한·영 기계번역을 위한 한국어 품사 분류

송재관, 박찬곤
청주대학교 전산정보공학과

Classification of Korean Parts-of-Speech for Korean-English Machine Translation

Jae-Gwan Song, Chan-Khon Park
Dept of Computer Science & Engineering, Chongju University.

요 약

본 논문에서는 한·영 기계번역을 위한 한국어 품사 분류를 한다. 한국어 표준문법에서 제시되는 품사 분류 기준은 의미, 기능, 형식의 세 가지 기준을 적용하고 있으며, 자연언어처리에서도 같은 분류 기준을 바탕으로 하고 있다. 품사 분류에 여러 가지 기준을 적용하는 것은 문법구조 이해 및 품사 분류를 어렵게 한다. 또한 한·영 기계번역시 품사의 불일치로 전처리가 필요하다. 이러한 문제를 해결하기 위하여 본 논문에서는 하나의 기준을 적용하여 품사 분류를 한다. 방법으로 한국어 표준문법에 의하여 말뭉치에 태깅하고 문제점을 찾아내며, 새로운 기준에 의하여 품사 분류를 한다. 본 논문에서 분류된 품사는 한국어 문장에서 통사적 역할이 동일하고, 영어에서의 사전 품사와 동일하다. 또한 품사 분류의 모호성을 제거하고, 한국어의 문장 구조를 명확히 표현하며, 한·영 기계번역시 패턴 매칭에 의한 목적언어 생성이 가능하다.

1. 서론

서구의 전통문법은 19세기 중엽부터 프랑스, 미국, 캐나다, 독일 등의 선교사들을 통해 우리말에 유입, 적용되었고, 서구문법을 받아들인 일본문법의 영향을 받으면서 국내학자들에 의하여 연구, 적용되었으며, 국어품사론 연구는 20세기초 국내학자들에에 의하여 본격적인 문법연구가 시작되면서부터 1963년 학교문법통일안이 이루어지기까지 문법연구의 주 대상이 되어왔다[1]

품사 분류의 유형은 단어의 개념 규정에 관계없이 단어를 문법적으로 분류한 것을 품사로 본다면 體辭(체언토, 용언토)를 모두 독립품사로 인정하는 체계, 허사 중 체언토(조사)만 독립품사로 인정하는 체계, 허사를 모두 독립품사로 인정하지 않은 체계로 분류하고 있다[2] 한국어 자연언어처리에서 품사 체계에 대한 많은 연구가 형태소 분석, 구문분석, 의미분석 등 각 단계에서 이루어져 왔다[3]. 각각의 단계에서 연구된 품사 체계는 국어학에서의 분류 기준을 바탕으로 하고 있으며, 각 단계마다 별도의 품사 체계를 가지고 있다.

본 논문에서는 한국어학에서의 품사 분류와 자연언어처리에서의 품사 분류를 알아보고 한·영 기계번역을 위한 품사 분류를 하며 그 유용성을 제시한다

2. 한국어의 품사

2.1. 품사의 정의

국어 초기문법에서 품사의 개념은 대부분 정확하게 드러나지 않고 단어와 동일시되었고 단어의 개념도 정확하게 드러나지 않았다. 安廓(1923)에서 처음으로 단어를 "문법적 성질의 근사함"에 따라 나눈 것이란 비교적 타당한 견해가 제시되고, 권영달(1941), 장하일(1947)을 거쳐 이희승(1955), 김민수(1956) 등에서 구체화된다[1]

金敏洙는 품사의 정의를 "單語는 品詞의 감(資料)이요, 單語의 職能(글월 속에서 행하는 임무)이 플러스된 것이 品詞다"[2]라고 하였고, 李熙昇은 "單語는 意味上의 分類的 單位요, 品詞는 單語를 統辭論的 범주에 의하여 分類한 것이다"[4]라고 하였으며, 남기심, 고영근[5]은 단어를 문법적 성질의 공통성에 따라 몇 갈래로 묶어 놓은 것이라고 하였다. 품사(品詞, parts of speech)는 단어를 기능, 형태, 의미 등이 같은 것끼리 분류한 것이다. 품사 분류는 언어 기술의 간편함을 기하여 필요한 일이다. 이것은 문법의 이해 또는 기술에 매우 유용하다. 수 많은 단어의 낱개에 대한 기술이 몇몇 어군에 대한 기술로 대치될 수 있기 때문이다[6]. 단어를 문법적 성질에 따라 몇 갈래로 나누어 이해하는 일은 한 언어의 문법 구조를 이해하는 데 큰 도움을 준다[7]. [렴종률]은 품사를 형태론적 표식과 문장론적 기능의 측면에서 공통점들을 가진 부류로서 분류하였으며 명사, 수사, 대명사, 동사, 형용사, 부사, 관형사, 감동사의 8품사로 분류하여 한국어 표준 문법에서 나타난 조사를 품사로 보지 않았다[8].

한국어 자연언어처리에서 품사에 대한 개념 및 분류는 한국어학의

학교 표준문법을 바탕으로 하고 있으며, 각각의 단계에서 필요에 따라 분류하고 있다. 품사라는 용어는 단독으로 쓰이기보다 품사 태그라는 용어로 주로 사용되어 왔고, 국어학에서 말하는 품사와는 많은 차이가 있다. 1996년 우리말 정보처리 규격 심포지움에 이르러 형태·통사 태그 집합이라는 용어로 바뀌었다.

2.2. 품사 분류 기준

품사 분류에 대한 기준은 언어관에 따라 달라지고 품사에 대한 개념 규정에 따라 달라질 수 있다. 전통 문법론에서는 의미, 형태, 기능의 세가지 기준을 적용하였고, 구조 문법론에서는 기능을 중시하였다[9].

최현배는 품사 분류의 표준을 "씨의 가름은 그 말본에서 구실(職能)을 주장으로 삼고, 그에 따르는 꼴(形式)과 뜻(意義)을 딸림(從)으로 삼아 결정하여야 한다"[10]고 하여 품사 분류의 기준을 처음으로 제시하였고, 金亨奎는 "(1) 職能, (2) 形態, (3) 語義의 순위로 비중을 두어 분류해야 한다"[11]고 하였다.

현재 학교 표준문법에서 취하고 있는 품사 분류 기준은 전통 문법론에 따라 기능, 형태, 의미의 세 가지 기준을 적용하고 있으며 품사 분류를 어렵게 하고 있다.

자연언어처리에서 품사 분류는 기존 국문법상의 품사 체계를 기본 바탕으로 하고 각 처리 단계에서의 필요에 따라 세분화하고 있으며, 품사를 어떤 부분에 응용할 것인가에 따라 다르게 분류하고 있다.

1996년 우리말 정보처리 심포지움에서 형태·통사 태그 집합을 위한 표준을 제시하였으며, 형태론적 분류체계에 근거하여 9가지의 상위 분류와 54개의 형태·통사 태그 집합으로 표준화하였다. 그러나 학교 문법을 그대로 반영하고 있으므로해서 여러 가지 기준에 따라 품사를 분류해야 하는 문제를 그대로 안고 있다.

품사 분류시 일반적으로 고려할 사항으로 품사 분류 단위, 품사 세분화, 학습방법 등이 있다[12].

본 논문에서는 記述文法(descriptive grammar)을 이용하여 한·영 기계번역을 위한 품사를 분류한다. 기술문법은 記述的(descriptive)인 방법으로 쓰여진 문법으로 기술문법이 목표로 하는 것은 주어진 언어자료를 처음부터 끝까지 일관성(consistent) 있고 기능한 한 완전하고(exhaustive) 가능한 한 간결하게 기술하는 것이다[9].

3. 한.영 기계번역을 위한 품사 분류

3.1. 한국어 표준 문법

품사의 개념 규정에 있어서 단어를 그 문법적 성질에 입각, 그 공통점에 의하여 몇 가지의 유형으로 분류한 것을 품사로 보는 것이 문법 기술상 편리하다. 따라서, 품사는 단어의 문법적 성질의 별칭으로 본다면, 언어의 분절적 단위면으로는 품사와 단어는 동일한 것이라 할 수 있다[13]. 그러므로, 단어를 어떻게 규정짓느냐에 따라 품사의 분류가 달라질 수 있다. 단어에 대한 견해로는 용언토는 단어에서 제외하고 체언토는 단어로 보는 전통문법적 견해[10,14,15]와 용언토·체언토 모두를 품사에서 제외하는 구조문법적 견해[16,17,18]가 있다.

한국어 표준문법에서는 전통문법의 방법론을 따르고 있으며, 본 논문에서는 구조문법의 방법론을 바탕으로 하여 한·영 기계번역에 유용한 품사를 분류한다.

본 논문에서 사용된 한국어 문장은 한국어 표준문법에 따라 태깅된 말뭉치[19]로부터 단문을 대상으로 하였고, 보조용언·관용어·연어는 연구 범위에서 제외하였다. 품사 태깅에 사용된 기호는 1997년 6월 우리말 정보처리 규격 심포지움에서 제시된 국어 형태 통사 태그의 표준안의 기호 중 9품사에 대한 분류 기호를 적용하였다.

한국어 표준문법에 따른 품사 분류는 말뭉치에서 이래와 같은 문제점을 가지고 있다.

첫째, 품사 분류 기준이 명확하지 않고 모호하다. 국어 표준문법에서 명사는 사물의 이름을 나타내는 단어, 대명사는 사람이나 사물, 장소의 이름을 대신하여 가리키는 단어, 동사는 상태나 성질을 나타내는 단어, 형용사는 상태나 성질을 나타내는 단어, 감탄사는 말하는 사람의 부름, 느낌, 놀람이나 대답을 나타내는 단어로 정의하고 있지만, 이러한 품사 분류 기준은 분류를 어렵게 한다.

① 공부, 건축

② 높이, 학교, 건물

①의 경우 동적인 의미를 지니고 있는데 불구하고 명사로 분류되고 있으며, ②는 상태의 의미를 지니고 있지만 명사로 분류되고 있다.

둘째, 품사 분류의 대상이 되는 단어는 문장구조 분석의 기본 단위이므로 문장 속에서 통사적 기능이 분류의 전제가 되어야 한다. 그러나 품사간 통사적 기능이 구별되지 않는다.

③ 먹는, 먹기로, 먹다

④ 아름다운, 아름답게, 아름답다

③은 동사이고, ④는 형용사로 서로 다른 품사이지만 관형어, 부사어, 서술어 여러 면에서 쓰이고 있다. 학교 문법에서는 동사와 형용사의 품사 설정에 있어서 기능보다는 어간의 의미에 따라 분류를 하기 때문이다. 이러한 분류 기준은 자연언어처리에서 품사 모호성을 증가시킨다.

3.2. 한.영 기계번역을 위한 품사

본 논문에서는 여러 기준에 따른 품사 분류의 모호성을 해결하기 위하여 유어(class word)와 기능어(function word)[9]의 두 가지 품사 분류법에 따라 품사를 분류하였다. 이 분류법은 하나의 단어가 문장 속에서 차지하는 문법적 기능은 하나라는 것을 전제로 하고 있다.

본 논문에서 한·영 기계번역을 위한 품사 태그 집합은 명사·대명사(n), 전치사(p), 부사(adv), 형용사(a), 동사(v), 감탄사(i),접속사(c)이며, 아래에서는 각 품사별 기능 및 말뭉치에 나타나는 문장을 예를 들었다.

3.2.1. 유어(class word)

(1) 명사

문장에서 주어, 목적어, 보어로서의 기능을 갖는다.

(2) 대명사

명사와 마찬가지로 문장에서 주어, 목적어, 보어의 기능을 한다.

(3) 부사

문장에서 다른 품사를 수식하거나 강조하는 기능을 하며, 국어학에서의 접속 부사는 본 논문에서는 접속사이다.

(4) 동사

문장에서 주어에 대하여 서술어의 역할을 한다.

(5) 형용사

문장에서 명사나 대명사를 수식하거나 설명하는 기능을 한다.

3.2.2. 기능어(functional word)

(1) 전치사

문장에서 명사, 대명사 뒤에 붙어 문법적인 관계를 나타낸다

(2) 감탄사

국어 표준 문법을 따른다

(3) 접속사

문장에서 각 부분을 연결하는 연결사로, 국어학의 접속 조사가 이에 해당한다.

아래의 예문들은 본 논문에서 분류된 품사 기준에 의해 태깅된 한국어 문장의 유형들로, 대명사는 명사와 문법적 기능이 동일하여 명사로 태깅하였다. 예문의 첫 줄은 학교 표준문법과 본 논문의 분류 기준에 따른 품사가 함께 태깅된 형태로 영문으로 바뀌었을 때의 패턴과 비교하여 기술하였다

⑤ 한_d_a 예쁜_a_a 소녀_n_n가_p 조그마한_a_a
마을_n_n에서_p_p 살았다_v
A pretty girl lived in a small village.
=> 국어문법: d+a+n+p / a-n+p / v
=> 본 논문: a+a+n+**p** / a-n+p / v
=> 영 문: a+a+n / v / p+a+a+n

⑥ 모든_d_a 발명_n_n과_p_c 발견_n_n은_p_p
힌_d_a 남자_n_n나_p_c 여자_n_n의_p_p
단순한_a_a 호기심_n_n에서_p_p 나왔다_v_v
Every invention and discovery resulted from the simple curiosity of one man or woman
=> 국어문법 : d+n+p+n+p / d+n-p+n+p / a+n-p / v
=> 본 논문 a+n+c+n+**p** / a+n+c+n-p / a+n+p / v
=> 영 문 a-n+c+n / v / p+a-a+n / p+a+n+c+n

⑦ 나_pn_n는_p_p 착한_a_a 소년_n_n이다_p_v
I am a good boy
=> 국어문법 : pn+p / a+n-p
=> 본 논문 : n+**p** / a+n / v
=> 영 문 n / v / a-n

⑧ (대부분_n의_p)_a 사람들_n_n은_p_p 실용적_d_a
이유_n_n로_p_p 영어_n_n를_p_p 공부한다_v_v
Most people study English for practical reason.
=> 국어문법 n+p / n+p / d+n+p / n+p / v
=> 본 논문 : a+n+**p** / a+n+p / n+p / v
=> 영 문 : a+n / v / n / p+a+n

위에 나타난 것처럼 본 논문에서 분류하고 있는 품사 분류는 영문으로 바꾸었을 때의 품사 패턴과 대부분 일치하고 있다 예문 ⑤~⑧에서 강조 표시된 "+p"부분이 영문패턴에서는 사라지고, 한국어 문장 패턴에서는 나타나지 않는 "a"와 같은 관사가 영문패턴에서 나타난다. 한국어 문장의 품사 패턴에서 불필요한 진치사를 제거하고 영어 문법에서 필요로 하는 관사에 대한 정보를 추가한다면 영어 문장 패턴으로의 변환이 가능하다

4. 결론

본 논문에서는 단어의 통사적 역할을 기준으로 하여 한국어 품사를 명사, 대명사, 전치사, 부사, 형용사, 부사, 동사, 접속사, 감탄사의 8품사로 분류하였다. 학교 표준문법에서 제시하고 있는 품사중 일부가 본 논문에서는 다른 품사로 바뀌어 나타나고 있다. 관형사는 명사를 한정하는 단어로서 형용사로 바뀌고, 조사는 전치사로 바뀌고, 부사중 접속부사는 접속사로 바뀌었고, 수사가 나타나지 않는다

본 논문에서 제시된 품사 분류는 하나의 기준에 의하여 정해지기 때문에 한국어 학교 표준 문법에서 나타나는 품사 분류의 애매성을 제거하며, 영어에서 사전의 품사와 일치하며, 또한 문장의 문법적 관계를 정확히 나타내므로써 한국어 문장의 구조를 이해하기가 쉽고, 품사 패턴에 명사의 의미소성을 가미한다면 한·영 기계번역에서 패턴에 의한 목적언어의 생성에 유용하다.

앞으로의 연구 과제로는 연어, 관용어, 보조용언 그리고 연결어미 등에 대한 품사 분류를 두고 있다

참고문헌

[1] 이광정, 국어품사분류의 역사적 발전에 관한 연구, 한신문화사, pp 2-5, 1991

[2] 金敏洙, 國語文法論研究, 通文館, p. 229. 1960

[3] 안미정, 김제한, 옥철영, "한국어 처리를 위한 품사 체계 연구," 제5회 한글 및 한국어 정보처리 학술발표 논문집, p 582, 1993.

[4] 李熙昇, 國語學概論, 民衆書館, p. 197-204, 1955.

[5] 남기심, 고영근, 표준 국어 문법론, p 54, 1988,

[6] 이주행, 현대국어문법론, p. 75, 1992

[7] 李翊燮, 張素鉉, 국어학개론, 한국방송통신대학출판부, p. 72, 1996.

[8] 렴종률, 조선어문법사, 한국문화사, pp 52-68, 1980

[9] 이정민, 배영남, 언어학 사전, 박영사, p. 651, 1993

[10] 최현배, 우리말본, 1955.

[11] 金亨奎, 國語學概論, 1968

[12] 이상주, 임희석, 임해창, "은닉 마르코프 모델을 이용한 두 단계 한국어 품사 태깅," 제6회 한글 및 한국어 정보처리 학술 대회 발표 논문집, pp. 305-312, 1994

[13] 李言鹿, 國語文法研究, 日新社, p. 71, 1991

[14] 李崇寧, 고등국어문법, 乙酉文化社, p. 35, 1965

[15] 橋本進吉, 國語法研究, 岩波書店, pp 11-45, 1968

[16] L. Bloomfield, Lauguage, p. 178. 1933

[17] Bloch & Trager, Outline of Linguistic Analysis, p 54 1642

[18] 張河 -, "낱말의 定義," 李熙昇先生 頌壽記念論叢, 一潮閣, p 614, 1958

[19] 송재관, 홍성웅, 박찬곤, "기계번역을 위한 한국어 문장패턴에 관한 연구," 제8회 한글 및 한국어 한국어 정보처리 발표 논문집, pp. 308-312, 1996