김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제11호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 62 No. 11 JUCHE105(2016).

류형패턴을 리용한 조선어 문장분류의 한가지 방법

리명일, 김동수

언어생활에서 쓰이는 무수한 문장들가운데는 단순한 구조의 문장도 있고 진술내용에 따라 여러 단계로 확대된 구조를 가지는 문장도 있다. 많은 문장들을 최대로 일반화하면 가장 기초적인 유한한 수의 문장을 얻을수 있다. 기초문장이란 최대로 일반화된 문장구조에 해당한 어휘를 채워넣은 문장을 말한다.[1, 3] 실례로 다음과 같은 세가지 형태의 문장들을 기초문장이라고 말할수 있다.

N주+V술,

Adj규+N주+V술,

Adj규+N주+Adv+V술

이 세가지 문장구조를 비교해보면 첫번째 구조가 일반화수준이 제일 높고 세번째 구조는 일반화수준이 제일 낮다.

문장구조는 문장의 골격으로서 단어는 문장구조속에 들어갈 때에만 문장속에서 자기의 위치를 차지하고 사상전달의 기능을 수행할수 있다.

선행연구[1, 2]에 제시된 기초문장들가운데서 대표적인 문장류형을 보면 다음과 같다.

1류형: 《N주+V술》문형의 문장(밀림이 설레인다.)

3류형: 《N주+N이술》문형의 문장(옥이는 모범학생이다.)

5류형: 《N주+N위+V술》문형의 문장(대렬이 청봉에 이르렀다.)

8류형: 《N주+N수+V술》문형의 문장(책상은 나무로 만들었다.)

⋮

14류형: 《N주+N대+N수+V술》문형의 문장(당이 혁명대오를 주체사상으로 일색화하였다.)

우의 류형들에서 기호 《N》은 명사, 《V》는 동사, 《A》는 형용사, 《주》는 주어, 《술》은 술어를 나타낸다. 또한 3류형의 《N이술》에서 《이》는 바꿈토를, 5류형의 《N위》에서 《위》는 위치(장소)를, 8류형의 《N수》에서 《수》는 수단을 나타낸다.

기초문장이 뜻을 가지자면 문장류형에 대응한 단어들이 옳게 선정되여야 한다. 실례로 1류형기초문장의 술어에는 일반적으로 자동사가 놓이는데 모든 자동사가 오는것이 아니라 주로 자연현상을 나타내는 자동사(개이다, 흐리다, 얼다, 여물다, 뜨다, 피다 등), 정신적 및 육체적작용을 나타내는 자동사(자다, 놀다, 곪다, 걷다, 노래하다 등), 사건의 발생을 나타내는 자동사(생기다, 일어나다, 발생하다 등), 피동형동사들(열리다, 시작되다, 건설되다 등) 등이 놓인다.

보는바와 같이 선행연구에서는 조선어문장을 14가지 류형으로 제한하였는데 이것은 콤퓨터에 의한 학습에서 여러가지 구조와 의미를 내포하고있는 문장에 대한 구체적인 분석을 불가능하게 한다. 여기로부터 론문에서는 기계사전에 등록할수 있는 류형패턴을 리용한 조선어 문장분류방법을 제기한다.

1. 구조요소목록을 리용한 조선어 문장류형작성

임의의 문장분석은 구조분석과 의미분석으로 나누어 론의할수 있다.

정의 문장의 구조를 표현하는 요소를 구조요소, 구조요소들의 목록을 구조요소목록이라고 한다.

일반적으로 문장이 주어지면 그것에 대응하는 문장류형이 결정되며 문장류형은 구조요소에 의하여 결정되는 류형패턴으로 표현할수 있다.[4]

조선어는 토교착어로서 문장구조는 토에 의하여 표현된다. 따라서 조선어문장의 구조요소에는 각종 토와 토결합들이 속한다.

우에서 론의한 대표적인 14가지 류형들가운데서 세가지 류형을 구조요소목록을 리용하여 변환하면 다음과 같다.

우선 《N주+V술》은 다음과 같이 변경된다.

《N주+V술》⇒《NP+Tzg+Vz》

여기서 Tzg는 주격토, Vz는 자동사이다.

다음 《N주+N이술》은 다음과 같이 변경된다.

《N주+N이술》⇒《NP+Tzg+NP+Tb+Tm》

여기서 Tb는 바꿈토, Tm은 맺음토이다.

끝으로 《N주+N대+N수+V술》은 다음과 같이 변경된다.

《N주+N대+N수+V술》⇒《NP+Tzg+NP+Td+NP+Tzo+VP》

여기서 Td는 대격토, Tzo는 조격토이다.

보는 바와 같이 문장류형은 토와 토결합을 요소로 하는 구조요소목록을 리용하여 표현할수 있는데 류형패턴작성과정은 다음과 같다.

① 입력문장을 형태단어로 분리한다.

② 분리된 형태단어에 대하여 구조요소목록, 단어 및 단어결합사전, 토결합규칙을 리용하여 토와 합성토를 추출한다. 이때 토 및 합성토의 추출은 뒤최장일치법에 기초하여 진행한다.

③ 토 및 합성토추출을 진행하여 얻어진 합성토와 어간이 토결합목록과 어간사전에 등록되여있는가를 확인하고 등록되여있지 않으면 등록한다.

④ 문장의 마지막에 위치한 동사에 대하여 동사원형을 얻고 그것에 대한 세부품사를 결정한다.

⑤ 토 및 합성토의 추출후 토들과 동사로 이루어진 문장류형패턴을 얻은 다음 문장류형패턴에 대하여 류형패턴사전에 등록되여있는가를 검사하고 등록되여있지 않으면 등록한다. 이때 류형패턴과 함께 입력문장을 동시에 등록한다.

우와 같은 조작을 반복하여 문장류형패턴을 얻고 그것에 대응한 조선어문장들을 정돈하여 문장자료기지를 구축한다.

작성된 문장류형패턴에는 순수 토와 토결합만이 아니라 문장의 의미해석에서 중요한 역할을 하는 동사의 세부품사도 들어있다.

2. 실험 및 분석

실험을 위하여 조선말대사전에 등록되여있는 5 000개 문장들을 대상으로 문장류형패턴을 작성하였다.

여기에 기초하여 제안한 체계와 이전체계의 성능을 비교한 결과는 표와 같다.

표에서 보는바와 같이 토를 기본으로 하여 문장류형을 분류하였으므로 문장류형의 개수는 종전 14개로부터 2 000개로 증가하였다.

한편 구조요소목록을 리용하여 문장류형패턴을 얻은 후 구조분석률은 70%로부터 77%로, 의미분석률은 66%로부터 77%로 개선되였다.

표. 지표별성능비교결과

지표	선행방법	현재방법
문장류형개수/개	14	2 000
문장구조분석률/%	70	77
문장의미분석률/%	66	68

맺 는 말

조선어문장들을 류형별로 분류하고 기계사전에 등록할수 있는 류형패턴을 작성한데 기초하여 조선어문장을 분류하기 위한 한가지 방법을 제기하였으며 실험을 통하여 그 효과성을 검증하였다.

참 고 문 헌

[1] 방금숙; 외국어로서의 조선어실천문법연구, **김일성**종합대학출판사, 102～120, 주체98(2009).

[2] 배광희; 콤퓨터응용언어학, **김일성**종합대학출판사, 32～43, 주체92(2003).

[3] Chris Callison-Burch; A Handbook for Language Engineers, Elservier, 30～46, 2003.

[4] Jan Strunk; A Comparative Evaluation of a New Unsupervised Sentence Boundary Detection Approach on Documents in English and Portuguese, Elservier, 44～56, 2011.

주체105(2016)년 7월 5일 원고접수

A Method for Classification of Korean Sentences using Pattern

Ri Myong Il, Kim Tong Su

We completed method for classification of Korean sentences using pattern of sentence form registered machine dictionary.

Key words: database, data classification