N-GRAM 한글 사전을 이용한 오인식 단어의 교정 알고리즘

*

이 종 연, 오 상 헌

현대전자 소프트웨어연구소

A Correction Algorithm for Misrecognized Words Using N-gram Hangeul Dictionary

JONGYUN LEE, SANGHUN OH

Software R&D Lab. HYUNDAI Electronics Co. Ltd.

요 약

본 논문은 온라인 한글인식 시스템에서 오인식된 단어를 교정하는 알고리즘이다. 교정 기법으로는 N-gram 한글사전을 이용하였다. 오인식된 단어는 후보키의 선정과 선정된 후보문자중 가장 유사한 단어로 대체된다. 오인식 단어는 사전에 수록된 단어의 형태소 정보 즉, 사전의 표제어, 이의 품사 및 접속 규칙을 활용하여 교정된다. 본 논문은 오인식 교정에서 필요한 한글의 형태소 분석기에 관한 선행연구를 전제한다.

I. 서 론

컴퓨터가 대중화되면서 사회의 모든 분야로 그 이용은 날로 확대되고 있다. 최근에는 컴퓨터에 익숙하지 못한 사용자들을 위해 펜컴퓨터(Pen-computer)에 관한 연구가 활발히 진행되고 있다. 펜컴퓨터의 연구개발의 주요 문제점은 온라인 이미지를 입력받아 코드화된 문자열로 변환하는 인식 기술이다. 온라인 문자인식은 사용자의 온라인 이미지를 입력하여 한글, 영문, 제스쳐, 특수기호의 정보로 인식하는 소프트웨어이다. 온라인 문자인식은 사용자의 스트록(Stroke) 이미지를 인식하여 처리하므로, 문

자인식의 결과는 항상 사용자의 요구기능을 만족시키지 못한다. 따라서, 온라인 문자인식의 결과는 별도의 처리 과정을 첨부하여 문자인식 시스템의 인식률 제고 및 오자(誤字) 교정이 필요하다. 이 과정을 문자인식 시스템의 후처리 과정 또는 오인식 교정이라 칭한다.

온라인 문자인식에서 오인식의 경우는 (1)띄어쓰기 오류, (2)형태소 오류로 분류된다. 형태소 오류에는 (1)불변어인 체언, 수식언, 독립언의 오인식과, (2)접미사의 오인식, (3)관계언의 오인식, (4)용언 어간의 오인식, (5)선어말어미와 어말어미의 오인식이 있다. 본 논문은 이들 형태소에 대한 오인식 교정 알고리즘의 기술로서 '한글의 형태소 분석에 관한 선행연구'를 전제로 하였다. 언어의 처리 수준은 기본적인 형태소 분석에서 구문 분석, 의미 해석으로 분류된다. 형태소 분석은 입력된 어절을 기본적인 형태소 단위로 분류하는 과정이다. 본 논문은 언어 처리의 기초적인 형태소 분석을 통해 오인식 문자를 교정한다. 따라서, 본 논문에서 설계한 오인식 교정기는 한글의 형태소 정보만을 이용한 오인식 교정이므로 그 교정범위가 제한된다.

본 논문의 전체적인 구성을 살펴보면 다음과 같다. 제Ⅱ장은 오인식교정기의 기능 및 시스템 개요를 기술하고, 제Ⅲ장은 교정기의 설계로 n-gram 한글사전과 오인식 교정의 기본 규칙, 띄어쓰기 교정, 체언/수식언/독립언/용언 어간의 교정, 관계언의 교정, 접미사의 교정, 선어말어미와 어말어미의 교정 방법에 관해 기술한다. 마지막으로 제Ⅳ장은 본 논문의 결론 및 향후 연구 방향을 기술한다.

Ⅱ. 오인식교정기의 기능

온라인 문자인식은 사용자의 스트록 이미지를 입력하여 한글, 영문, 제스쳐, 특수기호의 문자로 인식한다. 본 연구에서 언급되는 형태소분석기는 문자인식 시스템에서 출력된 문자열을 입력받아 형태소 단위의 분류와 용언의 원형을 복귀한다. 오인식 교정기는 형태소 분석된 문자열을 입력받아 사전의 형태소 정보를 이용하여 오인식된 문자열을 교정하는 프로그램이다.

2.1 어절의 상태 전이

한글에서 문장의 구성은 '주어+목적어+서술어'의 어순을 가지며, 한개의 문장은 여러개의 어절로 분리된다. 또한, 어절은 한개 또는 그 이상의 단어로 분리되며, 단

어는 다수개의 음절로 분리된다. 음절은 다시 '자음+모음+자음'의 음소들로 구성된다 [고등문법92]. 뜻을 가진 가장 작은 말의 단위를 형태소(形態素)라 하며, 이는 다시 독립적으로 사용할 수 있는 자립형태소와 다른 말에 의존하여 사용되는 의존형태소로 분류된다.

표 2.1과 같이 한개의 어절은 여러개의 형태소로 구성된다. 어절은 단일의 수식언, 독립언, 체언으로 형성되는 경우와 체언, 용언의 어간, 부사와 함께 접미사, 조사 및 서술격조사, 명사형어미, 선어말어미, 어말어미 등의 여러가지 결합 형태를 갖는다. 표 2.1의 결합형태는 문자인식 시스템의 후처리(Postprocessing) 관점에서 어절의 결합형태를 기술한 것이다. 표 2.1의 품사 정보는 한글 문법에서 품사 분류의 기준과 일치한다. 체언은 명사, 대명사, 수사 등을 포함하고, 수식언에는 관형사, 부사를 포함한다. 독립언에는 단독으로 이용되는 감탄사가 속한다. 조사는 일반 조사와 서술격조사로 구별한다. 일반적으로 서술격조사에는 체언과 결합하여 서술어를 형성하는 '이다, 아니다'가 해당되지만; 본 연구에서는 편의상 그 활용형이 유사한 접미사 '-하다'를 포함하여 어절의 결합 형태를 기술하였다.

한글에서 어절은
1)부사 (+조사)
:빨리 (+는)
2)독립언 / 체언 / 수식언
:아이구 / 현대전자 / 새, 헌
3)체언 (+접미사) +조사
:사람 (+들) + 을
4)체언(부사) +서술격조사 (+선어말어미) +어말어미 (+보조사)
:사랑 + 이 (+었) + 지만 (+은) / 사랑 + 하 + 였 + 지만 (+은)
5)체언(부사) +서술격조사 (+선어말어미) +명사형어미 +조사
:사랑 + 이 + ㅁ + 을
6)용언 어간 (+선어말어미) +명사형어미 +조사
:사랑하 (+시) + ㅁ + 을
7)용언 어간 (+선어말어미) +어말어미 (+보조사)
:사랑하 (+시) + ㅂ니까 (+요)
등의 결합 형태를 갖는다.

표 2.1 한글 어절의 결합형태

어절의 결합형태를 역순으로 분석하면 그림 2.1과 같은 형태소들로 구성된다. 그림 2.1과 같이, 어절의 구성은 조사가 없는 경우와 조사가 있는 경우로 구분된다. 조

사가 없는 경우는 수식언, 체언, 독립언이 단독으로 어절을 구성한다. 조사가 있는 경우는 여러가지 형태소들과 결합하여 하나의 어절을 형성한다.

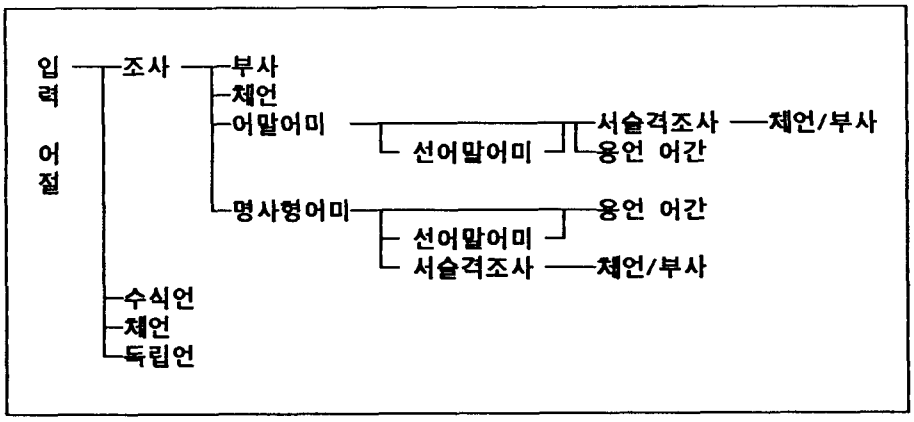

그림 2.1 한글 어절의 상태 전이도

2.2 시스템 개요

오인식 교정기는 기능별로 (1)N-gram 한글 사전, (2)형태소 오류의 교정, (3)띄어쓰기 교정의 세개 모듈로 구분되며, 이의 시스템 구성은 그림 2.2와 같다. 형태소 오류는 형태소 분석과정과 사전의 수록정보를 활용하여 교정된다. 형태소 오류의 교정은 형태소의 속성에 따라 불변어/용언어간, 접미사, 관계언, 선어말어미, 어말어미의 모듈로 세분된다.

불변어/용언어간의 교정은 단어의 활용 형태상 불변어인 체언(명사,대명사,수사), 수식언(관형사,부사), 독립언(감탄사)과 용언의 어간에 관한 오인식 교정이다. 관계언의 교정은 여러개의 조사가 복합된 복합조사와 단일조사 및 서술격조사(이다,아니다)에 관한 오인식 교정이다.

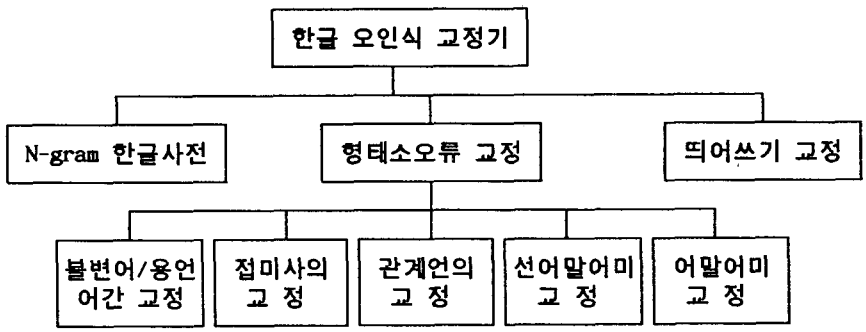

그림 2.2 한글 오인식교정기의 기능

2.3 설계상 제약조건

본 논문은 사전의 형태소 정보를 이용한 오인식 문자의 교정이다. 사전을 기초한 오인식 교정의 설계 조건은 다음과 같다.

첫째, 입력 어절의 오류는 형태소 분석기에 의해 판명된다. 오류의 기준은 입력 형태소가 한글사전에 표제어로 존재하지 않는 경우이다. 둘째, 본 연구에서 오인식 교정은 한글의 어절 정보와 사전의 표제어 정보 및 이의 좌우접속 정보를 이용한다. 따라서, 한글의 형태소 정보만을 이용한 오인식 교정은 그 교정범위가 제한된다. 예로 단어 '사명'를 잘못 인식하여 '자명'으로 오인식한 경우 '자명'이 사전에 존재하면 이는 옳바른 인식이라 판단한다. 셋째, 형태소 정보를 이용한 오인식 교정기는 한글 사전의 구축 및 형태소분석기의 구현이 선행돼야 한다.

Ⅲ. 교정기의 설계

문자인식 시스템에서 출력된 어절은 형태소 분석과정을 통해 형태소 단위로 분류된다. 오인식 교정기는 형태소 분석된 어절을 입력받아 오인식 문자의 존재 여부와 이의 위치를 파악하고, 파악된 오인식 형태소를 교정한다. 본 논문의 오인식 교정은 그림 3.1의 처리과정을 통해 처리된다. 다음은 오인식 교정의 기본 규칙과 한글사전의 구성 및 한글의 띄어쓰기 규칙, 그리고, 불변어/용언어간, 접미사, 관계언, 선어말어미, 어말어미 형태소에 대한 교정 알고리즘을 순차적으로 기술한다.

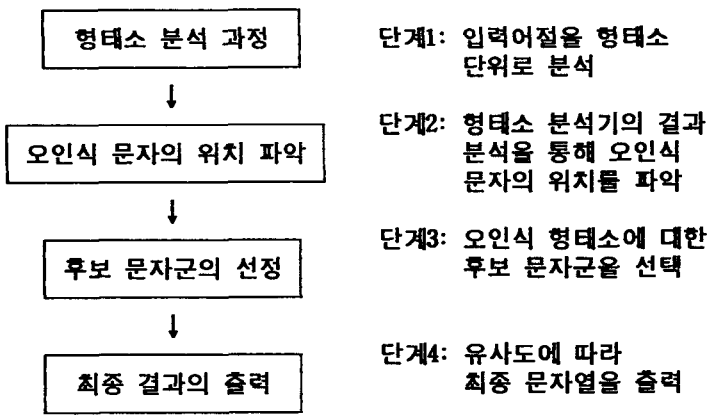

그림 3.1 한글 오인식 교정기의 처리 절차

3.1 오인식 교정의 기본 규칙

본 논문에서 한글의 형태소 정보를 이용한 오인식 교정의 기본 규칙은 다음의 표 3.1과 같다.

(1)문자인식 시스템에 의해 인식된 결과의 후보문자중 사전에 존재하는 문자열을 우선적으로 후처리의 결과로 산출한다. (2)입력 어절에서 좌우 접속정보의 분석 결과, 사전에 존재하는 형태소를 우선적으로 옳바른 인식이라 단정하고, 오인식 교정을 실행한다. (3)옳바른 인식의 경우도 한글 철자법상의 오류는 교정된다.

표 3.1 오인식 교정의 기본 규칙

3.2 N-gram 한글사전

본 논문에서는 오인식 교정을 위해 한글 사전을 (1)불변어 사전, (2)접미사 사전, (3)조사 사전, (4)선어말어미 사전, (5)어말어미 사전으로 분류 저장하였다. 불변어 사전에는 체언인 명사, 대명사, 수사와 감탄사, 용언의 어간이 수록된다. 불변어 사전에 수록되지 않는 단어는 오인식 형태소의 효율적인 교정을 위해 각각 별개의 사전으로 분리하였다.

사전의 구조는 사전 크기에 따라 인덱싱 화일과 데이타 화일로 구성하였다. 많은 형태소를 저장하는 불변어 사전에는 (1)체언인 명사, 대명사, 수사와 (2)용언 어간, (3)수식언인 관형사와 부사, (4)독립언인 감탄사가 수록되며, 한개 표제어에 대해 3 - 4 gram의 인덱싱 키를 갖는다. 인덱싱은 해싱(hashing) 구조로 설계하였다.

3.3 띄어쓰기 교정

(1)어절은 보통 띄어쓰기의 단위와 일치한다. (2)조사는 그 윗말에 붙혀 쓴다. (3)파생접사는 윗말에 붙혀 쓴다. (4)두개 단어가 합쳐서 한개의 복합어를 생성할 때는 붙혀 쓴다.

표 3.2 한글의 띄어쓰기 규칙들

띄어쓰기 교정은 한글의 띄어쓰기 규칙에 따라 교정되며, 한글의 띄어쓰기 규칙은 표 3.2의 내용과 같다.

3.4 오인식 문자의 교정

한글의 형태소 분석 관점에서 단어들은 불변어와 가변어로 분류된다. 불변어는 어형이 고정된 상태의 체언(명사,대명사,수사), 수식언(관형사,부사), 독립언(감탄사)과 별개의 형태소로 취급되는 관계언(조사)을 포함한다. 가변어는 용언과 선.어말어미, 접사, 서술격조사(조용사)등의 활용을 하는 단어이다. 본 논문에서 오인식 문자의 교정은 (1)불변어/용언어간의 교정, (2)접미사의 교정, (3)관계언의 교정, (4)선어말어미의 교정, (5)어말어미의 교정으로 분류되어 교정된다.

3.4.1 불변어/용언어간의 교정

불변어와 용언어간의 교정에는 활용을 하지 않는 체언, 수식언, 독립언과 용언의 어간이 포함된다. 체언에는 명사, 대명사, 수사를 포함하며, 체언의 교정은 다음에 기술한 '오인식 단어의 교정 알고리즘'에 따른다. 체언의 접속 규칙은 다음과 같다. 셈의 대상이 되는 명사와 대명사는 복수 표시의 접사 '-들'을 취할 수 있지만, 셀 수 없는 명사나 장소 표시의 대명사에는 '-들'을 붙일 수 없다. 또한, 의존명사인 '분, 것, 지, 따름, 채' 앞에는 반드시 꾸미는 말이 있어야 한다.

수식언에는 관형사와 부사가 있다. 관형사는 좌우측에 아무 것도 접속되지 않는다. 부사의 경우는 좌측에 접속정보가 없으며 우측에는 보조사(예:아무'도')가 접속 가능하다. 독립언에는 감탄사가 있으며, 좌우측의 접속정보를 갖지 않는다.

용언은 문장의 주체를 서술하는 기능으로 동사와 형용사가 포함된다. 용언은 단어의 끝이 여러 가지 형태로 바뀌는 특성을 갖으며, 이를 용언의 활용이라 한다. 용언이 활용할 때에 변하지 않는 줄기 부분을 어간(語幹)이라 하고, 변하는 부분을 어미(語尾)라 한다[고등문법92]. 용언의 활용은 규칙적인 활용과 어간과 어미의 형태가 달라지는 불규칙 활용으로 구별된다.

불변어/용언어간의 교정은 (1)문자인식 시스템에서 출력된 문자열 조합에 의한 교정과 (2)앞의 (1)에서 교정이 실패한 경우는 n-gram 한글사전을 이용한 교정 단계를 통해 처리된다. 본 논문에서 설계한 '오인식 단어의 교정 알고리즘'은 다음과 같다.

첫째, 문자인식 시스템의 인식 결과중 인식률이 큰 순서로 5개 문자열을 입력하여 이 문자열의 조합은 사전 검색을 통해 오인식 결과로 출력된다. 예로, '현대'를 문자

인식 결과로 '헌대, 현댸, 헐댜, 철래, 련내'로 인식한 경우 이들 단어를 사전에서 검색하여 존재하는 단어를 후처리 결과로 출력한다. 이때 모든 후보문자가 사전에 존재하지 않는 경우는 후보 문자 '헌, 현, 헐, 철, 련'과 '대, 댸, 댜, 래, 내'로 구성되는 문자열 조합을 생성한다. 생성된 문자열 '헌댸, 헌댜, ..., 현대, ..., 련내'중 사전에 있는 '현대'가 발견되면 이를 후처리 결과로 출력한다. 생성된 문자열 조합중 동시에 사전에 존재하는 경우는 인식률이 큰 단어를 후처리 결과로 출력한다.

둘째, N-gram 한글사전을 이용한 오인식 교정은 입력된 형태소에 대해 후보키를 생성하여 후보키를 갖는 사전의 표제어를 선정된다. 선택된 후보키는 입력된 후보키와의 유사도(similarity)를 계산하여 가장 큰 값의 단어를 후처리 결과로 출력한다. 선택된 단어중 동일한 유사도를 갖는 단어는 후보 문자군으로 출력하여 사용자의 선택에 따른다. 유사도 측정에는 표제어간의 유사도와 종성의 유무 일치, 품사 정보, 구문 정보 등이 활용된다.

셋째, 불변어/용언어간을 제외한 접미사, 조사, 선어말어미, 어말어미의 교정은 다음과 같다. 입력된 단어의 품사에 따라 각각의 사전 탐색을 통해 후보문자의 선정과 선정된 단어중 가장 큰 유사도를 갖는 단어를 후처리 결과로 출력한다.

3.4.2 접미사의 교정

1) 어근과 접미사 사이에는 조사가 끼일 수 없다.
2) 접미사 '-이,-기,-(으)ㅁ'이 붙어 명사를 파생하는 경우 어근이 자음이면 '-이'가 많이 쓰이고, '-ㅂ'받침을 가진 형용사에는 '-(으)ㅁ'이 붙는다. 또한, '-(으)ㅁ'가 붙는 말은 '-이'나 '-기'가 쓰일 수 없다.
3) 통사적 파생의 접미사에는
 가)용언을 명사로 파생하는 접미사 :-(으)ㅁ,-이,-기,애,-개,-게,-어지,-다리
 나)용언을 부사로 파생하는 접미사 :-오/우,-이,-히
 다)동사를 형용사로 파생하는 접미사 :-업/-압,-ㅂ-,-브-
 라)명사를 용언으로 파생하는 접미사 :-지-,-하-,-답-,-롭-,-스럽-
 마)명사를 부사로 파생하는 접미사 :-껏,-내,-로,-이,-히
 바)형용사를 동사로 파생하는 접미사 :-우-,-이-,-히-,-추-
 사)형용사를 관형사로 파생하는 접미사 :-(으)ㄴ
 아)부사를 동사로 파생하는 접미사 :-거리-,-이-,-히-
 자)관형사를 형용사로 파생하는 접미사 :-롭-

표 3.3 접미사의 결합 제약과 목록

한글에서 접미사는 어근에 붙는 접사로 그 어근에 뜻을 더하거나 품사를 바꾸는 형태소이다. 접미사는 일반적으로 (1)어근과 결합하여 그 뜻을 더해주는 어휘적 파생과 (2)새로운 품사를 파생시키는 통사적 파생으로 분류된다. 어휘적 파생의 접미사는 어근과 결합하여 새로운 의미의 단어를 생성하여, 사전에 표제어로 저장된다. 따라서, 본 연구에서 접미사의 교정은 통사적 파생의 접미사로 그 의미를 제한한다. 어근과 접미사간의 결합상의 제약과 통사적 접미사의 종류는 표 3.3과 같다. 접미사는 어절의 상태 전이 즉, '체언(부사) (+접미사) +조사'의 결합 형태를 이용하여 접미사의 위치를 파악한다. 복수표시의 접미사로 흔히 사용되는 것은 '-들, -희, -네' 등이 있으며, 명사와 대명사에만 표시되지만 수사에는 나타나지 않는다.

3.4.3 관계언의 교정

관계언에는 조사와 서술격조사('이다,아니다')가 있다. 조사는 기능 및 형식상의 특성에 따라 크게 격조사, 접속조사, 보조사(특수조사)로 분류된다. 보조사는 체언, 부사, 연결어미 등에 붙어 이의 뜻을 더해주는 역할을 한다(예:어머니는 나도 좋아한다). 보조사에는 '는, 도, 만, 부터, 까지, 조차, 마다, (이)나, (이)든지, (이)라도, 마저, (이)나마' 등이 있다.

(1) 은/는, 을/를, 야/아
(2) 과/와, 으로/로, 이고/고, 이며/며, 이나/나, 이든지/든지, 이나마/나마 이든지/든지, 이나마/나마, 인들/ㄴ들, 이랑/랑, 이라도/라도
(3) 이/가

표 3.4 음운론적 이형태를 갖는 조사들.

(1)사전에 존재하는 조사는 우선적으로 옳바른 인식이라 단정한다.
(2)위의 (1)번의 결과중 윗말의 종성 유무에 따라 사용가능 여부를 판별하여 오인식 교정 또는 적당한 단어로 교체한다.
(3)모든 후보문자열이 사전에 존재하지 않은 경우는 조사 사전에서 유사도가 가장 큰 문자열을 후처리 결과로 출력한다.

표 3.5 조사의 오인식 교정

조사는 윗말의 받침유무에 따라 형태를 달리하는 경우가 있으며, 표 3.4와 같다[

고영근92]. 표 3.4에서 각 쌍의 경우 전자(은,을,야,과...)는 윗말의 받침이 있는 경우에 사용되고 후자(는,를,아,와...)는 윗말에 받침이 없는 경우에 사용된다. 조사는 단어 윗말에 접속되며, 윗말의 속성에 따라 접속정보를 달리한다. 따라서, 조사의 교정은 윗말의 속성에 따라 적합한 단어의 선택 및 문자인식 시스템의 결과인 확률치에 따라 후처리 결과로 출력한다. 조사에 관한 오인식 교정 규칙은 표 3.5의 내용과 같다.

3.4.4 선어말어미의 교정

활용어의 어간에 붙어서 다른 말과의 관계를 나타내는 형태소를 어미(어미)라 한다. 이 어미는 크게 선어말어미와 어말어미로 분류된다. 선어말어미는 실질 형태소인 어간과 형식 형태소인 어말어미 사이에서 높임, 공손, 시제 등을 뜻하는 어미이다. 어말어미는 용언의 어간 또는 선어말어미 뒤에 붙는 어미이다. 어말어미는 그 자체만으로 단어를 형성하지만, 선어말어미는 뒤에 반드시 어말어미를 수반한다.

분리적 선어말어미	교착적 선어말어미
(1)주체높임법 :-시- (2)시제 :-는/ㄴ-(현재),-겠-(미래) -었/았-(과거) (3)공손 :-옵/오-	(4)상대높임법의 합소체 :-습/ㅂ- (5)서법 :-느-, -더-, -리- (6)강조법 :-니-, -것-

표 3.6 선어말어미의 분류

선어말어미는 용법상의 제약에 따라 표 3.6과 같이 분리적 선어말어미와 교착적 선어말어미로 분류한다. 분리적 선어말어미는 용법상의 제약이 뒤따르지 않으며, 어말어미에서 쉽게 분리될 수 있는 어미이다. 교착적 선어말어미는 분포가 극히 제한되어 용법상의 제약이 있으며, 어말어미와 밀착하여 사용된다. 따라서, 교착적 선어말어미인 '-것다, -렷다, -느냐, -는' 등은 용법상의 제약에 따라 선어말어미와 어말어미를 별도로 구별하지 않고 어말어미와 함께 사전에 등록한다. 둘 이상의 선어말어미가 결합할 때는 '높임-시제-공손'의 순서로 결합되며, 선어말어미의 결합순서는 일정하여 임의로 변경할 수 없다. 선어말어미의 결합순서는 분포의 넓고 좁음에 비례하여 표 3.6에서 표시된 '-시-, -는/ㄴ-, -겠-, -었-, -옵/오-'의 결합순서를 갖는다. 예로 선어말어미 '-시-'는 모든 선어말어미와 결합될 수 있지만 '-었-'은 그렇지 않

다[고영근92][박갑수91]. 지금까지 기술한 바와 같이 후처리 과정에서 나타난 선어말어미의 결합 조건과 규칙을 요약하면 표 3.7과 같다.

(1)본 연구에서 선어말어미는 분리적 선어말어미를 의미한다.
(2)교착적 선어말어미는 용법의 제약에 따라 어말어미와 결합된 형태를 사전에 저장한다.
(3)선어말어미는 뒤에 어말어미를 반드시 수반한다.
(4)입력된 선어말어미에 따라 다음과 같은 결합상의 제약을 받는다[강승식92].
가)매개모음 '으'는 어간의 끝음절에 받침을 요구한다.
나)선어말어미 '았/었'은 모음조화 규칙에 따라 결합된다.
다)선어말어미 '-사-/-셨-'은 어간의 끝음절에 받침이 없는 것을 요구한다.
라)선어말어미 '-였-'은 '-하다,-이다'로 끝나는 용언에만 결합한다.
마)선어말어미 '-ㅆ-'은 'ㅏ / ㅓ' 또는 'ㅐ / ㅔ'로 끝나는 어간과 결합한다.
바)선어말어미 '-ㅏㅆ/-ㅓㅆ-'은 어간의 끝음절이 'ㅗ,ㅜ,ㅡ'로 끝나는 어간과 결합한다. 이 조건을 만족하는 불규칙 용언은 '으 탈락, 우 불규칙,러 불규칙,르 불규칙,와 불규칙,워 불규칙' 등이 해당된다.
사)선어말어미 '-ㅕㅆ-'은 어간의 끝음절이 'ㅣ'로 끝나는 어간과 결합한다.
아)선어말어미 '-ㅐㅆ-'은 '-하다'로 끝나는 용언 또는 ㅎ불규칙 용언과 결합한다.

표 3.7 선어말어미의 결합제약

3.4.5 어말어미의 교정

어말어미는 한 문장을 종결형이 되도록 하는 『종결어미』와 문장의 종결 기능이 결어된 문장의 접속 또는 전성의 기능을 띤 『비종결어미』로 구분된다. '먹는다'의 활용을 예를 들면, '먹는다, 먹는구나, 먹느냐, 먹어라, 먹자' 등으로 표현된다. 예에서 '-다, -구나, -냐, -어라, -자' 등의 어미는 한 문장을 종결하는 종결어미이다. 비종결어미는 용언의 어간에 붙어 다음말에 연결하는 구실을 하는 연결어미(蓮結語尾)와 용언의 어간에 붙어 다른 품사의 성질로 바꾸는 전성어미(轉成語尾)로 분류된다. 연결어미에는 '-고, -면서, -든지 -든지'의 대등적 연결어미와 '-면, -니, -는데,'의 종속적 연결어미, '-어, -고, -게, -지'의 보조적 연결어미가 있다. 전성어미에는 '-ㄴ, -ㄹ'의 관형사형 연결어미와 '-ㅁ, -기'의 명사형어미가 있다. 이들 어말어미의 분류를 도시하면 그림 3.2와 같다.

이들 어미는 활용어(동사,형용사,서술격조사)의 활용 형태가 조금씩 다르며, 어미 결합상의 제약은 표 3.8과 일치한다.

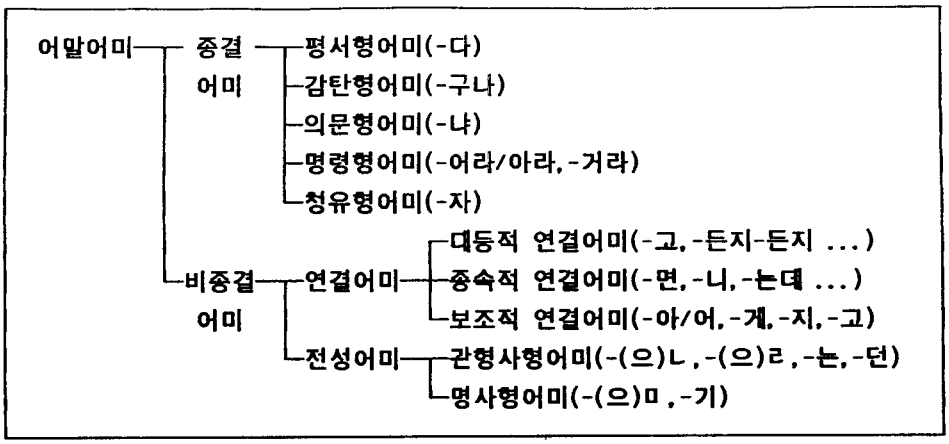

그림 3.2 어말어미의 분류

(1)동사에는 '-(는/ㄴ)다, -(느)냐, -(는)가, -(는)구나, -(는)도다, -아라/어라, -자' 등의 어미가 쓰이고,
(2)형용사에는 '-는/ㄴ, -느-'가 빠진 '-다, -(으)냐, -(으)ㄴ가, -구나, -도다, -아라/어라' 등의 어미가 붙는다.
(3)서술격조사에는 '-다, -냐, -ㄴ가, -(로)구나, -로다'가 쓰인다.
(4)명령형어미(-거라,-아라/어라)과 청유형어미(-자), 그리고 평서형어미 '-(으)마, -(으)려무나'는 형용사와 서술격조사에 사용되지 않고, 동사에만 나타난다.
(5)연결어미 '-면서, -고서, -자, -다가, -도록, -려고, -러, -고' 등은 동사에는 붙지만, 형용사와 서술격조사에는 붙지 못한다.
(6)보조적 연결어미는 동사나 형용사에만 붙고, 서술격조사에는 붙지 않는다.
(7)종속적 연결어미 '-라면,-라도'는 서술격조사에만 쓰인다(꽃이라도).
(8)어미 '-ㄹ세,-ㄹ시다,-로구나,-로다'등은 서술격조사(이다, 아니다)이외는 붙을 수 없다(일세, 이올시다, 이로구나, 이로다)

표 3.8 어말어미 결합의 제약

Ⅳ. 결 론

본 논문은 온라인 한글인식의 오인식 결과에 대한 교정으로, 한글의 형태소 정보를 이용한 오인식 교정기를 설계하였다. 오인식 교정기는 온라인 문자인식 시스템에 첨부되어 오인식 문자열의 교정을 통해 인식 시스템의 인식율 향상을 목적으로 한다.

결론적으로 N-gram 한글사전을 통한 오인식 교정은 오인식된 단어에 대한 후보문자의 선정은 거의 완벽하다. 단, 선정된 후보문자중 최종적인 한개 단어를 선정하는 문제가 존재하며, 한글의 형태소 정보만을 이용한 오인식 교정은 그 처리범위가 제한되었다. 향후 온라인 문자인식에서 완벽한 오인식 교정을 위해 한글의 형태소 정보 뿐만 아니라 구문 정보와 의미 정보를 이용한 오인식 교정기의 연구 개발이 요구된다.

본 연구에서 설계한 오인식 교정기는 문자인식 시스템에 통합 적용하여 오인식된 문자의 교정에 따라 인식율 향상이 기대된다. 또한, 본 연구에서 연구개발한 오인식 교정기는 음성인식(Speech Recognition) 및 오프라인 문자인식 시스템(Offline Character Recognition Systems)의 오인식 교정과 단독적인 철자교정기로서 활용될 수 있다.

참 고 문 헌

[고등문법92] "고등학교 문법," 성균관대학교 대동 문화 연구원, 1992.

[강승식92] 강승식외 1인, "한국어 형태소 분석기에서 선어말어미의 분석 모형," 정보과학회논문지 제18권 제5호, pp505-513, 1991.9.

[강승식92] 강승식외 1인, "한국어 형태소 분석기에서 불규칙 용언의 분석 모형," 정보과학회논문지 제19권 제2호, pp151-164, 1992.3.

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

[국어사전80] "동아 신크라운 국어사전," 동아교재사, 1980.

[고영근92] 고영근외 1인, "표준 국어 문법론," 탑출판사, 1992.

[김성용87] 김성용, "TABULAR PARSING 방법과 접속 정보를 이용한 한국어 형태소 분석기," 한국과학기술원 석사학위논문, 1987.

[박진규88] 박진규, "한글문서 인식 시스템의 오인식 수정에 관한 연구," 한국과학기술원, 석사학위논문, 1988.

[조규빈91] 박갑수, "하이라이트 고교문법," 지학사, 1991.

[TAKAHA89] H.TAKAHASHI외3인,"A Spelling Correction Method and its Apllication to an OCR System," Pattern Recognition,Vol.23, No.3/4, pp.363-377, 1990.