사전 성능개선을 통한 한국어 형태소분석기의 분석속도 향상

김영관,박민식,최진석,권혁철
부산대학교 전자계산학과

Improvement of Analysis Speed in Korean Morphological-Analyzer Using Ameliorated Dictionary

Young-kwan Kim, Min-sik Park, Jin-suk Choe, Hyuk-Chul Kwon
Department of Computer Science, Pusan National University

요 약

본 논문에서는 사전 구조와 탐색알고리즘을 개선하여 형태소분석기의 분석 속도를 향상시켰다. 형태소분석기의 분석시간은 사전탐색과 제약검사의 비중이 크다. 따라서 형태소분석기의 처리속도는 사전 탐색 기법에 많은 영향을 받는다. 본 논문에서는 한국어 형태소분석기에서 사용되는 사전의 탐색속도 향상과 한 문서에 나타나는 동일한 어절에 대해서 cache를 사용하여 형태소분석기의 처리 속도를 빠르게 하였다. 또한 기존의 형태소분석기에서 속도 증가를 위해 사용하는 어절-형태소분석결과 사전을 활용하여 더 발전시켰다. 본 논문에서는 어절-형태소분석결과 사전을 사용할 때, 분석 속도향상을 위한 새로운 가속기법인 '하이브리드(HyBrid)'방법을 사용하여 어절-형태소분석결과 사전의 적중률을 높였다.

1. 서 론

인터넷에 접속되는 서버의 수가 늘어나고 있으며, 홈페이지의 수도 기하 급수적으로 증가하고 있다. 정보화가 가속되어감에 따라 컴퓨터로 처리해야 하는 정보의 양이 급격히 증가했다. 우리 나라에서는 처리해야할 정보의 대부분이 한국어로 되어있다. 따라서 한국어 정보처리가 핵심 요소로 부각되고 있다. 한국어 정보를 처리하기 위해서 가장 기본적인 작업은 한국어 형태소 분석 작업이다. 형태소 분석 작업은 사전 탐색이 많고, 제약 검사를 위한 규칙도 많이 있어 분석 속도가 느리다. 따라서 빠른 형태소 분석기가 필요하게 되었다.

형태소 분석기의 분석 속도는 단위 시간에 분석한 어절의 수로 평가되어진다. 형태소 분석기를 빠르게 하기 위해서 많은 연구가 있었으며, 현재도 진행 중이다.[1][2]

형태소 분석기의 처리 속도 향상을 위한 방법은 형태소 분석기의 내부 알고리즘을 개선하는 방법과 원시 말뭉치에서 어절의 빈도 정보를 이용하는 방법이 연구되어 왔다.[3]

소설을 읽을 때 주인공의 이름이나 특정 지명이 반복해서 나타난다. 이처럼 일상생활에서 사용하는 문서에서도 같은 어절이 반복적으로 사용된다. 한 문서 내에서 반복적으로 나타나는 어절을 매번 새로 형태소분석을 하는 것보다는 이전에 분석한 결과를 다시 사용하면 형태소분석기의 분석속도는 빠를 것이다. 본 논문에서는 한 문서 내에서 반복 사용된 어절은 cache를 사용하여 중복 분석을 막고 분석 속도를 향상 시켰다. cache를 사용한 방법은 짧은 문서보다 긴 문서에서 더 효과적이다.

형태소 분석기에서 사용하는 기본 사전구조는 trie구조이다. trie구조는 어절을 찾아가는 과정에서 찾아야할 단어를 모두 찾을 수 있는 장점이 있다. 하지만 trie구조의 형재 노드 탐색 과정에는 순차탐색이 포함되어 있어 사전의 탐색속도를 느리게 한다. 본 논문에서는 기존의 trie구조를 개선하여 종성의 탐색과정에서 발생하는 순차탐색을 제거하였다.

형태소 분석기의 속도향상을 위해 지금까지 연구된 방법 중

빈도가 높은 어절을 미리 형태소분석하고 그 결과를 사전으로 만들어 사용하는 기법이 있다. 본 논문에서는 이 기법을 보완하여 더 빠른 형태소분석 기법을 개발하였다. 지금까지의 방법은 입력된 어절 전체가 어절-형태소분석결과 사전에 존재할 때만 형태소분석결과를 사전에서 가져왔다. 본 논문에서는 어절-형태소분석결과 사전에 어절이 존재하지 않을 때의 형태소분석 알고리즘을 개선하였다.

즉 어절-형태소분석결과 사전에 입력어절이 없어서 형태소분석을 하며, 분석할 때 기본사전에서 명사를 찾았다면, 찾아진 명사의 뒷부분만 가지고 다시 어절-형태소분석결과 사전을 탐색한다. 만일 어절-형태소분석결과 사전에서 남은 어절을 찾았다면 제약 검사만 하고 분석을 끝낸다. 이 방법은 복합명사인 어절을 빠르게 분석한다.

2. 형태소분석기의 개선

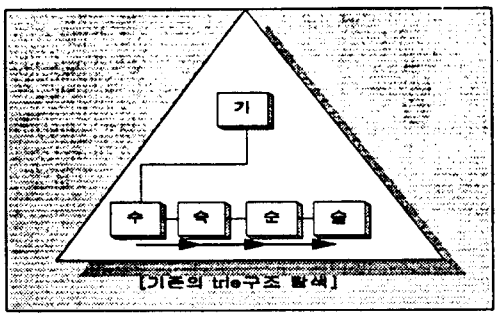
2.1 기존의 trie사전을 개선한 변형 trie사전

형태소분석기에서 사전탐색이 차지하는 비중은 높다. 따라서 형태소분석기에서 사용하는 사전 구조인 trie사전의 빠른 탐색을 위한 연구가 많이 있었다.

기존의 연구방법은 trie사전의 음절 인덱스 수를 늘려 사전을 탐색속도를 빠르게 하는 연구였다.

본 논문에서는 trie구조의 형제 노드 탐색과정에서 발생하는 순차탐색을 제거함으로써 사전의 탐색 속도를 빠르게 하였다. 본 논문에서는 기존의 trie구조를 개선하여 종성의 탐색과정에서 발생하는 순차탐색을 제거하였다.

[그림1]은 기존의 trie구조로 단어를 찾을 때의 과정이다.



[그림 1] 기본 TRIE구조 사전의 탐색 과정

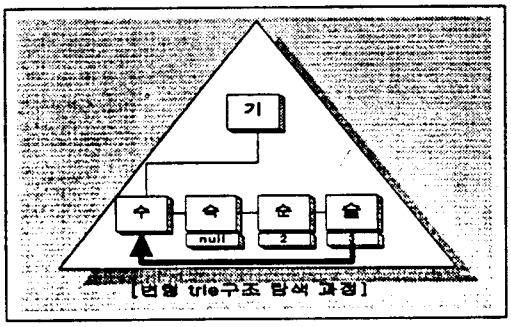
trie사전을 "기수", "기숙", "기순", "기술" 4개의 단어로 만들면 [그림1]의 구조가 된다.

사전의 구성은 아래와 같이 된다.

[기] ➡ [<수> - <숙> - <순> - <술>]으로 만들어진다.

찾고자 하는 단어가 "기술"이라면, '술'의 종성인 'ㄹ'은 종성조사이므로 '기수'와 '기술' 모두를 찾아봐야 한다. 따라서 '기수'를 찾은 다음에 다시 '숙', '순', '술'을 거쳐서 '기술'을 찾게된다. 즉, '기수'에서 시작해서 종성들을 순차 탐색하여 '술'을 만나거나 '술'보다 큰 값이 나올 때까지 계속 탐색을 하게 된다. 이처럼 기존의 trie구조에서는 마지막 단계에서 종성부분을 탐색하는 과정이 순차 탐색으로 되어 사전의 탐색 시간을 많이 소비한다.

본 논문에서는 기존의 trie구조를 개선하여 종성을 찾아가면서 발생하는 순차 탐색 과정을 제거하였다. 변형된 trie구조에서는 [그림2]에 나타나 있다.



[그림 2] 변형 TRIE구조 사전 탐색 과정

변형 trie구조에서는 '기술'을 찾을 때 '기수'와 '기술'을 따로 찾는 것이 아니라 먼저 '기술'을 찾는다. '기술'을 찾은 후 '술'에서 정보를 읽어 '수'의 위치를 계산하는 방법이다. '술'에 추가의 정보 필드가 '3'을 가지고 있는 것은 '술'에서 종을 제거한 음절이 자신의 앞쪽 3번째 노드라는 것을 의미하고 있다. 만약 종성을 제거한 음절에서 단어가 되지 않으면 'null'을 가지게 되어 종성을 제거한 음절을 찾는 과정이 생략된다. 따라서 변형 trie구조에서는 종성을 찾아가는 순차 탐색과정이 없어진다.

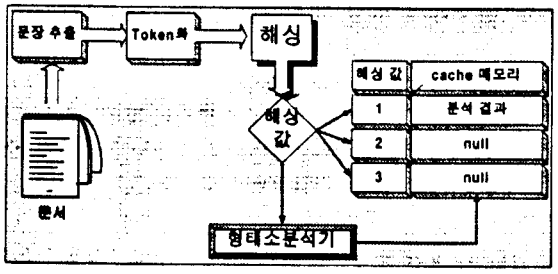
이 방법은 분석을 위해서 사전을 탐색할 때의 발생하는 종성의 순차 탐색과정을 사전을 만드는 과정으로 옮겨 놓은 것이다. 즉 사전을 만들 때 종성의 종류와 종성이 없을 때 단어가 되는지 등의 정보를 미리 넣어 두는 것이다.

2.2 한 문서 내에서 반복하여 사용된 어절의 cache적용

일반적인 data의 특징 중 하나가 좁은 지역에 밀집해서 나타나는 지역성(locality)이다. 이러한 특성은 어절도 마찬가지다. 소설책에서도 주인공의 이름은 많이 나타난다. 주인공 이외의 등장인물들도 일정 시간동안 계속해서 사전을 전개해 나간다. 뉴스에서도 일정시간동안은 동일 인물이나 같은 지역에 대한 기사가 집중적으로 나온다. 이처럼 일상생활의 문서에서도 시간의 지역성을 가지고 있다. 본 논문에서는 이러한 지역성의 특성을 반영하여 cache를 구현하였다. cache는 한 문서에서 반복적으로 나타나는 어절을 한 번의 형태소분석으로 처리한다. 어절을 분석하기 전에 cache에 어절의 분석결과가 있는지 확인해야한다. 그리고 이 과정은 빠른 탐색 시간을 요구한다. 본 논문에서는 이 과정을 해싱을 사용하여 구현하였다. cache의 알고리즘은 어절이 입력으로 들어오면 해싱을 하고, 해싱에서 반환된 번지에 분석결과가 있는지 확인한다. 분석결과가 있으면 분석결과를 가져오고, 없는 해당 번지에 입력어절의 분석결과를 기록하는 방법이다.

[그림3]은 cache를 적용한 형태소 분석기이다.

cache를 사용한 방법은 짧은 문서보다 긴 문서에서 더 효과적이다.

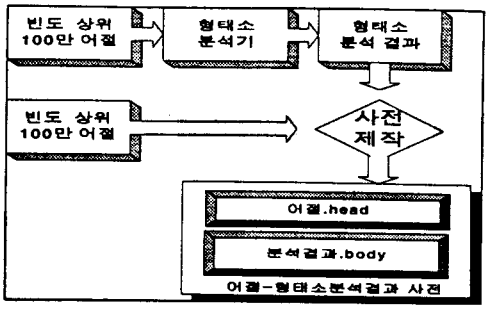


[그림 3] cache를 적용한 형태소 분석기

2.3 어절-형태소분석결과 사전의 응용(HyBrid)

대용량의 말뭉치와 어절의 빈도정보를 이용하여 형태소 분석기의 속도를 증가시키려는 연구가 진행되었다. [3]

어절의 빈도정보를 추출하고 빈도가 높은 어절 중에서 일부를 미리 형태소분석하고 그 결과를 사전으로 만들어 형태소 분석의 속도를 높이는 방법도 연구되었다. 100만 어절을 사용한 예를 들면 아래와 같다.



[그림 4] 어절-형태소분석사전을 만드는 과정

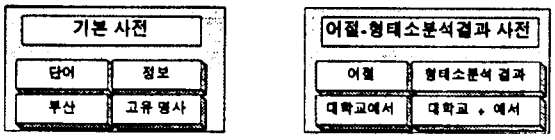
[그림4]는 어절-형태소분석결과 사전을 만드는 과정이다.

말뭉치에서 빈도가 높은 순서로 100만개의 어절을 추출한다. 그 100만개의 어절을 형태소분석을 하고, 분석 결과와 입력 어절 100만개를 가지고 어절-형태소분석결과 사전을 만든다. 어절-형태소분석결과 사전은 형태소분석기에서 입력어절을 분석하기 직전 탐색하여 사전에 결과가 있으면 형태소 분석 없이 바로 결과만 가져오는 기법이다.

본 논문에서는 이 기법을 개선하여 더 빠른 형태소분석 기법을 개발하였다. 지금까지의 어절-형태소분석결과 사전을 사용한 방법은 입력된 어절 전체가 어절-형태소분석결과 사전에 존재할 때만 형태소분석결과를 사전에서 가져왔다. 하지만 복합명사와 같이 긴 어절은 앞부분의 명사를 제거하고 나면 짧은 어절이 되고 빈도가 높은 100만 어절에 포함될 확률이 높아진다. 본 논문에서는 이 점에 착안하여 복합명사로 된 어절을 형태소 분석하는 과정에서 어절-형태소분석결과 사전을 사용하여 분석 속도를 가속화하였다.

다음의 예를 가지고 설명하겠다.

[그림 5]와 같이 어절-형태소분석결과 사전에는 "대학교에서"만 들어있고 "부산대학교에서"는 들어있지 않다고 가정한다.

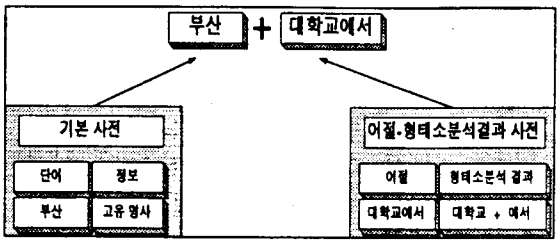


[그림 5] 기본사전, 어절-형태소분석결과 사전의 구성

입력어절은 "부산대학교에서"라고 가정하면, 형태소분석기는 "부산대학교에서"를 분석하게 된다.

"부산대학교에서"를 기본사전에서 찾으면 "부산(고유명사)-"로 시작하는 분석 후보가 생긴다. [그림6]에서와 같이 "부산"을 사전에서 찾은 후 "부산"을 제외한 남은 부분인 "대학교에서"를 다시 어절-형태소분석결과 사전에서 찾는다. "대학교에서"를 어절-형태소분석결과 사전에서 찾은 후 "부산"과

"대학교"의 제약 검사 단계로 진행한다.



[그림 6] 기본사전, 어절-형태소분석결과 사전의 구성

제약 검사에서 결합이 가능하면 분석은 끝나게 된다. 만약 결합이 불가능하게 되면 추가의 다른 분석이 가능한 지를 형태소분석 루틴이 분석한다.

어절의 빈도 정보를 이용하여 추출된 어절들은 대부분 짧은 어절로 되어 있다. 이 방법은 짧은 어절은 대부분 빈도가 높으므로 어절-형태소분석결과 사전에 들어있고, 복합명사인 긴 어절은 출현빈도가 상대적으로 낮아 어절-형태소분석결과 사전에 빠져있는 사실을 반영하였다. 따라서 빈도가 낮은 복합명사도 어절-형태소분석결과 사전을 사용하여 빠르게 분석할 수 있다. 부산대학에서는 이 방법을 "하이브리드(HyBrid)"라 이름 지었다.

3. 실험 및 결과

1. TRIE사전들의 메모리 사용량과 탐색 속도

결과 분석에서는 3가지 종류의 사전으로 비교한다.

Type A: 1음절 indexed TRIE 사전
Type B: 기존의 2음절 indexed TRIE 사전
(두 번째 음절의 중성까지를 인덱스로 사용)
Type C: 변형된 2음절 indexed TRIE 사전
(두 번째 음절 전체를 인덱스로 사용)

기존의 2음절 사전(Type B)은 분석 시간에 종성을 띄고 찾아보아야 할 지 판단하지만 변형된 2음절 사전(Type C)은 사전을 만드는 단계에서 미리 이러한 계산을 거쳐 정보를 역으로 찾아갈 수 있는 정보를 사전 안에 넣게 된다. 따라서 기존의 종성을 찾아가는 시간이 없어진다.

변형된 2음절 사전의 메모리 사용량은 아래 [표3]과 같다.

종류	인덱스로 사용하는 음절의 크기	사용된 메모리	
		인덱스 헤드 노드의 수	헤드 파일 크기(byte)
Type A	1음절	1357개	8,148
Type B	1음절 + 2음절의 중성까지	25,969개	207,754
Type C	1음절 + 2음절	37,323개	298,586

[표 3] 사전 종류별 메모리 사용량

(단위 : 초)

종류 (인덱스 크기)	9만 어절 탐색	18만 어절 탐색	27만 어절 탐색
Type A (1음절 인덱스)	0.98	1.98	3.01
Type B (2음절 중성)	0.49	0.95	1.44
Type C (2음절)	0.44	0.88	1.31

[표 4] 사전 종류별 탐색 속도

Type B와 Type C를 비교해 보면 변형된 2음절 TRIE사전이 기존의 2음절 사전보다 약 10%의 성능이 개선되었다.

위의 실험은 Pentium II 450MHz에서 테스트하였다. 사전의 내용은 88,247개의 단어로 된 기본사전이고, 입력된 사전의 단어 전체를 다시 탐색해 보는 방법으로 실험하였다. 탐색 시간이 짧아서 같은 내용을 2번(18만 어절), 3번(27만 어절) 반복하여 탐색 속도 측정하였다.

여기서 어절을 탐색 방법은 입력어절에서 생성 가능한 모든 단어를 찾아 주는 것을 말한다. 즉, 입력어절이 "학교"이면, 결과로 "학", "학교"가 찾아지는 것을 말한다.

2. HiBrid의 적용 어절 수

실험 어절은 ETRI에서 MATEC99를 위해서 대회 참가자에게 배포한 원시 말뭉치을 사용하였다. 빈도가 높은 100만어절을 분석한 사전은 부산대학에서 보유하고 것으로, 약 1억 어절에서 추출하였다.

입력어절 9만 어절 중에서 약 90%인 8만 어절은 빈도가 높은 100만 어절에 포함되었으며, 나머지 약 1만 어절 중에서 HiBrid방법으로 분석된 어절의 수는 23.7%인 2,376개였다. 따라서 HiBrid 방법은 100만 어절을 캐시로 사용할 때의 적중률

올 약 2.3%높인 결과이다.

실제 분석에서 HiBrid가 적용된 어절의 예를 보면 아래와 같다.

입력어절	기본사전의 명사		어절-말뭉치분석 결과 사전의 어절 (분석결과 내용)
신비사상가이자	신비	+	사상가이자
			사상가+이다+자
신앙공동체이다	신앙	+	공동체이다
			공동체 + 이다
순수과학에서의	순수	+	과학에서의
			과학 + 에서의
실제사건들이	실제	+	사건들이
			사건+들+이
학자관료들은	학자	+	관료들은
			관료+들+은
이익집단으로서의	이익	+	집단으로서의
			집단+으로서의
언어예술에서	언어	+	예술에서
			예술+에서

[표 5] 분석과정에서 HyBrid가 적용된 예

참고문헌

[1] 권혁철, "급증하는 대용량 한국어 문서의 색인 방법과 주제 탐색 기법 개발", 정보통신부 최종보고서, 1999

[2] 김남철,서영훈 "음절 기반 형태소 분석을 위한 효율적인 사전 구성", 제9회 한글 및 한국어정보처리 학술대회, pp.411-415 , 1997

[3] 김민정, "규칙과 말뭉치를 이용한 한국어 형태소 분석과 중의성 제거", 박사학위 논문 , 1997

[4] 강승식,"한국어 형태소 분석기 HAM의 형태소 분석 및 철자 검사기능", 제8회 한글 및 한국어 정보처리 학술발표 논문집, PP.246-252, 1996

[5] Edward A.Fox 외 3인, "Practical Minimal Perfect Hash Functions for Large Databases", CACM ,1992

[6] http://ourworld.compuserve.com/homepages/bob_jenkins/