한국어의 기계번역을 위한 용언 구조의 해석

한 공 록° 이 주 근
인하대학교 전자공학과

An Analysis of Korean inflected Word for Machine Translation

H. R. Han J. K. Lee
Dept. of Electronic Eng., Inha Univ.

Abstract

This paper proposes a method for analyzing the Korean inflected word in machine translation system. We define the processing rules which are useful of analyzing an irregular conjugation , pesent an parsing algorithm of noun and *specifed verb and reduce the space of dictionary* by the algorithm.

I.서론

용언 (동사, 형용사)은 한국어의 핵심 요소로서 기계번역 시스템의 설계에서 차지하는 비중이 크며 이들의 처리 방식에 따라 시스템 전반에 강한 영향을 미치게 된다[1, 2].

특히 한국어는 언어 분류학상 첨가어로서 [3] 용언의 어간에 어미가 결합하여 많은 활용을 하는 어휘적인 특징과 문장 상에서는용언이 명사와 보어들을 지배하는 역할을 한다[4]. 따라서 기계 번역은 물론 각종 한국어의 이해 시스템을 구성할 때에 용언의 구조 해석은 중요하다[5].

용언이 활용되는 기본형은 그림1과 같이 어간에 어미가 결합되는 형태이다.

VP = STM + SUFi
STM : 어간
SUFi : 어미 ($1 \leq i \leq n$)

그림1. 용언의 기본적 구성형 .

그러나 번역의 대상이 되는 한국어의 용언이 반드시 규칙적인 것만이 입력되는 것이 아니고 불규칙적으로 변형된 용언과 축약 또는 다른 품사에서 전성된 것들이 입력되는 경우가 많다.

그러므로 본 논문에서는 먼저 용언을 해석하기 위하여 규칙적인 어간과 어미의 분석을 논하고, 불규칙 용언의 해석과 축약어 처리 그리고 명사가 전성되어 동사와 형용사적으로 사용되는 것을 형태적으로 분류하여 처리하는 방법을 논한다.

II. 한국어 용언의 구조 해석

1. 용언의 기본 구조

한국어에서 용언이 될 수 있는 최소 단위는 앞 절에서 언급한 바와 같이 "STM + SUF"이다. 이 최소의 형태를 기본으로 하여 그외의 부가적인 요소들인 명사와 부사, 보조 어간등이 결합하여 표1과 같이 모든 한국어 용언을 만든다.

표1. 형태에 의한 용언의 구조 형식

	용언의 형식	출현빈도(%)
1)	STM + SUF	68.32
2)	STM + AUXSTM + SUF	11.7
3)	N + STM + SUF	16.24
4)	N + STM + AUXSTM + SUF	2.72
5)	AD + STM + SUF	0.93
6)	AD + STM + AUXSTM + SUF	0.08

AD : 부사, N : 명사, STM : 어간,
AUXSTM : 보조어간, SUF : 어미

표1은 1981년도에 발행된 국정교과서 (1 - 6학년)의 2만 단어와 1983 년도의 중학교 국어 교과서와 기타 사전에서 추출한 약 2만 단어인 총 4만 단어 중에서 용언을 제취하여 구조 형식을 유도한 것이고 위의 5 형태를 통합하면 다음과 같이 형식화 된다.

VP = (N) : (AD) : + STM + (AUXSTM) + SUF
(단, ()는 생략 가능, :은 선택을 나타낸다.)

표1의 형식 중에서 기본이 되는 것은 1)과 2)의 형태인데 STM의 경우에도 규칙적인 어간과 불규칙적인 어간이 있기 때문에 불규칙 어간을 사전에 있는 형태로 분석해 내는 과정이 중요하고 3)과 4)는 명사에 동사나 형용사를 만드는 문법적 요소가 결합하여 용언화된 것으로 이들 형태를 그대로 용언 사전에 기록하면 2 중의 부담이 되므로 명사(N)과 STM을 분리하여 처리하는 과정이필요하다. 또한 5)와 6)은 사용빈도가 작기 때문에 결합된 형태를 그대로 사전에 수록하여 분석을 용이하게 한다.

2. 불규칙 용언의 해석 규칙

사전에는 규칙 용언의 어간만이 수록되어 있기 때문에 입력된 불규칙 용언으로 부터 원래의 어간과 어미를 추출하기 위한 규칙을 정해야 한다.

한국어 용언의 변형이 발생되는 형태는 그림2와 같다[6].

- 어간이 변한 것 — ㄷ, ㄹ, ㅅ, ㅎ, 으 우 불규칙
- 어미가 변한 것 — 러, 어 불규칙
- 어간 어미가 모두 변한 것 — ㅂ, 르 불 규칙

그림 2. 용언의 불규칙 변형 형태

그림2의 변형된 성질과 탈락, 삽입된 요소들을 조합하여 원래의 어간을 찾아내는 규칙을 표2와 같이 정한다.

표2 불규칙 용언의 변환 규칙

규칙	기능	해당불규칙
+	탈락된 어간의 최종 음소를 속성값으로 하여 변형된 어간에 더한다.	ㄹ, ㅅ, 으 우, ㅎ
-	변형된 음소를 탈락시키고 첨가될 음소를 속성값으로 하여 더한다.	ㄷ, 르 러, ㅂ
0	"하"로 끝나는 어간 다음에 후속되는 "어-"로시작하는 어미를 "어-"로 바꾼다.	여

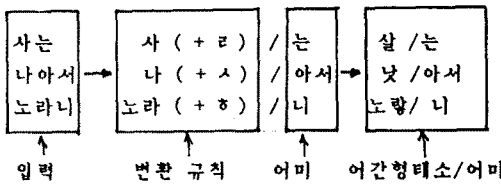
변형어간 (RULE ATTRIBUTE)

그림3. 변형어간의 사전 형태

따라서 불규칙어간의 지식 표현은 그림3과 같이 변형된 어간과 적용 규칙 그리고 첨가될 속성값들을 기록한다. 이들 규칙과 속성값으로서 원래의 어간을 도출한다.

이 규칙의 적용 예는 그림4와 같다.

1) + 규칙



2) - 규칙

깨달은 ⟶ 깨달(- ㄷ)/ 은 ⟶ 깨닫 / 은
흘러 ⟶ 흘ㄹ (- _)/ 어 ⟶ 흐르 / 어

3) 0 규칙

하여서 ⟶ 하 (0 아) / 여서 ⟶ 하 / 아서

그림4 불규칙 용언의 변환 규칙 적용 예

3. 축약의 처리

불규칙 활용 이외에도 모음이 축약되어 규칙과 불규칙 용언에 복합된 형태가 있다. 이때에 축약된 용언의 형태로 부터 원래의 어간과 어미를 생성해 내야한다. 한국어의 모음 축약은 5 형식이 있다[7]. 이 5 형식을 table로 하여 그림4와 같이 입력된 용언과 비교하여 치환한다.

왔다 ⟶ 오 + 았다 (ㅘ → ㅗ + 아)
가졌다 ⟶ 가지+ 었다 (ㅣ →]+ 어)
했다 ⟶ 하 + 였다 (ㅐ → }+ 여)
⟶ 하 + 었다 (0 규칙)

그림5 축약처리의 예

4. 명사(N) + 어간(STM)의 해석

1 절에서 논한 바와 같이 명사에도 어미가 결합하여 활용하는 경우가 있다. 이 때 명사와 어미의 사이에는 반드시 동사나 형용사의 역할을 하는 어간(STM)이 매개적으로 사용된다. 이 STM은 여러가지가 있으나 "명사+하다"의 결합이 전체 용언의 20%를 차지하므로 그 영향이 크다. 따라서 "명사+하다"를 하나의 어간으로 묶어서 사전에 포함시키면 처리는 용이하나 그 수가 방대해지므로 사전 용량에 부담을 준다. 그러므로 그림6과 같이 "하다"라는 어간화 문법소와 결합할 수 있는 명사에

MS의 의미소성을 주어서 명사에 기록하고 "명사"와 "하다"에서 "하다"를 분리하면 사전의 2 중 부담을 줄일 수 있다.

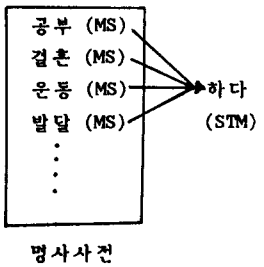


그림6 명사와 "하다"의 분리 처리

그러나 "명사+하다"를 분리할 경우, 명사는 단지 명사적인 기능만을 갖지고 있고 구문 및 의미의 역할은 "하다"라는 동사가 담당하게 되므로 원래의 "명사+ 하다"가 문장에서 지배하는 의미적인 역할을 ViQj라고 할때 이 의미적 역할을 상실하고 "하다"의 고유의 의미인 Vi'Qj'라는 새로운 의미가 파생되므로 구문, 의미해석에서 그림7과 같은 문제가 발생한다[8,9].

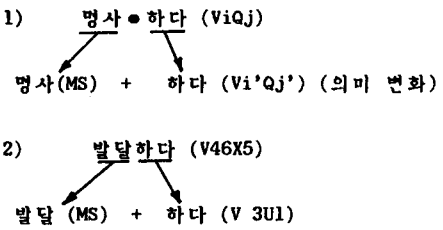


그림7 분리처리에 의한 의미 변화

이것을 해결하기 위하여 명사에 "하다"가 결합된 원래의 동사와 형용사의 의미를 명사에 기록해두고, 그 명사가 "하다"와 결합할 때는 그림8과 같이 "하다"의 의미 기호를 무시하고 명사에 부가된 의미 정보를 이용하여 문제점을 해결한다.

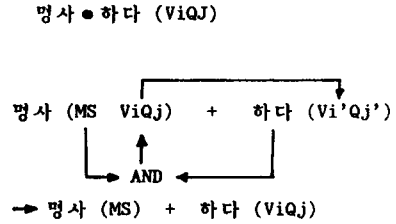


그림8 "명사●하다"의 의미 환원

III. 실험결과

```
KOREAN=>불렀다.$

===== morphological analysis =====
( BUREU: V/ ((V1W2 68) (V1U5 68)))
( EOSSDA: SUF / (EOSS T3) R1 F1)

-----------------------------------
KOREAN=>도왔다.$

===== morphological analysis =====
( DOB: V/ ((V25W2 272)))
( ASSDA: SUF / (ASS T3) R1 F1)

-----------------------------------
KOREAN=>도울.$

===== morphological analysis =====
( DOB: V/ ((V25W2 272)))
( R: SUF / (R T2) R1 F9)

-----------------------------------
KOREAN=>퍼서.$

===== morphological analysis =====
( PU: V/ ((V23W2 274)))
( EOSEO: SUF / T1 R1 F7)

-----------------------------------
KOREAN=>먹었다.$

===== morphological analysis =====
( MEOGI: V/ ((V6W2 74) (V8S4 74)))
( EOSSDA: SUF / (EOSS T3) R1 F1)

-----------------------------------
KOREAN=>하였다.$

===== morphological analysis =====
( HA: V/ ((V23U1 62)))
( EOSSDA: SUF / (EOSS T3) R1 F1)

-----------------------------------
KOREAN=>발달했다.$

===== morphological analysis =====
( BARDAR: N (MS V64X5))
( HA: V/ ((V23U1 62)))
( EOSSDA: SUF / (EOSS T3) R1 F1)

-----------------------------------
KOREAN=>압니다.$

===== morphological analysis =====
( AR: V/ ((V7Y5 75)))
( BNIDA: SUF / T1 (B R2) F1)
```

IV. 결론

본 논문에서는 한국어의 기계 번역을 위한 용언의 해석 방법을 논했다. 입력된 용언의 어간과 어미를 분석할때 변형된 부분들을 효과적으로 처리하기 위하여 불규칙의 해석 규칙을 정하고, 사전의 부담을 줄이기 위하여 명사와 "하다"를 분리 처리하는 algorithm을 제안했다. 이것을 한●일 양방향 기계 번역 시스템에 적용하여 한국어 해석을 간결하고 효율적으로 처리하는 것을 확인했다.

참 고 문 헌

1. S.H. LEE, K.R. HAN, J.K. LEE, "A BIDIRECTIONAL MT SYSTEM BETWEEN KOREAN AND JAPANESE BASED ON A PATTERN NET", Proc. 10th Symposium on Information Theory and its Application, Japan, Nov., 1987

2. S.K. HAN, S. H. LEE, J. K. LEE, "A KOREAN-ENGLISH MACHINE TRANSLATION SYSTEM BASED ON LEXICAL ASSOCIATION GRAMMAR", Proc. TENCON 87-IEEE, Aug., 1987

3. 서병국, "활용론 연구," 탑출판사, 1985.

4. 이 주근, 한 광록, "의미 frame에 의한 기계번역", 전자공학회 학술논문집, Vol.9, No.1, 1986, 6.

5. J.K. LEE, J.H. LEE, K.R. HAN, "Determination of Modificatory Scope and Inference for Korean", Proc. TENCON 87-IEEE, Aug., 1987

6. 서병국, "한국 교육 자료 총서," 형설 출판사, 1982.

7. 이 주근, 윤 면기, "자연어의 형태소 분석", 전자공학회 학술 논문집, Vol.17, No.1, 1984, 7.

8. B. Bruce , "Case System for Natural Language ," Artificial Intelligence, vol. 6,1975.

9. Kim, Nam-il, "Verb Phrase Complements in Korean", Linguistics in Morning Calm, SICOL-81, 1981