# 음절 기반 형태소 분석을 위한 효율적인 사전 구성

김 남 철, 서 영 훈
충북대학교 컴퓨터공학과

# An Efficient Dictionary for Syllable-based Korean Morphological Analyzer

Nam-Churl Kim, Young-Hoon Seo
Dept. of Computer Engineering, Chungbuk National University

## 요 약

형태소 분석기의 처리 속도는 분석 알고리즘과 형태소 사전의 탐색 기법에 따라 크게 좌우된다. 형태소 분석 성능의 향상을 위하여 많은 형태소 분석 방법이 제안되었으며, 음절 정보를 이용하는 형태소 분석기는 한국어 음절의 통계적 특성 정보를 이용함으로써, 분석 후보의 개수를 최대한 적게 하여 처리 속도를 향상시켰다. 본 논문은 형태소 분석시 발생하는 분석 후보들의 특성을 고려하여 사전 탐색 요구시 가장 많은 처리 시간을 필요로 하는 디스크 읽기 횟수를 줄일 수 있도록 음절별 블록 인덱싱한 사전 구성 방법을 제안한다. 이 방법은 형태소 사전을 첫 음절별로 블록화하고 인덱싱하여 3개의 추가적인 인덱스 테이블을 구축하는 사전 구성 방법으로, 인덱스 테이블을 모두 주기억장치에 적재하였을 때에는 평균 61.6%, 크기가 작은 두 개의 인덱스 테이블만 주기억장치에 적재하였을 때에는 평균 25%의 디스크 읽기 횟수를 줄일 수 있다.

## 1. 서 론

많은 자연 언어 처리 응용 시스템들이 전처리 단계로 형태소 분석 단계를 필요로 한다. 따라서 형태소 분석기의 성능 향상은 형태소 분석기 자체의 성능 향상 뿐만 아니라 이를 사용하는 응용 시스템의 전체적인 성능의 향상과도 관련되어 중요한 의미를 가진다.

형태소 분석기의 성능은 분석률과 처리 속도로 평가된다. 분석률은 전체 실험 단어 중에서 정확하게 분석된 단어의 비율로, 알고리즘의 효율성에 크게 의존한다. 한편 처리 속도는 단위 시간내에 분석하는 어절 수로, 알고리즘의 효율성과 함께 형태소 사전 탐색의 효율성에 크게 좌우된다.

형태소 분석의 처리 속도를 향상시키기 위해서는 효율적인 알고리즘으로 분석 후보의 생성시간을 줄이고, 불필요한 분석 후보의 생성을 최대한 줄이며, 사전 탐색 속도를 빠르게 하는 것이다.

형태소 분석의 성능 향상을 위해 많은 형태소 분석 방법이 제안되어져 왔으며, 그 중 음절 정보를 이용한 분석 방법[1,2]은 한국어 음절의 통계적 특성 정보를 이용하여 불필요한 분석 후보의 생성을 최대한 줄여 사전 탐색 횟수를 줄임으로써 처리 속도의 향상을 가져왔던 방법이다. 이 분석 방법으로 한 어절당 평균 사전 탐색 횟수를 Tabular Parsing법의 경우 33.91회보다 월등히 향상된 3.57회로 줄일수 있었다[1].

형태소 사전은 그 양이 방대하기 때문에 일반적으로 주기억장치에 적재하기보다는 보조기억장치에 적재된 채로 사전 탐색이 이루어진다. 그런데 보조기억장치인 디스크의 읽기나 쓰기의 연산은 다른 연산 처리속도보다 월등하게 느리다. 따라서 디스크 읽기를 최대한 줄이는 것이 형태소 분석 처리 속도를 향상시킬 수 있는 방법이 된다.

본 논문은 형태소 분석 처리 속도의 향상을 위해 분석시 발생하는 분석 후보들의 특성을 이용하

여 효율적인 사전 구성 방법을 제안한다. 이 사전 구성 방법은 일반적인 형태소 분석기에 적용될 수 있으며, 특정적인 사전 구조를 요구하는 형태소 분석 방법이 아니라면 어떤 형태소 분석기에라도 적용될 수 있다.

## 2. 형태소 분석과 형태소 사전

형태소 분석 알고리즘은 형태소 사전의 구성에 의존하게 되며, 또한 어떤 형태소 분석 기법을 사용하느냐에 따라서 형태소 사전의 구성도 달라지게 된다. 일반적으로 Two-level 모델에서는 입력문자열과 사전을 일치시키면서 two-level의 규칙에 따라 음운 현상을 처리하기 때문에 TRIE 구조로 사전을 구성하여야 한다[3,4].

그러나 음절 정보를 이용하는 형태소 분석기는 분석 후보를 모두 생성한 다음에 사전을 참조하게 되므로, 사전의 구조와 상관없이 처리 속도가 가장 빠른 방법을 사용하면 되기 때문에 비교적 사전 구성 방법의 제약이 없다. 일반적으로 디스크 읽기 속도를 고려하여 사전을 일정한 크기로 분할한 후, 그 분할된 영역의 첫 번째 단어들을 key로 인덱싱을 하고, 그 인덱스를 주기억장치에 적재하는 블록 인덱싱 기법을 사용하거나[1], 품사별 음절 길이별로 사전을 논리적으로 분할하는 방법을 사용하였다[2,5].(이후 '분할형 사전 구성'으로 칭하기로 한다.)

형태소 분석을 위한 사전은 크게 문법 형태소 사전과 어휘 형태소 사전으로 나눈다.

### 2.1 문법 형태소 사전

문법 형태소 사전은 형태소 분리 및 단어의 유형을 인식하는 등 형태소 분석에 미치는 영향이 매우 크며, 조사와 어미들에 대한 정보들로 이루어진다. 조사의 경우에는 하나 이상의 조사가 서로 결합 될 수 있으므로 이를 알고리즘으로 처리하는 경우와 서로 결합된 경우를 하나의 새로운 조사 항목으로 사전에 구성하는 방법이 있으나, 처리 효율상 후자의 경우가 선호된다.

문법 형태소 사전은 참조되는 횟수가 잦을 뿐만 아니라, 사전의 크기 또한 크지 않기 때문에 주기억 장치에 올려놓고 사용하는 것이 효율적이다.

### 2.2 어휘 형태소 사전

어휘 형태소 사전은 크기가 방대하기 때문에 형태소 분석 처리 속도에 매우 큰 영향을 미친다. 따라서 어휘 형태소 사전에 관한 논의는 주로 처리 속도와 관련된 사전의 구조에 집중되어 왔다[6,7].

영어의 경우 어휘 출현 빈도에 매우 자주 사용되는 사전(UFD : ultra-high frequency dictionary)과 매우 자주 사용되는 사전(HFD : high frequency dictionary)을 일반 어휘 사전과 구분하여 처리하는 방법이 많이 사용되지만, 한국어의 경우에는 형태소에 대한 출현 빈도가 계량화 되어있지 않으므로 음절 정보를 이용하는 형태소 분석기의 경우 분할형 사전 구조를 이용한다[1,2].

어휘 형태소 사전은 수록되는 어휘수에 따라 '12만 어휘 사전'과 소규모 사전을 중심으로 한 '6만 어휘 사전'을 구성하여 사용하기도 한다. 어휘 형태소의 빈도수 정보를 이용하면 어휘의 수가 많은 사전이 더 좋은 형태소 분석 결과를 낼 수도 있지만, 빈도수 정보를 이용하지 않는 형태소 분석기는 '6만 어휘 사전'을 이용한 형태소 분석기가 더 좋은 결과를 생성하기도 한다[5].

본 논문에서 제시하는 음절별 블록 인덱싱 어휘 형태소 사전 구성은 음절 정보를 이용한 형태소 분석기를 이용하여 '6만 어휘 사전'과 '12만 어휘 사전'을 대상으로 실험하였다.

## 3. 효율적인 사전 구성과 사전 탐색

형태소 분석 과정은 보통 한 어절에 대해, 문법 형태소의 검사, 원형 복원, 결합 조건 검사 등 분석 알고리즘에 의하여 분석 후보들을 생성하고, 그 후보들 중에서 어휘부에 대하여 어휘 형태소 탐색이 이루어지고, 그 결과 어휘가 사전상에 존재하면 올바른 분석 결과로 선택되어지고, 그렇지 않으면 분석결과에서 배제되는 과정을 거친다.

이 과정 중 어휘 형태소 사전 탐색 요구는 매우 빈번히 발생하며, 효율적인 형태소 분석기를 위해서는 이 사전 탐색 속도가 빨라야 한다. 그러기 위해서는 우선 형태소 분석기에서 생성되는 후보들의 특성을 분석하는 것이 필요하다.

### 3.1 형태소 분석 후보의 특성

다음은 '나는 학교에 갑니다'라는 문장의 각 어절에 대해 생성되는 형태소 분석 후보들이다.

1) '나는'에 대한 분석 후보:
   ① 나(noun) + 는(josa)
   ② 나(verb) + 는(eomi)
   ③ 날(verb) + 는(eomi)

* ④ 나늘(verb) + ㄴ(eomi)
* ⑤ 나는(etc)
** ⑥ 나느(verb) + ㄴ(eomi)
** ⑦ 나능(verb) + ㄴ(eomi)

2) '학교에'에 대한 분석 후보:
① 학교(noun) + 에(josa)
* ② 학교에(etc)
* ③ 학교에(noun)

3) '갑니다'에 대한 분석 후보:
① 가(verb) + ㅂ니다(eomi)
② 갈(verb) + ㅂ니다(eomi)
* ③ 갑니다(etc)
* ④ 갑니다(noun)
* ⑤ 갑니(noun) + 다(josa)
* ⑥ 갑니(verb) + 다(eomi)

위의 '나는'에 대한 7가지 분석 후보들 중 ⑥, ⑦은 음절 정보를 이용하는 형태소 분석기에서는 분석 후보에서 배제 되므로 최종 분석 후보로는 ①~⑤ 까지가 된다. 이 후보들 중에 최종 결과를 얻기 위해서 사전을 탐색하여야 하며, 각각의 분석 후보에 대해 한번의 디스크 읽기와 읽은 블록 내에서의 이진 검색이 필요하게 된다.

그런데 '나는'에 대한 분석 후보들 중 ①, ②, ④, ⑤는 모두 첫음절이 같다는 특성을 가진다. '학교에'의 경우는 3개의 분석 후보 모두가 첫 음절이 같고, '갑니다'의 경우는 6개의 분석 후보들 중 4개가 첫 음절이 동일하다. 또한 1)-①,②, 2)-②,③, 3)-③,④, 3)-⑤,⑥의 쌍들은 서로 같은 어절에 대해 서로 다른 품사 가능성을 위한 사전 탐색을 요구한다.

이와 같이 형태소 분석시 발생하는 분석 후보들은 분석 후보들 대부분의 첫 음절이 같은 특성을 갖는다. 이 특성을 이용하여, 형태소 사전을 첫 음절별로 블록화하고, 각각의 블록을 인덱싱하는 방법으로 구성하면, 사전의 각 블록은 첫 음절이 모두 같은 단어들로 구성된다.

그러면, 첫 음절이 같은 분석 후보들에 대해서 사전 탐색 요구시, 맨 처음 한 번만 해당되는 블록에 대해 디스크 읽기를 수행하여 주기억장치에 적재를 하고, 그 적재된 블록내에서 이진 검색만으로 나머지 후보들의 사전 탐색이 이루어진다. 이것은 첫 음절이 같은 분석 후보의 사전 탐색 요구시 반복적으로 발생하는 디스크 읽기의 횟수를 감소시켜 전체적인 형태소 분석 처리 속도를 향상시킨다.

위의 '나는'이라는 어절의 경우 5번의 디스크 읽기에서 2번의 디스크 읽기로 줄일 수 있다. 이는 내부적인 처리 속도보다 디스크 읽기 속도가 월등히 느리다는 점에 비추어 볼 때 많은 속도 향상을 가져온다.

### 3.2 어휘 형태소 사전의 구성

어휘 형태소 사전은 3개의 인덱스 테이블과 하나의 어휘 정보 파일로 구성된다.

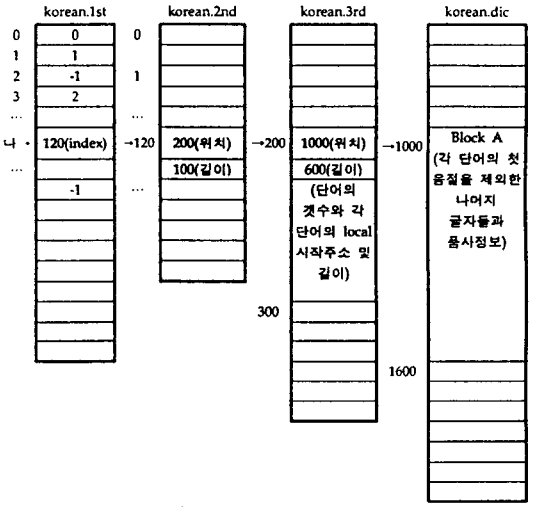


그림 1.어휘 형태소 사전 구성도

그림 1은 어휘 형태소 사전의 구성도와 함께 음절 '나'로 시작되는 어휘 형태소 블록을 찾아가는 과정을 나타낸 것으로서 korean.1st, korean.2nd, korean.3rd는 모두 인덱스 테이블이 되며, 이러한 테이블들은 사전 탐색시 주기억장치에 적재하며, 마지막 korean.dic만 보조기억장치에 적재한다.

위와 같이 사전을 구성함으로써, 첫 음절이 '나'인 어휘에 대한 사전 탐색시 Block A에 해당하는 부분을 한 번만 주기억장치에 적재한 후, 나머지 후보들에 대한 사전 탐색시에는, 이미 주기억장치에 적재되어있는 Block A내에서 이진 검색만 수행함으로써 사전 탐색을 수행한다.

korean.dic의 어느 한 블록을 읽기 위해서는 korean.1st, korean.2nd, korean.3rd등 3차에 걸친 인덱스 값을 구해야 한다. 이 인덱스 값을 구하는 시간을 줄이기 위해 korean.1st의 인덱스 값은 한글 코드 값 자체를 인덱스화한 값으로서, 사전 탐색을 요구하는 어절의 첫 음절로부터 쉽게 인덱스 값이 구해진다.

korean.2nd와 korean.3rd는 korean.dic의 각 블록에 대한 위치 및 길이뿐 만 아니라 그 블록내에 있는 각 단어의 지역적(local) 위치 및 길이에 대한 정보도 포함하며, 그 블록이 주기억장치에 적재가 된 후 이진 검색할 때 사용된다.

인데스를 위한 테이블은 개수가 적을수록 처리 시간상 효율적일 수 있지만 테이블 자체의 크기가 상대적으로 커지는 관계로 본 논문에서는 3개의 테이블로 구성하였다.

표 1은 '6만 어휘 사전'과 '12만 어휘 사전'의 경우 구성된 테이블 및 어휘 정보의 크기를 나타낸다.

표 1. 각 테이블 및 어휘정보의 크기 (단위:Kbytes*)

| 사전구분 | korean.1st의 크기 | korean.2nd의 크기 | korean.3rd의 크기 | korean.dic의 크기 |
|---|---|---|---|---|
| 6 만 어휘사전 | 30 | 9 | 129 | 440 |
| 12 만 어휘사전 | 30 | 12 | 262 | 1,004 |

* 1 Kbytes = 1,024bytes

## 4. 실험 및 결과

'6만 어휘 사전'으로 음절별 블록 인덱싱 사전을 구성하였을 경우, 그림 1에서 Block A의 최대 크기는 4 Kbytes 이내이며 같은 음절로 시작되는 단어의 최대 갯수는 약 650개가 된다.

분할형 사전 구성은 음절별 블록 인덱싱 방법으로 구성된 사전에 비해 디스크 읽기시에 읽는 블록의 크기가 작으며, 또한 블록 내의 단어 갯수가 적다. 하지만 일반적인 컴퓨터 하드웨어의 특성상, 디스크 읽기시 한 번에 읽는 블록의 크기는 일정하다. 예를 들어 컴퓨터 하드웨어 기종마다 다르지만, 일반적으로 Workstation기종 이상에서는 8 Kbytes, PC 기종에서는 1 Kbytes가 된다. 따라서 기종에 따라 1byte를 읽는 속도나 8Kbytes 또는 1Kbytes를 읽는 속도는 같게 된다. 그래서 음절별 블록 인덱싱 방법으로 구성된 사전의 경우 블록의 최대 크기가 4Kbytes 이므로 디스크 읽기시에 분할형 사전 구성과의 차이는 거의 없다. 또한 주기억장치내에서 이진검색의 검색 시간이 $O(\log_2 n)$이고 n의 최대크기가 650이므로, 분할형 사전 구성인 경우와 비교하면 탐색횟수는 거의 차이가 없을뿐 만 아니라, 디스크 읽기 시간에 비하면 무시 가능한 차이이다.

이와 같이 사전을 구성한 후, 무작위 추출된 61,391 어절에 대하여 형태소 분석 후보를 생성하고 최종적인 사전 탐색을 필요로 하는 횟수를 실험해 본 결과 표 2와 표 3의 결과를 얻었다.

표 2. 사전 탐색 횟수 실험 결과

| 처 리 어절 수 | 분석후보 갯 수 | 사전탐색 시 디스크 읽기횟수 | 첫음절이 동 일 한 분석후보 | 특수문자가 포함된 분석후보 |
|---|---|---|---|---|
| 61,391 | 299,275 | 109,820 | 184,370 | 5,085 |

표 3. 한 어절에 대한 평균 사전 탐색 횟수

| | 분석후보 갯 수 | 사전탐색 시 디스크 읽기횟수 | 첫음절이 동 일 한 분석후보 | 특수문자가 포함된 분석후보 |
|---|---|---|---|---|
| 한 어절당 평 균 값 | 4.87 | 1.79 | 3.00 | 0.08 |
| 분석후보에 대한 백분률 | | 36.8% | 61.6% | 1.6% |

이 실험에서 사용된 형태소 분석기는 범용 형태소 분서기로서 입력 어절에 대한 가능한 모든 분석 결과를 생성해내도록 구현된 시스템이며, 실험 결과에서도 볼 수 있듯이, 한 어절에 대한 평균 사전 탐색 요구 횟수는 4.87회가 발생한다. 이 횟수는 형태소 분석기에 사용된 알고리즘에 따라 다소 차이가 있을 수 있는 수치이다.

그 중 첫 음절이 동일하거나, 또는 단어 자체가 동일하여 디스크 읽기를 수행하지 않은 횟수는 전체의 61.6%에 해당하는 3.00회가 발생하며, 따라서 실제 디스크 읽기가 요구되는 것은 36.8%에 해당하는 1.79회 뿐이다.

따라서 본 논문에서 제안하는 방법으로 사전을 구성한다면 평균 61.6%에 해당하는 디스크 탐색횟수를 줄일 수 있으며, 디스크 읽기시 소요되는 처리시간이 다른 연산처리보다 월등히 많은 시간을 필요로 한다는 것을 감안한다면 이는 매우 효율적이라 할 수 있다.

여기에서, 표 1의 각 테이블이 크기를 살펴보면 korean.3rd의 크기가 각각 129Kbytes와 262Kbytes로, korean.1st와 korean.2nd에 비해 상대적으로 매우 큰 편이다. 따라서 주기억장치의 소모를 줄이기 위하여 korean.3rd 테이블도 디스크에 저장한 채로 사전 탐색을 할 경우, 약 40Kbytes의 비교적 적은

주기억장치만으로도 사전 탐색을 할 수 있으며, 이 경우 발생하는 디스크 읽기의 횟수는 표 2의 디스크 읽기 횟수의 2배인 3.58회가 되며, 이 때 또한 평균 25%의 사전 탐색 횟수를 줄일 수 있다.

## 5. 결 론

본 논문에서는 형태소 분석시 처리 속도의 향상을 위하여, 형태소 분석 후보들의 첫음절이 대부분 같은 특성을 이용하여, 형태소 사전을 첫 음절별 블록 인덱싱 방법으로 구성하였다.

실험 결과에서도 보았듯이 형태소 분석시 많은 시간을 필요로 하는 디스크 읽기의 요구 횟수는 평균 61.6%가 감소되었으며 세 번째 인덱스 테이블을 주기억장치에 적재하지 않은 경우에도 평균 25%가 감소되었다.

그러나 하드웨어에 따라 다른 연산처리 속도와 기억장치와의 처리 속도의 극심한 차이를 극복하기 위해 캐쉬(Cache) 메모리를 따로 두고 있기 때문에 디스크 읽는 횟수의 감소에 대해 정확히 비례적으로 처리속도도 감소하는 것은 아니다.

## 참 고 문 헌


[1] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 컴퓨터공학과 박사학위논문, 1993년 2월.

[2] 장동수, 음절에 기반한 한국어 형태소 분석기, 충북대학교 컴퓨터공학과 석사학위논문, 1994년 2월.

[3] K. Koskernnienmi, "Two-level Model for Morphological Analysis," Proceedings of the 8th International Joint Conference on Artificial Intelligence, pp. 683-685, 1983.

[4] D.B. Kim, S.J. Lee, K.S. Choi, G.C. Kim, "Two-level Morphological Analysis of Korean," Proceedings of the 15th International Conference on Computational Linguistics, Vol.1, pp.535-539, 1994.

[5] 강승식, "한국어 형태소 분석기 HAM의 형태소 분석 및 철자 검사 기능", 제 8회 한글 및 한국어 정보처리 학술발표 논문집, pp.246-252, 1996.

[6] 백대호, 이 호, 이임창, "Finite State Transducer를 이용한 한국어 전자 사전의 구조", 제 7회 한글 및 한국어 정보처리 학술발표 논문집, pp.181-187, 1995.

[7] 이승선, 송주원, 조완섭, 황규영, 최기선, "Compact TRIE index: 한국어 전자 사전을 위한 데이터베이스 색인 구조", 정보과학회 논문지, 22권 1호, pp.3-12, 1995.

[8] 금성출판사 사서부, 뉴에이스 국어 사전, 금성출판사, 1989.

[9] 민중서림 편집국, 국민학교 민중 새국어사전, 민중서림, 1992.