김일성종합대학학보
정보과학

주체107(2018)년
제64권 제4호

형태부분석모형을 리용한 사전자료기지구축의 한가지 방법

리 명 일

전자사전은 일정한 체계에 따라 작성한 단어, 문법 등을 물리적형태로 콤퓨터에 기억시켜놓은 자료로서 정보처리체계의 중요한 구성부분을 이룬다.

형태부전자사전[1]은 형태부를 올림말로 기억시킨 전자사전으로서 다음과 같이 표시된다.

$$D = (Pref, Tree, Suff, To)$$

여기서 D는 전자사전, *Pref*는 앞붙이, *Suff*는 뒤붙이, *Tree*는 말뿌리, *To*는 토를 나타낸다.

어간전자사전은 어간을 올림말로 기억시킨 사전으로서 단어《헛손질》은 하나의 올림말로 기억된다.

문법규칙사전[2]은 문법규칙에 따르는 내용들을 모두 기억시킨 사전이다.

실례로 조선어인 경우 명사와 명사는 토없이 결합할수 있고 동사와 동사는 토없이 말뿌리끼리 결합할수 있다. 즉 명사와 명사의 결합은 N+N/NP, 형용사와 명사의 결합은 Ad+N/NP, 동사와 동사의 결합은 V+V/VP로 표현할수 있다.

우의 3개 사전을 리용의 측면에서 다음과 같이 분석할수 있다.

① 형태부사전과 어간사전에서 형태단어들은 실지 언어생활에서 쓰이는 단어로서 빈도수 혹은 자모순서로 작성되는 경우 동적으로 끊임없이 갱신된다.

이것은 사전의 용량을 증대시키며 형태부해석과 탐색속도를 감소시킨다.

② 문법규칙사전에서는 조선어품사를 8개로 나누고 품사들사이의 결합관계, 품사와 토들사이의 결합관계를 규칙화함으로써 문장구조분석에 유용하게 리용할수 있다.

그러나 보다 정확한 문장구조분석을 위해서는 세부품사도 고려하여야 하며 세부품사들과 토들사이관계를 모두 규칙으로 반영하는것은 어려운 문제로 나선다.[3]

론문에서는 우의 제한성을 극복하고 형태부사전과 문법규칙사전자료기지를 구축하기 위한 다음과 같은 문제를 제기하였다.

첫째, 앞붙이, 말뿌리, 뒤붙이형태부에 기초하여 조선어품사기초형태부모형을 작성한다.

둘째, 기초형태부분석모형을 리용하여 사전자료기지를 구축하기 위한 방법을 확립한다.

1. 형태부분석모형을 리용한 사전자료기지구축방법

형태부분석에서는 문장속에서 쓰인 단어들이 어떤 의미를 가진 부분(기초형태부)들로 갈라지는가를 밝혀내야 한다. 이를 위하여 개별적인 품사들을 앞붙이, 말뿌리, 뒤붙이로 나누고 형태부를 결정할수 있다. 그러면 개별적인 품사들에 대한 기초형태부분석모형을 고찰하기로 하자.

① 동사

동사의 기초형태부는 말뿌리, 앞붙이로 이루어진다. 여기서 말뿌리를 동사의 원형으로 한다. 즉 동사를 *V*, 동사의 앞붙이기초형태부를 *PV*로, 동사말뿌리기초형태부를 *TV*로 표시하면 동사기초형태부분석모형은 다음과 같다.

$$VR=PV+\sum_{i=1}^{n}TV_i$$

실례로 동사《감돌다》의 기초형태부는 앞붙이《감》과 말뿌리《돌다》로 이루어진다.

② 형용사

형용사의 기초형태부는 앞붙이, 말뿌리, 뒤붙이로 이루어진다. 즉 형용사는 *AJ*, 형용사앞붙이기초형태부를 *PAJ*, 형용사뒤붙이기초형태부를 *SAJ*로 표시한다면 형태부분석모형은 다음과 같다.

$$AJ=PAJ+\sum_{i=}^{n}TAJ_i+SAJ$$

실례로《새빨갛다》의 기초형태부는 앞붙이《새》와 형용사말뿌리《빨가》, 뒤붙이《ㅎ다》로 이루어진다.

③ 명사

명사의 기초형태부는 앞붙이, 말뿌리, 뒤붙이로 이루어진다. 1개 음절로 이루어진 말뿌리는 독자적인 의미를 가르기 힘들기때문에 될수록 2개로 합치고 1개의 기초형태부로 한다.

또한 조선어에서는 명사말뿌리에 동사뒤붙이《하다》가 붙어 동사, 형용사로 되거나《같다, 답다》등과 같은 형용사뒤붙이들이 붙어서 형용사로 되는 명사,《껏, 히》등과 같은 부사뒤붙이들이 붙어서 부사로 되는 명사들은 모두 명사의 형태부로 취급한다.

명사를 *N*, 명사앞붙이기초형태부를 *PN*, 명사말뿌리기초형태부를 *TN*, 명사뒤붙이기초형태부를 *SN*으로 표시하면 명사형태부분석모형은 다음과 같다.

$$N = PN +\sum_{i=1}^{n}TN_i+SN$$

④ 부사

부사의 형태부는 말뿌리, 뒤붙이로 이루어진다. 여기서 부사는 일반부사, 상징부사, 상태부사, 접속부사, 부정부사로 나눈다.

부사를 *AV*, 부사의 말뿌리기초형태부를 *TAV*, 부사의 뒤붙이기초형태부를 *SAV*로 표시하면 부사기초형태부분석모형은 다음과 같다.

$$AV=\sum_{i=1}^{n}TAV_i+SAV$$

우와 같은 방법으로 나머지품사들에 대하여 기초형태부분석모형을 작성할수 있다.

다음은 우의 조선어품사에 대한 형태부분석모형을 리용하여 말뿌리사전, 앞붙이사전, 뒤붙이사전을 독자적으로 구축한다. 이때 8개 품사에 해당한 단어들이 말뿌리를 중심으로 하여 결합정보와 함께 등록된다.

앞붙이사전은 $PD=(Prex, Pos, Com)$, 뒤붙이사전은 $SD=(Suff, Pos, Com)$, 말뿌리사전은

TD = (*Wd*, *Pos*)로 표시된다. 여기서 *Prex*는 앞붙이, *Pos*는 품사, *Suff*는 뒤붙이, *Com*은 결합정보이다. 결합정보 *Com*은 앞붙이와 품사와의 결합관계를 나타낸다.

우의 기초형태부분석모형과 3개의 사전모형에 기초하여 사전자료기지를 구축하는 과정은 다음과 같다.

① 1개의 실례문장을 입력한다.

② 문장을 공백단위로 하여 형태단어를 분리한다.

$$S => (FW1,\ FW2,\ \cdots, FWn)$$

여기서 *S*는 문장, *FWi*는 형태단어를 나타낸다.

③ 형태단어에서 토를 분리하여 토결합사전에 등록한다.

$$FWi => BFW + To$$

여기서 *BFW*는 결합단어, *To*는 토형태부를 나타낸다.

④ 현재 구축된 형태부사전과 결합정보, 형태부해석기를 리용하여 형태부를 분리한다.

⑤ 분리된 기초형태부들을 해당한 품사에 따라 사전에 등록한다.

2. 실현 및 평가

우의 과정을 통하여 구축된 사전자료기지는 앞붙이, 뒤붙이, 말뿌리가 서로 독립적으로 구축되여있으므로 어간사전에서 단어들의 중복으로 하여 생기는 용량증가문제를 해결할수 있다.

형태부분석모형을 리용하여 사전자료기지를 구축하고 조선어문장에 대한 형태부해석과 문장구조분석을 진행한 결과는 표와 같다.

표. 형태부해석과 문장구조분석결과

번호	선행한 방법의 해석률/%		제안한 방법의 해석률/%		문장수/개
	형태부	문장구조	형태부	문장구조	
1	60.5	58.7	77.8	67	2 500
2	59.6	55	76	66	3 000
3	89	88	91	89	4 000
4	75	73	81	77	3 000
평균	71.025	68.675	81.45	74.75	3 125

표에서 보는바와 같이 제안한 방법의 형태부해석률은 81.45%, 문장구조해석률은 74.75%로서 선행한 방법보다 높다.

맺 는 말

앞붙이, 말뿌리, 뒤붙이형태부에 기초하여 조선어품사기초형태부모형을 작성하고 기초형태부분석모형을 리용하여 사전자료기지를 구축하기 위한 방법을 실현하였다.

참 고 문 헌

[1] **김일성**종합대학학보(자연과학), 56, 2, 39, 주체101(2012).
[2] Q. N. Rockiman; Natural Language Process, 2, 2, 32, 2015.
[3] Paul Piwek; Natural Language Process, 1, 2, 12, 2010.

주체107(2018)년 8월 5일 원고접수

A Method of Dictionary Database Construction Using Morphemic Analysis Pattern

Ri Myong Il

The dictionary database is an important part in information management system and is composed of words, sentences, and grammar.

In this paper we suggested a method of dictionary database construction using morphemic analysis pattern.

Key words: database, language process