한일기계번역시스템의 사전을 사용한 한국어 형태소분석시스템

姜 龍熙(동경대학), 田中光一(히타치정보네트워크), 松田純一(히타치중앙연구소)

The Korean Analysis Systerm by The Using of The Korean/Japanese Maching Translation's Dictionary

Yong Hee Kang(Tokyo University),Kouichi Tanaka,Junichi Matsuda(Hitachi, Ltd)

한일기계번역시스템의 형태소 해석 프로그램의 중간버퍼를 표준안에 맞추어 명사추출 및 품사태킹을 시도해 보았다. 기존의 모델을 유지하면서 사전의 표제어를 보충하여 출력의 형태를 바꾸는 방법으로 표준안의 출력에 가깝게 출력을 함으로써, 기존의 프로그램의 장점과 단점을 보완하는 것과, 표준안에 관한 문제제기가 본 연구의 목적이다. 특히 품사개념이 다른 사전에서 태킹 및 명사추출을 실시할 경우 표제어의 등록여부와 정확률의 인과관계는 높다고 판단된다. 그러므로 표준안의 품사기준은 그에 따른 시스템의 성패를 좌우한다.

목차

1. 序論

한일기계번역시스템의 형태소 해석 프로그램의 중간버퍼를 표준안[1]에 맞추어 명사추출 및 품사부착을 시도해 보았다. 연구방법으로는 크게 다음과 같이 나눌 수 있다

1)시험 제출 결과와 추가사전보수의 결과를 비교하는 방법을 사용한다.

2)오류발생의 원인를 사전과 프로그램의 문제로 나누어 규명하는 방법을 채택한다.

3)문제점의 접근방법은 표준안과 TU-HI의 품사개념과 대조해 보는 방법이다.

(이하 팀명은 TU-HI로 표시함.)

1.1. 시스템의 사양

1 표준안의 품사정의는 16개의 품사그룹 속에 28종의 품사로 이루어져 있다. 표준문법의 9품사와 다르다.

· 기본모델 :HICOM M/T(Hitachi)[2]

· 사전 1종 :knnew.nax(약10MB)

· 품사의 분류: 28종의 품사와 125개의 등록형으로 사전이 이루어져 있다.

· 사전의 검색방법은 순차검색방법과 램덤검색방식을 병용하고 있다.

사전의 기본 형식은 다음과 같다.

1표제어 2번역어 3품사 4품사 5의미정보 6의미정보
한국어 일본어 한국어 일본어

· 입력문의 제한

1)입력어절의 최대길이는 512문자

2)입력문장 최대길이도 512문자

3)입력문은 일반문장과 어절형 문장이 가능하나, 어절형 문장은 공기정보를 사용하는 부분과 連語型 형태소에서 오류가 발생한다. 본 대회에 제출한 결과는 어절형 단위문장을 문장형으로 고친 후 출력한 결과이다.

1.2. 품사

1. 명사류
   - 1.1 일반명사
   - 1.2 고유명사
   - 1.3 인명(姓)
   - 1.4 인명(名)
   - 1.5 지명
   - 1.6 법인명
   - 1.7 동사어간 가능명사(-하다型)
   - 1.8 대명사
2. 용언류
   - 2.1 동사
   - 2.2 형용사
3. 조사류
   - 3.1 조사
   - 3.2 終조사(어미)
   - 3.3 接續조사
4. 부사
   - 4.1 부사(일반부사)
   - 4.2 접속사 (접속부사)
5. 접사
   - 5.1 접두사
   - 5.1.1 數詞선두어
   - 5.2 접미사
   - 5.2.1 數詞用접미사
6. 관형사
7. 기호류
   - 7.1 숫자
   - 7.2 英字
   - 7.3 기호(4종)
8. 連語형
   - 8.1 조동사(어미+보조어간;10종)
   - 8.2 連用수식어
   - 8.3 連體수식어,
9. 기타류
   - 9.1 인사말
   - 9.2 제목
10. 등록형
    - 10.1 동사8형(어간+어미형 포함[3])
    - 10.2 변칙동사5형
    - 10.3 조사 6형
    - 10.4 형용사 7형
    - 10.5 접속조사 7형
    - 10.6 조동사 8형

또한 등록단위중 활용형과 동음이의어의 환경에 대비한 2중 등록형[4]을 포함하고 있다. 등록형은 파라다임[5](paradigme)으로 입력하며 동사, 동사, 형용사, 조동사(어미+보조용언)의 기본형과 그 활용형을 인덱스형으로 등록하는 방법이다. 그러나 용언의 명사형은 등록형이 아닌 단독형으로 처리하고 있다.

표제어 번역어 품사 품사 의미정보 의미정보
활용형(한) 어간(한) 인덱스

예)

기본 달리 はしる(번역어)

연용 달려 달리(어간) 인덱스정보

---

2 본 연구와 한일기계번역 시스템의 번역률과 무관하며 발표자의 개인 연구이다.

3 파라다임의 [달려]는 [달리+어]로 분리되나 번역상의 이점을 살려 단독형으로 처리한다.

4 예)나(대명사)+는(조사) : 나는(날다)

5 [합니다]형은 본 시스템에서는 번역어의 특성상 패러다임에 포함하지 않고 별도의 처리를 한다.

합니다 → します

예쁩니다 → きれいです

연체 달린 달리(어간) 인덱스정보
연용 달렸 달리(어간) 인덱스정보
연체 달리는달리(어간) 인덱스정보
연체 달릴 달리(어간) 인덱스정보
연용 달린 달리(어간) 인덱스정보
명령 달려라 달리(어간) 인덱스정보
*명사형 달림 ----------

위의 품사들과 표준안의 품사를 비교해 보면 표준안에는 [외국어], [파생접미사], [지정사]등이 채택되어 있지만 본 연구팀의 사전에는 [외국어], [파생 접미사], [지정사]등의 품사개념은 없다. 왜냐하면 외국어는 품사의 개념이 아니고, 파생접미사는 표제어 단위가 아닌 형태소의 일부분이며 지정사[이다]는 형용사의 품사처리를 하고 있다. 또한 번역어의 문법구조를 고려한 단일 형태소가 아닌 연어형과 조동사(어미+보조용언)등을 채택하고 있는 것이 가장 큰 차이점이다.

1.3. 사용모델

기본적으로 한일 기계번역의 중간버퍼를 사용하고 있으므로 품사부착과 명사추출의 프로그램은 동일하다. 표제어로부터 번역어선택의 기준은 n文節(어절)최장일치(n= 2 or 3)와 일부분의 환경에서는 별도의 기준을 채택하고 있으며, 동음이의어 환경에서는 공기정보에 의한 번역어의 선택을 하고 있으며 모든 명사에는 용언선택에 필요한 의미정보(의미 마커; 250종)가 포함되어 있다. 그리고 품사 접속 테이블(좌우 2항비교)도 참고 자료로 삼고 있다.

품사의 갯수(28) × 품사의 갯수(28)=784

1.4. 사전의 크기

1)표제어

기본적으로 표제어(한국어)와 번역어(일본어)의 숫자가 다르며, 하나의 표제어에 보통 여러가지 번역어를 가지고 있는 경우가 일반적이다.

2)기본적인 분류와 크기

사전 전체	표제어	189,747
	번역어	257,973
명사류	표제어	106,436
	번역어	146,916
용언류	표제어	8,561
	번역어	12,506
기타류	표제어	1,396
	번역어	2,181

3)세부적인 품사와 숫자

형식 : 품사	표제어	번역어
명사	83,474	118,740
한자어+하다의 명사	10,433	12,907
고유명사(地名등)	24,689	26,780
동사(어간)	6,005	7,969
동사(변칙)	44	79
부사	1,907	2,512
조사(조사+수식어포함)	225	285
어미	332	518
접사	759	981
보조용언[6]	545	818
연용수식連語[7]	173	194
연체수식連語[8]	73	81
접속조사	440	562

이하 생략

1.5. 시스템의 시간 복잡도

PC-Pentium 200MHZ에서 1만어절 처리속도는 600초이다.

1.6. 각 프로그램의 크기

수행시 최소 필요한 RAM크기는 500KB이상이다.

각 실행 화일의 크기 약 12KB이다.

1) 일반형(khitr)
2) 품사선택정보 동시출력형(khitrl)
3) 미등록어 동시검출출력형(khitrm)
4) 미등록어 병합프로그램(xemrg계열)
(약10KB~33KB)

1.7. 시스템의 장점

1) 사전의 수정 및 관리가 편리하다.

---

6 조동사(어미+용언)의 일부분
예) 먹고 싶다= 먹+고 싶다 -> 싶다
7 예)~에 관하여
8 예)~에 관한

2) 국어 문법구조에 가깝게 작성되어 있다.
3) 입력문이 어절형이 아닌 일반문이다.

2. 本論

2.1 대회참가 결과

대회측으로부터 품사부착정보 코파스와 원시 코파스를 제공받아 기존의 품사분류를 수정 사전의 보수작업을 했다. 그러나 품사부착 처리 프로그램의 처리에 많은 시간이 투여되어 사전 보수 작업은 거의 실시되지 못한 상태에 대회에 결과 보고를 마쳤다.

2.1.1 참가팀의 각 분야별 평균값

참가팀(18개팀)의 각 분야별 평균값은 표1)과 같다.(이하 R-재현률, P-정확률)

표1)

	재현률	정확률
뉴스	0.85	0.81
비소설	0.88	0.83
소설	0.83	0.75
전체	0.85	0.80

2.1.2 TU-HI의 대회제출결과

TU-HI가 대회기간중 제출한 결과의 평균치는 표2)와 같다.

표2)

	재현률	정확률	일탈률
뉴스	0.61	0.55	R 0.08 P 0.09
비소설	0.69	0.66	R 0.08 P 0.08
소설	0.64	0.54	R 0.08 P 0.08
전체	0.65	0.58	R 0.09 P 0.10

위의 결과로부터 다음과 같은 성향을 인식할 수 있다.

1) 뉴스와 소설(대화체)과 같은 인명이 빈번하게 출현하는 문장과 대화체 문장에 자주 쓰이는 축약형은 오류를 발생하게 하기 쉽고 비소설과 같이 문어형의 문장과 한자어의 사용이 상대적으로 높은 문장에서는 정확률이 높게 나타난다.

2) 다음의 연구방법에서 실제로 인명등의 사전보수를 한 후 결과(2.1.3참조)를 비교해 보면 위의 특성이 잘 나타난다.

2.1.3. 사전보충후 재평가

사전보수는 고유명사(인명)를 중심으로 대회측이 제공한 표준답안 명사를 6,000개를 최하위 순위로 보충한후, 프로그램은 고치지 않고 재실시 했다. 즉 명사미등록어를 배제한 단계에서의 평가이다. 결과는 표3)과 같다.

표3)

	재현률	정확률
뉴스	1.13	0.84
비소설	1.03	0.84
소설	0.99	0.69
전체	1.05	0.79

전체최저치의 파일을 재실시한 결과, 재현률과 정확률의 값이 크게 변했다.

전체최저 R 0.41 P 0.33 → R 1.05 P 0.88

평균값 이하의 파일은 인명 및 단체명의 빈도가 높은 파일이었고 사전의 표제어를 등록만 한 결과 평균치를 넘는 정확률을 보였다.

위의 결과의 원인으로는 기계번역에서 필요한 사전 정보로는 인명 및 단체명은 사전등록을 하지 않고 번역어의 생성과정에서 로마자 및 일본어의 가나문자로 처리하는 프로그램의 특성상, 형태소 분석과정에서는 정확히 도출하지 못한 점을 단순히 등록하는 방법으로 오류를 줄인 예라고 할 수 있다.

2.1.4. 대회참가를 위한 수정작업

1)前작업으로는, 대회측이 제공하는 시험용 입력문이 어절형이므로 일반문장으로 변형하는 작업을 해야했다. 그러나 품사부착한 출력이 일부분의 환경에서 표준안과 다른 관계로 대회측이 평가를 하지 못하는 문제가 발생한다.

2) 본 시스템의 본래의 출력형태는 입력문이 반각문자의 경우에도 사전구조상 전각문자로 출력하는 방법을 채택하고 있었으나 평가를 받기 위해 반각문자로 바꾸는 작업을

했다.

3)기계번역 시스템에서는 염두에 두고 있지 않는 축약형의 처리도 별도로 해야했다.

예를 들어 동사의 과거형의 경우 출현되는 예가 음운론적 이형태[9](allomorph)가 있고 형태소가 단수가 아니며(예 1), 2)) 기본형과 과거형의 형태소로 나누어 처리하는 과정까지 있어 시간적 부담이 되었다. 또한 조사의 축약형(contracted particle)의 경우 문자열 처리로 일부분의 조사만 처리하는 방법을 택했기에 [명사+조사]형의 축약형(예4)은 처리하지 못하는 문제점을 남겼다.

예1) 보다 봤다 → 보+았+다
보았다 → 보+았+다
예2) 던지다 던졌다 → 던지+었+다
던지었다 → 던지+었+다
예3) 그는 학교엘 갔다.
학교엘 갔다 → 학교+에+를
예4) 그는 기찰 타고 학교에 갔다.
기찰 → *기차+를

4)표준안과 품사개념이 다른 시스템의 품사를 표준안에 가깝게 출력할 수 있도록 처리 작업을 했다. 그러나 하나의 품사라도 기존의 프로그램의 품사개념이 표준안과 일치하지 부분에서 복수의 처리를 해야하는 문제가 있어 품사부착 오류의 결과로 나타나게 된다. 구체적으로 표준안에서는 [제1차 학술대회] 의[제]만을 유일하게 접두사로 인정하고 있다. 또한 [제일차 학술대회]와 같이 한글로 수사가 나타나는 경우에는 수사의 일부로 품사부착하고 있다. 그러나 TU-HI의 품사의 접두사에는 [제]를 포함하여 복수로 존재하며 명사와 복합되는 단어는 개별적으로 표제어로 등록하지 않고 복합하여 생성하는 방법을 채택하고 있기 때문에 기존에 접두사로 처리하던 단어에서 오류로 처리되는 예가 많았다.

예) 날고기 →날+고기 ,날음식→ 날+음식

특히 외국어와 같은 기존의 시스템에서는 사용하지 품사의 환경에서는 오류의 발생률(2.5 참조)이 높았다. 다음의 표4)는 표준안과 본 연구팀의 품사대조와 변환 적용 규칙이다.

표4)

계층1	계층2	계층3	TU-HI
1. s			7.1, 7.2, 7.3
2. f			NO (S)
3. n			
	3.1 nc		1.1-1.7
	3.2 nb		5.2.1
4. np			1.8
5. nn			5.1.1 (S)
6. pv			2.1, 10.1, 10.2
7. pa			2.2, 10.4
8. px			10.6
9. co			NO (2.2;S)
10. ma	10.1 mag		4.1
	10.2 maj		4.2
11. mm			6.
12. ii			NO(9.1;S)
13. x	13.1 xp		5.1.1(S)
	13.2 xs	13.2.1 xsn	1.1
		13.2.2 xsv	10.1
		13.2.3 xsm	10.4
14. j	14.1 jc		3.1, 10.3 S
	14.2 jx		3.1, 10.3 S
	14.3 jj		3.1, 10.3 S
	14.4 jm		3.1, 10.3 S
15. ep			NO 8.1,10.6
16. e	16.1 ef		3.2
	16.2 ec		3.3, 10.5
	16.3 et	16.3.1 etn	NO 1.1
		16.3.2 etm	10.1

NO = 해당하는 품사개념 없음.
S = 품사처리가 아닌 문자열처리임.

2.2. 명사추출의 오류분석

본 시스템은 품사부착과 명사추출을 동시에 출력하는 방법을 택하고 있다. 대회측에 제출한 결과와 사전 보수 작업을 행한 결과를 비교하며 문제점 및 개선 방안을 모색해 보았다.

9 됐다 → 되+었+다
했다 → 하+였+다
갔다 → 가+았+다

예문)1 파일 noun1000

01 1997년 1997/nn+년/nc
02 10월 10월/nc
03 01일 01/nn+일/nc
04 류근찬 류/xsn+근/f+차/pv+ㄴ/etm
05 앵커 앵커/f
06 : :/s
07 여러분 여러분/nc
08 안녕하십니까?안녕/nc+하/xsm+시/ep+ㅂ니까/px+?/s
09 KBS 9시 KBS 9/f+시/xsn
10 뉴스입니다. 뉴스/f+이/co+ㅂ니다/ef+./s

위의 결과에서 대표적인 명사추출의 오류와 그 원인을 모색하면 다음과 같다.

예 a) 1997년 1997/nn+년/nc
b) 10월 10월/nc
c) 류근찬 류/xsn+근/f+차/pv+ㄴ/etm
d) 앵커 앵커/f, 뉴스 뉴스/f

예a)의 경우는 수사 다음에 나타나는 [년]을 의존명사로 품사부착하지 않고 일반명사로 품사부착하여 과명사 처리된 오류이다. 원인으로는 기존 사전에 [년]이 일반명사로 입력되어 있기 때문에 품사변환의 적용규칙에 해당되지 않은 경우이다.

예b)의 경우에는 다음과 같은 경우를 상정할 수 있다. 기계번역의 특성상 번역을 우선하는 방법을 택하게 되므로 형태소 단위와는 무관한 일관형으로 등록하는 경우가 있을 수 있다. 이와 같은 경우가 사전 입력단위의 문제로 프로그램 처리와 별도로 오류로 연결되는 경우이다.

예c)의 경우는 인명을 번역어의 생성과정에서 변환하는 프로그램 구조상 사전에 미등록된 인명을 형태소 분석하는 경우이다. 이와 같은 예는 사전 보충으로 처리 가능하다.

예d)의 경우는 사전에 이미 명사로 등록되어 있으나, 표준안의 외국어(기존의 시스템에서는 비사용 품사)를 처리하기 위해 외래어를 외국어 처리하는 방법을 채택함으로써 유발된 오류이다. 이는 프로그램의 처리과정의 부적절한 판단에서 발생되었으며 표준안의 문제점으로도 지적할 수 있다.

예문2) noun2000

01 투르크족들의 투르크족/f+들/xsn+의/jm
02 민족 민족/nc
03 이동은 이동/nc+은/jc
04 언어 생활에도 언어 생활/nc+에/jc+도/jc
05 영향을 영향/nc+을/jc
06 끼쳐, 끼/nc+치/pv+어/ec+,/s

예 d-1) 끼쳐, 끼/nc+치/pv+어/ec+,/s
예 d-2) 끼쳐, 끼치/pv+어/ec+,/s

위의 d-1)의 경우는 명사를 과추출한 경우이다. 즉 [동사어간+어미]로 분석해야 환경에서 동음이의어의 처리가 불완전 했던 경우이며 동사 [끼치다]가 미등록되어 일어난 오류의 경우이다. 동사를 등록한 후 재실시 한 결과 (d-2) 오류가 발생하지 않았다.

예문3) noun2000

01 투르크족들은 투르크족/f+들/xsn+은/nc
02 오늘날 오늘날/nc
03 카자흐스탄에서 카자흐스탄/nc+에서/jc
04 트란스옥시아나 트/pv+란/jc+스/nc+옥/nc+시/nc+아/nc+나/jx
05 지역까지 지역/nc+까지/jc
06 이동하여 이동/nc+하/xsv+여/ec

예)e 트란스옥시아나트/pv+란/jc+스/nc+옥/nc+시/nc+아/nc+나/jx

위의 예 e)는 n문절(어절)최장일치와는 별도의 기준을 적용하는 경우에 고유명사라도 그 출현빈도의 영향으로 과분해해서 오류가 발생한 경우이다. 즉 시스템 처리중 품사 우선순위[10]에 밀려 오류가 발생한 경우이다.위와 같은 경우의 오류는 그 빈도가 적다.

2.3 명사추출 오류의 패턴

발생한 오류를 패턴으로 나누면 다음과 같이 5가지로 나눌 수 있다.

패턴1) 사전의 품사 입력 단위 오류
패턴2) 사전의 입력 형태소부족 오류
패턴3) 사전의 품사 출력 적용 처리 오류
패턴4) 동음이의어인한 과명사 출력오류

10 조사··>··어미··>··동사··>··명사···>···

패턴5) 사전의 등록여부와 무관한 시스템의 오류

위의 패턴을 사전과 시스템 프로그램의 오류로 분류하면 다음과 같다.

1) 사전오류 패턴: 1,2,4
2) 품사부착 프로그램의 오류패턴: 3
3) 시스템 선택기준 오류패턴 : 5

사전의 오류와 프로그램의 오류로 크게 나눌 수 있다. 위의 1)과 2)와 같이 그 원인을 알면 배제되는 오류가 있지만 3)과 같이 오류의 원인을 알아도 시스템의 선택기준에 의한 오류는 처리하기가 어렵다. 현 시점에서는 프로그램의 처리가 끝난 후 별도의 2차 프로그램에서 처리하는 방법등을 모색해야 한다.

2.4 품사부착 프로그램의 품사별 오류

본 연구팀의 품사부착의 출력형식이 대회 특이 준비한 형식과 일부 환경에서 다른 관계상 대회측의 평가는 없었다. 주로 오류를 중심으로 그 원인과 개선 방안을 모색해 본다

2.4.1 명사류

본 팀의 명사류 품사부착에 가장 큰 오류 발생환경은 다음의 품사간의 교체이다.

1) 일반명사 ↔ 의존명사
2) 일반명사 ↔ 수사
3) 일반명사 ↔ 부사
4) 의존명사 ↔ 수사

위의 1),2),4)의 경우에는 품사변환적용의 프로그램에 복합적인 원인이 많고 3)의 경우 사전에 표제어로 등록되어 있다 하더라도 조사와 결합된 환경에 나타나는 명사와 부사와의 동음이의어 환경에서 공기정보부족으로 인한 오류(2.3.2의 예문3)의 "한편" 참조)이다.

예문1) 표준안 tag10

1 벌금 벌금/nc
2 14억4천만원 14/nn+억/nn+4/nn+천만/nn+원/nb
3 추징금 추징금/nc
4 5억2천만원을 5/nn+억/nn+2/nn+천만/nn+원/nb+을/jc

예문2) TU-HI tag20

1 벌금 벌금/nc
2 14억4천만원 14/nn+억/nc+4/nn+천만/nc+원/xsn
3 추징금 추징금/nc
4 5억2천만원을 5/nn+억/nc+2천/nn+만원/nc+을/jc

표준안에는 [억], [천만][11]등의 단위가 수사와 의존명사로 품사부착하고 있다. TU-HI의 사전에는 일반명사와 접미사(數詞형)로 이중등록되어 있다. 그 결과 품사적용 규칙에 의해 명사와 의존명사로 품사부착이 되었고, 의존명사는 접미사로 품사부착되는 오류가 있었다. 또한 단위 기준의 억, 만, 천, 백, 십, 단위로 나누어 품사부착해야 되는 부분을 [만원]의 일반명사로 품사부착하는 오류를 범했다. 이와같은 예는 사전보수 및 문자열 처리등으로 수정 가능하다.

2.4.2 용언(보조용언 포함)과 어미류

용언오류의 원인은 사전의 미등록 상태를 배제하고 고려할 때 가장 복잡하고 품사적용의 애매성과 큰 관련이 있다.

예문1) 표준안 tag10

a) 카자흐스탄과 카자흐스탄/nc+과/jc
b) 비기고 비기/pv+고/ec
c) 말았습니다. 말/px+았/ep+습니다/ef+./s

예문2) TU-HI tag10

a) 카자흐스탄과 카자흐/f+스/nc+탄/nc+과/jc
b) 비기고 말았 비기/pv+고/ec+말/px+았/ep
c) 습니다. 습니다/ef+./s

위의 예문 2)의 c)와 같이 어절형 출력의 오류와 형태소의 일부분이 단절되는 오류가 있음을 알 수 있다. TU-HI의 가장 큰 문제점은 기존의 사전에 입력되어 있는 조동사(어미+동사, 어미+보조용언)의 처리이다. 조동사는 하나의 품사개념이므로 어미와 용언부분을 나누어 품사부착하는 단계(예문2)의 b)에 머물러 있다. 이와 같은 예는 번역어의 문법구조와 크게 관련이 있다. 2차 프로그램으로 표준안의 형식데로 출력하는 방법을 검

11 품사부착 말뭉치 구축 지침 p.16 참조

토 중이다. 보통 [ㅂ니다],[습니다]의 품사부착률은 99%이지만 예문2)의 c)와 같이 [습니다]가 분리되어 처리된 것은 예가 드물지만, [습니다]가 하나의 형태소인 관계로 어절형 출력단계에서 개행처리된 프로그램 버그로 인한 결과이다.

예문3) 표준안 tag20

a) 한편, 한편/maj+,/s
b) 이 이/mm
c) 시기에 시기/nc+에/jc

예문4) TU-HI tag20

a) 한편, 한편/nc+,/s
b) 이 이/mm
c) 시기에 시/pa+기에/ec

위의 예문4)의 c)는 형태소해석 단계에서 발생한 오류이다. 품사선택의 우선순위에 의한 오류이다. 표제어로 등록여부와 관계없고 공기정보와 의미정보를 기술하는 방법을 택해야 한다.

예문5) 표준안 tag20

a) 뿌리를 뿌리/nc+를/jc
b) 내렸던 내리/pv+었/ep+던/etm
c) 반면, 반면/nc+,/s

예문6)TU-HItag20

a) 뿌리를 뿌리/nc+를/jc
b) 내렸던 내리/pv+었/ep+던/ef
c) 반면, 반면/nc+,/s

위의 예문6)의 b)의 경우는 품사변환규칙의 오류이다. 샘플실험에서 어미류에 속하는 품사의 정확률은 80%정도이다.

2.4.3 조사류

위의 오류와 함께 빈번히 나타나는 조사의 오류는 공기정보 부족에 인한 것으로 예문2)의 a와 같은 예이다.

예문1) 표준안 tag20

a)인종과 인종/nc+과/jc
b)예술 예술/nc
c)같은 같/pa+은/etm

예문2) TU-HI tag20

a)인종과 인종/nc+과/jj
b)예술 예술/nc
c)같은 같/pa+은/etm

2.5.품사부착의 평균치

대회측에 제출한 품사부착 말뭉치의 출력형식에 문제가 있는 관계로 평가를 받지 못했으나 품사부착한 결과를 표준안의 형식으로 수동으로 고친 후 자체 평가를 한 결과 다음과 같은 평균치을 기록했다. 샘플의 재현률과 정확률을 계산하면 80%선을 유지한다.

대회참가팀의 평균치		TU-HI 표본 평균치	
	R	P	R	P
co	0.91	0.89	0.00	<none>
ec	0.93	0.92	0.60	0.64
ef	0.91	0.94	0.86	0.79
ep	0.97	0.99	0.83	0.86
etm	0.95	0.94	0.79	0.90
etn	0.95	0.86	<none>	0.00
f	0.94	0.97	0.67	0.06
ii	0.73	0.73	<none>	<none>
jc	0.96	0.97	0.82	0.75
jj	0.80	0.78	1.00	0.69
jm	0.98	0.99	0.95	1.00
jx	0.94	0.95	0.50	0.81
mag	0.93	0.90	0.67	0.75
maj	0.88	0.98	0.25	0.50
mm	0.79	0.77	0.67	0.67
nb	0.85	0.90	0.59	0.93
nc	0.87	0.81	0.81	0.89
nn	0.91	0.91	0.90	0.96
np	0.87	0.91	0.20	0.50
pa	0.93	0.88	0.50	0.38
pv	0.91	0.88	0.71	0.76
px	0.86	0.89	1.00	0.69
s	0.99	1.00	0.96	0.98
xp	0.86	0.86	0.00	0.00
xsm	0.84	0.86	0.79	0.79
xsn	0.93	0.92	0.67	0.76
xsv	0.93	0.95	0.85	0.93
전체	0.92	0.90	0.79	0.80

위의 결과는 기계번역의 번역률(참고문헌

[1],[2],[3],[4] 참조)과 비례하는 부분이 많으며 부사, 형용사, 특히 외국어의 경우는 등은 평균치를 밑도는 정확률로 나타났다. 의존명사의 재현률과 정확률의 수치로 품사적용 규칙이 기존의 품사(수사형 접미사)와 일치하는 부분과 불일치하는 비율(50%)을 알 수 있고 그 일단 품사부착된 정보의 정확률은 90%를 넘는다는 것을 알 수 있다.

2.6 표준안의 문제점

2.6.1.균등성의 문제

품사의 분류에 있어서 같은 품사로 분류되는 품사들이라면 품사의 성질이 같아야 한다. 예를들어 a,b,c가 같은 명사이라면 a에 적용되는 룰이라면 b와 c에도 동시에 적용이 되야 한다. 표준안의 품사부착 지침서의 내용중 균등성에 어긋나는 예로는 다음과 같다.

1) 품사[조용하다]의 처리문제(지침서 p.9)

표준안의 처리

a) 조용했다. 조용/nc+하/xsm+었/ep+다/ef+./s
b) 조용도 조용/nc+도/jc
c) 했다. 하/pa+었/ep+다/ef+./s

TU-HI의 처리

a) 조용했다. 조용/nc+하/xsm+였/ep+다/ef+./s
b) 조용도 조용/nc+도/jc
c) 했다. 하/pv+였/ep+다/ef+./s

본 팀에서는 [조용]을 명사형 등록하는 방법과 이미 사전에 등록되어 있는[~하다]를 [~+하다]분리하는 방법을 택했다.

표준안에서 [조용하다]를 [조용/n+하다/xsm]로 분리 처리하는 이유로 [조용]의 생산성을 예로 들고 있다. 그러나 명사로 분류된 [조용]이 다른 명사와 동일한 성질을 갖고 있는가 의문이다.

1) [조용]이 명사라면 보조사 이외의 조사와 공기할 수 있는가?

2) 표준안과 다른 해석은 불가능한가?

위의 내용을 적용해서 다음과 같은 처리를 가정해 본다.

예1)[조용하다]→조용하/pa

예2)[조용도 하다]→[조용/nc+도/jc+하/pv]

위와 같은 예는 [따뜻하다], [얌전하다]에도 적용이 가능하다. 즉 [도]와 같은 보조사와 공기하는 예문에서만 <명사>+<조사>+<동사>의 해석이 가능하고 출현빈도가 높은 [따뜻한]은 형용사로 품사부착이 가능 하다.

[따뜻]과 [하다]가 분리될 경우 다른 품사의 개념을 적용하는 것은 옳다. 그러나 [따뜻도 하다]의 형태를 근거로 [따뜻하다]의 형태소까지 <명사>+<형용사 파생접미사>로 해석할 필요는 없다고 판단된다.

위의 결과는 형태소분석 프로그램을 기계번역 프로그램으로 활용하게 될 경우에 번역어의 선택에도 큰 영향을 미칠 것이다.

또한 [한자어+하다]의 경우 예1)와 같이 대부분의 형태소가 [하다]와 분리 가능하다.

그러나 1음절 한자어의 경우 분리가능한 한자어는 그 수가 오히려 적다.

예1)공부하다 → 공부를 하다.
예2) 순서를 정하다/논하다.
예3) 마음이 약하다/강하다/순하다/족하다
예4) 그는 마음이 독하다.
cf: 독을 먹으면 죽는다.

위의 예2)와 예3)은 생산성이 없으니 하나의 품사 즉 동사 및 형용사로 품사처리 하는 것이 타당하지만 예4)의 [독]은 생산성이 있다고 보아야 할지 의문이다. [따뜻], [조용]의 처리는 분리되어 나타나는 예문에서만 명사처리 하면 해결될 문제이다.

2.6.2 적용규칙의 일관성

품사의 축약형의 처리 혹은 생략형의 처리는 모든 품사에 일관성 있게 적용되어야 한다. 예를 들어 표준안에서는 지정사[이다]의 생략형은 재현하지만 동사의 생략형은 재현하지 않는다.

(1) 지정사가 생략된 경우에는 지정사를 복원한다.(지침서 p.19)

(2) [하다]의 축약처리에서는 [하다]가 탈락하면서 두어절이 한 어절로 줄어드는 경우는 복원하지 않는다.(지침서 p.63 참조)

(3)별도의 기호"#"를 쓴다면 [하다]를 복원한다

예1) 학교다(학교/nc+이/co+다/ef)

예2) 먹어야겠다.

예2) 먹어야겠다.(먹/pv+어야/ec#하/px+겠/ep+다/ef)

위의 처리에는 문제점이 있다

1) 예2)와 같은 경우 어미 뒤에 선어말 어미가 나타타고 다시 어미가 뒤에나타난다.

2) 분리된 채로 나타나는 코파스가 존재한다. (출전 조선일보 93년도 기사)

1)내실을 찾아야 겠습니다.
2)자각과 반성이 있어야 겠습니다.
3)하나 하나 따져봐야 겠습니다.

또한 별도의 기호"#"를 쓴다면 [하다]를 복원한다고 지침서에 적혀 있지만, 코파스에서 찾아보면 분리되어 있는 경우가 있고 예외속에 다시 복원하기 위한 기호를 쓴다면 [이다]의 생략형도 기호를 사용해서 품사부착(예1-2))하거나 [하다]의 생략형에도 [하]를 기호 없이 재현(예2-3))해야 한다고 판단된다.

예1-2) 학교다(학교/nc#이/co+다/ef)
예2-2)먹어야겠다.(먹/pv+어야/ec#하/px+겠/ep+다/ef)
예1-3) 학교다(학교/nc+이/co+다/ef)
예2-3)먹어야겠다.(먹/pv+어야/ec+하/px+겠/ep+다/ef)

2.6.3 품사부착 형식의 문제

일본어의 대표적인 형태소 분석 시스템으로는 JUMAN과 CHASEN이 있다. 형태소 기준은 조금 다르나 출력형식은 같은 계층구조로 표시되어 있고 그 내용도 여러 정보를 동시에 병렬하는 방법을 택하고 있다. 다음은 JUMAN의 구체적인 형식과 품사부착 정보의 예이다.

형태소	발음정보	기본형	품사정보	활용정보	활용형태정보
お願い	(おねがい)	お願い	サ変名詞		
が	(が)	が	格助詞		
一つ	(ひとつ)	一つ	普通名詞		
あり	(あり)	ある	動詞	子音動詞ラ行	基本連用形
ます	(ます)	ます	動詞性接尾辞	動詞性接尾辞ま	基本形
が	(が)	が	逆語接続助詞		

위의 형식과 정보의 양을 비교해 볼 때 표준안의 형식과 내용에 좀 더 한국어의 특성을 살리는 품사부착 정보를 부가할 필요가 있다. 예를들면 발음, 교착, 굴적, 파생, 복합, 합성, 어원정보등을 들 수 있다.

예)그는 덧신을 신고 지붕으로 올라갔다.

표준안1)그/+는/ 덧/+신/+을/ 신/+고/ 지붕/+으로/ 올라가/+았/+다/./
제안1)그는/ 더씨늘/ 신꼬/ 지붕으로/ 올라가따/.
제안2) 집+웅(어원정보) 덧신(합성정보) 덧/+신/ 더/+ㅅ/+신/ 올라/르변칙(불규칙동사정보)

위의 제의를 포함하여, 과연 어떤 형식으로 품사부착을 하는 것이 보다 합리적이며 , 품사정보에 의한 보다 용이한 언어검색이 가능할 것인지는 충분한 시간과 토의를 거쳐 검토해 나가야 될 것이다.

3. 結論

기존의 기계번역 시스템의 출력형태를 바꾸어 형태소 분석대회에 임한 결과 기존의 프로그램의 문제점과 여러 해결방안을 강구할 수 있게 되었다. 명사 추출과 용언등의 정확률은 차후의 과제로 남겨놓고 어미류의 재현률이 높았던 이유로 파라다임형의 프로그램의 특성과 이점이 드러났다고 판단된다.

참고문헌

[1]강용희 [한일기계번역에 있어서의 오역 및 고찰] , 제8회 한글 및 한국어 정보처리 학술대회, pp351-366, 1996

[2] 姜龍熙 [韓日機械飜譯における助詞の誤譯の問題], 言語處理學會第3回年次大會發表論文集, pp.47-50, 1997(日本)

[3]강용희 [일본의 한일 기계 번역 시스템에 있어서의 오역과 그 언어환경], 제9회 한글 및 한국어 정보처리 학술대회, pp.303-310, 1997

[4]姜龍熙 [日本語と韓國語との言語の相異点と機械飜譯における問題点], AAMT Vol No22 April, pp.28-32,1998.(日本)

[5]松田純一,河野勝也 [構文ダイレクト方式による日韓機械飜譯システム],情報處理學會全國大會論文集, pp.139-140, 1993,(日本)

표준안의 품사기호와 명칭

1.s(기호)

2.f(외국어)

3.n(명사)

nc(일반명사)

nb(의존명사)

4.np(대명사)

5.nn(수사)

6.pv(동사)

7.pa(형용사)

8.px(보조용언)

9.co(지정사)

10.ma(부사)

mag(일반부사)

maj(접속부사)

11.mm(관형사)

12.ii(감탄사)

13.x(접사)

xp(접두사)

xs(접미사)

xsn(명사파생접미사)

xsv(동사파생접미사)

xsm(형용사파생접미사)

14.j(조사)

jc(격조사)

jx(보조사)

jj(접속조사)

jm(속격조사)

15.ep(선어말어미)

16.e(어말어미)

ef(종결어미)

ec(연결어미)

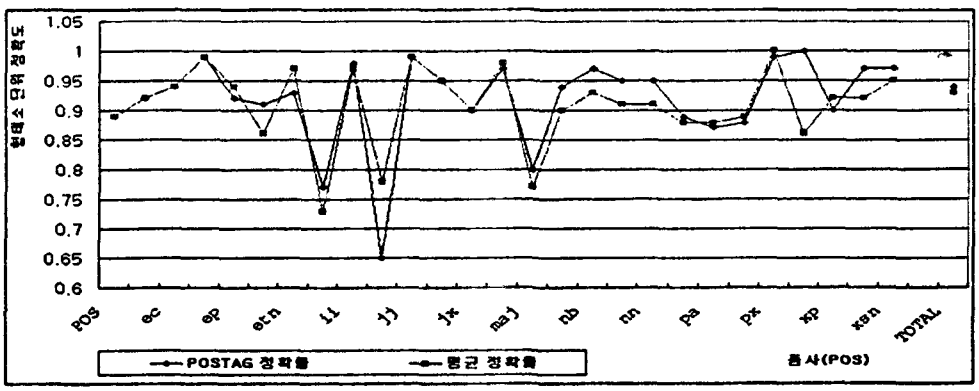
et(전성어미)

etn(명사형어미)

etm(관형사형어미)

정 오 표

p. 72-73 [그림 6], [표 9], [표 10] 의 올바른 모양



[그림 6] 품사별 형태소 단위 정확도 비교

정확도 (%)		
어절수	틀린 어절수	어절정확도
8000	682 개	91.48 %
형태소수	틀린 형태소수	형태소정확도
18063	764 개	95.77 %

[표 9] PtoS 태깅 정확도

	Mapping	Tagging
틀린 형태소수	298 개	466 개
틀린 에러 비율	1.65 %	2.58 %
상대 비율	39.01 %	60.99 %

[표 10] PtoS 매핑과 태깅 에러 비율

p. 106 제목오타

Korean/Japanese Maching Translation's

⇒ Korean/Japanese Machine Translation's