대용량 전자사전 구축을 위한 국어 대사전의 통계 정보

Statistical Information of Korean Dictionary to Construct an Enormous Electronic Dictionary

김철수, 김양범
서남대학교 컴퓨터정보통신학과

Cheol-Su Kim(chskim@seonam.ac.kr), Yang-Beom Kim(ybkim@seonam.ac.kr)

요약

언어 정보 처리 응용 분야는 정보검색, 형태소분석, 철자검색, 음성인식, 문자 인식 등 다양하다. 이러한 정보처리 과정은 전자 사전이 필수적이다. 본 논문에서는 국어대사전에 대한 기본적인 통계 정보들을 살펴보고, 전자사전 구축에 대하여 알아보았다. 대상 정보는 고어 및 불완전음절을 포함하는 단어를 제외한 표제어들에 대하여, 대사전의 표제어수, 전자사전의 엔트리수, 사용된 전체음절수, 서로 다른 음절수, 엔트리들의 평균 길이, 품사별 분포, 전자사전을 트라이로 구축할 때 사용되는 노드 수 등 이다. 전자사전의 전체 엔트리 수는 361,980개, 사용된 음절수는 1,289,659개로 엔트리들의 평균 길이는 3.56이었으며 서로 다른 음절수는 2,463개였다. 이러한 통계 정보들은 전자사전 구축 및 한국어 정보처리에 도움이 될 것이다.

■ 중심어 : | 국어 대사전 | 전자사전 | 통계정보 | 정보처리 | 음절 |

Abstract

There are various application areas of Language information processing such as information retrieval, morphological analysis, spell checker, voice recognition, character recognition, etc. In these language information processing areas, an electronic dictionary is essential. This thesis made researches on basic statistical information on the Korean dictionary and on the construction of electronic dictionary. The targets of analysis were the number of registered word in Korea dictionary, the entry number of registered word in electronic dictionary, the number of used syllables, the number of different syllables, the average length of entry, the distribution of part of speech and the number of used nodes to construct electronic dictionary using Trie, except for words including a archaic word or incomplete syllables. Total entry number of electronic dictionary is 361,980, the number of used syllables is 1,289,659, the average length of entries is 3.56 and the number of different syllables is 2,463. Theses informations would play a beneficial role in constructing an electronic dictionary and in processing Korean information.

■ keyword : | Korean Dictionary | Electronic Dictionary | Statistical Information | Information Processing | Syllable |

접수번호 : #070514-002
접수일자 : 2007년 05월 14일
심사완료일 : 2007년 06월 13일
교신저자 : 김철수, e-mail : chskim@seonam.ac.kr

I. 서 론

언어 정보 처리 분야는 기계번역, 정보검색, 철자검색, 음성인식, 문자 인식[1] 등 다양하다. 이러한 정보처리 과정은 응용 분야에 따라 형태소분석, 구문분석, 의미분석 단계를 거친다. 형태소 분석은 언어 정보 처리의 기본단계이자 필수 단계라 할 수 있으며 형태소 분석을 수행하기 위해서는 전자 사전이 필수적이다. 형태소 분석 과정에서 분석 후보들의 적합 여부를 판별하기 위하여 전자사전을 매우 많이 참조한다[2]. 전자사전은 응용 분야에 필요한 단어들과 그 단어가 가지고 있는 정보들을 가지고 있다. 전자사전에 저장되는 단어들의 규모, 단어들에 대한 관련정보 및 단어의 검색 방법은 응용 분야의 성격에 따라 달라진다.

전자사전에 저장되는 단어는 1개 이상의 글자들이 결합하여 구성되며, 한글 한 글자(음절)는 초성, 중성, 종성이 결합하여 구성된다. 정보처리를 위한 필수 단계라 할 수 있는 형태소 분석은 글자(음절) 단위의 처리 방법과 자소(초성, 중성, 종성) 단위의 처리 방법이 모두 가능하다. 응용분야의 성격 및 효율성을 고려하여 자소 단위의 처리를 하거나 음절 단위의 처리 방법을 결정한다. 한글 단어 구조는 다음과 같이 정의할 수 있다.

단 어 ::= {음절}$_1^*$
음 절 ::= 열린음절 | 단힌음절
열린음절 ::= 초성 · 중성
닫힌음절 ::= 초성 · 중성 · 종성

컴퓨터에서 표현 가능한 글자는 초성 19개, 중성 21개, 종성 27개가 결합하여 생성되며, 생성 가능한 음절수는 열린음절이 399(19×21)개, 닫힌음절이 10,773(19×21×27)개로 전체 11,172개이다. 11,172개의 음절 중 현대 국어에서 실제로 이용되고 있는 음절은 1/4이 못된다. 현대국어에서 실제로 사용되고 있는 음절은 어떤 음절이며 얼마나 자주 사용되는지 자주 사용하는 음절은 어떤 음절인지 등에 관한 정보는 언어처리 과정에 도움을 준다.

본 연구에서는 국립국어연구원에서 편찬한 50만개를 초과하는 표제어를 담고 있는 표준국어대사전에 대하여 기본적인 통계 정보들을 알아보고, 이것을 전자 사전으로 구축할 때 도움이 될 수 있는 음절정보들에 대하여 고찰한다.

II. 관련연구

사전에 저장된 표제어들에 대한 음절정보를 이용하여 형태소분석기의 성능을 향상시킨 연구[3]가 진행되었다. 형태소분석기에서 취급하는 서로 다른 음절수는 1,996개로, 형태소 분석기에서 일반적으로 사용되는 어휘 사전에 사용되는 음절은 1,734개, 어미에만 사용되는 음절 2개, 용언과 어미가 결합할 때 사용되는 음절이 260개로 전체 1996개의 음절이 사용된다. 이 연구에서는 중소 규모의 국어사전[4]을 기본으로 하였으며 전자사전에 등록된 단어(엔트리) 수는 101,601개였다. 형태소 분석을 위한 어휘 사전에 저장된 엔트리들의 음절길이별 단어 수는 [표 1]과 같다.

표 1. 음절 길이별 단어 수

음절길이	단어수	백분율(%)
1	996	0.981
2	44,946	44.238
3	37,440	36.850
4	14,192	13.968
5	2,903	2.857
6	951	0.936
7	131	4.129
8	34	0.033
9	5	0.005
10	2	0.002
11	1	0.001
합계	101,601	100.000

전자사전에 저장된 엔트리들은 2음절과 3음절인 단어가 82,386개로 전체 엔트리의 대부분인 80%를 차지한다. 101,601개의 엔트리를 나타내기 위해 필요한 음절수는 281,462개($\sum_{i=1}^{11} x_i \cdot y_i$ x_i=음절길이, y_i=단어수)로 평균 음절길이는 2.77이다. 이 단어들을 형태소 분석 환경에 맞도록 한 어휘사전이다. 품사별 표제어 수는 [표

2]와 같다.

표 2. 품사별 표제어 수

품사	개수	백분율(%)
명사	84,307	80.864
의존명사	455	0.436
대명사	188	0.180
수사	98	0.094
관형사	755	0.724
동사	7,961	7.636
형용사	3,306	3.171
부사	6,757	6.481
감탄사	394	0.378
보조용언	37	0.035
합계	104,258	100.000

[표 2]에서 보는 것처럼 명사가 80.6%로 전자사전의 대부분을 차지한다. 용언인 동사 및 형용사의 경우는 10.8%를 차지하여 많지 않음을 볼 수 있다. 품사별 표제어수가 음절길이별 단어 수 보다 많은 이유는 하나의 표제어가 2개 이상의 품사로 사용될 경우 각각의 품사 수에 포함하여 숫자를 계산했기 때문이다.

또한, 형태소 사전에서 제공하는 인접조건을 검사하여 형태소분석을 수행하는 연구가 진행되었다[5]. 이는 기존의 형태소 분석기보다 처리 속도가 매우 빠르다.

전자사전에 저장될 단어들에 대한 저장·검색 방법으로 트라이[6] 구조를 이용할 수 있다. 트라이는 단어들의 prefix를 공유하는 구조로 현재 노드에서 다음 노드로의 분기가 단어 전체가 아닌 단어를 구성하는 문자에 의해 결정되는 트리 구조의 특수한 형태이다. 트라이 구조의 최대 깊이는 트라이를 구성하는 단어들의 최대 길이+2(루트노드, 단어 종료기호 노드)이다. 트라이에 저장된 단어의 검색을 위한 노드 비교 수는 삽입된 단어 수에 무관하게 검색할 단어의 문자열길이+1 이다. 그러나 트라이 본래의 성질을 완벽하게 지원하도록 구현하기란 쉽지 않다. 트라이 구조 구현 방법의 하나로 2개의 배열을 이용하는 방법이 있다[7]. 이 방법은 트라이 성질을 완전하게 유지하면서 배열 공간을 매우 효율적으로 압축해 준다. 삽입할 단어(Key)들의 집합 K= $\{k_1, k_2, k_3 \cdots k_n\}$에 대한 트라이 구조를 다음과 같은 5개의 튜플 <S, I, g, s_1 F>를 가지는 디지털 탐색 트리로 정의한다.

디지털 탐색트리 M = <S, I, g, s_1 F>

S: 트라이에 존재하는 상태(노드)들의 유한집합

I: 단어를 구성하는 기호 또는 문자들의 유한집합

g: S× I→ S ∪ { FAIL }로 사상시키는 전이함수

s_1: 초기노드 또는 S 의 루트 노드

F: accept 노드들의 유한집합(F ⊆ S)

전이함수 g는 단어를 구성하는 임의의 입력기호 a(a ∈I)에 의해서 현재노드 r에서 다음노드 t로 가는 전이함수 $g(r,a)=t$로 표현한다. 전이함수 $g(s_r, a)=s_t$를 두 개의 배열 Base[], Check[]에 다음과 같은 원리를 적용하여 표현한다.

Base[r] + Digit(a) = t, Check[t] = r

전이함수 $g(s_r, a)=s_t$에 대하여 Base[r]+ Digit(a) = t, Check[t] = r 두 개의 조건을 만족하면 현재 노드 r에서 입력문자 'a'에 의하여 다음노드 t로 전이는 성공하지만, 2개의 조건을 만족하지 못하면 전이에 실패하여 현재 검색 중인 단어는 사전에 존재하지 않음을 의미한다.

전자사전에 저장될 단어들의 저장·검색 방법은 다양하다[8][9]. 전자사전은 응용 환경에 따라 저장·검색 방법이 달라진다. 단순 검색의 경우라면 해싱 방법이 유용하지만 음성인식과 및 자연어처리 과정을 위한 형태소 분석 환경은 트라이를 이용한 방법이 효율적이다. 트라이는 사전을 구성하는 엔트리 수에 무관하게 한 단어에 대한 검색시간이 일정할 뿐만 아니라 변형이 일어나는 단어들에 대한 검색에서 변형(활용) 과정을 동시에 조사하면서 검색할 수 있는 효과적인 전자사전 구조이다. 이 연구에서는 배열을 이용하여 트라이 구조를 구현한 방법을 한국어 사전에 적용하여 음절단위, 자소단위, 반음절 단위 방법으로 구축하여 비교·분석 하였다. 검색 시간은 음절단위 방법이 빠르고, 기억장소는 자소 단위방법이 작게 차지한다. 형태소 분석을 위한

전자사전이라면, 형태소 분석 방법에 따라 전자사전도 음절단위 및 자소단위 방법이 결정된다.

III. 사전 고찰

국어대사전[10]은 50만이 넘는 표제어를 7,328쪽에 담고 있는 대사전으로 방대하고 다양한 정보를 담고 있어, 백과사전의 기능을 겸하고 있다. 고어 및 불완전 음절을 포함하는 단어들을 제외한 표제어들 대한 닿소리별 표제어 수는 [표 3]과 같다.

표 3. 대사전의 닿소리별 표제어 수

닿소리	표제어수	백분율(%)
ㄱ	65,524	14.872
ㄴ	17,510	3.974
ㄷ	31,133	7.066
ㄹ	9,589	2.176
ㅁ	26,710	6.062
ㅂ	39,478	8.960
ㅅ	55,706	12.643
ㅇ	69,012	15.663
ㅈ	50,747	11.518
ㅊ	19,484	4.422
ㅋ	5,267	1.195
ㅌ	9,687	2.199
ㅍ	12,574	2.854
ㅎ	28,173	6.394
합계	440,594	100.00

[표 3]의 표제어 숫자는 동음이의어들을 각각 별개로 표제어로 하였다. 가장 많은 표제어를 가진 닿소리는 ㅇ, ㄱ, ㅅ, ㅈ 순서로 69,012, 65,524, 55,706, 50,747개로 각각 15.66%, 14.87%, 12.64%, 11,52%를 차지하여 4개의 닿소리(ㄱ, ㅅ, ㅇ, ㅈ)로 시작하는 표제어가 절반이상(54.69%)을 차지하였다.

표제어로 등록된 단어들의 품사별 분포는 [표 4]와 같다. 품사별 단어수 분포를 보면 명사가 가장 높으며 89%를 차지한다. 이는 표제어의 대부분이 명사임을 알 수 있으며 다른 연구에서보다도 약간 높은 비중을 차지한다.

표 4. 품사별 표제어 수

품사분류	표제어 수	비율(%)
명사	393,490	89.125
대명사	549	0.124
의존명사	1,230	0.279
수사	321	0.073
관형사	612	0.139
형용사	7,392	1.674
보조형용사	26	0.006
동사	17,120	3.878
보조동사	47	0.011
부사	16,455	3.727
감탄사	956	0.217
조사	372	0.084
어미	2,205	0.499
접두사	730	0.165
합계	441,505	100.000

[표 4]에서 보는 것처럼 품사별 표제어수의 합은 표제어로 등록된 단어 수의 합보다 약간 많다. 이는 하나의 표제어가 2개 이상의 의미를 가지고 있으면서, 2개 이상의 품사를 가지는 경우이다. 예를 들어 표제어 "굳다"의 경우, 동사로서의 "단단하게 되다."의 의미를 가지는 경우와 형용사로서의 "흔들림없이 강하다."의 의미를 가지는 경우이다.

1. 전자 사전 구성을 위한 가정

인쇄된 사전에 등록된 모든 표제어를 전자사전의 엔트리로 등록하는 것은 비효율적이다. 따라서 전자사전 구축 목적에 따라 전자 사전에 등록하는 표제어 숫자 및 대상도 달라진다. 본 연구에서는 다음과 같은 조건을 기준으로 전자사전을 구성하였다.

1. 동음이의어들을 1개의 엔트리로 한다. 예를 들어 단어 "가"의 경우 22개 다른 용도 및 뜻을 가지므로 사전에는 22개의 표제어를 가진다. 그러나 전자사전에서는 1개의 엔트리에 모든 정보들을 저장하므로 1개의 엔트리를 가진다.
2. 명사 단어에 "~하다", "~스럽다" 등이 결합하여 용언이 되는 경우 명사 단어를 전자사전의 엔트리로 하였으며 "~하다", "~스럽다" 등이 결합되어 가지는 정보는 정보부분에 표현하는 것으로 가정하였다. 그러나 명사로는 쓰이지 않고 "하다"가 결합한 형태로만 쓰이는 경우(섭섭-하다)는 엔트리에 포

함하였다.

3. 용언의 경우 단어의 끝에 "다" 음절은 사전에 저장하는 것으로 가정하였다.
4. 완전한 음절이 아닌 닿소리 1개만으로 구성된 불완전 음절을 포함하는 표제어는 제외하였다. 예를 들어 'ㄱㄴㄷ순', 'ㄹ변칙'과 같은 표제어들은 엔트리에서 제외하였다.
5. 본 본문의 연구 대상인 국어 대사전은 백과사전 성격을 띠고 있어 옛말, 방언, 전문용어, 오표기, 발음, 문형, 문법 등 국어의 온갖 정보를 포함하고 있어, 현대 국어에서 사용하지 않은 기호 및 고어(예를 들어 'ㄹㄹ디니')들이 표제어애 나타난다. 현대국어에서 사용되지 않은 고어 및 불완전 음절(ㄹㄹ,ㄱ)를 포함하는 표제어들은 제외하였다. 고어 및 불완전 음절이 포함되지 않은 단어, 방언, 옛말, 전문용어 등은 엔트리애 포함하였다.

2. 사전 분석

대사전에 등록된 표제어들을 3.1의 "전자사전 구성조건"을 기준으로 전자 사전을 구축했을 때 엔트리 수는 361,980개이며 닿소리별 엔트리 분포는 [표 5]와 같다.

표 5. 닿소리별 엔트리 수

닿소리	단어 수	백분율(%)
ㄱ	53,756	14.851
ㄴ	14,849	4.102
ㄷ	25,943	7.167
ㄹ	8,220	2.271
ㅁ	22,746	6.284
ㅂ	32,901	9.089
ㅅ	43,576	12.038
ㅇ	56,383	15.576
ㅈ	39,953	11.037
ㅊ	15,692	4.335
ㅋ	4,974	1.374
ㅌ	8,485	2.344
ㅍ	10,935	3.021
ㅎ	23,567	6.511
합계	361,980	100.000

[표 5]에서 보는 것처럼 사전을 구성하는 전체 닿소리 중 ㄱ과 ㅇ으로 시작하는 엔트리 수가 110,139개로 전체 엔트리 중 30.4%를 차지하며 ㄱ, ㅅ, ㅇ, ㅈ으로 시작하는 엔트리 수는 193,668개로 전체 엔트리 중 53.5%나 차지한다. 이는 특정 닿소리로 시작하는 단어가 매우 많음을 의미한다. 사전의 표제어에는 등록되어 있으나 전자사전의 엔트리에는 등록되지 않은 표제어 수는 [표 6]과 같다.

표 6. 닿소리별 동음이의어 및 불완전음절 표제어수

닿소리	표제어수	백분율(%)
ㄱ	11,768	14.87
ㄴ	2,661	3.97
ㄷ	5,190	7.07
ㄹ	1,369	2.18
ㅁ	3,964	6.06
ㅂ	6,577	8.96
ㅅ	12,130	12.64
ㅇ	12,629	15.66
ㅈ	10,794	11.52
ㅊ	3,792	4.42
ㅋ	293	1.20
ㅌ	1,202	2.20
ㅍ	1,639	2.85
ㅎ	4,606	6.39
합계	78,614	100.00

이 표제어들은 동음이의어 형태의 표제어이거나 고어 문자열이 포함된 표제어들이다. 동음이의어의 경우에는 1개의 동일한 엔트리에 표제어에 나타난 수만큼의 관련 정보들이 동일한 1개의 엔트리에 대한 정보들로 저장되어 있다. 따라서 동음이의어들의 모든 정보들은 검색가능하다. 전자사전 구축을 위한 기본 가정에 근거하여 대사전에 등록된 표제어들에 대하여 전자 사전을 구축했을 때 전자 사전에 저장되는 표제어들에 대한 음절별 엔트리 수는 [표 7]과 같다. [표 7]에서 보는 것처럼 3음절 단어와 4음절 단어가 전체의 59.47%를 차지한다. 6음절이상인 단어는 8.12%에 지나지 않는다. 이는 5음절 이하인 단어들이 90%이상임을 의미한다. [표 7]을 그래프로 표현하면 [그림 1]과 같다.

표 7. 음절길이별 엔트리 수

음절길이	단어 수	백분율(%)
1	1,360	0.376
2	77,878	21.514
3	114,981	31.765
4	100,271	27.701
5	38,096	10.524
6	17,785	4.913
7	6,824	1.885
8	2,873	0.794
9	1,129	0.312
10	445	0.123
11	195	0.054
12	86	0.024
13	35	0.010
14	11	0.003
15	3	0.001
16	3	0.001
17	4	0.001
18	1	0.000
합계	361,980	100.000

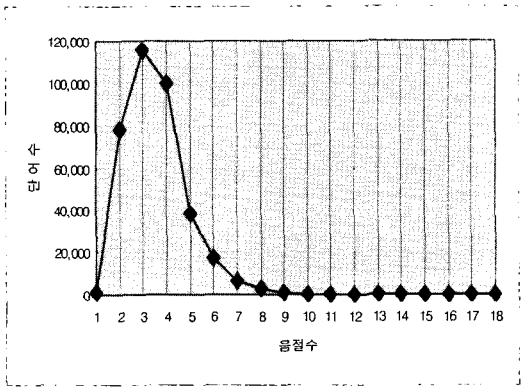


그림 1. 음절길이별 단어 분포 수

등록된 엔트리(단어)는 361,980개이며, 361,980개의 단어를 표현하기 위해 1,289,659개의 한글이 사용되어 단어에 대한 평균 문자열 길이는 3.56이었다[표 8]. 이는 기존의 방법들이 용언 중 품사가 동사인 경우, 단어의 끝에 오는 글자 "다"를 표현하지 않은 반면 본 논문에서는 포함시킨 요인과 등록된 단어들의 증가에 따라 복합명사가 상대적으로 많이 나타나기 때문이다.

표 8. 전체 엔트리에 대한 기본 정보

전체 엔트리 수	사용된 글자 수	단어 평균길이	서로 다른 글자 수
361,980	1,289,659	3.56	2,463

361,980개의 엔트리를 저장하는데 사용된 서로 다른 음절수는 2,463개로 나타났다. 이는 대사전답게 일반 사전에 비해 많은 음절이 나타남을 말해준다. 일반적인 사전의 경우 컴퓨터상에서 표현 가능한 한글 11,172개의 음절 중 2,000개 내외의 음절이 사용된다. 대사전의 경우 2,463개의 음절이 사용되어 500여 음절이 더 사용되고 있다. 이는 분야별 전문용어 및 방언의 사용에 따라 새로운 용어들이 추가되어 증가한 것으로 판단된다.

361,980개의 엔트리에 나타나는 2,463개의 음절에서 1번 출현하는 음절이 372개로 15.1%를 차지하였으며 대표적인 음절은 "힛힐힉흖휀~"이다. 2~10번 출현하는 음절이 592개로 전체의 24.04%를 차지하였다. 가장 많이 출현하는 음절은 "기"로 24,192번 나타난다. "기"이외에도 20,000번 이상 출현한 음절은 3음절 {리, 다, 이}이다. {기, 리, 다, 이} 음절의 출현 빈도는 90,099번 이었다. 이는 전체 음절 출현횟수 1,289,659개의 6.99%에 해당한다. 10,001~20,000 출현한 음절은 {도, 전, 대, 개, 수, 자, 사, 지} 8개였으며, 8개 음절의 출현횟수 합은 10,3476번으로 전체 음절 출현 횟수의 8.02%를 차지하였다. 10,001번 이상 출현하는 12개의 음절의 출현횟수는 전체음절 출현횟수의 15.01%나 차지한다.

표 9. 음절별 출현 빈도수 분포

출현빈도 범위	음절수	음절비율(%)	음절 예
1	372	15.104	힛힐힉흖휀~
2~10	592	24.036	힉헝힛흖흄~
11~100	682	27.690	흖흑펫팟톄~
101~1,000	510	20.706	킹룸눌꽁펀~
1,001~5,000	251	10.191	활김쇠균씨~
5,001~10,000	44	1.786	우적개원관~
10,001~20,000	8	0.325	도전대개수자사지
20,001~	4	0.162	리다이기
합계	2,463	100.000	

단어의 첫음절 위치에 나타나는 음절 종류는 2,194개였으며 가장 많이 출현하는 음절은 '사'로 첫음절 출현 빈도수 4,103 이었다[표 10]. 이는 '사'로 시작하는 엔트리수가 4,103개임을 의미한다. 2,194개의 음절 가운데 1,027개의 음절은 10번미만 나타난다. 즉, 나머지 1,000여개의 음절이 반복적으로 사용됨을 볼 수 있다.

표 10. 첫음절별 위치의 출현 빈도수

출현빈도	음절수	비율	예
4000~	1	0.046	사
3000~3999	10	0.456	가이자고~
2000~2999	15	0.684	무유소비~
1000~1999	70	3.191	일주중어~
100~999	451	20.556	포간로비~
10~99	620	28.259	떨탑혹톱~
1~9	1027	46.809	뗑숨끕깸~
합계	2,194	100.000	

표 11. 둘째음절 위치의 출현 빈도수

출현빈도	음절수	비율(%)	예
5000~	1	0.050	리
4000~4999	3	0.151	기사이
2000~3999	23	1.156	지자수상
1000~1999	191	9.603	조제마~
100~999	894	44.948	천형포~
2~99	549	27.602	찹룸굼~
1	328	16.491	젭석갊~
합계	1,989	100.000	

단어의 둘째 음절에 나타나는 음절의 종류는 1,989개로 나타났다[표 11]. 단어의 둘째 음절 위치에 가장 많이 나타나는 음절은 '리'로 5,636번 나타난다. 한번 나타나는 음절은 328개로 나타나는 전체 음절의 16.5%를 차지하였다.

음절별 출현 빈도수 분포 및 출현 음절 정보는 전자사전 구축하거나 형태소 분석 과정에서 참고 자료를 활용할 수 있다.

3. 트라이 전자사전 구성

361,980개의 엔트리를 트라이를 이용하여 전자 사전을 구축했을 때 트라이 사전의 각 레벨에서 분기 수, 각 레벨에서의 단말 노드수, 비단말 노드수와 현재 레벨을 통과하는 단어 수들을 알아보았다. 트라이 사전 구성에 따른 노드 수는 [표 12]와 같다.

표 12. 트라이 전자사전의 레벨별 노드 구성

레벨	음절	음절이상 단어수	비단말 노드수	단말 노드수	전체 노드수
1	0		1	0	1
2	1	361,980	2,194	0	2,194
3	2	360,620	114,258	1,360	115,618
4	3	282,742	234,377	77,878	312,255
5	4	167,761	155,672	114,981	270,653
6	5	67,490	64,999	100,271	165,270
7	6	29,394	29,689	38,096	67,785
8	7	11,609	11,438	17,785	29,223
9	8	4,785	4,742	6,824	11,566
10	9	1,912	1,901	2,873	4,774
11	10	783	778	1,129	1,907
12	11	339	336	445	781
13	12	143	142	195	337
14	13	57	56	86	142
15	14	22	21	35	56
16	15	11	10	11	21
17	16	8	7	3	10
18	17	5	4	3	7
19	18	1	1	4	5
20	19	0	0	1	1
합계			618,431	361,980	980,411

[표 12]에서 보는 것처럼 루트노드에서 첫 음절에 의해 레벨 2로 분기되는 노드 수는 2,194개, 둘째 음절에 의해 레벨 3으로 분기하는 노드 수는 114,258개, 셋째 음절에 의해 레벨4로 분기하는 노드 수는 312,255개, 넷째 음절에 의해 레벨 5로 분기하는 노드 수는 155,672로 레벨4까지는 노드수가 증가하지만 레벨 5부터는 노드수가 감소한다. 이러한 노드 분포는 전자 사전에 저장되는 엔트리들의 음절 길이 분포와 밀접한 연관을 가진다. 전체 엔트리 361,980개를 저장하는데 필요한 전체 노드 수는 980,411개로 한 개의 엔트리에 소요되는 평균 노드수는 2.71개였다[표 13].

표 13. 전자사전 구성 노드 수

전체 엔트리수	사용된 전체 노드수	단어당 평균 노드수
361,980	980,411	2.71

4. 비교 분석

연구 [2]에 의하면 뉴에이스 국어사전[4]에 수록된 어휘 형태소의 길이는 2.77음절이지만 "대사전"에 수록된 단어의 평균길이는 3.56음절로 0.79절 높게 나타났다. [4]는 2음절 단어와 3음절 단어가 81%로 대부분을 차지하고 있으나 "대사전"은 2음절 단어가 21.51%, 3음절 단어가 31.77%로 2~3음절인 단어는 192,860개, 53.28%로 상대적으로 낮은 편이며, 4음절 단어가 100,271개 27.70%, 5음절인 단어가 38,096개 10.524%를 차지하여 4~5음절인 단어가 138,367개 38.225%를 차지하여 4~5음절 단 단어가 상대적으로 높은 비중을 차지하고 있다.

이는 [4]는 용언의 경우 형태소 단위로 단어를 등록(예, 용언의 끝음절 "다"를 포함하지 않음)하고, 명사의 경우는 등록된 복합 명사 수가 적기 때문이다. [4]는 11절인 단어가 가장 길지만 "대사전"의 경우는 18음절인 단어까지 있다. 이는 대사전의 경우 전문 용어가 많고 복합명사가 상대적으로 많기 때문이다. [4]의 경우 101,601개의 단어 중 6음절 이상의 단어가 1,154개, 1.106%에 불과하다. 그러나 대사전[10]의 경우는 361,980개의 단어 중 29,394개, 8.12%나 되고, 11음절 이상인 단어가 338개 0.93%나 된다. 대사전의 경우 18음절인 엔트리도 있지만 11음절 이상인 엔트리 수는 많지 않음을 볼 수 있다. 이처럼 전자사전에 등록되는 엔트리들의 평균 음절수가 많은 이유는 전자사전에 등록되는 엔트리수 증가에 따른 복합명사 개수가 증가하기 때문이다.

[2]의 연구에서는 형태소 분석기에서 사용하는 음절이 1,996개였으나 "대사전"은 2,463개로 467개의 음절이 더 사용되었다. 이는 사전에 등록된 단어수의 증가로 인해, 일상생활에서 잘 사용되지 않은 새로운 단어들, 전문용어 및 신생어의 등록에 기인한 것으로 판단된다.

361,980개의 엔트리(사용 글자수: 1,289,659개, 평균길이: 3.56)를 가진 대사전을 트라이 구조를 이용하여 전자사전을 구축할 경우 980,411개의 노드 중 618,431개의 노드가 비단말 노드이므로 전형적인 리스트 구조를 이용하여 트라이 구조를 구현한다면 1개의 노드는 자식 링크필드(4Byte), 글자저장 필드(2Byte), 단어종료기호(1Byte), 형제 링크필드(4Byte)로 구성되므로 1개의 노드당 11바이트 필요하다. 따라서 단말노드를 제외한 618,431개의 노드를 표현하기 위하여 6.8MByte (618,431*11)가 필요하며, 모든 노드를 2개의 배열에 표현 한다면 7.8MByte (980,411*8)가 필요하다. 그러나 리스트 구조는 전자 사전에 저장되는 엔트리 수가 증가함에 따라 1개의 단어를 검색하기 위한 검색 시간이 증가하지만, 2개의 배열을 이용하여 트라이를 표현한 방법은 전자사전에 저장되는 엔트리 수에 무관하여 단어의 길이에만 의존하여 검색시간이 결정되므로 검색 시간 매우 빠르다.

Ⅳ. 결론

본 연구에서는 범용의 전자 사전으로 이용할 수 있는 전자 사전의 구축을 위한 "표준 국어 대사전"의 통계 정보에 대하여 살펴보았다. 50만개가 넘는 표제어에서 불안전 음절과 고어를 포함하는 음절을 제외한 표제어는 441,594개였으며, 전자사전 구축에 따른 엔트리 수는 361,980개였다. 전자사전 구축에 사용되는 엔트리 361,980개에 출현하는 서로 다른 음절수는 2,463 개였다. 전자사전에 출현하는 2,463개의 음절 중 단 1번 출현하는 음절은 372개였다. 361,980개의 엔트리에 사용된 전체 음절수는 1,289,659개이고, 단어의 평균 길이는 3.56이였다. 1,289,659개의 음절에서 가장 많이 출현하는 음절은 "기"로 24,192번 출현하였으며, 20,000번 이상 출현하는 음절이 4개였다. 길이가 1음절인 단어의 엔트리 수는 1,360개 이었으며, 가장 많은 음절수를 가진 단어는 "프로테스탄티즘의윤리와자본주의의정신"으로 18음절 이었다. 361,980개의 엔트리를 트라이를 이용하여 구축할 경우 980,400여개의 노드를 가진다. 이는 1개의 엔트리당 평균 2.71개의 노드를 사용한다.

361,980개의 모든 단어를 검색하기 위해서는 전체 1,651,641개의 노드를 검색하여 단어당 평균 4.56개의 노드를 방문한다. 이는 단어를 검색하기 위해 5개 정도의 노드를 검색이 필요함을 의미한다. 형태소 분석기에서 분석 대상이 되는 후보들에 대한 검색 과정에서 트라이 성질을 이용한다면 노드 검색 수는 이보다 줄어든다.

본 연구는 국어 대사전에 대한 기본적인 다양한 통계 정보 제공과 엔트리들을 트라이를 이용한 전자사전으로 구축할 나타나는 노드 관련 다양한 유용한 정보들은 제공하므로 써 한국어 처리를 위한 형태소 분석 과정 및 전자사전 구축에 도움을 줄 것으로 기대된다. 그러나 단어를 검색했을 때 필요한 정보들에 관한 연구는 언급하지 못했다. 전자사전 구축은 응용분야에 따라 저장·검색하는 방법과 검색했을 때 필요로 하는 정보들에 대한 기준들은 달라진다. 보다 객관적이고 완벽한 범용의 전자사전 구축을 위한 정보에 대한 표준과 정보처리 분야에서 공통적으로 사용할 수 있는 전자 사전 구축에 관한 연구가 필요하다.

참 고 문 헌

[1] 유진희, 이종혁, 이근배, "형태소 분석과 언어 평가를 이용한 문자인식 후처리", 정보과학회 논문지(B), Vol.22, No.6, pp.880-891, 1995.

[2] 강승식, *음절정보와 복수어 단어 정보를 이용한 한국어 형태소 분석*, 서울대학교 공학박사 학위논문, 1993.

[3] 강승식, 장병탁, "음절 특성을 이용한 범용 한국어 형태소분석기 및 맞춤법 검사기", 정보과학회 논문지(B), Vol.23, No.5, pp.530-539, 1996.

[4] 금성출판사 사서부, *뉴에이스 국어사전*, 금성출판사, 1989.

[5] 심광섭, 양재형, "인접 조건 검사에 의한 초고속 한국어 형태소 분석", 정보과학회 논문지: 소프트웨어 및 응용, Vol.31, No.1, pp.89-99, 2004.

[6] E. Fredkin, B. Beranek, and Newman, "Trie memory," CACM, Vol.3, pp.490-499, 1960.

[7] J. I. Aoe, "An Efficient Digital Search Algorithm by Using Double-array Structure," IEEE Transaction on S/W Eng., Vol.15, No.9, pp.1066-1077, 1989.

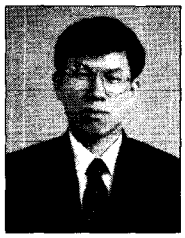
[8] 김철수, 배우정, 이용석, J. I. Aoe, "이중 배열 트라이 구조를 이용한 한국어 전자사전 구축", 정보과학회 논문지(B), Vol.23, No.1, pp.85-94, 1996.

[9] 김철수, *한국어 형태소 분석 환경을 효율적으로 지원하는 전자사전 구조*, 전북대학교 이학박사 학위논문, 1998.

[10] 국립국어연구원, *표준국어대사전*, 두산동아출판사, 1999.

저 자 소 개

김 철 수(Cheol-Su Kim) 정회원



- 1987년 2월 : 전북대학교전산통계학과(이학사)
- 1987년 2월 : 전북대학교전산통계학과(이학석사)
- 1998년 8월 : 전북대학교전산통계학과(이학박사)
- 1995년 3월 ~ 현재 : 서남대학교 컴퓨터정보통신학과 교수

<관심분야> : 자연어처리, 정보검색, 지식표현, U-Learning

김 양 범(Yang-Beom Kim) 정회원

- 1987년 2월 : 전북대학교전산통계학과(이학사)
- 1987년 2월 : 전북대학교전산통계학과(이학석사)
- 1995년 3월 ~ 현재 : 서남대학교 컴퓨터정보통신학과 교수

<관심분야> : 데이타베이스