한국어 형태소 분석을 위한 단어 유형 분류와 자료구조

강 승 식
한성대학교 정보전산학부
Email: kang@{ham, ice}.hansung.ac.kr

Word Classification and Data Structure for Korean Morphological Analysis

Kang, Seung-Shik
School of Information and Computer Engineering, Hansung University

요 약

한국어 정보처리 시스템은 유형별로 다양한 형태의 형태소 분석 정보를 필요로 하는데 이를 위하여 한국어의 단어 유형을 분류하고 형태소 분석 결과를 효율적으로 저장하는 자료구조를 제안한다. 형태소 분석에 필요한 단어 유형은 일반적인 유형과 단순화된 유형으로 구분하여 비교하였으며, 이를 기반으로 형태소 분석을 위한 새로운 단어 구성 전이도를 제시하였다. 형태소 분석 결과를 저장하는 자료구조는 HAM에서 사용되고 있는 자료구조를 기반으로 응용 시스템에서 필요로 하는 정보를 쉽게 사용할 수 있도록 보완하고 저장 공간의 효율성을 개선하였다.

1. 서 론

자연어 처리 시스템에서 형태소 분석기는 필수적으로 요구되고 있으며 그 응용 분야가 매우 넓다. 한국어의 경우 자동색인을 비롯한 일부 응용 분야에서 상용화되어 있으며, 한국어 정보처리와 관련된 모든 분야에서 응용 시스템을 구축하는데 가장 기본적인 모듈로서 사용된다. 즉, 모든 한국어 정보처리 시스템은 형태소 분석 단계를 거쳐야 하기 때문에 형태소 분석 결과를 먼저 파악하고 이를 기반으로 응용 시스템을 구축한다. 그런데 형태소 분석 결과에서 정의되는 단어 유형 및 품사 유형의 종류, 분석 결과를 저장하는 구조 등은 형태소 분석기마다 조금씩 다르기 때문에 응용 시스템의 구축은 형태소 분석 결과에 의존하게 된다.

이러한 문제점을 해결하고 한국어 정보처리 분야의 체계적인 연구를 위해 전자 사전 구축이나 형태소 분석 등 기초 기술 분야의 규격을 통일할 필요성이 제기되었다. 그 일환으로 한국과학기술원 인공지능연구센터의 주관하에 국어 정보베이스 축적과 국어 정보베이스 가공 규격에 관한 우리말 정보처리 규격 심포지움에서 이에 관한 논의가 있었다[1]. 이 심포지움에서 형태소 분석 관련 분야로 형태·통사 태그를 정의하고 있다. 형태·통사 태그는 형태소 분석 결과를 태깅 시스템이나 구문 분석에서 활용하는 관점에서 유용하게 활용될 수 있을 것이다.

형태소 분석기는 응용 분야에 따라 어형 인식 단계의 분석 결과만을 활용할 수도 있고 어형 확장 단계의 분석 결과를 요구하는 경우도 있으므로 두 단계 분석 과정으로 구성되는 것이 바람직하다[2]. 우리말 정보처리 규격 심포지움에서 제시된 형태·통사 태그는 형태소 분석의 어형 확장 단계의 분석 결과에 대한 태그 집합을 제안한 것이다. 이와 더불어 단어의 유형 분류와 단어 구성 전이도 등 형태소 분석에 필수적인 기본적인 부분과 어형 인식 단계의 형태소 분석 결과에 대한 논의가 필요하다.

현재 한국어 형태소 분석기는 주로 어형 인식 단계의 결과를 응용 시스템에서 사용하고 있는데 어형 인식 과정에서 사용하고 있는 한국어의 단어 유형과 단어 구성 전이도는 형태소 분석기마다 독자적으로 정의하고 있다. 따라서 한국어의 형태론적 특성과 관련된 기본적인 부분을 전산언어학의 관점에서 정의해야 할 필요가 있다. 본 논문은 한국어 형태소 분석기 HAM(Hangul Analysis Module) version 2.0[1]을 중심으로 어형 인식 단계의 형태소 분석 결과를 응용 시스템에서 활용하는데 필요한 단어 유형과 단어 구성 전이도, 분석 결과를 저장하는 자료구조 등을 제안한다.

---

1) HAM은 한글 분석 모듈로서 형태소 분석기 CORAN(Korean morphological analyzer)를 기반으로 자동 색인 기능과 철자 검사 및 교정 기능이 구현되어 있으며, 모호성 해결 기능과 구문 분석 등 한국어 분석에 필요한 기능으로 확장될 예정이다. HAM version 2.0은 ham.hansung.ac.kr에 연구 및 실험용으로 사용할 수 있도록 library가 공개되어 있으며, anonymous ftp에 의하여 down받을 수 있다. HAM 1.5는 1995년도에 Hitel과 천리안 등 PC 통신 자료실에 DOS용 실행 파일이 공개된 바 있다.

1. PTN_AID 단일어(ADV/INT/DET)
2. PTN_N 체언(N/PN/NM/XN/CN/UN/AN/HJ/ET)

3. PTN_NJ 체언 + 조사
4. PTN_NS 체언 + 접미사
5. PTN_NSJ 체언 + 접미사 + 조사

6. PTN_NSM 체언 + 용언접미사 + 어미
7. PTN_NSFM 체언 + 용언접미사 + 선어미 + 어미
8. PTN_NSMJ 체언 + 용언접미사 + 'ㅁ/기' + 조사
9. PTN_NSFMJ 체언 + 용언접미사 + 선어미 + 'ㅁ/기' + 조사

10. PTN_NCM 체언 + '이' + 어미
11. PTN_NCFM 체언 + '이' + 선어미 + 어미
12. PTN_NCMJ 체언 + '이' + 'ㅁ/기' + 조사
13. PTN_NCFMJ 체언 + '이' + 선어미 + 'ㅁ/기' + 조사

14. PTN_NSCM 체언 + 접미사 + '이' + 어미
15. PTN_NSCFM 체언 + 접미사 + '이' + 선어미 + 어미
16. PTN_NSCMJ 체언 + 접미사 + '이' + 'ㅁ/기' + 조사
17. PTN_NSCFMJ 체언 + 접미사 + '이' + 선어미 + 'ㅁ/기' + 조사

18. PTN_NSMXM 체언 + 용언접미사 + '아/어' + 보조용언 + 어미
19. PTN_NSMXFM 체언 + 용언접미사 +'아/어'+ 보조용언 + 선어미 + 어미

20. PTN_NJCM 체언 + '에서/부터/에서부터' + '이' + 어미
21. PTN_NJCFM 체언 + '에서/부터/에서부터' + '이' + 선어미 + 어미

22. PTN_VM 용언 + 어미
23. PTN_VFM 용언 + 선어미 + 어미
24. PTN_VMJ 용언 + 'ㅁ/기' + 조사
25. PTN_VFMJ 용언 + 선어미 + 'ㅁ/기' + 조사
26. PTN_VMCM 용언 + 'ㅁ/기' + '이' + 어미
27. PTN_VMCFM 용언 + 'ㅁ/기' + '이' + 선어미 + 어미

28. PTN_VMXM 용언 + '아/어' + 보조용언 + 어미
29. PTN_VMXFM 용언 + '아/어' + 보조용언 + 선어미 + 어미
30. PTN_VMXMJ 용언 + '아/어' + 보조용언 + 'ㅁ/기' + 조사
31. PTN_VMXMJ 용언 + '아/어' + 보조용언 + 선어미 + 'ㅁ/기' + 조사

32. PTN_ADVJ 부사 + 조사 /* "빨리도" */

33. PTN_CNOUNJ 복합명사 + 조사
34. PTN_CNOUNSJ 복합명사 + 접미사 + 조사
35. PTN_CNOUNSM 복합명사 + 용언접미사 + 어미
36. PTN_CNOUNCM 복합명사 + '이' + 어미
37. PTN_CNOUNSCM 복합명사 + 용언접미사 + '이' + 어미
38. PTN_CNOUNSFM 복합명사 + 용언접미사 + 선어미 + 어미

39. PTN_ASC_J Ascii + 조사
40. PTN_ASC_CM Ascii + '이' + 어미
41. PTN_ASC_CMJ Ascii + '이' + 'ㅁ/기' + 조사
42. PTN_ASC_CFM Ascii + '이' + 선어미 + 'ㅁ/기'
43. PTN_ASC_CFMJ Ascii + '이' + 선어미 + 'ㅁ/기' + 조사

44. PTN_ASC_SM Ascii + 용언접미사 + 어미
45. PTN_ASC_SMJ Ascii + 용언접미사 + 'ㅁ/기' + 조사
46. PTN_ASC_SFM Ascii + 용언접미사 + 선어미 + 어미
47. PTN_ASC_SFMJ Ascii + 용언접미사 + 선어미 + 'ㅁ/기' + 조사

48. PTN_NVM 체언 + 동사 + 어미
49. PTN_NVFM 체언 + 동사 + 선어미 + 어미

50. PTN_NK_J 미등록명사 + 조사
51. PTN_NK_S 미등록명사 + 접미사
52. PTN_NK_SJ 미등록명사 + 접미사 + 조사

그림 1. 한국어 단어의 유형 분류

2. 단어의 유형 분류

영어와 같은 굴절어는 단어의 구성 원리가 단순하고 굴절 현상이 다양한데 비해 한국어와 같은 교착어는 한 단어를 구성하는 형태소의 수가 여러 개이므로 단어의 유형이 다양하게 나타난다. 단어의 유형을 분류하는 이유는 형태소 분리 과정을 체계적으로 구현하기 위한 목적과 분리된 형태소를 저장하여 구문 분석 등 형태소 분석 결과를 전달하기 위한 두 가지 목적이 있다. 그런데 형태소 분리 과정은 형태소 분석기 내부에서 일어나는 과정이므로 중요치 않으며, 분석 결과를 응용 시스템에 쉽게 전달하는 목적을 만족시킬 수 있도록 단어의 유형을 분류하여야 한다.

강승식(1993)은 32가지의 단어 유형을 제시하고 있으나 모든 유형을 나열하고 있지는 않으며, HAM에서는 이를 확장하여 그림 1과 같이 52가지 유형을 사용하고 있다[3]. 그림 1에서 조사와 어미는 복합조사와 복합어미를 포함한 것이며, 특히 어미뒤에 조사가 결합된 유형도 어미로 간주하였다. 어미에 명사형 어미 'ㅁ/기'도 포함되어 있으나 명사형 어미만 따로 지칭할 때는 어미라는 용어 대신에 'ㅁ/기'를 사용하였다. 용언접미사는 '하다/되다/시키다'이다. 접두사가 결합된 체언과 용언은 접두사를 따로 분리하지 않았다.

그림 1은 모든 한국어의 모든 단어 유형을 포괄하고 있지는 않으며 서로 유사한 유형들이 중복되어 있다. 실제로 HAM에서도 52가지 유형을 모두 사용하지는 않는다. 유형 6~9와 유형 10~13의 예를 보면 두 그룹의 차이는 단지 용언화 접미사와 서술격 조사 '이'의 차이만 있을 뿐이다. 그런데 서술격 조사를 용언화 접미사로 간주하더라도 문제가 발생하지 않기 때문에 HAM에서 유형 10~13은 사용되지 않고 있다. 유형 14~17도 접미사가 결합된 것만 제외하면 유형 6~9와 유사하며, 복합명사와 관련된 유형 33~38과 영문자 및 숫자가 포함된 유형 39~47은 체언을 기술하는 유형 3~9로 대치 가능하다.

그림 1의 유형 분류는 복합명사와 미등록어, alphanumeric 문자가 포함된 것 등을 별개의 유형으로 분류함으로써 형태소 분리 과정의 용이성을 추구한 것으로 체언과 관련하여 중복된 단어 유형을 설정하였기 때문에 일관성이 결여되어 있다. 즉, 단일어 유형에서 체언을 명사(N), 대명사(PN), 수사(NM), 의존명사(XN), 복합명사(CN), 미등록어(UN), alphanumeric(AN), 한자(HJ), 기타(ET) 등을 하나의 유형으로 처리한 것과 문법형태소가 결합된 유형에서 이를 다른 유형으로 간주하는 것은 옳지 않다.

형태소 분석을 위한 단어의 유형 분류는 가능한 한 단순화시키는 것이 좋다. 왜냐 하면 구문분석기 등 응용 시스템에서 형태소 분석 결과를 사용할 때 필요한 정보만 쉽게 사용할 수 있어야 하기 때문이다. 그러나 이 때 형태소 분석 정보가 손실되지 않아야 한다. 응용 시스템에서 형태소 분석 결과를 쉽게 해석할 수 있도록 한국어의 단어 유형을 단순화시키면 그림 2와 같이 14가지 유형으로 분류된다. 단어의 유형을 단순화하기 위해 적용한 단순화 규칙은 다음과 같다.

① 체언에 대한 접두사와 체언접미사는 체언부로 포함시키고 체언에는 이에 관한 정보를 부여한다. 즉, 체언접미사가 분리된 체언은 분리된 접미사의 유형에 대한 정보를 갖는다. 접두사가 분리되는 경우도 체언접미사와 같은 방법으로 처리한다.
② 서술격 조사 '이'는 용언화 접미사로 간주한다.
③ 선어말어미는 어말어미와 통합하여 어미부로 한다.

```
N1. PTN_N       체언          /* N/PN/NM/XN/CN/UN/AS/HJ/ET */
N2. PTN_NJ      체언 + 조사
N3. PTN_NSM     체언 + 용언화접미사 + 어미
N4. PTN_NSMJ    체언 + 용언화접미사 + 'ㅁ/기' + 조사
N5. PTN_NSMXM   체언 + 용언화접미사 + '아/어' + 보조용언 + 어미
N6. PTN_NJCM    체언 + '에서/부터/에서부터' + '이' + 어미

V1. PTN_VM      용언 + 어미
V2. PTN_VMJ     용언 + 'ㅁ/기' + 조사
V3. PTN_VMCM    용언 + 'ㅁ/기' + '이' + 어미
V4. PTN_VMXM    용언 + '아/어' + 보조용언 + 어미
V5. PTN_VMXMJ   용언 + '아/어' + 보조용언 + 'ㅁ/기' + 조사

A1. PTN_AID     단일어          /* ADV/INT/DET */
A2. PTN_ADVJ    부사 + 조사     /* "빨리도"  */

X1. PTN_NVM     체언 + 동사 + 어미
```

그림 2. 단순화된 단어 유형의 분류

단순화된 단어 유형을 사용한 형태소 분석 결과에서 필요한 정보를 모두 추출하려면 단순화 규칙을 역으로 적용하면 된다. 즉, 체언이 분리된 것은 접두사와 체언접미사를 검사하고 어미가 분리된 것은 선어말어미 정보를 검사함으로써 분리된 형태소를 순서대로 복원할 수 있다. 어형 인식 후에 어형 확장에 의한 형태소 분석 결과에서는 복합명사로부터 단위 명사를 추출하고 복합조사와 복합어미로부터 단위 조사와 단위 어미를 추출하는 과정 등을 추가해야 한다. 또한 각 단위 형태소에 대한 세부 정보를 부여하여 이를 응용 시스템에서 사용할 수 있도록 한다.

3. 단어 구성 전이도

단어 구성 전이도(word transition network)는 단어의 유형에 따라 구성되는 것이 바람직하다. 그런데 그림 2에서 분류한 단어의 유형은 접미사와 선어말어미가 단어의 유형에 직접 반영되지 않고 체언과 어말어미에 첨부된 정보로 간주되고 있다. 따라서 그림 2의 단어 유형을 기반으로 하고 접미사와 선어말어미의 분리를 고려하여 단어 구성 전이도를 구성한다.

한국어의 단어 형성 원리(word formation rule)와 한글 맞춤법의 띄어쓰기 규칙에 따라 단어 구성 전이도를 기술하면 그림 3과 같다. 그림 3의 단어 구성 전이도는 단어의 유형을 인식하기 위한 관점에서 구성한 것으로 형태소의 품사 체계와 밀접한 연관성이 있다. 권혁철(1991)은 어말어미 뒤에 오는 명사와 조사, 선어말어미 등을 반복적으로 기술하고 있다[4]. 그러나 그림 3에서는 어형 인식의 효율성 측면에서 조사 또는 어말어미의 결합형을 사전에 수록한다고 가정하여 반복부를 제거하였고, 어말어미 뒤에 오는 특수 조사의 경우에도 어말어미의 일부로 간주함으로써 단어 구성 전이도를 단순화시켰다.

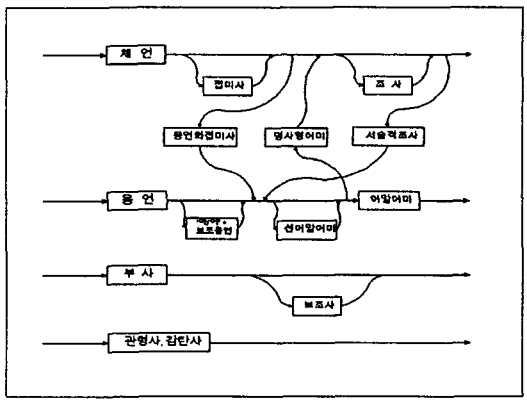

그림 3. 한국어 단어 구성 전이도

김덕봉(1990)은 접두사를 별개의 토큰으로 기술하고 있다[5]. 그런데 접두사를 별개의 토큰으로 간주하면 접두사를 취하는 조건과 취할 수 없는 조건이 품사 정보에 의해 구분되지 않는다. 그림 3에서는 모든 명사에 대하여 접두사 결합 조건을 기술해야 하는 문제를 극복하기 위하여 체언의 일부로 간주하였다. 또한 김덕봉(1990)은 서술격 조사를 체언 뒤에 오는 형태로 구성하였는데, 이 경우에 '합법적이다', '학교에서였다' 등과 같이 조사 뒤에 오는 서술격 조사를 처리하지 못하는 문제가 있다. 김병희(1993)는 체언 뒤에 오는 부사화 접미사를 단어 구성 전이도에 포함하였으나 그림 3에서는 형태소 사전에 수록되는 것으로 가정하였다[6].

그림 3의 단어 구성 전이도는 형태소끼리의 결합 제약 조건이 반영되지 않았기 때문에 형태소 분리시에 결합 제약 검사가 선행되어야 한다. 예를 들어, 체언 뒤에 오는 용언화 접미사는 일반적으로 접미사를 허용하지 않지만 '구체화되다', '단순화시키다'와 같이 용언화 접미사 앞에 올 수 있는 체언 접미사도 있다. 비슷한 예로 '학교에서였다'와 같이 서술격 조사 '-이' 앞에 조사가 올 수도 있다. 이 때 사용되는 조사는 '-에서', '-부터' 등 매우 제한적이므로 서술격 조사 '-이' 앞에 올 수 있는 조사는 '-에서', '-부터', '-에서부터'라는 제약 조건을 부여해야 한다.

용언화 접미사 뒤에는 '아/어'+보조용언이 올 수 있다. 이를 허용하려면 그림 3에서 용언화 접미사의 다음 단계를 '아/어'+보조용언으로 수정해야 한다. 그러나 용언화 접미사 뒤에 보조용언이 오는 것은 보조용언을 띄어쓰는 것이 일반적이므로 허용하지 않았다. 즉, 그림 3의 전이도는 한국어 단어의 형성 원리를 기반으로 하고 있으나 단어의 심층구조를 반영하기 않고 띄어쓰기 단위로 구분되는 표

층형을 대상으로 구성한 것이다.

4. 분석 결과의 저장구조

4.1 HAM의 분석 결과 저장 구조

HAM에서 형태소 분석 결과를 저장하기 위한 구조는 그림 4와 같이 단어에 대한 분석 결과를 저장하는 ham_result를 기반으로 한다. ham_result는 그림 1의 단어 유형 분류에 따라 분리되는 형태소들을 저장하는 frame 구조를 취하고 있다. 단어의 유형 정보인 pattern 항목에 의하여 단어의 유형을 파악한 후에 각 유형에 따라 필요한 항목에 저장되어 있는 정보를 사용한다.

ham_result에서 선어말어미 정보를 저장하는 pomi 항목은 8 비트 중에서 오른쪽 4 비트에 '시/었/였/겠'을 flag 정보로 표현한다. 즉, pomi가 2진수 00001100이면 입력 단어에서 선어말어미 '시'와 '었'이 분리되었음을 의미한다.[2)]

```
typedef struct ham_result {
  HAM_UCHAR pattern;          /*    word pattern    */
  HAM_UCHAR pos;              /* P.O.S. of the stem */
  HAM_UCHAR dinfo;            /* dic. info. of stem */
  HAM_UCHAR suffix;           /* index of suffix tab*/
  HAM_UCHAR pomi;             /* prefinal Eomi info.*/
  HAM_UCHAR vtype;            /* irreg type of verb */
  HAM_UCHAR phon[STEMSIZE]; /* stem of input word */
  HAM_UCHAR josa[JOSASIZE]; /* string of Josa     */
  HAM_UCHAR eomi[EOMISIZE]; /* string of Eomi     */
  HAM_UCHAR xverb[XVERBSIZ];/* string of xverb    */
} HAM_RESULTSTR, *HAM_PRESULTSTR; /* RESULT STRUCT*/
```

그림 4. HAM에서 형태소 분석 결과의 저장 구조

ham_result에서 어말어미를 저장하는 항목이 eomi이다. 그런데 '먹어보다'와 같이 붙여쓴 보조용언이 포함된 단어에서는 어말어미가 '-어'와 '-다' 두 개가 분리되고 있다. 다행히도 한글 맞춤법에서 보조용언의 붙여쓰기는 본용언의 어말어미가 '-어'인 경우에만 허용되므로 '-어'는 따로 저장하지 않고 보조용언에 대한 단어 유형에서는 '-어'가 분리된 것으로 간주함으로써 해결된다. 다만, 맞춤법에는 어긋나지만 일반적으로 '하고싶다', '하게되다'와 같이 본용언의 어말어미가 '-고'와 '-게'인 경우에 붙여쓰기도 한다.

'-고/-게' 뒤에 보조용언을 붙여쓰는 것은 띄어쓰기 규칙에 어긋나기 때문에 HAM version 1.5는 형태소 분석을 실패하도록 하였으나, version 2.0에서는 분석이 가능하도록 수정하였다. 그런데 이 때 '-고'와 '-게'가 분리되었다는 정보를 저장할 항목이 없으므로 선어말어미 항목 pomi에서 사용되지 않고 있는 2비트를 활용하고 있다. 또한, '-[에서][부터]였다'와 같이 서술격 조사 '이' 앞에 조사가 오는 유형을 처리하기 위하여 pomi 항목의 2 비트를 flag로 사용한다. 즉, '에서'가 분리되었으면 첫번째 비트, '부터'가 분리되었으면 두 번째 비트를 1로 한다. 이 때 josa 항목을 사용하지 않은 이유는 '-에서부터였음은'과 같이 조사가 분리될 수도 있기 때문이다.

모호성이 내포된 단어는 두 가지 이상의 분석 결과가 가능하므로 한 단어에 대하여 두 가지 이상의 분석 결과를 저장할 수 있도록 해야 한다. 그림 5는 단어 및 문장 단위의 형태소 분석 결과를 저장하는 구조체인데 단어 수준의 분석 결과인 ham_word에 ham_result를 최대 15개까지 저장할 수 있도록 하고 있다. 또한 문장 수준의 분석 결과는 각 단어에 대한 분석 결과를 저장하여 모호성 해결이나 구문 분석기에서 직접 활용할 수 있도록 한다.

```
typedef struct ham_word {
  HAM_UCHAR phon[WORDSIZE];/* string of input word*/
  HAM_SHORT ret_code;      /* HAM return code     */
  HAM_SHORT nresults;      /* # of HAM results    */
  struct ham_result result[MAXRESULT];/* HAM res. */
} HAM_WORDSTR, HAM_FAR * HAM_PWORDSTR;/* WORD STR.*/

typedef struct sentence {
  HAM_UCHAR phon[SENTSIZE];/* string of input sent*/
  HAM_SHORT senttype;      /* type of a sentence  */
  HAM_SHORT nwords;        /* # of words in sent  */
  struct ham_word words[MAXWORDS];/* word struct. */
} HAM_SENTSTR, HAM_FAR * HAM_PSENTSTR;/* SENT STR.*/
```

그림 5. 단어 및 문장 단위 분석 결과

4.2 형태소 분석 결과의 해석 방법

형태소 분석 결과를 해석하는 방법은 미리 정의된 단어 유형에 따라 필요한 항목을 검사하는 방법(어형 해석법)과 그림 3의 단어 구성 전이도에 의해 어근의 유형에 따라 필요한 항목을 검사하는 방법(어근 해석법)이 있다. 어형 해석법은 단어 유형마다 각각 분석 결과를 생성하는 routine이 필요하므로 단어 유형이 다양할수록 분석 결과의 해석 방법이 복잡해지는 단점이 있다. 따라서 이 방법을 사용할 때는 그림 1처럼 단어의 유형을 세분화하기 보다는 그림 2와 같이 단어의 유형을 단순화시키는 것이 바람직하다.

어근 해석법은 모든 단어에서 분리되는 어근(stem)의 유형을 파악하여 그림 3의 단어 구성 전이도에 따라 형태소 분석 결과를 해석한다. 이 방법은 단어 유형에 무관하게 형태소 분석 결과의 생성 과정이 하나의 routine으로 구성되는 장점이 있다. 어근 해석법에 의한 형태소 분석 결과의 생성 알고리즘은 그림 6과 같다. 이 알고리즘의 형태소 해석 과정을 추적하면 단어의 유형을 파악할 수 있다.

HAM에서 사용되고 있는 frame 구조는 형태소 분석 결과를 표현하는데 적합하며 어형 해석법이나 어근 해석법을 적용하는 것이 모두 가능하다. 그러나 형태소 분석 결과를 쉽게 활용할 수 있도록 하고 정보를 더욱 효율적

2) HAM version 1.5에서는 선어말어미의 분석 결과를 16진수로 출력하고 있다.

으로 저장하기 위하여 보완되어야 한다.

HAM의 형태소 분석 자료구조(그림 4)를 기반으로 단어 구성 전이도를 활용한 어근 해석법을 적용하기 쉽게 재구성한 자료구조는 그림 7과 같다. 이 자료구조는 어근과 체언 접미사, 조사와 어미를 각각 구조체로 독립시켰다. 조사와 어미를 구조체로 독립시킨 이유는 어형 확장시에 조사와 어미에 대한 세분화된 정보를 쉽게 추가할 수 있도록 하기 위한 것이다.

```
1. 어근 및 체언접미사 생성;
2. if (stemtype == 체언) {
     용언화접미사 확인 및 생성;
     if (v-suffix_is_found) goto eomipart;
   josapart:
     조사 확인 및 생성;
     if (Josa == "[에서][부터]" and
         copula_is_found) goto eomipart;
   }
3. else if (stemtype == 용언) {
     '아/어'+보조용언 확인 및 생성;
   eomipart:
     선어말어미 확인 및 생성;
     어말어미 확인 및 생성;
     if (nominal_Eomi_is_found) goto josapart;
   }
4. else (stemtype == 부사)
     보조사 확인 및 생성;
```

그림 6. 어근 해석법에 의한 알고리즘

어근과 체언 접미사를 구조체로 독립시키려면 접미사 항목을 체언 접미사와 용언화 접미사로 구분해야 한다. HAM에서는 하나의 접미사 항목을 공유하여 단어 유형에 따라 어떤 table의 인덱스인지를 구분하고 있다. 그러나 "과학적인", "현실화되다"와 같이 체언 접미사와 용언화 접미사가 동시에 오는 단어도 있으므로 접미사 index 항목은 체언 접미사와 용언화 접미사를 각각 설정하는 것이 바람직하다.

```
typedef struct {
  HAM_UCHAR phon[STEMSIZE]; /* stem of input word */
  HAM_UCHAR dinf;           /* parts of speech:dic*/
  HAM_UCHAR nsfx;           /* index of noun sfix */
} STEM_STR, *PSTEM_STR;     /* structure of stem  */

typedef struct {
  HAM_SHORT josa;           /* index of Josa tab  */
  HAM_SHORT eomi;           /* index of Eomi tab. */
  HAM_UCHAR pomi;           /* prefinal Eomi info.*/
} JOEM_STR, *PJOEM_STR;     /* struct of Josa/Eomi*/

typedef struct ham_result {
  HAM_UCHAR patn;           /*    word pattern    */
  HAM_UCHAR type;           /* type of input word */
  STEM_STR  stem;           /* stem and suffixes  */
  HAM_UCHAR vsfx;           /* index of verb sfix */
  JOEM_STR  joem;           /* Josa, Eomi, p-Eomi */
  HAM_UCHAR xverb;          /* index of xverb tab.*/
} HAM_RESULTSTR, *HAM_PRESULTSTR; /* RESULT STRUCT*/
```

그림 7. 형태소 분석 결과의 저장 구조

이 자료구조에서 단어의 유형을 표시하는 patn 항목은 분석 결과를 해석하는 알고리즘에 의하여 추론이 가능하다. 이 때 patn의 단어 유형은 그림 2에서 제시한 단순화된 분류법을 사용하는 것이 효율적이다. 또한 어근의 유형(체언/용언/기타)을 나타내는 type 항목은 어근 구조체에서 사전에 수록된 정보인 dinf의 분석에 의하여 알 수 있으나 응용 시스템에서 쉽게 사용할 수 있도록 하기 위한 것이다. 또한 조사와 어말어미를 저장하는 항목은 string이 아니라 조사/어미 table에 대한 index로 구성하여 저장 공간의 효율성을 높이고 있다.

5. 결 론

한국어 형태소 분석기에서 형태소 분석 결과를 효율적으로 표현하는 방법을 살펴 보았다. 분석 결과를 저장하는 새로운 자료구조를 제안하기 위하여 형태소 분석에 필요한 단어의 유형을 일반적인 유형과 단순화된 유형으로 구분하여 비교하였는데 단순화된 유형을 사용하는 것이 바람직함을 알 수 있었다. 또한 한국어의 단어 구성 전이도는 형태소 분석기마다 다른 형태로 구성되고 있는데 이를 비교하여 문제점을 파악함으로써 새로운 단어 구성 전이도를 제시하였다.

단어의 유형 분류와 단어 구성 전이도를 기반으로 형태소 분석 결과를 효율적으로 저장하기 위한 자료구조를 제안하였다. 본 논문에서 제안한 자료구조는 HAM에서 사용되고 있는 자료구조를 보완한 것으로 응용 시스템마다 다양한 요구 사항이 발생하는 특성을 고려하여 형태소 분석 결과를 쉽게 활용하는 측면과 저장 공간의 효율성을 개선하였다.

참고문헌

[1] 제1회 우리말 정보처리 규격 심포지움, 한국과학기술원 인공지능 연구센터, 1996년 7월.

[2] 강승식, 장병탁, "음절 특성을 이용한 범용 한국어 형태소 분석기 및 맞춤법 검사기", 정보과학회 논문지(B), 23권 5호, pp.530-539, 1996.

[3] 강승식, *음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석*, 서울대학교 컴퓨터공학과 박사학위 논문, 1993년 2월.

[4] 권혁철, 채영숙, 김재원, 김민정, "한국어 철자 검색을 위한 형태소 분석 기법", 국어정보학회 학술발표 논문집, pp.179-186, 1991.

[5] 김덕봉, 최기선, 강재우, "한국어 형태소 처리와 사전 - 접속정보를 이용한 한글 철자 및 띄어쓰기 검사기 -", 어학연구, 26권, 1호, pp.87-113, 1990.

[6] 김병희, 임권묵, 송만석, "형태소 접속 특성과 인접 말마디 정보를 이용한 형태소 분석기", 제5회 한글 및 한국어 정보처리 학술발표 논문집, pp.395-404, 1993.