응용을 위한 품사 태깅 시스템의 매핑[1]

김 준석, 차 정원, 이 근배
포항공과대학교 컴퓨터공학과 자연어 처리 연구실
경북 포항시 남구 효자동 산 31 번지

Application portable Part-Of-Speech tagger mapping

Junseok Kim, Jungwon Cha, Geunbae Lee
Natural Language Processing Lab.
Dept. of Computer Science & Engineering, POSTECH
{johan, jwcha, gblee}@nlp.postech.ac.kr

요 약

품사 태깅 시스템은 자연 언어 처리의 가장 기본이 되는 부분으로 상위 자연 언어 처리 분야인 구문분석, 의미분석의 전처리로 사용되거나, 기계번역, 정보검색이나 음성인식 및 합성등과 같은 많은 응용 시스템을 위해서도 필요하다. 이렇게 여러 가지 목적을 위해 품사 태깅 시스템은 존재하는데, 각각의 응용을 위해서 최적화된 태깅 시스템을 따로 구성하기도 하고, 하나의 태깅 시스템을 여러 가지 응용을 위해서 사용하기도 한다. 이때, 문제가 되는 것 중에 하나는 각 응용마다 요구하는 품사 태그 세트가 다르다는 것이다. 품사 태그세트가 고정되어 있다면 어떤 응용을 위해서는 사용되는 품사 태그세트가 너무 적어서 문제가 되고, 반대로 품사태그세트가 너무 많아서 시스템의 수행속도가 중요시되는 응용에서 성능저하의 요인이 되기도 한다. 본 논문에서는 하나의 태깅 시스템의 품사태그세트를 조절할 수 있도록 하여 몇 가지 응용시스템에 맞게 최적화 시킬 수 있는 방법론을 제시하고 실험을 통해서 시스템의 성능, 유지보수 및 시스템의 여러 리소스 관리 측면에서도 가장 효율적인 방법론임을 입증하고자 한다.

1. 서 론

품사 태깅에 사용되는 기본 단위들의 집합을 품사 태그 세트라고 하는데, 각 태깅 시스템들은 응용 목적에 적합한 서로 다른 문법 해석과 사전 정보를 바탕으로 고유의 품사 태그세트를 정의하고 있다[1][2][3]. 품사 태깅 시스템은 그 자체로도 의미가 있지만, 구문분석, 의미분석, 정보검색, 기계번역 및 음성처리 등 다른 응용 시스템을 위한 전처리 시스템으로도 많이 사용되고 있고, 이렇게 다양한 목적의 한국어 처리를 위해서 현재까지 많은 품사 태깅 시스템들이 개발되어 왔다. 품사 태그 세트에 속한 품사의 수를 품사 태그세트의 크기라고 할 때, 태깅 시스템의 사용목적에 따라서 요구되는 품사 태그세트의 크기는 달라질 수 있다. 예를 들어 구문분석기나 의미분석기를 위한 전처리로 태깅 시스템을 이용하고자 한다면 보다 상세하게 구분된 품사 태그들이 필요하므로 태그세트

[1] 본 연구는 과학재단 특정기초(1997.9-2000.8 #97-0102-03-01-3) 와 BK21(교육부 두뇌 한국사업)의 연구비 지원으로 수행되었습니다.

의 크기가 커지게 된다. 한편, 태깅 시스템이 정보검색을 위한 전처리로 사용될 경우에는 질의분석이나 색인을 위한 키워드 추출에 이용되기 때문에 위의 구문분석이나 의미분석과 같은 상세한 분류까지는 필요하지 않을 수 있다. 오히려 검색시스템의 수행속도와 메모리 사용등과 같은 효율성을 위해서는 보다 작은 크기의 태그세트를 원하게 된다. 품사 태깅 시스템의 오류를 분석해 보면 오류의 상당 부분은 품사의 세 분류에 따른 중의성 증가 때문이므로 태그세트의 크기를 줄이면 태깅 시스템의 정확도 향상에도 도움을 줄 수 있다[4][5]. 한편, 음성 처리 중에 TTS(Text to Speech) 나 기계 번역등과 같은 응용시스템에서는 그 나름대로 필요한 적당한 크기의 태그세트가 존재할 수 있다.

위와 같은 요구를 만족시키기 위해서 몇 가지 방법을 생각할 수 있다. 첫번째 방법으로 아주 큰 품사태그세트를 사용하는 태깅 시스템을 만들고 적은 품사태그세트만을 필요로 하면 태그를 병합해서 출력하여 사용하는 방법이 있을 수 있다. 두 번째 방법으로 각 응용에 맞는 태깅 시스템을 따로 구축하여 사용하는 방법을 들 수 있다. 그런데 위의 두 가지 방법은 각각 문제점을 안고 있다. 첫번째 방법은 작은 품사태그세트를 사용하는 응용시스템의 경우 큰 태그세트로 분석을 한 다음에 이를 이용하므로 처리 시간이 많이 소요되는 문제점이 있고, 두 번째 방법의 경우는 여러 가지 태깅 시스템을 구축하고 관리 해야 하는 오버헤드를 가지게 된다. 따라서 본 논문에서는 이러한 문제점을 해결하기 위해서 태깅 시스템을 사용하는 응용시스템에서 간단한 태그필터만 정의해 주면 원하는 태그세트를 가질 수 있는 태깅 시스템을 구현할 수 있는 방법론을 제시하고자 한다. 태깅 시스템이 서로 다른 크기의 태그세트를 가질 수 있으므로 다양한 응용에 사용될 수 있고 동시에 시스템이나 리소스 관리 측면에서도 유용한 장점을 가질 수 있게 되어 응용에 대한 태깅 시스템의 이식성(portability)를 높임을 논문의 목적으로 한다. 본 논문의 구성은 다음과 같다. 2 장에서는 방법론의 전체적인 구조를 살펴보고, 3 장에서는 태그세트 매핑에서의 주요 고려 사항들을 살펴보고, 4 장에서는 타당성을 입증하기 위한 여러 가지 실험 및 결과에 대해 고찰해 보고, 마지막으로 5 장에서 결론을 내린다.

2. 관련 연구

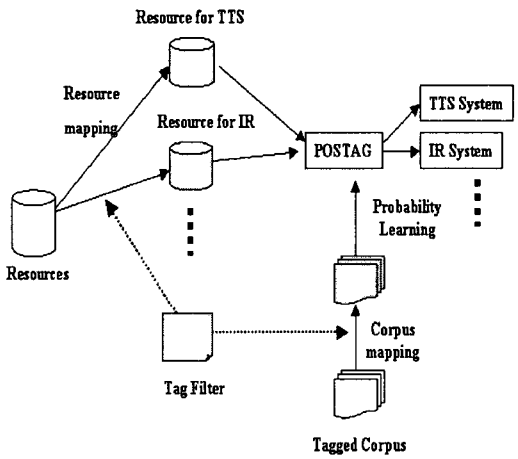
외국에서는 논문[8]에서와 같이 영어/프랑스어/스웨덴어에 대해서 태그세트와 태깅 정확도 간의 관계에 대한 실험을 수행한 것이 있었다. 태그세트 선택 시에 출력 코퍼스에서 필요로 하는 linguistic 한 품사나 구문의 구별을 할 수 있어야 하는 external criteria 와 가능하면 태깅을 효율적으로 수행할 수 있는가에 초점을 맞춘 internal criteria 로 구분하였다. 태그세트의 크기를 몇 가지 형태로 바꾼 다음에 태깅 결과와 미등록어 추정의 정확률을 3 가지 언어에 대해서 실험을 한 결과 반드시 태그세트의 크기가 작다고 좋은 성능은 보장되지 않았지만 언어에 따라서는 다른 결과도 있음을 보였다. 또한 태그세트를 변화시켜서 최적의 성능을 보이는 찾아내는 것보다는 태깅 시스템을 사용하는 응용에 최적화된 태그세트를 찾아야 함을 주장하였다.

한국어에 대해서는 [9]에서 기계번역을 위한 품사태그세트를 정의하였고, [5]에서 단순화된 태그세트와 세분화된 태그세트로 나누어서 성능을 평가하여 단순화된 태그세트가 성능이 더 높은 정확도를 보인 예가 있으나 아직 한국어에 대해서 본 논문에서와 같이 태그세트를 응용시스템에 맞게 변경 시키는 것에 관련된 연구는 아직 없었다.

3. 전체 구조

[그림 1]은 본 논문에서 제안하는 태그필터를 이용하여 응용에 알맞은 태그세트로의 매핑을 하는 방법론을 보여주는 그림이다. 우선, 품사 태깅 시스템을 사용하는 TTS 나 정보검색과 같은 각 응용 시스템들은 어떠한 태그세트를 사용하는지를 명시하는 간단한 태그필터를 작성해야 한다. 태그 필터를 이용해서 크게 2 가지 일을 하는데 첫번째가 품사가 부착된 코퍼스로부터 태그필터와 확률 학습기를 이용해서 매핑된 코퍼스를 만들고 이로부터

POSTAG[2]가 필요로 하는 어휘 확률 및 문맥 확률을 비롯하여 미 등록어 추정을 위한 음절 트라이그램등과 같은 확률정보를 학습한다 [6]. 다음으로, 품사 태깅 시스템에서 사용되는 다양한 사전류, 접속정보 테이블 등의 여러 리소스들이 태그 필터와 매핑 프로그램을 이용하여 자동으로 바이너리 형태로 컴파일 시킨다. 태깅 시스템은 매핑된 리소스와 학습된 확률 값을 이용하여 각 응용에 최적화된 태그세트를 가지는 태깅 시스템의 결과를 이용할 수 있게 된다. 위의 과정을 원하는 태그세트를 찾을 때까지 어느 정도의 반복 과정을 수행하여 응용 시스템에 최적화된 태그세트를 찾으면 된다. 태깅 시스템을 이용하는 응용 시스템의 측면에서 보면 태그필터만 수동으로 제작해 주면 나머지 과정을 모두 완전 자동으로 처리할 수 있도록 시스템을 제작하여 태깅 시스템의 사용을 쉽도록 하였고, 하나의 태깅 프로그램만을 사용하여 여러 가지 응용 시스템에 맞는 품사 태깅 시스템을 구현하여 프로그램 소스레벨에서 여러 버전의 태깅 시스템을 구현하는 번거로움을 해소 했고, 하나의 리소스에서 모두 매핑을 통해서 각 응용 시스템에서 사용하도록 하여 리소스 관리의 효율을 높였다.



[그림 1] 전체 구조도

3. 주요 고려사항

3.1 태그필터

POSTAG는 총 73개의 태그로 구성된 태그세트를 가지는데 계층적인 구조를 가진다.[3] 99년 표준태그로 사용된 태그세트와는 매핑을 통해서 잘 호환됨이 입증된 품사 태그세트 이다 [7]. 태깅 시스템을 사용하는 응용 시스템의 경우 태그필터를 어떻게 정의 하느냐가 중요하다. [표 1]는 본 연구실의 정보검색 시스템에서 사용하는 태그필터의 일부분 인데 구성은 표에서와 같이 왼쪽에 전체 태그들이 있고 오른쪽에 매핑 될 태그들이 위치한다. 가장 먼저 고려 되어야 할 것은 응용 시스템을 위해서 반드시 구분이 필요한 태그들은 품사 태그세트에 포함이 되어야 한다는 점이다. 나머지는 태깅 시스템의 수행속도와 정확도를 고려하여 태그세트의 크기를 줄일 수 있는데 어떤 품사 태그를 줄였을 경우 시스템의 성능에 나쁜 영향을 줄 수 있음을 고려해야 한다는 점이다. 예를 들어서 어미의 경우는 계층을 너무 올렸을 경우 시스템의 전체 정확도가 떨어지는 결과를 초래하였다. 각 응용 시스템을 위해서 꼭 필요한 태그들과 몇 가지 조정할 수 있는 태그의 계층을 조정하면서 시스템의 수행속도와 정확도를 고려하여 현재 본 연구실에서는 구문분석 시스템은 73개 전체 태그를 사용하고 TTS는 42개, 정보검색 시스템은 35개의 태그를 각각 사용하고 있다.

[표 1] 정보검색 시스템용 태그필터 중 일부

MCC(한자어 보통명사)	MC(보통명사)
MCK(한국어 보통명사)	MC(보통명사)
MCF(외래어 보통명사)	MC(보통명사)
GC(성상 관형사)	G(관형사)
GP(지사관형사)	G(관형사)
GS(수량관형사)	G(관형사)
BD(문장연결 부사)	B(부사)
BT(시간 부사)	B(부사)

[2] 포항공대 자연언어처리 연구실에 사용하는 품사 태깅 시스템의 이름

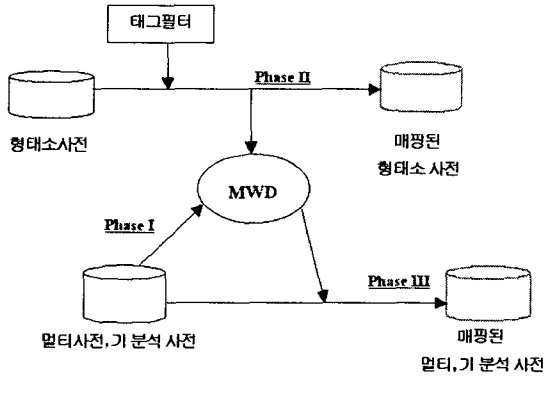
[3] [부록 1] 에 POSTAG의 전체 품사 태그세트와 현재 정보검색 시스템에서 사용하는 태그세트를 표시하였다.

BM(양태 부사)	B(부사)

3.2 리소스

태깅 시스템을 위해서는 많은 리소스들이 사용되는데 크게 사전, 접속 테이블, 확률 테이블로 구분할 수 있다.

우선, 사전의 경우 형태소 사전을 비롯하여 미 등록어 추정을 위한 패턴사전, 다 형태소 분석의 정확도를 높이기 위해서 도입된 다 형태소 사전, 분석된 결과를 가지고 있는 기 분석 사전등과 같은 여러 가지 목적의 사전들이 있는데 다 형태소 사전 및 기 분석 사전의 경우 형태소 사전을 바탕으로 구성이 되어 있다. 사전의 매핑의 경우 [그림 2]와 같은 방법을 이용하는데 우선 다 형태소 사전 및 기 분석 사전에 있는 모든 엔트리를 MWD(Multi Word Dictionary)에 저장을 하고 태그 필터를 이용하여 기본 형태소 사전을 매핑을 한다. 이때 MWD를 참조하여 MWD에 속하는 경우 그 변화를 MWD에 기록한다. 다음에 다 형태소 사전 및 기 분석 사전을 매핑할 때에 MWD를 참조하여 매핑을 수행하면 된다. 또한 매핑 전에는 서로 다른 엔트리였는데 매핑 후에는 같은 엔트리가 될 경우에 각각이 가지는 접속정보를 서로 더해 주는 것이 필요하다.



[그림 2] 멀티, 기 분석 사전의 매핑 예

현재 POSTAG에서는 전체 73개의 태그세트로 구성되어 있고 형태소 사전의 중의성이 있는 엔트리의 형태소 당 평균 중의 율이 2.34인데 35개의 태그세트를 사용하는 정보검색용 태그필터를 사용한 후 형태소 사전의 평균 중의 율은 2.18으로 떨어졌고, 형태소 엔트리 개수 또한 많이 줄어들었다. 예를 들어 MCC(한자어 보통명사), MCK(한국어 보통명사)는 정보검색에서는 굳이 구별할 필요가 없어서 모두 MC(보통명사)로 매핑 한다. 이때, 품사가 MCC인 사내(社內)와 MCK인 사내(남자)는 모두 MC 사내가 되면서 중의성은 낮아지게 된다. 또한 사내(社內)가 가진 접속정보 "[무>H 답>H 스럽>H 같>]" 과 사내(남자)의 "[무>적>]"은 서로 결합되어 "[무>H 답>H 스럽>H 같>적>]"이 된다.[4] 형태소 사전 외에도 미등록어 추정을 위한 패턴 사전이 사용되는데 형태소 분석 시 사전에 없는 미등록어가 나왔을 때 사용되는데 추정 되어야 할 품사가 매핑한 후에는 줄어들기 때문에 수행 속도면에서 빨라지게 된다.

두 번째 리소스로 접속테이블을 들 수 있다. 형태소 분석 과정을 크게 나누어 보면 Segmentation, 원형 복원, 접속 검사로 나눌 수 있는데, 사전이나 Rule을 이용하여 Segmentation 및 원형 복원을 수행하고 접속 테이블을 이용하여 여러 후보 형태소간의 접속을 검사하여 후보노드의 수를 줄인다. 접속 테이블은 태그정보, 어휘 정보 및 형태소 사전에 기록된 각 형태소가 가지는 여러 가지 접속 feature 들로 구성 되어 있다. 태그 필터를 이용하여 자동으로 매핑을 수행하여 매핑된 접속테이블을 만드는데, 태그정보의 경우 매핑 되면서 기존에 가졌던 specific 한 정보를 잃어 버릴 수 있기 때문에 더 이상 사용할 수 없게 된다. 이러한 접속 정보는 테이블에서 제거되어야 한다.

확률 테이블은 [그림 3]의 수식을 사용하는데, 응용 시스템에서 새로운 품사 태그세트를 정의하면 품사가 부착된 코퍼스로부터 우선 태그필터를 이용하여 코퍼스를 매핑한 후에 태깅 시스템에서 필요로 하는 어휘 및 문맥 확률 정보를 학습한다. POSTAG의 태깅 코퍼스의 형태는

[4] 무: 조사결합에 이용, H 답: 오른쪽에 '답'이 와서 형용사가 될 수 있음(사내답다), 적: 뒤에 '적'이라는 접미사가 붙을 수 있다.

'태그<주형태>(변이형태)'의 형식인데, 고려해 줄 것은 같은 변이 형태를 가지는 형태소의 서로 태그가 태그필터를 통해서 하나의 태그로 매핑이 될 때 주형태가 다른 것을 학습 시에 고려해 주어야 한다는 것이다. 예를 들면 다음과 같다. 만약 태그필터에서 DR(규칙동사)와 DI(불규칙 동사)를 모두 D(동사)로 매핑 한다고 할 때, '감기가 나으니'의 DI<낫>(나)와 '나는 새'에서의 DI<날>(나), '여름을 나다' 에서의 DR<나>(나)은 모두 같은 어휘로 매핑이 되므로 품사가 D(동사)인 '나'에 대한 어휘 확률 계산 시에 이를 반영해 줘야 한다.

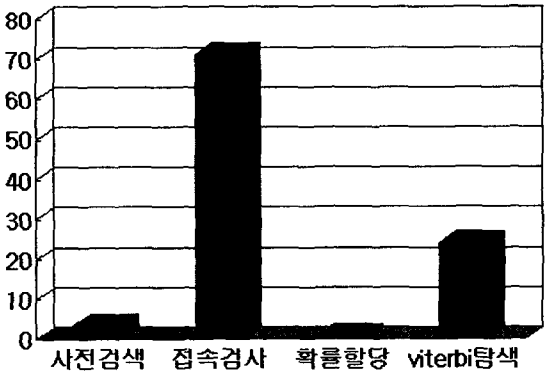
$$T^* = \arg\max_{T} \prod_{i=1}^{n} \Pr(t_i \mid t_{i-1})^{\alpha} \left(\frac{\Pr(t_i \mid m_i)}{\Pr(t_i)} \right)^{\beta}$$

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

[그림 3] POSTAG 확률 모델

4. 실험

실험은 크게 태깅 시스템의 분석속도 측면과 분석 정확도 부분으로 나누어서 시행하였다. 우선 분석속도 측면부터 살펴 보겠다. [그림 4]는 POSTAG의 수행 중 어느 부분에서 가장 load가 심한지를 측정하여 상대적인 비율을 그래프로 나타낸 결과이다.



[그림 4] POSTAG 실행 속도 분석표

그림에서 보는 것처럼 형태소 후보들간에 접속 유무를 검사하는 접속검사 부분에서 가장 많은 시간이 소요되었다. 이것은 형태소 후보가 너무 많이 만들어지는 것과 밀접한 관련이 있다. 따라서 수행속도를 향상시키기 위해서 형태소 후보의 수를 줄이는 것이 필수적이다. 실험은 4000문장, 33,700어절의 한국어 문장에 대해서 수행하였고, 본 연구실에서 연구하는 TTS와 정보검색 시스템에서 사용되는 태그필터에 대한 POSTAG의 수행결과를 분석하였다. [표 2]는 TTS와 정보검색응용 시스템에서 사용하는 태그필터를 이용하여 매핑을 했을 때의 수행속도 및 형태소 노드의 수를 보여주는데 태그세트의 크기에 비례함을 볼 수 있다. 이는 형태소 중의성이 낮아지고, 추정하는 미등록어의 개수가 적어지기 때문에 후보 그래프 노드의 수가 줄어 들어서 접속검사의 횟수가 많이 줄어 들었기 때문이라고 분석할 수 있다.

[표 2] 태그필터 별 실행 시간 분석

응용분야	구문분석	정보검색	TTS
사용 태그 수	73	35	42
전체 수행 시간 (초)	4640.05	1973.96	2928.86
사전 탐색 (초)	139.25	104.46	95.09
접속검사 (초)	3364.96	1469.81	2185.45
Viterbi 탐색 시간(초)	1007.63	356.95	610.57
그래프 노드 수 (개)	2,481,291	1,296,430	1,907,622
접속 검사 횟수 (회)	8,654,288	4,575,937	6,458,750

두 번째 실험으로 대용량의 문장에 대한 정확도 실험을

수행하였다. 실험은 MATEC[5]에서 사용되었던 25,000 문장, 240,000 어절에 대해서 내부실험과 10 개로 나누어서 9 개에 대해서 학습하고 1 개에 대해서 실험을 하는 외부실험으로 진행하였다. [표 3]의 내부 실험 결과로 가장 작은 태그세트를 이용하는 정보검색 시스템용 태그필터를 사용하는 태깅 시스템의 정확도가 전체 태그세트를 모두 사용하는 것에 비해 어절 단위로 약 0.67%, 형태소 단위로 0.85%의 정확도 향상을 보였다. 형태소 중의성이 낮아지고 추정할 미등록어의 개수도 적어져서 정확도는 향상 되었지만 속도면에서와 같이 크게 향상되지는 않았다. [표 4]는 외부실험결과를 보여주는데 정확도 향상 정도가 크지는 않지만, 작은 태그세트를 사용하는 응용 태깅 시스템의 정확도가 조금 높음을 확인 하였다.

[표 3] 내부 실험 결과

응용분야	구문분석	정보검색	TTS
사용 태그 수	73	35	42
어절 정확도	92.21	92.88	92.47
형태소 정확도	95.34	96.21	95.64

[표 4] 외부 실험 결과

응용	구문분석	정보검색	TTS
정확도	어절/형태소	어절/형태소	어절/형태소
실험 1	92.12/95.28	92.87/96.13	92.44/95.52
실험 2	92.07/95.13	92.56/95.89	92.48/95.33
실험 3	91.98/94.52	92.34/95.66	92.18/94.93
실험 4	91.86/94.48	92.01/95.45	91.95/94.88
실험 5	92.05/95.01	92.28/95.53	92.11/95.25
실험 6	92.11/95.17	92.33/95.66	92.29/95.48
실험 7	91.95/94.44	92.00/95.49	91.99/95.04
실험 8	92.05/94.77	92.09/94.96	92.06/94.81
실험 9	91.77/94.32	92.12/95.33	92.09/95.02
실험 10	91.95/94.72	92.29/95.46	92.17/95.03
평균	91.99/94.78	92.28/95.55	92.17/95.12

[5] 한국전자통신원(ETRI)에서 주관한 제 1 회 한국어 형태소분석, 태깅, 명사추출 대회

5. 결 론

실험을 통해서 태깅 시스템을 이용하는 응용 시스템에서 태그필터를 이용해서 원하는 태그세트로 출력 결과를 냄으로써, 전체적인 시스템의 속도와 정확도 두 가지 측면에서 유용함을 보였다. 그 결과, 정밀한 품사 태그들의 구분을 필요로 하는 구문분석에서는 큰 태그세트를 이용하고, 수행속도를 중요시하는 정보검색과 같은 응용시스템에서는 키워드를 추출하는데 필요한 작은 태그세트를 간단하게 태그필터 만을 정의해 주면 나머지는 모두 자동으로 매핑을 통해서 태깅 시스템에서 원하는 결과를 얻을 수 있도록 하였다. 이와 시스템을 이용하여 응용 시스템에서는 각자에 맞는 최적화된 태그세트를 얻는데 도움을 받을 수 있을 것이다. 앞으로 해야 할 일은 태그세트를 변화 시킬 때에 따른 태깅 시스템 자체의 성능도 중요하지만 태깅 시스템을 사용하는 응용 시스템의 성능 변화도 중요하므로 그에 대한 실험이 필요 하고, 각 품사별로 응용 시스템에 미치는 영향에 대한 보다 상세한 분석이 필요하다.

참고 문헌

[1] 안 미정, 김재한, 옥철영 "한국어 처리를 위한 품사 체계 연구", 제 5 회 한글 및 한국어 정보처리 학술발표 논문집, 1993

[2] 시스템 공학 연구소, "한국어 품사 태그 세트 검토 및 보완", 우리말 정보처리 규격 심포지움, 1997

[3] 한국전자통신연구원 컴퓨터·소프트웨어 기술 연구소 지식정보연구부, "품사 부착 말뭉치 구축 지침서", 1999,:

http://aladin.etri.re.kr/~nlu/STANDARD/

[4] 남지순, 최기선 "어절 정보 사전을 이용한 형태소 분석의 중의성 해결", 제 9 회 한글 및 한국어 정보처리 학술대회, 1997

[5] 김진동, 이상주, 임해창 "어절 띄어스기를 고려한 형태소 단위 품사 태깅 모델", 제 10 회 한글 및 한국어 정보처리 학술대회, 1998.

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

[7] 김준석, 심준혁, 이근배 "품사 태그 세트의 매핑을 이용한 한국어 품사 태거 이식", 제 11 회 한글 및 한국어 정보처리 학술대회, 1999.

[8] David Elworthy, "Tagset Design and Inflected Languages", Proceeding of the ACL SIGDAT Workshop, 1995

[9] 송재관, 박찬곤 "한 영 기계번역을 위한 한국어 품사분류", 한국정보과학회 가을 학술발표집, 1998

<table>
<tr><td colspan="2">기호(s)</td><td colspan="4">쉼표(s,) 문장종결(s.) 여는 따옴표(s`) 닫는 따옴표(s') 이음표(s-) 단위(su) 한자어(sh) 외국어(sf) 기타(so)</td></tr>
<tr><td rowspan="6">체언</td><td rowspan="3">명사(M)</td><td colspan="2">보통명사(MC)</td><td colspan="2">한국어(MCK) 한자어(MCC) 외래어(MCF)</td></tr>
<tr><td colspan="4">고유명사(MP)</td></tr>
<tr><td colspan="4">의존명사(MD)</td></tr>
<tr><td rowspan="3">대명사(T)</td><td colspan="2">정칭(TC)</td><td colspan="2">사람(TCH) 사물(TCT)</td></tr>
<tr><td colspan="2">재귀(TS)</td><td colspan="2">사람(TSH) 사물(TST)</td></tr>
<tr><td colspan="2">의문/부정(TO)</td><td colspan="2">사람(TOH) 사물(TOT)</td></tr>
<tr><td>수사(S)</td><td colspan="4">양수사(SC) 서수사(SO)</td></tr>
<tr><td rowspan="5">용언</td><td>동사(D)</td><td colspan="4">규칙(DR), 불규칙(DI)</td></tr>
<tr><td>형용사(H)</td><td colspan="4">규칙(HR), 불규칙(DI)</td></tr>
<tr><td rowspan="3">보조용언
(b)</td><td colspan="2">동사(bD)</td><td colspan="2">규칙(bDR) 불규칙(bDI)</td></tr>
<tr><td colspan="2">형용사(bH)</td><td colspan="2">규칙(bHR) 불규칙(bHI)</td></tr>
<tr><td colspan="2">존재사(bE)</td><td colspan="2"></td></tr>
<tr><td>존재사(E)</td><td colspan="4"></td></tr>
<tr><td>지정사(I)</td><td colspan="4"></td></tr>
<tr><td rowspan="2">수식언</td><td>관형사(G)</td><td colspan="4">성상(GC) 지시(GP) 수량(GS)</td></tr>
<tr><td>부사(B)</td><td colspan="4">문장연결(BJ) 과정(BP) 부정(Bb) 공간(BS) 시간(BT) 양태(BM) 수량(BN) 정도(BD)</td></tr>
<tr><td>독립언</td><td>감탄사(K)</td><td colspan="4"></td></tr>
<tr><td>관계언</td><td>조사(j)</td><td colspan="4">격조사(jC) 보조사(jS) 기타조사(jO)</td></tr>
<tr><td rowspan="5">어미(e)</td><td rowspan="5"></td><td rowspan="2">문법어미(eG)</td><td>어말어미(eGE)</td><td colspan="2">서술(eGEs) 의문(eGEu) 명령(eGEm)
청유(eGEc) 약속(eGEy) 미정(eGEx)</td></tr>
<tr><td>선어말어미(eGS)</td><td colspan="2">시제(eGSt) 존재(eGSp) 양태(eGsm)</td></tr>
<tr><td rowspan="3">복합문(eC)</td><td rowspan="3">내포문(eCN)</td><td>동사구내포(eCND)</td><td>인용(eCNDC)
보조적(eCNDI)</td></tr>
<tr><td>명사구내포(eCNM)</td><td>명사형전성(eCNMM)
관형사형전성(eCNMG)</td></tr>
<tr><td>부사구내포(eCNB)</td><td></td></tr>
<tr><td></td><td>접속문(eCC)</td><td colspan="2"></td></tr>
<tr><td rowspan="3">접사</td><td>접두사(+)</td><td colspan="4">명사결합(+M)</td></tr>
<tr><td>접미사(-)</td><td colspan="4">고유명사(-N) 복수(-b) '적'(-J) 수량(-S) 기타(-O)</td></tr>
<tr><td rowspan="2">조용사(y)</td><td colspan="2">동사(yD)</td><td colspan="2">규칙(yDR) 불규칙(yDI)</td></tr>
<tr><td colspan="2">형용사(yH)</td><td colspan="2">규칙(yHR) 불규칙(yHI)</td></tr>
</table>

[부록 1] POSTAG 품사 전체 태그 세트와 정보검색 태그 세트