GOVP1199223159

'90 첨단 요소 기술 과제

100

한국어 철자 및 띄어쓰기 교정 시스템에 관한 연구

Research on Spelling and Word-spacing Corrector for Korean Language

연 구 기 관
한 국 과 학 기 술 원

과 학 기 술 처

제 출 문

과학기술처 장관 귀하

본 보고서를 "한국어 철자 및 띄어쓰기 교정 시스템에 관한 연구" 사업의 최종 보고서로 제출합니다.

1991. 6. 15.

주관연구기관 : 한국과학기술원

연구책임자 : 최기선 (한국과학기술원 교수)

연구원 : 박혁로, 김덕봉, 권철중,
조영환 (한국과학기술원 박사과정)

연구조원 : 강중빈, 정진성 (한국과학기술원 석사과정)

요 약 문

I. 제 목

한국어 철자 및 띄어쓰기 교정 시스템에 관한 연구

II. 연구 개발의 목적 및 중요성

컴퓨터의 대중화와 편리한 문서편집기의 보급으로 컴퓨터를 이용한 문서의 관리가 일반화되고 있다. 컴퓨터를 이용한 문서 작성은 문서의 재사용과 수정이 용이하기 때문에 유용할 뿐만 아니라 컴퓨터를 이용한 정보의 가공 측면에서도 필수적이다. 저장된 문서에 존재하는 오류들은 입력전의 문서에 내재되어 있었건, 혹은 입력 도중에 발생하였건 간에 고쳐져야 하는 것이다. 저장된 문서를 이용하는 시스템이 문서에 대한 전적인 신뢰를 하고 있다면 문서 내의 오류는 가공 시스템의 신뢰성 저하에 원인이 될 수 있다. 그 반대로 문서에 오류가 있다고 가정하여, 이를 해결하며 처리하는 시스템은 오류 처리에 대한 추가적인 부담을 안게 된다. 응용 시스템을 문서 내의 오류에서 자유롭게 동작할 수 있도록 하기 위해서는 입력 문서에 대한 오류 수정 작업을 별개로 여겨서 처리하는 방법을 생각할 수 있다. 그러므로 정보 가공 시스템과 별도로 입력 문서에 존재하는 오류를 제거하기 위하여 입력된 문서에서 오류를 발견하여 교정하는 방법에 대한 연구가 필요하다.

III. 연구 개발의 내용 및 범위

본 연구에서는 한국어 철자 및 띄어쓰기 교정 시스템을 개발한다. 본 연구의 최종 목표는 형태소 해석과 구문 해석과 의미 해석을 통한 단어의 철자 및 띄어쓰기 오류의 교정이다. 본 연구의 1차년도 연구에서는 형태소 해석 방식을 채택하고 있으므로 문서에서 철자 오류와 띄어쓰기 오류를 발견하고 교정할때 형태론적인 오류 여부만을 관심 둔다.

IV 연구 개발 결과 및 활용에 대한 건의

1. 연구 결과

- 한국어 형태소 해석기 개발

  * 한국어 좌우접속 테이블 작성 * 한국어 형태소 분리 시스템 구축 * 사전 인터페이스 개발

- 한국어 사전 개발

  * 기능어 사전 구축 * 동사 및 명사 사전의 구축

2. 기대 효과

- 한국어 형태소 해석 개발 기술 축적

  * 한국어 처리 시스템의 기초 시스템 구축

- 한국어 처리 기술 축적

  * 일반적인 언어처리를 요구하는 응용에 본 기술을 사용

- 각종 입력기의 작동오류 제거 기술 축적

SUMMARY

I. Title

Research on Spelling and Word-spacing Corrector for Korean Language

II. Purpose and Importance of R&D

The errors in textual documents can be major causes for the critical degradation of the reliability of the document-related systems. For example, such system includes wordprocessor, machine translation, chracter recognition, text database, intermeeting telephone, and so on. Handling textual errors mean two kinds of jobs; first, identifying errors and suppling suggestions. This system should use several kinds of knowledge and allow learing mechanism through user feedback. Here, Knowledge means liguistic, statistic, and heuristic knowledge. This R&D focuses on the liguistic knowledge: morphological, syntactic, and semantic ones. Such knowledge building needs long-term basic study on natural language and careful knowledge engineering and extraction. As every AI systems fails deu to knowledge storage, the importance of this R&D should be viewd from knowledge perspectives.

III. Contents and Range of R&D

In this first stage of R&D, we approached the problem based on morphological knowledge, whereas the final outcome of this study will also make use of syntactic and semantic information of texts. What we have developed in this first stage deals with the errors of morphological level.

CONTENTS

차 례

I. 서론

본 보고서는 "한국어 철자 및 띄어쓰기 교정 시스템에 관한 연구" 과제의 1차년도 연구 결과에 대하여 기술한다. 본 과제는 1990년 과기처 첨단 과제로 시작되어 1993년까지 3년동안 연구가 계속될 예정이다. "한국어 철자 및 띄어쓰기 교정 시스템에 관한 연구"과제는 한국과학기술원이 주관 연구 기관이다.

본 과제는 한국어 철자 및 띄어쓰기 교정 시스템을 개발하기 위한 과제로 다음과 같은 목표를 추구하고 있다.

	1차년도	2차년도	3차년도
수준	형태소해석의 수준으로 문검사	구문해석의 수준으로 문검사	의미해석의 수준으로 문검사
사전	기능어사전 완성 명사 및 동사 사전 5,000단어	기능어 사전 확충 명사 및 동사 사전 50,000단어 입력 및 코드 부여	기능어 사전 개량 명사 및 동사 사전의 실용화 및 확충, 개량
운영체제	UNIX 상에서 작동	MS-DOS상에서 작동	이식성 확보
처리방식	일괄 처리 방식	개별 처리 방식	문서처리기를 내장

본 과제는 한국어에 대한 철자 검사 및 띄어쓰기 오류의 발견과 교정을 위하여 한글 맞춤법[문교부90]을 기본으로 띄어쓰기 오류, 조사/어미 오류, 표준어 오류, 철자 오류 등으로 분류하였다. 각각의 오류를 발견하고 처리하기 위하여 오류를 유형별로 정리하여 처리하는 방법을 고안하였다. 본 과제의 접근 방법은 다음과 같다.

- 어절을 기본 단위로 해석
- 좌우접속정보에 의한 어절 구조 모델 수용
- 최장일치법에 기인한 형태소 해석
- 한글 맞춤법에 따른 오류의 분류
- 양방향접근법에 의한 오류 판별
- 사전의 이용과 어절 분석을 통한 교정의 병행
- 사용자와 독립적인 코드 체계

한국어 형태소 해석은 한국어에 대한 해석 영역이 넓은 접속정보를 이용한 어절 구조 모델[강재우89]을 수용하였고 처리의 용이성을 위하여 최장일치법에 기인한 형태소 해석 방법을 사용하였다. 오류로 판명된 어절에서 오류의 종류와 위치를 판별하기 위하여 양방향접근법을 고안하였다. 오류의 교정은 어절의 형태소 구조에 대한 분석을 통하여 처리하는 방법과 사전을 이용하는 방법을 병행하였다. 한글의 처리에 있어서 코드 체계는 심각한 문제를 야기한다. 사용자 코드 체계에 융통성을 부여하기 위하여 내부 코드 체계를 3바이트 체계로 고정하고 외부 코드와 변환 라이브러리를 통하여 처리하는 방법을 사용하였다.

본 보고서의 구성은 다음과 같다. II장에서 어절의 구조와 형태소 해석의 방법론, 그 수준의 철자 검사기에 대한 관련연구에 대하여 간략한 소개를 한다. III장에서는 철자와 띄어쓰기에 대한 검사와 교정을 위한 한국어의 특성에 대하여 고찰한다. IV장에서는 본 시스템에서 가정하는 오류의 유형을 분류하고 오류의 위치와 종류를 판별하기 위한 양방향접근법에 대하여 설명한다. 본 시스템의 설계 및 구현에 대한 설명으로 V장에서는 사전 시스템과 코드 변환 라이브러리, 오류 검사기, 오류 교정기에 관하여 설명한다. VI장에서는 구현된 시스템의 성능에 관한 실험에 대하여 설명을 하고, VII장에서 본 보고서의 결론을 맺는다.

II. 관련 연구

본 시스템은 한글 문서에 대한 맞춤법 오류를 교정하기 위하여 어절단위의 형태론적인 분석을 통한 처리를 한다. 주어진 어절에 대하여 형태소 해석이 가능하면 맞는 어절이고, 그렇지 않다면 틀린 어절이라고 판명할 수 있다. 오류로 판명된 어절은 적절한 처리를 하여 올바른 어절로 교정하여야 한다. 그러므로 본 교정 시스템은 오류를 포함한 입력 문서에 대한 형태소 해석기의 응용이다. 본 장에서는 형태소 해석을 위하여 본 시스템에서 가정하는 어절의 구조와 현재까지 발표된 철자법의 검사 시스템과 교정 시스템에 관하여 설명한다.

2.1 어절의 구조

한국어의 문장 구성은 형태소, 어절, 구, 문장의 순으로 이루어 진다. 말들의 계열관계와 통합관계에 따라 마디 지어지는 한덩이의 말을 어절이라고 한다. 맞춤법의 띄어쓰기는 어절을 단위로 하고 있는데 이는 말에서의 끊어짐이 글에서 사이띄기로 바뀌어진 것이다. 단어는 어절과 밀접한 관계를 맺고 있다. 한 어절이 그대로 한 단어가 되는가 하면 두 단어가 모여서 한 어절이 되기도 한다. 단어는 하나의 형태소로 이루어진 단일어와 둘 이상의 형태소로 이루어진 복합어로 나뉜다. 복합어는 파생법과 합성법에 의해 형성되는데 파생어란 실질 형태소에 형식 형태소가 결합된 말이고, 합성어란 두 개 이상의 실질 형태소가 결합된 말이므로 형태소들의 복잡한 구조가 어절을 구성한다. 형태소는 일정한 음성에 일정한 뜻이 결합되어 있는 말의 가장 작은 단위이며 이는 형태론의 최소단위가

되는 동시에 음운론의 상위 단위가 되기도 한다[남기심85].

어절은 형태소들의 복잡한 유형의 접속으로 구조를 이룬다. 어절은 홀로 사용 가능한 자립어에 부속어들이 붙어 있는 모양을 한다. 부속어들도 이것에 붙는 자립어의 성격에 따라 세분화되어지며 복잡한 접속 양상을 갖는다. 그림 2-1은 어절과 단어, 형태소의 관계를 나타낸다.

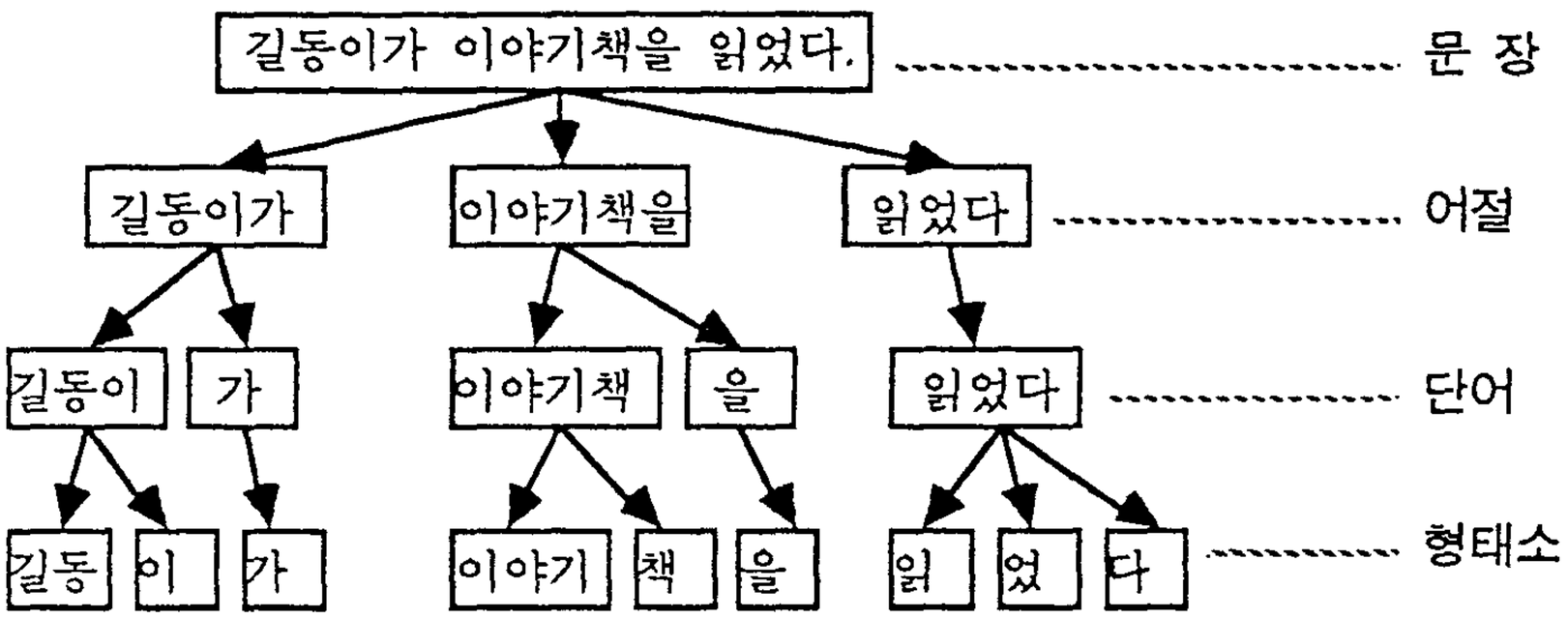


그림 2-1 문장의 구성 - 문장, 어절, 단어 형태소의 관계

2.2 좌우접속정보를 이용한 형태소 해석

형태소의 좌우접속정보는 파생법과 합성법에 근거한 단어의 형성을 표현하는 방법의 일종이다. 이 방법의 기본 개념은 형태소를 두 개의 독립적인 분류 체계에 따라 나누고 두 분류 체계간의 접속성을 조사하여 이를 형태소간의 접속 모델로 하자는 것이다. 형태소의 분류는 우선 자기 자신의 특성에 관한 분류가 있겠고 자기의 오른쪽의 형태소에 대한 제약적 의미의 분류가 있다. 예를 들면 고유명사는 오른쪽에 붙을 수 있는 형태소에 따라 인명, 지명, 국명, 사건명, 서적명, 기타

등으로 나뉘고 각각은 또 종성의 성격에 따라 유종성과 무종성, ㄹ종성으로 나뉘게 된다. 우리는 여기서 자기 자신의 특성에 관한 분류를 좌접속 분류라 하고, 오른쪽의 형태소에 대한 제약적 성격의 분류를 우접속 분류라고 하였다[강재우89]. 좌우접속모델의 기본 개념은 그림 2-2와 같다.

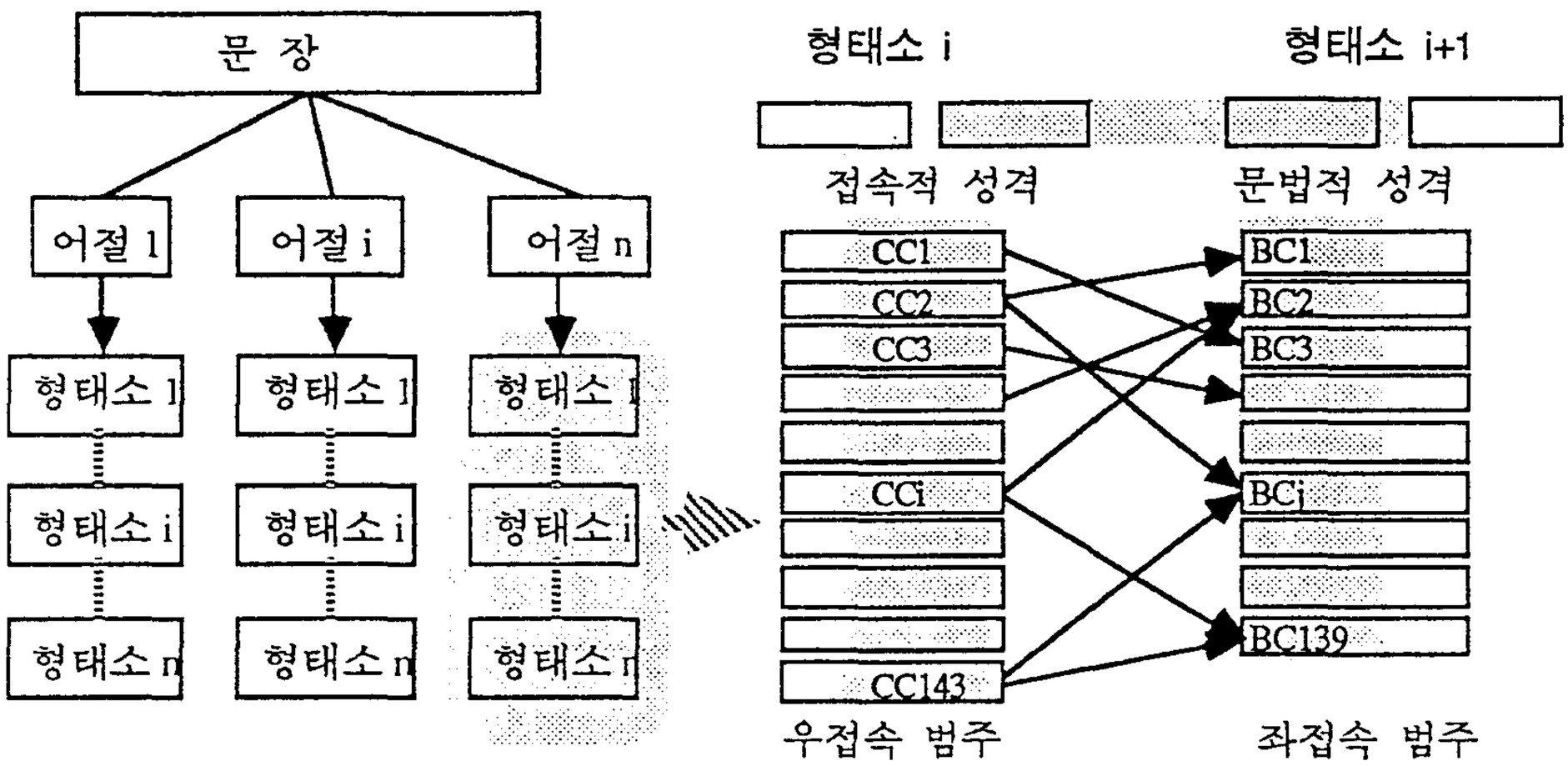


그림 2-2 좌우접속모델

이 방법은 형태소를 좌우접속 양상에 따라 범주화하고 각 범주 사이의 접속 가능성을 조사하여 도표화하여 어절내의 형태소에 대한 접속 구조를 알아내자는 것이다. 본 시스템은 형태소를 사전의 표제어로 하고 어절 단위로 처리하므로 한국어에 대한 해석의 영역이 넓은 형태소의 좌우접속정보[강재우89]를 이용한 어절 구조 모델을 수용한다. 본 시스템에서 사용하는 좌우접속정보표는 단지 입력 어절에 대한 오류 여부의 판별 뿐만 아니라 오류 어절에 대한 교정이 접속

가능해야 하므로 기존의 것보다 좀 더 섬세할 필요가 있었다. 복합 명사와 복합 용언의 경우 사전에 존재한다고 가정하고 명사끼리나 용언끼리의 접속의 가능성을 배제하였다. 접미사와 접두사의 경우에 제한을 완화시킨다면 의미적으로 어울리지 아니하는 경우가 발생하므로 일반적으로 허용 가능한 최소한의 접미사와 접두사를 제외하고는 모두 명사에 붙여져 있는 형태를 가정하였다.

2.3 Head-Tail구분을 기초로 한 해석

이 방법은 어절을 변형되지 않는 부분(Head)과 변형되는 부분(Tail)으로 구분하여 어절의 뒷 부분에서부터 가능한 모든 Tail을 찾아 가면서 Tail 테이블을 구성한다. 또한 어절의 앞 부분에서부터는 Head를 찾아내어 그 Head와 각 Tail들과의 접속 가능성을 Tail 테이블에 들어있는 정보를 이용해서 검사한다. 이러한 방법으로 접속 가능한 Head를 계속 찾는 방법이다. 이 방법은 음운 현상이나 불규칙 현상을 고려하여 용언과 어미를 세분하여 이들 사이의 결합 관계를 나타내는 표를 만들어 이용한다[최형석84]. 본 시스템의 오류 판별을 위한 양방향접근법은 어절의 양쪽 방향에서 각각 해석을 함으로써 위의 해석 방법과 비슷한 처리를 거친다.

2.4 한글 철자법 교정 시스템

이 시스템은 문서 편집 중에 철자를 교정하는 대화식 시스템으로 COBOL언어로 NEC/S100상에서 구현되었다. 접사를 찾는 방법은 최장일치법을 사용하였으며, 어절의 뒤에서부터 접사를 찾기 위해 조사/어미 테이블을 역순으로 구성하였다. 조사/어미 테이블을 이용하여 어절의 뒤에서 조사/어미를 잘라내고

나머지를 사전에서 찾아 없으면 오류로 간주하여 사용자에게 교정 여부 및 사전 등록 여부를 묻는다. 교정된 어절은 교정 단어 사전에 등록하여 다시 오류가 발생할 때 자동으로 교정한다. 따라서 일단 에러를 범한 단어들에 대해서 같은 오류를 범할 경우 자동적으로 교정할 수 있도록 하는데 촛점을 맞추었다[김영웅84].

이 시스템은 단지 철자법의 교정만을 처리하였으며 입력 문서가 한글 맞춤법에 맞게 띄어 쓰여야 하는 등의 많은 가정을 하고 있어 실용적이지 못했다.

2.5 한글 철자 및 띄어쓰기 검사기

본 시스템의 근원이 되는 시스템으로 접속정보를 이용한 형태소 해석으로 철자 및 띄어쓰기 오류를 검사한다. 이 시스템은 최단일치법을 이용하여 사전에서 일치하는 형태소를 찾는다. 사전에는 형태소를 기본으로 하는 사전 표제어와 표제어에 대한 좌우접속 정보를 담는다. 띄어쓰기 검사는 우선 삭각의 어절에 대해서 철자 오류를 검사하고 철자 오류가 발생한 경우에 이 오류가 띄어쓰기가 잘못되어서 생긴것이 아닌가를 검사한다[강재우90]. 그림 2-3은 이 시스템의 개략적인 구성이다.

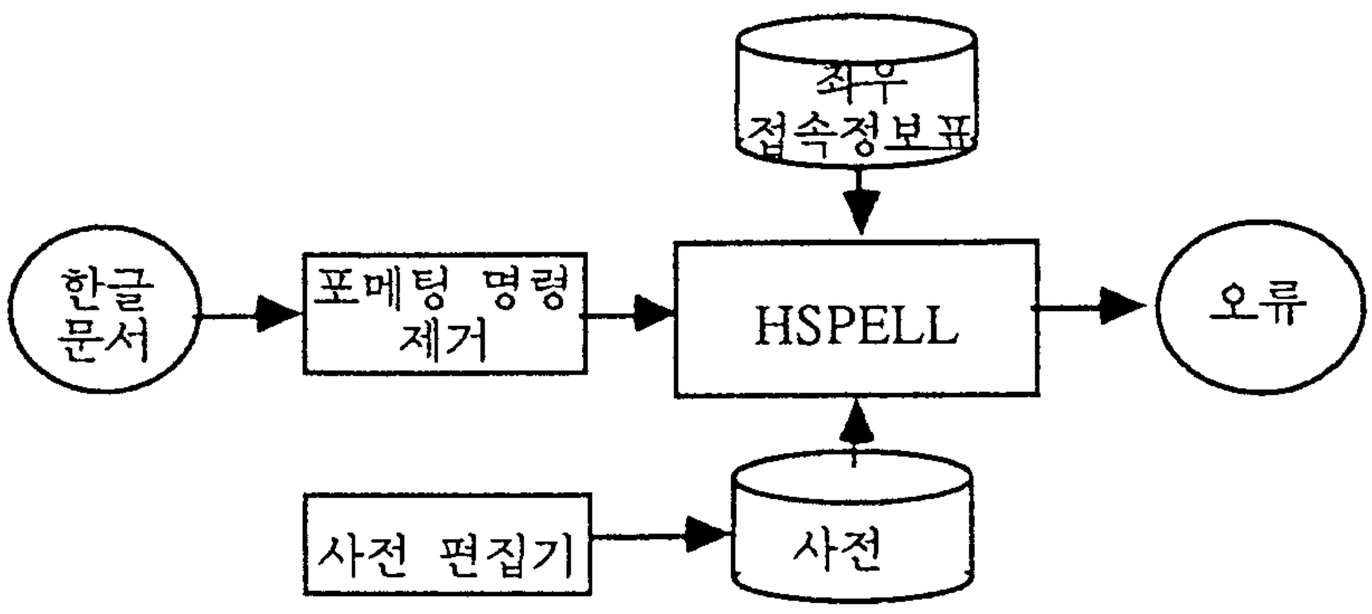


그림 2-3 한글 철자 및 띄어쓰기 검사기

이 시스템은 어절의 오류 여부를 검사하기 위하여 주어진 어절을 계속해서 최단 일치를 적용하여 사전을 찾고 이 과정에서 발견된 형태소들의 접속 정보를 이용하여 접속 가능성을 검사한다. 접속 가능성의 검사는 좌접속정보와 우접속정보를 index로 하는 2차원 bit-map을 이용한다. 한 어절은 몇개의 형태소로 구성되기 마련이고, 최단 일치가 잘못 적용된 경우에는 back-tracking을 해야 하기 때문에 스택을 사용한다.

본 시스템은 접속정보를 이용한 어절 모델을 따른다. 이 시스템의 최단 일치법은 한글 N바이트 코드 체계를 사용하여 구현된 시스템으로 다른 코드 체계에서는 적용이 힘들다. 그러나 N바이트 코드 체계는 처리가 단순한 반면에 사전의 구성이 비효율적이고 자소의 위치를 알아내기 힘들기 때문에 시스템에서는 적합하지 않다.

III 한국어에 대한 고찰

자연 언어의 기계적인 처리를 하기 위해서는 대상 언어에 대한 체계적인 연구가 필요하다. 본 보고서에서는 특히, 한글 철자 및 띄어쓰기 검사기를 구현하는 데 필요한 한국어의 특성 및 어절의 형태에 대해 살펴보고, 그 오류의 종류를 제시한다. 그리고 좌우접속정보표를 만들기 위해서 필요한 품사의 분류와 한국어 처리를 하는 데 필요한 불규칙 현상 및 음운 현상의 처리에 대해 살펴보겠다.

3.1 한국어의 특성

한국어는 계통상 우랄 알타이 어족에 속한다. 우랄 알타이 어족은 첨가어(교착어)로서 의미를 나타내는 실질 형태소에 조사와 어미 같은 어법적 관계를 나타내는 형식 형태소가 붙음으로써 문법 기능을 한다. 한글 철자 및 띄어쓰기 검사를 위해 고려해야 할 한국어의 특성은 다음과 같다[조규빈88, 성균관88, 미승우88, 김성용87].

○ 음절의 특징으로서 한글은 모아쓰기 특징을 가진다.

한글은 본질적으로 자음과 모음으로 구성된 언어이다. 이 자모가 조합되어 단음절을 형성한다. 따라서 컴퓨터 처리시 한글 코드가 조합형이 바람직하다. 본 시스템에서 가정하는 입력 문서는 한글과 영문이 혼용된 N 바이트 조합형 한글 코드이다.

○ 띄어쓰기가 일정하지 않고 복잡하다.

따라서 띄어쓰기 오류를 범하기 쉽다.

○ 한국어는 첨가어로서 각 낱말의 어미 변화에 의해 문장의 성분을 결정하며, 첨용과 활용이 자유롭다.

○ 한국어는 불규칙 현상 및 음운 현상이 발달하였다.

따라서 이러한 현상에 대한 처리가 필수적이다.

○ 한글과 한문 및 영문을 혼용하여 사용하는 경우가 많다.

한글과 영어가 붙어서 하나의 어절을 형성할 경우에는 다음과 같이 처리한다. 혼용 형태가 '영어+한글+영어', '영어+한글', '한글+영어'인 경우에는 '한글'만을 대상으로 하고, '한글+영어+한글'인 경우에는 '영어'가 '[]', '{}', '<>', '()'로 둘러싸인 경우에만 '한글+한글'을 대상으로 하여 처리한다.

3.2 품사의 분류

한국어에 대한 국어학자의 품사 분류는 다양하지만, 컴퓨터에 의한 한국어의 처리를 위해서 기존의 논문들에서는 각각 독자적인 품사 분류를 하고 있다[안동언87, 박상규84, 손우형86, 김성용87].

본 보고서에서는 좌우접속정보표의 작성과 사전 구성을 위하여 필요한 품사 분류를 그림 3-1과 같이 하였다.

이 분류는 학교 문법 체계에 따르되 본 시스템에서의 처리가 편리하도록 한 분류이다. 여기서 조용사는 명사나 의성어 또는 의태어에 붙어서 이것들을

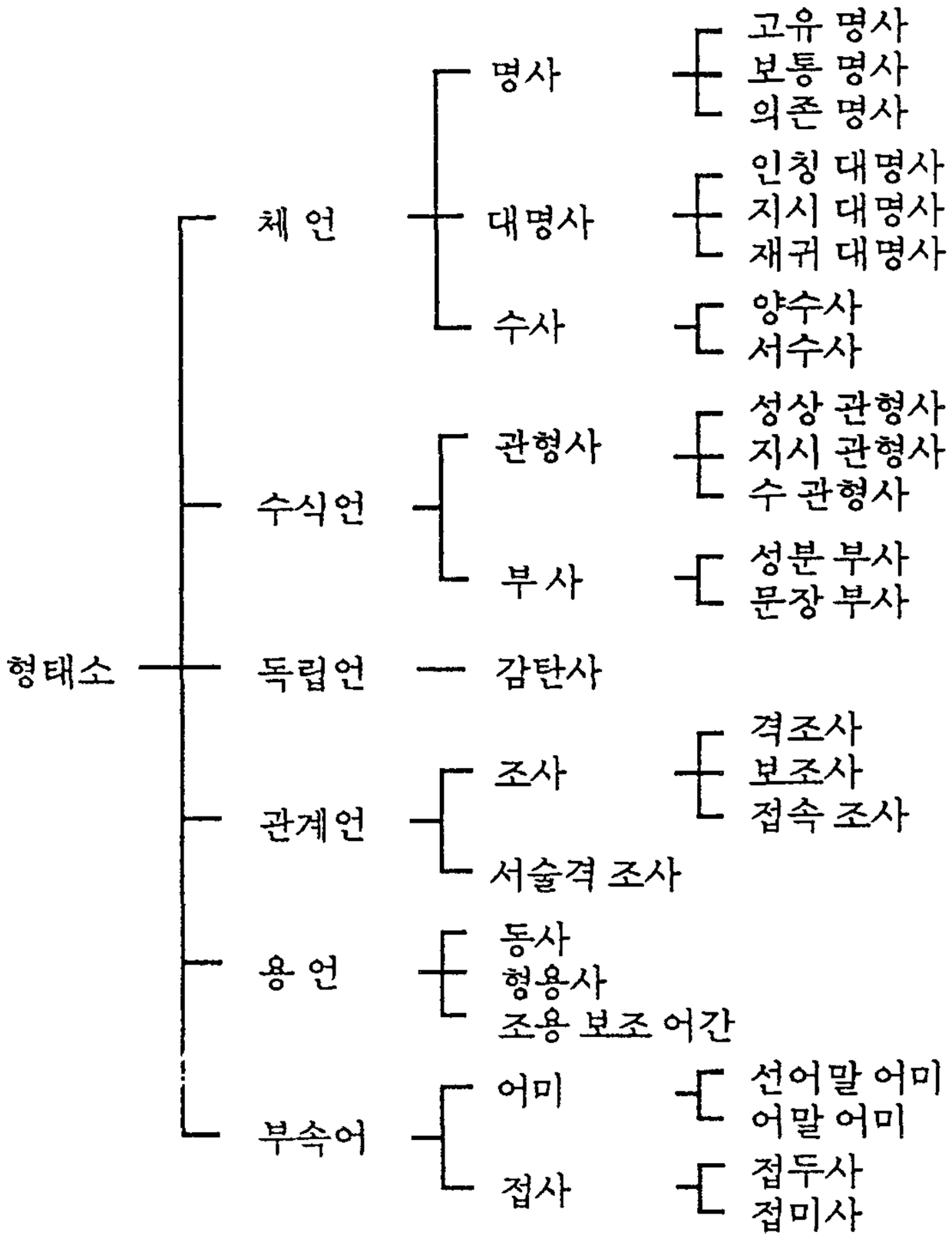


그림 3-1. 접속정보표 구성을 위한 한국어 품사 분류

용언화시키는 역할을 해 주는 어근을 말한다[김성용87]. 조용사로는 '하다', '이다', '되다' 등이 있는데, 조용사는 따로 분류하지 않으면 등록해야 할 표제어의 수가 상당히 많이 늘어난다. 따라서 사전의 크기를 줄일 수 있는 방법으로 조용사를 따로 분류한 것이다.

3.3 어절의 형태

한국어의 문장 구성은 형태소, 어절, 구, 문의 순으로 이루어진다. 가장 작은 의미 단위인 형태소(morpheme)가 모여 어절을 형성하고, 어절이 모여 구가 되며, 구가 연결되어 문장을 이룬다. 그런데 한국어는 어절 단위의 띄어쓰기를 하므로, 한 어절내에서 형태소를 파악하여 오류를 찾는 것이 철자 및 띄어쓰기의 과제이다[한국과88].

형태소들의 접속 양상을 알기 위해서는 한국어 어절의 구조를 파악하여야 한다. 한국어 어절의 개략적인 구조는 그림 3-2와 같이 나타낼 수 있다[성균관88, 조규빈88].

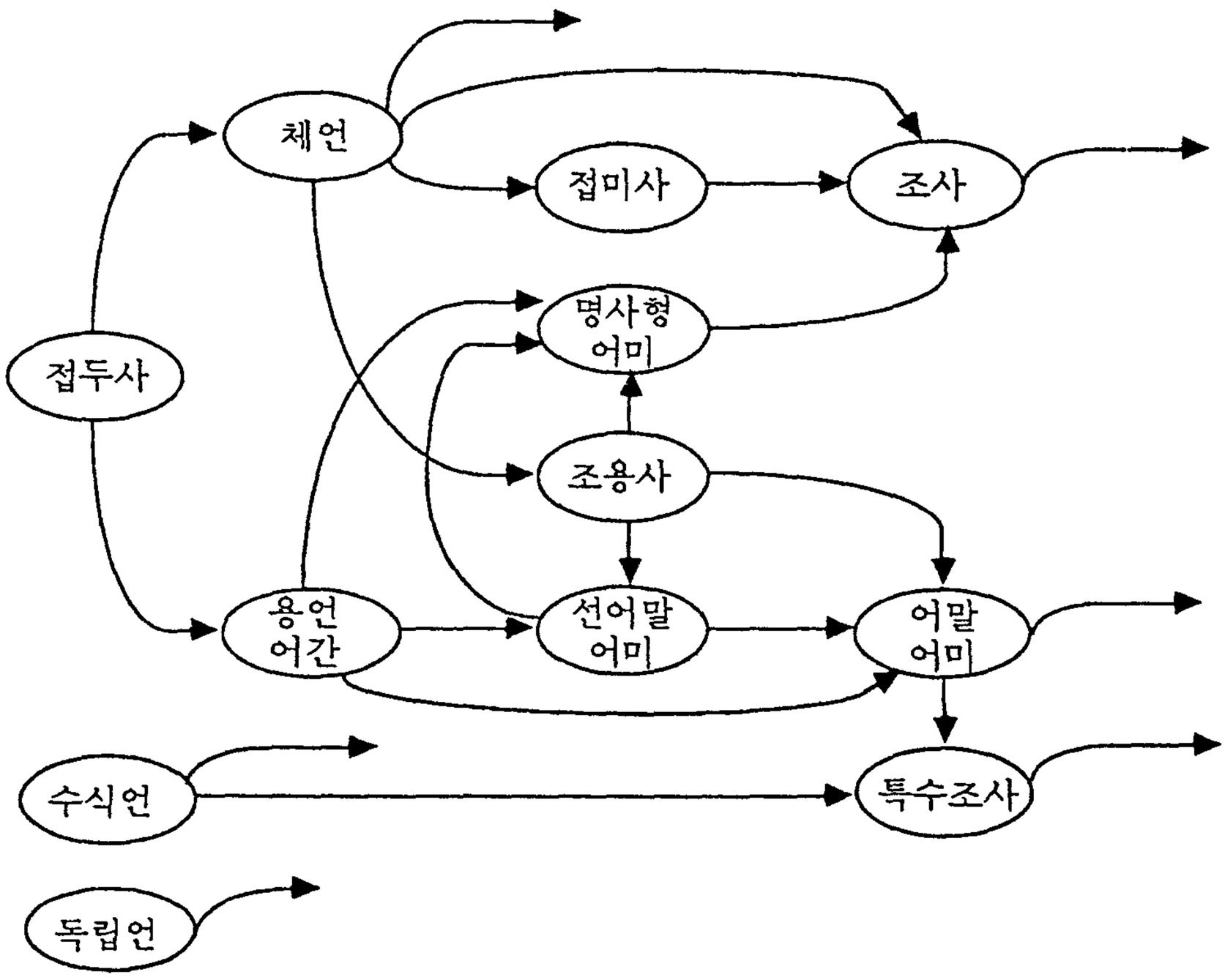


그림 3-2. 한국어 어절의 형태에 대한 네트워크

화살표(->)는 화살표의 좌측의 범주와 우측의 범주가 접속 가능함을 의미한다. 예를 들면, 어절 '과학자는'은 '과학+자+는'과 같이 명사 '과학'에 접미사 '자'가 붙고, 여기에 다시 조사 '는'이 붙은 형태이다. 또한 어절 '먹었다'는 '먹+었+다'로 동사 '먹다'의 어간 '먹'과 과거시제 보조어간(선어말어미) '었'이 붙고, 여기에 평서형 종결어미(어말어미) '다'가 붙은 형태이다.

3.4 각 품사별 처리

이 절에서는 3.2절에서 분류한 각 품사에 따라서 좌우접속정보표(부록 A,B)에서 형태소를 어떻게 처리하는지 간단히 설명한다.

3.4.1 체언

체언에는 명사, 대명사, 수사가 있다. 명사는 고유명사, 보통명사, 의존명사, 외래어로 나누었으며, 보통명사는 용언어간 겸용과 용언어간 비겸용으로 나누고, 이들을 각각 한자형과 순한글형으로 나누었다. 또한, 우접속 범주에서는 각 체언의 종성 유무에 따라서 무종성과 유종성(ㄹ종성 제외), 그리고 ㄹ종성으로 나누었다. 고유명사에는 좌측에 접속되는 것이 없고, 보통명사에는 접두사나 보통명사가 접속된다. 보통명사가 접속되어 복합명사를 만드는데, 복합명사는 처리하기 어려운 것 중의 하나이다. 모든 복합명사를 사전에 등록하면 잘못된 복합명사를 쉽게 검사할 수 있으나 사전의 크기가 문제이다. 본 시스템에서는 기본적으로 모든 보통명사끼리 접속되어서 복합명사를 만들 수 있다고 가정하고 있다. 의존명사는 단위성 의존명사인 경우는 좌측에 숫자나 양수사가 접속될 수 있다. 체언의

우접속은 접미사와 조사가 종성 유무나 체언의 특성에 따라서 달리 접속된다.

3.4.2 수식언

수식언에는 관형사와 부사가 있다. 수식언은 좌측에는 아무 것도 접속될 수 없고, 우측에는, 관형사는 아무 것도 접속될 수 없고, 부사인 경우는 특수조사가 접속될 수 있다.

3.4.3 독립언

독립언으로는 감탄사가 있는데, 좌우에 아무 것도 접속될 수 없다.

3.4.4 관계언

관계언에는 조사와 서술격조사가 있다. 조사는 격조사, 보조사, 접속조사가 있는데, 종성유무에 따라서 좌측에 체언, 접미사, 용언의 명사형, 부사, 종결어미, 보조적 연결어미가 접속될 수 있다. 우측에는 접속되는 것이 없다. 한국어에서 조사는 몇 개 안되지만 조사들이 결합되어 복합조사를 형성할 수 있어 많은 복합조사가 있다. 그러나 복합조사의 접속 형태가 규칙적이지 못하고 복잡하여 조사끼리의 접속 양상을 규칙으로 나타내더라도 잘못된 조사를 허용할 수 있다. 그런데 한국어에서 모든 가능한 복합조사를 조사하면 500여개가 되는데, 이를 모두 등록하면 잘못된 조사의 허용을 막을 수 있다. 따라서 본 보고서에서는 모든 복합조사를 등록하였다.

3.4.5 용언

한국어에서 동사와 형용사가 차지하는 비율이 30% 정도 된다[유재원85]. 특히, 용언들 대부분이 불규칙 활용 및 음운 현상을 일으키고 있어 이에 대한 처리가 필수적이다. 활용이란 용언이 일정한 문법적 관계를 표시하기 위하여 그 끝을 여러 가지로 바꾸는 현상을 말한다[조규빈88]. 용언이 활용할 때에 어간과 어미의 모습이 달라지는 일이 있는데, 이들 중 국어의 일반적 음운 규칙으로 설명할 수 없는 것을 불규칙 활용이라 한다. 한국어에서 불규칙 현상은 표 3-1과 같이 분류할 수 있다.

표 3-1. 한국어 불규칙 현상의 분류

구분	불규칙 현상	활용의 형태	예
어간의 바뀜	ㅅ 불규칙 *	ㅅ + 모음어미 -> 모음어미	짓 + 어 -> 지어
	ㄷ 불규칙 *	ㄷ + 모음어미 -> ㄹ + 모음어미	깨달 + 아 -> 깨달아
	ㅂ 불규칙	ㅂ + 모음어미 -> 오/우 + 어미 첫음절의 마지막 음소	돕 + 아 -> 도와
	르 불규칙	르 + 모음어미 -> ㄹ+ㄹ+모음어미	흐르 + 어 -> 흘러
	우 불규칙	ㅜ + 모음어미 -> 어미 첫음절의 마지막 음소	푸 + 어 -> 퍼
어미의 바뀜	여 불규칙	어간 + 아/어 -> 어간 + 여	하 + 아 -> 하여
	러 불규칙	어간 + 아/어 -> 어간 + 러	이르 + 어 -> 이르러
	거라 불규칙 *	어간 + 어라/아라 -> 어간 + 거라	가 + 아라 -> 가거라
	너라 불규칙 *	어간 + 어라/아라 -> 어간 + 너라	오 + 어라 -> 오너라
어간과 어미의 바뀜	ㅎ 불규칙 (형용사에만 해당)	ㅎ + 모음어미 -> 모음어미 ㅎ + ㄴ -> ㄴ	파랗 + 어 -> 파래 파랗 + ㄴ -> 파란

*는 동사에만 해당

표 3-2. 한국어 음운 형상의 분류

구분	음운 현상	활용의 형태	예
어간의 바뀜	으 탈락	으+모음어미 -> 모음어미	쓰+어 -> 써
	ㄹ 탈락	ㄹ+어미(ㄴ,ㅂ,ㅅ,오)->모음어미	살+니 -> 사니
어미의 바뀜	어미 아/어 (모음 조화)	양성어간+아/어 ->양성어간+아 음성어간+아/어 ->음성어간+어	잡+아/어 -> 잡아 먹+아/어 -> 먹어
	으 삽입	어간+어미(ㄴ,ㄹ,ㅂ,오,시,ㅁ) -> 어간+으+어미	잡+며 -> 잡으며
어간과 어미의 바뀜	간음화	아/어 탈락, 이+어 -> 여 오+아 -> 와, 외+어 -> 왜 우+어 -> 워, 오+이 -> 외 우+이 -> 위, 으+이 -> 의	차+아 -> 차 서+어 -> 서 즐기+어 -> 즐겨 보+아 -> 봐

또한 일반적인 음운 규칙으로 설명할 수 있는 활용의 불규칙성은 표 3-2와 같다

본 보고서에서는 불규칙 활용 및 음운 현상의 처리는 다음과 같이 한다.

(1) 규칙 어간

규칙 어간은 그대로 사전에 등록한다.

(2) 어간이 바뀌는 불규칙 어간

불규칙 어간 중 변하지 않는 부분만 사전에 등록하고 좌우접속정보표로 그들의 접속 형태를 제한한다.

(3) 어미가 바뀌는 불규칙 어간

규칙 어간을 사전에 등록하고 좌우접속정보표로 어미가 접속되는 형태를 제한한다.

(4) 어간과 어미가 바뀌는 불규칙 어간

불규칙 어간 중 변하지 않는 부분을 사전에 등록하고 좌우접속정보표로 어미의 접속되는 형태를 제한한다.

3.4.6 부속어

부속어로는 어미와 접사가 있는데, 접사는 접속형태가 복합명사의 문제와 같이 접사가 붙은 형태를 모두 사전에 넣든가, 그렇지 않으면 모든 보통명사에 접속 가능하도록 할 수 있는데, 본 보고서에서는 사전의 크기를 줄이기 위해 후자를 택했다. 어미는 다양하고 복잡하지만 수가 제한되어 있고 접속 형태가 잘 알려져 쉽게 처리할 수 있다.

3.5 오류의 종류

철자 및 띄어쓰기 검사시에 발생할 수 있는 오류는 다음과 같이 나눌 수 있다. 검사기가 가지는 시스템 오류와 이 시스템에 들어오는 입력문이 가지는 철자 및 띄어쓰기 오류인 입력문 오류이다.

3.5.1 시스템 오류

시스템이 가지는 오류로는

(1) 제 1 형 오류 : 시스템이 철자 및 띄어쓰기가 맞는 어절을 틀렸다고 취급하는 오류

(2) 제 2 형 오류 : 시스템이 철자 및 띄어쓰기가 틀린 어절을 맞았다고

취급하는 오류

등이 있다.

제 1 형 오류의 감소는 제 2 형 오류의 증가를, 제 2 형 오류의 감소는 제 1 형 오류의 증가를 가져 올 수 있다. 제 1 형 오류는 어절 중 미등록어가 있는 경우에 발생될 수 있고, 사전에 정보가 잘못된 경우에도 있을 수 있다. 제 2 형 오류도 사전 정보가 잘못된 경우에 발생할 수 있다. 시스템 오류를 모두 다 줄여야 하겠지만, 특히 제 2 형 오류를 줄이는 것이 실용적인 시스템을 만들기 위해 중요하다.

3.5.2 입력문 오류

입력문의 오류로는

(1) 철자 오류 : 한글 맞춤법에 맞지 않게 철자를 쓴 오류

(2) 띄어쓰기 오류 : 띄어쓰기를 잘못한 오류

가 있다.

철자 오류는 올바른 맞춤법을 알지 못하고 잘못 쓴 경우, 또는 타자가 잘못된 경우가 있다. 맞춤법을 바로 알지 못하고 쓴 경우는 일관되게 잘못 쓰는 일이 많다. 타자가 잘못된 경우는 일관되게 틀리지는 않지만, 많은 부분을 차지할 것으로 생각된다.

띄어쓰기 오류는 가장 많이 틀리는 오류 중 하나이다[미승우88]. 한국어는 영어나 일본어와는 달리 복잡한 띄어쓰기 규칙을 가지고 있다. 또한 이러한

띄어쓰기 규칙을 많은 사람들이 제대로 알지 못하고 있어서 문제가 더욱 심각하다 실례로서, 중고교 교과서에서의 오류를 조사한 자료를 보면 이들 오류 중 띄어쓰기 오류가 약 29.4%의 높은 비율을 차지하고 있는 것을 알 수 있다[미승우88] 띄어쓰기 오류의 종류를 다음과 같이 두 종류로 정의한다.

(1) 띄붙 오류 : 띄어 써야 할 것을 붙여 쓰는 오류

(2) 붙띄 오류 : 붙여 써야 할 것을 띄어 쓰는 오류

우선, 띄붙 오류는 하나의 어절을 형성하기 때문에 그 어절의 분석시 곧바로 밝혀질 수 있다. 따라서 그 어절 내에서 교정을 볼 수 있다. 그러나 붙띄 오류는 하나의 어절을 2개 이상의 어절로 띄어 쓰는 오류이다. 따라서 한 문장 안에 있는 모든 어절에 대해서 붙여 써야 하는지를 검사해야 한다. 한 문장 안에 n개의 어절이 있다면, 모든 가능한 접속의 경우의 수 $a sub n$은

이다. 그러나 붙띄 오류는 하나의 어절을 2개로 띄어 쓰는 경우가 대부분이므로, 인접한 두 어절에 대해서만 고려하였다.

IV. 맞춤법 오류의 유형과 양방향 접근법

어절에 존재하는 오류의 종류를 판별하고 적절한 대처를 하기 위해서는 어절구조 모델에 입각한 오류의 모델이 필요하다. 이 장에서는 맞춤법 오류의 유형을 정리한 후 각 오류 모델에 대한 발견과 교정의 방법론을 설명한다.

4.1 오류의 분류

어절의 옳고 그름의 기준은 맞춤법에 있다. 한글 맞춤법은 문장의 단위인 어절의 구성과 형태에 대한 규정을 말한다. 이러한 맞춤법을 염두에 두고 한국어 문서에서 자주 발생되는 오류를 분류하여 보기로 한다. 컴퓨터에 입력된 문서에는 보통 두가지 종류의 오류가 존재한다. 하나는 집필자가 범하는 오류인 문법적 오류와 상식이나 지식이 부족해서 빚어지는 것들이다. 이러한 오류는 문서에 대한 교정자들이 문서의 내용을 완전히 이해하여야 고칠 수 있는 의미 오류와 어절의 수준에서 맞춤법에 어긋나는 맞춤법 오류가 있다. 다른 하나는 원 문서를 컴퓨터에 입력되는 도중에 발생하는 오류로서 키보드의 타이핑 오류나 문자인식, 음성인식의 처리 과정에서 발생하는 미인식, 오인식 오류이다.

오류 유형	유형별 합계
띄어쓰기 오류	808
어미와 조사의 오용	689
적절하지 못한 낱말	379
맞춤법 오류	132
외래어의 한글 표기 오류	69
한자 표기 오류	16

그림 4-1 중, 고 교과서 오류 조사표 (일부)

그림 4-1는 일반 출판물보다 오류가 적은 교과서들을 자료로 하여 얻은 결과[미승우90]를 일부만 나타낸 것이다. 여기에서 교과서에 나타난 오류 들을 소개하는 이유는 교과서 교열에는 대체로 맞춤법도 알고 교열 경력이 많은 사람들이 참여하고 있는데, 그러한 사람들이 저지른 오류라면 다른 사람들도 저지르기 쉬운 것들이라고 생각했기 때문이다. 그러나 본 시스템이 관심을 가지고 해결하려는 것은 문장이나 문서내의 의미적인 관계의 부정확성에 기인한 오류보다는 단지 문서내의 맞춤법 오류에 국한하였으므로 관심을 가지고 처리하는 범위는 어절내의 맞춤법 오류인 띄어쓰기 오류, 조사와 어미의 오용, 단어 구성의 오류와 입력시 발생하는 오류(mis-typed) 등이다.

어절내의 오류가 있는 경우, 그 어절은 다음의 유형중에 어느 하나에 속하게 된다.

- 표준어 오류 : 표준어가 아닌 단어를 사용한 경우.

  룡궁을 가는 -> 용궁을 가는, 띠어쓰기를 -> 띄어쓰기를

- 띄붙 오류 : 어절내에 독립적인 두개 이상의 어절이 존재하는 경우.

  잘사는 -> 잘 사는, 할수가 -> 할 수가

- 붙띄오류 : 어절이 형식 형태소로만 구성되어서 독립된 어절로 여겨지지 않는 경우

  먹 으려고 -> 먹으려고, 학교 를 -> 학교를

- 조사/어미 오류 : 어절내의 형식 형태소가 실질 형태소와 서로 어울리지 아니하는 경우.

학교을 -> 학교를, 알맞는 -> 알맞은

- 철자 오류 : 어절내의 어느 부분이 사전에 존재하지 아니하는 경우.

고등학고에서는 -> 고등학교에서는, ㄱ을하늘을 -> 가을하늘을

위의 유형을 도식화하여 정리하면 그림 4-2와 같다. 이 오류에 대한 분류는 발생의 원인과 유형을 위주로 분류한 것인데, 오류의 어휘 분석이나 문법적 속성에 따라 분류[강승식90]를 하기보다는 오류의 발견과 교정 방법에 중심을 두어 분류하였다.

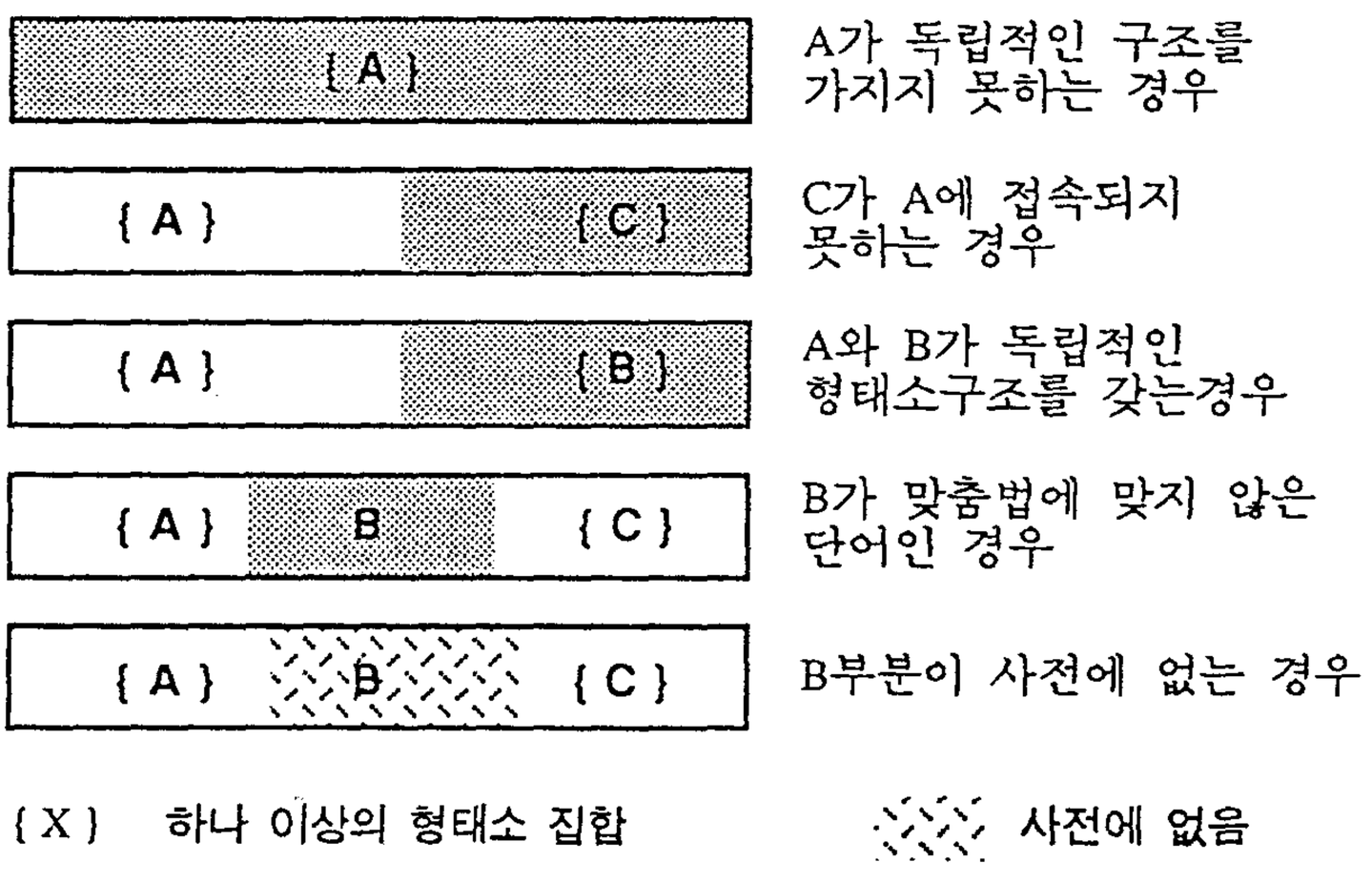


그림 4-2 어절에서 발생하는 오류의 유형

4.2 오류 유형과 처리 방법

이 장에서는 오류 유형에 대하여 각각을 정리하고 본 시스템에서 처리 하고자 하는 범위에 관하여 설명한다.

4.2.1 철자 오류

문서의 입력 과정에서 실수에 의해 발생되는 철자 오류는 어절을 구성하는 형태소 중에 철자의 일부분이 추가되거나 삭제, 변경되어 그 형태를 잃어버린 것이다. 그러나 의도하지 않은 실수에 의해 형태소가 변경되었지만 다른 형태소로 인정이 되는 철자인 경우는 철자 오류라고 할 수 없다. 그림 4-3의 철자 오류의 분포는 교정을 하지 않은 실제의 입력 문서를 대상으로 하여 조사한 결과이다. 철자 오류의 특징은 키보드로 문서를 입력하는 조작 과정에서 발생하며 대개 한 글자가 잘못되는 경우가 많다 영어와 비슷하게 한 글자가 추가되거나 삭제, 변경되어 발생되는 오류가 주류를 이루고, 두 자 이상의 오류는 거의 발견되지 않는다. 또한 특이한 점은 한 자소가 삭제되거나 추가된 경우에 미완성 자소가 자주 발견되었다.

유형	발견 횟수	비율	미완성 자소
1자소 삭제	80 / 203	39.4 %	23 / 80
1자소 추가	59 / 203	29.1 %	17 / 59
1자소 변경	47 / 203	23.2 %	0 / 47
기타 삭제	14 / 203	6.9 %	10 / 14
기타 추가	2 / 203	0.9 %	0 / 2
두자 뒤바뀜	1 / 203	0.4 %	0 / 1

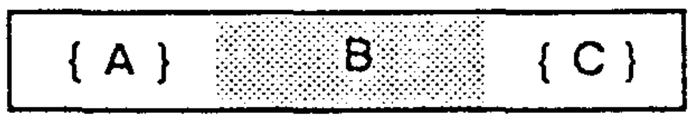
그림 4-3 철자 오류의 분포

철자오류의 교정은 사전 표제어가 아닌 부분을 사전의 표제어로 고치는 것으로, 한 자소를 삭제하거나 추가, 변경함으로써 이를 행하고 있다. 본 시스템에서는 입력 어절에 대하여 미완성 자소가 있는 가를 먼저 검사하여 발견되는 경우는 오류 검사기를 거치지 않고 미완성 자소에 대한 교정을 한다. 교정 과정에서 사전의 표제어로 가능한 단어는 보통 복수개로 나타나는데, 이는 키보드 상의 자소위치정보와 좌우접속정보를 이용하면 이들 후보의 수를 줄이는데 효과적이다. 본 시스템은 여러 후보 단어들을 가능성이 높은 순서로 나타내기 위하여 단어의 발견 빈도를 이용하였다.

4.2.2 표준어 오류

한글 맞춤법은 표준어를 소리대로 적되, 어법에 맞도록 함을 원칙으로 하고 있다. 즉, 어절을 구성하는 형태소는 소리대로 적은 표준어이면서 어법에 맞아야 한다. 이때 어법이란 뜻을 파악하기 쉽도록 각 형태소의 본모양을 밝히어 적는 것을 말한다. 본 시스템에서는 표준어 오류를 맞춤법에 맞지 않은 단어를 사용한 경우로 정의한다. 그러므로 표준어 오류는 어절을 구성하는 형태소나 단어가 어법에 맞지 아니하는 비표준어로 그 자체에 오류가 있는 것을 말한다. 이러한 오류는 반복적으로 나타나는 것이 특징이며, 특정한 음소를 고쳐본다든가 하는 방법으로는 해결할 수 없고, 사전에 오류 단어를 등록시키고 이에 대응하는 교정 단어를 두고, 이를 이용해야 한다.

사전은 표제어로 자주 발견되는 비표준어에 대한 정보를 가지고 있다. 형태소 해석 과정에서 형태소는 사전의 표제어와 대응되는데, 표제어가 비표준어일 경우



B가 맞춤법에 맞지 않은 단어인 경우

예) 용궁 + { 에서 + 는 } -> 용궁에서는
{ 안 } + 사돈 + { 의 } -> 안사돈의

그림 4.4 표준어 오류의 유형

사전에 지시된 같은 의미의 표준어로 대치한다. 하지만 어절 내의 실질 형태소가 교체되면 바로 뒤의 형식 형태소에 영향을 미치므로 실질 형태소와 형식 형태소와의 관계를 고려해서 교체어를 사전에 담고 형식 형태소의 교체도 실시해야 한다. 이러한 과정은 형태소 해석시에 처리가 가능하므로 표준어 오류를 범한 어절이 다른 오류를 포함했다면 이를 뒤에서 처리한다. 표준어 오류의 교정은 문서에서 일정 단어를 다른 단어로 대치하는 응용에도 사용되어질 수 있다. 이 때는 사용자의 요청에서 사전에 원래의 단어와 대치될 목적 단어를 임시적으로 기록하면 된다.

4.2.3 띄어쓰기 오류

단어는 독립적으로 쓰이는 말의 단위이기 때문에, 글은 단어를 단위로 하여 띄어쓰는 것이 가장 합리적인 방식이다[문교부90]. 형태소는 단어의 기초 단위가 되는 요소인 실질 형태소와 접사나 어미, 조사처럼 실질 형태소에 결합하여 보조적 의미를 덧붙이거나 문법적 관계를 표시하는 요소인 형식 형태소로 나뉜다. 그리고 형태소들의 적절한 모임으로 띄어쓰기의 단위인 어절이 구성된다. 띄어쓰기 오류는

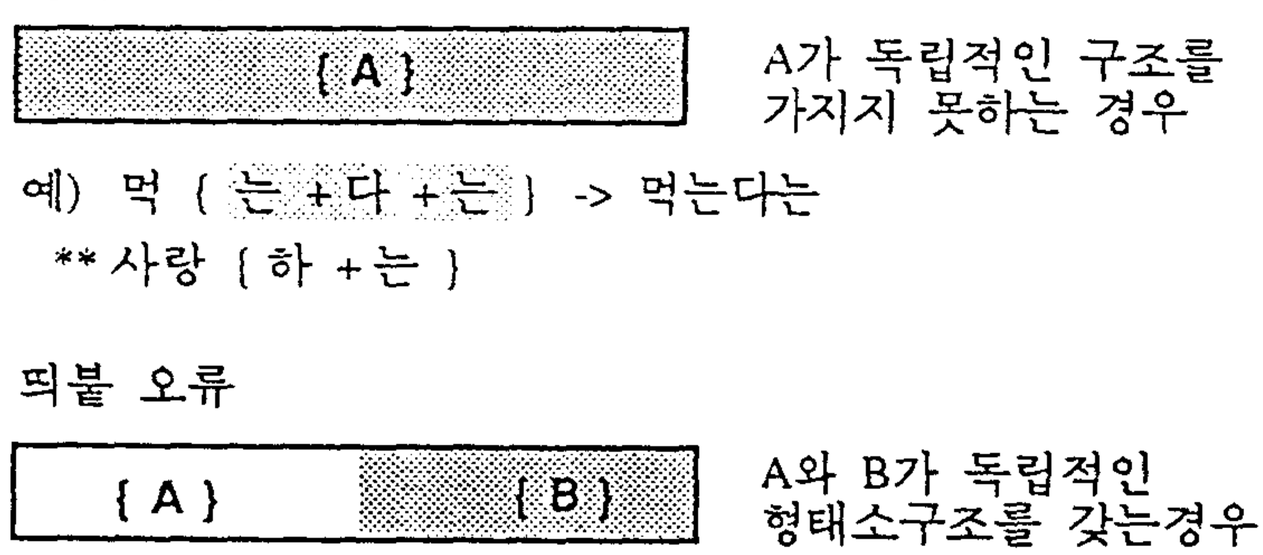


그림 4-5 띄어쓰기 오류의 유형

다음과 같은 두가지가 있다[강재우90].

띄붙 오류의 경우는 하나의 오류 어절에 여러 어절이 붙어 있는 경우이므로 적당한 곳을 띄어야 한다. 그러나 붙띄 오류는 하나의 형태소 구조가 두개 이상의 어절로 띄어져서 발생한 오류이므로 근접한 어절에 붙여 주어야 한다. 띄어쓰기 오류는 어절내에 형태소 구조가 넘치거나 모자라는 경우에 발생한다. 우선, 띄붙 오류는 하나의 어절에 둘 이상의 독립적인 형태소 구조를 형성하기 때문에 오류 어절을 양방향에서 해석하면서 두 구조가 만나는 지점을 추적하여 띄어 놓으면 된다. 예를 들면, 그림 4-5에서 "사랑하는사람의"란 어절은 왼쪽에서 해석하여 "사랑하는"이란 독립적인 형태소 구조를 갖는 어절을 발견할 수 있고, 마찬가지로 오른쪽에서 "사람의"란 어절을 발견할 수 있으므로 그들의 만나는 지점을 띄어주면 된다. 이 때 오른쪽 부분이 완벽하지 못한 구조를 가지고 있다면 다음의 어절과

어절과 붙여 보는 시도를 하게 된다. 그러나 붙띄 오류는 하나의 어절을 2개 이상의 어절로 띄어 쓰는 오류로 어절의 구조가 독립적이지 못하다. 따라서 전후의 어절들에 대하여 붙여 써야 하는 가를 검사해야 한다. 앞의 어절의 정보를 보관하기 위해서는 임시 보관 장소인 버퍼의 사용이 필요하다.

4.2.4 조사/어미 오류

한글 맞춤법은 체언과 조사, 어간과 어미를 구별하여 실질 형태소의 형태를 파악하기 쉽도록 되어 있다[문교부90]. 어절은 실질 형태소와 형식 형태소가 일정한 문법적 구조를 갖고 있어야 하는데, 발음의 문제나 저자의 무지로 인하여 형식 형태소가 실질 형태소와 어울리지 않는 경우가 발생한다. 이 때, 어절의 실질적 의미를 나타내는 것은 실질 형태소인 체언과 어간이므로 어울리지 않는 체언과 조사, 어간과 어미가 발견된다면 이것은 형식 형태소가 잘못되었다고 봐야 한다. 그러므로 본 시스템에서는 조사/어미 오류를 어절내의 두 형태소 구조중에 오른쪽의 구조가 독립적이지 못하며 단지 접속이 불가능한 경우를 말한다.

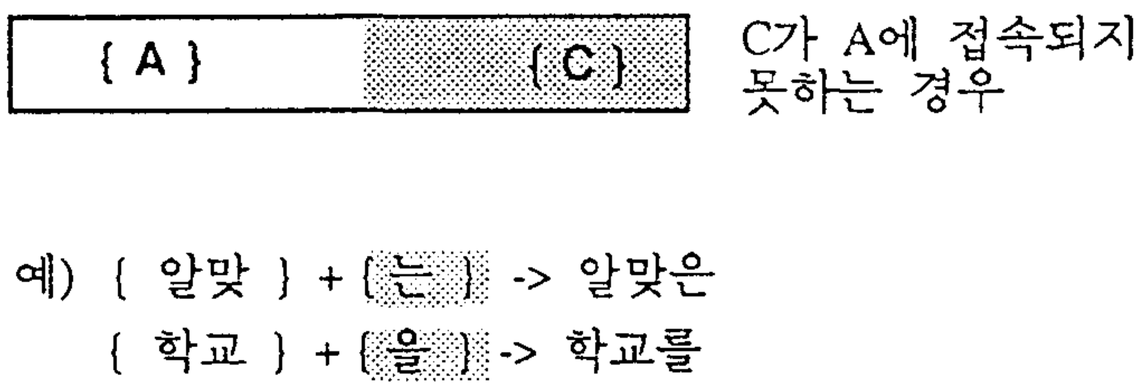


그림 4-6 조사/어미 오류의 유형

어절 내의 모든 형태소가 사전의 표제어이면서 서로의 접속이 불가능하다면

형식 형태소 중에서도 조사나 어미가 잘못 사용되었다고 판단할 수 있다. 이를 해결하기 위해서는 사전에 조사와 어미에 대한 수정 규칙을 둘 필요가 있다. 이 규칙의 조건에 따라 다른 것으로 교체하여 오류를 교정한다.

4.3 시스템이 범하는 오류

시스템이 오류의 발견과 교정시에 범할 수 있는 오류의 종류는 다음과 같다.

4.3.1 오류 검사시 발생하는 오류

입력 어절에 대한 오류 여부는 형태소 해석이 올바르게 되는 가에 대한 대답과 일치한다. 즉 오류의 검사시에 시스템이 범하는 오류는 다음과 같다.

제 1형 오류 : 올바른 어절을 틀린 어절이라고 판단하는 경우
제 2형 오류 : 틀린 어절을 올바른 어절이라고 판단하는 경우

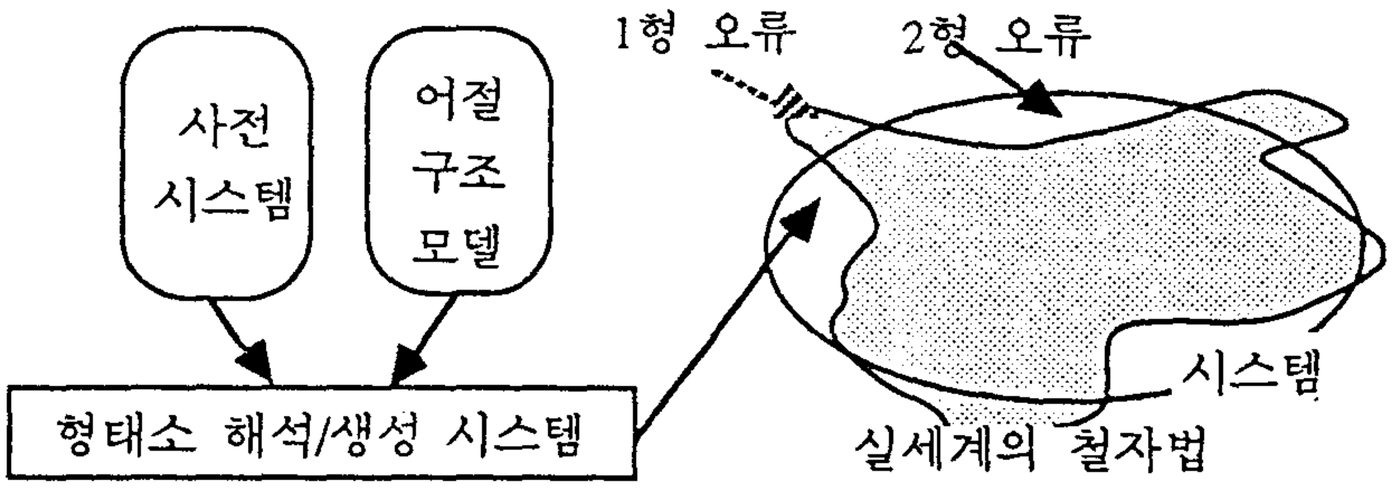


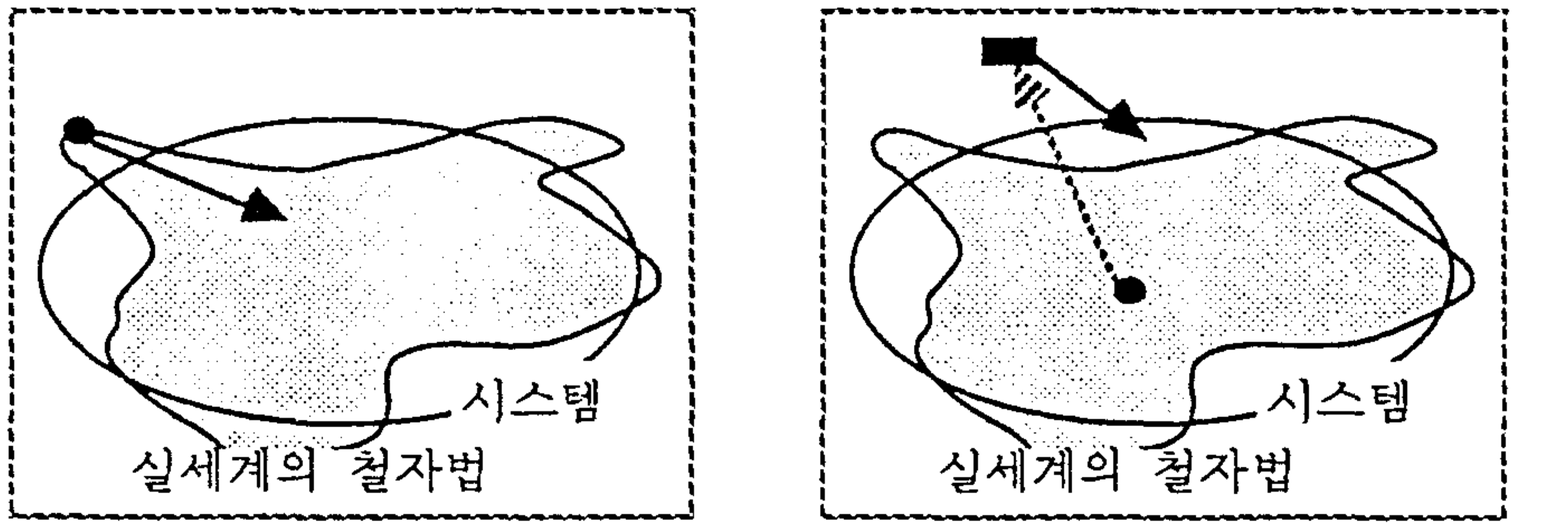
그림 4-7 오류 검사시 발생하는 시스템 오류

위의 그림에서 제 1형오류는 시스템이 맞춤법에 합당한 어절을 틀렸다고 취급하는 오류로서 사전에 등록되지 아니한 단어를 포함하거나 좌우접속정보가

틀리게 기입된 경우이다. 제 2형오류는 시스템이 맞춤법에 합당하지 아니한 어절을 맞았다고 취급하는 오류로서 사전의 내용에 잘못이 있는 경우이거나 좌우접속정보 체계에 문제가 있는 경우이다. 제 1형 오류의 감소는 제 2형 오류의 증가를, 제 2형 오류의 감소는 제 1형 오류의 증가를 가져 올 수 있다[강재우90]. 철자 검사의 입장에서는 제 2형 오류가 사용자에게 숨겨지므로 제 2형 오류를 줄이는 것이 실용적인 시스템을 만드는데 중요하다.

4.3.2 오류 교정시에 발생하는 오류

오류로 판명된 어절은 시스템이 받아 들일 수 있는 형태의 것으로 복구를 해야 하는데 이 때 범하는 오류는 다음과 같다.



제 3형 오류 : 올바른 어절을 잘못으로 인식하고 교정하는 경우

제 4형 오류 : 오류 어절을 실제의 어절로 복구하지 못하는 경우

그림 4-8 오류 교정시 발생하는 시스템 오류

제 3형오류는 시스템 오류 제 2형을 교정하려고 하였을때 발생하는 것으로써 아무리 치밀한 교정을 하더라도 피할 수 없는 오류이다. 제 4형오류는 사용자의

실수를 교정하는 과정에서 사용자의 의도와는 별개로 다른 답을 만들어 내는 것이다. 이러한 오류는 본 시스템이 형태소 해석을 통하여 처리하기 때문이며 완전한 처리를 위하여는 구문 해석과 의미 해석등의 처리가 불가피하다.

4.4 양방향접근법

입력 어절에 오류가 있는 것이 형태소 분석기에 의해 판명되면 오류의 위치와 종류를 알아내는 작업이 필요하다. 예를들면 오른쪽에서 왼쪽으로 해석하는 시스템의 경우, 어절 W에 M1, M2, M3, M4의 형태소 구조가 있다고 하자. 만일 M1, M2 부분이 사전 미등록어이거나 접속이 불가능하여 더 이상의 해석을 할 수 없을때 형태소 해석기는 M3, M4의 발견에서 그치므로 입력 어절 W는 오류가 있다고 판정한다. 그러므로 M1, M2부분에서 발생한 구체적인 오류의 종류와 위치를 알아내기 위해서 추가적인 해석 즉, 왼쪽에서 오른쪽으로의 해석을 해보자는 것이다.

{A}는 왼쪽에서 오른쪽으로 해석한 결과를 의미하고 {C}는 오른쪽에서 왼쪽으로 해석한 결과를 의미한다고 하자. 이때 {A}와 {C}는 각각 접속 가능한 어절의 배열이고 {A}의 첫 형태소는 어절의 처음에 쓰일 수 있는 형태소 이고 {C}의 마지막 형태소는 어절의 마지막 형태소로 쓰일 수 있는 경우가 된다. 또 B는 어절내의 {A}와 {C}사이의 부분이라고 하자. 5.2절에서 정리한 각 오류의 유형별로 오류의 위치와 종류를 파악하는 규칙과 처리 방법은 다음과 같다.

○ 철자 오류 : {A}와 {C}가 만나지 않는 부분 B가 어절 내에 존재하는 경우. B부분의 각 자소를 바꾸어 사전을 검색하여 사전 표제어가 되는 경우 {A}와 {C}에 모두 접속 가능하면 찾은 것으로 한다.

○ 표준어 오류 : {A}와 {C} 사이의 B가 비표준어로써 사전에 등록되어 있고 B'이라는 후보가 있는 경우. B'가 {C} 부분에 접속이 가능한 가를 검사하고 불가능할 경우 {C}의 시작 형태소가 조사/어미인 경우에 조사/어미 오류의 처리를 한다.

○ 띄붙 오류 : {A}와 {C}가 각각 독립적으로 사용 가능한 경우. 즉 {A}의 마지막 형태소가 어절의 마지막 형태소로 적당하고 {C}의 첫 형태소가 어절의 첫 형태소로 적당한 경우. {A}와 {C}를 띄어 준다.

○ 붙띄 오류 : {C}부분이 어절 전체이면서 {C}의 첫 형태소가 어절의 첫 형태소로 부적합한 경우. {C}와 앞 어절의 결합성을 조사하여 결합이 가능한 경우에 앞 어절에 붙이어 준다.

○ 조사/어미 오류 : {A}와 {C}가 접속 불가능하고 {C}의 첫 형태소가 어절의 첫 형태소로 부적합한 경우. {C}의 첫 형태소가 조사나 어미 등일 경우이므로 사전에서 대체어를 찾아 {A}와 접속이 가능한 것들에서 {C}의 나머지 부분과 결합이 가능한 경우 대체시킨다.

양방향에서 행하는 형태소 해석의 결과는 가능한 모든 경우의 가능성을 유지하여야 한다. 형태소 해석의 중간 결과는 스택에 그 구조가 저장되는데,

양방향 해석 결과를 어절 내의 위치에 맞게 정렬하는 처리가 필요하다. 이러한 중간 결과들의 정렬은 스택에서의 pop과 해석후의 처리를 통하여 해결된다. 즉, 현재 확인된 형태소를 pop시 알맞는 위치에 삽입한다. 양방향 해석이 끝나게 되면 각각의 해석에서 결과를 유지해야 하는데, 이 때 어절내의 일정 위치에 중복된 하나 이상의 해석 결과가 있을 수 있으므로 이들을 하나로 통합시키는 작업을 하여야 한다.

양방향 분석을 위해서는 어절의 정형화된 틀에 어절을 일치시키는 방법으로는 처리가 불가능하고 어절 내의 형태소 각각에 대한 접속성이 기본이 되는 해석 모델이 필요하다. 좌우접속분류는 기본적으로는 어절내의 i번째 형태소와 i+1번째 형태소의 결합 가능성을 가지는 분류 체계이므로 역으로 i+1번째와 i번째의 형태소가 접속 가능성을 판별할 수 있다. 그러므로 좌우 접속분류에 의한 어절 구조의 모델에서는 양방향 접근이 가능하다.

V. 시스템의 설계 및 구현

5.1 시스템의 개요

우리는 맞춤법 오류를 검사하고 교정하는 시스템을 그림 5-1과 같이 좌우접속정보표와 사전, 오류 검사기, 오류 교정기로 구성한다. 시스템의 내부에서는 한글 처리에 있어서 음소 단위의 처리가 가능하고 각 음절의 길이가 일정한 3바이트 코드 체계를 사용하였다. 그러나 이 코드 체계는 시스템의 내부에서만 사용이 되며, 입출력시 사용자의 여러 코드 체계를 받아들여 변환하는 방법으로 사용 터미날에 독립성을 부여했다.

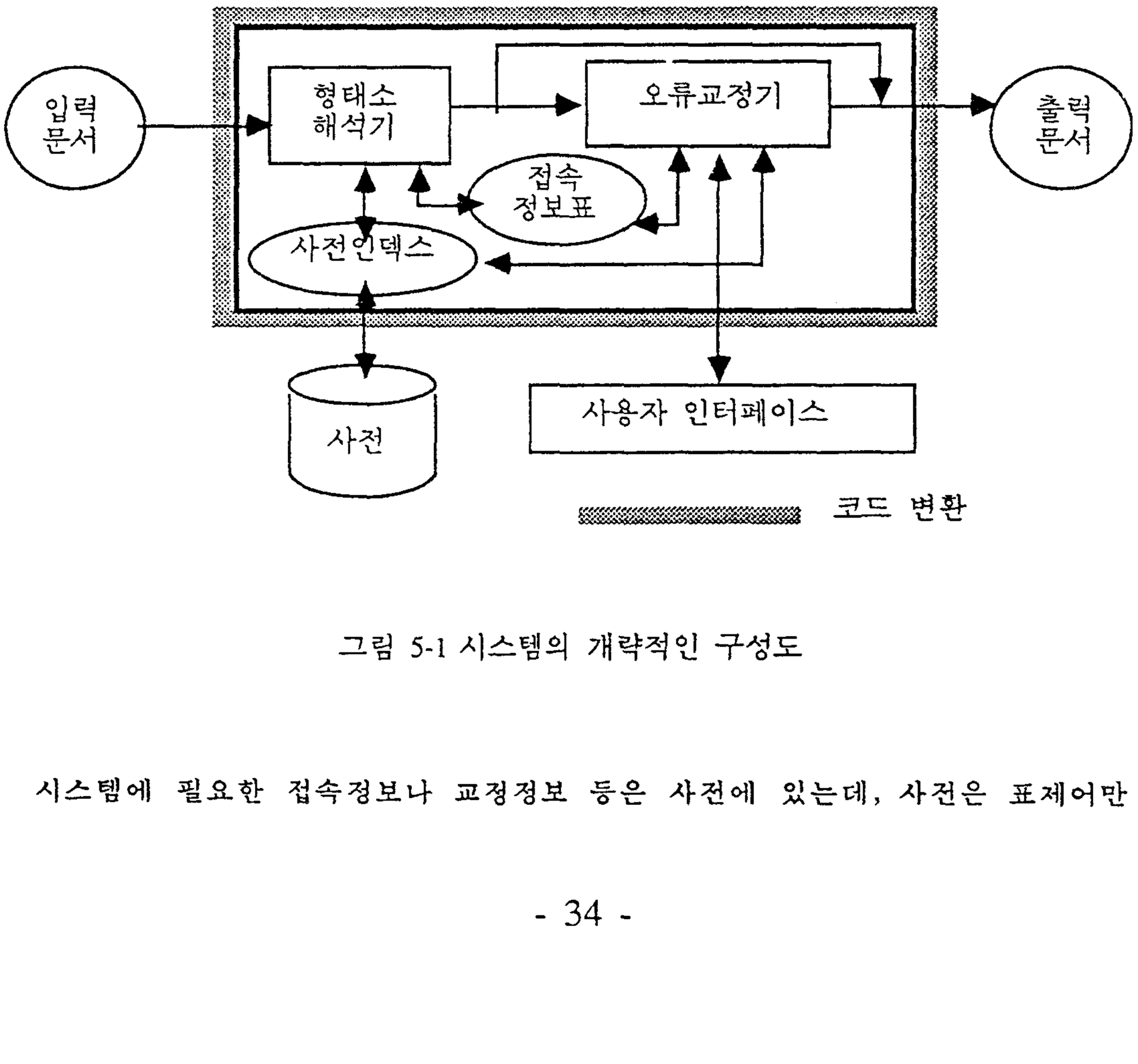


그림 5-1 시스템의 개략적인 구성도

시스템에 필요한 접속정보나 교정정보 등은 사전에 있는데, 사전은 표제어만

시스템 내부에 저장되고, 기타 정보들은 외부 사전에 저장된다. 단위 어절이 시스템 내부로 들어오면 사전을 이용하여 형태소 해석을 행하고, 해석이 불가능한 경우에는 오류 교정기에 해석 결과와 함께 보내져 교정을 한다. 여기에서 해석 결과는 여러개가 나올 수 있으므로 우리는 두 가지의 출력 방식을 채택하였다. 즉, 가능한 교정 후보들을 생성하여 사용자에게 선택을 의뢰하는 대화 모드(Interactive mode)와 모든 후보를 화일에 출력하는 일괄처리 모드(Batch mode)를 함께 제공한다.

5.2 사전 시스템

사전은 실용적인 시스템을 만드는데 중요한 역할을 한다. 사전의 표제어는 형태소와 동일한 의미를 갖는다. 본 시스템에서는 표준어가 아닌 형태소에 대해서도 표제어로 삼고, 대신 올바른 표준어 단어로의 매핑(mapping)을 하도록 하였다. 사전에는 발견과 교정을 위한 대부분의 정보를 갖고 있는데, 이 정보의 정확도가 시스템의 성능을 가늠하는 기준이 될 수 있다. 본 시스템에서는 앞 장에서 정의한 오류의 발견과 교정을 위하여 표제어, 표제어 성격, 좌우접속정보, 수정규칙의 정보를 하나의 사전 정보로 삼는다.

5.2.1 사전의 종류

본 시스템에서는 사용자가 자신의 고유한 사전을 보유하면서도 기본적인 어휘에 관해서는 고려하지 않아도 되도록 하기 위하여 시스템 제공 사전과 사용자 정의 사전을 가정한다. 시스템이 제공하는 사전을 마스터 사전이라 정의한다. 마스터 사전에는 기본적인 어휘인 조사, 어미, 보조어간, 접사, 관형사, 대명사, 수사,

용언 어간 등으로 구성하여 이들의 접속 정보를 미리 작성하였다. 사용자 정의 사전은 명사만으로 구성되도록 제한을 두는 대신에 표제어에 대한 접속 정보를 시스템이 자동적으로 생성하기 위하여 제한된 몇 가지의 표시를 하도록 하였다.

5.2.2 사전의 내부 구조

빠른 속도의 처리와 철자 교정을 위하여 사전 표제어는 내부에 TRIE의 형태로 적재된다. 이때 TRIE의 노드는 초성, 중성, 종성이 각각 다른 의미를 가지므로 초성 노드와 중성 노드, 종성 노드는 서로 다른 형태를 갖는다. 시스템의 내부에는 사전의 표제어만을 적재하고 실제의 정보를 가리키기는 포인터를 유지한다. 이 때 표제어가 비표준어인 경우에는 교체될 단어에 대한 정보를 보관하는 것이 유리하다. 즉, 사용자의 요구에 따라 문서 전체에서 어떤 단어를 다른 단어로 바꾸려 한다면 사전 인덱스에 이를 임시로 기억시켜 놓으면 추가적인 작업 없이도 쉽게 처리될 수 있다. 본 시스템에서 사용하는 사전 시스템의 구조는 다음의 그림 5-2와 같다.

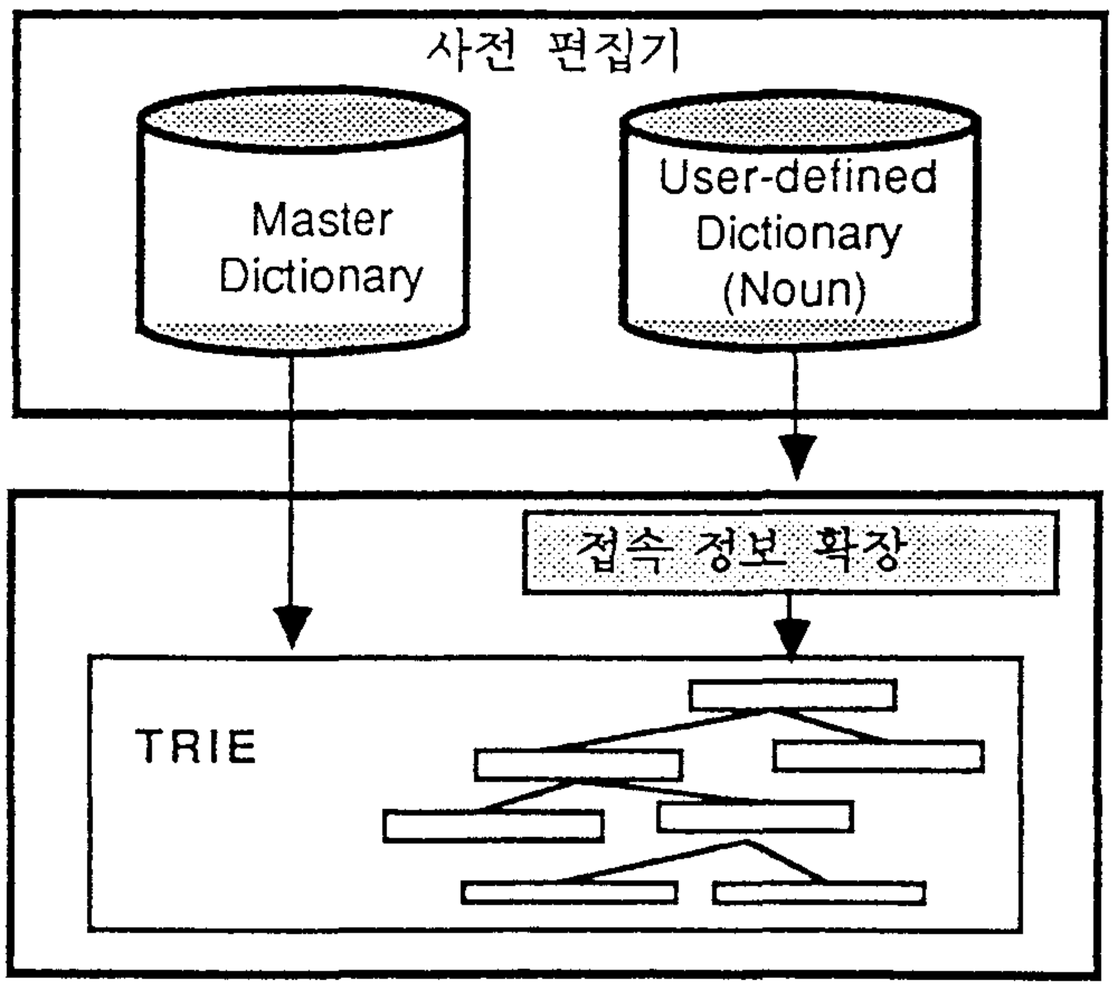


그림 5-2 사전 시스템의 개략적인 구조

5.3 코드 변환 라이브러리

사용자의 코드 체계에 따라서 시스템을 구성한다면 사용자 인터페이스, 사전 등에 따라 여러가지 중복된 작업을 하여야 한다. 외부의 코드 체계와 독립적인 코드 체계를 사용하고 인터페이스가 필요한 각 부분에 코드 변환을 한다면 하나의 내부 처리로 여러 코드 체계에 대한 적절한 처리를 할 수 있다. 한글은 글자 단위로 쓰는 언어이므로 한글의 자소를 글자로부터 알아 내기 위해서는 시스템은 초성, 중성, 종성을 글자로부터 분리해낼 수 있어야 한다. 이 때 한글의 글자를 마치 영어의 알파벳처럼 자소의 일차원 배열로 생각하면 각 글자를 인식하기 위해서 매번 문자열의 처음부터 한글 오토마타를 돌려야 한다. 그래서 본 시스템에서는 내부 구조로 3바이트 코드 체계를 지원하기 위한 라이브러리를 구성하였다. 3바이트 체계는 글자 단위의 인식이 가능하기 때문에 한글에 대한 글자단위의 처리와 더불어 오류 교정시 어절내의 정확한 위치 파악을 가능하게 해준다.

3바이트 코드체계를 지원하기 위한 라이브러리의 function들은 다음과 같다.

- 각 코드 체계간의 문자열 변환
- 3바이트에서 중성코드값과 초성코드값간의 변환
- 3바이트 문자열 복제, 비교
- 3바이트 문자열 임의 구간의 변경, 복제, 첨가
- 입출력 function

현재 시스템에서 제공하고 있는 외부 코드 체계는 2바이트 완성형과 N바이트 풀어쓰기 코드체계이다. 이들 외부 코드와의 변환은 사용자 입력 문서에 대한 입력과 출력, 사용자 인터페이스가 한 체계를 이루고 사전 시스템 인터페이스가 별개의 체계를 이루도록 하였다.

5.4 오류 검사기

오류 검사기는 어절내의 오류가 있는가를 확인하기 위하여 형태소 해석을 통해 어절의 형태소 구조를 분석하려는 것이다. 이 때, 하나 이상의 분석이 가능한 경우에 형태소 해석기 자체는 어절 내의 형태소 구조에 대한 애매성을 해결할 수 있는 방법이 없다. 그러므로 만일 한 어절에 여러 가지 형태소 구조가 존재하더라도 형태소 분석의 방법에 의해서 그 중 하나의 분석이 성공하면, 더 이상의 분석 없이도 그 어절이 형태론적으로 올바른 것이라고 생각할 수 있다. 이것은 형태소 해석 수준으로는 어절이 문장에서 어떠한 문법적/의미적 위치를 갖는가를 알 수 없다는 한계에 기인 한다. 그림 5-3은 오류 검사기의 대략적인 구성도이다.

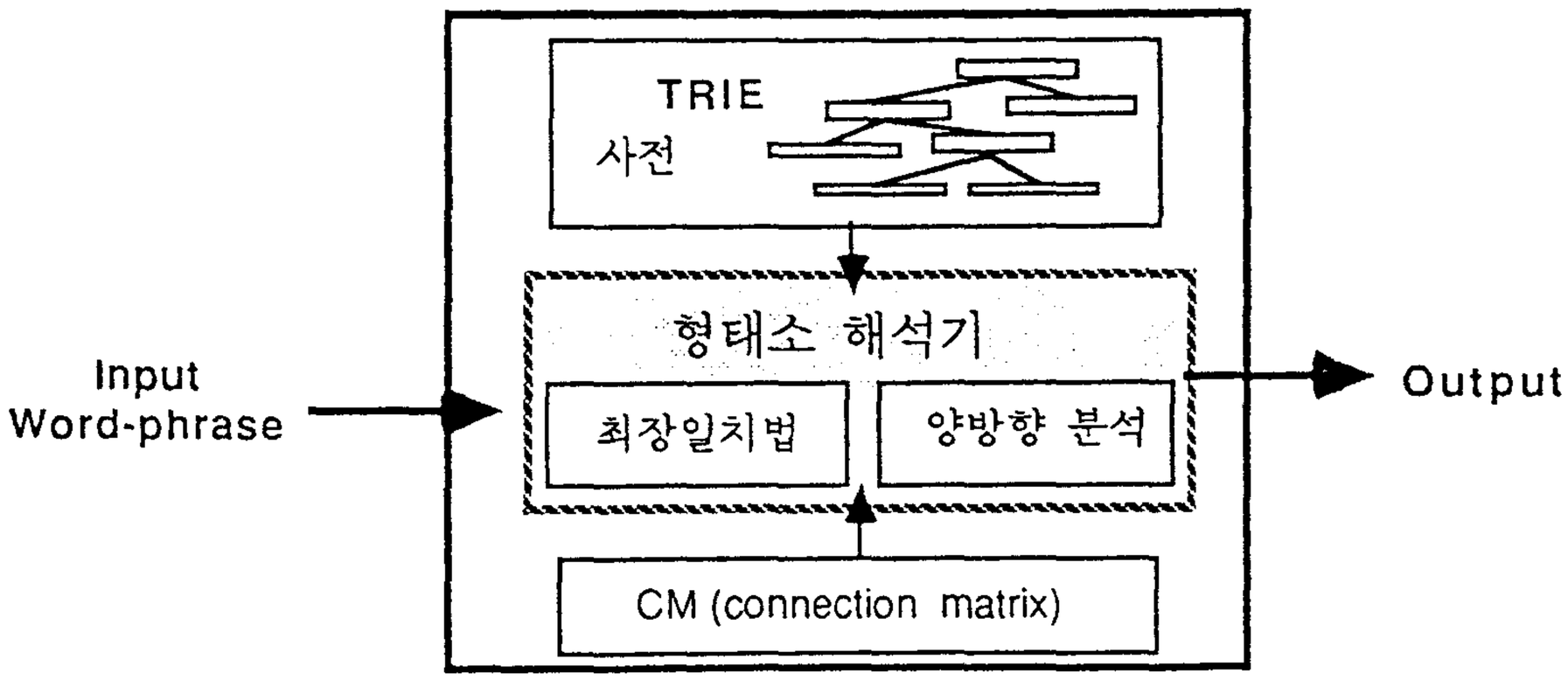


그림 5-3 오류 검사기의 개략적인 구조

본 보고서에서 구현한 오류 검사를 위한 형태소 해석기의 특성은 다음과 같다.

- 사전 구조에 대한 투명성(transparency)

본 시스템에서 주장하는 사전과의 인터페이스는 사전의 내부 구조에 대한 자율성이다. 즉, 사전 구조에 독립적으로 동작함으로써 주어진 한글 스트링에 대한 사전 엔트리가 존재하는가와 그에 대한 정보는 무엇인가만을 전달하는 사전 인터페이스를 제공하자는 것이다. 이러한 접근 방법은 사전 구조에 맞춰진 형태소 해석 알고리즘을 사용하는 것보다 효율성은 다소 떨어지지만 이식성과 확장성을 유지할 수 있다.

- 범주별 처리 (category driven)

어절 내의 형태소들에 대한 직접적인 발견 즉, 해석시 형태소 각각에 대한 처리가 아닌 대신에, 그 형태소의 접속범주에 의한 처리와 발견을 한다. 즉 시스템에서 형태소에게 요구하는 정보는 어휘로서의 형태소가 아니고 그 형태소가 가진 접속 분류 범주이다. 이 방법은 형태소가 가지고 있는 독특한 접속 양상들을 개별적으로 처리하는 것을 피함으로써 한국어 어절에 대한 단순화된 모델에서의 처리가 가능하다.

- 스택을 이용한 알고리즘 (stack based algorithm)

어절에서 형태소를 분리를 하기 위해서는 어절을 개념적으로 두 부분으로 나누어 생각할 수 있다. 한 부분은, 어절 내에서 현재까지 해석이 되어진 부분이고 다른 하나는 아직 해석이 이루어지지 않은 부분이다. 아직 해석되지 않은 부분은

현재 해석된 부분의 연장선상에서 해석되어야 한다. 따라서 현재의 해석 결과를 저장할 필요성이 생기는데, 그런 구조로 가장 적당한 것이 스택이다. 스택에는 항상 현재까지 올바르게 해석되어진 중간 결과가 기록된다. 스택에 새로운 형태소가 들어 갈 때는 이전의 마지막 형태소 정보인 top과 접속성을 조사하여 자신의 접속 정보를 조정하여야 한다. 즉, 스택에 push되는 것은 이전 상태의 스택에 새로 발견된 형태소가 접속 가능한 경우이고, 스택에서 pop되는 것은 미해석 부분이 더이상 해석 가능하지 못하는 경우이다.

- 최장일치법 적용 (longest morpheme first)

형태소의 추출은 가장 긴 것을 우선으로 하는 최장일치법을 사용한다. 표준어 오류의 교정은 사전에 비표준어에 대하여 교체될 단어에 대한 정보가 있는 것이므로 형태소 처리 단계에서 발견하여 처리를 해주는 것이 효율적이기 때문이다. 이 때 최장일치법을 사용하지 않으면 비표준어 어휘가 잘게 나뉘어 분석될 가능성이 있다. 최단일치법을 사용한다면 사용자가 의도하는 오류 단어는 작은 단위의 해석으로 인해 무시될 수 있다.

- 우에서 좌로의 해석 (right to left analysis)

본 시스템이 우에서 좌로의 해석 즉, 형식 형태소인 조사, 어미, 접미사 등을 우선 처리하는 이유는 표준어 오류의 발견과 교정시 수반되는 접사와의 불일치를 쉽게 제거하기 위함이다. 즉 어절의 의미를 갖는 실질 형태소가 바뀌게 될 때 그에 대한 현재까지의 해석인 형식 형태소의 교체를 한다. 예를 들면, "한국어"를 "한글"로 바꾸는 것을 표준어 오류로 처리했다고 하자. 단순한 단어의 교체는

다음과 같이 "한국어를"을 "한글를"로 바꾸어 놓게 된다. 이러한 불일치를 없애기 위해서는 형식 형태소인 어절의 오른쪽부분부터의 해석이 필요하다.

5.5 오류 교정기

오류 검사기에 의해서 어절에 오류가 있음이 판명되면 오류의 종류와 위치를 찾는 과정이 필요하다. 이 때 오류의 종류와 위치를 알아내기 위하여 어절의 오른쪽에서 왼쪽으로 분석을 하고, 왼쪽에서 오른쪽으로 분석을 하여 두 분석 내용을 조사하는 양방향 접근법을 사용한다. 오류의 발견과 교정을 위한 본 시스템의 전체적인 흐름도는 그림 5-4와 같다.

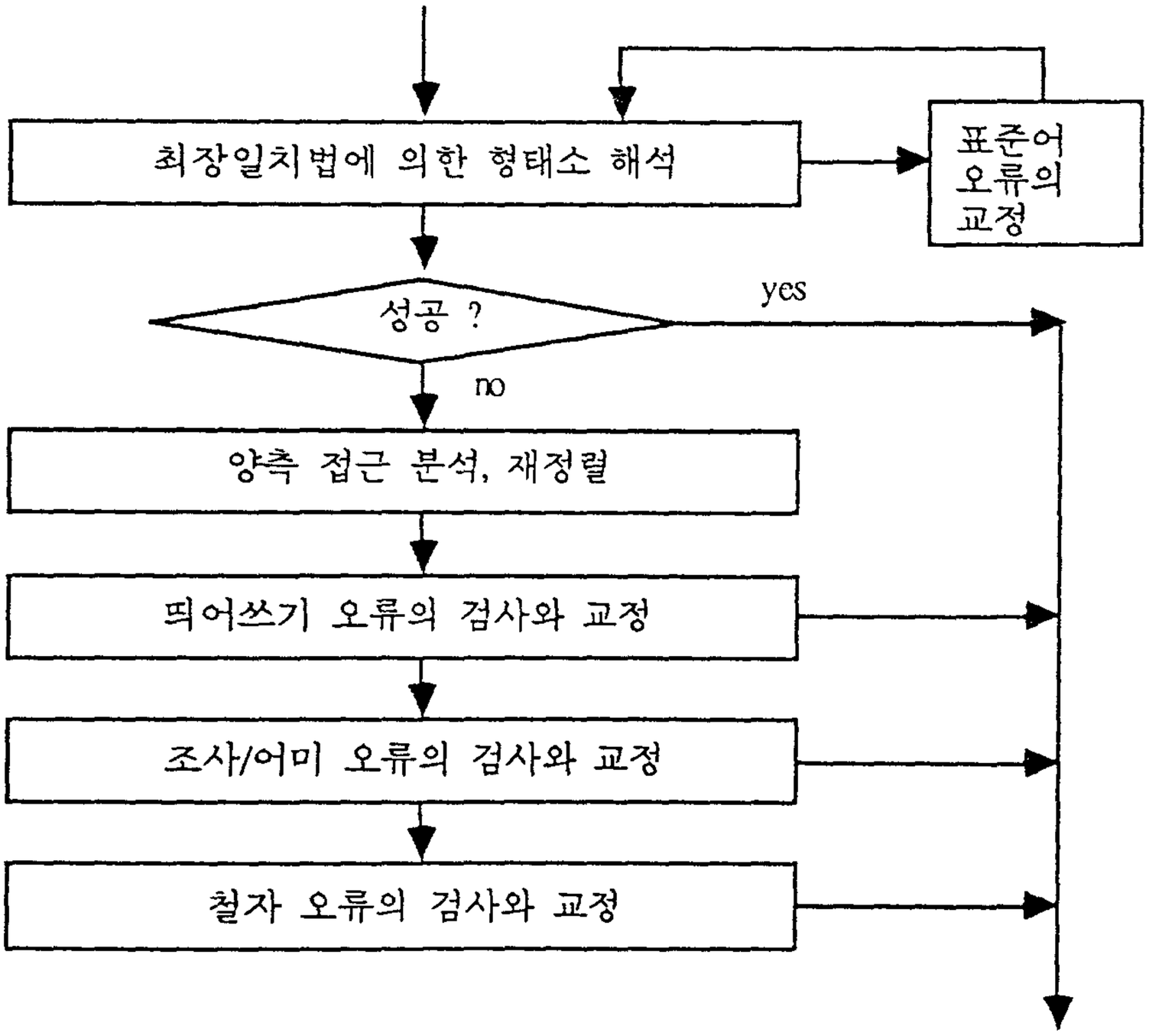


그림 5-4 오류 교정의 흐름도

입력 어절에 대하여 맨처음 발견하고 교정할 수 있는 오류는 표준어 오류이다. 표준어 오류는 오류의 검사를 위한 형태소 해석시 교정되며, 이 경우 오류의 검사와 교정을 다시 거치게 된다. 다음으로는 양방향접근분석에 의하여 정렬된 결과 리스트의 형태에 두 결과 리스트가 마주치는 지점을 가지고 있다면 띄어쓰기 오류이거나 조사/어미 오류일 가능성이 높으므로, 먼저 띄어쓰기 오류인가를 검사하고 다음으로 조사/어미 오류인가를 검사한다. 위의 두 오류는 서로 상이한 조건을 가지고 있으므로 처리의 우선 순서는 의미가 없지만, 띄어쓰기 오류의 발견 빈도가 조사/어미 오류보다 높으므로 띄어쓰기 오류를 먼저 검사하고 처리한다. 마지막으로 양방향 해석 결과 리스트가 서로 만나는 지점이 없다면 이는 사전에 나타나지 않은 단어라고 추정하여 철자 오류라고 가정하고 처리한다.

VI. 실험

본 시스템은 맞춤법 오류의 교정을 위하여, 오류의 검사를 위한 형태소 해석기와 발견된 오류 어절에 대한 오류 교정기의 두 부분을 거치게 된다. 그러므로 오류 검사기의 정확도와 오류 교정기의 복구성이 본 시스템의 성능을 결정하게 된다. 본 장에서는 오류 검사를 위한 형태소 해석기의 정확도를 시스템이 범하는 오류의 빈도와 사전 시스템에 독립적인 정확도의 측면에서 살펴본다. 또한, 오류 교정기의 성능을 알아 보기 위하여 시스템이 오류를 포함하고 있는 어절에 대하여 각 맞춤법 오류의 교정도를 실험하였다.

6.1 오류 검사기의 실험

오류 검사기는 오류가 있는 어절에 대하여 오류가 있다고 판단하고, 오류가 없는 어절에 대하여는 오류가 없다고 판단하는 것을 목적으로 한다. 그러므로 오류 검사기의 성능은 전체 입력 어절 중에서 얼마나 많은 어절에 대하여 바른 판단을 하는가에 대한 대답이라 할 수 있다. 본 보고서에서는 오류 검사기의 성능을 실험하기 위하여 다음과 같은 변수를 정의한다.

T : 입력 문서 내의 어절의 갯수

E1 : 입력 문서 중 맞는 어절을 오류로 판단한 횟수

E2 : 입력 문서 중 틀린 어절을 오류로 판단하지 못한 횟수

U : 입력 문서 중 미등록어를 가진 어절의 수

M : 입력 문서 중 의미상 오류이지만 문법적으로는 맞는 어절의 수

입력 문서에 따라서 시스템 오류의 수가 달라질 수 있으므로 입력문 의존적 성능과 입력문 독립적 성능을 정의한다. 아래의 함수는 오류 검사기의 성능을 평가하기 위한 평가 함수이다.

입력문 의존적 성능(Pd) :

$$Pd = \{ 1 - (E1 + E2) / T \} * 100$$

입력문 독립적 성능(Pi) :

$$Pi = \{ 1 - (E1 + E2 - U - M) / (T - U - M) \} * 100$$

즉, 입력문 의존적 성능은 전체 입력 어절에서 시스템 오류를 제외한 어절의 비율로 정의할 수 있다. 입력문 독립적 성능은 입력 어절 중에 미등록어를 포함한 오류를 제외한 경우의 성능이라고 생각할 수 있다. 여기에서 미등록어를 제외시킨 이유는 미등록어의 경우, 시스템보다는 입력 문서에 의존적으로 발생하기 때문이다.

위 평가 함수에 입각하여 본 보고서에서는 4개의 입력 문서에 대한 실험을 하였다. 그 결과는 다음과 같다.

입력 문서	T	E1	E2	Pd	U	M	Pi
논문 1	1158	12	2	98.79	5	1	99.30
매뉴얼 번역문	1361	26	5	97.72	6	2	98.31
논문 초록	1590	17	18	97.80	10	7	98.85
논문 2	1732	42	2	97.46	18	0	98.48
합계	5841	97	27	97.88	39	10	98.19

그림 6-1 오류 검사기의 실험 결과

본 실험은 총 5841어절에 대하여 7331개의 표제어를 갖는 사전으로 실험하였으며 대부분의 전문 용어는 사전에 포함시켰다. 위의 결과는 본 시스템이 97.9%의 확률로 맞는 어절을 맞았다고, 틀린 어절을 틀렸다고 판정할 수 있음을 말한다. 또한, 만약에 본 시스템이 가지고 있는 사전이 완벽하다고 가정한다면 그 확률은 98.2%가 된다. 실험을 통하여 위의 문서들에서 발생한 시스템 오류의 대부분은 중복되는 것을 알 수 있었다. 이것은 사전의 내용에 잘못이 있는 경우와 좌우접속정보의 분류 체계와 접속표에 문제가 있는 경우이다. 그러므로 자주 반복적으로 발견되는 시스템 오류의 해결을 위하여 좌우접속정보표와 사전의 내용을 보강할 필요가 있다.

6.2 오류 교정기의 실험

오류의 교정은 오류 검사기에 의하여 오류로 판명된 어절에 대하여 실시한다. 오류 교정기에서 시스템 오류에 의하여 맞는 어절을 틀린 것이라고 판명하는 경우에 교정기는 잘못된 결과를 만든다. 마찬가지로 오류 어절에 대하여 시스템이 맞는 어절로 판정한 경우에는 오류 교정을 할 수 없다. 그러므로 오류 교정기의 성능은 오류 검사기의 성능에 의존적이라 할 수 있다. 다음의 결과는 6.1절과 같은 입력 문서들에 대한 교정의 결과를 나타낸다.

위의 실험의 결과로 실험 대상 문서에는 띄어쓰기 오류가 가장 빈번하게 나타났으며, 첨단 분야의 논문에 대한 초록을 일차로 입력한 문서 3의 경우는 철자 오류가 많이 나타났다. 나머지 문서들은 학위 논문을 대상으로 한 것이어서 철자 오류의 발견 빈도가 낮았지만 보통 같은 종류의 오류가 반복되는 것을 발견할 수

입력 문서	T	띄어쓰기 오류			철자 오류	조사/어미 오류
		붙띄	띄붙	실패		
논문 1	1158	1	9	3	2	0
매뉴얼 번역문	1361	34	13	5	11	1
논문 초록	1590	11	12	7	34	2
논문 2	1732	8	8	22	2	1
합계	5841	54	42	37	49	4

그림 6-2 오류 교정기의 실험 결과

있었다. 실험을 하면서 같은 오류가 반복적으로 나타나는 것을 발견하였다. 이는 문서 집필자가 범한 것으로, 특이한 점은 몇 단어에 대하여 자주 오류를 범하는 것이다. 이것은 각 문서마다 고유하게 나타나는데 어떤 이는 띄어써야 할 특정한 어절들을 붙여 쓰고, 어떤 이는 붙여 써야 할 특정 어절을 반복적으로 사이띄기를 하는 등의 경우이다. 역시 특정 단어에 대한 반복적인 실수를 범하는 현상도 볼 수 있었다.

VII. 결 론

본 보고서에서는 과기처 첨단 요소 과제인 "한국어 철자 및 띄어쓰기 교정 시스템에 관한 연구" 과제의 1차년도 연구 수행 결과를 기술 하였다. 연구의 주제를 제시하고 연구의 방향과 그 접근 방법을 소개하였으며, 실제 시스템을 제작하였다. 연구의 주제에 따라 한국어 문장에서 나타나는 철자 오류와 띄어쓰기 오류의 검사와 교정의 방법에 관하여 기술하였고, 조사/어미 오류와 표준어 오류의 개념과 처리방법을 기술하였다. 사전의 내용은 아주 단순하게 어휘에 대한 접속정보와 교체 후보에 대한 정보만으로 구성할 수 있음을 보였다.

본 연구 과제의 1차년도 목표는 "형태소 해석 수준의 시스템"으로 형태소 해석기와 사전 시스템, 오류 검사와 교정기로 구성 된다. 입력 어절의 오류 검사를 위하여 형태소 해석 방법을 사용하였고 오류의 종류와 위치를 알아내기 위하여 어절의 양쪽 방향에서 형태소 해석을 하는 양방향접근법을 고안하였다. 맞춤법 오류로 띄어쓰기, 조사/어미, 표준어 오류를 정의하였고 입력시에 발생되는 오류를 철자 오류로 분류하여 처리하였다. 오류의 교정은 사전을 이용하여 해결하는 방법과 문법적 지식을 이용하여 어절의 구조를 분석함으로써 오류를 처리하는 방법을 병행하였다.

사전이 차지하는 주기억장치 공간을 줄이기 위한 방법으로 사전의 표제어만을 주기억장치에 두는 방법을 채택하였지만 확장성과 이식성을 위하여 사전 구조와 독립적으로 시스템이 동작하는 방법을 취하였다. 또 각종 코드체계를 갖는

터미날을 지원하기 위하여 내부 코드 체계를 3바이트로 채택하고 외부의 코드 체계를 독립적으로 처리 가능하도록 코드 변환 라이브러리를 작성하였다.

2차년도에서는 문서에 대한 해석 수준을 구문 해석 수준으로 하고 대량의 사전을 갖추는 연구를 수행할 계획이다. 1차년도의 개발 시스템이 UNIX상에서 형태소 해석의 수준으로 단어를 검사하는 수준의 시스템으로 구현되었지만, 2차년도는 개인용 컴퓨터에서 대화형식의 처리로서 문장내에서 단어의 쓰임새를 검사하고 교정하는 방법을 적용할 것이다.

참 고 문 헌

[강재우89] 강재우, 좌우접속정보표, 한국과학기술원, CS Lab. Memo, 1989.

[강재우90] 강재우, 접속정보를 이용한 한글 철자 및 띄어쓰기 검사기의 설계 및 구현, 한국과학기술원 석사학위논문, 1990.

[강승식90] 강승식, 김영택, "한국어 문장의 문법 오류에 관한 연구," 인지과학회 춘계학술발표대회논문집, 1990.

[국어연89] 국어 연구소, 국어 오용 사례집, 학술원 부설 국어 연구소, 1989.

[권재욱90] 권재욱, 조성배, 김진형, "계층적 신경망을 이용한 다중크기의 다중활자체 한글 문자 인식," 한글 및 한국어정보처리 학술대회, 제 2회, 1990.

[김영웅84] 김영웅, 한글 철자법 교정 시스템, 한국과학기술원 석사학위논문, 1984.

[남기심85] 남기심, 고영근, 표준 국어문법론,탑출판사,1985.

[문교부90] 문교부, 국어 어문 규정집, 대한교과서출판사, 1990.

[미승우90] 미승우, 새 맞춤법과 교정의 실제, 어문각, 1990.

[미승우86] 미승우, 맞춤법과 문장작법, 어문각, 1986.

[백산출89] 백산 출판사, 한글 맞춤법에 따른 붙여쓰기/띄어쓰기 용례집, 백산

출판사,1989.

[송춘환90] 송춘환, 원형 복원 방법을 이용한 한글 철자 검사기의 설계 및 구현, 한국과학기술원 석사학위논문, 1990.

[이관용90] 이관용, 이일병, "획 추출에 의한 한글 문서 인식 시스템의 설계 및 구현," 한글 및 한국어정보처리 학술대회, 제 2회, 1990.

[이재홍89] 이재홍, 오상현, "한글 음절의 초성, 중성, 종성 단위의 발생확률, 엔트로피 및 평균상호정보량,"전자공학회논문집, 제26권, 제9호, pp.1299-1307, 1989.

[조영환90a] 조영환, 한글 문서의 타이핑 오류 유형 조사, 한국과학기술원, CS Lab. Memo, 1990.

[조영환90b] 조영환, 김덕봉, 최기선, 김길창, "한글 맞춤법 오류 검사 및 교정 시스템," 한국정보과학회 가을 학술발표논문집, 17권 2호, 1990.

[조용덕90] 조용덕, 음절인식을 위한 순환구조 신경회로망, 한국과학기술원 석사학위논문, 1990.

[최형석84] 최형석, 이주근, "자연어 어절 처리 알고리즘," 한국정보과학회 가을 학술발표논문집, 11권 2호, 1984.

[Cho90] Y. H. Cho, J. W. Kang, K. S. Choi, G. C. Kim, "Hangul Spelling and Word-spacing Checker Using Connectivity Information," proceedings of PRICAI , November, pp.334-338, 1990.

[Durham82] Ivor Durham, "Spelling Correction in User Interfaces," Carnegie-Mellon University, 1982.

[Muth77] Frank E. Muth Jr., Alan L. Tharp, "Correcting Human Error in Alphanumeric Terminal Input," Information Processing & Management, Vol 13. pp.329-337, Pergrmon Press, 1977.

[Peterson80] James L. Peterson, "Computer Programes for Detecting and Correcting Spelling Errors," CACM, Vol.23, No.12, December, pp.676-687,1980.

[Price89] Gayle Dawn Price, "Grammatik IV User's Guide," Reference Software International, 1989.

부록 A. 좌접속 정보표

문법 범주				예	번호	우 접 속
…				…	…	…
명사	보통명사	용언 어간 겸용	한자형	무종성	설계, 연구, 처리	19	한자 보통명사형 접미사, 조사, 서술격 조사, 보통명사 조용보조어간
				유종성	분석, 가난, 번역	20	
				ㄹ종성	청결, 거절, 자살	21	
			순한글형	무종성	노래, 빨래, 이야기	22	순한글 보통명사형 접미사, 조사, 서술격 조사, 보통명사 조용보조어간
				유종성	밥, 사랑, 자랑	23	
				ㄹ종성	일, 절, 말	24	
		용언 어간 비겸용	한자형	무종성	권리, 자유, 우유	25	접미사, 조사, 서술격 조사, 보통명사
				유종성	학생, 책상, 병원	26	
				ㄹ종성	글, 사실, 지하철	27	
			순한글형	무종성	다리, 나무, 거리	28	접미사, 조사, 서술격 조사, 보통명사
				유종성	손, 도장, 땅, 바닥	29	
				ㄹ종성	달, 건달, 발, 터울	30	
…				…	…	…	…
숫자				1, 2, 3, 4, 5	370	숫자,숫자형 접미사, 단위성 의존명사
기호 (특수문자)				@, #, S, %, &, *, !	380	체언, 조사, 기호, 숫자
파생어 (위 표에 예외적인 것)					390	

부록 B. 우접속 정보표

문법 범주					예	번호	우 접 속
…					…	…	…
명사	보통명사	용언 어간 겸용	한자형	무종성	설계, 연구, 처리	19	한자 보통명사형 접미사, 조사, 서술격 조사, 보통명사 조용보조어간
				유종성	분석, 가난, 번역	20	
				ㄹ종성	청결, 거절, 자살	21	
			순한글형	무종성	노래, 빨래, 이야기	22	순한글 보통명사형 접미사, 조사, 서술격 조사, 보통명사 조용보조어간
				유종성	밥, 사랑, 자랑	23	
				ㄹ종성	일, 절, 말	24	
		용언 어간 비겸용	한자형	무종성	권리, 자유, 우유	25	접미사, 조사, 서술격 조사, 보통명사
				유종성	학생, 책상, 병원	26	
				ㄹ종성	글, 사실, 지하철	27	
			순한글형	무종성	다리, 나무, 거리	28	접미사, 조사, 서술격 조사, 보통명사
				유종성	손, 도장, 땅, 바닥	29	
				ㄹ종성	달, 건달, 발, 터울	30	
…					…	…	…
숫자					1, 2, 3, 4, 5	370	숫자,숫자형 접미사, 단위성 의존명사
기호 (특수문자)					@, #, S, %, &, *, !	380	체언, 조사, 기호, 숫자
파생어 (위 표에 예외적인 것)						390	