문단내 유사 단어수의 조사 및 문단 기반 한국어 형태소 분석*

이 은 주 · 이 무 상
(대전산업대학교)

Rhee, Eunjoo and Lee, Moosang. 1999. **Survey of the Number of Similar Words in Paragraph and Korean Morphological Analysis Based on Paragraph.** *Studies in Language,* 15-1, 203-219 In this paper, a new morphological analysis model of paragraph base has been presented to solve the problem decreasing a number of word, a target to search in dictionary, which is a fault in phrase or sentence unit morphological analysis. Korean paragraphs have been analyzed to prove the validity of paragraph unit analysis which is not traditional analysis method of sentence unit by sentence-grammar base, over 2,166 paragraphs and 91,668 phrases in the book of middle school, high school, and vocational institute. The validity of paragraph unit analysis has been proved by the analysis result that the appearance frequency of similar words is 34.35% in confidence interval of 95%±0.701. The experiments of Korean morphological analysis have been executed by using the proposed model based upon proved results. The efficiency of the proposed model is shown by the result of the analysis that the number of dictionary search is decreased 45.1% than the traditional method of phrase unit analysis, at the appearance frequency of similar words is 41.2% in a paragraph. (**Taejon National University of Technology**)

1. 서 론

형태소(morphology)란 의미가 있는 최소의 단위 또는 문법적, 관계적인 뜻을 나타내는 단어 또는 단어의 구분으로 정의할 수 있다. 형태소는 자립성의 유무에 따라 자립 형태소와 의존 형태소로 구분할 수 있으며, 기능에 의해 어휘 형태소와 문법 형태소로 구분할 수 있다(김영택, 1994. Ishizaki, 1995).

형태소 분석(morphological analysis)은 자연 언어 분석의 첫 단계로서 단어 단위로 분리된 입력 문자열로부터 각각의 형태소를 분리하고, 용언의

* 이 논문은 1996학년도 대전산업대학교 연구비에 의하여 연구되었음.

불규칙 활용이나 굴절 현상이 일어난 단어에 대해서는 원형을 복원하는 과정이다. 따라서, 형태소 분석은 자연 언어의 제약조건인 문법 규칙에 맞는 '분석 후보들을 어떻게 생성할 것인가?' 하는 문제와, 공기제약(concurrence restriction)과 같은 상호 제약 조건과 사전에 의해 '분석 후보로부터 옳은 결과를 어떻게 선택할 것인가?' 하는 문제를 해결하는 것이라 할 수 있다(김영택, 1994).

지금까지 이러한 형태소 분석을 위해 여러 가지 방법론들이 연구되었고, 보편적으로 한국어 형태소 분석시 입력 단위로는 어절 또는 문장으로, 형태소 분리 방법으로는 음절 단위 분석법이 많이 사용되고 있다. 이는 한글의 특성에 맞게 음절 단위로 함으로써 통계적 음절 특성을 이용하는 것이다(김재훈, 1996. 강승식a, 1996).

그런데 형태소 분석기의 성능평가 요소는 단어의 분석률과 처리 속도이다. 이중에서 형태소 분석기의 처리속도는 분석에 사용되는 알고리즘의 효율과 어휘사전의 탐색 속도에 좌우된다. 분석 알고리즘의 효율은 형태소를 분리하고 그 원형을 복원하는 과정의 복잡도에 의존한다. 어휘사전 탐색속도는 사전의 참조 횟수와 탐색 시간에 의하여 결정되며, 이 탐색속도가 시스템의 성능을 좌우한다. 이것은 입출력 연산을 해야 하는 사전 탐색은, 비교 연산을 주로 하는 분석 알고리즘에 비해 훨씬 많은 시간이 소요되기 때문이다. 그래서 형태소 분석의 성능 향상을 위해서는 비교연산의 수를 줄이는 것 뿐 만 아니라 사전 탐색 시간을 줄이는 것이 매우 중요하다. 또 사전 탐색 시간을 줄이는 문제에서도 사전 탐색 속도보다는 사전 탐색의 대상이 되는 단어 자체의 수를 줄이는 것이 분석기 성능 향상에 보다 효과적이다(김영택, 1994. 강승식b, 1996).

그러나, 현재까지 연구된 어절 또는 문장 단위의 형태소 분석, 즉 상황이나 장면에서 문을 독립적으로 취급하는 문문법(sentence grammar) 기반의 문장 단위 연구에서는 이러한 사전 탐색의 대상이 되는 단어수를 감소하는데 비효율적이라는 문제점이 있다. 따라서 본 논문에서는 이러한 문제점을 해결하기 위하여 문보다 큰 단위층인 이야기의 틀 내에서 우리가 통상 사용하는 말과 글 속에는 앞 · 뒤가 서로 밀접한 관련성을 가지고 있으며 또한 유사한 개념의 어절이 빈번히 출현하는 특성을 이용하여 문단 단위 형태소 분석법을 제안한다(김영택, 1994. Ishizaki, 1995).

제안한 문단 단위 형태소 분석법은 주어진 문단에 존재하는 유사 어절들로부터 대표 단어와 단어 유형을 추출하고, 추출된 대표 단어 정보를 이용하여 형태소 분석을 한다. 그러므로 형태소 분석을 위한 사전 탐색 대상이 되는 단어수를 크게 줄일 수 있고, 단어 사전 탐색 횟수도 감소된다.

2. 일반적인 한국어 형태소 분석

2.1 한국어의 특징

한국어는 형태상으로 교착어, 계통상으로는 알타이 어족에 속한다. 교착어란 의미를 나타내는 어휘 형태소에 조사와 어미 같은 어법적 관계를 나타내는 문법 형태소가 붙음으로써 문법 기능을 한다고 하여 첨가어라고도 하는데, 우리말을 비롯해서 몽골어, 일본어, 터기어, 고대 만주어 등이 있다.

한국어의 특징을 요약하면 다음과 같다(김영택, 1994. 이용석, 1996).

(1) 우랄알타이 어족이며 교착어(첨가어)로 말의 순서나 어형 변화보다 조사, 조동사의 부속어로서 문법적인 관계를 나타낸다.
(2) 용언의 불규칙 현상이나 음운 현상이 발달하였다.
(3) 명사에 의한 띄어쓰기가 자유롭다.
(4) 어근에 파생 접사나 어미가 붙어서 단어를 구성한다.
(5) 모든 문법적 형태소는 반드시 어근 또는 어간 뒤에 쓰인다.

굴절어인 영어의 경우는 형태소 분석 과정에서 큰 오류를 범할 우려가 적으나, 교착어인 한국어는 단어 구분의 애매성으로 인해 여러 가지로 분석할 가능성 있다. 예로서 그림 1에서 "나는" 이라는 단어에 대한 형태소 분석 결과는 3가지의 서로 다른 분석 후보들로 결과가 나타나는 것을 볼 수 있다.

```
나는 (NOUN '나') + (JOSA '는')
      (VERB '나') + (EOMI '는')
      (VERB '날') + (EOMI '는')
```

그림 1. 한국어 형태소 분석 결과

실제로 "나는"이라는 단어에 대한 분석 후보수는 7가지이며, "소나무라고?" 라는 어절은 무려 51개나 되는 분석 후보수를 나타낸다. 이러한 결과는 자연 언어 처리 시스템의 상위 과정인 품사 태깅 및 구문 해석 단계에 아무런 도움을 주지 못한다(김재훈, 1996).

결국, 한국어에서는 올바른 자연 언어 처리 시스템을 위해서는 반드시 올바른 형태소 분석이 선행되어야만 차후 단계의 올바른 수행이 가능한 것을 알 수 있다.

2.2 한국어 문장의 구성

하나의 문장은 그림 2와 같이 형태소, 단어, 어절, 문장의 순으로 구성되며, 문법적으로 가장 큰 단위는 문장이다. 그런데 일반적인 형태소 분석에 관한 방법론으로는 언어 독립적인 연구인 이러한 문문법 위주로 시도하지만, 실제로 우리가 사용하는 말과 글은 무의미한 글자의 조합이 아니라, 그 구성 요소인 전후의 문장들이 서로 밀접한 연관성을 가지고 있다. 즉 말과 글 속에는 앞 · 뒤 화두 및 문장들의 밀접한 관련이 있으며, 이러한 관계를 고려하지 않으면 올바른 자연 언어 처리를 할 수가 없다.

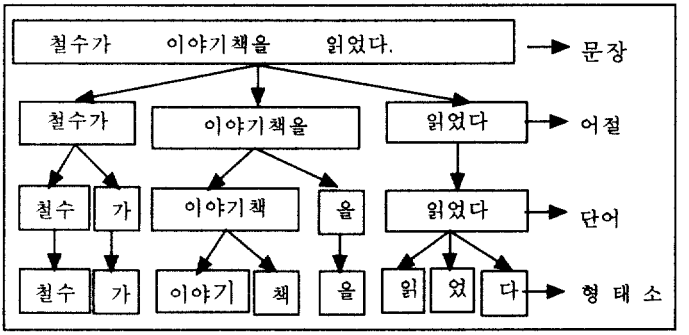

그림 2. 문장의 구성

따라서, 하나의 문장의 실질적인 기능이나 의미는 한 문장이 실현되는 구체적인 맥락을 고려해야 비로소 정확하게 파악되는데 이 맥락의 단위를 이야기라 한다. 이 이야기는 담화 문법의 화에 해당하고, 텍스터 문법의 텍스터 개념과 유사하며, 텍스터는 문장보다 한 단계 높은 개념이다(Ishizaki, 1995. 박갑수, 1992).

그림 3의 텍스트 문법에서의 이야기 틀을 통상 문단이라고 하며 문단과 문단의 구분은 들여 쓰기로 한다. 한 문단 또는 인접한 문단에 포함된 문장들에서는 동일 개념의 대표 단어가 빈번히 출현하고, 이 단어들은 일반적으로 같은 품사 및 의미를 가지고 있다.

2.3 기존의 한국어 형태소 분석법의 문제점

한국어 형태소 분석 방법에는 최장 일치법, Tabular 파싱법, 양방향 최장 일치법, 음절 단위 분석법, 접속 정보표를 이용한 방법 등 여러 가지 방법론들이 연구되어 왔다. 그러나, 지금까지 연구된 형태소 분석 방법론들은

다음과 같은 몇 가지 문제점들이 있다.

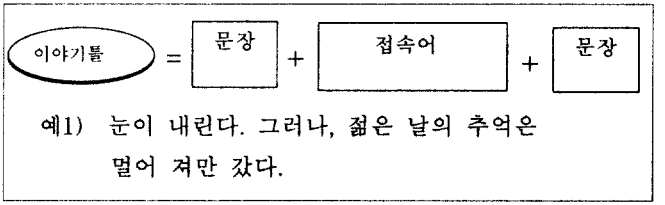

그림 3. 텍스터 문법에서의 이야기 틀

첫째, 관련된 상황이나 장면에서 문을 독립하는 문문법 연구에만 집중되어 있다. 이것은 이상화된 독립상황에서 문의 총어구조나 의미구조를 연구하는 것이다. 우리가 말과 글을 사용할 때는 이와 같은 독립된 상황은 비현실적이다.

둘째, 사전 참조 대상 단어수를 줄이는 연구가 미비하다. 양방향 최장일치법 등에 의한 형태소 분석 방법으로 어휘 형태소 사전을 참조하는 횟수를 줄이려는 노력도 있었다. 그러나, 이러한 방법들도 입력 단위가 어절 또는 문장 단위이기 때문에 효과적이지 못하다. 형태소 분석기의 성능 평가 요소는 단어의 분석률과 처리속도이다. 형태소 분석기는 주어진 입력 자료에 대해서 분석 알고리즘을 통해 형태소를 분리하고 그 원형을 복원하는 과정을 반복한다. 이 과정에서 올바른 분석을 하기 위하여 끊임없이 사전 정보를 필요로 하기 때문에 사전의 참조 횟수와 탐색 시간이 전체 시스템의 성능을 좌우한다.

따라서, 형태소 분석 알고리즘의 효율을 높이기 위해서는, 비교 연산의 수를 줄이는 것뿐만 아니라 사전 탐색 횟수를 줄이는 것이 매우 중요하다. 사전의 탐색 횟수를 줄이는 문제에서도 사전 구축 방법과 탐색 알고리즘에 의한 탐색 시간 감소보다는, 사전 탐색의 대상이 되는 단어수 자체를 줄이는 것이 보다 효과적이다. 결론적으로, 문보다 큰 단위에서 전후의 문맥을 고려하면서, 사전 참조의 대상이 되는 단어수 자체를 줄이는 방법이 한국어 형태소 분석기의 성능을 향상시키는 방법이라고 할 수 있다.

3. 한국어 문단 분석

기존의 한국어 형태소 분석 기법들의 문제점을 해결하기 위해 문보다 큰 단위인 문단내의 유사 단어 출현 빈도와, 대표 단어 및 단어 유형의 품사 및 의미를 고찰하기 위하여 2개의 문단을 (한판암. 1987)에서 발췌하였고,

그림 4에 보였다. 또 대표 단어와 단어 유형에 관한 정의는 정의 1, 정의 2와 같다.

그림 4. 한국어 문단 분석을 위한 예문

[문단 1]
정보의 전달은 언어나 문자 이외에 여러 가지 방법으로 가능한 것이다, 하나의 예로서 기호들도 정보의 표시를 하고 있기 때문에 전달이 가능한 것이다.

[문단 2]
그러나, 이와 같이 정보 전달에 사용되는 기호들은 그 의미를 아는 사람에게는 의미가 전달되지만 전혀 그 의미를 모르는 사람에게는 단순한 기호 이외의 다른 아무런 가치가 없는 것임은 틀림없는 사실이다.

[정의 1] 대표 단어
대표 단어란 2개 이상의 유사 어절들에서 공통으로 포함하고 있는 문자열을 말한다.

[정의 2] 단어 유형
단어 유형이란 대표 단어로 추출된 후 남은 각각의 유사 어절들의 문자열을 말한다.

그림 4의 예문에 대하여 1개 및 2개 문단별 단어출현 유형 분석 결과는 각각 표 1, 표 2와 같다. 표 1에서 구분은 그림 4에서 설정한 예문에서의 문단 구분을 의미한다. 어절 구분은 추출된 대표 단어의 순차적인 순서이다. 대표 단어 추출에 사용한 분석 방법은 좌-우 분석법을 사용하였다. 단어 유형 및 출현 빈도에서 어절구분 1의 "의(3)"은 대표 단어로 "정보-"를 공유하고, 유형으로 "-의"를 사용하는 단어가 3개 있다는 뜻이다. 그래서 문단 1에서는 대표 단어가 5개, 단어 유형이 13개이며 문단 2에서는 대표 단어가 4개, 단어 유형이 9개이다.

표 3은 1 개 및 2 문단 단위에서 대표 단어의 출현 빈도와 대표 단어당 포함된 평균 유사 단어수를 보인다.

표 1. 1개 문단별 단어출현 유형 분석

구 분	어절 구분	대표 단어	단어유형 및 출현 빈도	계
문단 1	1	정보-	의(3), 는(1)	4
	2	전달-	은(1), 을(1), 이(1)	3
	3	가능한	-(2)	2
	4	것이다	-(2)	2
	5	기호-	들도(1), -(1)	2
	소 계			13
문단 2	2	전달-	에(1), 되지만(1)	2
	5	기호-	들은(1), -(1)	2
	6	의미-	를(2), 가(1)	3
	8	사람-	에게는(2)	2
	소 계			9

표 2. 2개 문단별 단어출현 유형 분석

어절구분	대표 단어	단어 유형 및 출현 빈도	계
1	정보	-의(3), -는(1), -(1)	5
2	전달	-은(1), -을(1), -이(1), -에(1), -되지만(1)	5
3	가능한	-(2)	2
4	것이다	-(2)	2
5	기호	-들도(1), -(2), -들은(1)	4
6	의미	-들(2), -가(1)	3
7	이외	-에(1), -의(1)	2
8	사람	-에게는(2)	2
누계			25

표 3. 문단별 단어의 출현 빈도 분석

구 분		전체 어절수	대표 단어수	단어 유형	출현 빈도	평균단어 수
1개 문단	문단1	30	5	13	43.3%	2.6단어
	문단2	28	4	9	32.1%	2.3단어
	소계	58	9	22	37.7%	2.5단어
2개 문단	문단1	58	8	25	43.1%	3.1단어

3.1 문단 단위 분석 결과 검토 및 신뢰도

대표 단어 출현빈도 조사에 대한 보다 확실한 검토를 위해 (한국교육개발원, 1996), (한국교육개발원, 1995), (김재면, 1994)를 대상으로 문단 분석을 하였다. 문단 분석에 사용된 자료는 2,166개 문단의 91,668 어절이었다.

표 4는 조사 대상을 1개 문단씩을 단위로 하여, 표 5는 이웃하는 2개 문단씩을 단위로 하여, 문단 내 존재하는 대표 단어와 유사 단어 출현에 대한 분석 결과이다. 표 6에서는 표 4와 표 5를 종합하여 비교하였고, 유사 단어 출현 빈도는 1개 문단이 34.35%, 2개 문단이 42.66%로, 이웃하는 2개 문단씩을 단위로 하여 분석한 것이 8.31% 높게 나타났다. 문단당 평균 어절수는 1개 문단 단위에서 43.53개, 2개 문단 단위에서 84.72개로 약 2배정도 많으나, 대표 단어 당 유사 단어 개체수는 2.62와 2.80개로 약 0.18개 차이밖에 없다.

표 4. 1개 문단별 분석결과

구분		총문단수	총어절수	문단당 평균 어절수	대표 단어수	단 어 유 형	출현 빈도	대표단어당 개체수
계		2,166	91,668	43.5	12,030	31,490	34.35	2.62개
중학교	소계	644	20,913	32.5	1,861	4,345	20.77	2.33개
	1 장	182	5,620	30.9	473	1,094	19.4	2.31개
	2 장	226	6,607	29.2	588	1,359	20.5	2.31개
	3 장	236	8,686	36.8	800	1,892	21.7	2.37개
노동부	소계	549	26,539	48.3	3,833	10,493	39.53	2.74개
	1 장	225	10,175	45.2	1,481	3,881	38.1	2.62개
	2 장	60	2,530	42.2	353	927	36.6	2.62개
	3 장	124	8,344	67.3	1,263	3,776	45.2	2.98개
	4 장	140	5,490	39.2	736	1,909	34.7	2.59개
고등학교	소계	973	44,216	48.4	6,336	16,652	37.66	2.63개
	1 장	39	1,974	50.6	292	815	41.2	2.79개
	2 장	102	4,975	48.8	652	1,678	33.7	2.57개
	3 장	118	5,580	47.3	760	1,988	35.6	2.61개
	4 장	149	6,624	44.5	932	2,518	38.0	2.70개
	5 장	177	8,054	68.8	1,287	3,555	44.1	2.76개
	6 장	135	5,862	43.4	873	2,217	37.8	2.53개
	7 장	122	5,421	44.4	836	2,155	39.7	2.57개
	8 장	131	5,726	43.7	704	1,736	30.3	2.46개

그러므로 형태소 분석기에서 입력 단위를 2개 문단으로 하는 것은 전체 어절수가 2배로 증가하므로 메모리 사용의 효율성이 떨어지나, 대표 단어당 유사 단어 수에서는 별다른 차이가 없기 때문에 전체적인 시스템의 효율성이 떨어진다고 볼 수 있다. 그러므로 형태소 분석 단위는 1개 문단을

기준으로 하는 것이 효과적이라고 할 수 있다.

표 5. 2개 문단별 분석 결과

구분		총문단수	총어절수	문단당 평균 어절수	대표 단어수	단 어 유 형	출현 빈도	대표 단어당 개체수
계		2,151	182,240	84.7	27,226	77,744	42.66	2.86개
중학교	소계	641	41,624	64.9	4,809	11,960	28.7	2.49개
	1 장	181	11,179	61.8	1,171	2,913	26.0	2.48개
	2 장	225	13,135	58.4	1,546	3,794	28.8	2.45개
	3 장	235	17,310	73.7	2,092	5,253	30.3	2.51개
노동부	소계	545	52,757	96.8	8,434	25,259	47.9	2.99개
	1 장	224	20,259	90.4	3,233	9,462	46.7	2.92개
	2 장	59	5,009	84.9	808	2,332	46.5	2.89개
	3 장	123	16,567	132.3	2,735	8,810	53.1	3.22개
	4 장	139	10,922	78.6	1,658	4,655	42.6	2.80개
고등학교	소계	965	87,859	91.05	13,983	40,525	46.13	2.90개
	1 장	38	3,898	102.6	623	1,914	49.1	3.07개
	2 장	101	9,825	97.3	1,478	4,126	41.9	2.79개
	3 장	117	11,102	94.9	1,676	4,860	43.7	2.90개
	4 장	148	13,177	89.0	2,117	6,265	47.5	2.96개
	5 장	176	16,028	91.1	2,670	8,324	51.9	3.12개
	6 장	134	11,657	87.0	1,923	5,315	45.5	2.76개
	7 장	121	10,781	89.1	1,831	5,291	49.0	2.89개
	8 장	130	11,391	87.6	1,665	4,430	38.8	2.66개

표 6. 1개 및 2개 문단 단위별 분석 결과

구분	교재	총 문단수	총 어절수	문단당 평균 어절수	대표 단어수	단 어 유 형	출현빈도	대표단어당 개체수
1	소계	2,166	91,668	43.53	12,030	31,490	34.35	2.62
개	중학교	644	20,913	32.5	1,861	4,345	20.77	2.33
문	노동부	549	26,539	48.3	3,833	10,493	39.53	2.74
단	고등학교	973	44,216	48.4	6,336	16,652	37.66	2.63
2	소계	2,151	182,240	84.72	27,226	77,744	42.66	2.80
개	중학교	641	41,624	64.9	4,809	11,960	28.70	2.49
문	노동부	545	52,757	96.8	8,434	25,259	47.90	2.99
단	고등학교	965	87,859	91.05	13,983	40,525	46.13	2.90

또한 형태소 분석에서 대표 단어 정보를 이용하면, 대표 단어에 대한 사전 검색을 한번만 하고, 나머지 단어 유형에 대한 사전 검색만 하면 되므

로 사전 검색 횟수 자체를 줄일 수 있는 장점이 있다.

표 7은 표 6에 대한 신뢰도 검증 결과로, 표준 편차가 1개 문단 단위가 2개 문단 단위보다 큰 이유는, 2개 문단 단위 분석시 입력 문단의 총 어절 수가 2 배정도 증가하여, 유사 단어 출현 빈도의 최저치인 0에 가까운 문단수가 1개 문단에 비해 상대적으로 감소하였기 때문이다(최종석, 1985).

표 7. 한국어 문단 단위 분석의 통계치

항목	실험결과 통계치	
	1개 문단 단위	2개 문단 단위
실험 모집단수	2,166개	2,151개
모집단내 어절수	91,668	182,240
평균	34.35 %	42.66 %
표준 편차	16.65	14.05
신뢰도	95% ± 0.701	95% ± 0.594

4. 문단 단위 형태소 분석기 모델

4.1 문단 단위 형태소 분석기

일반적으로 형태소 분석은 분석의 대상은 단어로 하고, 문서로부터 형태소 분석의 대상이 되는 단어를 추출하고, 문장 부호를 분리하며, 숫자나 특수 문자열을 처리하는 전처리 단계를 수행하여 단어 또는 어절 단위의 입력 문자열을 추출하고, 이어서 분석 후보 생성 단계에서는 형태소 분리 과정과 원형 복원 과정을 거쳐 가능한 모든 분석 후보들을 생성한다. 그리고, 생성된 모든 분석 후보들 중에서 어휘 사전의 검색과 단어 형성 규칙(word formation rule), 결합 제약 조건 등에 의해 최종적으로 옳은 후보를 선택하는 분석 후보 선택 과정으로 이루어진다.

형태소 분석 대상이 되는 후보 생성의 경우 단어의 구성 자소에 해당하는 만큼의 조합이 발생한다. 이러한 조합의 수만큼 사전을 검색하고 또한 문단 내에 동일한 단어가 빈번히 출현하더라도 매번 사전의 자료를 검색하여야 하므로 시스템 효율성이 떨어진다.

현재, 사용되고 있는 형태소 분석 모델들은 입력 단위를 어절 또는 문장으로 하면서 입력된 문자열 자체에서만 조합의 수를 줄이려는 노력들을 하고 있다. 그러나, 한국어 문단 분석의 결과가 보인 것과 같이, 우리가 일상적으로 사용하는 글 속에는 일정한 범위 즉, 문단 내에서는 동일 또는 유

사 단어들이 빈번히 출현하는 것을 알 수 있다.

따라서 본 논문에서는 이러한 특성을 이용하여 그림 5와 같은 문단 단위 형태소 분석 모델을 제시하고자 한다. 이 모델은 기존의 형태소 분석기처럼 어절 또는 문장 단위의 입력으로부터 직접 형태소 분석을 수행하지 않고, 문단 단위의 입력 문자열에 대하여 문단 처리를 선행한다. 이 문단 처리기는 문단 단위의 입력 자료들에서 대표 단어와 단어 유형의 정보를 추출해 내는 역할을 한다.

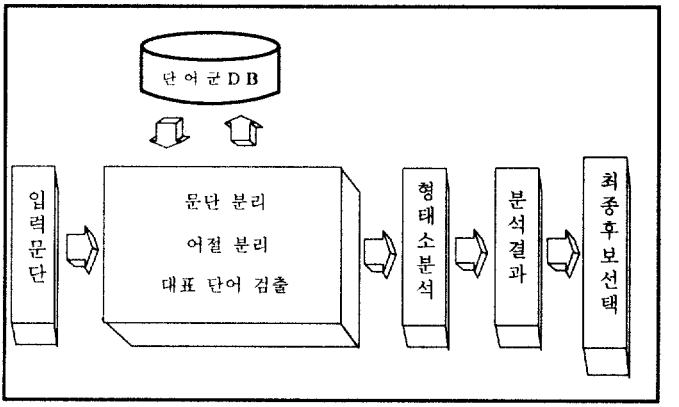

그림 5. 문단 단위 형태소 분석 모델

4.2 문단 처리기

문단처리기의 수행 과정을 그림 6에서 보면 입력 문단으로부터 어절을 분리하고, 동일 또는 유사 어절들로부터 대표 단어와 단어 유형 및 원문내에서 위치 등의 자료를 추출해 낸다.

입력 문단에서 "정보의", "정보의", "정보는", "정보의"라는 어절들이 문단 처리기에 의해 유사 단어 유형들을 가진 후보들로 추출된다. 추출된 각 어절에서 동일한 부분인 "정보-" 라는 단어를 추출하여 대표 단어로 정하고, 대표 단어로 추출되지 않은 부분인 "-의" 와 "-는" 등은 "정보-"라는 대표 단어에 속한 단어 유형으로 간주한다.

한국어 특징중의 하나인 모든 문법적 형태소는 반드시 어근, 어간 뒤에 쓰인다는 것을 이용하여 "정보"라는 단어는 어근 즉, 어휘 형태소 간주되고 "-는", "-의" 등은 문법 형태소로 간주된다. 이때 대표 단어와 해당 대표

단어에 속한 단어 유형들이 각 원문 내에서의 위치에 대한 정보도 포함시켜 문단 분석 결과를 만들어 낸다. 문단 처리기의 통하여 추출된 문단 분석 결과는 이후 단계인 형태소 분석 과정에서 분석을 수행하는데 필요한 정보로 제공되어 진다.

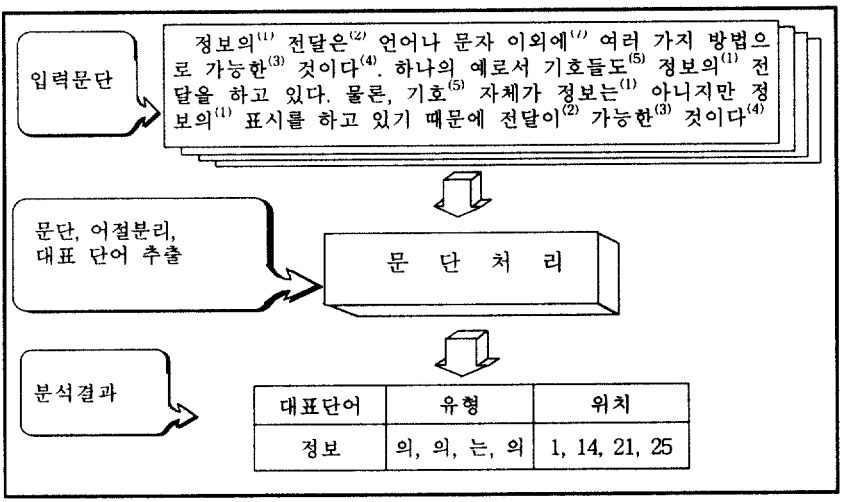

그림 6. 문단 처리기의 처리과정

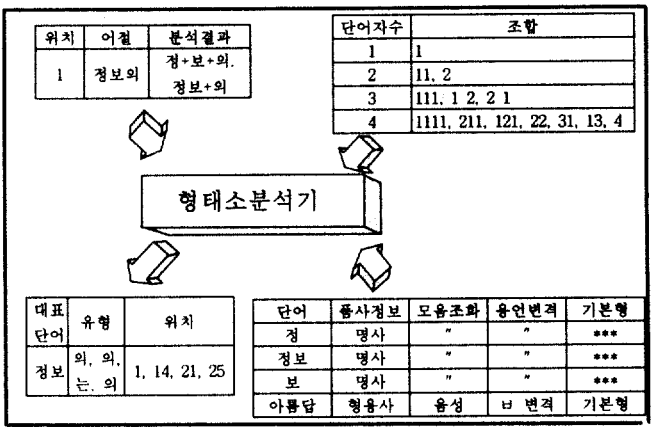

그림 7. 문단 단위 형태소 분석 과정

4.3 형태소 분석 과정

문단 단위 한국어 형태소 분석 과정을 그림 7에서 살펴보면, 입력 문단으로부터 그림 6의 문단처리기를 통해 추출되어져서 분석기에 제공되는 문단 분석 결과를 토대로, 각 대표 단어에 대한 정보는 사전을 탐색하여 얻는다. 예로서 그림 6에서 "정보의", "정보의", "정보는", "정보의"라는 어절들에서 추출된 대표 단어 "정보"는 단어 사전을 탐색을 위한 대상 단어가 된다. 만약 대표 단어 정보가 없다고 가정한다면 앞의 각 어절들이 출현하는 원문 위치인 1, 14, 21, 25에서 각각 한번씩, 전체 4회의 사전 탐색하여야만 한다.

그러나 제안한 그림 7의 문단 단위 형태소 분석기에서는, 대표 단어 "정보-"를 어휘 형태소 사전에서 탐색한 후, 각각의 원문 위치로 탐색 결과를 되돌려 준다. 또 이때 단어 유형인 "-의", "-의", "-는", "-의"는 문법 형태소 사전을 참조하여 탐색 결과를 얻는다. 결국, 제안한 문단 단위 형태소 분석 방법은 기존의 형태소 분석보다 (식 1)만큼 사전 참조 횟수를 줄일 수 있다.

$$DIC\text{-}COUNT = \sum_{i=1}^{n} W_{ci} - \left(\sum_{i=1}^{k} G_{ci} + \sum_{i=1}^{m} S_{ci} \right) \quad \text{--- (식 1)}$$

여기서, W : 한 문단의 각각의 단어,
G : 한 문단에서 출현한 대표 단어,
S : 한 문단에서 출현한 단어유형, I : 변량(count),
C : 조합 가능한 경우의 수, n : 단어의 개수,
k : 대표 단어의 개수, m : 단어유형의 개수,
DIC-COUNT : 사전 참조 횟수이다.

4.4 실험 및 고찰

본 논문에서 제안한 문단 단위 형태소 분석기의 성능을 평가하기 위하여 기존 어절 단위 형태소 분석기와 본 논문에서 제안한 문단 단위 형태소 분석기를 각각 구현하여 형태소 분석 실험을 하였다.

표 8은 그림 4의 예문에 대한을 실험한 결과이고, 여기서 어절 단위와, 제안한 문단 단위로 구분하여 실험하고 분석하였다. 실험결과 문단 단위 형태소 분석에서 유사 단어 출현 빈도가 43.3%, 32.1% 일 때 어절 단위에 의한 형태소 분석에서 보다 각각 44.16%, 59.76%의 사전 참조 횟수가 감소

되었다. 표 9에서는 동일한 유사 단어 출현 빈도에서의 사전 탐색 효율성의 차이를 보였다.

표 8. "그림 4"의 예문에 대한 형태소 분석 결과

문단	단위	총어절수	유사단어 출현빈도	사전참조 횟 수	평균 개체수	효율성
1	어절	30	-	231	-	44.16%
	문단		43.3%	129	2.6	
2	어절	28	-	338	-	59.76%
	문단		32.1%	136	2.3	

표 10은 (한국교육개발원, 1995)를 대상으로 행한 형태소 분석 결과이다. 어절 단위 형태소 분석과 제안한 문단 단위 형태소 분석방법에서의 총 사전 참조 횟수를 조사하였고, 이들을 비교하여 제안한 방법의 효율성을 확인할 수 있었다.

표 9. 동일한 유사 단어 출현 빈도에서 사전 탐색

어절 유형	대표단어	단어 유형	평균개체수	사전탐색회수
정보는, 정보가 정보에, 정보를	정보-	-는, -가 -에, -를	4	7
정보가, 정보를	정보-	-가, -를	2	10
우리는, 우리가	우리-	-는, -가		

실험 결과, 어절수 1,974, 대표 단어수 292, 단어 유형이 815개인 자료에서, 기존의 어절 단위 형태소 분석기는 21,378회의 사전 참조를 하였고, 제안한 문단 단위 형태소 분석기는 9,643회의 사전 참조를 하여 유사 단어 출현 빈도 41.2%에서 45.1%의 사전 참조 감소율을 보였다. 그런데 유사단어 출현 빈도와 사전 참조의 효율성이 정비례하지 않는 것은 표 9에서 볼 수 있듯이 문단 내에서 유사 단어 출현 빈도가 동일하다고 하더라도 대표 단어당 평균 개체수에서 차이가 원인이다. 또 다른 이유는 어절 자체의 길이와 어절에서 대표 단어와 단어 유형으로 분리된 단어의 구성 음절의 길이가 각각 다르기 때문이다

5. 결 론

표 10. 형태소 분석 결과

문단번호	총 어절수	대표단어수	단어 유형	출현 빈도	어절단위 처리	문단단위 처리	효율성
1	26	2	4	15.3	296	147	49.6
2	76	13	32	42.1	909	394	43.3
3	28	3	7	25.0	422	184	43.6
4	36	4	10	27.7	346	173	50.0
5	59	8	25	42.3	680	326	47.9
6	26	2	5	19.2	225	116	51.5
7	71	9	29	40.8	750	334	44.5
8	135	21	71	52.5	1403	635	45.2
9	39	5	10	25.6	395	186	47.0
10	22	3	10	45.4	314	119	37.8
11	10	1	2	20.0	136	67	49.2
12	164	25	88	53.6	1914	805	42.0
13	50	7	16	32.0	571	266	46.5
14	27	4	12	44.4	331	165	49.8
15	29	2	5	17.2	317	172	54.2
16	33	3	8	24.2	380	166	43.6
17	26	4	8	30.7	319	142	44.5
18	52	7	19	36.5	493	232	47.0
19	23	3	7	30.4	400	169	42.2
20	44	5	11	25.0	353	209	59.2
21	35	8	19	54.2	371	150	40.4
22	85	11	34	40.0	947	437	46.1
23	101	14	43	42.5	852	399	46.8
24	40	7	15	37.5	456	203	44.5
25	47	2	7	14.8	368	214	58.1
26	43	10	24	55.8	523	190	36.3
27	104	12	41	39.4	899	446	49.6
28	35	8	22	62.8	401	142	35.4
29	43	11	29	67.4	743	235	31.6
30	34	4	8	23.5	310	170	54.8
31	61	10	33	54.0	574	234	40.7
32	51	9	28	54.9	482	220	45.6
33	26	5	10	38.4	289	146	50.5
34	56	13	29	51.7	554	255	46.0
35	64	10	24	37.5	680	326	47.9
36	62	11	29	46.7	606	306	50.4
37	44	9	21	47.7	608	233	38.3
38	43	6	18	41.8	563	221	39.2
39	24	1	2	8.3	198	109	55.0
계	1974	292	815	41.2	21378	9643	45.1

한국어 형태소 분석기의 성능은 분석 알고리즘과 사전 탐색 시간에 좌우되는데 이중 사전 탐색 시간이 분석기 성능 향상에 훨씬 더 큰 영향을 미친다. 사전 탐색 시간을 줄이는 방법으로는 사전의 구조나 탐색 알고리즘보다는 대상 단어수를 줄이는 것이 효과적이다. 그러나, 현재까지 연구된 분석 방법은 입력 단위를 어절 또는 문장 단위 형태소 분석을 함으로써 사전 탐색 대상 단어수를 감소시키는 연구가 미비한 실정이다.

우리가 통상 사용하는 말과 글은 무의미한 조합이 아니고, 그 전후가 서로 밀접한 관계를 가지고 있는데, 이러한 관계의 묶음을 문단이라고 한다. 한 문단 내에서는 유사한 단어들이 빈번히 출현하고, 또한 앞·뒤 문장들도 서로 밀접한 관계를 가지고 있다.

따라서, 본 논문에서는 이러한 특성을 이용하여 문단 단위 한국어 형태소 분석기를 제안하고 이를 검증하기 위해서 (한국교육개발원, 1996), (한국교육개발원, 1995), (김재면, 1994)를 선택하여, 약 9만여 어절과 2000여 개의 문단을 조사한 결과 1개 문단 단위에서 34.35%의 유사 단어 출현 빈도를 조사하였고, 이 조사된 자료를 통계 처리하여 95% 유의 수준에서 ± 0.701의 신뢰도를 구하였다.

위와 같은 검증 자료를 토대로 기존의 어절 단위 형태소 분석기와 본 논문에서 제시한 문단 단위 형태소 분석기를 각각 구현하여, 본 논문에서 설정한 예문을 이용하여 실험한 결과, 유사 단어 출현 빈도 43.3%, 32.1%에서 각각 44.16%, 59.76%의 사전 탐색 횟수가 감소됨을 확인하였다. 또 (한국교육개발원, 1995)의 제1장, 어절수 1,974개, 대표 단어수 292개, 단어 출현 유형이 815개, 유사 단어 출현 빈도 34.35%인 자료를 대상으로 형태소 분석 실험을 한 결과, 기존의 어절 단위 형태소 분석기에 비해 45.1%의 사전 참조 횟수가 감소되었다.

참 고 문 헌

강승식, 장병탁. 1996a. "음절 특성을 이용한 범용 한국어 형태소 분석기 및 맞춤법 검사기," 한국정보과학회 논문집, 제23권, 제5호, 531-532.

강승식, 이하규. 1996b. "한국어 형태소 분석기 HAM의 형태소 분석 및 철자기능검사," 제8회 한글 및 한국어 정보처리 학술대회, 249-250.

김재면, 박병수. 1994. 정보처리론. 서울: 한국산업인력관리공단.

김영택. 1994. 자연 언어 처리. 서울: 교학사.

김재훈, 김길창. 1996. "언어지식을 이용한 형태소 해석의 모호성 감소," 제8회 한글 및 한국어 정보처리 학술대회, 231-234.
박갑수, 조규빈. 1992. 고교문법. 서울: 지학사.
이용석, 이기오, 이근용. 1996. "효율적인 한국어 분석을 위한 확장된 최장일치법," 제8회 한글 및 한국어 정보처리 학술대회, 255-261.
최종석 외 3인. 1985. 통계학. 서울: 정익사.
한국교육개발원. 1996. 중학교 사회2, 서울: 대한교과서주식회사.
한국교육개발원. 1995. 고등학교 정치경제. 서울: 대한교과서주식회사.
한판암. 1987. 전자계산학개론. 서울: 상조사.
Ishizaki S. 1995. The Natural Language Processing. Tokyo: Sohwadang.

대전광역시 동구 삼성2동 375
대전산업대학교 전자계산학과
300-717
Email: ehrhee@hyunam.tnut.ac.kr
Fax: +82-042-636-3554

접수일자: 98. 4. 30
게재결정: 99. 5. 20