# 품사 태그 세트의 매핑을 이용한 한국어 품사 태거 (POSTAG) 이식 [1)]

김 준석, 심 준혁, 이 근배

포항공과대학교 컴퓨터공학과 자연어 처리 연구실

경북 포항시 남구 효자동 산 31 번지

# Porting POSTAG using Part-Of-Speech TagSet Mapping

**Junseok Kim, Junhyuk Shim, Geunbae Lee**

*Natural Language Processing Lab.*

*Dept. of Computer Science & Engineering, POSTECH*

*johan@nlp.postech.ac.kr, nikkie@nlp.postech.ac.kr, gblee@nlp.postech.ac.kr*

## 요 약

품사 태그세트 매핑은 서로 다른 품사 태그세트로 태깅되어 있는 대량의 코퍼스들로부터 정보를 얻고 또한 제공함을 통해 코퍼스의 재사용성(reusability)을 높이는데 유용하게 사용된다. 본 논문은 포항공대 자연언어처리 연구실의 자연언어처리 엔진(SKOPE)의 품사 태거(POS TAG)에서 사용되는 태그세트와 한국전자통신연구원의 표준 태그세트 간의 양방향 태그세트 매핑을 다룬다. 매핑을 통해 표준태그세트로 태깅된 코퍼스로부터 POSTAG 를 위한 대용량 학습자료를 얻고 POSTAG 가 두 가지 태그세트로 결과를 출력할 수 있다. 특히 한국어 태그세트 매핑에서 발생할 수 있는 여러 가지 문제점들, 즉 사전 표제어 차이 (형태소 분할 차이), 태그 할당 차이, 축약 처리 차이 등과 그것들의 기계적인 해결책을 살펴보고, 태그세트 매핑의 정확도를 측정하기 위해서 매핑 전과 후의 태깅 시스템의 정확도를 서로 비교함으로써 매핑의 정확도를 측정하는 실험을 수행하였다. 본 자동 매핑 방법을 반영한 POSTAG 는 제 1 회 형태소 분석기 평가 대회(MATEC'99)에 적용되어 성공적으로 사용되었다.

## 1. 서 론

최근에 다양한 목적의 한국어 처리를 위해서 많은 품사 태깅 시스템들이 개발되었다. 품사 태깅에 사용되는 기본 단위들의 집합을 품사 태그세트라고 하는데, 각 태깅 시스템들은 응용 목적에 적합한 서로 다른 문법 해석과 사전 정보를 바탕으로 고유의 품사 태그세트를 정의했다. 품사 태그세트를 기준으로 학습을 위한 대용량의 태깅된 코퍼스를 구축하는 과정은 한국어의 통계 기반 어휘 정보와 규칙 기반 어휘정보를 추출하는 연구의 기본이 되고 있다. 그 결과, 각 태깅 시스템 별로 태깅

1) 본 연구는 정보통신부 대학 기초 연구 (1998. 7 - 2000. 6) 연구비 지원으로 수행되었습니다.

된 결과를 전문가의 지시에 따라 반자동으로 수정하여 대용량 태깅된 코퍼스를 구축하고 있으나 그 과정이 매우 느리며, 많은 비용이 요구되는 문제점이 있다[7].

품사 태그세트 매핑은 서로 다른 두 태그세트의 정의와 분류의 차이에 대한 매핑 관계를 설정하여 변환시킬 수 있게 함으로써 서로의 태깅된 코퍼스를 공유하는 기술이다. 품사 태그세트 매핑에 따른 장점은 다음과 같다[3,4].

- **코퍼스의 재사용성(Reusability)** : 품사 태깅 시스템 간의 태그세트 매핑은 태깅된 코퍼스 정보를 공유하고, 각 품사태거의 태깅 결과를 다른 품사태거에 배포하여 코퍼스의 재사용성(reusability)을 높이며, 대용량의 태깅된 코퍼스를 쉽게 구축할 수 있게 해준다.

- **표준 태그세트의 기준 강화** : 각 태깅 시스템의 태깅된 결과를 태그세트 매핑을 이용해 표준의 태깅된 코퍼스로 수렴하는 과정에서 태깅 기준에 관한 의견을 수렴함으로써 표준 태그세트의 기준을 타당하고 명확하게 강화할 수 있다.

- **품사 태깅 시스템 보강과 평가** : 현재의 품사 태깅 시스템의 사전 보강 및 태그세트 비교 작업을 통해서 시스템에 대한 객관적인 평가가 가능하다.

다른 태그세트로 태깅된 코퍼스를 이용하기 위해서 그 태그세트 정의 기준을 따라 시스템 전체를 바꾸는 것도 고려할 수 있는데 이 경우 태깅 기준의 변화에 따라 기존에 연구되어 온 많은 시스템들에 대한 수정이 불가피하다. 따라서, 기존의 시스템을 유지하면서 동시에 다른 태그세트로 태깅된 정보를 이용하고 자료 제공을 위해서는 태그세트 매핑 방식이 유용하다.

본 논문에서는 포항공대 SKOPE 시스템 내의 태깅 시스템인 POSTAG 에서 사용되는 태그 세트와 표준 태그세트 간의 매핑을 다룬다. 매핑의 주요 고려 사항 으로는 세그먼트 차이(사전 표제어 차이), 원형과 이형태의 차이, 품사 차이, 그리고 형태소 원형 복원 문제 및 축약 차이 처리를 들 수 있다. 본 논문에서는 점진적인 방식의 접근을 시도하였는데, 우선, 한 어절 내에서 품사, 원형, 세그먼트 차이 및 출력 형식과 문장 구분 문제를 차례대로 해결하고 어절간에 생기는 매핑 문제의 해결을 도모한다. 또한 피드백 과정을 통해 매핑 오류를 감소시키는 방식을 이용한다. 태그세트 매핑의 정확도를 측정하기 위해서 매핑하기 전의 정확도와 매핑 후의 정확도를 서로 비교함으로써 매핑의 정확도를 측정하는 실험을 수행 하였다. 본 자동 매핑 방법을 반영한 POSTAG 는 MATEC'99[2]에 적용되어 성공적으로 사용되었다. 본 논문의 구성은 다음과 같다. 2 장에서는 기존에 태그세트 매핑에 대한 연구를 살펴보도록 한다. 3 장에서는 품사 태그세트 매핑 방법론과 알고리즘 대해 소개한다. 4 장에서는 실험 및 결과를 소개하며 마지막으로 5 장에서 결론을 맺는다.

## 2. 기존의 연구

태그세트 매핑은 단일 언어 내에서의 품사 태깅 결과를 태그세트를 기준으로 분석하는 Mono-lingua TagSet Mapping 과 다국 언어간의 품사 태깅 결과를 태그세트를 기준으로 분석하는 Multi-lingua TagSet Mapping 이 있다. 최근에는 단일어 내에서 태그세트 간의 품사 하위 분류 기준 설정 과정에서 발생하는 세그먼트의 차이를 문법적인 매핑 규칙 리스트로 만드는 방법[5,6]과 대용량의 Parallel-Tagged-Corpus 에서 빈도수에 따라 자동으로 매핑 관계를 추출하는 방법[4]이 연구되어 왔다. 또, 다국어 간의 문법을 반영한 태그세트 표준의 구축에 관한 연구[3]도 매핑과 관련되어 국외에서 활발하게 진행중이다.

영어 태깅 시스템 AUTASYS 는 영어권에서 많이 사용되는

---

2) 한국전자통신연구원에서 주관한 제 1 회 형태소분석 , 태깅, 명사추출 대회

LOB(Lancaster/Oslo Bergen), ICE (International Corpus of English) 두 가지의 태그세트로 태깅 결과를 출력할 수 있다[5,6]. 매핑 방식은 일반적인 영어 단어(Word) 분류에 대해서는 직접적인 단어간의 매핑을 하고, 품사의 하위 분류 기준의 차이 발생하는 보다 세부적인 문제들은 예외사전을 이용하여 매핑을 수행하였다. 그 결과, LOB 태그세트로 태깅된 결과를 매핑을 이용해서 ICE 태그세트 태깅된 결과를 만들어 내는 과정에서 96% - 97%의 태그세트 매핑 성공률을 보였다.

한편, Parallel-Annotated 코퍼스로 태그세트 매핑 규칙을 자동으로 찾아주는 연구를 들 수 있다[4]. 이 연구는 AMALGAM Project(Automatic Mapping Among Lexico-Grammatical Annotation Models Project)의 일환으로 진행되었는데, 하나의 원시코퍼스를 SEC (Spoken English Corpus) 태그세트와 ICE 태그세트 두 가지로 태깅된 Parallel 코퍼스로부터 자동으로 매핑 규칙을 추출해 내는 방법을 사용하였다. 그러나 Parallel 코퍼스를 생성하는데 비용이 많이 들고, 영어와는 달리 형태소에 대한 많은 세그먼트의 차이가 존재하는 한국어에서는 Parallel 코퍼스로부터 매핑 규칙을 찾기 위해서 선행되어야 하는 형태소 alignment 문제 해결이 쉽지 않다는 문제가 있다.

한편 유럽 언어들에 대한 표준 태그세트 제작을 위한 EAGLES(Expert Advisory Group on Language Engineering Standard) 프로젝트에서는 Specification Language를 이용하여 수동으로 작성된 매핑 Rule을 이용하여 Upenn 태그세트와 SUSANNE 태그세트 간의 매핑을 다루는데, 이 연구에서는 매핑에서 발생하는 에러(noise)들을 자동으로 report 해 주는 Tool 을 제작하였다[3].

위와 같은 연구에도 불구하고 한국어 내에서의 태그세트 매핑에 관한 연구는 아직 미비하다. 이러한 필요성에 따라 본 논문은 한국어에서 서로 다른 태그세트로 태깅된 코퍼스간의 변환 알고리즘을 제안한다.

## 3. 태그세트 매핑 방법

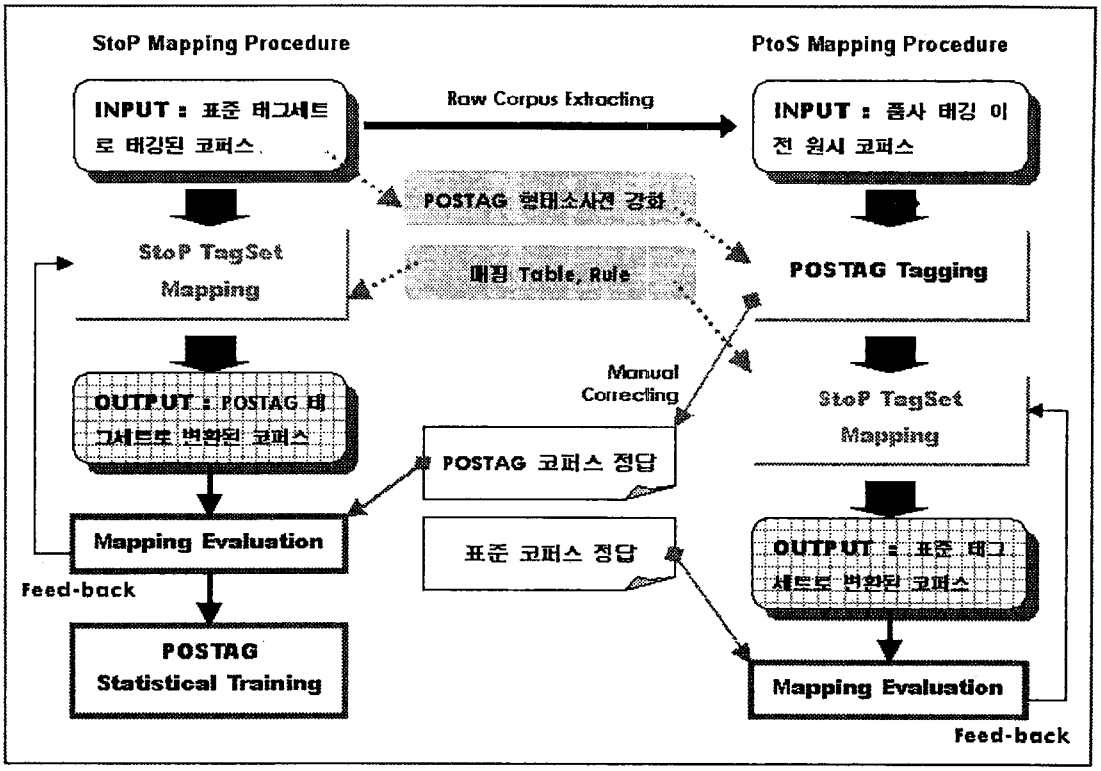
### 3-1. 태그세트 소개

SKOPE 시스템에서 사용되는 태그 개수는 총 97 개이며 이 중에서 POSTAG 시스템 즉, 태깅에서 사용되는 태그의 개수는 41 개로 계층적(hierarchical)인 태그세트 구조를 가진다. 이에 반해 표준 태그세트는 태그의 수가 총 27 개인데 비계층적인 구조로 이루어져 있다[2]. 태그세트의 크기는 POSTAG 태그세트가 크지만, 표[1]에서 보는 것처럼 태그의 종류가 많다고 단순한 포함관계는 아니다. 표준 태그세트를 제작할 때에 POSTAG 태그세트가 고려대상 중 하나였기 때문에 큰 차이를 보이지는 않는다.

| 표준 품사 tag set | POSTAG tag set |
|---|---|
| s | s, s', s'' su so |
| f | f |
| nc | MC/MP |
| nb | MD |
| np | T |
| nn | S |
| pv | DR,DI,E |
| pa | HR,HI,LE |
| px | b |
| co | I |
| Mag/maj | B |
| mm | G |
| ii | K |
| xp | + |
| xsn | - |
| xsv | y |
| xsm | y |
| jc | JC |
| ji | JS |
| jx | JO |
| jm | JO |
| ep | eGS |
| ef | eGE |
| ec | eCC, eCNDI, eCNDC, eCNB, eCNMG |
| etn | eCNMM |
| etm | eCNMG |

**표(1) 태그세트 비교 테이블**

표[1]에서 **1:1** 이나 **N:1** 의 태그세트 관계는 매핑에서 크게 문제가 되지 않으나 pv(동사)를 DR(규칙동사), DI(불규칙동사), E(존재사)로의 1:3 의 태그세트 관계나 표준의 ec(연결어미)에서의 eCC(연결어미), eCNDI(보조적연결어미), eCNDC (인용어미), eCNB(부사형 전성어미), eCNMG(관형사형 전성어미)로의 1:5 의 태그세트관계등과 같은 1:n 태그세트 관계는 매핑에 어려움을 가진다.




[그림 1] 매핑 수행 방법론

### 3-2. StoP & PtoS 매핑 구조

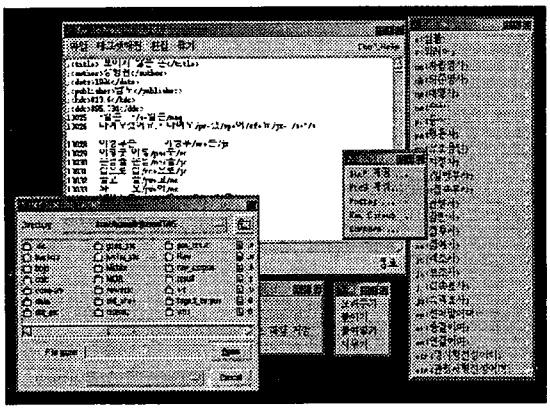
매핑은 크게 표준(ETRI Standard) 태그세트에서 POSTAG 태그세트로의 StoP 매핑과 POSTAG 태그세트에서 표준 태그세트로의 PtoS 로 나눈다. 표준 태그세트로 태깅된 코퍼스를 POSTAG 태그세트로 태깅된 코퍼스로 만들어서 POSTAG 시스템에서 확률학습을 하기 위해 StoP 매핑을 한다. 코퍼스로부터 코퍼스로의 매핑이므로 StoP 매핑은 off-line 으로 수행한다. 한편 PtoS 매핑은 POSTAG 에 출력 옵션을 주어 양쪽 태그세트로 출력이 가능하게 하기 위해서 online 매핑을 시도한다. 이때, POSTAG 내부의 태깅 결과를 담아 두는 graph 자료구조를 입력으로 받아서 PtoS 매핑을 수행한다(그림 1 참조).

### 3-3. 매핑 수행방법론

가장 먼저 매핑을 위해서 각 태그세트에 대한 특성 파악을 하였다. POSTAG 태그세트와 표준 품사 태그세트에 대한 지침서를 통해서 품사 태그세트 설정에 대한 기준에 대한 학습을 하였다. 그 연구로부터 기본적인 매핑 규칙과 예외 사항들에 대한 매핑 테이블을 제작한다. 또한, 표준 품사태그세트로 태깅된 코퍼스로부터 각 품사별로 형태소를 추출하여 우리쪽 사전을 보강하는 작업을 병행하였다. 다음은 매핑 규칙과 테이블을 이용하여 표준 태그세트로 태깅된 코퍼스로부터 POSTAG 태그세트로 매핑을 수행한 결과를 POSTAG 시스템의 결과를 수동으로 검사하여 만든 정답문서와 자동으로 비교하여 발견된 오류를 자동으로 알려주는 Mapping Tool([그림 2] 참조)을 사용하여 매핑 규칙과 테이블을 보강 해 주는 피드백(feedback)과정을 반복한다([그림 1] 참조).

PtoS 매핑은 StoP 매핑과 유사한 방식을 사용하였다. 우선, 표준 태그세트로 태깅된 코퍼스의 Raw Text 를 POSTAG 시스템의 입력으로 하여 POSTAG 태그세트로

결과를 메모리에 graph 형태로 얻게 된다. 매핑 규칙과 테이블을 사용하여 POSTAG 태깅 결과 graph 로부터 표준태그세트로 매핑을 한다. 매핑된 결과를 표준태그세트로 태깅된 정답 코퍼스와 비교를 통해 오류들에 대해서는 역시 Feedback 과정을 반복한다. [그림 3]은 **"준혁이가 흔들어 보였다."**라는 문장에 대한 POSTAG 의 태깅된 코퍼스 결과를 표준으로 변환시킨 결과이다. POSTAG 에서는 형태소 단위로 품사, 원형 이형태를 출력해주는 반면에 표준에서는 어절 단위 표층정보와 어절 내 형태소 분석을 품사, 원형 단위로 출력한다.



(그림 2) TagSet Mapping Tool

| POSTAG Format | ETRI Standard Format |
|---|---|
| s<문장시작>(I)<br>MP<준혁>(준혁)<br>-<이>(이)<br>jC<는>(는)<br>DI<흔들>(흔들)<br>eCNB<어>(어)<br>s<ㅅ>(ㅅ)<br>DR<보이>(보이)<br>eGS<었>(ㅆ)<br>eGE<다>(다)<br>s.<.>()[등].]<br>s<문장끝>() | 준혁이는 준혁/nc+이/xsn+는/jc<br>흔들어 흔들/pv+어/ec<br>보았다. 보이/px+았/ep+다/ef+./s |

(그림 3) 태깅된 코퍼스의 변환 결과

### 3-4 매핑 알고리즘

매핑은 한 문장에 대해서 처리하고, 한 문장에서 각 어절을 구성하는 형태소의 품사에 따라 매핑을 수행한다. StoP 매핑을 예로 들어 표준 태그세트로 태깅된 문장의 임의의 한 형태소 X 의 품사가 'pv'(동사) 일 때 예를 들어 보겠다. POSTAG 태그세트로 태깅된 코퍼스는 품사<주형태>(이형태)로 구성되므로 품사 결정을 해주고, 표층정보를 이용해서 이형태를 복원하면 된다.

```
Procedure_Map_pa(surface, S_MOR(X))
{
    P_MOR(X) = P_MOR(X);
    P_POS(X) = Pattern(Last_S_MOR(X));
    P_ALLO(X)=Restore_ALLO(surface, S_MOR(X))
}
```

우선 입력으로 표층(surface) 정보와 형태소 X 의 주형태(S_MOR(X))이 들어간다. 형태소 X 의 원형은 그대로 이용하고, 원형의 마지막 음절(Last_S_MOR(X))을 이용하여 규칙동사(DR) 인지 불규칙동사(DI) 인가를 음절 패턴사전을 이용하여 결정한다.[1] 마지막으로 표층정보와 원형정보를 이용하여 이형태 복원과정을 수행하면 매핑이 끝난다. 한편, 세그먼트가 달라지는 경우는 가장 먼저 세그먼트 문제부터 해결해야 한다. 예를 들어 표준 태그에서는 "청소를 하던(던/etm:관형사형 전성머미) 사람"에서 '던' 을 관형사형 전성어미로 보는데 POSTAG 태그세트에서는 '던' 을 eGS<더>(더) + eCNMG<ㄴ>(ㄴ) 으로 태깅 한다(여기서 eGS 는 시제선어말어미 이고, eCNMG 는 관형사형 전성어미 임). 이와 같이 세그먼트의 차이를 보이는 형태소들은 매핑 테이블에 그 정보를 저장하고 있고, 매핑 시에 테이블을 참조 한다. 테이블의 구성은 [표 2]와 같다.

(표 2) 매핑 테이블

| key | Seg | Mapping 정보 |
|---|---|---|
| 던/etm | 2 | eGS<더>(더) + eCNMG<ㄴ>(ㄴ) |

Seg 정보는 세그먼트의 개수를 의미하는 데 0 ~ 3 의 값을 가진다. "시었겠/ep" 의 경우는 "eGS<시>(시)+eGS<었>(었)+eGS<겠>(겠)" 과 같이 매핑 되므로 3 이라는 Seg 값을 가진다. 한편, "낡아빠진 운동화"에서 표준 태그세트 기준에서는 빠지/px(보조용언)으로 태깅하는데 반해 POSTAG 에서는 "DR<낡아빠지>(낡아빠지)"(규칙동사)로 태깅하므로 매핑 시에 "빠지/px"는 같은 어절내의 바로 앞의 형태소가 용언이면 그 용언에 붙어서 출력이 되어야 하므로 Seg 값이 0 이 된다. PtoS 매핑 알고리즘은 기본 구조는 StoP 와 유사한데, 주요 고려 대상이 표준과 POSTAG 의 태그세트 관계가 N:1 인 것이 차이점이다.

## 4. 실험 및 결과 분석

본 논문에서 제안한 알고리즘을 구현하여 두 가지의 매핑 유형을 MATEC'99 의 학습용 코퍼스와 테스트 코퍼스에 대해 실험하였다. StoP 매핑은 실험 문장에 대한 정답 파일을 수작업을 통해 구축하여 매핑 결과와 비교하였고, PtoS 매핑은 MATEC'99 에서 제공한 평가대회 정답과 매핑 결과를 비교하였다.

### 4.1. StoP Mapping 결과

StoP 매핑은 정답의 태깅된 코퍼스를 Off-line 으로 매핑하므로, 실험에서 매핑의 정확도를 측정하였다. 이 실험을 위하여 MATEC'99 를 위해 주어진 정답의 태깅된 코퍼스 중 소설 7178 어절에 대하여 수작업으로 정답 파일을 구축하였다.

(표 3) StoP 매핑 정확도

| 정확도 (%) | | |
|---|---|---|
| 어절수 | 틀린 어절수 | 어절정확도 |
| 7178 | 119 개 | 98.40 % |
| 형태소수 | 틀린 형태소수 | 형태소정확도 |
| 18744 | 178 개 | 99.05 % |

이 과정에서 나타난 매핑 에러의 정확도를 평가한 결과, 위의 [표 3]의 내용과 같이 형태소 단위 정확도가 99.05% 이었다.

### 4.2. PtoS Mapping 결과

PtoS 매핑은 POSTAG 의 형태소 분석 및 태깅의 결과를 On-line 으로 받아 매핑 하므로, 실험에서 태깅 정확도와 매핑 정확도를 측정하였고, 매핑에서 발생하는 에러와 태깅에서 발생하는 에러의 비율을 비교하였다. PtoS 매핑과 태깅 실험에 사용된 문장은 MATEC'99 테스트 코퍼스 소설 9012 어절, 21051 개 형태소이며, 결과는 다음과 같다. 태그세트 매핑을 포함한 품사 태깅 결과, 형태소 단위 정확도가 94.86% 이었고([표 4],[표 5]참조), PtoS 매핑에서 발생하는 에러를 제외한 순수한 품사 태깅 정확도는 96.82 % 이었다([표 6] 참조).

(표 4) PtoS 태깅 정확도

| 정확도 (%) | | |
|---|---|---|
| 어절수 | 틀린 어절수 | 어절정확도 |
| 9012 | 953 개 | 89.43 % |
| 형태소수 | 틀린 형태소수 | 형태소정확도 |
| 21051 | 1082 개 | 94.86 % |

(표 5) PtoS 매핑과 태깅 에러 비율

| | Mapping | Tagging |
|---|---|---|
| 틀린 형태소수 | 424 개 | 658 개 |
| 틀린 에러 비율 | 2.01 % | 3.13 % |
| 상대 비율 | 39.15 % | 60.85 % |

(표 6) POSTAG 결과비교

| | Mapping 반영 | Mapping 미반영 |
|---|---|---|
| 정확도 (%) | 94.86 % | 96.82 % |

### 4.3 Mapping 오류 분석

위의 실험 결과에서 나타나는 태그세트 매핑 에러는 예외 처리가 반영되지 않은 에러들과 품사 태깅 레벨의

매핑에서 다루기 힘든 예외들로 나뉜다. 전자의 경우는 지속적인 매핑 정보의 추가를 통해 반영될 수 있으나 후자의 경우는 다음의 경우에서처럼 반영이 힘들다.

예를 들면, 접속 조사가 공동격 조사로 사용되는 (예 1)과 (예 2)의 경우를 보는 것처럼 표준 태그세트에서는 격조사 '와/과, 하고, 랑' 뒤에 '만나다, 부딪치다, 헤어지다, 사귀다, 함께, 같이, 동행하다.' 등의 단어가 나타날 경우에 이들을 공동격조사로 구분하고 있다. 하지만, 이러한 분석은 어휘들의 공기 정보를 알아야 분석이 가능하다. 반면에 POSTAG 시스템은 두 경우 모두 jO(기타 조사)로 태깅 하므로 PtoS 매핑이 어렵게 하고 정확도를 떨어뜨리는 원인이 된다.

**(예 1)** 사과와(와/jj) 배를 먹었다.
**(예 2)** 그 애와(와/jc) 싸우지 말아라.

### 4. 결 론

실험 결과를 통해서 태그세트 매핑을 통해서 하나의 태그세트로 태깅된 문서를 다른 태그세트로 태깅된 문서로 출력할 수 있다는 것을 보였다. 그 결과, 표준 태깅된 코퍼스와의 매핑에서 양방향으로 98%의 성능을 얻었다. 한편, 이 과정을 통해서 대규모 확률 학습 코퍼스를 얻을 수 있었고, 또한 사전 강화작업과 MATEC'99 대회를 통하여 POSTAG 시스템의 객관적 평가 또한 수행할 수 있었다. 다른 학교나 연구소에서도 표준 태그세트로의 매핑을 고려하여 코퍼스를 제작한다면 표준태그세트는 표준 그 자체로서의 의미 뿐만 아니라 여러 가지 태그세트로 태깅되어 있는 코퍼스들간의 매핑을 위한 'interlingua' 가 될 수 있을 것이다. 또한 하나의 언어 내에서 품사태그, 세그먼트등 서로 다른 feature 를 가지는 코퍼스의 변환을 매핑을 통해서 해결 가능함을 보였으므로, 서로 다른 언어에서 사용되는 태그 세트간의 매핑으로의 확장을 고려하여 볼 만할 것이다.

## 참고 문헌


[1] 차정원, "일반화된 미등록어 처리를 이용한 혼합형 품사 태거", 석사학위 논문, 포항공과대학교 컴퓨터공학과, 1998.

[2] 한국전자통신연구원 컴퓨터·소프트웨어 기술 연구소 지식정보연구부, "품사 부착 말뭉치 구축 지침서", 1999, [URL : http://aladin.etri.re.kr/~nlu/STANDARD/]

[3] S. TEUFEL 1995a, "A support tool for tagset mapping," *Proceedings of Special Interest Group for linguistic data and corpus-based approaches to NLP 1995. Workshop in cooperation with EACL 95.*

[4] John Hughes, Clive Souter and Eric Atwell, 1994., "Automatic Extraction of TagSet Mapping from Parallel-Annotation Corpora," In *Center for the Computer Analysis of Langage And Speech School of Computer Studies,* Leeds University..

[5] Fang, A.C. 1996. AUTASYS : Automatic Tagging and Cross-Tagset Mapping. In *Comparing English Worldwide: The International Corpus of English*, ed. by S. Greenbaum. Oxford University Press. 110-124.

[6] Fang, A.C. and G. Nelson, 1994. "Tagging the SEU Corpus: a LOB to ICE Experiment Using AUTASYS." In *Oxford Literary and Linguistic Computing*, 9(2) 189-194.

[7] Gaoffrey Leech and Roger Garside. 1991. "Running a grammar factory: The production of syntactically-annotated corpora or 'treebanks'." In Stig Johannsson and Anna-Brita Stenstrom(eds.), *English Computer Corpora*. Berlin : Mouton de Gruyter. 15-32.