

한국어 오류 교정 시스템의 구현

최재혁[†] · 김권양^{††}

요 약

기존 워드프로세서의 맞춤법 검사기는 여러 오류 단어 후보군 중에서 1개를 선택하는 오류 작업의 불편함과 60%대의 낮은 교정률 그리고 늦은 처리 속도 등의 단점을 갖고 있다. 본 연구에서는 이러한 단점들을 해결하기 위하여 1개의 교정 단어와 페이지 단위의 일괄 교정으로 교정의 불편함을 해소하고, 높은 오류 교정률과 빠른 처리 속도를 가능하게 하는 방안을 제시한다. 이를 위하여 형태소 분석 시 처리 속도를 향상시키기 위한 방법으로 양방향 최장일치법을 적용하며, 교정 시의 정확성보장과 처리 속도의 향상을 위한 사전과 여러 알고리즘(복합명사 분리, 보조용언 분리, 오타 교정 등) 등을 개발하여 적용하였다. 특히 한국어에서 모호성이 많이 발생하는 의존명사 및 접미사와 조사/어미의 구분 방안, “로써/로서, 되다” 등의 구분 처리 방안을 제시하여 교정 시스템의 신뢰성을 높였다.

Implementation of Korean Error Correction System

Jae-hyuk Choi[†] · Kweon-yang Kim^{††}

ABSTRACT

Korean error detectors of word processors have defects such as inconvenience that users choose one of error groups, lower detecting rate of 60%, and slow processing time. In this study, I proposed a resolution method of these defects. For these, I applied bidirectional longest match strategy for morphological analysis to improve processing time. I suggested dictionaries and several algorithms such as separation of compound noun and assistant declinable words, correction of typing error to improve processing time and to guarantee correction accuracy. I also suggested a distinguishable method for dependent noun/suffix and Josa/Eomi where many ambiguities are generated, and a distinguishable method for Korean “로써/로서, 되다” to improve the reliability of the correction system.

1. 서 론

우리는 현재 오류가 난무하는 문서 속에서 살고 있다. 상당히 많은 사람들이 어떤 단어가 표

준어인지 혼동하기 일쑤이고, 학창 시절에 배웠던 추억 속의 맞춤법을 고집하는 경우가 지금도 종종 있다. 언어의 속성을 보면 자주 사용하는 말이 결국 표현으로 바뀐다는 사실이 그렇듯이, 한글 맞춤법은 매번 수정되었지만, 특별히 관심을 갖지 않는 한 정확히 알고 쓰기는 거의 불가능하다. 또한 많은 분량의 글을 쓰다 보면 눈에 보이지 않는 오류가 많이 발생한다.

[†] 중신회원: 신라대학교 컴퓨터교육과 교수
^{††} 정회원: 경일대학교 컴퓨터공학과 교수
 논문접수: 2000년 9월 1일, 심사완료: 2000년 11월 10일

초창기의 워드프로세서는 문서 작성의 기능만이 요구되었고, 최근에 들어 개발된 워드프로세서에는 맞춤법 검사 기능이 추가되어 있지만, 이는 컴퓨터가 오류 단어만을 탐색하고 여러 오류 대치 후보 군을 제시하는 수준으로 사용자가 모든 오류 단어에 대해 일일이 수정을 행해야 하는 번거로움을 초래할 뿐만 아니라, 이러한 수정작업에 많은 시간을 낭비하게 된다.

따라서, 정확하고 빠르면서도 오류를 자동으로 교정해 주는 오류 교정 시스템의 개발이 절실히 요구되고 있지만, 현재의 기술로는 이러한 모든 오류에 대한 100%의 교정은 거의 불가능하다. 이는 한국어 자체의 처리도 어렵거니와 글쓰는 사람의 의도에 따라 오류가 정당일 수도 있기 때문이다.

기존 워드프로세서 맞춤법 검사기는 오류 단어 하나하나에 대해 여러 개의 교정 단어 후보군에서 1개를 선택하도록 하는데 이는 사용자에게 처리의 불편함과 많은 처리 시간을 요구하게 한다. 특히 교정률은 60%대에 머물고 있으며, 오류단어를 오류가 아닌 단어로 인식되어 오류 교정을 행하지 않는 단어가 전체 오류단어의 약 20% 가까이 되고, 워드프로세서마다 복합명사 및 보조용언의 분리 방법에 일정한 규칙이 없어 교정의 신뢰성에 의문을 가지게 되어 실제적으로 워드프로세서의 사용자가 많이 이용하지 않고 있다[1].

본 연구에서는 기존의 맞춤법 검사기의 단점인 오류 수정 작업의 불편함을 각각의 오류 어절에 대해 1개의 교정 단어를 페이지 단위로 제시함으로써 어느 정도 해소시키면서, 오류 수정을 위한 처리 시간의 감소 및 높은 오류 교정의 정확률을 보장하는 것을 주목표로 하였다. 본 연구에서는 자연어 처리 시 처리 속도에 결정적으로 영향을 미치는 사전 참조 횟수의 감소를 피하여 시스템의 처리 속도를 향상시키고자 한다. 이를 위하여 한국어 형태소 분석시의 가장 사전 참조 횟수가 적은 양방향 최장일치법을 적용하며[2], 교정을 위해서도 정확성을 보장하면서 사전 참조 횟수를 감소시키는 여러 알고리즘(복합명사 분리, 보조용언 분리, 오타교정 등)을 개발하여 적용하였다[3,4]. 또한 교정의 정확률을 높이기 위하여 모호성이 발생하는 의존명사, 접미사와 조사어미

와의 구분 방안, 일반적으로 한국 사람이 많이 틀리는 단어의 교정, 특히 “로써/로서, 데, 뿐, 만” 등의 구분 처리 방안을 제시하여 교정 시스템의 신뢰성을 높였다[5].

기존의 한국어 오류 단어 교정에 관한 연구는 80년대와 90년대 초의 집중적 연구 결과로 철자 검사/교정기는 현재 상용화되어 있지만, 이는 한 어절 내의 오류에 대한 처리가 대부분이다. 부산 대학교에서 개발한 시스템은 연어정보를 이용하여 여러 어절 간의 오류 검증 및 교정이 가능하도록 실험실 수준에서 구현하였다[6]. 또한 고려 대학교에서 개발한 코퍼스용 철자 및 띄어쓰기 오류 교정 시스템은 형태소 분석기를 이용한 교정 시스템의 단점을 기술하고 코퍼스의 특성을 효과적으로 이용하였으나, 이 시스템은 방대한 양의 코퍼스를 이용하기 때문에 이를 검색하고 처리하는데 걸리는 시간에 대한 오버헤드를 고려하지 않았고 코퍼스에 존재하지 않은 단어는 결국 미처리 되므로 실용화되지 않았다[7]. 따라서 1개의 교정 결과만을 제시해 주는 교정 시스템은 아직까지 개발되지 않았다. 본 연구는 한국 최초의 실용화가 가능한 오류 교정 시스템이라는 데 큰 의의를 가진다.

2. 양방향 최장일치법을 이용한 오류 교정 시스템

일반적으로 한국어 단어에서 오류라 함은 비표준어 오류, 띄어쓰기/붙여쓰기 오류, 조사/어미 오류, 외래어 표기 오류, 철자 오류 등을 말하는데, 이를 맞춤법 교정 시스템에서는 교정 기준에 따라 의미 오류, 맞춤법 오류, 기타 오류와 같이 크게 3가지로 분류할 수 있으며, 이는 다음과 같다.

① 의미 오류 : 정확한 의미를 파악해야만 교정할 수 있는 오류. 문장내의 일부분이 집필자의 실수나 무지에 의해 잘못 쓰여졌지만, 집필자가 아닌 타인이 봐서는 무엇이 잘못되었는지, 또는 어떤 부분을 고쳐야 본래의 뜻에 맞는 문장이 될지 등을 알 수 없는 오류. 심지어 사실과 다른 오류(예를 들면, 이순신 장군이 상해 임시정부와

관련돼 독립투사로 묘사하는 것 같은)도 이 유형에 포함된다

② 맞춤법 오류 : 맞춤법을 구성하는 한글 맞춤법, 표준어 규정, 외래어 표기법, 띄어쓰기 등에 위반되는 오류

③ 기타 오류 : 체언이나 용언 뒤에 부적절한 받침(있을 수도 있고 없을 수도 있는)또는 활용 패턴에 맞지 않는 조사나 어미를 사용한 오류 문장 부호의 부적절한 사용, 한자 표기오류 등이 있을 수 있으며, 조금은 광범위한 표현이지만 일부 자소를 잘못 사용한 철자 오류 등도 이에 포함된다.

이 중 의미 오류는 그 문장의 본 뜻과 작성자의 의도를 충분히 알아야 교정할 수 있으므로 교정 시스템에서 이러한 의미 오류에 대한 교정은 거의 불가능하다. 다만, 간단한 형태의 숙어 개념이 의미 오류나 단순한 구문 분석 정도로 교정이 가능한 의미 오류에 대해 본 연구에서는 이러한 형태에 속하는 말뭉치를 조사하여 처리할 예정이며, 본 연구에서 처리해야 할 주된 오류 대상이 맞춤법 오류와 기타 오류이다.

일반적으로는 ②와 ③을 총칭해서 맞춤법 오류라 할 수 있지만 실제의 오류발견과 교정 기준이 다르고, ③의 오류 유형은 한글 맞춤법에서는 다루지 않으므로 이를 분리하였다.

2.1 띄어쓰기 교정 시스템

본 연구에서는 한국어 실 문서상에서 나타나는 맞춤법 오류 중에서 80% 이상을 차지하는 띄어쓰기/붙여쓰기 오류에 대한 교정 시스템을 먼저 개발하였고, 이 시스템을 기초로 하여 나머지 오류에 대한 교정 시스템을 개발하였다.

띄어쓰기 오류는 가장 빈번히 나타나는 오류 유형이다. 특히 신문에서는 단순히 지면을 절약한다는 관점에서 띄어쓰기 규정을 무시하지만, 신문에 익숙해진 독자는 자기가 사용한 문서에도 이를 그대로 받아들여 사용하고 있다. 띄어쓰기는 국어에 대한 문법 사항, 즉 품사에 대한 이해의 표현이므로 이러한 띄어쓰기 오류는 우리 말, 우리 글을 올바르게 사용해야만 하는 관점에서 볼 때, 대단히 심각한 문제로 받아들여야 할 것

이다. 이러한 띄어쓰기 오류를 검사하는 방법은 개발자가 국문법이나 맞춤법 규칙을 잘 알고 있는 상태에서 정리한 어절 규칙과 사전 정보를 이용하는 방법이 주로 사용되고 있다.

일반적으로 띄어쓰기에 실패한 단어는 사전에 존재하지 않으므로 이를 분리 또는 합성하기 위하여 상당히 많은 사전 참조를 행해야 한다.

본 연구는 정확한 띄어쓰기 교정을 하면서도 이러한 사전참조 횟수를 줄여 전체적인 처리 속도를 향상시키는 것을 가장 큰 목표로 두고 띄어쓰기 교정 시스템을 개발하였으며, 이를 위해 아래의 연구 결과를 적용한다. <표 1>은 띄어쓰기 교정 시스템에서 교정 가능한 띄어쓰기 오류 세부 유형을 제시한다. <표 1>에서의 붙뺄오류는 붙여써야 할 것을 띄어쓴 오류이며, 띄붙오류는 띄어써야 할 것을 붙여쓴 오류이다.

1) 형태소 분석시의 사전 참조 횟수 감소

띄어쓰기 교정을 위해서는 문서의 모든 어절에 대해 한국어 형태소 분석을 반드시 행해야 하므로, 본 연구는 기존의 한국어 형태소 분석 기법 중 정확하면서도 가장 사전 참조 횟수가 적은 양방향 최장일치법을 적용하여 1차적으로 어절 단위의 형태소 분석시의 사전 참조 횟수를 줄여 전체 시스템의 처리 속도의 향상을 꾀하였다. 부수적으로 형태소 분석에 실패한 단어에 대해서도 양방향 최장일치법을 적용하였다. 즉, 형태소 분석에 실패한 단어가 띄어쓰기를 하지 않은 단어라면, 이 단어를 둘 이상의 단어로 분리하여 다시 분리된 단어들을 가지고 다시 형태소 분석을 행해야 하는데, 이때 보다 더 효율적인 분리 방법과 양방향 최장일치법을 적용하게 된다. 기존의 형태소 분석 기법 중에서 가장 많이 사용하는 Tabular Method 방법이 어절당 10.6회의 어휘사전 참조를 행하나, 양방향 최장일치법은 어절당 2.1회의 어휘사전 참조만을 행하므로 약 80% 이상의 어휘사전 참조 횟수를 감소시킨다[2].

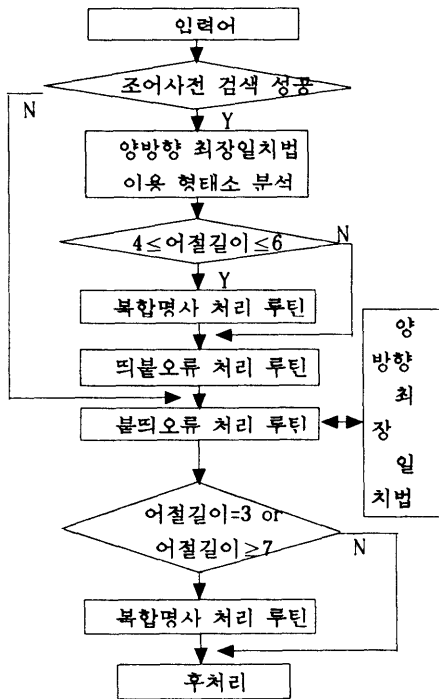
2) 오류 형태 세부류

띄어쓰기 오류는 크게 붙뺄오류와 띄붙오류로 나눌 수 있다. 이러한 붙뺄오류와 띄붙오류의 형태에 대한 말뭉치를 조사하여 오류 어절의 구성

6) 한국어/외국어 처리 방안

한국어와 외국어를 구분하는 루틴을 적용하면 교정의 질을 높일 수 있다. 예를 들어, “낙동강”, “미시시피강”의 경우 강을 중심으로 앞 어절이 한국어이면 “낙동강”으로, 마지막 음절이 ‘강’이고 앞단어가 외국어일 경우 “미시시피 강”으로 교정되어야 한다[8]. 또한 외국어는 사전에 미등록어일 경우가 많은데, 미등록어가 외국어로 판별될 경우 오류 단어 교정 루틴을 실행하지 않아도 되므로 잘못 교정된 단어나 불필요한 사전 참조를 행하지 않아 전체 수행 시간을 감소시킬 수 있다.

(그림 1)은 띄어쓰기 오류 교정 시스템의 개략적인 흐름도이다.



(그림 1) 띄어쓰기 교정 시스템의 개략적인 흐름도

모든 오류 단어를 교정해 주는 시스템을 의미한다. 일반적인 이러한 오류 교정 시스템은 사전에 상당히 의존적이며, 지나친 세밀한 오류 교정은 오히려 잘못된 교정 결과를 초래할 가능성이 있으므로, 오류 교정의 범위에 대한 설계는 시스템의 성능 및 정확도에 결정적으로 영향을 미치게 된다. 따라서 본 연구는 오류 교정의 정확도와 처리 속도의 향상에 주안점을 두면서 다음과 같은 방법으로 연구하였다.

1) 오류교정 사전 구축

오류 단어에 대한 말뭉치를 조사한 후, 이를 그룹별로 세밀히 분류하고 자주 사용되어지는 오류 단어에 대한 오류 교정 사전을 구축하였다. 특히 조사/어미 오류 사전, 선어말어미 오류 사전, 비표준어 오류 사전, 외래어 표기 오류 사전을 따로 구축하여 오류 교정에 대한 효율성을 높였다. 한국어는 80% 이상이 조사나 어미와 결합한 형태의 어절을 구성하고 있으므로 오류가 있는 어절에서 조사나 어미를 분리할 수 있다면, 교정의 효율성을 상당히 향상시킬 수 있다.

본 연구는 위의 오류 사전과 ‘쓰’, 한국어의 조사나 어미에 나타나는 글자의 빈도 수와 한국어 조사나 어미의 첫 글자에만 나타나는 글자 등을 이용하여 오류가 있는 어절에서 조사/어미 분류를 가능하게 하였다.

2) 자주 틀리는 철자 오류 처리 방안

한국어의 조사 ‘로서’와 ‘로써’, ‘음으로’와 ‘으므로’를 구분할 수 있는 방안을 마련하고자 한다. 이는 한국어를 사용하는 대부분의 사람이 잘 모르고 사용하는 것이 대부분으로 오류의 상당한 부분을 차지한다. 이를 구분하기 위하여 이들 단어에 대한 말뭉치를 조사한 후, 체언의 의미 정보를 이용하여 해결하였다. 이 중 ‘음으로’와 ‘으므로’의 구분은 만족할 만한 결과가 나오지 않아, 현재 연구 중에 있으며, 추후 시스템에 보완할 예정이다. 자세한 내용은 3장에서 기술한다.

3) 오타 교정 알고리즘 개발

오타 교정을 위한 테이블 및 사전을 구성한다. 즉, 오타 교정을 위한 확률을 도입한 자소 대치

2.2 오류 단어 교정 시스템

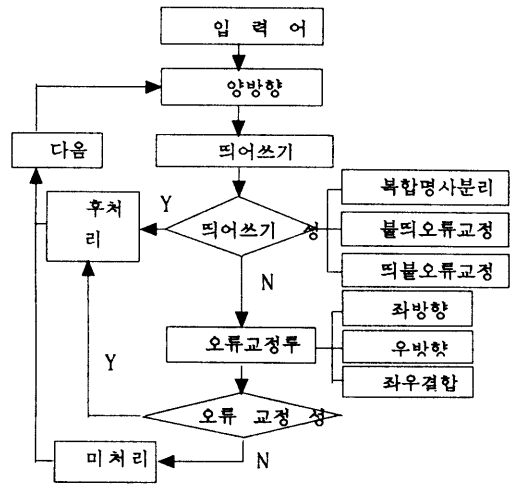
오류 교정 시스템은 띄어쓰기 오류를 포함한

테이블 및 음절 대치 테이블을 구성하여 오타 발생 음절에 대해 음절과 자소를 대치하면서 오타 교정을 행한다. 이때 확률적으로 오타일 가능성이 가장 높은 음절을 추측할 수 있는 방법을 개발하여 시스템에 적용하였다. 또한 자소 대치 테이블 이외에 한국어와 외국어를 구분할 수 있는 한국어 외국어 구분 테이블을 이용하여 외국어로 판명된 미등록 외국어에 대해서는 오류 교정을 행하지 않고 외국어로 표시하여 처리한다.

(그림 2)는 한국어 오류 교정 시스템의 개략적인 흐름도이다.

위의 흐름도에서 입력 어절에 대해 양방향 최장일치법을 이용한 형태소 분석을 먼저 행한 후, 형태소 분석에 실패하면 띄어쓰기 교정 루틴을 먼저 호출한다. 그 이유는 오류중 약 80%가 띄어쓰기 오류이므로, 이를 먼저 처리하는 것이 시스템의 효율성을 높일 수 있다. 띄어쓰기 교정 루틴은 불 띄오류 교정 루틴과 띄불오류 교정 루틴, 복합 명사 분리루틴으로 나누어진다. 그리고 형태소 분석에 성공하더라도 불 띄오류 교정루틴을 실행한다. 그 이유로 예 “어느 해 보다”에서 ‘보다’와 같이 형태소 분석에 성공하더라도 앞 어절 “해”와 붙여서 다시 분석하여 “해보다”로 분석해야 할 가능성이 있기 때문이다. 이 경우는 모호성이 발생하므로 모호성 처리가 필요하다.

띄어쓰기 교정이 실패하면 오류 교정 루틴을 호출하는데 이는 크게 좌방향 교정 루틴, 우방향 교정 루틴, 그리고 좌우 결합 루틴으로 나뉜다. 좌방향 교정 루틴은 어절 내에 받침이 ‘쓰’이 있는지를 먼저 찾은 후, 조사어미의 오류에 대한 처리를 행한다. 이는 ‘쓰’이 존재할 경우 오류 어절 내에서 조사어미 부분을 쉽게 추출할 수 있으므로 시스템의 처리 속도를 향상시킬 수 있다. 좌우 결합 루틴은 한 글자의 초성 중성 종성에서 한 가지만 나타날 경우를 처리하는 루틴이다. 예를 들면, ‘발전의’라는 글자를 ‘발전으|’로 잘못 타이핑 쳤을 때 교정시켜주는 루틴이다. 우방향 교정 루틴은 체언과 용언 부분(실질형태소)에 대한 오류 교정 루틴으로 오류 사전을 이용한 비표준어 오류, 외국어 표기 오류의 교정과 음절 대치 테이블, 자소대치 테이블을 이용한 오타 교정 등을 처리한다.



(그림 2) 한국어 오류 교정 시스템의 개략적인 흐름도

위의 흐름도에 근거하여 오류 단어를 처리하는 과정은 다음과 같다.

예 1) 이땅에 :

- ① 조사 어미 분리 : 이땅+에
- ② 띄불오류 교정 루틴 호출 : “이 땅에”출력

예 2) 태어났었다 :

- ① 띄어쓰기오류 교정 루틴 실패
- ② 좌방향 오류 교정 루틴 호출
‘쓰었다’를 조사어미 부분으로 분리
- ③ 조사어미 오류사전 및 교정루틴을 통해 ‘쓰었다’로 교정한 후, “태어났었다” 출력

예 3) 태어났다 :

- ① 띄어쓰기 교정 루틴 실패
- ② 좌방향 오류 교정 루틴 호출
→ ‘다’를 조사어미 부분으로 분리
- ③ 선어말어미 처리 루틴을 통해 ‘쓰다’로 교정 후 “태어났다” 출력

예 4) 농홍의 :

- ① 띄어쓰기 교정 루틴 실패
- ② 좌방향 오류 교정 루틴 호출
→ ‘의’를 조사어미부분 분리
- ③ 선어말 어미 처리 루틴에서의 교정 실패
- ④ 우방향 오류 교정 루틴 호출
→ 체언 용언 교정 루틴 호출을 통해 음절대치 테이블 이용하여 ‘농’을 ‘중’으로 교정 후, “중홍의” 출력

3. 선택 오류 교정 방안

기존의 교정 시스템에서는 주로 자판 오류나 맞춤법 지식 부족으로 인한 오류에 대한 전체적인 교정을 시도해 왔다. 그러나 사람들이 자주 틀리는 특별한 어휘들에 대한 처리는 미흡한 실정이다. 형태가 비슷하거나 같은 형태가 다른 기능을 함으로써 혼동을 일으키는 어휘들은 다른 오류 어휘들과는 달리 출현할 때마다 틀릴 가능성이 아주 높음에도 불구하고, 이들에 대한 명확한 구분 방법을 제시하기 어렵기 때문이다.

<표 2>는 형태가 비슷하거나 같은 형태가 다른 기능을 함으로써 두 개 중에 하나를 선택하는데 있어서 발생하는 철자 오류나 띄어쓰기 오류의 유형들을 나타낸 것이다. 본 논문에서는 이러한 오류 유형들을 위해 선택 오류라는 명칭을 사용하기로 한다.

<표 2> 선택 오류의 유형과 용례

	유형	용례
철자 오류	의존형태소	'로써/로서', '(음)으로/므로', '던지/든지' 등
	자립형태소	'반드시/반듯이', '지그시/지긋이' 등
띄어쓰기 오류	의존명사/어미의 일부	데, 듯, 바, 지, 걸
	의존명사/접미사	넋, 것, 나름, 체, 채, 적, 이래, 노릇, 무렵, 빨, 쪽, 차, 통, 품
	의존명사/조사	뿐, 만큼, 대로, 만, 족족
	동사/접미사	되다

맞춤법 오류는 크게 철자 오류와 띄어쓰기 오류로 구분할 수 있는데, 본 논문의 대상이 되는 철자 오류는 서로 다른 용법으로 쓰이지만 비슷한 형태를 가짐으로써 생기는 사용자의 선택 오류를 말한다. 철자 오류는 의존형태소 오류와 자립형태소 오류로 나눌 수 있다. 예를 들어, '로써/로서', '(음)으로/므로' 등의 의존형태소는 그 선행

요소가 문장내에서 어떤 의미로 사용되었는가에 따라 구별이 되며, '반드시/반듯이', '지그시/지긋이'와 같은 자립형태소는 그 문장의 다른 성분들과의 구문관계에 따라 하나를 선택해야 하는 것이다. 한국어는 조사나 어미와 같은 기능어가 발달되어 있으므로 의존형태소 오류가 자립형태소 오류보다 더 빈번히 발생하기 때문에 본 논문에서는 철자 오류 중에서 의존형태소 오류의 교정 방법을 제시하고자 한다.

띄어쓰기 오류는 가장 빈번하게 나타나는 오류의 형태로서[3] 띄어 써야할 것을 붙여쓴 오류와 붙여 써야할 것을 띄어쓴 오류의 형태로 나타난다. 즉, 본 논문에서 의존명사의 경우에는 앞말과 띄어야 하나 붙여 쓴 띄붙오류에 속하고, 어미의 일부, 접미사, 조사 등은 붙여 써야 할 것을 띄어 쓸 경우 붙띄오류라고 할 수 있다. 예를 들어 '데'는 단독 혹은 조사나 어미와 함께 쓰여 한 어절을 이룰 때 의존명사로 사용되고, 앞말과 함께 어미 '-는데'의 일부로 사용되면 앞말에 붙여 쓴다.

'의존명사/접미사'는 의존명사와 접미사의 역할을 공유하는 유형들을 모은 것이다. 예를 들어 '넋'은 "해가 질 넋에 출발했다."에서 어떤 형편·처지·상황에 놓여 있는 시공간의 방향을 나타내는 뜻을 가질 때는 의존명사의 역할을 하므로 앞말과 띄어쓴다. 그러나 "들넋, 아랫넋"에서는 어떤 쪽(방향)을 나타내는 접미사로 사용될 때는 앞말과 붙여 써야 한다.

'의존명사/조사'는 의존명사와 조사의 역할을 공유하는 유형들을 모은 것이다. 예를 들어 "들리는 것은 소리뿐"에서 '뿐'은 체언 뒤에 나타나 조사로 사용되었으므로 앞말과 붙여 써야 한다. "바라볼 뿐"에서 '뿐'은 용언의 관형사형 어미 뒤에 나타나 선행문의 형편, 처지에 한정되어 의존명사로 사용되었으므로 앞말과 띄어써야 한다. 마찬가지로 "좋을 대로 해라"에서 '대로'는 내포문이 한정하는 형편·처지·상황과 동일한 의미로 의존명사로 사용되었으므로 앞말 '좋을'과 띄어 써야 한다. 그러나 "마음대로 하십시오"에서는 조사로 사용되어 앞말과 붙여써야 한다.

'되다'는 동사가 되는 경우가 있고, 접미사가 되는 경우가 있다. "결정하다", "해결하다"에서 처

럼 '-하' 접미사가 붙어 동사가 되는 말에 '-하' 대신 '-되'가 붙는 "결정되다, 해결되다"의 '-되'는 동사 파생 접미사이므로 붙여쓴다. 그러나 "문제 되는"과 같이 명사 뒤에 '되다'가 나오는 형태는 위의 "결정되다, 해결되다"와 같은 형태를 가지지만, "문제 되는"에서의 '문제'는 동사인 '되다'의 보어로서 조사가 생략된 형태이므로 '문제'와 '되다'는 띄어써야 한다[2].

본 논문에서 교정 방법을 찾기 위해 이용한 코퍼스는 97년 한국과학기술원(KAIST)에서 만든 대한민국 국어 정보베이스의 한국어 텍스트 코퍼스 5만 어절과 동아일보 사설 1만 6천 어절이다.

3.1 '로써/로서'의 교정 방법

이 절에서는 일반적으로 틀리기 쉬운 철자 오류의 예인 '로써/로서'의 교정 방법에 대하여 설명한다.

- 방법 1. 용언의 명사형 + '로써'
- 방법 2. 사람의 의미자질을 가지는 명사 + '로서'
- 방법 3. 상용구

방법 1에 나타난 모든 형태에 조사 '로서'가 잘못 사용된 경우 '로써'로 교정한다. 방법 2는 자격을 나타낼 수 있는 사람의 의미자질을 갖는 명사가 선행될 때 '로서'로 교정하는 방법이다. 모든 명사에 대한 의미자질을 부여하는 데에는 한계가 있다. 따라서 본 논문에서는 모든 명사에 의미 자질을 부여하지 않고 사람의 의미자질을 가지는 명사에 대한 정보를 사전에 두고 이를 이용한다. 방법 3은 국어사전에는 나타나지 않는 어휘이나 일반적으로 많이 사용하는 형태를 묶은 것이다.

예 1)

- ① 저 돌을 움직임으로써 문을 열 수 있다.
- ② 교육자로서 일생을 보내다.
- ③ 현재로서, 당시로서 등

3.2 '-는데/-ㄴ 데'의 교정 방법

이 절에서는 띄어쓰기 오류 유형 중 '의존명사/어미의 일부'의 예인 '-는데/-ㄴ 데'의 교정 방법에 대하여 설명한다.

- 방법 1. '데[조사]' + '있다/없다'의 활용형 또는 그 파생어
- 방법 2. '데' + 격조사 혹은 어미

방법 1에서 '데'는 의존명사로 사용되어 앞말과 띄어 쓴다. '데'가 '것에, 곳에'로 바꾸어 사용해도 말이 될 때 의존명사로 사용된 경우인데 '-에 있다/없다'가 선행되는 말은 주로 장소에 관련된 말이 오기 때문이다.

예 2)

- ① 그것의 목적은 지식을 얻는 데 있다.
- ② 그를 설득하는 데에 며칠이 걸렸다.

3.3 '듯'의 교정 방법

이 절에서는 띄어쓰기 오류 유형 중 '의존명사/어미의 일부'의 예인 '듯'의 교정 방법에 대하여 설명한다.

- 방법 1. 용언의 관형사형 어미 아래에서 띄어쓴다.

'듯'은 의존명사와 어미의 일부로 함께 사용될 수 있기 때문에 띄어쓰기에 혼란을 가져온다. 의존명사 '듯'은 용언의 관형사형 어미 '-ㄴ, -ㄹ' 아래에 쓰여 '그럴 것 같기도 하고, 그렇지 않을 것 같기도 하다'는 뜻으로 사용된다. '듯'이 어미의 일부로 쓰일 때는 '듯이'의 준말로 각 어간에 붙어, '비슷하거나 같은 정도'의 뜻을 나타낸다.

예 3)

- ① 있는 듯 없는 듯하다.
- ② 새가 하늘을 날듯이 ...

방법 1을 적용했을 때, '날다, 살다, 길다'와 같

이 어간의 끝음절 종성이 'ㄹ'인 단어들은 형태소 분석시 모호성이 발생한다. 예를 들어 예 ②의 '날듯이'는 '나(出)+ㄹ듯이'와 '날(飛)+듯이, 날(飛)+ㄹ듯이'로 형태소 분석된다. 어간에서 발생하는 형태소 분석 결과의 모호성은 구문 분석을 이용할 경우 해소될 수 있다. 그러나 아래 예 ③, ④에서 각각 '날(飛)+듯이'인지 '날(飛)+ㄹ듯이'인지를 결정하기 위해서는 앞뒤 문맥이나 상황을 고려해야 한다.

- ③ 독수리가 날듯 날렵하게 비행한다.
- ④ 새끼새가 곧 날 듯 날개를 파닥거린다.

3.4 '뿐, 만큼, 대로'의 교정 방법

이 절에서는 띄어쓰기 오류 유형 중 '의존명사/조사'의 예인 '뿐, 만큼, 대로'의 교정 방법에 대하여 설명한다.

방법 1. 용언의 관형사형 어미 아래에서 띄어쓴다.

'뿐, 만큼, 대로'는 의존명사와 조사의 역할을 공유한다. 의존명사인 '뿐, 만큼, 대로'는 용언의 관형사형 어미 '-ㄴ, -ㄹ'이 선행되면 반드시 띄어 쓰고, 체언 뒤에 나타날 경우 조사로서 체언과 붙여 써야 한다. 예를 들면 다음과 같다.

예 4)

- ① 끊임없는 투쟁만이 있을 뿐이다.
- ② 유일하게 살아남은 사람은 그뿐이다.
- ③ 그녀를 연상할 만큼 닮았다.
- ④ 어머니라는 호칭만큼 그리운 말도 드물다.
- ⑤ 그가 바라는 대로 행동한다.
- ⑥ 네 뜻대로 해라.

예 5)

- ① 그건 하나의 표현일 뿐이야.
- ② 온통 집안일뿐이야.
- ③ 재산 공개가 주 이슈인 만큼...
- ④ 미술 애호인만큼 미술을 사랑하는 사람들

그러나 예 5)의 ①과 ②의 '표현일'과 '집안일'에서 '일'이 형태는 같지만 각각 '이-(서술격 조사)+-ㄹ'과 '일(事)'로 사용된다. 마찬가지로 예③, ④에서도 '인'이 '이-(서술격 조사)+-ㄴ'과 '인(人)'으로 사용될 수 있기 때문에 띄어쓰기에 모호성이 발생한다. 그런데 '집안일일 뿐이야'나 '그녀는 미술 애호인인 만큼'과 같이 '일'이나 '인'이 연속되어 나타나면, 두 번째 '일, 인'은 분명히 '이-(서술격 조사)+관형사형 어미(-ㄴ, -ㄹ)'이므로 '뿐, 만큼'과 띄어 쓰면 된다. 그러나 예 5와 같은 문장들의 정확한 띄어쓰기를 위해서는 구문 분석이 필요하다.

'대로'는 위에서 언급한 형태들 외에 명사 '대로(大路)'와 중의성이 발생할 수 있다.

예 6)

- ① 올림픽대로를 따라 내려가세요.
- ② 너 같 대로 가라.

예6의 ①에서처럼 특정 지역을 나타내는 '올림픽대로, 천호대로' 등과 같이 '대로'가 명사 '대로(大路)'로 쓰일 수 있기 때문에 '대로(大路)'가 포함된 복합어를 사전에 포함시켜 처리한다. 그런데 예6의 ②에서 '같 대로'는 동사의 어간 '가-'에 관형사형 어미 '-ㄹ'과 의존명사 '대로'로 분석될 수 있고, 명사 '같대'와 조사 '로'의 분석도 가능하다. 이러한 경우에 구문 분석을 이용하여 교정해야 한다.

3.5 '되다'의 교정 방법

이 절에서는 띄어쓰기 오류 유형 중 '동사/접미사'의 예인 '되다'의 교정 방법에 대하여 설명한다.

방법 1. 명사 + 동사 파생 접미사 '되'

'되다'는 그 자체로 동사이면서 '하다'처럼 동사 파생 접미사로도 쓰인다. 이때 '되다'가 동사일 경우 앞 명사와 띄어 쓰고, 동사 파생 접미사일 경우 앞 명사와 붙여 써야 한다.

‘결정되다, 해결되다’가 ‘결정하다, 해결하다’처럼 ‘-하’ 접미사가 결합될 수 있는 명사 ‘결정, 해결’에 ‘-되’가 붙은 경우, 이 ‘-되’는 접미사이다. 그러나 ‘-하’가 붙을 수 없는 명사 뒤에 ‘되다’가 결합된 경우, 이 ‘되다’는 보어인 앞 명사의 동사가 된다. 따라서 이를 처리하기 위해 사전에 ‘-하’가 결합될 수 있는 명사에 대한 정보를 두어 이용한다. 본 시스템에서는 양방향 최장일치법의 품사 분류 방법을 이용하여 이를 해결한다[2].

예 7)

- ① 문제하다(X) -> 문제 되는
- ② 주목하다(O) -> 주목되는

예 7)의 ①과 ②에서처럼 ‘하다’ 와 붙을 수 있는 단어일 때 붙여 쓰고, 그렇지 않으면 띄어 쓴다.

4. 실험 및 고찰

본 논문에서 제안한 선택 오류의 교정 방법을 이용한 정확률 실험과 오류 교정 시스템의 정확률 실험, 2가지 실험에 대한 결과를 제시한다. 먼저 선택 오류 교정 방법에 대한 실험을 위해 97년 한국과학기술원(KAIST)에서 만든 대한민국어 정보베이스의 한국어 텍스트 코퍼스 20만 어절, 동아일보 사설 20만 어절과 기타(통신 기사판의 문서) 20만 어절 등 총 60만 어절을 대상으로 하였다. <표 3>은 선택 오류 유형과 교정률을 나타낸 것이다.

<표 3> 선택유형과 교정률

유형	교정률 (%)
‘로써/로서’	68.6
‘-ㄴ 데/-ㄴ 데’	72.9
‘dot’	70.0
‘뺨’	86.9
‘만큼’	86.5
‘대로’	100
평균	80.8

위 실험 결과 평균 80%의 교정률을 보였으며, 추후 구문 분석 및 의미 분석 등의 방법을 보완

하면 교정률이 향상될 것으로 생각된다.

오류 교정 시스템의 성능을 검증하기 위하여 본 논문에서 구현한 오류 교정 시스템(HEC : Hanguk Error Corrector), H사 워드프로세서, M사 워드프로세서의 맞춤법 검사/교정 시스템 등 3가지 시스템에 대하여 실험한 결과는 <표 4>와 같다.

<표 4>의 분석 결과, M 워드프로세서의 맞춤법 검사기는 보조용언, 복합명사의 분리 원칙이 없고, 복합명사 및 오류 단어 분리 시 후보 답이 너무 많으며 이들 간의 우선 순위가 없어 사용하기 조금 불편함이 있었고, 오류 단어 교정 후 제대로 처리되었는 지에 대한 형태소 분석을 하지 않음으로 인해 잘못된 결과가 출력되는 경우가 종종 있었다. M 워드프로세서가 H 워드프로세서보다 교정률이 상대적으로 높게 나타난 이유는 교정 단어의 후보자 수가 상대적으로 많기 때문으로 분석된다. 그러나 교정의 신뢰도에 결정적으로 영향을 미치는 검색하지 못하는 오류 단어(비교정 오류 단어)의 수가 가장 많이 나타나 시스템의 성능이 가장 떨어지는 것으로 분석되었다. H 워드프로세서는 보조용언은 무조건 붙여쓰고 복합명사 분리를 하지 않았으며, 두개의 워드프로세서 모두 모호성 처리를 하지 않은 것으로 분석되었다.

본 논문에서 구현한 HEC는 1개의 교정 결과만을 제시하는 교정 시스템임에도 불구하고 상대적으로 많은 교정 후보자를 출력하는 두 개의 워드프로세서(66%, 63%)에 비해 91%라는 높은 교정률을 나타내었으며, 검색되지 아니한 오류 단어의 수도 상대적으로 가장 적게 나타나 시스템의 신뢰도가 가장 높은 것으로 판명되었다. 또한 교정 결과가 2개 이상을 가지는 모호성 문제에 대해서도 90% 정도까지 해결하였다. 예를 들어 ‘먹는가’, ‘먹는가가’, ‘먹는데있다’는 “먹는가’, ‘먹는가가’, ‘먹는 데 있다’로 결과를 출력한다. 그러나, HEC에서 나타나는 오류의 대부분은 사전에 미등록된 단어에 대한 것과 복합명사 분리 시 대부분의 연구 및 시스템에서 처리하지 않는 3음절 복합명사에 대한 분리를 행함으로써 상대적으로 많은 분리 모호성이 발생하여 잘못 교정된 오류로 나타났다.

<표 4> 기존의 시스템과 HEC와의 교정률 비교

	M 워드프로세서	H 워드프로세서	HEC
1. 총 실험 어절 수	47,182	47,182	47,182
2. 오류 추출 단어 수 (동일 단어 수 제외)	853	1,082	2,074
3. 잘못된 오류 교정 단어 수 (외래어, 특수 미 등록어 제외)	126	281	142
4. 오류 미검색 단어 (복합명사, 보조 용언 제외)	243/317	188/317	57/317
5. 비교정 오류 단어 수 ②+③	369	469	199
6. 오류 단어의 교정률 (1-(3+4)/(2+4))*100	66%	63%	91%
7. 특성	1) 복합명사, 보조용언 분리 원칙 없음 2) 모호성 처리 안함	1) 보조용언 무조건 붙여서 교정 2) 복합명사 분리 안함 3) 모호성 처리 안함	1) 복합명사 분리 2) 보조용언 분리 (반드시 붙여 써야 하는 것 제외) 3) 모호성 문제 어느 정도 해결

특히 HEC는 <표 5>와 같은 모호성 단어를 해결하고 있다.

<표 5> HEC에서 처리 가능한 모호성 단어의 예

부산)은 => 부산은 부산)대구는 => 부산) 대구는	어느해보다 => 어느 해보다 정리해보다 => 정리해 보다
훈련만하고 => 훈련만 하고 그만한 => 그만한	나이가들어감에 => 나이가 들어 감에 방에들어감에 => 방에 들어감에
이유중하나는 => 이유 중 하나 출장중이다 => 출장중이다 그중에서 => 그 중에서	삼만오천 => 삼만 오천 예순 여섯 => 예순여섯 한두개 => 한두 개
직접들은 => 직접 들은 사람들은 => 사람들은	몇수천만명 => 몇수천만 몇사람 => 몇 사람명
최씨는 => 최씨는 최재혁씨는 => 최재혁 씨는	뉴욕시 => 뉴욕 시 부산시 => 부산시 추출시는 => 추출 시는

5. 결론

본 논문에서 구현한 오류 교정 시스템(HEC)은 기존의 맞춤법 검사기의 단점인 오류 수정 작업

의 불편함과 처리 시간의 감소, 높은 교정률을 가진 최초의 한국어 오류 단어 자동 교정 시스템 이라는데 가장 큰 의의를 가진다.

본 시스템은 기존의 맞춤법 검사기의 단점인 오류 수정 작업과 처리 시간을 감소시키면서, 높은 오류 교정의 정확률을 보장하는 자동 오류 교정 시스템의 개발을 위한 첫 단계로써 한국어 오류의 80% 이상을 차지하는 띄어쓰기 오류에 대한 자동 교정 시스템이다. 본 시스템은 오류 어절에 대해 한 개의 교정 결과만을 제시하고 사용자에게 페이지 단위로 교정 결과를 보여준 후 일괄 처리하였으며, 확인된 결과는 그 다음부터는 출력되지 않고 그 파일이 끝날 때까지 자동 교정 되도록 하여 매번 오류 단어가 나타나는 것을 방지하였다.

본 논문에서는 우리가 사용하는 일반 문서에서 띄어쓰기가 잘못된 단어에 대한 교정과 오류 단어에 대한 교정을 행하기 위하여, 띄어쓰기 교정 및 오류 단어 교정 시스템의 개발 시 교정 정확률 및 처리 속도를 높이기 위해 본 시스템에서 구축한 각종 오류 교정 방안을 제시하였다. 특히 선택 오류를 처리할 수 있는 교정 방안을 제시하였다. 본 시스템의 처리 결과, 약 81% 정도

의 선택 오류 단어에 대한 교정률과 47182 어절에서 57개의 비교정 오류 단어를 포함한 총 2131개의 오류 어절 중에서 1932개의 오류 어절을 올바르게 교정하여 91%의 높은 교정 정확률을 보였다.

앞으로 오류 교정 시스템의 교정률을 높이기 위해 일반 사용자가 가장 많은 오류를 범하는 선택 오류 단어에 대한 보완과 미등록어 추정에 대한 연구가 이루어지면 교정률의 향상을 꾀할 수 있으며, 오류 어절 사전의 보완 등이 이루어지면 상용화도 가능할 것이다.

본 시스템의 개발 의의를 다음과 같이 정리할 수 있다.

1. 정확한 문서 작성의 환경을 통하여 올바른 의미의 문서 전달이 가능하다
2. 반복되는 오류 발생의 예방 효과와 맞춤법 학습의 교육적인 효과를 볼 수 있다.
3. 문자 인식이나 음성 인식 결과의 후처리에 매우 효과적이므로 한국어 응용처리 시스템의 질을 한 차원 높일 수 있는 계기가 될 수 있다.

참고 문헌

[1] 최 재혁(1996). 양방향 최장일치법을 이용한 한국어 띄어쓰기 자동 교정 시스템. 제 9회 한글 및 한국어 정보 처리 학술발표 논문집.

[2] 최재혁(1994). 양방향 최장일치법에 의한 한국어 형태소 분석기의 구현. 경북대학교 공학박사학위 논문.

[3] 최재혁(1996). 한국어 복합명사 분리 방안. 제 8회 한글 및 한국어 정보 처리 학술발표 논문집.

[4] 최재혁(1996). 키워드 추출을 위한 한국어 보조용언의 처리 방안. 교육과학연구소 논문집, 창간호 부산여자대학교.

[5] 최재혁외 4인(1997). 한글 문서에서 선택 오류에 대한 교정. 제 5회 한국정보과학회 영남지부 학술 발표 논문집.

[6] 채영숙, 이영식, 권혁철(1992). 한글 철자 오류 교정 시스템. 한글 및 한국어 정보

처리학술발표논문집.

- [7] 박영환, 김경선, 송만석(1991). 말뭉치에 기반한 형태소 분석기 및 철자 검사기의 구현. 한국정보과학회 가을 학술발표 논문집, 18권 2호.
- [8] 최재혁외 2인(1996). 맞춤법 교정을 위한 한국어/외국어 구분 방안. 한국정보과학회 춘계학술발표논문집, 제 23권 1호
- [9] 최재혁(1995). 양방향 최장일치법에 의한 한국어 키워드 추출시스템. 자연과학연구소 논문집, 제 1집, 부산여자대학교.
- [10] 강승식, 김영택(1990). 한국어 문법 검사/교정 시스템의 설계. 한국정보과학회 봄 학술 발표논문집, 18권 2호.
- [11] 김민정, 권혁철(1994). 한국어 특성을 이용한 자동 색인 기법. 한국정보과학회 가을 학술발표논문집, 19권 2호.
- [12] 심광섭(1996). 음절간 상호 정보를 이용한 국어 자동 띄어쓰기. 한국정보과학회 논문집, 23권 9호.
- [13] 이은철, 이종혁(1992). 계층적 기호 접속 정보를 이용한 한국어 형태소 분석기의 구현, 제 4회 한글 및 한국어 정보처리 학술발표 논문집.
- [14] 정한민, 이근배, 이종혁(1993). 자판 특성을 이용한 Neuro-fuzzy 한국어 철자 교정기의 구현, 제 5회 한글 및 한국어 정보처리 학술발표 논문집.
- [15] 심철민, 김현진, 김영진, 권혁철(1995). 언어 정보를 이용한 한국어 철자 검사기의 기능 개선, 제 7회 한글 및 한국어 정보처리 학술발표 논문집.
- [16] 최기호(1994). 한글 맞춤법 새 길라잡이. 토담.
- [17] 장석진(1993). 정보 기반 한국어 문법. 언어와 정보.
- [18] 원영섭(1992). 띄어쓰기 맞춤법 용례. 세창출판사.
- [19] 미승우(1993). 새 맞춤법과 교정의 실제. 어문각.
- [20] 김정숙(1994). 정보 사회와 우리글 바른 표기. 일진사.

[21] 손세모들(1996). 국어 보조용언 연구. 한국
문화사.



최 세 혁

1984 경북대학교 전자공학과
(공학사)

1986 경북대학교대학원
전자공학과(공학석사)

1994 경북대학교대학원 전자공학과(공학박사)

1988 한국전자통신연구원(ETRI) 위촉연구원

1999~2000 미국 UCI 방문 교수

1989~1994 신라대학교 전자계산학과 조교수

1995~현재 신라대학교 컴퓨터교육과 부교수

관심분야: 한국어 정보처리, 자연어처리, 정보검색, 컴파일러.

E-Mail: jhchoi@silla.ac.kr



김 권 양

1983 경북대학교 전자공학과
(공학사)

1990 경북대학교대학원
전자공학과(공학석사)

1998 경북대학교대학원 컴퓨터공학과(공학박사)

1983~1988 한국전자통신연구원(ETRI) 연구원

1999~2000 미국 University of Central Florida
방문 교수

1991~현재 경일대학교 컴퓨터공학과 부교수

관심분야: 자연어처리, 인공지능, 정보검색, 한국어 정보처리.