기계 번역을 위한 한국어 문장 패턴에 관한 연구

송재관, 홍성웅, 박찬곤
청주대학교 전자계산학과

A Study on the Sentence Pattern of the Korean Language for Machine Translation

Jae-Gwan Song, Sung-Woong Hong, Chan-Khon Park

요 약

본 연구에서 말뭉치를 이용하여 기계 번역을 위한 한국어 문장 패턴을 추출하였다. 문장 패턴은 해당 언어의 기본 문법 구조를 가지고 있기 때문에 언어 습득을 위해서 유용하다. 기계 번역을 위해서는 기본 문법 구조뿐만 아니라 각 단어간의 의미 관계를 나타낼 수 있어야 한다. 본 연구는 품사 태깅 및 명사에 의미 소성을 태깅하여 한국어의 문장 패턴을 추출하였다, 추출된 문장 패턴은 구문분석시 애매성을 해소할 수 있으며, 동음다의어의 해석이 가능하며, 의미의 부정합 판정이 가능하다.

1. 서론

자연언어가 갖는 문장 패턴은 언어분석과 기계번역을 위해 중요하다. 한국어의 문장에 나타나는 문장 패턴을 파악하려는 연구는 지금까지 많은 국어학자들에 의해 연구되어 왔다[1,2,3,4,5,6]. 그러나 주로 국어학적인 입장에서 연구로서 기계번역을 위한 문장패턴으로 적용하기에는 미흡하였다.

본 연구는 술어 기반의 문형추출을 통한 모호성해결에의 접근 방식[7]이나 자연질의어를 술어-인자 구조로 변환하는 방식[8]과 같은 맥락의 술어 중심구도에서 말뭉치를 이용한 문형추출을 통해 기계번역을 위한 문장패턴을 접근, 설정하고자 한다.

본 연구에서는 현재 제안되어 되어 있는 한국어의 일반적인 문장패턴에 대하여 분석하였고, 말뭉치를 구축하여 1차적으로 한국어의 기본 문장 패턴을 추출하였고, 2차적으로 명사의 의미소성을 부여하여 기계번역을 위한 문장패턴을 추출하였으며, 추출된 문장패턴의 유용성을 제시하였다.

2. 한국어의 문장패턴

2.1 문장패턴

문장패턴이란 문장의 구조 유형으로서 수많은 구체적인 문장으로부터 추출하여 구조적형식의 공통성에 따라 공식화한 틀을 가리킨다.

실제 언어 행위에서 표현되는 무한한 수의 구체적인 문장들은 일정한 구조적형식에 따라 제한된 수의 유형으로 공식화될 수 있으며 이렇게 공식화된 하나의 유형들은 일반적으로 수많은 구체적인 문장을 대표하게 되는데 이와 같이 수많은 개개의 구체적인 문장 가운데서 같은 종류의 문장을 대표할수 있는 틀을 문장패턴이라 한다[2].

2.2 문장패턴에 대한 연구

한국어에서의 문장패턴연구는 미국이나 일본에서의 문장패턴연구보다는 뒤늦게야 시작되었으며 또 뒤떨어져있다[9].

미국이나 일본의 경우에서 보면 미국에서는 1920년대초부터, 일본에서는 1940년대부터 문장패턴에 대한 연구가 본격적으로 진행되기 시작하여 이미 수십편의 논문과 저서들을 통하여 그 연구성과들이 발표되고 있다.

우리말에서는 60년대 이전 시기에도 문장패턴에 대한 논의가 있었던 것만은 사실이나 모두가 다 문법서의 서술에서 문장성분을 분류하고 기술하는 한 방편으로 제시한것에 지나지 않는바 우리말 문장패턴에 대한 본격적인 연구라 보기는 어렵다. 우리말의 경우 문장패턴이 문장론의 한 독자적인 분과로 등장되어 본격적으로 연구되기 시작한 것은 1960년대 후반기부터이다[2].

이와 같이 우리말에서의 문장패턴연구는 다른 언어의 문장패턴연구에 비해 퍽 뒤늦게 시작되었을뿐만 아니라 그 연구성과 역시 크지 못하다.

또한 연구가 주로 국어학적인 입장에서 기술된 것이어서 전산학적인 측면에서의 재조명이 필요한 실정이다.

2.3 한국어 문장패턴 조사분석

국내외의 문장패턴을 살펴보면 영어의 경우는 학교 문법의 5형식, Jeperson의 8형식, Ross와 Doty의 9형식, 구조주의의 Fries의 5형식, Stryker의 6형식, 변환생성주의의 촘스키의 문장패턴 등 그 구현이 학자와 유파에 따라 심한 차이를 보인다.

일본어의 경우 3형, 4형, 12형, 30형, 20형 등이 있다.

한국어의 문장패턴은 3형, 4형, 5형, 6형, 7형, 12형, 41형 등 설정된 수가 학자에 따라서 각각 다르게 나타나고 있다[2, 9]. 한국어의 문장패턴 설정 기준을 살펴보면 첫째, 각이한 문장성분의 부동한 배합방식을 전면적으로 고려해야 한다는 견해와, 둘째, 근간성분만 고려하되 그중에서도 서술어를 중심으로 고려해야 한다는 견해로 나누어 볼 수 있다.

지금까지 나타나 있는 문장 패턴들중 기계 번역을 위한 문장 패턴에 가장 접근한 것으로 41형을 들 수 있다. 41형[2]은 자동사 술어 문장패턴 10형, 타동사 술어 문장 패턴 9형, 양면동사 술어 문장 패턴 9형, 형용사 술어 문장패턴 8형, 명사 술어 문장패턴 5형으로 술어로 되는 단어들의 의미론적인 특성을 고려하여 분류하였다.

41형은 양면동사 9형이 자동사와 타동사 술어 문장 패턴에 중복되어 나타나고 있을뿐더러 기계번역을 위해서 필수적인, 의미해석에 따른 문제점을 해결할 수가 없다.

3. 한국어의 말뭉치 구축과 문장패턴의 추출

3.1 말뭉치 구축

한국어의 문장 패턴을 조사하기 위한 자료의 선택은 문장패턴의 추출에 있어서 중요한 문제중의 하나이다. 근본적으로 이 문제를 해결하기 위해서는 한국어로 이루어진 모든 문장을 조사하여야 하지만 이것은 엄청난 시간이 필요할뿐더러 실제적으로 불가능하다. 그러므로 일단은 조사를 진행하고 그 결과의 정확성 여부를 판단하여야 한다.

본 연구에서는 문장패턴을 추출하기 위하여 문법적으로 가장 표준이 될 수 있는 초·중·고등학교의 국어 책으로, 국민학교 국어책 6권, 중학교 국어책 3권, 고등학교 국어책 2권을 분석 대상으로 삼았다. 말뭉치는 1차적으로 품사와 격조사에 대한 정보를 태깅하였고, 2차적으로 의미 소성을 태깅하였다.

말뭉치의 구성을 위하여 품사는 "학교 문법 통일안"에 의거한 9품사 이외에 용언에 "지정사(이다,아니다)"를 추가하였다. <표 3-1>은 한국어의 품사 분류표이며 말뭉치 구성에 필요한 품사 기호이다.

<표 3-1> 한국어의 품사 분류표

구 분	품 사	기 호
체 언(N)	명 사	A NA
	대 명 사	B NB
	수 사	C NC
용 언(V)	동 사	A VA
	형 용 사	B VB
	지 정 사	C VC
수식언(A)	관 형 사	A AA
	부 사	B AB
관계언(S)	조 사	A SA
독립언(I)	감 탄 사	B IB

또한 여러 형태로 나타나는 조사는 <표 3-2>과 같이 대표격 조사로 표기한다.

<표 3-2> 조사 분류표

대표조사	코드	조사 그룹
가	1	이, 가, 께서, 에서, 서
을	2	르, 을, 를
에	3	에, 에게, 한테, 께, 더러, 보고
와	4	과, 하고
로	5	로,에게로,으로
보다	6	과/와, 처럼, 만큼, 보다, 하고
에서	7	에서, 으로부터, 로부터, 서
를 위해	8	을 위해, 을 위하여, 을 위하여서
에 의해	9	에 의해, 에 의하여, 에 의해서
라고	10	라고, 이라고, 고
에 대해	11	에 대해, 에 대하여

3.2 문장패턴의 추출

3.2.1 기본 문장패턴 추출(32형)

자연 언어 이해 시스템이나 중간언어 방식을 이용한 기계번역시스템에서 자연 언어 문장을 이해하고 번역을 하기 위해서는 대량의 지식을 이용한 의미구조의 생성이 요구된다. 한국어 문장의 내부 의미 구조를 생성하기 위해서는 각 단어에 해당하는 개념과 개념들 사이의 개념적 관계를 나타내는 지식들이 요구된다[7]. 본 연구에서는 말뭉치로부터 문장패턴을 추출하여 자연어 처리에서 나타나는 몇가지 문제점을 해결하고자 한다. 한국어의 문장패턴은 동사와 형용사 같은 술어 자체의 어휘적 의미를 실현시키기 위해 문장을 구성하는 데 없어서는 안될 필수적 성분과 생략이 가능한 수의적 성분으로 구성되는 특징이 있다[2].

이러한 특징을 이용하여, 구축된 한국어 말뭉치로부터 문장패턴을 추출하였다.

본 연구에서 추출된 문장패턴은 편의상 서술어에 따라 동사형, 형용사형, 지정사형(N+이다, N+이 아니다)으로 분류하였다.

말뭉치로부터 추출한 결과는 <표 3-3>와 같다.

<표 3-3> 추출된 문장패턴의 유형

종 류	유 형
동 사 형	19형 문장패턴
형용사형	8형 문장패턴
지정사형	5형 문장패턴
합 계	32형 문장패턴

위에서 제시된 문장패턴은 구문 분석 결과 모호성을 해결할 수 있다. 가령 문장 "나는 책상에 있는 연필을 보았다,"에서 체언구 "책상에"는 구문 분석 결과 모호성이 발생하게 된다. 즉, 체언구 "책상에"는 술어 "있다"와 "보다"에 개념관계로 연결될 수 있지만 각 술어의 문장패턴 정보를 이용함으로써 "책상에"는 술어 "있다"와 개념관계로 연결됨을 알 수 있다. "있다" 동사는 위에서 제시된 동사문장패턴 1형의 문장 구조를 가지며 "보다" 동사는 11형의 문장 구조를 갖는다.

① 나는 연필을 보았다. = N1 + N2 + V
= 동사 11형

② 연필이 책상에 있다. = N1 + V
= 동사 1형

이와 같이 조사 정보와 술어 정보로 쓰인 문장패턴을 이용하여 구문 모호성을 해결할 수 있다[12]. 그러나 기계번역을 위해 반드시 필요한 것이 의미 모호성의 해결로 위에서 제시된 문장패턴으로는 의미 모호성을 해결할 수 없다. "고양이가 야구공을 창문에 던졌다"와 같은 경우 제시된 문장패턴에 의하면 12형에 해당하는 것으로 구문상 문제가 없다. 하지만 의미상으로는 부적합한 문장이다.

③ 고양이가 야구공을 창문에 던졌다.
= N1 + N2 + N3 + V

문장의 궁극적인 목적이 의미전달인데 문장의 분석시 의미 파악이 제대로 되지 않아서는 안된다.

3.2.2 기계번역을 위한 문장패턴 추출

본 연구에서는 기계번역을 위한 문장패턴을 추출하기 위하여 1단계에서 추출된 32형 문장패턴에 명사의 의미소성을 부여하여 조사, 분석하였다. 이를 위하여 부여된 명사의 의미소성은 <표 3-4>와 같이 10종으로 분류한다[10].

<표 3-4> 명사 의미 소성 분류

소성	코드	의미
abs	NA	추상적인 명사
act	NB	행위
ani	NI	동물
con	NC	구상체
div	ND	종류
hum	NH	인간
loc	NL	장소
num	NN	수량
mat	NM	물질
time	NT	시간

명사 의미 소성이 부여되어 추출된 문장패턴은 <표 3-5>과 같다.

<표 3-5> 기계번역을 위한 문장패턴 유형

종 류	유 형
동 사 형	153형 문장패턴
형용사형	41형 문장패턴
지정사형	19형 문장패턴
합 계	213형 문장패턴

3.2.3 추출된 기계번역용 문장패턴의 유용성

명사의 의미소성을 부여하여 추출한 기계번역용 문장패턴은 다음과 같은 유용성을 가진다.

(1) 의미의 부정합을 판정할 수 있다.

고양이가 야구공을 창문에 던졌다.

앞에서 문제시 되었던 위의 문장은 명사에 의미소성을 부여함으로써 아래와 같이 해결할 수 있다.

고양이(동물), 야구공(구상체), 창문(구상체)로 분류 되고 동사 "던지다"가 가질 수 있는 문장패턴은 동사형 95형인 <N(인간) 이 N(구상체) 을 N(구상체) 에 V>로 된다. 따라서 제안된 문장패턴에 명사의미소성을 부여하고 용언이 가질 수 있는 문장 패턴을 매칭시키면 의미가 잘못되었음을 알 수 있다.

(2) 동음 다의어의 해석이 가능하다.

그는 그 노래의 가사를 읽는다.

한국어에서는 동음다의어가 다수 나타나므로서 자연어 처리를 어렵게 한다. 위의 예문에서 <가사>란 단어도 여러 의미를 가지고 있는데 명사의 의미소성에 따라 분류하면 다음과 같다.

ㄱ. 가사(加賜) (행위) - 더 보태어 줌
ㄴ. 가사(家舍) (장소) - 집
ㄷ. 가사(假死) (행위) - 정신을 잃어 죽은 것과 같은 상태
ㄹ. 가사(歌詞) (구상체) - 노랫말 등 여러 형태가 있다.

위의 문장은 동사가 결합할 수 있는 패턴에 따르면 동사형 75형 <N(인간) 이 N(구상체) 을 V>과 같다.
따라서 표기된 <가사>의 의미는 노랫말을 뜻한다는 것을 알 수 있다.

(3) 애매한 문의 해석

문장의 표현 형식을 보면 주어와 술어의 관계가 불분명한 경우가 있다.

학생에게 테니스를 칠 것을 권하였다.

위의 문장은 구문적으로 (학생에게)는 (칠)과 (권하였다)의 양방향에 연결되는 것이 가능하다. 그러나 (권하였다)의 결합패턴을 보면

치다 : 동사형 75형으로 N(인간) 이 N(구상체) 을 V로 되고,
권하다 : 동사형 96형으로 N(인간) 이 N(구상체) 을 N(인간) 에게 V와 같이 되므로 위의 문장에서 (학생)은 (권하였다)에 연결되는 것을 알 수 있다.

3.2.4 기계번역용 문장패턴의 적용

앞에서 추출된 기계번역용 문장패턴은 말뭉치로부터 추출된 32형의 문장패턴에 명사의 의미소성을 부여하여 추출함으로써 문장패턴의 수가 213형으로 늘어났다. 문장패턴의 검색시 모든 문장패턴을 대상으로 할 경우 검색속도가 느려진다. 이 문제는 사전구축시 동사 사전에 각 단어

가 취할 수 있는 문장패턴의 번호를 기록하여 해당 패턴만 검색하므로써 해결할 수 있다.

4. 결론

한국어에서의 문장패턴에 대한 연구는 미국이나 일본에 비하여 뒤늦게 시작 되었으며 또한 뒤떨어져 있다. 또한 연구의 관점이 국어학적인 것이어서 기계번역을 위한 연구가 필요했다.

이를 위하여 기존의 문장패턴에 대하여 조사 분석한 후 말뭉치를 구축하였다. 말뭉치는 다양한 문장들의 모음으로서 언어에 대한 지식이나 정보를 포함하고 있다. 한국어의 문장패턴은 동사와 형용사 술어 자체의 어휘적 의미를 실현시키기 위해 문장을 구성하는데 없어서는 안될 필수적 성분과 생략이 가능한 수의적 성분으로 구성되는 특징이 있다. 이러한 특성을 기반으로 1단계, 2단계에 걸쳐 문장패턴을 추출하였다.

1단계에서는 동사형 19형, 형용사형 8형, 지정사형 5형의 32형이 추출되었다. 2단계에는 기계번역용 문장패턴을 추출하기 위하여 앞에서 추출된 32형 문장패턴에 명사의 의미소성을 부여하여 추출한 결과 동사형 153형, 형용사형 41형, 지정사형 19형이 추출되었다.

본 연구에서 추출된 문장패턴들은 한국어 문법의 기본 구조를 표현하고 있기 때문에 언어습득을 위해서는 물론이거니와 기계번역을 위하여 유용하다.

첫째, 구문분석시 애매성을 해소할 수 있으며,
둘째, 동음다의어의 해석이 가능하며,
세째, 의미의 부정합 판정이 가능하다.

앞으로의 연구는 복합문에 대한 문장패턴의 확장과 빈도수 사전의 구축 그리고 추출된 문장패턴을 실질적으로 기계 번역에 적용하는 연구등이 필요하다.

참 고 문 헌

[1] 정인승, 인문계 고등학교 표준문법, 1968, pp 8~11.

[2] 강은국, 조선어 문형 연구, 서광학술자료사, 1993.

[3] 이용주 · 구인환, 중학교 국어문법, 1967, pp 63~64.

[4] 한국국어교육연구회, 고등문국문법, 1966, pp 175~177.

[5] 최현배, 고등말본, 1955, pp 16~17.

[6] 강복수 · 유창균, 인문계 고등학교 문법, 1969, pp 119~120.

[7] 박인철 ,배우정, 안동언, 이용석, "술어 기반 문형 정보를 이용한 한국어의 의미 구조 생성에 관한 연구", 한국어정보과학회 학술발표논문집, 1995, pp. 43 ~49.

[8] 윤성희, "한글 자연어 질의어 처리를 위한 문장 분석 기법에 관한 연구", 1996.

[9] 정교환, "국어문형고", 국어국문학회, 국어국문학 15, 1982, pp 137~155.

[10] 김진한, "한국어 결합가 패턴에 의한 기계번역에 관한 연구", 청주대학교 석사학위논문, 1987.