문장패턴을 활용한 형태소 분석기의 설계

The Design of morphological analyzer using a sentence-patterns

홍성웅*, 연제용*, 박찬곤*
*청주대학교 컴퓨터정보공학과
e-mail:leoking@naver.com

Sung-woong Hong*, Che-Yong Yon*, Chan-Khon Park*
*Dept of Computer Information & Engineering, Chong-Ju University

요 약

본 논문에서는 한국어의 문장패턴을 활용한 형태소 분석기를 설계하였다. 어절기반의 형태소 분석기들이 갖는 형태소 분석 정보는 어절의 품사 등의 기초적 정보만을 포함한다. 본 논문에서 제안하는 문장패턴을 활용한 형태소 분석기는 문장단위의 형태소 분석을 제안하였고 형태소 분석단계에서 구문분석과 문장패턴이 갖는 의미정보를 포함함으로서 분석결과의 활용도를 높이도록 하였다.

제안된 형태소 분석기의 결과를 활용하여 질의 응답시스템, 정보 검색 등의 분야에서 구문분석, 의미분석의 단계를 최소화 하여 결과를 얻을 수 있을 것으로 기대한다.

1. 서론

모든 자연어 처리 분야에서 가장 근본이 되는 것은 형태소 분석이며, 형태소 분석의 목적은 주어진 문을 형태소·말의 나열로 분해하고, 각각의 형태소·말의 품사 등을 결정하는 것이다[1].

기존의 형태소 분석기는 어절단위의 형태소 분석기이다. 이는 근접한 형태소간의 결합조건을 검사하여 형태소의 일부를 교체, 삽입, 삭제하는 방법을 많이 사용하고 있다. 어절 단위 형태소 분석기는 가능한 모든 후보를 생성하여 문법 형태소와 비교한 후 형태론적 변형을 일으킨다.

또 다른 형태소 분석방법으로는 음절단위 형태소 분석기가 있다. 음절단위 형태소 분석기는 어절 단위 형태소 분석기 보다 효율적인 후보 생성을 할 수 있다. 그러나 음절 하나하나를 분리하여 결합규칙을 비교해야 하는 제약이 있다.

본 연구의 형태소분석기에서는 문장패턴을 적용함으로써 기존의 형태소 분석기의 분석과정을 최소화하여 질의응답 시스템을 보다 효율적이고 빠른 결과를 얻는데 목적이 있다.

2. 관련연구

형태소분석기는 한국어를 처리함에 있어 여러 사전을 기반으로 한 여러 단계를 거쳐 실행된다. 기존의 형태소 분석기들은 사용되는 용도에 따라 많은 성능차이를 보이고 있다. 그것은 각기 사용되는 용도에 따라 사용되는 알고리즘의 차이가 존재하기 때문이다. 예를 들어 맞춤법 검사를 하는 형태소분석기와 정보검색을 위한 형태소 분석기는 확연한 차이점을 보인다. 맞춤법 검사를 위한 형태소 분석기는 최우선을 고려되어야 할 것은 정확성이 이다. 이에 비해 정보검색용 형태소 분석기는 정확성과 빠른 속도를 최우선으로 하고 있다[2].

현재의 형태소 분석 시스템은 형태소분석, 구문분석, 의미 분석 단계를 별개의 과정으로 처리하고 있으며, 이로 인해 각 단계에서 문장 분석이 별도로 발

생하고 있다. 또한 구문분석 단계에서 구문 규칙의 적용 후 의미 분석 단계에서 구문규칙의 재적용이 발생가능 하다. 이러한 문제점을 해결하기 위해 본 연구에서 제시하는 형태소 분석기는 형태소분석, 구문분석, 의미 분석을 하나의 단계에서 적용가능하게 하는 것이다. 구문 분석과 의미분석을 문장패턴을 이용하여 한번에 처리하고 문장패턴을 활용한 사전검색 후보를 통해 형태소 분석 단계에서 사전검색에 대한 효율성을 기대할 수 있다.

3. 문장패턴

문장패턴이란 문장의 구조 유형으로서 수많은 구체적인 문자으로부터 추출하여 구조적 형식의 공통성에 따라 공식화한 틀을 가리킨다[3].

한국어의 문장에서는 수많은 문장 패턴이 다양한 형태로 나타나고 있다. 이러한 다종다양한 문장에서도 일정한 규칙의 구조적 형식을 따르고 있으며, 모든 문장에서 나타나는 일정한 틀을 제한된 수의 유형으로 나누었다.

한국어 문장패턴의 연구는 미국이나 일본에 비해 뒤늦게 시작되었고 뒤떨어져 있다. 미국은 1920년대 초부터 이미 문장의 패턴을 연구하기 시작하였고, 이웃한 일본에서도 1940년대부터 연구를 시작하였다[4,6]. 국내에서는 1960년부터 문장패턴에 관한 연구가 활발히 이루어졌다.

한국어의 문장패턴을 보면 3형, 4형, 5형, 6형, 7형, 12형, 41형 등 설정된 수가 학자에 따라서 각각 다르게 나타나고 있다[5]. 본 논문에서 제시된 문장패턴은 위의 문장 패턴을 이용하여 그 특징을 분석하여 만들어 졌다.

여기서 추출된 문장은 일반 문장 패턴으로 분류할 수 있는데 총 32개의 기본 문장 패턴을 가지게 되었다. 또한 일반문장패턴을 제시한 후 의문질의어 패턴을 위한 문장 분석을 자연어처리에 적합한 구문 및 의미 분석 단계에서 사용 가능하도록 의미적인 결합관계가 적용된 새로운 의문 질의어 패턴을 생성하였다.

3.1 일반문장패턴

일반문장 패턴은 서술어에 따라 동사형, 형용사형, 지정사형(N+이다, N+이 아니다)으로 분류하였고, 여러 형태로 나타나는 조사를 분류하여 [표 1]과 같이 정리하였다[6].

기본구성		패턴구성	수
동사형	N1+N3+V	NH1+NH3+V NH1+NL3+V :	24
	N1+N2+V	NH1+NA2+V NH1+NH2+V :	22
	N1+N2+N3+V	NH1+NH2+NL5+V :	22
형용사형	N1+N3+A	NA1+NH3+A NH1+NH3+A :	14
	N1+N4+A	NA1+NA4+A NH1+NH4+A :	4
	N1+N7+A	NH1+NH7+A NA1+NC7+A :	4
지정사형	N1+N	NI1+NI NL1+NT :	8
	N1+N3+N	NH1+NA3+NA NH1+NH3+NB :	3
	N1+N4+N	NH1+NH4+NA NH1+NH4+NB :	2

[표 1] 일반문장 패턴

3.2 질의어 문장패턴을 위한 문장 분석[5]

기본 구성		패턴 구성	수
동사형	N1+V	WNI1+V WNH1+V :	7
	N1+N3+V	WNH1+NH3+V NH1+WNL3+V :	24
	N1+N5+V	WNH1+NL5+V NC1+WNC5+V :	5
지정사형	N1+N	WNI1+NI NL1+WNT :	8
	N1+N3+N	WNH1+NA3+NA NH1+WNH3+NB :	3
	N1+N4+N	WNH1+NH4+NA NH1+WNH4+NB :	2
형용사형	N1+A	WNA1+A NH1+WA :	7
	N1+N3+A	WNA1+NH3+WA NH1+WNH3+A :	14
	N1+N4+A	WNA1+NA4+A NH1+WNH4+A :	4

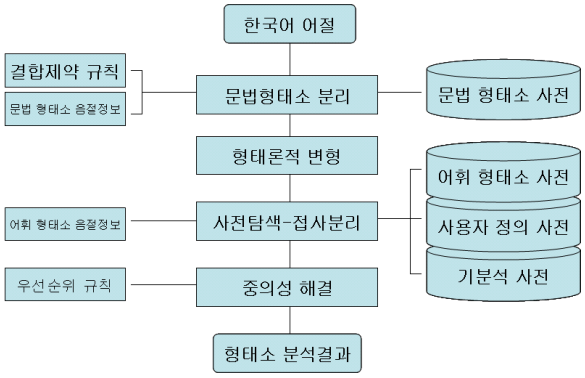
[표 2] 질의어 문장패턴

의문질의어 패턴을 추출하기 위하여 기존의 질의응답 시스템에서 사용된 자연어 질의어를 표에서 제시한 기준에 따라 태깅하였고, [표 2]와 같이 정리하였다.

4. 문장패턴 활용한 형태소 분석기

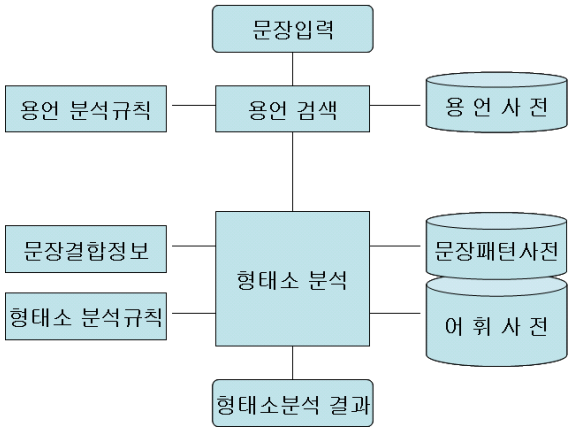
4.1 형태소 분석

기존에 제안 된 형태소분석기의 분석은 어절을 기준으로 분석하며, 문법형태소 분리, 형태론적 변형 복원, 사전탐색 및 접사분리, 중의성 해결의 과정을 거친다. 문법형태소는 결합제약 규칙과 문법 형태소 음절정보를 활용하고, 문법 형태소 사전을 검색하여 분리하고 형태론적 변형을 복원한다. 사전탐색과정에서는 어휘 형태소 음절 정보를 활용하여 어휘 형태소 사전, 사용자 정의 사전, 기분석 사전을 검색하고 접사를 분리한다. 중의성 해결은 사전 탐색을 통해 얻어진 결과로 부터 우선순위 규칙을 적용하여 형태소 분석결과를 얻는다.



[그림 1] 기존의 형태소분석[7]

본 논문에서 제안한 방법은 어휘 형태소 사전에서 용언사전을 분리하고 용언 분석규칙을 우선적으로 적용하여 용언의 원형을 찾아 용언 사전을 검색함으로써, 문장패턴 사전에서 검색된 용언이 가질 수 있는 문장결합 정보로부터 형태소 분석이 가능한 후보를 생성하고 이 후보들로부터 형태소 분석 결과를 얻는다.



[그림 2] 제안된 형태소 분석기

어절기반 분석의 형태소 분석결과는 품사정보와 앞뒤 어절의 결합 관계정도의 정보를 가진다. 따라서 형태소 분석결과의 사용 분야에 따라 구문분석과 의미분석을 거치는 번거로움을 갖는다. 그러나 본 시스템에서 제안된 방식으로 얻어진 형태소 분석결과는 구문분석단계를 거치지 않더라도 문장전체의 유기적 관계와 문장패턴 정보에서 갖고 있는 의미소성까지 갖게 된다.

예) <입력문장>

과학자들은 대기중에서 염화플루오르화탄소를 검출하였고 이것이 오존층을 파괴한다는 것 을 입증하였다.

<적용된 문장 패턴>

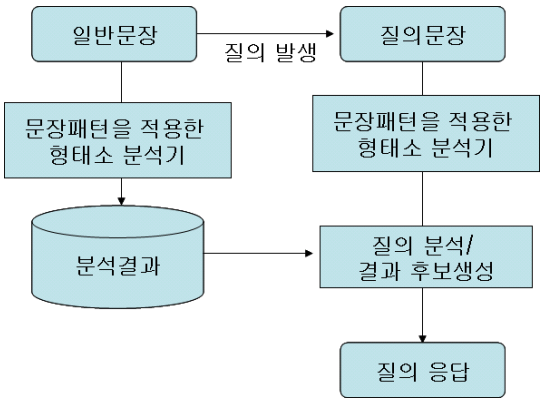
검출하다 - 동사 141형(NH1+NL7+NM2+V) 파괴하다 - 동사 66형 (NC1+NC2+V) 입증하다 - 동사 73형 (NH1+NB2+V)

<분석결과>

과학자들은(N+s+j) 대기중에서 (N+j) 염화플루오르화탄소를 (N+j) 검출하였고(V) + 이것이 (N+j) 오존층을 (N+j) 파괴한다는 (V) 것을 (N+j) 입증하였다 (V) ⇒ NH1(+ NL7 + NC2 +V)+ (NB2 (=NC1+NC2+V) + V) ⇒ NH1 + NL7 + NC2 = 과학자들은 대기중에서 염화플루오르화탄소를 검출하였다. NC1+NC2+V = 이것(염화플루오르화탄소)이 오존층을 파괴한다. NH1+NB2+V = 과학자들이 것(이것(염화플루오르화탄소)이 존층을 파괴한다)을 입증하였다.

4.2 형태소분석 결과의 활용

문장들의 패턴은 서술어가 결정짓고, 같은 서술어 또는 같은 의미를 갖는 서술어들은 같은 패턴을 갖게 되며, 일반 문장에서 도출된 의문문의 문장패턴과 도출대상의 일반 문장의 문장패턴이 일치한다. 의문문의 의문요소와 일치되는 문장패턴의 요소를 추출함으로써 응답후보의 생성이 가능하다[3,5].



[그림 3] 제안된 형태소 분석기 활용의 예

예) <일반문장>

메소드는 컴퓨터 명령어를 규정하고, 클래스 객체의 특성은 관련 데이터를 규정한다

<적용된 문장 패턴>

규정하다 - 63형(NA1+NA2+V)

<일반문장 분석결과>

메소드는(N+j) 컴퓨터 명령어를(N+j) 규정하고(V), 클래스 객체의 특성은(N+j) 관련 데이터를(N+j) 규정한다(V) (NA1+NA2+V)+(NA(=NA3+NA)1+NA2+V)

<질의문장>

무엇이 컴퓨터 명령어를 규정하는가?

<질의문장 분석결과>

WNA1+NA2+V

<결과 후보생성>

WNA = 메소드, 특성(클래스 객체의 ~)

5. 결론 및 향후연구과제

본 논문에서는 기존에 연구 되었던 어절 분석기반의 형태소 분석시스템에서 제시되었던 분석결과보다 발전된 형태의 정보를 갖는 형태소 분석결과를 제시한다.

문장성분에서 서술어가 갖는 문장패턴 정보를 형태소 분석에 적용함으로써 구문분석과 의미분석 단계를 거쳐야만 얻을 수 있는 정보 중 일부를 형태소 분석의 결과에 적용함으로써 질의응답 시스템, 정보검색 등의 분야에 활용할 수 있다. 형태소 분석단계에서 더 많은 정보를 얻어 내기 위해 처리 시간이 길어 질 수 있으나, 기존의 구문분석, 의미 분석을 거치지 않더라도, 활용할 수 있는 정보를 이전단계에서 생성함으로써 시간을 절약 할 수 있다.

향후 과제는 질의응답시스템, 정보 분석 시스템에 적극적으로 활용 가능한 정보를 분류하고, 문장패턴에 적용하여 다양한 패턴연구와 다양한 형태의 질의를 처리할 수 있는 시스템을 구축해야 할 것이다.

참고문헌

[1] Makoto Nagao, 자연언어 처리, pp.125

[2] 손소현, 정보검색용 한국어 형태소 분석기 구현, 2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호, pp379-382

[3] 송재관, “기계번역을 위한 문장패턴에 관한 연구” pp.30-39, pp.14

[4] 김진한, “한국어 결합가 패턴에 의한 기계번역에 관한 연구”, 청주대학교 석사학위논문 1987.

[5] 우근신, 문장패턴을 이용한 자연어 질의 시스템에 대한 연구 제15회 한글 및 한국어 정보처리 학술대회

[6] 송재관, “ATP-KSP를 이용한 기계번역 시스템에 관한 연구” pp.21-29

[7] 강승식, 한국어 형태소 분석과 정보검색, pp.121

[8] 이석호,홍봉희, “한글 질의어 시스템의 설계 및 구현”,한국정보과학회 논문지, Vol.11,No. 1,84.5.

[9] 윤성희, “한국어 자연 언어 질의 문장분석에 서 스키마 도메인 정보를 이용하는 중의성 해결 기법”, 상명여대 산업 과학 연구,1996.