Journal of Korean Institute of Intelligent Systems, Vol. 24, No. 2, April 2014, pp. 117-122
http://dx.doi.org/10.5391/JKIIS.2014.24.2.117

사전 기반 최소대립쌍 검색 도구

A minimal pair searching tool based on dictionary

김태훈* · 이재호** · 장문수*†
Tae-Hoon Kim, Jae-Ho Lee, and Moon-Soo Chang†

*서경대학교 컴퓨터과학과, **고려대학교 국어국문학과
† Dept. of Computer Science, Seokyeong University, **Deft. of Korean Language and Literature, Korea University

요 약

최소대립쌍이란 한 음소의 차이만으로 다른 의미를 갖는 단어의 쌍을 말한다. 본 논문은 최소대립쌍을 이용한 국어음운학 연구의 효율성을 위해 최소대립쌍 검색도구를 제안한다. 검색 도구 개발에 앞서 기존 프로그램과 몇 가지 비교 분석을 통해, 개발해야할 한국어 최소대립쌍 검색 도구의 방향을 제시한다. 제안하는 검색도구는 컴퓨터 사용에 익숙하지 않은 국어학자를 위해 키보드 입력을 최소화한 사용자 친화적인 인터페이스를 제시한다. 효율적인 최소대립쌍 연구를 위해 분류 검색 기능을 제공함으로써 더욱 면밀한 최소대립쌍 연구가 가능하도록 한다. 그리고 성능 향상을 위해 유니코드 분석으로 음소를 분리하여 사전 로딩 속도를 향상시키고, 검색의 효율성을 위해 사전 구조를 최적화한다. 검색 알고리즘은 음절 개수를 이용한 해시 탐색으로 검색 속도를 높인다. 제안하는 도구는 초기 버전에 비해 사전 변환 속도는 5배, 검색 속도는 3배 향상되었다.

키워드 : 최소대립쌍, 음절, 음소, 음소 분리, 사전

Abstract

The minimal pairs mean the pairs that have same phonotactics except just one sound in the sequences cause different lexical items. This paper proposes the searching tool of minimal pairs for efficiency of phonological researches with minimal pairs. We suggest a guide to develop Korean minimal pair searching programs by comparing to other programs. Proposing tool has user-friendly interface, minimizing key inputs, for linguistics who are not fluent in computer programs. And it serves the function which classifies the words in dictionary for the detailed researches. And for efficiency, it increases speed of dictionary loading by separating syllables through Unicode analysis, and optimizes dictionary structure for searching efficiency. The searching algorithm gains in speed by hashing algorithm using syllable counts. In our tool, the speed is improved more than earlier version about 5 times at converting dictionary and about 3 times at searching.

Key Words : Minimal pair, Syllable, Phoneme, Phoneme separation, Dictionary

1. 서 론

컴퓨터를 이용한 데이터 구축이 용이해지면서 말뭉치(corpus)를 이용한 많은 연구가 이루어지고 있으며, 이에 따라 말뭉치를 알맞게 가공하는 프로그램들이 많이 개발되고 있다. 이에 비하여 사전을 가공해서 유용한 정보를 추출하는 프로그램은 많지 않다. 말뭉치는 언어 생활에 관하여 이루어지는 연구를 수행하는 데에는 적합한 자료이지만, 언어 체계에 대한 연구를 진행하는 데에는 적합하지 못한 자료이다.

언어 사용자들이 실제 언어 생활 속에서 주로 쓰는 어휘는 사실 그렇게 많지 않다. 세종계획 말뭉치에 포함된 형태소는 총 301,472개로, 이 중에서 출현빈도가 높은 형태소 1위~500위의 출현 빈도를 합하면 79.03%가 된다[1]. 상위 0.16%의 형태소가 말뭉치의 79.03%를 차지하고 있다는 말이다.[1] 즉, 언어 생활에서 사용하는 어휘들은 주로 쓰는 것들만 많이 쓰고 자주 쓰지 않는 어휘들은 거의 쓰지 않는다는 것을 뜻한다. 따라서 언어 상에 존재하는 모든 어휘를 대상으로 하는 연구를 위해서는 사전을 이용한 정보 검색 시스템이 필요하다.

접수일자: 2013년 9월 1일
심사(수정)일자: 2013년 10월 12일
게재확정일자 : 2014년 4월 1일
† Corresponding author
-감사의 글
발음 사전을 제공해주신 고려대학교 국어국문학과 신지영 교수님께 감사드립니다.
본 논문은 2011년도 서경대학교 교내학술연구비의 지원에 의한 연구결과임.

1) 세종 계획 말뭉치에서 빈도를 추출하는 작업은 '꼬꼬마'를 이용하였다. (http://kkma.snu.ac.kr/)

사전을 통해서 얻을 수 있는 정보는 최소대립쌍 유무, 음소배열정보, 발음형 관찰 등이 있다. 본 논문에서는 우선으로 최소대립쌍의 유무를 판별하는 기능에 대해서만 집중적으로 다룬다. 다른 기능 역시 추가될 필요가 있으나 이는 추후 연구로 미룬다.

최소대립쌍은 같은 자리에 있는 한 소리의 음성적 차이 때문에 서로 다른 의미를 갖게 되는 단어의 쌍을 의미한다[2]. 최소대립쌍에서 음성적 차이란 음소(phoneme)의 차이를 의미한다. 간단히 예를 들어, '가방'과 '나방'은 최소대립쌍을 이룬다. 그러나 '물'과 '몸'은 중성과 종성 두 개의 음소가 다르므로 최소대립쌍을 이루지 못한다. 최소대립쌍은 음운론 영역에서 음소를 설정하기 위한 테스트로 중요한 의미를 지닌다. 또, 음운과 변이음을 구분하는 기준으로 최소대립쌍을 제시하고 있다.

이를 종합해보면 최소대립쌍은 한 언어에서 음소 체계를 세우는 데 중요한 역할을 한다고 할 수 있다. 즉, 최소대립쌍은 언어 체계에 대한 연구인 것이다. 따라서 최소대립쌍을 추출하는 작업은 말뭉치가 아니라 사전을 통해서 이루어져야 한다.

그러나 최소대립쌍을 사람이 찾을 경우 사전을 가지고 있더라도 매우 힘든 작업이다. 어휘력이 좋은 사람이라도 수 만개의 단어를 암기하기 어렵고, 이미 알고 있는 단어라고 해도 필요한 순간에 떠올리기가 쉽지 않기 때문이다.

한 단어의 최소대립쌍들을 찾기 위해서는 단어의 음소 위치에 일일이 다른 음소를 대입하고, 이 단어의 존재 유무를 사전을 참고하여 찾을 수 있다. 그러나 이 방법은 시간이 오래 걸리고 번거롭다는 단점이 있다. 게다가 특정 두 음소의 최소대립쌍들, 예를 들어 음소 'ㄱ'과 'ㅂ'이 대립하는 수많은 최소대립쌍들을 찾는 것은 더욱 어렵다. 본 논문에서는 국어학자들의 이와 같은 어려운 점들을 해결하기 위해 최소대립쌍 검색 도구를 개발하고자 한다.

검색 도구의 핵심은 빠른 검색 속도와 편리한 사용법이다. 본 논문에서는 최소대립쌍과 관련하여 기존에 어떤 프로그램들이 있는지 살펴보고, 제안하는 최소대립쌍 검색 도구와 어떤 차이점이 있는지 비교한다. 그리고 도구의 성능 향상을 위해 해시 탐색 기법을 사용하여 기존 최소대립쌍 검색 도구의 검색 속도를 높이고, 한글의 유니코드 패턴을 이용한 음소 분리 알고리즘을 사용하여 발음 사전을 음소 분석 사전으로 빠르게 변환한다. 그리고 키보드 입력을 최소화하고 마우스를 이용한 인터페이스로 빠르고 간편하게 검색 도구를 이용할 수 있도록 한다.

2. 기존 연구

국내에서는 공개된 한국어 최소대립쌍 검색 도구를 찾을 수 없었기 때문에 해외에 공개된 영어 최소대립쌍 검색 프로그램을 비교 대상으로 삼았다. 그 중 LessonPix.com[2]의 웹 기반 프로그램인 Minimal Pairs는 최소대립쌍이 성립하는 두 단어를 입력하면 그와 동일한 패턴의 최소대립쌍 목록들을 찾아준다. 또 하나의 최소대립쌍 검색 프로그램인 Minimal pair finder[3]는 응용프로그램으로 영어의 음소를 입력하여 최소대립쌍 목록을 찾아주는 검색 도구이다. 본 논문에서는 후자인 Minimal pair finder와 자세한 비교를 통해 제안하는 최소대립쌍 검색 도구 개발에 참고한다.

2.1 Minimal pair finder

Minimal pair finder는 UCLA 언어학 교수인 Bruce P. Hayes가 제공하는 최소대립쌍 검색 프로그램이다. Minimal pair finder는 사전을 기반으로 최소대립쌍을 검색하는 애플리케이션 프로그램이다.

이 프로그램에서 사용하는 사전 파일은 일반 사전의 표제어를 소리나는대로 표기한 것으로 구성되어 있다. 예를 들어 사전 파일에 표기되어있는 "AA1 B JH EH2 K T" 라는 단어에서 숫자를 제외하고 공백을 없앤 후 연결하면 "AABJHEHKT" 라는 단어가 된다. 하지만 발음에 유의하여 천천히 읽어 보면, 그 단어의 표제어를 알 수 있다. 예시의 단어는 영어단어 "OBJECT"를 의미한다. 'AA'는 알파벳 'O'의 발음을 표기한 것이고 'B'는 'B', 'JH'는 'J', 'EH'는 'E', 'K'는 'C', 'T'는 'T'와 같다. 알파벳 뒤에 붙는 숫자는 강세(stress)를 의미하고, 문자들 사이에 공백은 한 음소를 구분하기 위한 것이다.

표 1. Minimal pair finder 사전 내 영단어의 발음표기 예
Table 1. Phonetic transcription example of english word in Minimal pair finder dictionary

Phonetic transcription	Headword
AA1 B JH EH2 K T	OBJECT
W EH1 T IH0 NG	WETTING
V IH1 K T ER0	VICTOR
T EH1 L IH0 NG	TELLING
IH0 M Y UW1 N AH0 T IY0	IMMUNITY

최소대립쌍이 음성적 차이에 기초하기 때문에, 단어를 발음으로 표기하는 것이 일반적인 단어를 표기하는 것보다 최소대립쌍 연구에 더욱 도움이 될 수 있다. 따라서 Minimal pair finder의 기반 사전을 수정하여 한국어 최소대립쌍을 찾는 것도 가능하다.

그러나 Minimal pair finder는 한국어 최소대립쌍을 찾는 데 몇 가지 문제점이 있다.

검색 기반 사전을 구성하는 단어들을 한글이 아닌 영어로 표기해야 하며, 모음을 표기하는 방식이 일반적인 영어의 모음 표기와 달라서 이 프로그램을 전문적으로 사용하는 사람이 아니라면 사전 파일에 표기된 단어가 어떤 의미인지 알기 어렵다. 게다가 최소대립쌍 검색 도구의 기반 사전을 구축하기 위해서 사람이 직접 단어를 타이핑해야 하는 수고를 피할 수 없기 때문에, 사전을 구축하는데 영어로 한국어 단어들을 기입하는 것은 시간도 오래 걸리고 표기 오류 가능성이 높아진다. 이러한 문제점들은 사전 구축에 큰 영향을 미치는 문제이다. 그리고 프로그램의 가장 핵심 기능인 검색 속도가 느린 것은 가장 큰 문제인데, 즉각적인 결과 출력을 원하는 사용자 입장에서 10초 이상의 느린 검색 속도는 반드시 개선해야 할 점이다. 그 밖에 최소대립쌍 연구에 있어 중요한 정보가 될 수 있는 각 단어의 품사와 어종 정보가 생략되어 있고, 최소대립쌍을 한번 찾으면 프로그램이 종료되어 다시 실행시켜야 하는 등 지나치게 단순한 인터페이스로 인해 프로그램을 사용하는 데 불편한 점이 있다.

2) 최소대립쌍을 비롯해 말소리를 이용하여 다양한 언어 교육을 컨텐츠를 제공한다.

이를 토대로 새롭게 개발하는 최소대립쌍 검색 도구의 충족 조건은 처음 사용하는 사람도 사전의 내용과 프로그램 사용 방법을 쉽게 이해할 수 있어야 하며, 최소대립쌍 연구에 필요한 부가적인 정보도 추가해야 한다. 그리고 최소대립쌍 연구에 필요한 모든 정보는 가독성과 사전 구축의 용이함을 위해 한글로 구성되어 있어야 한다. 마지막으로 최소대립쌍 검색 속도를 최대한 빠르게 하여 결과를 확인함에 있어 불편함이 없도록 해야 한다.

3. 제안하는 최소대립쌍 검색 도구

3.1 프로그램 구성

제안하는 최소대립쌍 검색 도구는 크게 세 가지 모듈로 구성된다. 첫 번째는 사용자가 최소대립쌍을 찾기 위해 입력한 데이터를 검색에 적합한 형태로 변환시키는 입력 처리 모듈이고, 두 번째는 발음 사전을 음소 분석 사전으로 변환시키고, 메모리에 로드시키는 사전데이터 처리 모듈이 있다. 마지막 검색 모듈은 사전에서 최소대립쌍을 찾아내는 역할을 한다.

입력 처리 모듈은 사용자가 단어를 입력하면 이 단어를 더 작은 단위인 음소로 나누는 음소 분리 모듈과 분리한 음소들을 선택하여 최소대립쌍을 찾을 수 있도록 버튼형식의 인터페이스로 출력하는 인터페이스 구성 모듈로 나뉜다.

사전 데이터 처리 모듈은 사전 변환 모듈, 사전 로딩 모듈 두 가지로 구성된다. 사전 변환 모듈은 발음 사전을 최소대립쌍 검색에 적합한 형태인 음소 분석 사전으로 변환시키는 역할을 하며, 사전 로드 모듈은 변환한 음소 분석 사전을 메모리에 할당시켜 최소대립쌍을 검색할 수 있게 한다.

검색 모듈은 단어 검색과 음소 검색으로 구성된다. 단어 검색은 입력한 단어와 특정한 위치의 음소가 다른 단어를 사전에서 찾는 모듈이다. 음소 검색은 두 음소를 입력하도록 하여, 단어 검색에서 사용했던 단어 검색 모듈을 활용하여 입력한 음소가 포함된 모든 단어들을 찾아낸다. 두 음소를 입력했기 때문에 단어 리스트 또한 두 개가 생성되며, 두 리스트에서 최소대립쌍을 성립하는 짝(pair)을 골라내기 위한 분별 작업을 시행한다. 이 기능을 필터링 모듈이 담당한다. 마지막으로 두 검색 방법 모두 공통적으로 결과를 리스트로 출력하는 결과 출력 모듈이 있다.

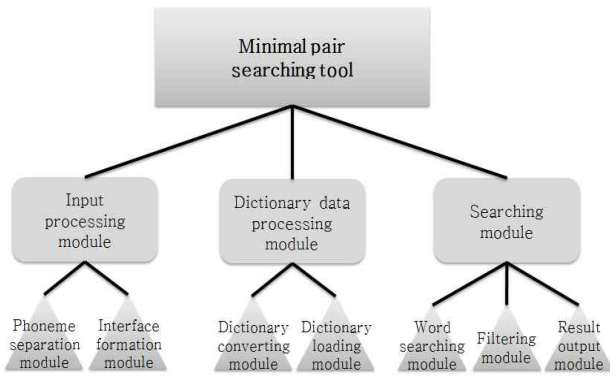

그림 1. 최소대립쌍 검색 도구 구성

Fig. 1. A structure of Minimal pair searching tool

3.2 검색 도구의 기반 사전

본 논문에서 제안하는 최소대립쌍 검색 도구는 사전을 기반으로 한다. 사전의 내용은 Minimal pair finder와 마찬가지로 단어의 발음을 표기한 내용으로 구성한다. 이러한 사전을 발음 사전[3]이라고 하며, 최소대립쌍 검색 도구의 기반이 되는 사전이다.

발음 사전의 내용은 전부 한글로 작성하기 때문에 사전을 새롭게 구축해야 할 때도 편리하다. 게다가 사전 내용이 매우 단순한 구조로 이루어져 있기 때문에 프로그램을 처음 사용하는 사람들도 이해하기 쉽다.

그러나 영어와 달리 음절로 기록되는 완성형 한글로는 음절을 이루는 음소의 분석이 불가능하며[4], 최소대립쌍을 찾기 위한 음소 단위의 데이터 비교를 위해 음절들을 음소 분리한 음소 분석 사전[4]이 필요하다. 음소 분석 사전은 발음 사전의 내용을 그대로 유지하되, 단어를 음소 분리한 결과를 추가로 저장한다.

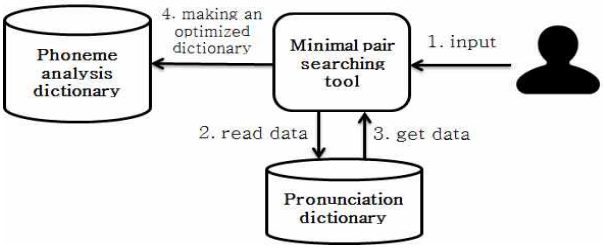

그림 2. 사전 변환

Fig. 2. Dictionary converting

음소 분석 사전으로 변환할 때 이중모음에 대한 처리가 필요하다. 예를 들어 '활성화'(활성화의 발음표기)는 '사' 이중모음을 포함하고 있기 때문에 이것을 'ㅗ'와 'ㅏ' 두 개의 모음으로 볼 것인지, 아니면 'ㅘ'로 하나의 모음으로 볼 것인지 연구 목적에 따라 달리할 수 있다. 사용자는 프로그램 실행 후 나타나는 윈도우에서 이중모음 처리 방식을 선택할 수 있다.

표 2. 음소 분리 결과

Table 2. A result of phoneme separation

Word	Phoneme separation
가공업	ㄱ-ㅏ- ㄱ-ㅗㅇ ㅇ-ㅓㅂ
손자비	ㅅ-ㅗㄴ ㅈ-ㅏ- ㅂ-ㅣ-
어렴푸시	ㅇ-ㅓ- ㄹ-ㅕㅁ ㅍ-ㅜ- ㅅ-ㅣ-
칠쩐팔기	ㅊ-ㅣㄹ ㅉ-ㅓㄴ ㅍ-ㅏㄹ ㄱ-ㅣ-
활썽화	ㅎ-ㅘㄹ ㅆ-ㅓㅇ ㅎ-ㅘ-
	ㅎㅗㅏㄹ ㅆ-ㅓㅇ ㅎㅗㅏ-

3.3 단어 종류의 세분화

최소대립쌍 연구를 위해 제안하는 사전은 표제어와 함께 한국어 단어를 소리 나는대로 표기한 것과 어종, 품사 정보를 저장한다. 어종은 '고유어', '한자어', '외래어' 로 구성되

3) 발음 사전은 한국어 음성학을 연구하시는 고려대학교 신지영 교수님이 제공해주신 발음 사전을 사용했다.

4) 최소대립쌍을 찾는데 적합한 형태의 사전. 한글의 가장 작은 단위인 음소가 표기되어 있다.

어 있으며, 품사는 그림 3과 같이 가장 큰 카테고리인 '자립형태소'와 '의존형태소' 하위에 여러 가지 항목들로 구성되어있다.

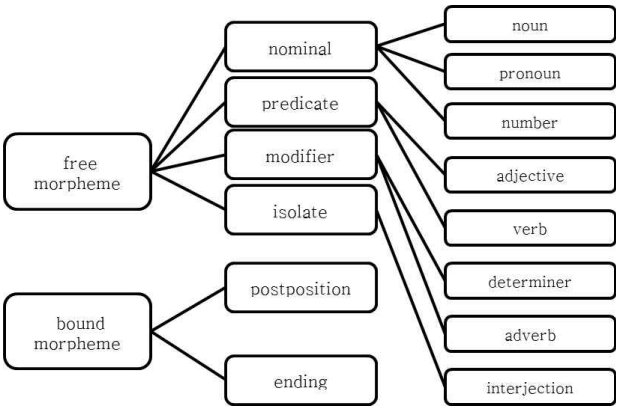

그림 3. 품사 정보
Fig. 3. Word class information

사전에 포함된 어휘들을 그림 3처럼 나누어서 보는 이유는 언어 사용자가 인식하고 있는 어휘 사전의 구조가 이와 비슷하기 때문이다. 인지적 관점[5]에서 볼 때, 유사한 것들은 비슷한 범주로 묶이는 경향이 있다.

최소대립쌍은 이름에서 알 수 있듯이 둘 이상의 쌍(pair)으로 존재한다. 그리고 최소대립쌍은 유사한 두 단어이다. 이들은 인지적으로 비슷한 범주로 묶일 수 있다. 따라서 같은 품사 내에 최소대립쌍이 존재할 가능성이 품사 범주를 넘어서 최소대립쌍이 존재할 가능성보다 더 높다. 같은 범주 내에 존재하는 최소대립쌍들은 의미에 따라 일정 부분 관련성이 있을 가능성이 있다. 예를 들어, '다방'과 '사방' 등은 '어떤 공간'을 지칭한다는 의미를 공유하고 있다. 공유하는 의미는 '-방'의 형식에서 나오는 것으로 보인다. 이처럼 품사별로 나누어서 최소대립쌍을 검색하는 것은 검색된 어휘 간의 의미적 관계를 살펴보기 쉽게 한다. 따라서 사전의 단어들을 종류에 따라 분류시켜 검색함으로써 최소대립쌍 연구에 편리함을 준다는 이점이 있다.

3.4 사용자 친화적 인터페이스

본 논문에서 제안하는 최소대립쌍 검색도구는 두 가지 검색 방법을 제시한다. 하나는 단어 기준 검색이고 다른 하나는 음소 기준 검색이다. 사용자는 각각의 검색 기능을 검색 인터페이스를 통해 접근한다. 본 논문에서는 검색 기능에 맞춰 사용자 친화적인 인터페이스를 제안한다.

첫 번째로 단어 기준 검색 인터페이스는 사용자에게 최소한의 키보드 입력만을 요구하고 나머지는 사용자에게 익숙한 마우스를 사용하게 한다. 그림 4는 이 과정을 나타내고 있으며, 사용자가 '가방'을 입력하면 입력 처리 모듈은 이 단어를 음소로 분리하여, 각 음소를 마우스로 선택할 수 있는 버튼 위에 나타낸다. 사용자가 'ㄱ'을 선택하면 이에 대한 최소대립쌍 결과가 '나방', '다방' 순으로 우측 리스트에 나타난다. 만약 사용자가 어종별, 품사별로 검색을 원한다면 콤보박스 인터페이스를 통해 선택한 카테고리별로 검색할 수 있다.

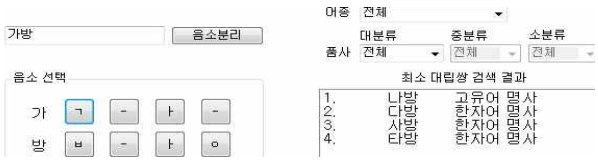

그림 4. 단어 기준 최소대립쌍 검색
Fig. 4. Minimal pair searching for word criterion

두 번째로 음소 기준 최소대립쌍 검색은 사용자가 한 단어의 최소대립쌍이 아닌 두 음소가 대립하는 최소대립쌍을 찾고자 할 때 사용한다. 음소 기준 검색 인터페이스는 사용자가 두 개의 텍스트 입력란에 각각 서로 다른 두 개의 음소를 입력하고, 이 음소의 종류를 선택하는 것만으로 검색을 가능하게 한다. 그림 5는 이 과정을 나타내고 있으며, 사용자가 'ㄱ'과 'ㄴ'을 텍스트박스에 입력하고 음소타입을 초성으로 선택한 후 검색하게 되면, 우측 리스트에 글자 순서대로 사전에 있는 'ㄱ'과 'ㄴ'의 모든 최소대립쌍 리스트가 출력된다. 음소 기준 검색은 사전 내에서 사용자가 입력한 두 음소의 최소대립쌍을 모두 찾기 때문에 최소 수 백가지 이상의 결과가 나타난다.

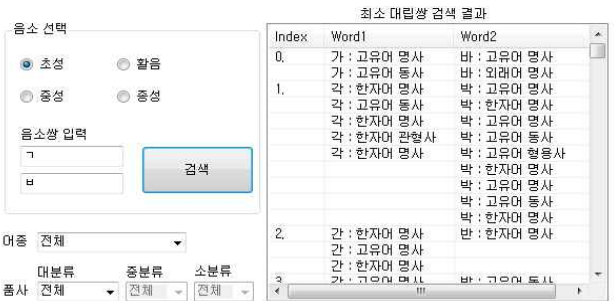

그림 5. 음소 기준 최소대립쌍 검색
Fig. 5. Minimal pair searching for phoneme criterion

제안하는 최소대립쌍 검색 도구의 인터페이스는 Minimal pair finder와 몇 가지 차이점을 가지고 있다.

첫 번째로 가장 큰 차이점은 검색 방법의 다양성이다. Minimal pair finder는 오직 한 가지 방법으로 최소대립쌍 검색이 가능한 반면, 제안하는 최소대립쌍 검색 도구는 단어 검색과 음소 검색 두 가지 검색 인터페이스를 제시하고, 뿐만 아니라 어종, 품사를 선택하여 사용자가 원하는 세부적인 검색이 가능하다.

두 번째는 결과 출력 인터페이스의 차이점이다. Minimal pair finder는 결과를 파일로만 출력하여 프로그램 내에서 최소대립쌍 검색 결과를 확인할 수 없다. 그러나 제안하는 도구는 리스트 인터페이스를 통해 사용자에게 즉각적으로 결과를 보여주며, 필요에 따라 결과를 저장할 수 있도록 한다.

4. 검색 성능향상

검색을 주요 기능으로 하는 프로그램에서 검색 속도는 가장 중요하다. 최소대립쌍 검색은 많은 음소 단위의 비교

5) 이를 비트겐슈타인은 가족닮음(family resemblance)이라는 개념으로 설명했고, 인지언어학에서는 이웃(neighborhood)이라는 용어를 통해서 설명한다. 즉, 형태나 음운, 의미 등이 비슷한 단어들은 인지적인 거리가 가까움을 말하는 것이다.

가 이루어지기 때문에 단순 비교 알고리즘으로는 빠른 검색 속도를 보장하기 어렵다. 이 장에서는 최소대립쌍 검색 성능을 높이기 위해 다음 세 가지 부분, 즉 사전변환, 데이터 구조, 검색 알고리즘에서 속도 향상 방안을 제안한다.

4.1 사전 변환 알고리즘

최소대립쌍을 찾기 위해서는 앞서 설명했듯이 수많은 음소 단위의 비교가 이루어져야 한다. 그러나 발음 사전의 단어는 완성형 한글로 이루어져 있어 음소 분석이 불가능하므로, 단어를 음소 분리하여 구성 음소들을 알아야 한다. 최소대립쌍을 검색할 때마다 단어의 음소를 분리하는 일은 검색 속도 측면에서 매우 비효율적이므로 발음 사전의 단어들을 미리 음소 분리한 음소 분석 사전으로 변환한다.

기존에 사용했던 최소대립쌍 검색 도구의 음소 분리 방법[6]은 미리 정의된 음소 테이블을 참조하여 각 음절의 음소를 분리하는 방법이었다. 즉, 한 음절을 음소 분리하기 위해 수많은 비교를 필요로 했다. 그러나 이 방법은 음소 분석 사전을 만들기 위해 매우 많은 시간이 걸린다.

본 논문에서는 0xAC00부터 0xD7A3 영역에 표현된 한글의 유니코드 패턴을[5] 이용하여 발음 사전의 단어들을 음소 분리하는 방법을 제안한다. 유니코드 패턴을 이용한 음소 분리 알고리즘을 이용해 최소대립쌍 검색에 필요한 몇 가지 중요한 정보를 얻을 수 있다. 음소를 분리하여 각 음절을 이루는 음소의 개수와, 한 단어를 이루는 총 음소의 개수, 이중모음의 유무 등을 알 수 있다. 유니코드 테이블에 한글 초성은 자음, 쌍자음 순서로 19개, 중성은 21개, 종성은 28개(없음 포함)가 규칙적으로 나열되어 있으며, 이 패턴을 이용해 각 음소에 순번을 부여할 수 있다. 표 3은 그 순번을 나타낸 표이다.

표 3. 초/중/종성 인덱스 표
Table 3. Initial/Final consonant and vowel index table

Index	0	1	2	3	4	5	...	26	27
initial consonant	ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	...		
vowel	ㅏ	ㅐ	ㅑ	ㅒ	ㅓ	ㅔ	...		
final consonant		ㄱ	ㄲ	ㄳ	ㄴ	ㄵ	...	ㅍ	ㅎ

유니코드에서 한글은 0xAC00부터 시작되므로 이 정보와 초성, 중성, 종성의 개수를 이용해 한 음절의 음소를 분리하는 공식을 도출할 수 있다. 표 4에 음소 분리 공식을 나타낸다.

4.2 검색 구조 최적화

검색 대상이 되는 사전을 적절한 구조로 메모리에 할당하는 것은 최소대립쌍 검색 속도를 비롯해 도구의 전체적인 효율성에 큰 영향을 미친다. 제안하는 도구에서는 4.1절에서 설명한 음소 분석 사전을 메모리에 효율적으로 할당하여 검색 속도를 향상시키고자 한다.

본 논문에서 제안하는 사전의 구조는 그림 6에 나타나 있으며, 단어-음절-음소 순으로 정보가 자연스럽게 이어질 수 있도록 top-down방식으로 되어 있다. 모든 정보를 포함하는 가장 큰 리스트는 음절 개수가 같은 단어들만 연결시켜 음절 개수를 기준으로 해시탐색이 가능하도록 한다.

표 4. 한글 음소 분리 공식
Table 4. Hangul phoneme separation formula

Value	Formula
syllable unicode	((initial consonant index*21)+vowel index)*28+final consonant index+0xAC00
initial consonant index	((syllable unicode-0xAC00)/28)/21
vowel index	((syllable unicode-0xAC00)/28)%21
final consonant index	((syllable unicode-0xAC00)%28

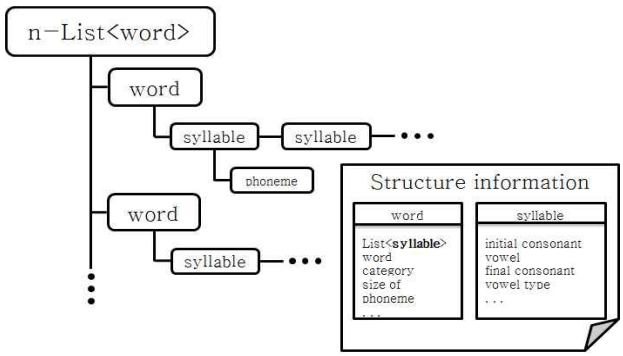

그림 6. 사전 데이터의 계층 구조
Fig. 6. A hierarchical structure of dictionary data

제안하는 최소대립쌍 검색 도구는 Minimal pair finder와 비교했을 때 구조화된 음소 분석 사전의 영향으로 속도 면에서 우위를 보인다. 사전의 표제어가 약 4만 2천 단어인 본 도구는 400쌍 이상의 최소대립쌍을 찾는데 약 10초가 걸린다. 이는 표제어가 약 2만 6천 단어인 Minimal pair finder가 평균 187.6개의 최소대립쌍을 찾는데 약 75초가 소요됨과 비교했을 때, 제안하는 검색 도구가 속도 면에서 월등히 빠른 것을 알 수 있다.

표 5. 최소대립쌍 검색 도구와 Minimal pair finder 비교
Table 5. Comparison of Minimal pair searching tool and Minimal pair finder

	Minimal pair finder	Minimal pair searching tool
number of word	25,953	42,359
number of results	187.6	434.7
avg. time (sec)	75	10

4.3 음소 기준 검색 속도 향상

기존 도구에서는 음소 기준으로 최소대립쌍을 찾기 위해서는 음절의 개수와 상관없이 기입한 음소가 포함된 모든 단어를 상호 비교하여 검색했다. 결국 불필요한 비교로 인하여 계산량이 크게 늘어나게 된다.

본 논문에서는 해싱(hashing) 알고리즘을 사용하여 이러한 문제를 완화하여, 검색 속도를 향상시킨다. 제안하는 알

6) 본 연구실에서 2012년도에 개발하였으나 미공개된 최소대립쌍 검색 도구에서 사용한 알고리즘.

고리즘은 키 값을 음절의 개수로 하여 해시 테이블을 구성한다. 사용하는 발음 사전에서 최장음절단어의 음절 수는 7이다. 따라서 해시 테이블은 1~7 까지 키 값을 가진다. 그림 7은 해시 테이블의 구조를 나타내고 있다.

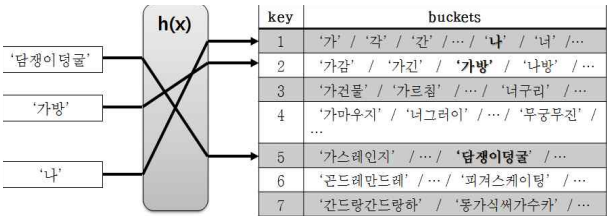

그림 7. 해시 테이블
Fig. 7. Hash table

음소 기준 최소대립쌍 검색은 해싱 알고리즘을 적용하기 전인 초성 검색을 기준으로 평균 434단어를 찾는데 10초 정도의 시간이 걸렸지만, 해싱 알고리즘을 적용하면 3초대의 검색 속도를 나타내어 3배의 속도 향상이 이루어졌다.

표 6. 최소대립쌍 검색 도구의 효율성 개선 전후
Table 6. Before and after about improvement which Minimal pair searching tool efficiency

	Before	After
Searching time(sec)	10	3
Converting time(sec)	10	2

5. 결론

본 논문에서는 국어학자, 특히 음성학 관련 연구자를 위해 효율적인 최소대립쌍 검색 도구를 제안하였다. 공개된 한국어 최소대립쌍 검색 도구가 없는 상황에서 영어 최소대립쌍 검색 프로그램인 Minimal pair finder를 비교 대상으로 하여 새로운 검색 도구를 제안하였다. 검색 도구 개발의 주안점은 편리한 검색 인터페이스와 검색 속도의 향상이었다. 인터페이스는 마우스를 많이 사용하는 오늘날의 컴퓨터 환경에 맞춰 사용자의 키보드 입력을 최소화하는 방향으로 구성했다. 그리고 개발된 도구의 성능을 개선하기 위해 발음 사전에서 음소 분석 사전으로 변환하는 방법을 유니코드 음소 분리 공식을 사용하여 사전 변환 시간을 2초 미만으로 하여 5배 빨라지도록 개선하였고, 해싱 기법을 사용하여 최소대립쌍 검색 속도도 3배 향상되었다

References

[1] Ho-Sung, Jung, “Statistical analysis of Standard Korean Dictionary,” *New Korean Linguistic Life*, Vol. 10 No. 1, pp. 55-72, 2000.
[2] Ji-Young Shin, *Speech Sound of Korean*, 2011. p79-83.
[3] Bruce Hayes, "Small utility programs for phonology," Available: http://www.linguistics.ucla.edu/people/hayes/UtilityPrograms/, 2011, [Accessed: December 10, 2013]
[4] Tae-Jin Kang, "How should we revise new Hangul code?; The standard Hangul code should be Johab code," *Korean Linguistic life*, Vol. 21, No. 2, pp. 116-123, 1990.
[5] Jong-Hoon Kim, *Operation principles of computer*, 2004. p33.
[6] Soon-Bum Lim, Wu-Jin Sim, Yong-Je Lee, Dong-Sun Nam, Hyun-Young Kim, and Yong-Ho Lim, "Requirements for Hangul Text Layout and Typography," Available: http://www.w3.org/TR/klreq/korean/#hangulcodes, 2013, [Accessed: December 10, 2013]

저 자 소 개

김태훈(Tae-Hoon Kim)
2014년 : 서경대학교 컴퓨터과학과 학사
2014년~현재 : ㈜이노와이어리스 연구원
관심분야: HCI, Network system, Natural Language Processing
E-mail : chsk0510@naver.com

이재호(Jae-Ho Lee)
2012년: 고려대학교 국어국문학과 문학사
2012년: 고려대학교 뇌및인지과학연계전공 공학사
2013년~현재 : 고려대학교 국어국문학과 석사과정
관심분야 : Phonology, Phonetics, Korean Linguistics, Natural Language Processing, Speech Recognition/Synthesis System
E-mail : thanatos8023@gmail.com

장문수(Moon-Soo Chang)
1992년 : 고려대학교 전자전산공학과 공학사
1994년 : 고려대학교 전자공학과 공학석사
2001년 : 동경공업대학 지능시스템과학전공 공학박사
2000년~2003년 : 한국전자통신연구원 선임연구원
2003년~현재 : 서경대학교 컴퓨터과학과 부교수
관심분야 : Natural Language Understanding, Knowledge Mining, HCI
Phone : +82-2-940-7754
E-mail : cosmos@skuniv.ac.kr