쳐 혼동 문자 테이블로 구성할 수 있다.

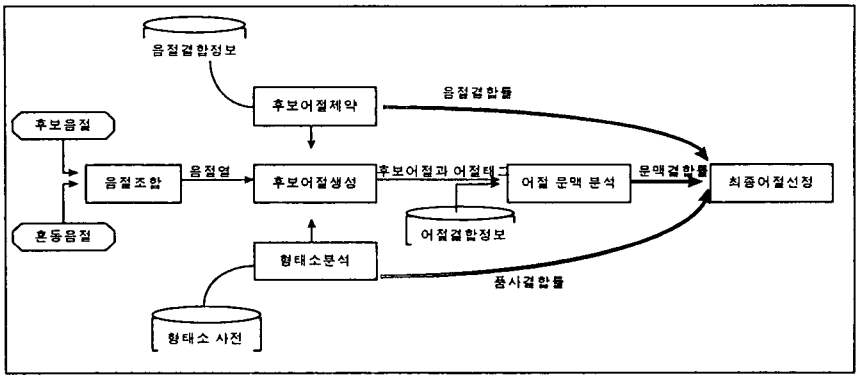
후처리 과정 중 비단어 오류 문제를 다룰 때 고려해야 하는 중요한 이슈 중의 하나는 미등록어 문제인데, 이를 해결하기 위한 방법 중의 하나는 주어진 일정 문자열이 단어로서 가능한 문자열인지를 판단할 수 있는 기준을 제시하는 것이다. 단어 형성에 영향을 미치는 언어학적 요인으로는 음운론적 특성과 형태론적 특성이 있을 수 있다. 표음 문자인 한국어의 경우 모음조화 현상이나 용언의 규칙, 불규칙 활용등에 따른 형태론적 변형 특성이 단어의 표층부에서 음절 특성으로 나타난다. 그러므로 음절 특성과 형태론적 특성을 결합하면 단어로서 가능한지를 결정할 수 있는 기준을 제시할 수 있게 된다.

또한, 비문맥 오류를 다루기 위해서는 한국어의 문장 구조 특징을 고찰하고 그에 따른 올바른 어절 선택 기준을 제시하여야 한다. 한국어의 문장구조는 하나이상의 어절들로 구성되고 어절간의 통사적 연결 관계에 의해 문장이 이루어진다. 즉, 어절은 통사 구조로 이루어지고, 통사적 기능의 형식 형태소는 의미상 그것이 속하는 어절내에서의 최소 자립 형태에만 관계하지 않고 문장내 인접 어절의 의미 내용에 대한 판단 형식을 결정함과 같은 문법적 의미도 갖게 되어 통사적 문법 구조를 이루게 된다. 그러나 통사 구조에 나타난 통사 정보만으로는 문법을 만족스럽게 설명할 수 없으며 이 구조에 잠재되어 있는 형태정보를 이용하여야 할 경우가 문법 현상에는 많이 나타난다[시정곤94]. '철수는 어제 산 시계를 잃어 버렸다.'라는 예문을 살펴보자. '-는, -ㄴ, -를, -어, -었다'의 형태소는 형태적으로 각각 통사적 기능의 형식 형태소이면서 각각 선행 어간에 의해 영향받아 이형태가 결정된 것이다. 이렇듯 형태적 정보에 의해 어절의 표면형이 결정되는 것을 살펴볼 때, 표면형이 다른 다수의 후보 어절을 다루는 후처리 방법에서 형태론적 속성에 기반한 좌우 어절간의 통사적 문맥 구조 정보는 올바른 어절을 선택할 수 있는 하나의 기준으로 제시될 수 있다.

이와 같은 점들을 고려하여 본 연구에서는 인식 시스템의 혼동 정보와 한국어의 음운적, 형태적, 통사적 문맥 특성 정보를 이용한 후처리 방법을 제안한다. 제안된 시스템에서 사용하는 인식기의 특성 정보는 인식기로부터의 후보 음절 정보와 인식 결과를 학습하여 얻은 혼동 음절 정보이며, 한국어의 특성 정보는 코퍼스에서 추출한 음절 결합 및 형태소 품사 결합 정보와 어절간 통사적 문맥 정보이다. 다음의 <그림1>은 제안 시스템의 전체적인 구성을 보여준다.

<그림1>에서 보는 바와 같이 시스템 구성은 인식기의 특성이 반영된 후보음절과 혼동음절들을 조합하여 후보 문자열을 생성하는 단계, 음절 결합 제약과 형태소 결합 제약에 따라 후보 어절을 제약하는 단계, 그리고 좌우 어절간 통사적 문맥 결합률을 이용하여 최적 후보 어절을 선정하는 단계로 이루어진다. 최적 후보 어절 선정을 위한 모델은 다음 식 (1)과 같이 표현할 수 있으며, 음절 결합률, 형태소 결합률, 그리고 좌우 어절간 결합률을 결합하였을 때 값이 최대가 되도록 하는 어절을 가장 올바른 어절로 선정한다. 식 (1)에서 $ST(T_{i-1}, T_i, T_{i+1})$은 중심 어절과 좌우 어절간의 결합률로 어절 태그간 상호 결합률이며, $\Pr(W_i)$와 $\Pr(T_i)$은 각각 중심 어절의 음절 결합 확률과 형태소 품사열의 결합률이다.

$$argmax_i \Pr(W_i) + \Pr(T_i) + ST(T_{i-1}, T_i, T_{i+1}) \quad (1)$$



<그림 1> 문자인식 후처리 시스템 구성도

음절 결합 확률은 표층 문자열의 음절간 결합 확률이며 생성된 후보 문자열이 가능한 음절 결합으로만 이루어진 문자열인지 검사하고 모든 음절 결합이 존재하면 올바른 후보 어절로 선정한다. 또한 음절 결합적으로 가능한 후보 어절들일지라도 형태소 결합이 가능하지 않은 어절들이 존재할 수 있다. 그러므로 형태소 분석을 수행하여 2차 후보 어절 제약을 가한다. 이때 미등록어가 형태소 분석에 실패하여 제거될 수도 있으므로 음절 결합 확률이 일정값 이상되는 어절들에 대해서는 미등록어로 간주하고 제거 대상에서 제외한다. 마지막 최종 후보 어절을 선정하는 단계에서는 좌우 어절간의 통사적 문맥 결합률을 함께 적용하는데, 후보 어절들의 어절 태그 사이에 발생가능한 동품사 중의성을 감소시키기 위해 어말 어휘를 부착한 어절 태그를 사용한다. 그리고 형태소 분석이 되지 않은 문자열이 후보 어절로 포함되는 경우를 고려하여 음절 결합률, 형태소 품사 결합률, 어절간 결합률 각각을 더한 값을 적용하도록 한다.

2.1 인식기 특성을 이용한 후보어절 생성

인식 오류는 유사 문자간의 혼동에 의해 주로 발생한다고 앞서 기술한 바 있다. 확률적 방법을 사용하여 인식기의 유사 문자 혼동에 대한 모델을 표현하면 다음 식(2)와 같이 표현할 수 있다. 즉, Noisy-channel 모델에 따라 인식 결과 x가 주어졌을 때 본래 인식하고자 했던 음절이 s일 확률로 나타내는 것이다. 이러한 모델은 다시 Bayes의 법칙을 적용하여 전체 모집단 내에서의 s의 확률과 s가 주어졌을 때 x로 인식할 확률들로 나타낼 수 있다. 이는 인식 결과를 보고 원래 의도했던 문자를 알아내는 확률을 구하는 것 보다는 원래 의도했던 문자가 다른 유사문자로 인식되는 결과들을 통계적으로 계산하여 확률을 구하는 것이 쉽기 때문이다.

$$\Pr(s|x) = \frac{\Pr(x|s)\Pr(s)}{\Pr(x)} \simeq \sum \Pr(x_i|s)\Pr(s) \qquad (2)$$

인식기는 인식기 자체의 후보 음절 신뢰도 계산법에 따라 각 후보 음절들의 인식 신뢰도를 구하고 후보 음절과 신뢰도를 함께 제시한다. 후처리기는 신뢰도 우선순위 3순위 이내의 후보 음절들을 입력으로 받으며 그 가운데 신뢰도가 0이상인 후보 음절들만을 취한다. 또한 인식기에서 출력되는 후보 음절들 가운데 원래 의도했던 음절이 포함되지 않는 경우들도 있을 수 있는데, 이러한 경우에는 인식기의 인식 결과를 학습하여 혼동 음절 테이블을 구성하며 혼동 음절들에 대한 인식 신뢰도는 식 (2)에 따라 구한다.

후보 어절 생성은 인식기로 부터 입력받은 후보 음절과 혼동 음절을 조합하여 생성하며 후보 음절과 혼동 음절의 집합은 예 3)에서 보는 바와 같다. 이 때 인식 신뢰도 값은 사용되지 않으며 후보음절들의 우선 순위만이 후보 어절 생성에서 후보 어절들의 우선 순위를 결정하는데 사용된다. 왜냐하면 혼동 음절들의 확률은 후보 음절들과 동등한 방식으로 구해진 것이 아니기 때문이다.

예 3) 지[0.77] 치[0.57] 기[0.55]
금[0.27] 굼[0.20] 곰[0.20]
까[099]
기[0.60] 치[0.58] 지[0.58] 계[0.50]
의[0.93]
씬[0.99] 인[0.66] 민[0.66]
식[1.00] 릭[0.66] 실[0.67]
기[0.97] 미[0.68] 계[0.50]
술[0.83] 줄[0.63] 출[0.59]
이[0.98]
아[0.99]
직[1.00] 픽[0.67] 질[0.67]
도[0.99] 모[0.54]
실[1.00] 틸[0.67] 힐[0.67]
용[0.99] 욕[0.68] 응[0.67]
화[0.97]

2.2 음절 결합 제약

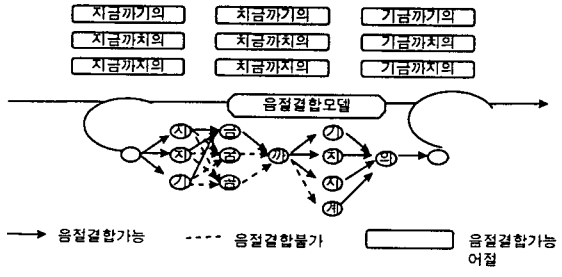
하나 이상의 후보 음절들과 혼동 음절들의 조합으로 후보 문자열들을 생성하고 그 중 올바른 어절 하나를 선택하는 문자인식 후처리를 고려할 때, 생성된 후보 문자열 중에는 어절로서 인정될 수 있는 음절열도 존재하지만 어절로서 인정될 수 없는 음절열들도 존재한다. 문자인식 시스템의 입력은 실세계에서 사용하는 어절들로 구성되었다고 보는 것이 보편적이므로 어절로서 가능하지 않은 음절열들은 배제하고 가능한 어절 중에서도 어절 생성률이 높은 음절열을 올바른 어절로 선정하는 것이 바람직하다.

언어의 음운적 특징은 언어 성립 과정시 형식적인 면에서 중요하게 작용한 요소로, 한국어의 어절 구조내에서의 음운적 특성은 음절열들의 결합 형태로 표출된다. 즉, 어절 형성에서의 음운적 특징은 모음조화 현상이나 형태소 결합 과정에서 용언의 규칙, 불규칙 활용 등이 표층적 어휘 형태로 나타난다. 그러므로 하나의 음절열이 주어졌을 때 음절열이 어절로서 가능한가를 판단하는 기준으로 연속적인 음절 결합 가능성을 살펴보는 것은 의미가 있다.

본 연구에서 사용한 음절열 W_i가 어절로서 가능한지 판단하는 기준은 음절 결합들의 전이 확률값이며 다음 식(3)과 같이 표현된다.

$$\begin{aligned} \Pr(W_i) &= \Pr(c_1, \ldots, c_n) \\ &\simeq \Pr(c_1|\#) \prod_{j=1,n-1} \Pr(c_{j+1}|c_j) \Pr(\#|c_n) \end{aligned} \quad (3)$$

식(3)에서 $\Pr(c_1|\#)$은 음절 c_1이 어절의 첫 음절로 사용될 확률을 말하고, $\Pr(\#|c_n)$은 음절 c_n이 어절의 마지막 음절로 사용될 확률을 말한다. 이는 모든 음절들이 어절의 첫 부분이나 마지막 부분에 나타날 수 있는 것이 아니라 한국어의 음운적 제약 특성을 따르게 된다는 것을 반영한다. 어절 내에서의 음절들의 결합은 임의의 음절 c_i, c_j에 대해 $\Pr(c_j|c_i)$의 조건 확률로 나타내고, 조건 확률은 c_i의빈도와 c_i, c_j의 동시 발생 빈도에 의해 구한다.



<그림 2> 음절 결합 제약

<그림 2>를 살펴 보면 음절들의 결합으로 만들어진 후보 문자열들의 종류는 36 가지이나 실제 음절 결합이 가능한 후보 어절은 모두 9가지로만 제약됨을 볼 수 있다.

2.3 형태소 결합 제약

음절 결합 제약이 어절의 표층 형태에 대한 제약이었다면 형태소 결합 제약은 어절구조 내부의 통사론적 제약으로 형태소가 연결될 때 특정 품사끼리만 연결될 수 있는 것을 의미한다. 형태소 결합 제약은 후보 어절에 대해 형태소 분석을 수행하여 분석이 가능한 어절들이 올바른 후보 어절로 포함되도록 하고 분석을 통과한 어절에 대해서는 형태소 품사열의 발생 확률을 구한다.

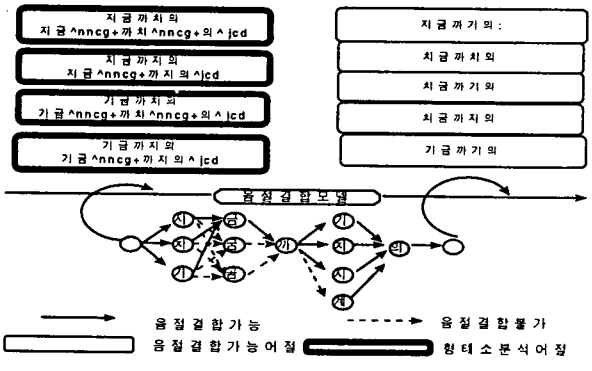
본 연구에서 사용한 형태소 분석기는 체언(명사, 대명사, 수사), 용언(동사, 형용사), 수식언(관형사, 부사), 독립언(감탄사), 관계언(조사)등을 중심으로 총 51개의 품사로 구성되었고, 품사 Bigram 확률 모델을 이용하여 품사열의 발생 확률을 구한다. 확률 모델에서 i번째 형태소 품사열의 발생 확률을 구하는 방법은 다음 식(4)와 같이 표현된다.

$$\Pr(t_{i1}, \ldots, t_{im}) = \prod_{j=1}^{m} \Pr(t_{ij}|t_{ij-1}) = \prod_{j=1}^{m} \frac{f(t_{ij-1}, t_{ij})}{f(t_{ij-1})} \quad (4)$$

식 (4)에 의하면 품사열의 길이가 길수록 발생 확률이 작아지게 되므로 확률값이 아주 작아 0으로 수렴하는 것을 막기 위하여 로그를 취하는 방법을 사용한다. 결과적으로 식 (4)를 변형하여 식 (5)를 사용한다.

$$\begin{aligned} \Pr(T_i) = \log \Pr(t_{i1}, \ldots, t_{im}) &= \sum_{j=1}^{m} \Pr(t_{ij}|t_{ij-1}) \\ &= \sum_{j=1}^{m} \frac{f(t_{ij-1}, t_{ij})}{f(t_{ij-1})} \end{aligned} \quad (5)$$

후보 어절들에 대한 형태소 분석 결과는 <그림 3>에서의 예와 같다.



<그림3> 형태소 결합제약

2.4 통사적 어절 문맥 및 최종 후보 어절 선정

음절 결합이나 형태소 결합 제약을 통과한 올바른 후보 어절이 하나 이상 존재하는 경우, 그 가운데에서 최적 후보 어절 하나를 선정하여야 한다. 최적 후보 어절 선정을 위한 문맥 구조로써 어절간의 의미적 관계까지 고려한 구문 구조를 보는 것이 이상적인 방법이지만, 어절간의 의미적 연관 관계 정보를 자동으로 구축하는 것이 어렵고, 또한 문장 전체에 대해 구문 분석을 수행할 경우 처리적 부담이 가중된다는 문제가 있다.

이러한 점을 고려하여 본 연구에서는 형태론적 속성을 고려한 통사적 구조의 부분 어절 문맥 정보를 사용한다. 형태론적 속성이라 함은 통사적 문맥 구조 속에 잠재해 있는 형태 정보로 통사적 기능을 갖는 조사나 어미의 속성, 즉 어말 어휘의 속성을 말한다. 통사적 기능의 형식 형태소들은 의미상 어절내에서의 최소 자립 형태로서 뿐만 아니라 인접 어절의 의미 내용에 대한 판단 형식을 결정함과 같은 문법적 기능을 갖기도 한다. 한 문장에 대한 '-는, -ㄴ, -를, -었다.' 와 같은 통사적 기능 어말 어휘들은 선행 어간의 음운론적 조건에 따라 이형태가 결정된 것으로, 어절의 표면형인 음절열과 밀접한 관계가 있다. 그러므로 후보 음절과 혼동 음절들의 조합에 의해 상이한 표층 형태를 보이는 후보 어절들 사이에서의 올바른 통사적 문맥 구조 결정시 형식 형태소와 같은 어말 어휘 정보는 중요한 의미를 갖는다.

이에 통사적 문맥 구조에서의 어절 정보는 형태소 품사열에 통사적 기능의 어말 어휘를 결합한 어절 태그를 사용한다 (예 4). 형태소 품사열은 형태소 분석 결과인 품사열을 단순화시킨 것이며, <표 1>의 품사 집합을 따른다. 품사 집합은 동일 단어 내에서의 이품사 중의성을 감소시키는 방향에서 축약된 27개의 단순 품사들로 구성된다. 단순화된 품사의 주 대상은 어절에서 주로 실질 형태소를 이루는 체언, 용언, 그리고 수식언 부분이며 그외 품사는 어절태그의 분별력을 높이기 위하여 세분화한다. 또한 어절을 구성하는 실질 형태소들의 결합 형태에 있어서도 파생 접사(접두사, 접미사)에 의한 파생어는 전성된 하나의 품사로써 취급하며 복합어의 경우도 동일 품사가 복합된 경우에는 하나의 품사로써 축약한다.

예 4) 지금까지의 인식 기술이 아직도 실용화
NN^JCD^까지의 NN NN^JC^이 MA NN

통사적 문맥 구조는 전체 문장 구조를 보기보다는 좌우 어절 태그들의 결합 형태에 따른 부분적 통사 문맥 구조를 살핀다. 즉, 현재 어절을 중심으로 좌우 어절 문맥의 결합 강도를 고려하는 것이다. 통사적 구조의 좌우 어절 문맥 결합 강도는 어절 태그 사이의 상호 결합률을 사용하여 구하며 다음 식(6)에 따른다.

$$ST(T_{i-1}, T_i, T_{i+1}) \simeq MI(T_{i-1}, T_i) + MI(T_{i+1}, T_i) \quad (6)$$

<표 1 > 어절태그에 사용된 품사집합

품사명	품사기호	품사명	품사기호
명사	NN	격조사 (주격,보격,목적격)	JC
대명사	NP	관형격조사	JCD
수사	NU	부사격조사	JCA
동사	VV	접속격조사	JN
보조동사	VVX	서술격조사	JCP
형용사	VJ	보조사	JX
보조형용사	VJX	호격조사	JCV
부사	MA	종결형어미	EFF
관형사	MD	연결형어미	EFC
감탄사	IE	명사형어미	EFN
명사형접미사	XSN	관형형어미	EFD
관형사형접미사	MD	선어말어미	EP
부사형접미사	MA	기호	SS
용언화접미사	XSV		

$$MI(T_i, T_{i-1}) = \log \frac{f(T_i, T_{i-1})N}{f(T_i)f(T_{i-1})} \quad (7)$$

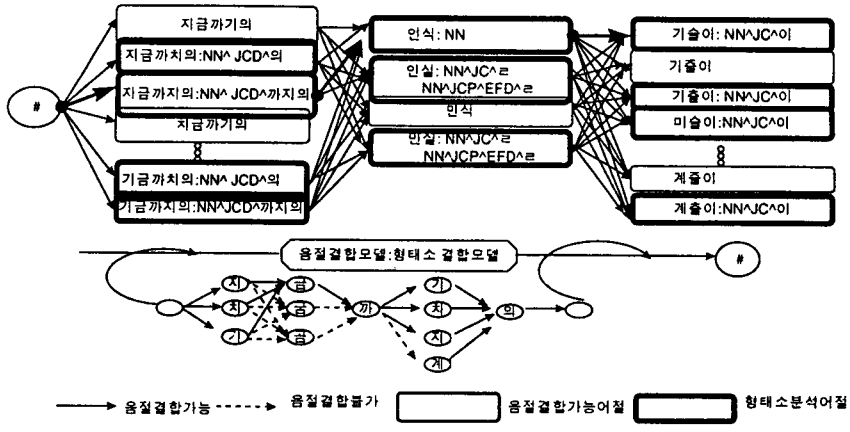
$$MI(\#, T_1) = \log \frac{f(\#, T_1)N}{f(\#)f(T_1)} \quad (8)$$

$$MI(T_n, \#) = \log \frac{f(T_n, \#)N}{f(T_n)f(\#)} \quad (9)$$

식 (6)에서 어절간 연결 문맥의 최소 단위는 어절 태그들의 Bigram이며 어절 W_{i-1}과 어절 W_i간의 상호 결합률은 각각의 어절 태그 T_{i-1}, T_i에 대해 발생 빈도와 동시 발생 빈도를 구하고 식 (7)의 정의에 따라 구한다. 그리고 문장의 시작 부분에 올 수 있는 어절과 문장의 마지막 부분에 올 수 있는 어절들에 대한 제약이 있으므로, 식 (8)과 식 (9)를 사용하여 문장의 처음과 마지막에 나타나는 어절 태그들에 대해서도 각각 별도의 값을 계산한다. 여기서 "#"은 문장의 시작과 마지막을 의미하는 기호이며 f(#)은 전체 문장수를 의미한다.

상호 결합률을 사용하는 목적은 어절 태그간의 의존적 결합 정도를 좀더 정확하게 표현하기 위한 것으로 현재 어절이 이전 어절과 강한 결합 관계를 갖는 경우, 이전 어절의 빈도에만 조건하는 전이 확률을 사용했을 때 상호간의 결합 정도를 제대로 나타내지 못하는 점을 보완하고자 함이다.

최종적으로 올바른 후보 어절 하나를 선정하는 과정은 앞서 언급한 식 (1)에 따르며, 음절 결합률, 형태소 결합률, 그리고 좌우 어절간 문맥 결합률을 결합하였을 때 값이 최대가 되도록 하는 어절을 가장 올바른 어절로 선정한다. <그림 4>를 예로 하여 최종 어절을 선택하는 과정을 살펴보면 문장의 처음은 #에서부터 시작하고 문장의 첫 어절로 나올 가능성이 크면서 다음 후보 어절들과 가장 결합률이 높은 어절은 'NN^JCD^까지의' 라는 어절 태그를 갖는 것이다. 그리고 후보 어절들의 음절 결합률 및 형태소 결합률을 어절간 결합률과 결합하여 더하면 '지금까지의'라는 후보 어절이 가장 높은 것으로 나타나 첫 번째 어절 위치에 올 수 있는 가장



< 그림4> 최종어절 선택

올바른 후보 어절로 선정된다.

III. 실험 및 평가

3.1 실험방법

제안된 시스템의 성능평가를 위해 우선순위 3위 이내의 후보 음절들을 인식 결과로 내 보내는 실제 인식기의 결과를 사용하였다. 인식기의 특성 정보로는 인식결과 3순위 이내에 들어오지 않아 인식 오류를 발생시킨 음절들을 <표 2>와 같이 혼동 음절 테이블로 구성하여 사용하였다. 혼동 음절 테이블은 원문과 인식 결과를 비교하는 반자동 학습 방법에 의해 학습하였으며 53쌍의 혼동 음절쌍이 학습되었다.

<표 2> 혼동 음절 테이블

	원음절(1)	원음절(2)
게	그(0.13)	계(0.23)
겨	그(0.25)	기(0.43)
기	계(0.23)	
꼿	못(0.11)	
꾼	문(0.18)	

후보 어절 제약 및 선택을 위해 사용된 통계적 언어 특성 정보인 음절 결합 정보와 어절 태그간 결합 정보는 각각 1000만 원시 코퍼스와 17만의 품사 태깅된 코퍼스에서 추출하였다. 음절 결합 정보와 어절 태그간 결합 정보에 대한 내용은 <표 3>, <표 4>와 같다. 어절 태그간의 결합 정보는 어말 어휘가 어절간 문맥 결합에 기여하는 정도를 실험하기 위하여 어말 어휘를 부착한 경우와 부착하지 않은 경우를 모두 구하였다.

<표3> 음절 결합 정보

고유어절	고유음절	첫음절	마지막음절	음절결합
1,073257	2,382	2,095	1,762	155,310

<표 4> 어절태그 결합정보

	문장	고유 어절태그	첫 어절태그	마지막 어절태그	어절태그 Bigram
A	15,848	2,486	462	615	22,396
B	15,848	497	165	153	5,136

A: 어말어휘를 부착한 경우
B: 어말어휘를 부착하지 않은 경우

형태소 분석 및 형태소 품사열 발생 확률을 구하기 위해 17만 품사 태깅된 코퍼스로부터 품사 Bigram을 추출하고 그들에 대한 확률을 구하여 사용하였다..

3.2 실험결과 및 평가

후처리 시스템의 성능은 인식기에서 출력한 제 1 순위 후보 문자가 틀린 경우 이를 바로 교정한 정도와 후처리에 의해 제1순위 인식률이 향상된 정도, 올바르게 인식된 제 1 순위 문자를 틀리게 교정하는 정도로 평가할 수 있다. 또한 어절 단위에서 제 1 순위 인식 결과 어절 중 잘못 되었던 어절을 올바로 교정하는 정도와 올바른 어절을 틀리게 교정하는 정도로 평가할 수도 있다.

교정률과 오교정률은 각각 아래 식 (10), 식 (11)과 같이 정의하여 평가하며, 전체 교정률은 식 (12)과 같이 정의하여 평가하였다.

$$\text{교정률} = \frac{a}{c} \times 100(\%) \qquad (10)$$

$$\text{오교정률} = \frac{b}{d} \times 100(\%) \quad (11)$$

$$\text{전체교정률} = \frac{a-b}{c} \times 100(\%) \quad (12)$$

a: 올바로 교정한 갯수 b: 틀리게 교정한 갯수
c: 오인식된 갯수 d: 올바로 인식된 갯수

957 개의 어절로 구성되고, 그 중 33개의 어절이 미등록어를 포함하고 있는 일반 문서에 대해 실험한 결과는 다음 <표 5>, <표 6>과 같다. <표 5>의 결과를 살펴보면 어절간 통사적 문맥 관계에 기반하여 제안한 후처리 방법이 문자 인식의 성능을 문자 단위 40%이상, 어절 단위 65.3%이상 향상 시키고 있음을 알 수 있다. 또한 어절간 통사적 문맥 관계 정보에 통사적 기능의 어말 어휘를 부가하여 사용한 경우, 문자 단위 48.1%(+8.1%), 어절 단위 72.0%(+6.7%)의 성능 향상을 보여, 어말 어휘를 사용한 것이 그렇지 않은 것보다 더 우수한 후처리 성능을 나타내고 있음을 보여 주고 있다.

<표5> 문자인식 후처리 결과 (%)

		후처리전 인식률	후처리후 인식률	교정률	오교정률	전체 교정률
A	문자	94.1	97.4	56.0	0.60	48.1
	어절	87.6	96.6	83.9	1.66	72.0
B	문자	94.1	97.1	51.0	0.66	40.0
	어절	87.6	95.7	78.8	1.90	65.3

<표 6> 비단어 오류와 비문맥 오류 처리 결과(%)

	비단어 오류	비문맥 오류	교정률		전체교정률	
			비단어	비문맥	비단어	비문맥
A	79.7	20.3	81.9	91.6	78.7	45.8
B			78.7	79.1	75.5	25.0

A: 어말어휘를 고려한 통사적 문맥 정보 사용
B: 어말어휘를 고려하지 않은 통사적 문맥 정보 사용

인식 오류 유형을 비단어 오류와 비문맥 오류로 나누었을 때, <표 6>은 통사적 기능의 어말 어휘를 고려한 문맥적 후처리 방법 A가 어말 어휘를 고려하지 않은 문맥적 후처리 방법 B보다 비단어 오류 교정 뿐만 아니라 비문맥 오류 교정에 있어서 후처리 성능이 더 우수함을 보여주고 있다. 즉 A는 전체 오류의 79.7%, 20.3%를 차지하는 비단어 오류와 비문맥 오류 각각에 대해 81.9%(+3.2%)와 91.6%(+12.5%)의 교정률, 78.7%와 45.8%의 전체 교정률을 보여 줌으로써 B에 비해 오교정률을 적게 하면서 비문맥적 오류를 효과적으로 다루고 있음을 보여준다.

음절 결합 정보와 형태소 분석을 함께 사용하여 미등록어를 처리한 내용을 분석한 결과는 전체 어절의 3.4%를 차지한 미등록어에 대해 음절 결합 결과 97%가 통과하고 통과된 어절중 올바로 선택된 미등록어는 90%를 차지한다. 이로써 음절 결합 정보 사용으로 전체 미등록어의 87.5%를 올바로 인식하여 미등록어 문제를 어느 정도 해결할 수 있음을 보여 주었다.

또한 인식결과 후보음절과 혼동음절들을 결합하여 생성한 평균 28.1개의 후보어절들에 대해, 음절결합 제약을 가하여 어절 평균 14개를 제거하고 형태소 분석대상 후보어절수를 50%정도 감소시키는 효과를 얻었다. 후보 어절 당 평균 2.4개의 형태소 분석 후보를 내보내는 형태소 분석기를 생각할 때 좌우 어절간 문맥결합을 처리하는데 소요되는 처리 시간적 부담은 그리 크지 않는 것으로 나타났다.

IV. 결론 및 향후 연구

본 논문에서는 한글 문자인식 후처리 분야에서 아직까지 잘 해결되지 못하고 있는 문제인 미등록어와 비문맥 오류를 고려한 문자 인식 후처리 시스템을 제안하였다. 제안된 시스템은 인식기의 특성이 반영된 후보 음절과 혼동음절을 사용하여 후보 어절을 생성하고, 단어로서 가능한지를 결정하는 기준으로 확률적 음절 결합 정보와 형태소 품사 결합 확률을 사용하여 형태소 분석 기법만을 사용했을 때 발생할 수 있는 미등록어 문제를 해결하고자 하였다. 그 결과 형태소 분석만을 수행 했으면 배제될 수도 있었던 미등록어를 배제 대상에서 제외하여 처리할 수 있었다. 그리고 최적의 올바른 후보 선정 과정에서 한국어의 음절 특성, 형태론적 특성, 통사적 특성 정보를 종합적으로 사용함으로써 미등록어와 비문맥적 오류를 효과적으로 처리하였다.

실제 인식 결과를 가지고 실험한 결과, 제안된 시스템이 인식기의 성능을 문자 단위 94.1%에서 97.4%로, 어절 단위 87.6%에서 96.6%로 향상시켰다. 그리고 전체 실험 어절의 3.4%인 미등록어에 대해서 87.5%를 올바로 인식하고, 전체 오류의 20.3%를 차지했던 비문맥 오류를 91.6% 교정할 수 있었다. 또한 인식 성능 향상에 기여한 정보는 단독 정보에 의한 성능 향상보다는 단어 형성에 관여하는 음운적, 형태론적, 통사적 정보가 종합적으로 작용함에 의해 이루어진다는 점을 보여 주었다.

그러나 이러한 성능 향상에도 불구하고 후처리에서 실패하는 경우들이 발생하였는데, 이는 영숫자 혼용 문서 인식시 인식기의 영상 분할 오류에서 파생된 띄어쓰기 오류와 이상 문자 출력으로 혼동 음절 조합 실패에 따른 것들이었다. 그러므로 이러한 문제를 극복하기 위한 향후 연구로써 문자인식 후처리에 적합한 띄어쓰기 교정 알고리즘을 고안하고, 영숫자들의 혼용에 따른 효과적인 혼동 문자 학습 알고리즘을 개발하는 것이 필요하다 하겠다.

참고문헌

[김민정97] 김민정, 권혁철, "언어적, 경험적 제약을 이용한 한국어 문자 인식 후처리 기법", 정보과학회 논문지, 제24권, 제1호, 1997년 1월, pp. 25-31

[김윤호92] 김윤호, 이종국, 김항준, 이상조, "형태소 분석을 이용한 문자인식 에러의 검출", 한글 및 한국어 정보처리 학술발표 논문집, 1992년, pp. 555-566

[민병우91] 민병우, 이성환, "문자인식을 위한 오인식 수정기술", 정보과학회 논문지, 제9권,제1호, 1991년 2월, pp.7-13

[박진우94]박진우, 이일병, "통계적 방법에 의한 후처리", 한글 및 한국어 정보처리 학술발표 논문집, 19945년, pp. 518-526

[시정곤94] 시정곤, "국어의 단어 형성 원리", 국학자료원, 1994

[유진희95] 유진희, 이종혁, 이근배, "형태소분석과 언어평가를 이용한 문지인식후처리", 한국정보과학회 논문지 제22권 제6호, 1995년 6월, pp. 880-890

[이종연93] 이종연, 오상헌, "N-GRAM 한글 사전을 이용한 오인식 단어의 교정 알고리즘", 한글 및 한국어 정보처리 학술발표 논문집, 1993년, pp. 271-283

[홍남희93] 홍남희, 이원일, 이종혁, 이근배, "어절정보와 문자열정보를 이용한 문자인식에서의 오인식 수정기법에 관한 연구", 제1회 문자인식 워크샵 발표 논문집, 1993년, pp. 109-113

[황호정94] 황호정, 도정인, 권혁철, "한글문자인식을 위한 후처리기의 개발과 속도개선", 제2회 문자인식 워크샵 발표논문집, 1994 년, pp. 180-189

[Hisamitsu95] Toru Hisamitsu, Katsumi Marukawa, Yoshihiro Shima, "Optimal Techniqueues in OCR Error Correction for Japanese Texts", Proceedings of The 3rd International Conference on Document Analysis and Recognition, Aug. 1995, pp. 1014-1017

[Kukich92] Karen Kukich, "Techniques for Automatically Correcting Words in Text", ACM Computing Survey. Vol. 24, No. 4, 1992, pp.377-439

[Mays91] Eric Mays, Fred J. Damerau, Robert L. Mercer, "Context Based Spelling Correction", Information Processing Management, Vol. 27, No.5, pp. 517-522

[Tong96] Xiang Tong, David A. Evans, "A Statistical Approach to Automatic OCR Error Correction in Context", Proceedings of the 4th Workshop on Very Large Corpora, Aug. 1996, pp. 88-100