통계와 규칙을 이용한 강인한 품사 태거 [1)]

심 준혁, 김 준석, 차 정원, 이 근배

포항공과대학교 컴퓨터공학과 자연어 처리 연구실

경북 포항시 남구 효자동 산 31 번지

Robust Part-of-Speech Tagger using Statistical and Rule-based Approach

Junhyuk Shim, Junseok Kim , Jongwon Cha, Geunbae Lee

Natural Language Processing Lab.

Dept. of Computer Science & Engineering, POSTECH

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

요 약

품사 태깅은 자연 언어 처리의 가장 기본이 되는 부분으로 상위 자연 언어 처리 부분인 구문 분석, 의미 분석의 전처리로 사용되고, 독립된 응용으로 언어의 정보를 추출하거나 정보 검색 등의 응용에 사용되어 진다. 품사 태깅은 크게 통계에 기반한 방법, 규칙에 기반한 방법, 이 둘을 모두 이용하는 혼합형 방법 등으로 나누어 연구되고 있다. 포항공대 자연언어처리 연구실의 자연 언어 처리 엔진(SKOPE)의 품사 태깅 시스템 POSTAG 는 미등록어 추정이 강화된 혼합형 품사 태깅 시스템이다. 본 시스템은 형태소 분석기, 통계적 품사 태거, 에러 수정 규칙 후처리기로 구성되어 있다. 이들은 각각 단순히 직렬 연결되어 있는 것이 아니라 형태소 접속 테이블을 기준으로 분석 과정에서 형태소 접속 그래프를 생성하고 처리하면서 상호 밀접한 연관을 가진다. 그리고, 미등록어용 패턴사전에 의해 등록어와 동일한 방법으로 미등록어를 처리함으로써 효율적이고 강건한 품사 태깅을 한다. 한편, POSTAG 에서 사용되는 태그세트와 한국전자통신연구원(ETRI)의 표준 태그세트 간에 양방향으로 태그세트 매핑을 함으로써, 표준 태그세트로 태깅된 코퍼스로부터 POSTAG 를 위한 대용량 학습자료를 얻고 POSTAG 에서 두 가지 태그세트로 품사 태깅 결과 출력이 가능하다. 본 시스템은 MATEC '99[2)]에서 제공된 30000 어절에 대하여 표준 태그세트로 출력한 결과 95%의 형태소 단위 정확률을 보였으며, 태그세트 매핑을 제외한 POSTAG 의 품사 태깅 결과 97%의 정확률을 보였다.

[1)] 본 연구는 정보통신부 대학 기술 (1998. 7 - 2000. 6.) 지원으로 수행되었음.

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

1. 서 론

한 문장에서 단어는 문맥에 따라 다른 품사를 가진다. 그 문장에 가장 적당한 품사를 선택하는 과정을 품사 태깅이라 한다. 품사 태깅은 구문분석, 의미분석등에 사용되어 자연어 처리의 초기 단계로서 중요한 역할을 한다. 품사 태깅 과정은 크게 단어에 대하여 가능한 모든 품사를 생성하는 '모호성을 생성하는 부분'과 여러 품사 중에서 가장 적절한 품사를 선택하는 '모호성을 해소하는 부분'으로 구성된다*([11])*.

한국어는 교착어로 형태소 단위의 품사 태깅을 위하여 다양한 결합 후보를 생성하는 형태소 분석 과정을 거쳐야 한다. 이처럼 형태소 분석을 거치는 과정이 모호성을 생성하는 부분에 해당한다. 품사 태깅은 문장의 모호성을 해소하는 과정으로 크게 통계에 기반한 방법 *([2],[12],[13])*, 규칙에 기반한 방법*([1])*, 이 둘을 모두 이용하는 혼합형 방법*([6],[11],[14])* 등이 있다.

포항공대 자연언어처리 연구실의 자연언어처리 엔진(SKOPE)의 품사 태깅 시스템 POSTAG 는 매우 강화된 미등록어 추정을 이용한 혼합형 품사 태깅 시스템이다. 본 시스템은 형태소 분석기, 통계적 품사 태거, 에러 수정 규칙 후처리기로 구성되어 있다. 이들은 각각 단순히 직렬 연결되어 있는 것이 아니라 형태소 접속 테이블을 기준으로 분석 과정에서 형태소 접속 그래프를 생성하고 처리하면서 상호 밀접한 연관을 가진다. 또, 미등록어용 패턴사전에 의해 등록어와 동일한 방법으로 미등록어를 처리함으로써 효율적이고 강건한 품사 태깅을 한다*([14])*.

POSTAG 의 형태소 분석기는 형태소 사전과 다어절어 사전, 그리고 형태소 패턴 사전을 이용하여 등록어와 미등록어 형태소를 동일한 방법으로 분석한다. 이 과정에서 형태소 원형을 복원하며, 접속 검사표를 이용하여 품사간 접속을 검사한다. 접속이 이루어진 형태소들은 부분적인 형태소 접속 그래프를 형성하고, 이 그래프 노드들의 학습된 확률 정보를 이용하여 Viterbi 탐색 결과 최적의 품사열을 결정한다. 이때 미등록어의 확률 정보는 음절 Tri-gram 을 이용하여 구해진다. 한 문장에 대하여 최적의 품사열이 정해지면 에러수정 후처리기가 통계 정보 태거의 오류들을 올바르게 수정하여 최종적으로 가장 올바른 결과를 출력한다.

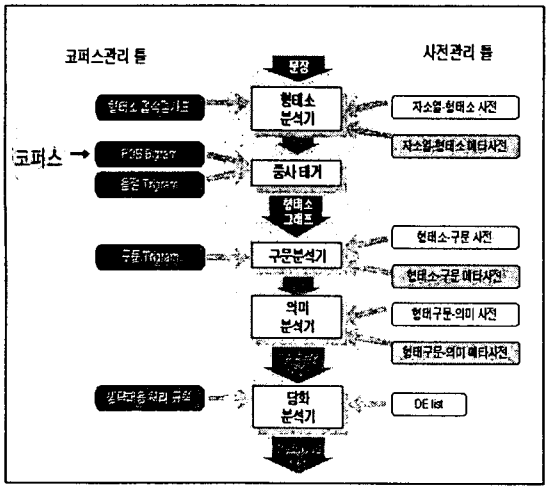
한편, POSTAG 에서 사용되는 태그세트와 한국전자통신연구원의 표준 태그세트*([15])*간의 양방향 태그세트 매핑을 통해 표준태그세트로 태깅된 코퍼스로부터 POSTAG 를 위한 대용량 학습자료를 얻었고 POSTAG 가 두 가지 태그세트로 결과를 출력할 수 있도록 하였다. 표준 태그세트와 POSTAG 태그세트 간의 매핑에서는 서로 다른 태깅 기준과 태그세트의 차이로 인해 몇 가지 해결해야 할 문제점들을 지닌다. 우선, 두 태그세트는 서로 다른 형태소 세그먼트를 가진다. 따라서 매핑에서 세그먼트 문제의 해결은 가장 우선적으로 해결해야 할 문제이다. 서로 다른 품사 태그세트 체계로 인한 품사 태그들 간의 매핑 문제 역시 해결해야 한다. 또, 표준 태그세트에서는 형태소의 원형과 품사만을 출력하지만, POSTAG 태그세트는 형태소의 원형과 이형태, 그리고 품사를 출력해 주므로, 형태소 이형태 복원 문제 또한 해결해야 한다*([10])*.

본 시스템은 MATEC '99 에서 제공된 30000 어절에 대하여 표준 태그세트로 출력한 결과 형태소 단위 정확률 95%의 성능을 보였으며, 태그세트 매핑을 제외한 원래 POSTAG 의 품사 태깅 결과 형태소 단위 정확률 97%의 성능을 보였다. 특히 명사 추출기 분야에서는 형태소 단위 정확률 95%의 우수한 성능을 보여 정보 검색 시스템으로의 응용과정에서 본 시스템의 우수함을 입증하였다. 본 논문의 구성은 다음과 같다. 2 장에서는 형태소 접속 그래프를 이용한 형태소분석 및 태깅에 관하여 기술한다. 3 장에서는 태깅 과정에서의 미등록어 추정 모듈에 대하여 설명한다. 4 장에서는 POSTAG 태그세트와 표준 태그세트 간의 매핑 방법론을 소개한다. 5 장에서는 실험 및 평가 결과를 분석하며 마지막으로 6 장에서 결론을 맺는다.

2. POSTAG : 형태소 그래프를 이용한 형태소 분석 및 태깅

2.1. 자연어처리 엔진(SKOPE)

자연어 처리 엔진은 자연어 처리를 수행하는 핵심적인 소프트웨어 프로그램으로 형태소 분석기 및 품사 태거, 구문분석기, 의미분석기 등의 3 단계 모듈로 나누어진다. 본 연구실에서 개발한 한국어를 위한 대용량 사전을 가진 자연어 처리엔진 SKOPE (Standard KOrean Processing Engine)는 [그림 1]과 같은 구조를 가졌다. SKOPE 는 크게 품사 태깅과 결합된 형태소분석기 모듈과 구문분석기와 결합된 의미분석기 모듈의 두 부분으로 나뉘어진다. 형태소분석기 모듈과 구문분석기 모듈은 하나의 형태소 열이 아닌 형태소 그래프를 사용하여 연결된다.



[그림 1] 자연어 처리 엔진 SKOPE

형태소 분석과 품사 태깅은 형태소 패턴 사전을 이용하는 일반적인 미등록어 추정이 가능하고 통계적 방법과 규칙을 이용한 방법을 혼합하여 입력된 한국어 문장을 강건하고 정확하게 처리한다. K-CCG(Korean-Combinatory Categorical Grammar)를 이용하는 구문분석은 무제한의 문장에 대하여 강건한 구문분석이 가능하다. 마지막으로 의미분석은 구문분석의 결과로 나온 문법적으로 맞는 문장이 한국어 내에서 어떠한 의미를 갖는지를 해석한다.

현재는 형태소 분석기의 전처리로 띄어쓰기 및 철자 오류 교정을 반영하여 사용자의 타이핑 오류가 포함된 문장들도 문제없이 처리할 수 있도록 연구하고 있으며, 보다 많은 학습 말뭉치를 활용하여 태깅의 정확도를 높이고 있다. 구문분석기는 강건함을 유지하면서 한국어에서 문법적으로 올바른 모든 문장에 대한 구문분석이 가능하도록 연구하고 있다. 한편, Tree Bank 구축을 통하여 구문 분석기의 학습으로 정확도를 향상시킴과 동시에 부분 구문 분석 기법을 도입하여 전체 문장의 구문 분석에 드는 과도한 부하를 줄일 수 있는 방안을 연구하고 있다. 의미 분석기에서는 의미 중의성 해소 기법을 도입하여 의미 제약과 의미 확률을 학습하여 의미 분석기의 성능을 향상시키고, 생략과 대용의 정확한 복원을 통하여 사용자가 말한 문장의 정확한 의도를 파악하여 응용 시스템에서 그러한 정보를 이용할 수 있도록 하는 시스템을 개발 중에 있다.

2.2. POSTAG 개요

POSTAG 는 본 연구실에서 제작한 형태소 분석 및 품사 태깅 시스템이다. POSTAG 는 음소, 음절, 어절 단위의 확률 정보를 기반으로 통계적 접근 방법을 기본으로 품사 태깅을 한다. 한편, 품사 태깅 결과의 에러들을 입력으로 받아 어절간 형태소 관계를 분석한 결과를 학습한 규칙들을 활용하는 에러 수정 후처리 방법에 의해 품사 태깅을 한다. 따라서, POSTAG 는 형태소 분석과 품사 태깅의 효율성을 증대시킨 혼합형 품사 태깅 시스템이다. 그리고, 패턴 사전을 기반으로 미등록어 추정을 함으로써 실제 무제한의 텍스트에서 동작하는 강건성을 가진 품사 태거이다.

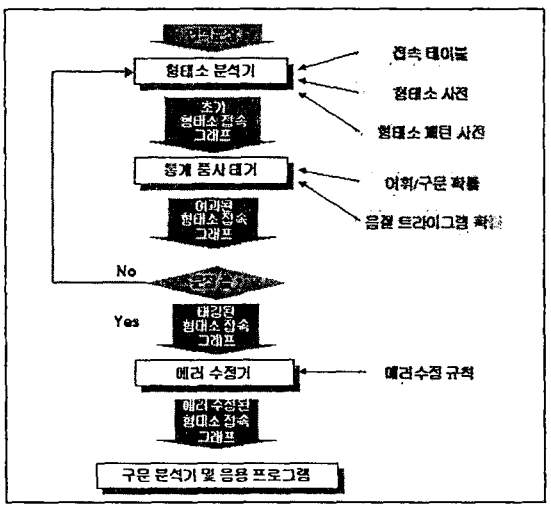
POSTAG 는 현재 복합 명사 합성과 분할을 이용한 명사 추출기에 응용되어 정보 검색 분야에 적용되고 있고, 구문 분석기, 언어의 중의성 해소 연구, 언어 모델 제작과 자연어 질의 시스템 구축에 기반 시스템으로 이용되고 있다.

POSTAG 의 사전은 형태소 사전, 다어절어(복합어) 사전, 패턴 사전으로 구성되어 있다.

2.2.1. POSTAG 의 특징

POSTAG 는 자연어 문장을 입력으로 받아 형태소 분석

기와 품사 태거가 동일한 분석에서 문장 입력 순서에 따라 순방향으로 분석한 후, 에러 수정 후처리 과정을 거쳐 결과 그래프와 Tagged Corpus 를 출력한다([그림 2] 참조).



[그림 2] 품사 태깅 시스템 POSTAG

형태소 분석기에서는 자소열-형태소 사전과 자소열-형태소 메타 사전 (패턴사전) 정보를 바탕으로 음소단위 품사검색에 의해 모든 형태소 단위 후보를 생성한다. 자소열-형태소 사전의 각 형태소 엔트리는 41 개의 기본 품사에 대한 품사간 접속 정보를 나타내는 접속 검사표를 사용하고 있다([표 1] 참조). 이러한 접속 정보는 오류 단일 후보가 과 생성되는 것을 방지한다. 한편, 자소열 형태소 메타 사전은 사전에 미등록 되어 분석 되지 않은 형태소들의 자소열 패턴을 분석해 품사를 추정하여 등록어와 동일한 방법으로 처리된다.

품사 태거에서는 학습된 통계적 문맥 확률 정보와 어휘 확률 정보를 이용하여, Viterbi Search 를 거쳐 이상적인 품사 태그열을 찾아 분석 결과 그래프를 만든다. 통계적 어휘 확률 정보는 어절 내 품사 접속 Bi-gram 과 어절간 품사 접속 Tri-gram 에 관한 확률 정보로 나뉘어 학습용 Corpus 를 통해 학습된다. 이 정보는 미등록어 확률사전에 학습되어 품사 태깅의 일관성을 유지한다.

결과로 나온 품사 태그열에 에러수정 후처리기를 사용함으로써, 언어적 특성을 반영하지 못하는 통계적 태깅 방법의 오류를 수정한 후, 결과 그래프와 Tagged Corpus 가 만들어 진다.

POSTAG 에서 사용하는 품사

No	품사	품사이름	No	품사	품사이름
1	MC	보통명사	22	eGS	선어말어미
2	MPN	인명고유명사	23	eCND	보조적연결어미
3	MPC	국명고유명사	24	eCNDC	인용어미
4	MPP	지명고유명사	25	eCNMM	명사형전성어미
5	MCO	기타고유명사	26	eCNMG	관형사형전성어미
6	MD	의존명사	27	eCNB	부사형전성어미
7	T	대명사	28	eCC	연결어미
8	G	관형사	29	y	조용사
9	S	수사	30	b	보조용언
10	B	부사	31	+	접두사
11	E	감탄사	32	-	접미사
12	DR	규칙동사	33	su	단위문장기호
13	DI	불규칙동사	34	so	기타문장기호
14	HR	규칙형용사	35	s`	문장기호 `
15	HI	불규칙형용사	36	s'	문장기호 '
16	I	지정사	37	s.	문장기호 .
17	E	존재사	38	s-	문장기호 -
18	jC	격조사	39	s,	문장기호 ,
19	jS	보조사(한정사)	40	sl	영문자 및 외국어
20	jO	기타조사	41	sh	한자어
21	eGE	종결어미			

[표 1] POSTAG 태그 세트 구성

2.2.2 사전 정보 관리

POSTAG 에서는 41 개의 품사를 계층적으로 분류하여 기본 품사로 사용하고 있다. 이중에서 [표 1]의 보통명사, 고유명사, 관형사, 동사, 형용사에 계층적으로 분류된 11 개의 품사에 대해서는 미등록어 추정 품사로 사용하고 있다.

POSTAG 사전은 크게 형태소 사전, 다어절어 사전으로 구성된 자소열 형태소 사전과 패턴 사전으로 구성된 자소열 메타 사전으로 이루어져 있다. 자소열 형태소 사전의 형태소 사전은 품사 태그, 형태소원형, 형태소 이형태, 접속정보를 가지는 형태소 Entry 11 만 5000 개가 등록되어 있다. 자소열 형태소 사전의 다어절어 사전은 우리말에서 다어절에 걸쳐 자주 사용되는 축약어, 연어 Entry 3 만 7000 개를 형태소 단위로 함께 등록하여 성능 향상을 도모하였다. 한편, 자소열 형태소 메타 사전은 미등록어 처리를 위한 형태소 패턴 사전이며, 형태소의 자소열 패턴 분석 과정에서 주로 마지막 음절

정보를 기준으로 Entry 를 분류하여 품사를 할당한다 ([표 2] 참조).

자소열 형태소 사전 형식		
품사태그<원형>	(이형태)	[접속 정보]
MCC<가감>	(가감)	
MCC<가결>	(가결)	[ㄹ>D 하>D 되>]
bDI ㄹ<지#말>	(지_말)	[동<무유<Hps<시<규>아>]
(가도누워)	DI M<가로놀>(가로누워)[축약>] eCC<어>l-축약-)[품형지<축약<] %	
(와_같)	jO<와>(과)[무<] s<#>(#)[어절끝<어절시작>] HR<같>(같)[유>아>] %	
자소열 형태소 메타 사전 형식		
품사태그<원형>	(이형태)	[접속 정보]
MFO<IVz>	(IVz)	[유>]
MFK<IV*V ㄹ>	(IV*V ㄹ)	[ㄹ]
G*<IV*IVz>	(IV*IVz)	[MC>MK>MF>]
I*<IV*IV>	(IV*IV)	[무>jS>]
DI ㄷ<IV*묻>	(IV*묻)	[규>]

[표 2] POSTAG 형태소 사전의 구성

형태소 분류정보는 형태론적, 구문론적인 정보를 포함한 품사 분류 정보이고 접속 자질은 그 형태소가 다른 형태소와 가지는 결합에 대한 정보로 '유>'는 앞의 형태소와 결합 과정에서 사용되는 유종성 정보를 의미한다. 형태소 접속 검사표는 wild card 를 포함하는 표현으로 예를 들어 "M*<*>[*유>*] ⇔ j*<*>[*유>*]"는 유종성 명사와 유종성 명사와 결합하는 조사간의 결합을 위한 표현이다.

2.3. 형태소 분석기와 품사 태거

2.3.1. 형태소 분석기

형태소 분석기는 Character-Synchronous Dynamic Programming Algorithm 을 기반으로 구성되어 있다. 입력된 문장은 한 어절 단위로 기본적인 형태소 분석을 한다. 이때, 어절에 포함된 음소들의 합성과 분할을 통해 자소열 형태소 사전내의 기본 형태소 사전에서 해당하는 Entry 정보를 찾아 가능한 모든 후보를 만든다. 한편, 어절의 맨 마지막 음절 혹은 음소가 기본 형태소 사전에서 발견되지 않을 경우, 다음 어절의 첫 음절 혹은 음소의 집합에 걸쳐 Collocation 을 발생하는 형태소들을 추정하여 다어절어 형태소 사전에서 가능한 모든 후보를 찾는다. 이처럼, 형태소 분석기는 품사 접속 정보 테이블에 근간하여 자소열 형태소 사전, 자소열 형태소 메타 사전을 가지며, 자소열 형태소 사전에는 여러 어절에 걸친 형태소를 분석하기 위한 다어절어 사전을 포함하고 있다.

자소열-형태소 메타 사전을 이용한 미등록어 추정은 형태소 위치와 개수에 무관하게 추정하기 위해 음절 정보를 이용하여 미등록어 패턴 사전을 통해 이루어 진다. 이때, 어절 내에서 어미나 조사 등의 정보로 추정함으로써 오는 모델의 복잡함을 확률모델로 끌어들여 확률정보를 통해 자연히 해결함으로써 오류를 해결하였다. 그 결과, 미등록어를 등록어와 동일하게 처리할 수 있어 형태소 분석기의 구조를 간단하게 하였다.

2.3.2. 어휘 확률과 문맥 확률의 관리

여기에는 등록어와 미등록어 모두에 대해 음절 단위의 어휘 문맥 확률 정보인 학습된 Bi-gram, Tri-gram 확률 사전을 가지고 있어야 한다. 대량의 코퍼스에서 학습된 어휘 문맥 확률 정보는 각각의 Entry 별로 상대적인 빈도수를 나타낸다. 미등록어에 대한 어휘확률 정보 역시 미등록어 품사를 하나의 Entry 로 처리하여 등록어와 동일하게 처리하고 있다.

2.3.3. 품사 태거

통계적 처리에 의한 품사 태깅 방법을 기본으로 하는 POSTAG 는 주어진 문장에 대한 가장 확률이 높은 해당 품사열을 찾는 은닉 마르코프 모델의 Viterbi Search Algorithm 을 적용하여 누적 확률값을 계산하며, 빔(beam) 을 적용하여 후보를 줄여 다음의 계산량을 줄이고 그 시간 프레임까지의 최적의 패스를 구한다. 그 결과, 형태소 분석기에서 생성된 품사열 후보 결과에 대해서 어절 내 품사 Bi-gram, 어절 간 품사 Tri-gram 확률 사전 결과를 기반으로 후보 품사열 중에서 최적의 품사열을 찾아 품사를 할당하는 역할을 한다.

$$T^* = \arg\max_T \prod_{i=1}^{n} \Pr(t_i \mid t_{i-1})^{\alpha} \left(\frac{\Pr(t_i \mid m_i)}{\Pr(t_i)} \right)^{\beta}$$

[식 1] POSTAG 의 통계 태깅 모델

[식 1]은 은닉 마르코프 모델에 가중치를 부여하여 어절 내 품사 bi-gram에 의하여 문맥 확률과 어휘 확률을 구하는 식이다. 여기서 T*는 최적의 품사열을 나타내고 Pr ($t_i \mid t_{i-1}$)은 bi-gram의 문맥 확률값을 의미하며, Pr ($t_i \mid m_i$) / Pr (t_i)는 수정된 어휘 확률값을 의미한다*([7])*.

접속이 이루어진 형태소들은 부분적인 형태소 그래프를 형성하고, 이 그래프에 대해 통계 정보를 적용하는 Viterbi 탐색을 이용해 최적의 품사열을 찾는 1-best 태깅을 한다. 이 때, 미등록어의 통계 정보는 음절 tri-gram을 이용하여 구해진다. 한 문장에 대하여 최적의 품사열이 정해지면 에러수정 후처리기가 통계 정보 태거의 오류를 올바르게 추정하여 최종적으로 가장 올바른 결과를 출력한다.

2.4. 규칙에 의한 품사 태거의 에러 수정 후처리

POSTAG의 에러 수정 후처리기는 현재 어절 혹은 형태소 분석 결과, 규칙 틀, 참조 대상 형태소 혹은 품사를 입력으로 받아 올바른 분석 결과를 뽑아낸다. 이 과정은 통계적 품사 태거가 가지는 확률 기반 결과 추출의 문제점을 보완하기 위하여 후처리기로 추가되었다([식 2] 참조).

- **Nn : 현재 어절의 n번째, Next 어절.**
- **Pn : 현재 어절의 n번째, Previous 어절.**
- **F : 참조 어절의 어휘 형태소.**
- **L : 참조 어절의 기능 형태소.**
- **M : 참조대상에서 형태소의 참조.**
- **T : 참조 대상에서 품사를 참조.**

에러 수정 후처리기 규칙 틀		
규칙틀	설명	
N1FT	다음 첫번째 어절 (N1)	첫번째 형태소 품사 (FT)
N2FT	다음 두번째 어절 (N2)	첫번째 형태소 품사 (FT)
N3FT	다음 세번째 어절 (N3)	첫번째 형태소 품사 (FT)
P1LM	이전 형태소 어절 (P1)	마지막 형태소 (LM)
P1FM	이전 형태소 어절 (P1)	첫번째 형태소 (FM)

[식 2] POSTAG의 에러 수정 후처리 규칙

오류 수정 규칙의 학습에서는 먼저 통계 태거의 결과와 올바르게 태깅된 말뭉치를 입력으로 받아서 통계 태거의 오류 특성을 수정할 수 있는 규칙을 만들어 낸다. 이때, 과다하게 생성된 규칙이 적용될 경우, 올바르게 태깅된 결과를 거꾸로 오류로 변환시킬 수 있으므로, Scoring을 통해 빈도수가 높고, 적정 기준 값 이상의 규칙만을 적용한다*([1])*. 그리고 나서, 학습 말뭉치에서 획득한 규칙을 다른 실험 말뭉치에 적용이 가능한 보다 일반적인 규칙을 생성하는 과정에서 양질의 규칙을 만들기 위해 한국어의 어휘적인 특성을 반영한 휴리스틱을 이용하여 [식 3]과 같은 일반 규칙을 생성한다.

- **일반화 휴리스틱 1 :**
  용언(D,H)+명사형 어미(eCNMM)는 앞에 주격기나 목적격이 있을 경우에 해당하고 수식을 받으면 전체 어절이 명사(MC)이다.
- **일반화 휴리스틱 2 :**
  '다른'과 같이 "다르/형용사+관형형어미"와 "다른/관형사 "로 될 수 있는 형태소는 앞 문맥의 용언으로 사용되면 "형용사+관형어미"이고, 아니면 "관형사"이다.

[식 3] POSTAG의 에러 수정 후처리 규칙

3. POSTAG의 미등록어 추정

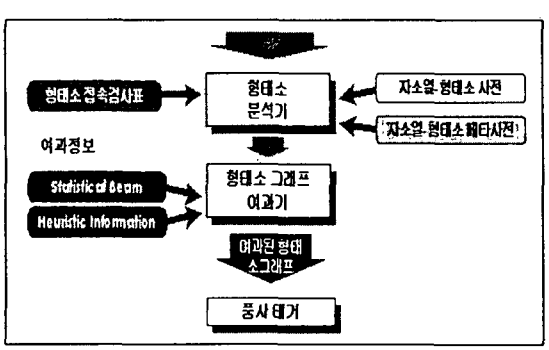
미등록어는 사전에 없어서 분석을 할 수 없는 단어라고 정의할 수 있는데, 이는 언어의 개방성을 나타내주는 쉬운 예가 된다. 미등록어 문제의 자연스러운 접근방법은 미등록어가 가지는 자체적인 특징과 그 주변의 특징으로 품사를 예측하는 것이다. 영어권의 예를 들면, 우리는 'rakishly' 라는 단어를 처음 보아도 'ly'를 보고 부사일 가능성이 높다고 생각한다. 이것은 임의의 단어 어근과 접사의 조합으로 이루어졌다는 것에서 유추된 접근 방법이다.

한국어의 경우 하나의 어절이 의미 형태소와 형식 형태소의 결합으로 이루어져 있고, 이때 의미 형태소가 미등록어가 되므로, 미등록어의 주변 정보인 형식 형태소를 참고하여 미등록어 품사를 예측할 수 있다. 또, 의미 형태소가 가지는 자체적인 특징, 예를 들어 '김준석'은 '김'이라는 음절이 형태소의 처음이고 3음절로 이루어졌다는 정보를 참고하여 품사를 결정하는데 도움을 줄 수 있다.

3.1. 미등록어 추정 모듈을 통합한 형태소 분석기

기존의 많은 연구들은 형태소 분석에서 실패한 어절에 대해서 별도의 모듈을 두어서 등록어와 다른 방법으로 형태소 분석을 시도한다. 그러나, 이 경우 미등록어가 포함되어 있음에도 불구하고 미등록어로 분석을 하지 못하고 사전에 있는 다른 분석을 함으로써 미등록어를 찾지 못하는 경우가 있다. 예를 들어 '빔'은 "레이저 빔(beam)"의 명사와 "좌석이 빔(emptiness)"의 '비/형용사+ㅁ/명사형 전성어미'로 분석이 가능하다. 만약 '빔/명사'는 미등록어이고 '비/형용사+ㅁ/명사전성어미'는 등록어일 경우, 분석 가능한 형태소가 있기 때문에 미등록어를 찾지 못하는 단점이 있다. 또한 별도의 미등록어 처리 모듈을 둠으로써 형태소 분석기가 복잡해지며, 미등록어를 분석하기 위해서 오른쪽에서 왼쪽으로 분석을 시도하기 때문에 실제 응용에서는 처리가 아주 복잡해질 수 있다.

POSTAG 시스템의 형태소 분석기는 입력 문장을 왼쪽에서 오른쪽으로 읽어 가며 사전을 참조하여 형태소를 분리한다. 분리된 형태소들은 접속정보 테이블을 참조하여 접속 검사 과정을 거치게 된다. 위치와 개수에 무관한 미등록어 추정을 위해서 음절 정보를 이용하여 미등록어 패턴 사전을 구성하였다. 이 사전을 이용하여 입력어절에 대해서 일반적인 사전과 미등록어용 패턴 사전을 동시에 찾음으로써 등록어와 미등록어를 동일한 방법으로 처리할 수 있다. 미등록어 추정 모듈을 형태소 분석기에 통합한 모습은 [그림 3]와 같다.



[그림 3] 미등록어 추정 모듈을 통합한 형태소 분석기

3.2. 미등록어 패턴 사전 (자소열-형태소 메타 사전)

형태소 패턴 사전([표 2]참조)을 이용하면 미등록어를 마치 등록어인 것처럼 처리할 수 있어 미등록어 처리를 위해 별도로 시스템을 구성하지 않아도 되는 장점을 가지게 된다. 그리고 어절 내에서 어미나 조사 등의 정보로 추정함으로써 오는 모델의 복잡함을 확률모델 내로 끌어들여 확률정보를 통해 자연히 해결함으로써 오류를 해결하고 모델자체의 간단함도 이룰 수 있다. 한국어가 품사별로 가지는 음절들의 제약에 대한 많은 연구를 바탕으로 하고 그들을 더욱 확장하여 POSTAG 시스템에서는 미등록어를 추정하기 위한 형태소 패턴 사전(자소열-형태소 메타사전)을 이용한다.

[표 3]는 형태소 패턴사전의 또 다른 보기이다. 여기서 'Z'는 자음을 의미하고 'V'는 모음을 나타내며 '*'는 자음과 모음을 포함하여 여러 개의 자소를 나타내는 기호이다. 예를 들어, "고마워"라는 형태소는 패턴사전에서 "ZV*워"와 형태소 이형태가 매치되며 이때 형태소 원형은 "ZV*ZV ㅂ"을 이용하여 "고맙"으로 복원되고 접속정보는 "축약>"이 된다.

기존의 많은 미등록어 추정 방법은 조사나 어미를 이용하여 미등록어를 추정했기 때문에 추정 품사를 '명사', '동사'로 한정하는 경향이 많았으나 POSTAG 시스템에서는 여러 응용에서 사용할 수 있고, 한국어의 개방적 특성을 고려하여 미등록어 추정품사를 '명사(보통명사, 고유명사)', '관형사', '부사', '동사(규칙동사,불규칙동사)', '형용사(규칙형용사,불규칙형용사)'등 8 개의 품사를 대상으로 한다.

[품사<원형>]	[이형태]	[접속정보]
HI ㅂ<ZV*ZV ㅂ>	(ZV*워)	[축약>]
DI ㅅ<ZV*ZV 젓>	(ZV*젓)	[규>]
DI ㄷ<ZV*듣>	(ZV*들)	[불>어>]

[표 3] 미등록어 패턴 사전

3.3. 어휘확률 계산

한국어 미등록어 추정의 기존의 연구에서는 미등록어

빈도수를 한 번으로 간주하거나 규칙을 이용하여 처리하였기 때문에 $Pr(W_i|t_i)$를 계산할 필요가 없었다. 그러나, POSTAG 시스템과 같이 등록어와 같이 추정하고 등록어와 같은 확률모델 내에서 Viterbi 탐색을 이용하여 처리하고자 한다면 어휘 확률값($Pr(W_i|t_i)$)을 계산하는 방법이 필요하다.

하지만 미등록어에 대한 어휘확률은 말뭉치로부터 직접 구할 수는 없기 때문에 이를 계산하는 특별한 방법이 필요하다. 초기에는 한국어 음소 Trigram을 이용하여 미등록어 어휘확률을 계산하는 방법을 사용하였으나 음소는 그 하나만으로 의미를 가질 수 없기 때문에 이러한 정보로부터 형태소의 어휘확률을 추출하기 위해서는 많은 양의 태깅된 말뭉치 필요하게 된다. 따라서 음소 Trigram을 음절 Trigram으로 확장하여 미등록어 어휘확률을 계산하는 방법을 사용한다. 왜냐하면, 음절은 그 하나로서 의미를 가질 수 있는 단위가 될 수 있으므로 특별한 음절로 시작하는 형태소에 대해서 정보를 줄 수 있다. 예를 들어 '김'이 형태소의 처음에 나타나면 인명일 가능성이 높기 때문에 '김'으로 시작하고 3 음절인 미등록어는 인명 고유명사의 어휘확률이 높아진다. 구체적인 계산식은 [식 4]와 같다.

$$\frac{Pr(\ell|m)}{Pr(\ell)} \approx Pr_1(e_1|\#,\#)Pr(e_2|\#,e_1)\left[\prod_{i=3}^{n} Pr(e_i|e_{i-2},e_{i-1})\right]Pr(\#|e_{n-1},e_n)$$

[식 4] 어휘 확률 계산식

여기서 'm'은 형태소를 나타내며, 'e'는 음절을 't'는 품사를 나타낸다. 따라서, $m = e_1e_2\ldots e_n$를 의미한다. 또한 '#'은 형태소 표지를 나타낸다. [식 4]는 데이터 부족 문제를 해결하기 위해서 [식 5]로 평탄화 과정을 거친다. 여기서 $f_t(e_i|e_{i-2},e_{i-1})$는 품사 't'가 음절 e_{i-2}, e_{i-1}, e_i 과 함께 나타나는 횟수이다.

$$Pr_t(e_i|e_{i-2},e_{i-1}) \approx f_t(e_i|e_{i-2},e_{i-1}) + f_t(e_i|e_{i-1}) + f_t(e_i)$$

[식 5] 평탄화 과정

예를 들어 '김준석'이 인명 고유명사가 될 어휘 확률은 다음과 [식 6]같이 계산된다.

$$\frac{Pr(MPN|\text{김준석})}{Pr(\text{MPN})} \approx Pr_{MPN}(\text{김}|\#,\#)\times Pr_{MPN}(\text{준}|\#,\text{김}) \times Pr_{MPN}(\text{석}|\text{김},\text{준})\times Pr(\#|\text{준},\text{석})$$

[식 6] 어휘확률 계산과정 예

이 경우 '김'이 형태소의 처음에 나오는 경우는 인명 고유 명사일 때가 다른 품사 경우보다 많으므로 인명 고유명사의 어휘확률이 상대적으로 높아진다.

모든 음절 Trigram 은 학습 시에 미리 계산되어 미등록어 추정과정에서 이용된다.

3.4. 미등록어 여과기

미등록어 패턴 사전에서 추정된 미등록어들은 접속 검사를 거치면서 축소가 된다. 여기서 통과된 미등록어들은 음절정보와 여러 가지 휴리스틱과 Viterbi 탐색과정에서 계산된 누적 확률 값을 이용하는 beam을 통해서 여과과정을 거친다. 몇 가지 휴리스틱을 살펴보면 다음과 같다.

- 시제 선어말 어미의 활용형

한국어에서 어휘형태소는 선어말어미를 가지지 못한다. 이러한 성질을 이용하여 미등록어로 추정된 형태소에서 시제 선어말 어미가 포함된 형태소를 제거한다. 시제 선어말 어미는 기본 형태인 '셨', '았', '었', '였', '겠' 과 '웠', '왔'등 총 97 개를 등록하여 사용한다*([9])*.

- 체언류

한국어에서는 '는', '를' 그리고 "가을", "고을", "마을"로 끝나는 복합어, "갑을", "넨장맞을", "노갑이을", "노을", "리을", "빌어먹을", "올", "젠장맞을", "태을"을 제외한 '을'로 끝나는 체언은 없다. 미등록어로 추정된 체언 중에서 이 정보에 어긋나는 형태소들은 제거한다.

4. 표준 품사 태그세트 및 표준 코퍼스 변환

품사 태그세트 매핑은 서로 다른 두 태그세트의 정의와 분류의 차이에 대한 매핑 관계를 설정하여 변환시킬 수 있게 함으로써 서로의 태깅된 코퍼스를 공유하는 기술이다. 품사 태깅 시스템간의 태그세트 매핑은 태깅된 코퍼스 정보를 공유하고, 각 품사태거의 태깅 결과를 다른 품사태거에 배포하여 코퍼스의 재사용성을 높이며, 태깅 기준에 관한 의견을 수렴함으로써 표준 태그세트의 기준을 타당하고 명확하게 강화할 수 있는 이점을 가지고 있다*([3],[4],[5],[8])*.

본 연구실에서는 MATEC99를 준비하는 과정에서 POSTAG의 태그세트와 표준 태그세트 간의 매핑을 시도하였다. 그 결과, 표준으로의 근본적인 수정을 배제하고도 POSTAG로부터 표준의 태깅된 코퍼스 결과를 추출하는 효과를 얻을 수 있었다*([10])*.

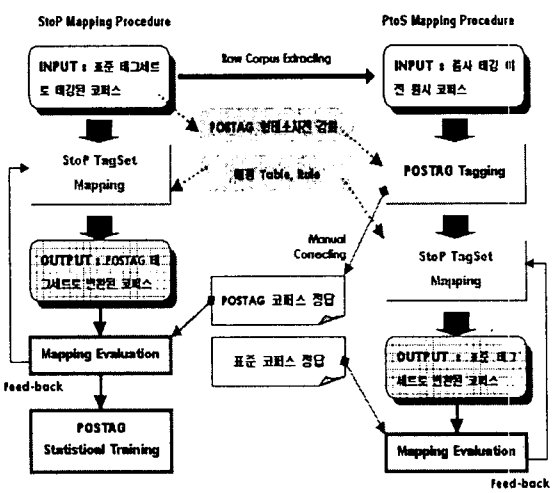
표준으로의 매핑 과정에서는 주로 세그먼트 차이(사전 표제어 차이), 원형과 이형태의 차이, 품사 차이, 그리고 형태소 원형 복원 문제 및 축약 차이의 처리를 고려하였다. 한편, 이 과정에서 점진적인 방식의 매핑을 시도하였는데, 우선, 한 어절 내에서 품사, 원형, 세그먼트 차이 및 출력 형식과 문장 구분 문제를 차례대로 해결하고 어절간에 생기는 매핑 문제의 해결을 도모하였다. 또한 피드백 과정을 통해 매핑 오류를 감소시키는 방식을 이용한다. 태그세트 매핑의 정확도를 측정하기 위해서 매핑하기 전의 정확도와 매핑 후의 정확도를 서로 비교함으로써 매핑의 정확도를 측정하는 실험을 수행하였다.

4.1. 품사 태그세트 매핑 구조

POSTAG의 품사 태그세트 매핑은 크게 표준(ETRI Standard) 태그세트에서 POSTAG 태그세트로의 StoP 매핑(Standard to POSTAG Mapping)과 POSTAG 태그세트에서 표준 태그세트로의 PtoS(POSTAG to Standard Mapping)이 있다. PtoS 매핑 결과 표준의 태깅된 코퍼스를 POSTAG 태그세트로 태깅된 코퍼스로 변환하여 확률학습에 필요한 대용량의 코퍼스를 습득할 수 있다. 이 과정은 태깅된 코퍼스들간의 매핑이므로 Off-line으로 수행한다. 한편 StoP 매핑은 POSTAG에 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. 이러한 매핑은 한 문장에 대해서 처리하고, 한 문장에서 각 어절을 구성하는 형태소의 품사에 따라 매핑을 수행하게 된다.

4.1.1. StoP 매핑

POSTAG 태그세트와 표준 품사 태그세트에 대한 지침서를 기준으로 기본적인 품사 분류 비교, 매핑 규칙, 예외 사항들에 대해 매핑 테이블을 제작했다. 이 과정에서 품사별로 POSTAG의 사전을 보강하고, 표준 태거의 태깅 원칙과 비교하여 POSTAG의 태깅 기준을 수정하는 작업을 병행하였다. 매핑 규칙과 테이블로 PtoS 매핑을 수행한 결과는 수동으로 만들어진 정답 코퍼스와 비교되어 보강해주는 피드백과정을 반복한다.



[그림 4] 품사 태그세트 매핑 알고리즘

4.1.2. PtoS 매핑

PtoS 매핑은 표준의 태깅된 코퍼스의 Raw-Text에 대한 POSTAG 결과를 매핑 규칙과 테이블을 사용하여 태깅 결과 그래프로부터 표준 태그세트로 매핑 한다. 매핑된 결과를 표준태그세트로 태깅된 정답 코퍼스와 비교를 통해 오류들에 대해서는 역시 자동의 피드백 과정을 반복한다([그림 4] 참조).

4.1.3. 세그멘트 차이에 대한 매핑 테이블 규칙

아래의 [표 4]은 PtoS 와 StoP 의 품사별 매핑 규칙을 설명한 것이다. Stag 는 표준 품사 태그 세트를 의미하며, Ptag 는 POSTAG 의 품사 태그 세트를 의미한다.

- Stag : ETRI 표준 태그세트
- PTAG : POSTAG 태그세트

품사	규칙
심볼	Stag 의 s 가 Ptag 의 s, s. su so 같이 세분화 된다. POSTAG 의 symbol dictionary 를 참조하여 주형태와 이형태의 구분하여 변환한다.
외국어	'예스' ,'노'와 같은 한국어로 발음된 외국어의 경우에 Stag 의 f 를 Ptag 의 MC 로 매핑한다.
명사	Stag 에서는 명사를 자립명사(nc)와 의존명사(nb)로 나누는데, Ptag 에서는 명사를 보통명사(MC), 고유명사(MP), 의존명사(MD)로 구분한다. nb 는 MD 로 매핑 되지만, nc 경우 POSTAG 의 보통명사 dictionary 를 참조하여 MC,MP 를 구분하여 변환한다
동사 형용사	Stag 의 pv 가 Ptag 의 규칙동사(DR)와 불규칙 동사(DI)로 구분하여 변환된다. 한편, 형용사도 pa 에서 규칙형용사(HR) 과 불규칙 형용사(HI)로 변환된다. 규칙 및 불규칙 태그세트 구분은 패턴 사전을 이용하여 처리하고, 태깅된 표준 코퍼스를 참조하여 이형태 복원 문제를 해결한다.
보조용언	Stag 의 px 가 Ptag 의 보조용언(b)로 매핑 된다. 보조용언의 경우 보조용언의 정의 범위에서 차이가 많이 나므로 Stag 의 정의에 따라서 POSTAG 의 태깅 지침을 변경하였다.
지정사	Stag 의 co 가 존재사(E)와 지정사(I)로 매핑된다. POSTAG 에서는 '이'와 '아니'만을 지정사로 보는데 '아니'의 경우에 표준 tag set 으로 태깅될 때에는 형용사(pa)로 태깅된다.
부사	Stag 의 일반부사(maj)와 접속부사(maj)는 Ptag 의 접속부사(BJ)와 그 외 부사로 구분하여 매핑한다.
관형사	Stag 의 mm 은 Ptag 의 관형사(G)로 매핑된다. 하지만, mm 에서의 '두세, 서너, 두서너, 네, 스무, 몇'는 Ptag 에서 수사(S)로 태깅되는 예외를 지닌다.
접사	Stag 는 접사를 접두사(xp)와 접미사로 명사 파생 접미사(xsn), 동사 파생 접미사(xsv), 형용사 파생 접미사(xsv)로 구분하여 태깅한다. 이는 Ptag 의 접두사(+), 접미사(-), y(조용사)로 각각 매핑된다. 접두사의 경우에는 양쪽 모두 '제-' 만을 인정하므로 쉽게 매핑된다. 파생 접미사는 양쪽의 품사정의차이와 세그먼트차이가 많이 존재하여, 비교 작업과 사전 튜닝 작업이 요구된다.
조사	Stag 의 조사는 격조사(jc), 보조사(jx), 접속조사(jj), 속격조사(jm)으로 구분된다. 이는 Ptag 의 격조사(jC), 보조사(jS), 기타조사(jO)로 각각 매핑된다. 격조사를 제외한 조사의 경우 품사 분류 정의 차이와 주형태와 이형태 정의 차이가 많으므로 1:1 매핑을 하였다. 한편, 어미와 함께 사용되는 '요' 경우, '-어요/eGE'를 '-어/ef+요/jx'로 변환시키는 1:N 매핑을 하였다. 속격조사의 경우에는 Stag 에 '의' 와 '엣' 두가지가 존재하는데 Ptag 에서는 '의'뿐이므로 '엣'의 사전에의 추가 작업이 필요하다.
선어말어미	Stag 의 ep 는 Ptag 의 선어말어미(eGS)로 매핑된다. 표준에서는 '었었','-쓰었','시었'등을 'ep+ep'로 태깅하는데, POSTAG 에서는 하나의 'eGS'로 태깅하므로 이에 대한 처리가필요하다.
어말어미	Stag 에서는 어말어미를 종결어미(ef), 연결어미(ec) 및 전성어미로 명사형 전성어미(etn) 과 관형사형 전성어미(etm)으로 구분한다. 이는 Ptag 의 eGE(종결어미), eCC(연결어미), eCNDI(보조적 연결어미), eCNDC(인용어미), 명사형 전성어미(eCNMM), 관형사형 전성어미(eCNMG)로 각각 매핑된다. 복합 어미, 어미의 축약 복원, 생략 등에서 많은 차이가 나고, POSTAG 에 미등록어가 많아서 비교 작업과 사전 튜닝 작업이 요구된다.

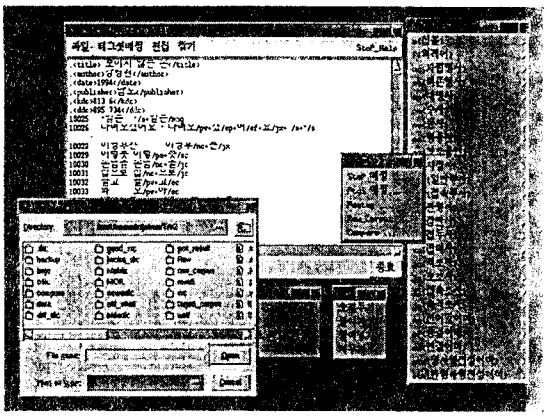
[표 4] 품사별 세그멘트 차이에 대한 매핑 테이블 규칙

4.2. 태그세트 매핑 관리 도구 (Mapping Tool)와 품사 태그세트 매핑의 수행

본 연구실에서는 POSTAG 의 코퍼스를 분류하여 관리하고, 자동화된 표준 태그세트와의 매핑 수행 결과를 출력 받아 검사하고 피드백 하여 재수행 하는 과정을 원할히 하기 위하여 태그세트 매핑 관리 도구를 제작하였다. 한편, 품사 태그세트 매핑은 4.1 에서 소개한 품사별 세그먼트 차이에 대한 매핑 테이블 규칙을 일반화 하는 과정을 거쳐서 태그세트들간의 매핑 입출력을 자동으로 수행할 뿐만 아니라, 형태소들의 태깅 결과 차이를 새로운 규칙으로 추출해 준다.

4.2.1. 태그세트 매핑 관리 도구

POSTAG 태그세트 매핑 관리 도구는 태깅된 코퍼스를 정답의 코퍼스, 실험을 위해 태깅된 코퍼스, 코퍼스의 원문 등으로 나누어 관리함과 동시에 POSTAG 와 ETRI 표준 태거로 태깅된 코퍼스 결과를 양방향으로 변환한다. 이 도구에는 태그세트 매핑 함수를 만드는 과정에서 코퍼스를 가공하기 위해 만들어진 각종 파일 처리 함수들이 추가 기능으로 들어가 있다. 또, 표준 태거의 태깅 지침서와 학습 코퍼스를 이용하여 구축한 표준 형태소 사전과 POSTAG 의 형태소 사전 정보를 제공하고 있다.



[그림 5] 태그 세트 매핑 도구

이 도구는 POSTAG에 PtoS 매핑과 StoP 매핑 Library 와 GUI Framework을 추가하여 제작되었으며, POSTAG의 사전 관리 도구 및 전처리기와 합쳐져 POSTAG Workbench로 확장될 예정이다([그림 5] 참조).

4.2.2. 태그세트 매핑 수행 예

다음은 StoP 매핑에서 표준 태그세트로 태깅된 문장의 임의의 한 형태소 X의 품사가 'pv'(동사) 일 때, POSTAG 태그세트로 태깅된 코퍼스 "품사<주형태>(이형태)"로 자동으로 매핑하는 함수의 예이다.

```
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));
}
```

이 경우, 표층정보와 형태소 X의 주형태가 입력으로 들어간다. 형태소 X의 원형은 그대로 이용하고, 원형의 마지막 음절(Last_S_MOR(X))을 음절 패턴 사전을 참조여 규칙동사(DR)와 불규칙동사(DI) 여부를 구분한다. 마지막으로 표층정보와 원형 정보를 이용하여 이형태 복원과정을 수행하면 매핑이 끝난다.

PtoS 매핑에서는 기본 품사 비교 테이블로 품사를 매핑하고 용언과 선어말어미를 제외하고는 모든 품사의 원형을 이형태와 일치시킨 후 매핑에서 품사별로 나타나는 중요한 세그멘트 차이들을 각각의 Rule을 통해 처리한다. 이때, 품사별 세부적인 예외들에 대하여 1:1 n:1 1:m n:m 세그멘트 차이로 분류하여 매핑 테이블을 처리한다. 예외 처리 과정 중에서 Lexical 정보를 참고해야 하는 예외들에 대해서 마지막으로 처리해 준다.

예를 들어 표준태그에서는 "청소를 하던(던/etm:관형사형 전성어미) 사람"에서 '던'을 관형사형 전성어미로 보는데 POSTAG 태그세트에서는 '던'을 eGS<더>(더)+eCNMG<ㄴ>(ㄴ) 으로 태깅 한다. (여기서 eGS는 시제선어말어미 이고, eCNMG는 관형사형 전성어미 임). 이와 같이 세그먼트의 차이를 보이는 형태소들은 매핑 테이블에 그 정보를 저장하고 있고, 매핑 시에 테이블을 참조 한다. 테이블의 구성은 [표 5]와 같다.

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

[표 5] 매핑 테이블

Seg 정보는 세그먼트의 개수를 의미하는 데 0~3의 값을 가진다. "시었겠/ep" 의 경우는 "eGS<시>(시)+eGS<었>(었)+eGS<겠>(겠)" 과 같이 매핑 되므로 3 이라는 Seg 값을 가진다.

5. MATEC99 평가 결과 및 분석

5.1. MATEC99 참가 개요

제 1 회 형태소 분석기 평가 대회에서 POSTAG의 품사 태거는 POSTAG의 결과를 태그세트 매핑 과정을 거쳐 표준과 동일한 태깅된 결과를 제출하였다. 명사 추출기의 경우, 표준으로 태그세트 매핑된 결과에서 명사를 추출하였다. 반면에, 형태소 분석기 결과에는 태그세트 매핑을 적용할 수 없었기 때문에, POSTAG의 출력 결과를 제시하였다.

평가대회에서 사용된 사전의 크기는 약 15만개의 등록어로 MATEC 99를 준비하는 과정에서 학습 코퍼스와 표준 태그세트 태깅 지침서를 참고하여 새로운 형태소

들을 추가로 등록하였다. 그 결과, 형식 형태소의 경우, 등록어 수가 기존의 사전 등록어보다 2 배 이상 증가하였고, 불필요한 축약어 및 연어 등록어에 대해 표준의 세그먼트와 일치시키며 변경, 추가하였다([표 6] 참조).

- 형태소 사전 크기

명사 사전	7 만개 (복합명사 포함)
동사 사전	2 만 7000 개
형용사 사전	9500 개
기타 품사	8500 개
형태소 사전 합계	11 만 5000 개

- 다어절어(복합어) 사전 크기

동사 사전	2 만 8000 개
형용사 사전	6700 개
연어 사전	900 개
기타 품사	1400 개
다어절어 사전 합계	3 만 7000 개

- 패턴 사전 크기

형태소 사전	1600 개
다어절어 사전	100 개

- 추가, 변경된 사전 정보 (형태소 사전, 다어절어 사전)

실질 형태소	약 20000 만개 (보통명사제외)
형식 형태소	어미: 약 250 개 조사: 약 100 개(복합어미포함) "어미 + 보조용언" : 약 85 개 [다어절어 등록어]

[표 6] POSTAG 사전 규모와 MATEC99 준비과정에서 추가 및 변경된 사전 정보

사용환경	SunOS 5.6 (UNIX)
시스템 사양	Sun Sparc Station Ultra 2
CPU	233Mhz * 2 개 (Synchronous)
RAM	128Mbyte
프로그램 수행시 최소 필요한 RAM 크기	약 12 Mbyte (Object 파일크기+프로그램 내에서 사용하는 크기+사전 크기)
프로그램 파일 크기	약 70Mbyte (형태소 사전 및 확률사전 포함 전체)
참여 인원 및 기간	총 60 일 : 품사 태그 매핑을 위하여 50 일간 2 명(1 명:100%, 1 명 50%) 참가, 10 일간 매핑과 튜닝을 위해 5 명 (5 명:100%) 참가.

[표 7] POSTAG 시스템 사양

평가대회에 사용된 환경과 시스템 사양은 [표 7]과 같다. 위의 시스템에서 실험한 결과 형태소 분석기 및 품사 태거(POSTAG 만을 사용한 경우)를 테스트하는 데에는 약 40 분이 소요되었다. 한편, 품사 태거와 태그세트 매핑(표준으로의 태그세트 매핑을 사용한 경우)에는 45 분이 소요되었다. 명사 추출기는 308 개의 파일을 품사 태깅 및 매핑한 후에 추출하였기 때문에 약 55 분의 시간이 소요되었다. 다른 품사 태거 시스템과 비교해서 테스트 시간이 오래 걸리는 이유는 일반화된 방법의 미등록어 추정 시간이 매우 오래 걸리고, 태깅된 결과에 대하여 태그세트를 매핑하고 출력 형식을 변경하기 때문이다.

5.1.1. 명사 추출기 준비 작업

MATEC'99 를 위한 명사 추출 프로그램을 새롭게 작성하였다. 기존의 명사 추출기는 정보 검색의 복합명사의 합성과 분할에 사용된 명사 추출기를 MATEC'99 의 목적에 맞게 변형하였다. 이 과정에서, 명사 추출 기준 안건이 "Type 2"로 결정되면서, 파일 내에서 연속된 명사는 단일 명사로 취급하여 추출하였고, 중복된 명사는 1 번만 추출되도록 하였다.

5.1.2. 품사 태거

POSTAG 의 품사태그세트와 ETRI 의 표준 태그세트를 양방향으로 변경시켜주는 매핑 프로그램을 제작하였다. 표준에서 POSTAG 로의 매핑은 결과 파일을 통해 Off-line 으로 이루어졌고, 그 결과를 문맥 확률과 어휘 확률을 계산하는데 다시 반영하였다. 이 과정에서 발생되는 미등록어들을 처리하기 위해, Training Corpus 의 등록어를 분석하여 사전 등록어 강화 작업을 수행하였다. 이 과정에서 기존의 사전 미등록어 리스트를 포함하여 7 만개의 등록어를 추가하는 작업을 하였다.

POSTAG 의 결과에서 표준으로의 매핑은 On-line 으로 이루어져 형태소 그래프의 결과에서 바로 표준 그래프로 결과를 변환시켰다. 이 과정에서 태그세트 간의 Annotation Scheme 이 다르고 반영된 접속 정보가 달라서, 매핑에서 반영하기 어려운 차이들에 대해서는 구문 분석과 의미 분석에 영향을 미치지 않은 범위 내에서 사전 정보를 변경시켰다. 대부분의 의미 형태소에 대해

서는 별다른 차이가 없었기 때문에 변경된 양으로는 많지는 않았지만, 대부분의 기능 형태소나 관련된 등록어(조사, 어미, 보조용언, 조용사, 연어 정보, 축약의 복원)에 대한 수정이 불가피하여 변경 과정에서 각각에 대한 변경 문제를 판단하느라 많은 시간이 소요되었다. 한편, 태그세트 매핑 과정에서 1-best 품사 태깅 결과만을 매핑하기 때문에 형태소 분석 결과 나온 N-best 후보들에 대해서는 POSTAG 의 결과를 그대로 출력 결과로 제출하였다.

5.1.3. 학습 코퍼스의 활용

표준으로 태깅된 학습 코퍼스 26 만 어절에 대해, 명사 추출기와 품사 태거의 확률 학습에 사용했고, 품사별로 등록어를 추출하여 사전을 강화하는데 사용하였으며, 빈번하게 발생하는 품사 태깅 에러를 수정하기 위한 후처리에서 규칙학습의 정답의 코퍼스로 사용하였다. 한편, 태그세트 매핑을 위하여 양쪽 코퍼스들 간의 태그세트 차이, 등록 정보 차이, Annotation Scheme 의 차이에 의한 Segment 의 차이등의 매핑 규칙을 자동으로 추출하는데 사용하였다. (여기에는 양쪽의 품사 지침서와 학습 코퍼스가 함께 기준으로 사용되었다.)

5.2. MATEC99 테스트 평가 결과

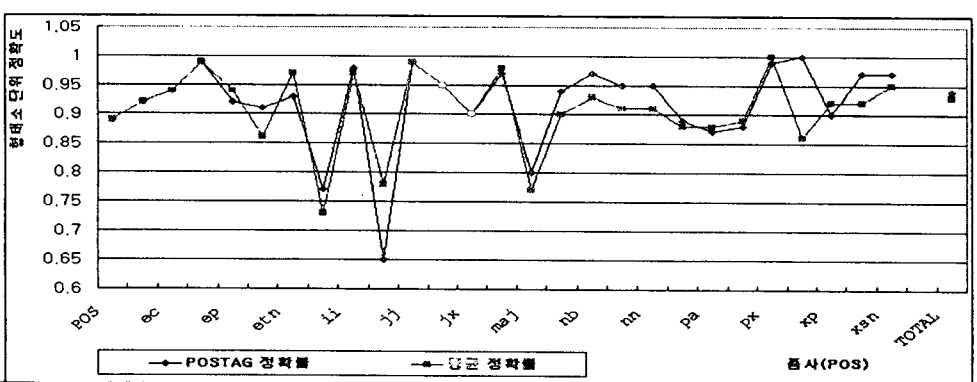
MATEC99 에 적용한 POATAG 는 POSTAG 의 형태소 분석 및 태깅의 결과를 On-line 으로 받아 PtoS 매핑을 한다. MATEC99 의 테스트 결과, 명사 추출기에서 94.5%의 형태소 단위 정확률을 얻었고, 품사 태거에서 94%의 형태소 단위 정확률을 얻었다([표 8] 참조).

[표 8] MATEC99 평가 결과

	명사 추출기		품사 태거	
	재현률	정확률	재현률	정확률
POSTAG 결과	94 %	95 %	95 %	94 %
평균 결과	85 %	80 %	93 %	93 %

위의 표에서 나타난 결과에서 실제 POSTAG 의 태깅 에러의 비율을 알기 위하여 테스트 코퍼스를 대상으로 태깅 정확도와 매핑 정확도를 함께 측정하였고, 매핑에서 발생하는 에러와 태깅에서 발생하는 에러의 비율을 비교하여 보았다.

이 실험을 위하여 MATEC'99 테스트 코퍼스의 소설, 비소설, 뉴스 스크립트에서 임의로 8000 어절, 18063 개의 형태소를 선택하여 정답 코퍼스와 비교하여 에러를 조사한 결과 총 682 개의 어절 764 개의 형태소 단위 에러가 발생하였는데, 이중에서 268 개의 어절 298 개의 형태소 단위 에러가 매핑에 서 발생한 것이었다. 이 결과, 태그세트 매핑을 포함한 품사 태깅 결과, 형태소 단위 정확도는 95.77% 이었고([표 9],[표 10] 참조), PtoS 매핑에서 발생하는 에러를 제외한 순수한 품사 태깅 정확도는 97.42 % 이었다([표 11] 참조).



[그림 6] 품사별 형태소 단위 정확도 비교

[표 9] PtoS 태깅 정확도

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

	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

	Mapping 반영	Mapping 미반영
정확도 (%)	95.77 %	97.42 %

[표 11] POSTAG 결과비교

5.3. MATEC99 결과 분석 및 문제점

5.3.1. POSTAG 매핑 에러 분석

[그림 6]에서 보는 바와 같이, POSTAG의 테스트 결과 가장 많은 에러를 발생하는 품사들은 외래어, 접속조사, 감탄사, 부사, 관형사, 보조용언과 동사였다. 이들 품사 중에서 관형사와 보조용언과 동사의 경우는 대부분 태깅 과정에서 여러가지 가능한 후보가 발생하면서 오류를 발생시켰다. 또한, 그 외의 품사들 중에서 발생하는 오류의 대부분은 품사 태깅 과정에서 미등록어의 추정을 실패하거나, 올바른 형태소 분석이 실패하여 발생하였었다. 하지만, 접속조사, 감탄사, 부사, 외래어에서 발생하는 에러들은 태그세트 매핑에서 발생하는 에러이며, 이들은 다음에서 보는 바와 같이 매핑에서 반영하기 힘든 문제들 이었다.

5.3.2. POSTAG 매핑 에러에 관련된 문제점

표준 태그 세트로의 변환 과정에서 발생하는 에러 중에는 POSTAG 태깅 과정이나 태그세트 매핑 과정에서 반영할 수 없는 에러들이 존재하였다. 이러한 예들은 주로 품사 태깅의 수준에서 그 정보를 얻을 수 없는 것임에도 불구하고 표준 태그세트의 태깅 지침으로 설정되어 있었기 때문에 발생한 매핑 에러들이었다. 차후의 표준 태그세트의 태깅 지침을 변경하는 과정에서 다음과 같은 매핑 에러에 관한 의논이 있어야 할 것이다.

(예 1) 그는 "아임 쏘리." 라고 말했다.

품사 태깅을 하는 과정에서 위의 (예 1)에서 아임/f 쏘리/f를 외국어로 추정할 수 있는 어휘 정보나 문맥정보 없다. 따라서, 품사 태깅이 가능하기 위해서 외래어를 우리말로 발음한 어휘들에 대하여 사전에 외래어로 모두 등록시키는 과정이 불가피하다.

(예 2) 그랬더니/maj 그녀는 화를 냈다.

(예 3) 내가 그랬더니(그러/pv+었/ep+더니/ec) 너도 그랬잖아.

"그러-, 이러-, 저러-" 등의 활용형이 접속 부사로 쓰이는 경우, POSTAG에서는 "그랬더니" 분석을 항상 (예 3)와 같이 한다. 문장 내에서 (예 2) 과 같이 쓰이는 경우, 특정한 문장 내 위치 정보 없이 문법적인 관계를 알아야 부사인지 동사구인지 구분이 가능하므로, 표준 태그세트의 기준에 맞게 태깅 수준에서의 매핑이 불가능하다. 이러한 오류의 경우 규칙에 의한 에러 수정 후처리기에서 일부분에 대해서는 처리가 가능하나 일반적인 해결이 불가능하다.

(예 4) 그만 하지 그래/ii.

(예 5) 네가 한다 그래(그러/pv+어/ef).

위의 예는 특정한 어휘가 감탄사로 쓰이는 경우와 서술어 역할을 하는 경우를 품사 태깅에서 구분할 것을 요구한다. POSTAG에서는 "그래, ..."와 같이 쉼표와 함께 나올 경우를 제외하고는 항상 (예 5)와 같이 분석을 한다. 하지만, 표준 태그세트에서는 '그래'가 문장에서 빠질 경우 전체 문장의 의미가 달라는 경우에만 (예 5)로 분석하고 그 외에는 (예 4) 로 처리한다. 이러한 분석은 특정 문장 내에서의 어휘간의 문법적 관계를 알아야 구분이 가능하므로 역시 일반적인 해결이 불가능하다.

(예 6) 사과와(와/jj) 배를 먹었다.

(예 7) 그 애와(와/jc) 싸우지 말아라.

위의 예에서와 같이 표준 태그세트에서는 접속조사 '와/과, 하고, 랑' 뒤에 '만나다, 부딪치다, 헤어지다, 사귀다, 함께, 같이, 동행하다.' 등의 단어가 나타날 경에는 이들을 공동격조사로 구분하고 있다. 하지만, 이러한 구분은 문장 내에서의 특정한 의미의 단어들이 함께 사용되는지를 판단해야 하므로 의미 분석 관계를 알아야 구분이 가능하다. 따라서 태깅 수준에서의 매핑이 불가능하다.

5.3.3. 그 밖의 문제점들

본 연구실에서는 이번 대회를 위한 형태소 분석기를 따로 제작하지 않고, 태그 세트 매핑을 거쳐 실제 품사 태깅 시스템과의 일관성을 유지하는데 노력하였다. 대부분의 다른 품사 태거에서는 MATEC 을 위한 시스템을 따로 구축하거나 표준 태거와 동일하게 변형시켜 참가하였다. 다음 평가에서는 대회에 참가하는 시스템이 실제 시스템과 일관성이 떨어지는 팀에 대한 Penalty 가 있어야 한다. 마지막으로, 다음 평가에서는 테스트 시간을 각 시스템의 태깅 시간에 비례하여 가능한 적은 시간에 태깅하고 바로 결과를 제출함으로써 테스트의 공정을 유지할 수 있도록 해야하겠다. 이에 대한 사전 협의와 평가가 필요하다.

6. 결 론

본 논문에서는 입력 문장에 대해서 형태소 접속 그래프를 생성하면서 분석하는 통계적 품사 태거와 이것의 단점을 보완하는 에러 수정 규칙을 이용하는 강건한 품사 태깅 시스템을 소개하였다. 포항공대 자연어 처리 연구실에서는 혼합형 품사 태깅 시스템을 사용하여 MATEC 99 에 참가하였다. 본 연구실에서는 시스템의 일관성을 유지하면서, 표준 태그세트에 의해 제공되는 대량의 양질의 코퍼스를 사용하기 위하여 POSTAG 태그세트와 표준 태그세트와의 양방향 매핑을 구현하였다. 이번 대회에 참가하여, 표준 태그세트로 제공되는 태깅된 코퍼스를 POSTAG 에서 사용할 수 있도록 하였고, POSTAG 의 결과를 표준으로 출력함으로써 현재의 POSTAG 의 기준에서 대량의 코퍼스를 사용할 수 있는 환경을 구축한 것이 가장 큰 수확이었다. 다음으로 매핑으로 통해서 양쪽의 태깅 기준을 차이를 비교하면서 기존의 POSTAG 의 태깅 기준을 많이 튜닝 할 수 있었다. 마지막으로 학습 코퍼스와 표준의 형태소 분석 지침서를 참고로 하여 사전 미등록어를 많이 추가할 수 있었다. 이번 대회를 통해서 얻은 많은 결과에 대하여 다양한 태깅 안건에 대한 정확한 분석을 통해 능동적으로 표준 태그세트와 표준 형태소 분석 지침에 반영되기를 기대한다. 또한, 향후의 자연어 처리 평가 대회에서 본 연구실의 태그세트 매핑 방법이 상위레벨의 분석기 평가 과정에서 널리 이용되기를 기대한다.

참고 문헌

[1] E. Brill, " A simple rule-based part-of-speech tagger", Proceedings of the conference on applied natural language processing, Trento, Italy, pp. 153-155, 1992.

[2] E. Charniak, C. Hendrickson, N. Jacobson, and M. Perkowitz, "Equations of Part-of-Speech Tagging," Proceeding of Nat'l Conf. on Artificial Interlligence(AAAI-86) pp.784-789, 1993.

[3] 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.

[4] 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.

[5] J. Hughes, Clive Souter and Eric Atwell, 1994., "Automatic Extraction of Tagset Mapping from Parallel-Annotation Corpora," Center for the Computer Analysis of Langage And Speech, School of Computer Studies, Leeds University, [http://agora.leeds.ac.uk/amalgam/].

[6] J.P. Chanod, P. Tapananinen, "Statistical and constraint-based taggers for French," Technical Report MLTT-016, Rank Xerox Research Centre, Grenoble.

[7] Kenneth Ward Church, " A Stochastic Parts Program and Noun Phrase Parser for Un-restricted Text," Proceedings of applied natural language processing, Austin, Texas, 1988.

[8] 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.

[9] 강승식, “음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석,” 박사학위 논문, 서울대학교 컴퓨터 공학과, 1992.

[10] 김준석, 심준혁, 이근배, "품사 태그세트의 매핑을 이용한 한국어 품사 태거(POSTAG) 이식", 제 11 회 한글 및 한국어 정보 처리 학술발표논문집, 1999.

[11] 신상현, “TAKTAG: 통계와 규칙에 기반한 혼합형 한국어 품사 태깅 시스템,” 포항공과대학교, 전자계산학과, 석사 학위 논문, 1996.

[12] 이운재, “한국어 문서 태깅 시스템의 설계 및 구현,” 한국 과학 기술원, 전산학과, 석사 학위 논문, 1992.

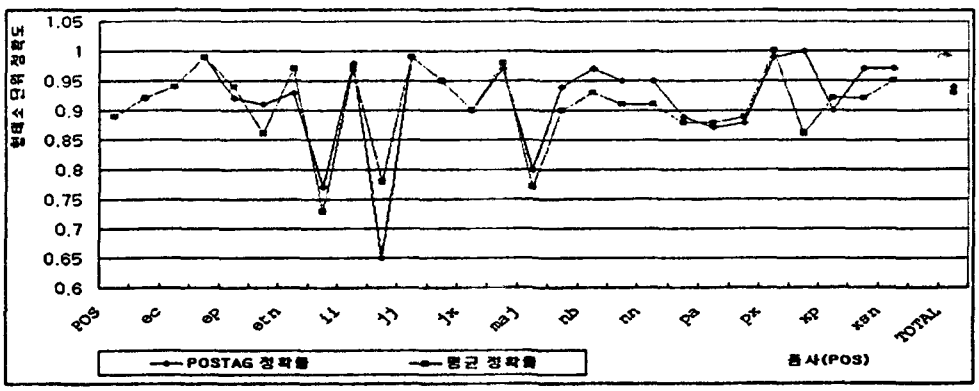
[13] 임철수, “HMM 을 이용한 한국어 품사 태깅 시스템의 구현,” 한국 과학 기술원, 전산학과, 석사 학위 논문, 1994.

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

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

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