의미소를 이용한 한국어 오류 문자 교정 시스

The error character Revision System of the Korean using Sememe

박 현 재, 박 해 선, 강 원 일, 손 영 선

동명정보대학교 정보통신공학과

Hyun-Jae Park, Hae-Sun Park, One-Il Kang and Young-Sun Sohn

Department of Information & Communication Engineering,

Tongmyong University of Information Technology

(yssohn@tmic.tit.ac.kr)

요 약

현재 구현되어 있는 한국어 철자 교정 시스템은 문장의 문법 정보나 연어 관계로부터 문장의 오류를 처리하는 방식을 쓰고 있다. 본 논문에서는, 홑문장에서 의미소 사이의 관계를 이용하여 오타 문자를 수정하고 오타에 의한 의미적인 오류가 있을 때에는 의미에 해당하는 적절한 단어를 대체하여 제공하는 시스템을 제안한다.

단어의 뜻에 따라 체언은 의미 트리를 형성하고, 서술어는 주어 및 목적어의 체언과 의미 관계를 정의한다. 오류가 포함된 문장에서, 의미 관계를 비교, 분석하여 주어 및 목적어의 체언이 틀렸을 경우에는 서술어로부터, 서술어가 틀렸을 경우에는 주어 및 목적어의 체언으로부터, 수식어가 틀렸을 경우에는 체언 또는 서술어로부터 정의된 상호 의미 관계를 이용하여 한 문자에 대한 오타를 수정하고 오타에 의한 의미적 오류가 발견될 때에는 상기와 같은 철자 교정 방법을 적용하였다.

Key words : 자연어 , 한글 , 의미소 , 문자교정

I. 서론

오늘날 컴퓨터의 발달로 인해 수많은 정보들을 전자 문서화하여 수정, 저장 및 관리하는 시스템이 일반화되어 가고 있다. 이에 따라 사용자가 좀 더 빠르고 정확하게 전자 문서를 처리할 수 있도록 해주는 편집 시스템의 필요성이 대두되고 있다[1~8].

편집 기능 중, 잘못된 단어를 올바른 단어로 교정해 주는 기능은 언어의 특성상 상당한 어려움이 따른다.[3~5] 특히 한국어는 하나의 어근에 여러 개의 형식 형태소들이 결합되어 다양한 단어들의 생성이 가능한 교착어이므로 영어와 같은 비교착어보다 문서 교정이 어렵다[4~6].

글자의 교정은 크게 철자 교정과 의미 교정으로 나눌 수 있다. 철자 교정은 문서상에서 단어가 틀렸다고 판단될 때 이를 적절한 단어로 교정하거나 교정 후보를 제시하고, 의미 교정은 글의 흐름에 있어 뜻이 통하지 않거나 어색하다고 판단될 때 문맥을 이용하여 단어를 교정하거나 교정 후보를 제시한다[9].

본 논문에서는, 서술어와 의미적으로 함께 쓰일 수 있는 주요 문장 성분들 및 각 문장 성분들이 가지는 의미를 이용하여 문장의 오타 및 의미적인 오류를 교정하는 시스템을 구현하였다.

II. 시스템

본 논문에서는 그림 1에서 알 수 있듯이 문장을 입력 받아 어절을 분리한 뒤 DataBase (DB)를 이용하여 단어를 분리한다. 단어의 품사 정보와 문장 성분의 위치 정보 및 개수 정보를 이용하여 어절의 문장 성분을 분석 및 추정한다. 어절의 문장 성분 정보와 유사 단어 정보를 얻고, 서술어의 문장 형식과 문장 성분 간의 의미 관계를 이용하여 오류 단어를 교정한다. 교정된 문장들은 문서 편집기에 제시되어진다.

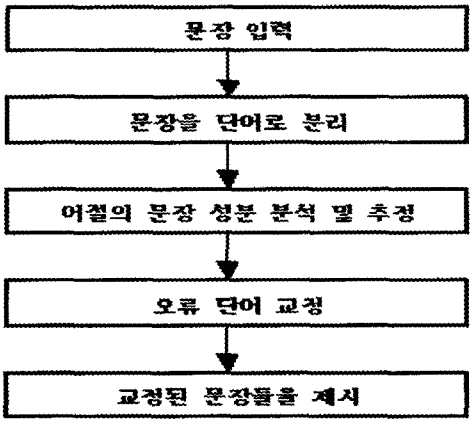

그림 1. 시스템 흐름도

사용자가 입력하는 문장 조건은 다음과 같다.

1. 주요 문장 성분들이 아래의 형식들과 같이 이루어져 있거나 이들을 꾸며주는 수식어를 하나씩 가지는 문장이어야 한다. [10, 11]

(형식1)주어+서술어
(형식2)주어+직접목적어+서술어
(형식3)주어+간접목적어+직접목적어+서술어

2. 한 문장으로 규정한다.
3. 띄어 쓰기가 맞아야 한다.
4. 오타가 하나를 초과해서는 안된다.
5. 틀린 단어 중, 글자의 탈락 및 삽입에 대해서는 고려하지 않는다.

III. 문장 성분 분석 및 추정

입력된 문장을 어절로 분리한 후, DB를 검색하여 각 어절을 단어로 분리한다.

문장은 각 어절의 단어 개수 정보에 따라서 분류하고, 단어의 품사 정보를 이용하여 문장 성분을 분석 및 추정한다. 오타가 포함되어 있어 분석되지 않은 어절은 분석된 문장 성분의 분포에 따라서 분류하고, 틀린 어절의 위치 정보와 단어 개수 정보 그리고 나머지 어절들의 분석 결과를 이용하여 문장 성분을 추정한다.

3.1. DataBase

단어 분리를 위한 DB 내의 table 구성은 체언, 서술어, 관형어, 부사어, 조사 등이 있다. 체언 table은 명사, 대명사, 수사로 분류되어 있고, 서술어 table은 동사, 형용사로 분류되어 있다. 조사 table은 주격 조사, 목적격 조사, 보격 조사로 분류되어 있다. 각 table에는 640, 100, 35, 10개의 임의의 단어들이 각각 포함되어 있으며 조사 table에는 일반적으로 많이 사용하는 조사들을 추출하여 포함시켰다.

형식에 따라 문장 분석을 하기 위한 table로는 형식1, 형식2, 형식3이 있고 각 table에는 서술어의 의미와 의미 관계가 맞는 주어의 의미들이나 목적어의 의미들이 포함되어 있다.

3.2. 어절 분석

한 어절은 그림 2의 (a)처럼 체언과 조사가 함께 구성된 경우와 그림 2의 (b)처럼 서술어, 관형어 또는 부사어가 단독으로 구성된 경우로 구분된다.

어절의 구성이 그림 2의 (b)와 같은 경우, 어절을 서술어, 관형어, 부사어 table과 차례로 비교하여 단어를 찾는다. 단어가 존재하지 않으면 그림 2의 (a)인 경우로 간주하고 어절 끝의 글자들을 조사로 판단하여 조사 table과 비교한다. 조사로 판단된 글자들을 제외한 어절의 나머지 부분은 체언 table과 비교한다. 결과가 존재하지 않는 단어는 틀린 단어로 간주된다.

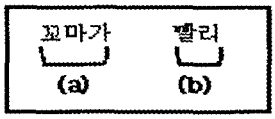

그림 2. 어절의 구성에 따른 분류

IV. 오류 단어 교정

문장 성분들 간의 의미 관계를 비교하기 위해서 각 단어들을 의미에 따라 분류한다.

체언은 그림 3의 예와 같은 트리 형식으로 분류하였고 서술어와 관형어는 의미 범주로 분류하였다. 부사어는 자신이 꾸며줄 수 있는 서술어의 의미 범주에 따라 분류하였다.[12~14]

오류 단어는 어절의 문장 성분 분석 결과와 추정 결과를 비교하여 두 결과가 다르면 글자가 틀린 경우로 간주하고, 같으면 의미가 틀린 경우로 간주하여 단어를 교정한다.

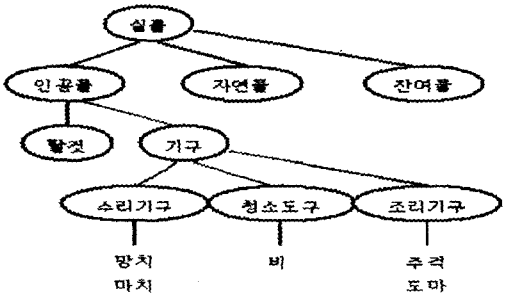

그림 3. 체언 의미 트리의 예

4.1. 오타를 포함하는 경우

그림 4를 보아 알 수 있듯이 글자가 틀린 경우에 대한 교정은 문장을 단어로 분리하는 과정에서 얻은 오타 정보를 이용한다.

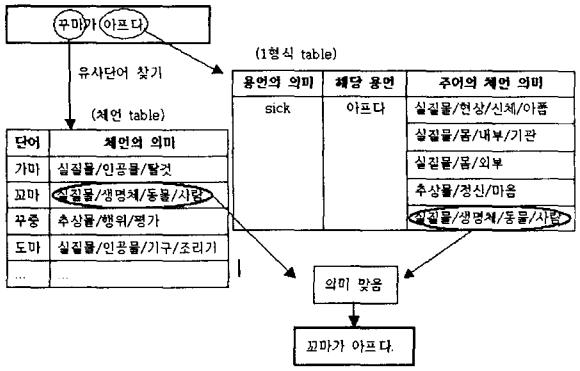

그림 4. 형태적 오류를 포함하는 예

먼저, 오타가 조사 위치에서 발생하는 경우에는, 오타를 포함하는 어절 내의 체언을 찾고, 그 체언에서 마지막 글자의 종성 유무 정보를 이용하여 어절 추정 결과에 만족하는 조사로 교정한다.

오타가 관형어 위치에서 발생하는 경우에는, 틀린 관형어와 유사한 단어들을 관형어 table에서 선택한다. 선택된 관형어들 중 오류 어절의 바로 뒤에 오는 체언을 수식할 수 있는 단어들을 선택한다. 주어를 수식하는 관형어가 틀린 경우에는, 서술어와 의미가 중복되지 않는 단어들을 선택한다.

오타가 부사어 위치에서 발생하는 경우에는, 틀린 부사어와 유사한 단어들을 부사어 table에서 선택한다. 선택된 부사어들 중 서술어를 수식할 수 있는 단어들을 선택한다.

오타가 체언 위치에서 발생하는 경우에는, 틀린 체언과 유사한 단어들을 체언 table에서 선택하여 체언들의 의미를 찾아낸다. 입력된 문장 형식에 따라 형식 table이 결정되므로, 결정된 형식 table에서 서술어의 의미를 검색한다. 선택된 체언들 중 검색의 결과로 찾아진 서술어와 의미 관계가 맞는 단어들을 선택한다. 틀린 체언을 수식하는 관형어가 있을 경우에는, 관형어의 수식을 받을 수 있는 단어들을 선택한다.

오타가 서술어 위치에서 발생하는 경우에는, 틀린 서술어와 유사한 단어들을 서술어 table에서 선택하여 서술어들의 의미를 찾아낸다. 선택된 서술어들 중 입력된 문장에서 서술어를 제외한 주요 문장 성분들과 의미 관계가 맞는 단어들을 선택한다. 틀린 서술어를 수식하는 부사어가 있을 경우에는, 부사어의 수식을 받을 수 있는 단어들을 선택한다. 주어를 수식하는 관형어가 있을 경우에는, 관형어와 의미가 중복되지 않는 단어들을 선택한다.

4.2 의미적 오류를 포함하는 경우

그림 5의 예에서 알 수 있듯이 의미가 틀린 경우에 대한 교정은 문장 성분들 간의 의미 관계를 이용한다.

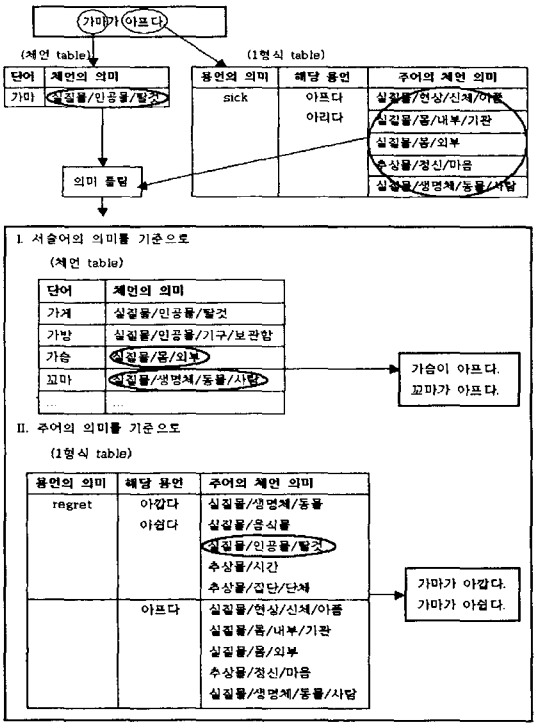

그림 5. 의미적 오류를 포함하는 예

주요 문장 성분들 중 하나인 서술어는 자신의 의미에 따라 나머지 주요 문장 성분들의 구성이 달라지므로 문장의 기본 형식을 한정시킬 수 있다.

의미적 오류에 대한 교정은 어절의 추정이 올바른 경우와 올바르지 못한 경우로 나눈다. 어절의 추정이 올바른 경우에는, 서술어와 나머지 주요 문장 성분의 의미를 비교한다. 서술어를 제외한 모든 주요 문장 성분과 서술어의 의미 관계가 맞는 경우에는, 조사, 관형어 그리고 부사어를 검사한다.

조사를 검사하는 경우에는, 주요 문장 성분 중 체언에서 마지막 글자의 종성 유무에 따라 어절 추정 결과에 만족하는 조사인지 비교한다. 만족하지 않는 조사인 경우 조사의 위치에서 오타가 발생한 경우의 방법과 동일하게 교정한다.

관형어를 검사하는 경우에는, 관형어가 바로 뒤의 체언을 수식할 수 있는지 비교한다. 수식할 수 없는 경우, 관형어의 위치에서 오타가 발생한 경우의 방법과 동일하게 교정한다.

부사어를 검사하는 경우에는, 부사어가 서술어를 수식할 수 있는지 비교한다. 수식할 수 없는 경우, 부사어의 위치에서 오타가 발생한 경우의 방법과 동일하게 교정한다.

1개의 주요 문장 성분의 의미가 틀린 경우에

는, 서술어를 제외한 나머지 주요 문장 성분을 기준으로 하여 교정하고, 서술어의 의미를 기준으로 하여서도 교정한다. 2개 이상의 주요 문장 성분의 의미가 틀린 경우에는, 서술어의 의미를 기준으로 하여 교정한다.

어절의 추정이 올바르지 못한 경우에는, 추정이 잘못된 어절의 위치를 파악할 수 없으므로 주요 문장 성분의 개수에 따라 입력 문장을 형식별로 분류하여 검사한다.

추정된 주요 문장 성분이 2개일 때는 형식1인 경우로 분류한다. 3개일 때는 형식1에서 1개의 부사어나 관형어가 주요 문장 성분으로 잘못 추정되었거나 형식2인 경우로 분류한다. 4개 이상일 때는 형식2 또는 형식3에서 1개의 부사어나 관형어가 주요 문장 성분으로 잘못 추정되었거나 형식 3인 경우로 분류한다.

각각의 경우에서 검사 결과가 틀렸다면, 오타가 발생한 경우의 방법과 동일하게 단어를 교정한다.

V. 문장 교정 실행의 예

의미적 오류가 있는 문장의 실행창과 교정된 문장들의 예가 그림 6에서 보여진다.

그림 6. 의미적 오류가 있는 문장의 실행창과 교정 예

VI. 결론 및 향후 과제

본 논문에서는 단일 문장의 입력 과정에서 발생할 수 있는 오류를 문장 성분 간의 의미 관계를 이용하여 자동으로 교정해 주는 시스템을 구현하였다. 종전의 문법적 요소만을 고려하는 방법에서는 고칠 수 없었던 의미적 오류도 교정 가능함을 알 수 있었다.

향후 과제로는 겹문장에서의 오류 교정 및 여러 문장 간의 의미 파악을 이용한 문단의 오류 교정이 고려되어 진다.

참고 문헌

[1] 김현진, "어절 간 의존 관계와 부분 문장 분석을 이용한 한국어 문법 검사기 구현", 부산대학교 대학원 전자계산학과 석사 학위 논문, 1997

[2] 한창우, "개선된 N-Gram 기법을 이용한 철자교정", 한양대학교 대학원 컴퓨터공학과 석사 학위 논문, 1996

[3] 김동주, "형태소 결합 제약을 강화한 맞춤법 검사기", 한양대학교 대학원 전자계산학과 석사 학위 논문, 1997

[4] 김재원, "한글 맞춤법 오류의 교정 기법에 관한 연구", 부산대학교 대학원 전자계산학과 석사 학위 논문 ,1992

[5] 김광영, "문맥에 의한 중의성 제거와 문장 분석을 이용한 한국어 문법 검사기", 부산대학교 대학원 전자계산학과 석사 학위 논문, 2001

[6] 이영식, 박영자, 송만식, "어절 생성 사전을 이용한 한국어 철자 교정", 정보처리학회 논문지, 제8-B권 제1호, pp.98-104, 2001

[7] 이영식, "사전 근사탐색과 Heuristics를 이용한 한국어 철자 오류 교정 시스템 구현", 부산대학교 대학원 전자계산학과 석사 학위 논문, 1993

[8] 채영숙, "한글 철자 검색기와 교정기의 구현 개발 환경", 부산대학교 대학원 계산통계학과 석사 학위 논문, 1990

[9] 채영숙, "연어 규칙에 기반한 한국어 문서 교정 시스템의 구현", 부산대학교 대학원 전자계산학과 박사 학위 논문, 1998

[10] 남기심, 표준 국어 문법론, 탑출판사, 1998

[11] 백봉자, 외국어로서의 한국어 문법 사전, 연세대학교 출판부, 2002

[12] 성익호, "국어 섭취 동사의 의미 연구", 계명대학교 교육대학원 국어 교육학과 석사 학위 논문, 2002

[13] 최경봉, "명사의 의미 분류에 대하여", 한국어 학회, [한국어학] 4권, pp.11-45,1996

[14] 이재윤, 김태수, "WordNet과 시소러스", 제11회 언어정보 연찬회 발표 논문집, pp.232~269, 1998