단순화된 어절을 단위로 하는 한국어 품사 태거

이의현[O+], 김영길[#], 신재훈[+], 권홍석[+], 이종혁[++]
포항공과대학교 컴퓨터공학과[+], 한국전자통신연구원[#]
{eh_lee[O+], rave0206[+], hkwon[+], jhlee[++]}@postech.ac.kr, kimyk[#]@etri.re.kr

A Korean Part-of-Speech Tagger using Simplified Eojeol-based unit

Eui-Hyeon Lee[O+], Young-Gil Kim[#], Jaehun Shin[+], Hong-Seok Kwon[+], Jong-Hyeok Lee[++]
Pohang University of Science and Technology, Department of Computer Science & Engineering[+]
Electronics and Telecommunications Research Institute[#]

요 약

영어권 언어가 어절 단위로 품사를 부여하는 반면, 한국어는 굴절이 많이 일어나는 교착어로서 데이터부족 문제를 피하기 위해 형태소 단위로 품사를 부여한다. 이러한 구조적 차이 안에서 한국어에 적합한 품사 태깅 단위는 지속적으로 논의되어 왔으며 지금까지 음절, 형태소, 어절, 구가 제안되었다. 본 연구는 어절 단위로 태깅함으로써 야기되는 복잡한 품사 태그와 데이터부족 문제를 해소하기 위해 어절에서 주요 실질 형태소와 주요 형식 형태소만을 뽑아 새로운 어절을 생성하고, 생성된 단순한 어절에 대해 CRF 태깅을 수행하였다. 실험결과 평가 말뭉치에서 미등록 어절 등장 비율은 9.22%에서 5.63%로 38.95% 감소시키고, 어절단위 정확도를 85.04%에서 90.81%로 6.79% 향상시켰다.

주제어: 한국어 품사 태거, 태깅 단위, 데이터 부족 문제, 교착어

1. 서론

품사태그 부착 혹은 품사 태깅(POS tagging)은 여러 가지 품사를 가질 수 있는 단어에 대해 그 애매성을 해소해주는 작업으로 자연어 처리 기술의 가장 기본 단계이다. 따라서 기계번역, 정보검색 등 응용 정보처리시스템의 기반기술로서 높은 정확도가 요구된다.

그간 주 연구 대상이었던 영어는 굴절형 자체에 형태소 정보가 접합된 품사태그를 부여한다. 예를 들어 'ate'에는 VBD(동사, 과거시제), 'apples'에는 NNS(명사, 복수형) 태그를 부여한다. 그럼에도 영어는 형태적으로 빈약한(morphologically-poor) 언어로서 문법정보가 단어보다는 어순으로 표현되기에 어절 내부 구조가 간단해 품사가 48~135가지로 상대적으로 적다[1]. 또한 파생되는 어절 수가 많지 않아 데이터 부족(data sparseness) 문제에서 비교적 자유롭다. 이 때 품사 태깅은 전형적인 시퀀스 레이블링 문제가 된다.

반면 형태적으로 풍부한(morphologically-rich) 언어에서는 빈번한 굴절현상으로 품사태그에 접합되는 형태소 정보가 많고, 이는 품사 집합의 크기를 과도하게 키워 전체적 성능을 떨어뜨린다. 반대로 품사 가짓수를 제한하면 많은 문법정보를 잃게 된다. 대표적으로 독일어는 Stuttgart-Tubingen Tagset에서 54개의 품사를, 형태론적 변형은 783종류를 가지며[2], 비슷하게 불가리아어와 체코어는 BulTreeBank와 Prague Dependency Treebank에서 각각 680, 1400여개 품사를 가진다[1]. 또한 파생되는 어절수가 많아 심각한 데이터 부족 문제를 겪는다.

형태적으로 풍부한 언어에 속하는 한국어에서도 같은 문제점이 지적되어왔다[3,4]. 하지만 한국어는 교착어(agglutinative language)로서 어근으로부터 접사를 분리하기에 용이해, 형태소태그를 정의하고, 어절이 아닌 형태소에 태그를 부착한다. 엄밀히 말해 형태소태그는 전통적 품사태그와 구분되지만 관습적으로 품사태그라 불리어왔고, 본 논문에서도 이후 편의상 품사태그라고 부르도록 한다.

따라서 한국어 품사 부착기는 일반적으로 모든 가능한 형태소열을 생성하는 형태소 분석기(Morphological Analyzer: MA)와 문맥을 고려해 최적 형태소열을 선택하는 품사 부착기(POS tagger)로 구성된다.

이러한 영어권 언어와 한국어 분석의 구조적 차이 때문에 기존 방법론은 한국어에서 문제를 일으킨다. 예를 들어 형태소 단위 태깅은 좁은 문맥, 낮은 문맥 확장성을 가지며, 후보열 간 차이로 인한 여러 편향 문제를 겪는다.

따라서 품사 태깅 시 처리 단위에 대한 논의는 지속적으로 이루어졌다. 지금까지 음절, 형태소, 어절, 구 기반 방식이 제안되었고, 그 중 어절 기반 방식 연구는 어절을 단위로 삼아 생기는 복잡한 품사 태그와 데이터부족 문제를 해소하는데 목적을 둔다.

본 연구도 이러한 관점에 초점을 같이한다. 기존 방법론은 임의 형태소 길이를 가지는 어절을 그대로 사용했지만, 본 연구에서는 언어학적 지식을 동원해 핵심 실질 형태소와 핵심 형식 형태소, 2형태소만을 가지는 어절로 단순화하고 태깅하여 데이터부족 문제를 해소하였다.

2. 관련 연구

2.1. 형태소 단위 태깅

형태소 단위 태깅은 데이터부족 문제를 완화시키고 간단한 품사 태그를 갖는다. 하지만 현실적으로 확률 모델은 마르코프 가정(Markov assumption)으로 참조 문맥 길

이를 제한하는데, 더 작은 단위인 형태소는 더 좁은 문맥을 갖고, 확장성이 떨어지며, 어절 정보를 무시한다. 또한 어절의 분석 형태소열들은 상이한 분기와 길이를 가질 수 있어 후보 간 경로 분기와 길이가 다를 때 선택이 편향되는(label & length bias) 문제가 일어날 수 있다[5]. [5]는 일본어 문장을 대상으로 사전 기반 lattice를 구성하고, 형태소 단위 품사 태깅을 수행했으며, global normalization을 하는 CRF가 레이블 및 길이 편향 문제를 해소한다고 주장했다.

하지만 [5]의 방법론을 한국어에 적용해본 결과, 대신 기능어를 포함한 경로를 선택하는 경향이 있었다. 예를 들어 “현물 거래”는 “현무를 거래”의 준말이라고도 볼 수 있는데, 전자가 후자에 비해 학습데이터에 압도적으로 많이 등장함에도 불구하고 기능어인 ‘를’의 발생빈도 및 ‘를 위하’, ‘명사+목적격조사’ 등의 발생빈도가 내용어에 비해 압도적으로 많아 다른 자질로 극복하지 못하게 된다. 일본어는 띄어쓰기가 없는 대신 내용어와 기능어에 다른 문자체계(한자&가타카나/히라가나)를 사용해 구분이 명백하지만 한국어는 애매해 기능어 편향 문제를 갖는다.

2.2. 음절 단위 태깅

음절 단위 태깅은 형태소 단위 태깅에서 나타나는 문제를 해결하려는 목적보다는, 형태소 분석 단계를 배제하기 위해 연구되었다. 형태소 분석은 사전 및 접속정보 등 구축비용이 드는 언어 자원을 요구하고, 오류를 전파한다[6]. 특히 SNS나 메신저 등 띄어쓰기 정확성의 신뢰도가 떨어지는 입력문에 대해서 심각한 성능 하락을 일으킨다[7].

그 외에도 음절 단위 태깅은 미등록어를 잘 추정하고, 비문법적인 입력문에도 강하며, 모든 후보군의 표층형이 동일해 문맥 확장성이 크다[8]. 하지만 언어자원을 활용하지 않아 인위적인 수정이 불가능해 유지보수가 어렵다. 단적으로 사전 추가로 해결되는 형태소 단위 태깅과 달리 해당 어휘가 포함된 학습말뭉치를 추가해 재학습시켜야 한다. 또한 음절 자체는 언어학적 정보를 가지고 있지 않아 형태소 단위보다 넓은 문맥참조와 다양한 자질을 요구하고, 매 음절마다 모든 태그에 대해 확률값 계산을 하기 때문에 일반적으로 다른 단위 태깅 방법보다 매우 느린 속도를 가진다.

2.3. 구 단위 태깅

본 논문에서 말하는 구(phrase)는 임의의 형태소열을 의미하며, 언어학에서 말하는 구와 차이가 있다.

[9]는 품사 태깅을 문장에서 품사 열로의 번역으로 보고 구 기반 통계번역(Phrase-based Statistical Machine Translation, PBSMT) 모델을 사용했다. 하지만 이 경우에도 여전히 길이 및 기능어 편향 문제를 갖는다.

2.4. 어절 단위 태깅

어절 단위 태깅은 문맥 확장성이 좋고, 빠른 속도와 높은 정확도를 가지며, 어절 정보를 활용할 수 있다[3,8,10]. 또한 기존 영어권의 연구와 같은 구조를 가져 기타 단위 태깅에서의 부작용에서 자유롭다. 하지만 앞서 언급한 것처럼 어절의 구성이 다양해 큰 태그집합을 가지고, 데이터부족 문제에 민감하다.

[3]은 복잡한 어절 품사 태그에서 어절 사이의 구문적 의존 관계에 무관한 품사를 제거해 단순화했다. 이 때 태그의 복잡도는 줄어들지만 어절이 그대로 유지돼 데이터부족 문제는 해소되지 않는다.

[11]는 복잡한 태그를 유지하되, 데이터가 부족할 경우 단계별로 전이 모델을 단순화했다. 예를 들어 어절 대 어절 전이 정보가 학습되지 않았을 경우 어절 대 시작 형태소, 마지막 형태소 대 시작 형태소 순으로 단순화한다. 이 과정에 전이확률의 가중치 등 반복 실험을 통해 결정되는 파라미터가 많아 재연이 어렵다.

본 논문의 방법론과 가장 비슷한 연구는 [10]이다. 본 연구는 [10]와 같이 어절을 형태소 분석해 핵심 실질 형태소와 핵심 형식 형태소를 추출 및 결합해 길이 2짜리 단순한 어절 및 어절 품사 태그로 변환해 태깅을 수행한다. 하지만 본 논문은 핵심 실질 형태소 및 핵심 형식 형태소에 대한 정의가 다르며, 시퀀스 레이블링 모델인 CRF를 사용했다.

3. 실험 방법

3.1. 전체 구성도

실험에서 구현한 프로그램은 크게 형태소 분석기, 어절 단순화 모듈 그리고 품사 태거로 이루어졌다. 그림 1은 전체 프로그램의 도식도이다.

3.2. 형태소 분석기

형태소 분석기에는 축약 및 불규칙 활용이 빈번한 한국어에 적합하다고 알려진 음절단위 CYK 알고리즘을 사용했다[12]. 태그집합은 세종 태그집합과 다른 언어철학을 가지고 설계된 KLE 태그집합을 사용했다. KIBS 태그집합으로부터 만들어졌으며 어근, 지정사 등을 인정하지 않은 것이 차이점이다. 표 1와 같이 세종 태그집합보다 세분화되어 총 68개의 태그를 갖는다.

표 1 세종 및 KLE 태그집합 비교(일부)

세종 태그	KLE 태그
보통명사	{동작서술성, 상태서술성, 비서술설} 보통명사
선어말어미	{과거시제, 미래시제, 피동, 사동, 회상, 겸양, 존칭}형 선어말어미
종결어미	{평서, 의문, 명령, 청유, 약속}형 종결어미

사용된 언어자원은 형태소 사전 및 음절 축약 패턴, 접속정보이다. 형태소 사전은 약 42만 항목을 가지며, 이는 일부 복합 형태소 및 접속정보와 KLE 태그집합을 표현하기위해 중복된 형태소들의 합이다.

3.3. 어절 단순화 모듈

어절 단순화 모듈은 주어진 임의 길이 형태소 열에서 핵심 실질 형태소와 핵심 형식 형태소를 선택해 길이 2 형태소 열로 바꿔주는 모듈이다. 핵심 실질 형태소와 핵심 형식 형태소는 다음과 같이 정의된다.

- 핵심 실질 형태소: 어절 마지막 체언, 용언 및 상당어구.

상당 어구란 굴절 등의 현상으로 특정 문법 범주가 되는 형태소열을 말한다. 예를 들어 체언 상당 어구는

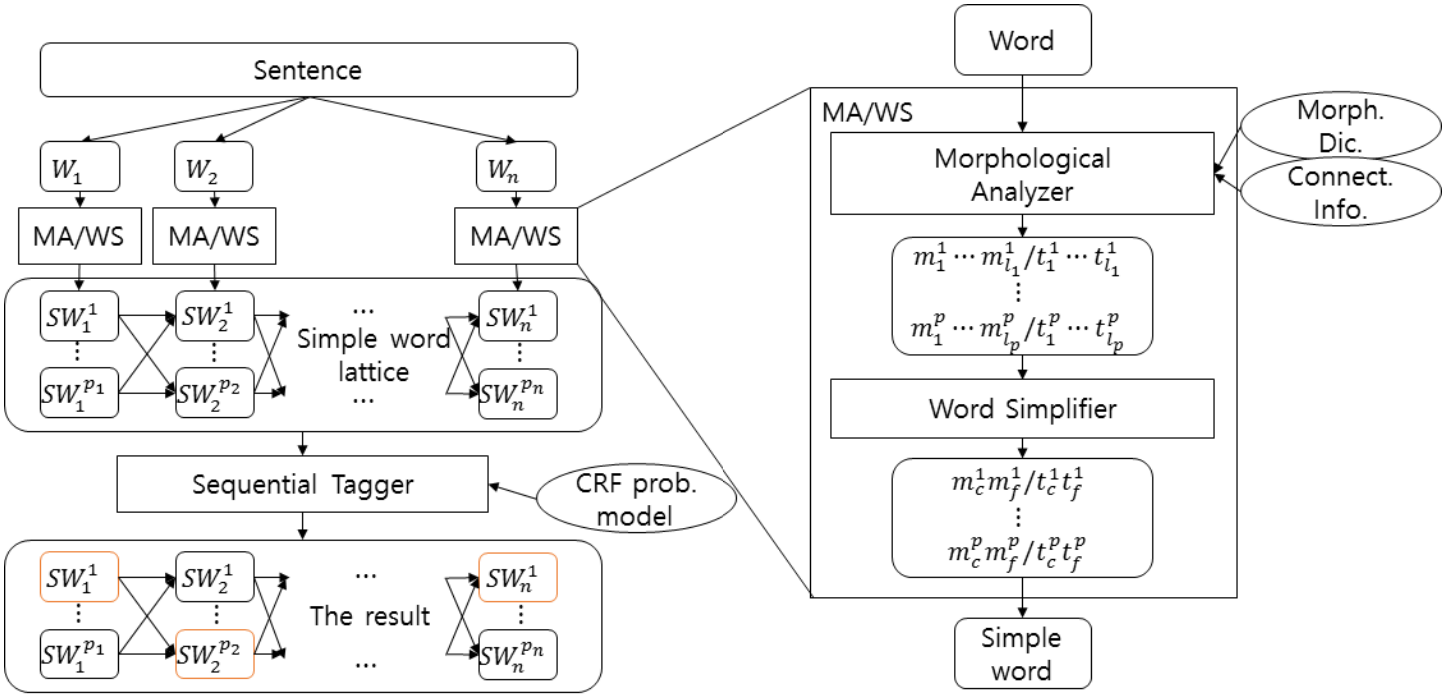

그림 1 전체 구성도

'용언+명사파생접사', '체언+접미사', 용언 상당 어구는 '체언+동사/형용사/부사 파생접사'를 포함한다.

[10]에서는 접미사에 대해 '친구+들+과'에서의 '들'처럼 태깅성능에 영향을 미치지 못한다고 보아 제외했으나, '한국+산'이나 '신문+사'와 같은 예에서는 오히려 '한국'과 '신문'보다 더 중요한 문법기능을 갖는다.

표 2 핵심 실질/형식 형태소 예

형태소열	핵심 실질 형태소	핵심 형식 형태소
미술+학원	학원	null
미술+학원+,	학원	,
미술+학원+에서+,	학원	에서
미술+학원+에서+부터+는	학원	에서
미술+학원+부터+는	학원	는
미술+학원+이+다	학원이	다
운동+하+다	운동하	다
운동+하+시+었+다	운동하	다
운동+하+어+주+다	운동하어주	다
살+시+어+요	살	어
사+시+어+요	사	어

- 핵심 형식 형태소: 보조사를 제외한 어절 마지막 기능어. 단 어절에 형식 형태소가 없을 경우 null을, 보조사뿐일 경우 마지막 보조사를 취한다. 기호뿐일 경우 기로를 핵심 형식 형태소로 삼는다.

[10]에서는 어절에 여러 형식 형태소가 있을 경우 선행하는 형태소를 취했는데, 이는 head-final language인 한국어에 부적합하다. 예외적으로 보조사는 단순히 보조의미를 추가하기 때문에 후순위를 갖는다. 표 2은 핵심 실질/형식 형태소 예이다.

이렇게 단순화하면 데이터부족 문제는 다소 해소되지만, 태깅 시 히스토리가 달라지므로 어절 단위 태깅의 장점인 문맥확장성은 잃게 된다.

3.4. 품사 태거

문장은 형태소 분석기와 어절 단순화 모듈을 거쳐 lattice로 표현되고, 태깅은 이 lattice 상에서 수행된다. 태깅 프레임 워크는 독립성 가정 없이 자질을 선택할 수 있는 판별모델 CRF를 사용했으며, 학습에는 Le Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI)에서 개발한 공개 CRF 태깅도구인 Wapiti v1.5.0을 사용하고, 디코더는 새롭게 구현했다.

3.5. 학습 및 평가

학습 및 평가 말뭉치로 21세기 세종계획 최종 성과물(2011.12, 수정판 2쇄 기준)을 활용했다. 문제를 간단히 하기 위해 직접인용문(예: 그가 "그래"라고 말했다), 부연설명문(예: 학교(school)에 가다), 나열형(예: 서울-대전-대구-부산) 등이 없다고 가정해 682,212 문장을 추출하고, 9:1 비율로 학습 및 평가 데이터로 균일하게 분리했다.

제안된 방법이 효용이 있으려면 미등록어를 줄여주면서 단순화로 인한 의존 관계 정보 손실을 최소화해야한다. 따라서 평가는 미등록 어절 비율과 어절단위 precision($\frac{\text{맞힌 어절 수}}{\text{전체 어절 수}}$)를 측정했다. 특히 미등록 어절 비율은 학습 데이터를 10만 문장 단위로 나누어 학습데이터 크기에 따른 변화 추이를 측정했다.

학습 자질은 베이스라인(model_complex)와 제안된 모델(model_simple)에 대해 똑같이 $u(t_0|e_0)$, $u(t_0|e_{-1},e_0)$, $b(t_{-1},t_0)$을 부여했다(u:unigram, b:bigram, t:tag,

표 3 미등록 어절 비율 및 뼌화 추이

#.학습 문장	10,000	20,000	30,000	40,000	50,000	60,000
#.학습 어절	1,270,653	2,349,448	3,484,059	4,541,701	5,515,218	6,416,586
#.평가 문장	68220					
#.평가 어절	727550					
미등록 어절 비율 (%, baseline)	21.91	15.85	13.11	11.46	10.11	9.22
미등록 어절 비율 (%, simplified)	16.70	11.27	8.89	7.54	6.38	5.63
미등록 어절 감소 비율 (%)	23.76	28.85	32.16	3.42	36.87	38.95

e:eojeol, 첨자는 상대적 위치를 나타냄). model_simple_tune은 model_simple에서 $u(t_0|cm_{-1}e_0)$, $u(t_0|fm_{-1}e_0)$, $u(t_0|ft_{-1}e_0)$ 자질을 추가해 학습한 모델이다(cm:실질 형태소, fm:형식 형태소, ft:형식 형태소 태그).

4. 실험 결과 및 분석

4.1. 미등록어 비율

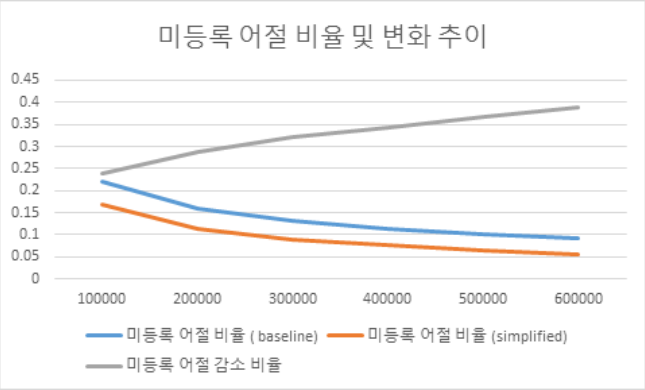

그림 2 미등록 어절 비율 및 변화 추이 그래프

단순화 어절의 경우 베이스라인에 비해 항상 미등록 어절 비율이 낮으며, 학습 말뭉치 크기가 커질수록 감소 비율이 커져 38.95%의 미등록 어절을 감소시키는 것을 확인할 수 있다. 표 3는 미등록 어절 비율 및 변화 추이에 대한 구체적 수치이다.

4.2. 어절단위 precision

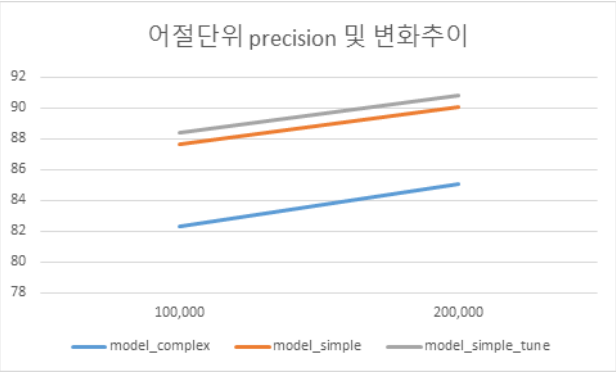

그림 3 어절단위 precision 및 변화추이 그래프

실험 환경 여건상 학습말뭉치 중 20만 문장밖에 활용하지 못하였다. 같은 자질을 사용한 제안된 모델(model_simple)이 베이스라인(model_complex)보다 약 5% 높은 성능을 보이며, 추가 자질을 활용할 경우 최고 ??%의 성능을 보이는 것을 확인했다. 표 4는 어절단위 precision 및 변화추이에 대한 구체적 수치이다.

표 4 어절단위 precision(%) 및 변화추이

#.학습 문장 / 모델명	100,000	200,000
model_complex	82.32	85.04
model_simple	87.67	90.06
model_simple_tune	88.42	90.81

베이스라인의 성능(82.32%)이 [8]에서 보고된 67.64%보다 높은 데, 그 이유는 본 연구에서 사용한 형태소 분석기는 정교한 접속정보를 사용하고, 따라서 오답 후보들을 크게 줄여줬기 때문으로 보인다.

다양한 조건(학습 말뭉치, 평가 말뭉치, 형태소 분석기 등) 차로 정확한 성능비교가 불가능하지만 표 5에서 어절단위 precision를 비교했다[1].

표 5 모델 별 어절 단위 precision

태깅 단위	모델 종류	어절 precision(%)
형태소 단위	김진동 외[14]	95.80
음절 단위	심광섭[6]	96.31
어절 단위	조민희 외[10]	98.03
구 단위	나승훈 외[9]	96.14

어절 단순화 처리를 했을 때 성능(90.81)은 표 5에서와 같이 다른 태거에서 보고된 성능보다 크게 낮은데 다음과 같은 원인이 있을 수 있다.

- 작은 학습 말뭉치 크기(20만 문장)
- 규칙기반 형태소 분석기와 말뭉치의 철학차이

규칙기반 형태소 분석기는 세종말뭉치에 독립적으로 만들어졌고, 형태소 분석기가 생성한 후보가 학습말뭉치에 존재하는 태그열과 달라 확률정보가 제대로 학습되지 않았기 때문일 수 있다. 예를 들어 '주세요'의 경우, 세종 말뭉치는 '주+시+어요'로 분석하지만 본 연구에서 사용된 형태소 사전은 '어요'를 가지지 않고 분리한다.

1) 각 논문에서 언급되는 정확도는 precision을 의미한다고 가정.

5. 결론

본 논문에서는 태깅 단계에서 선택할 수 있는 여러 처리 단위(음절, 형태소, 어절, 구)에 대해 각각 장단점을 고찰하고, 높은 정확도와 문맥 확장성이 높으나 데이터 부족 문제를 가진 어절 단위에 대해 핵심 실질 형태소와 핵심 형식 형태소로 단순화시킬 경우 미등록 어절의 등장 비율을 감소시키는 등 데이터부족 문제를 완화시켜 실재 품사 태깅 성능 향상에 도움을 줌을 확인할 수 있었다. 이 방법론은 한국어 외에도 일반적인 교착어의 태깅 성능 향상에 도움을 줄 것으로 기대된다.

하지만 어절을 단순화함에 의해 문맥확장 구현이 다소 어려워지게 되는 한계를 지닌다. 또한 기존 형태소 분석기 및 태거의 성능에 미치지 못했는데, 이는 적은 학습말뭉치 크기와 규칙기반 형태소 분석기에 기인한 문제라고 추측된다.

이 경우 [13]과 같은 방법으로 학습말뭉치와 유사한 후보군을 생성하는 형태소 분석기를 사용한다면 추가 성능향상이 있을 것으라 생각된다. 또한 형식 형태소의 이형태를 기본형으로 변환하는 등 추가 단순화를 꾀할 수 있다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신 · 방송 연구개발사업(R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발), ICT명품인재양성사업(R0346-16-1007) 및 (주)시스트란인터내셔널의 지원을 바탕으로 수행하였습니다.

참고문헌

[1] Georgiev, Georgi, et al. "Feature-rich part-of-speech tagging for morphologically complex languages: Application to bulgarian." Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012.

[2] Maier, Wolfgang, Daniel Dakota, and Daniel Whyatt. "Parsing German: How Much Morphology Do We Need?." SPMRL-SANCL 2014 (2014): 1.

[3] 이상주, 임희석, and 임해창. "은닉 마르코프 모델을 이용한 두단계 한국어 품사 태깅." 1994 년 제 6 회 한글 및 한국어정보처리 학술대회 (1994): 305-312.

[4] Han, Chung-Hye, and Martha Palmer. "A morphological tagger for Korean: Statistical tagging combined with corpus-based morphological rule application." Machine Translation 18.4 (2004): 275-297.

[5] Kudo, Taku, Kaoru Yamamoto, and Yuji Matsumoto. "Applying Conditional Random Fields to Japanese Morphological Analysis." EMNLP. Vol. 4. 2004.

[6] 심광섭. "형태소 분석기 사용을 배제한 음절 단위의 한국어 품사 태깅." 인지과학 22.3 (2011): 327-345.

[7] 이창기. "Structural SVM 을 이용한 한국어 띄어쓰기 및 품사 태깅 결합 모델." 정보과학회논문지: 소프트웨어 및 응용 40.12 (2013): 826-832.

[8] Lee, Do-Gil, and Hae-Chang Rim. "Probabilistic models for Korean morphological analysis." Companion to the proceedings of the international joint conference on natural language processing. 2005.

[9] 나승훈, and 김영길. "구기반 통계적 모델을 이용한 한국어 형태소 분할 및 품사 태깅." 한국정보과학회 2014 한국컴퓨터종합학술대회 논문집 (2014): 571-573.

[10] 조민희, et al. "최대 엔트로피 모델 기반 품사 태거의 성능 향상 기법." 제 16 회 한글 및 한국어 정보처리 학술대회 발표자료집 제 16 권 제 1 호 16.1 (2004): 73-81.

[11] 신준철, and 옥철영. "한국어 품사 및 동형이의어 태깅을 위한 단계별 전이모델." 정보과학회논문지: 소프트웨어 및 응용 39.11 (2012): 889-901.

[12] 권오욱, et al. "음절단위 CYK 알고리즘에 기반한 형태소 분석기 및 품사태거." 1999 년도 제 11 회 한글 및 한국어 정보처리 학술대회 및 제 1 회 형태소 분석기 및 품사태커 평가 워크숍 (1999): 76-87.

[13] 신준철, and 옥철영. "기분석 부분 어절 사전을 활용한 한국어 형태소 분석기." 정보과학회논문지: 소프트웨어 및 응용 39.5 (2012): 415-424.

[14] 김진동, 임희석, and 임해창. "Twoply HMM: 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델." 정보과학회논문지 (B) 24.12 (1997): 1502-1512.