"의미적 한 단어" 유형 분석 및 형태소 분석 기법[1)]

허윤영, 권혁철
부산대학교 전자계산학과

Korean Morphological Analysis Considering a Term with Multiple Parts of Speech

Yun-Young Hur, Hyuk-Chul Kwon
Department of Computer Science, Pusan National University

요 약

한국어 문서중 신문이나 시사지, 법률관련문서, 경제학관련문서, 국문학관련문서와 같은 전문분야 문서에는 한글, 한자, 영어, 문장부호와 같은 기호들의 결합으로 이루어지면서 하나의 뜻으로 나타내는 "의미적 한 단어"가 많이 존재한다. 이러한 단어들은 이를 고려하지 못한 형태소 분석기의 분석률을 감소시키고, 오분석율을 증가시킨다. 본 논문은 "의미적 한 단어"의 유형과 분석과정에 따른 유형을 분류하였으며 그에 적합한 형태소 분석기법을 제시하였다. 유형 분류과 제시된 형태소 분석기법으로 구현된 형태소 분석기는 기존의 형태소 분석기보다 분석률이 증가되었으며 오분석률은 감소되었다.

I. 서론

한국어 문서에서는 동음이의어를 나타내기 위해서, 중요한 단어임을 나타내기 위해서, 또는 한글로만 표현이 불가능한 신조어를 표현하기 위해서 등과 같은 여러가지 이유로 많은 부분이 한글, 한자, 영어, 숫자 등의 결합으로 이루어진 단어로 구성된다. 또한 문장부호 등과 같이 혼동될 수 있는 기호(., -)가 단어를 구성하는 요소가 되기도 한다. 이러한 단어들은 복합명사적 성격을 띄기 때문에 일반 한국어 복합명사처럼 여러 어절에 걸쳐 띄어 쓰여지기도 한다. [6]

기존의 형태소 분석에서는 기호와 함께 쓰여진 어절이나 여러 어절에 걸쳐 나타나는 한 단어, 한자, 영어, 숫자와 혼합된 한 단어 등의 처리에 대한 거의 고려가 없다. [3, 4, 5] 때문에 이러한 어절은 다른 어절로 분리되고, 전혀 연관이 없는 단어들로 잘못 분석되기도 한다. 너구나 이러한 단어가 전체 문서에서 많은 부분을 차지하는 신문기사, 전문 분야의 문서 등에서는 형태소 분석 오류가 많이 발생한다. 이러한 형태소 분석 오류는 형태소 분석의 정확도 뿐 아니라 형태소 분석을 기반으로 하는 검색시스템, 철자 검사 시스템, 문서 분류 시스템 등의 여러 응용시스템의 효율에 큰 영향을 미친다.

본 논문에서는 형태소 분석에 영향을 미치는 이러한 어절 및 단어들에 대한 유형을 분류하고, 이를 위한 형태소 분석 기법을 제안하였다.

II. "의미적 한 단어"

분석하면 기호와 함께 쓰여진 단어, 한 어절 또는 여러 어

1) 이 논문은 1993년도 한국 학술 진흥 재단의 공모 과제 연구비에 의하여 연구되었음.

절에 걸쳐서 나타나는 한글, 한자, 영어, 숫자 등이 혼합된 단어들이 빈번히 나타난다. 조사된 빈도수는 다음과 같다.

신문	전체어절수	혼합된 어절	비율(%)
1월	822877	175056	21.27
3월	927986	192669	20.76
*4월	612694	126493	20.64
7월	891119	179411	19.83
*8월	573603	113751	20.57
합 계	3828279	787380	20.57

< 표 1 > 혼합된 어절 수(* 전체 분량이 아님)

본 논문에서는 어절에 기반한 형태소 분석에 영향을 미치는 단어들을 다음과 같이 정의하였다.

[정의]
"의미적 한 단어" : 한글, 한자, 영어, 기호의 결합된 형태로 한 어절 또는 여러 어절에 걸쳐서 구성되면서 하나의 뜻으로 사용되는 단어

신문 기사 말뭉치로부터 추출된 '의미적 한 단어'를 의미에 따라 분류해 보면 다음과 같이 분류된다. 이 유형분류는 지식베이스 구축시 이용된다. [1, 2]

유형	유형내용	유형 예	비율
1	행정구역	釜田 2동, 부산직할市	19.6
2	나라 명칭	中·러, 서방선진 7개국	1.46
3	지역, 장소	落東江상수원, 金海공항	6.35
4	성명	루이15세, 빌 클린턴	9:56
5	연속단위	5판4승제, 1할3푼4리	5.05
6	단체명칭	기독교방송社, 119구급대	6.13
7	행사명칭	제1회동아시아대회, APEC정상회담	3.65
8	직업, 직위	FIFA회장, 亞·太재단이사장	0.94
9	행정관서	中部경찰서, 이북5도청	3.12
10	사람	클린턴美대통령, PLO의장	29.5
11	기타의미	벙커C油, 헬機, 7·4공동성명	14.6

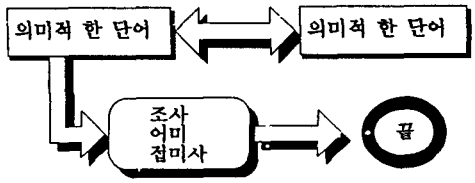
< 표 2 > '의미적 한 단어' 의미적 유형표

<표 2>에 나타난 '의미적 한 단어'에는 고유명사가 많다. 또한 특정 단어와 결합하는 경우가 많으며 일반 명사와 다시 복합 명사를 만들 수 있고, 일반 명사와 동일하게 이용되므로 조사, 접미사 등도 올 수 있다.

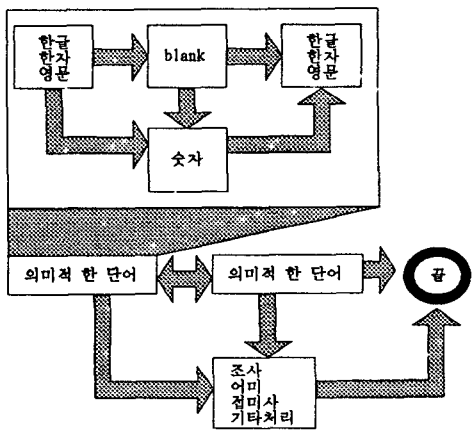
분석 과정에 따른 유형은 다음과 같은 세 가지 유형으로 분류된다.

1) 한 어절로 구성된 "의미적 한 단어" 유형
2) 여러 어절로 구성된 "의미적 한 단어" 유형
3) "의미적 한 단어"과 일반 어절이 결합된 유형

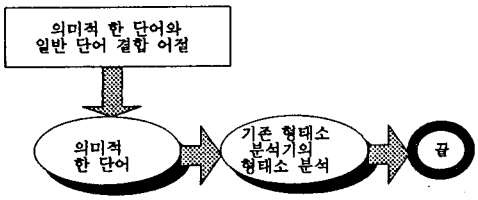
한 어절로 구성된 '의미적 한 단어'는 지식베이스를 이용하여 형태소 분석이 되며 분석 유형 A라고 하였다. 여러 어절로 구성된 '의미적 한 단어'은 어절 간 오토마타로서 형태소 분석이 되며 분석 유형 B라고 하였다. 마지막으로 '의미적 한 단어'와 일반 어절의 결합으로 이루어진 유형이 있다. 이러한 유형은 기존 형태소 분석기의 분석결과를 이용하여 분석된다. 이렇게 분석되는 유형을 분석 유형 C라고 하였다.



< 그림 1 > 분석 유형 A



< 그림 2 > 분석 유형 B

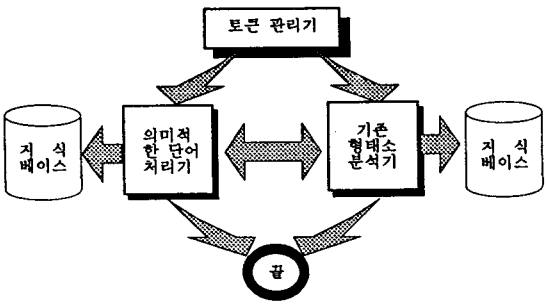


< 그림 3 > 분석 유형 C

III. "의미적 한 단어" 형태소 분석기

의미적 유형 분류와 형태소 분석 과정에 따른 유형을 이용하여 '의미적 한 단어'의 형태소 분석이 고려된 형태소 분석기를 구현한다.

의미적 유형 분류는 '의미적 한 단어' 형태소 분석을 위한 기본 단위가 된다. 분석 과정에 따른 유형 A, 유형 B, 유형 C에 의해서 '의미적 한 단어'를 분석한다. 전체 형태소 분석기는 '의미적 한 단어'의 형태소 분석을 고려한다. 전체 형태소 분석기는 토큰 관리기와 '의미적 한 단어' 형태소 분석 처리기, 기존 형태소 분석기로서 <그림 4>와 같이 구성된다. <그림 4>와 같이 '의미적 한 단어' 형태소 분석 처리는 기존 형태소 분석기와 서로 독립적으로 구성된다. 따라서 기존 형태소 분석기가 분석에 실패한 어절에 대해서만 '의미적 한 단어' 형태소 분석을 할 수 있고, 그와 반대로 먼저 '의미적 한 단어'의 형태소 분석을 한 후 분석에 실패한 어절에 대해서 기존의 형태소 분석을 할 수 있다. 본 논문의 형태소 분석기는 기존 형태소 분석기가 분석에 실패된 어절에 대해서만 '의미적 한 단어' 처리한다.



< 그림 4 > 전체 형태소 분석기 구성도

다음은 '의미적 한 단어'가 분석되는 과정이다.

< 분석 유형 A 예 >

의미적 한 단어	분석
D전자산업에서는	D전자산업(CW) + 에서는(조사)
美국방차관補는	美국방(CW) + 차관補(CW) + 는(조사)

(CW : 의미적 한 단어)

< 분석유형 B 예 >

의미적 한 단어	분석
제1회 東아시아게임을	{제(어절)+1(숫자)+/blank/+東아시안게임(CW)} (CW) + 을(조사)
로버트 갈루치차관補는	{로버트(어절)+/blank/+갈루치(어절)} (CW) + 차관補(CW) + 는(조사)

* 여기서 '/blank/'은 blank가 있거나 없거나 된다는 의미임.

< 분석 유형 C 예>

의미적 한 단어	분석
UR협상결과는	UR협상(처리과정) + 결과(기존분석) + 는
巫·太재단이사장은	巫·太재단(처리과정) + 이사장(기존분석) + 은 (조사)

IV. 실험 및 평가

실험은 신문말뭉치 중 월별 1000라인씩 표본 추출한 표본 말뭉치로 하였다. '의미적 한 단어' 처리가 고려되지 않은 경우 '의미적 한 단어'의 74%가 미분석되었고, 24.5%가 오분석되었다. 오분석은 대부분 띄어 쓰여진 '의미적 한 단어'에서 발생하였다. 올바르게 분석된 경우(수사처리)를 고려하더라도 오분석률이 매우 크다. [5]

본 논문에서 구현한 형태소 분석기의 분석 결과는 <표 3>과 같다. <표 3>에서 '의미적 한 단어'의 구성비는 서론에서 조사한 혼합어절의 전체 어절에 대한 구성비인 20%보다 낮은 12%였다. 이는 서론에서 조사한 혼합 어절에는 '의미적 한 단어'가 아닌 어절이 상당수 포함되어 있으며, 토큰분리기가 완벽한 '의미적 한 단어'를 추출하지 못하였기 때문이다.

본 형태소 분석기는 기존의 형태소 분석기가 거의 분석하지 못하는 '의미적 한 단어'를 97%의 분석율을 가지며 분석하

였다. 미분석률 3%의 '의미적 한 단어'는 '의미적 한 단어' 형태소 분석을 위해 구축한 지식베이스에 없는 단어로 구성된 '의미적 한 단어'였다.

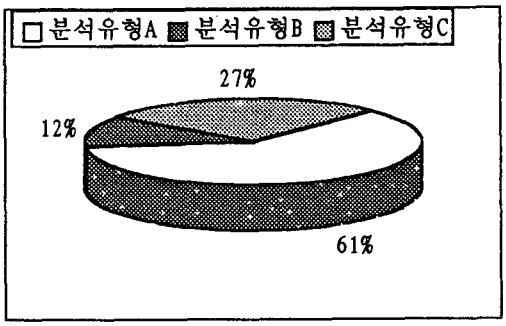
[참 고]

가) 월별 표본말뭉치
나) 전체 어절 수
다) 의미적 한 단어 수
라) 구성비 = 다)/나) * 100
마) 의미적 한 단어의 형태소 분석 시도한 의미적 한 단어 수
바) 분석 성공한 의미적 한 단어의 수
사) 의미적 한 단어 분석률 = 바)/마) * 100

가	나	다	라 (%)	마	바	사(%)
1월	15174	1901	12.53	1074	1021	95.1
3월	17354	2054	11.84	1236	1210	97.9
4월	16711	2000	11.97	1062	1027	96.7
7월	17880	2156	12.06	1233	1208	97.8
8월	17781	2022	11.37	1182	1165	98.6
계	84900	10133	11.94	5787	5631	97.3

< 표 3 > 표본 말뭉치 분석 결과표

'의미적 한 단어' 형태소 분석이 성공한 결과를 분석 과정에 따르는 유형별로 분석하면 다음과 같다.



< 그림 5 > '의미적 한 단어' 분석 유형별 구성비

V. 결론

본 논문에서는 한글, 한자, 영어, 기호의 결합된 형태로 한 어절 또는 여러 어절에 걸쳐서 구성되면서 하나의 뜻으로 사용되는 단어를 '의미적 한 단어'라 정의하였다. 한국어 문서에서 나타나는 이러한 '의미적 한 단어'의 유형을 분류하였으며 그에 대한 형태소 분석 기법을 제시하였다. 구현된 형태소 분석기는 '의미적 한 단어'에 대해 97%의 분석률을 가졌다.

'의미적 한 단어' 처리를 위한 지식베이스를 확장시키기 위해, 더 많은 자료에서 '의미적 한 단어'를 추출하여 각 분석 유형 분류를 하여야 할 것이다. 또한, 처리 속도의 저하를 최소화하기 위해서 '의미적 한 단어'의 분석시 불필요한 후보어절 생성을 최대한 제약하기 위한 연구가 필요하다.

참고 문헌

[1] 이영식, 권혁철, 채영숙, '한글 철자 검사기에서의 사전', '92 우리말 큰잔치, pp. 36-44, 1992

[2] 이승선, 송주원, 황규영, 최기선, 'TRIE구조를 이용한 한국어 전자사전을 위한 데이타베이스 인덱스 구조', 한국정보과학회 봄 학술발표논문집 Vol 21. No1, pp. 849-852, 1994

[3] 채영숙, 김재원, 김민정, 권혁철, '한국어 철자 검색을 위한 형태소 분석 기법', '91 우리말 정보화 잔치, 국어정보학회, pp. 179-186, 1991

[4] 강승식, '음절 정보와 복수어 단위정보를 이용한 한국어 형태소 분석', 서울대학교 컴퓨터공학과 박사학위논문, 1993

[5] 이은철, 이종혁, '계층적 기호 접속정보를 이용한 한국어 형태소 분석기의 구현', 한글및 한국어 정보처리 학술발표 논문집, pp. 95-104, 1992

[5] 김민정, 권혁철, '한국어 형태소 분석에서의 수사처리', 제3회 한글 및 한국어 정보 처리 학술발표논문집, 한국인지과학회, pp. 813-862, 1991

[6] 부산일보사 Style Book