형태소 기분석 사전과 DBMS를 이용한 형태소 분석 말뭉치 구축의 한 방법

조진현 강범모
고려대학교 언어과학과
belus@ikc.korea.ac.kr bmkang@korea.ac.kr

The Method for the Construction of POS Tagged Corpus based on Morpheme Ready Made Dictionary and RDBMS

Jin Hyun Cho Beom-mo Kang
Department of Linguistics, Korea University

요 약

본 논문은 1999년도에 구축된 '150만 세종 형태소 분석 말뭉치'를 바탕으로 형태소 기분석 사전을 구축하고, 이를 토대로 후처리의 수작업을 고려한 반자동 태거를 구축하는 방법론에 대해 연구한 것이다. 분석말뭉치 구축에 있어 기존 자동 태거에 의한 자동 태깅의 문제점을 분석하고, 이미 구축된 형태분석 말뭉치를 이용해 후처리 작업이 보다 용이한 1차 가공말뭉치를 구축하는 반자동 태거의 개발과 그 방법론을 제시하는데 목적을 두고 있다.

이와 같은 논의에 따라 분석 말뭉치의 구축을 위한 태거는 일반적인 언어 처리를 위한 태거와는 다르다는 점을 주장하였고, 태거에 전적으로 의존하는 태깅 방식보다는 수작업의 편의를 제공할 수 있는 태깅 방식이 필요함을 강조하였다. 본 연구에서 제안된 반자동 태거는 전체적인 태깅 성공률과 정확도가 기존의 태거에 비해 떨어지지만 정확한 단일 분석 결과를 텍스트의 장르에 따른 편차 없이 50% 이상으로 산출하고, 해결이 어려운 어절 유형에 대해서 완전히 작업자의 판단에 맡김으로써 오류의 가능성을 줄인다. 또한 분석 어절에 대해 여러 표지를 부착함으로써 체계적이고 단계적인 후처리 작업이 가능하도록 하였다.

1. 서 론

우수한 품질을 지닌 대용량의 분석 말뭉치를 구축하는데 어려움을 주는 요인이 몇 가지 존재한다. 이와 같은 문제점에는 태그 세트의 설정에 관한 문제, 분석 단위의 설정 문제, 구축을 위한 효과적인 응용 도구의 선택 문제, 분석 말뭉치에 대한 일관성의 문제 등이 있다. 특히 마지막에 언급한 일관성 획득의 문제는 구축을 담당하는 작업자들에게 큰 어려움으로 다가오는 요소이다.

본 논문에서는 기존에 사용되어 온 말뭉치 구축용 자동 태깅 시스템의 문제가 후처리 작업의 효율을 떨어뜨리고, 전체적인 일관성의 유지를 어렵게 하여 분석 말뭉치의 신뢰성을 떨어뜨리는 요인 중의 하나라는 점을 중시하고자 한다. 분석 말뭉치 구축용 태거는 정확한 분석이 중요하지만 후처리 단계의 수작업을 용이하게 하기 위한 사용자 환경이 중요하기 때문에 일반적인 자동 태거와 약간의 성격을 달리한다고 볼 수 있다. 정확한 분석이 가능한 어절에 대해서는 자동으로 태깅을 수행하되 규칙에 의해 해결할 수 없는 중의성 문제가 발생하면 무리한 단일분석의 산출보다는 후처리 단계로 넘기는 것이 오히려 수작업에서의 노력과 수고를 덜 수 있는 방법일 것이다.

이에 따라 분석 말뭉치 구축에 있어 기존 자동 태거에 의한 자동 태깅의 문제점을 분석하고, 이미 구축된 형태분석 말뭉치를 이용해, 후처리 작업이 보다 용이한 1차 가공 말뭉치를 구축하는 반자동 태거의 개발과 그 방법론을 제시하고자 한다.

2. 연구 방법

보다 안정적이면서 효율적인 후처리의 작업 환경을 제공할 수 있는 분석 말뭉치 구축용 반자동 태거의 개발에 대한 방법론을 제시하기 위해 문화관광부 주관으로 1999년에 제작된 150만 어절의 '21세기 세종계획 형태소 분석 말뭉치'를 이용하였다.

또한 반자동 태거 개발의 전 과정을 단계별로 제시하고 중의성 해결을 위한 문맥 규칙의 구성 방법, 단일분석 어절의 선택 기준과 미분석 어절에 대한 처리 방안을 모색하였다. 이를 위해 관계형 데이터베이스 관리 시스템 (Relational Database Management System)을 이용하여 형태소 기분석 사전의 구축을 관계형 데이터베이스로 구성하였다.

3. 자동 태거에 의한 말뭉치 구축의 문제점

세종 분석 말뭉치는 자동 태거를 이용하여 자동 품사 부착을 시행한 후 이를 수작업에 의해 수정하는 2단계 방식으로 구축되었다. 후처리 단계에서의 수작업은 자동 태거가 확정하지 못하는 중의성의 문제를 해결하고 태거의 오류를 수정하기 위해 필요한 작업이다. 2000년에 구축된 200만 어절의 세종 분석 말뭉치에서 후처리 작업을 수행하기 전에 발생된 오분석의 유형을 살펴보면 다음과 같다. 괄호 안의 분석은 세종 분석 말뭉치의 구축 지침에 의한 바람직한 분석결과로 이를 반영한 자동 태깅 시스템이 그것을 제시하지 못한 것이다. (단, JKB는 부사격 조사, JKG는 관형격 조사, JC는 접속조사, VV는 본동사, VX는 보조용언, NNP는 고유명사, NNG는 일반명사, XR은 비자립 어근, EC는 연결어미, ETM은 관형형 어미, NF는 명사추정범주이다)

Ⅰ. 다중 분석에서의 오류

(1) 중의성 해결에 실패한 예. [()속이 올바른 분석임]
<몰이해와 편견> 몰이해/NNG+와/JKB (와/JC) 편견/NNG
(2) 잘못된 추정에 의한 분석
<학교측이> 학교측이/NF (학교/NNG+측/NNB+이/JKS/이/JKC)

Ⅱ. 단일 분석에서의 오류

(1) 말뭉치 전체적으로 잘못 태깅된 예
<미세한> 미/NNP+세한/NNG (미세/XR+하/XSA+ㄴ/ETM)
(2) 말뭉치에서 부분적으로 잘못 태깅된 예
<작품 속의> 작품/NNG+속의/NNG (작품/NNG+속/NNG+의/JKG)
(3) 잘못된 추정에 의한 분석
<2백억씩> 2/SN+백억/NF+씩/XSN (2/SN+백억/NR+씩/XSN)

Ⅲ. 잘못된 원문에 의한 분석 오류
<못이치고> 못이치/NF+이/VCP+고/EC

Ⅰ,Ⅱ와 다르게 Ⅲ의 유형은 분석 대상 자료가 입력시 오타, 띄어쓰기 등에 의해 이미 오류가 발생될 수 밖에 없는 환경을 제공하고 있기 때문에 전처리 작업에서 모두 교정되지 못한 어절에 대해서는 후처리 작업의 필요성이 인정된다.

Ⅰ-(2),Ⅱ-(3)의 유형은 자동 태깅을 위한 형태소 분석 단계에서 사전에 해당 어휘가 존재하지 않거나 문맥 규칙과 확률 정보에서 획득될 수 없는 유형들이 대부분이므로 수작업에 의한 분석이 필요하다. 주로 인명, 기관명 등의 고유명사와 외국어를 한글로 표기할 때의 문제, 방언, 구어, 태거 자체의 사전 부족 등이 원인이 되어 발생된다.

그리고 Ⅰ-(1)의 유형도 현재 중의성에 대한 문제를 자동 태거가 완벽하게 해결할 수 없는 현실을 감안할 때 인정될 수 있는 오류로 보인다. 예를 들어 '-와/과'는 세종 분석 말뭉치에서 '접속조사(JC)'나 '부사격 조사(JKB)' 2가지로 분석될 수 있는 중의성을 지닌 형태소이다. 이 형태소들은 문맥 안에서 규칙화하기가 쉽지 않기 때문에 대부분 수작업에 의한 처리 대상이 된다. 그런데 기존의 자동 태거들은 주로 단일 분석의 생성을 원칙으로 하기 때문에 확률 정보를 이용하여 2가지 분석 중 확률이 높은 것을 추정, 선택하게 되는데, 이 때 잘못된 선택을 일으킬 수 있는 가능성이 존재한다. 만일 이러한 처리 결과에 대한 어떠한 표시도 주지 않는다면 수작업에 의한 처리 과정에서 오분석 어절을 놓치게 될 수 있다. 왜냐하면 태거의 사용자는 그러한 통계적 기준이나 처리 방식에 대해 알 수 있는 정보가 부족하기 때문에 처리 유형을 경험적으로 추정할 수밖에 없기 때문이다.

이와 같은 문제가 Ⅱ의 유형에서도 발생한다. 중의성이 없는데도 단일 분석 자체가 잘못 분석된 경우는 규칙의 오류, 통계 정보의 오류, 학습 데이터의 오류 등 태거 자체의 오류로 볼 수 있기 때문에 후처리에서의 수작업을 매우 어렵게 만든다. 특히 Ⅱ-(2)와 같이 부분적으로 잘못 태깅된 경우에는 잘못 분석된 유형이 다양하게 발생해 오류를 추정하기가 쉽지 않고, 잘못 태깅된 어절을 찾기 위해 맞게 분석된 어절도 일일이 검토해야 한다.

중의성 해결이 잘못 되어 오분석을 일으킨 어절을 분석 결과로부터 모두 추정하여 수정하기란 쉬운 작업이 아니다. 발견된 '와/과'의 오류를 수정하기 위해 이것이 결합된 유형을 모두 다시 살펴야 하며, 만약 이러한 오분석이 발견되지 못한 경우에는 그대로 분석 오류의 형태로 남게 될 것이다.

지금까지 살펴본 문제들은 말뭉치 구축시 사용되는 자동 태거가 통계적 방법에 의해 한 가지 분석만을 제시하는 태거와는 다른 성격임을 보여준다. 중의성이 없이 확실한 - 본 논문에서 사용하는 '확실한', '결정적' 등의 의미는 '대량의 말뭉치에서 용례를 추출한 결과로부터 중의적으로 분석될 가능성이 없다'는 것을 의미한다 - 분석이 가능한 어절과 결정적 규칙을 사용하여 중의성 해결이 가능한 유형에 대해서는 수작업이 필요하지 않을 수준의 확실한 분석이 이루어져야 한다. 반면에 중의성의 해결이 어려운 어절이나 분석이 어려운 어절에 대해 무리한 단일 분석보다 의미있는 분석 가능 후보를 제시함으로써 후처리 작업에서 선택할 수 있도록 하는 것이 바람직하다. 이를 위해 분석된 어절과 미분석 어절에 대한 구분을 확실히 줄 수 있는 적절한 표시 형태가 필요하다. 따라서 지금부터는 형태소 기분석 사전과 형태소 분석, 문맥 규칙 정보를 이용하여 오류가 비교적 적은 단일분석을 제시하고, 자동으로 처리하기 어려운 다중 분석 어절에 대해서는 적절한 정보를 제공함으로써 후처리 작업에서 편의를 제공할 수 있는 작은 규모의 반자동 태거를 시험적으로 개발하고 그에 대한 방법론과 문제점에 대해 알아보고자 한다.

방법론에 대한 고찰에 앞서 본 논문에서 제안하는 반자동 태거에 대한 명칭을 CETtagger라고 부르기로 한다[1].

4. 반자동 태거 개발을 위한 방법론

4.1 반자동 태거를 위한 DBMS의 구축 방법

CETtagger 반자동 태거는 처리를 위해 MySQL (ver3.22)이라는 DBMS를 이용하였고, C언어로 작성되었다. MySQL을 이용하여 각 작업 단계에서 필요한 테이블들을 구축하였다.

각 단계에서 사용하는 테이블들을 관계형(relation)으로 구성하였는데, 그 이유로는 데이터베이스의 구조를 단순화시킬 수 있고 저장공간의 낭비를 줄일 수 있기 때문이다. 또한 데이터에 대한 개념적이고 논리적인 접근을 쉽게 할 수 있으며, 저장된 데이터의 검색을 효과적으로 만든다.

이와 같은 이유로 형태소 기분석 사전은 CETtagger라는 이름의 데이터베이스 안에 테이블(table) 형태로 구축되었다. **[표-1]**은 각 테이블의 구성과 용도를 표로 나타낸 것이다.

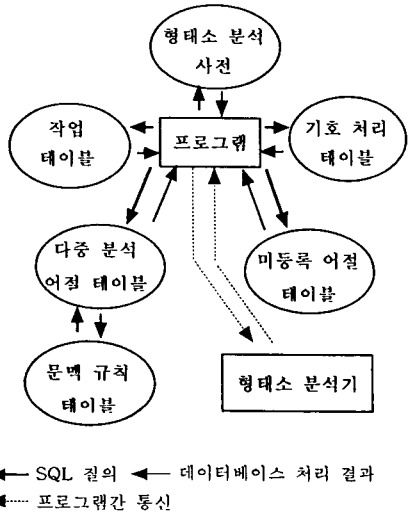
CETtagger	
종류	TABLE 용도
형태소 기분석 사전	기본 테이블
	분석 유형 사전
	유형 빈도 사전
	형태소 단위 문맥 규칙 사전
	형태소 단위 비교값 저장
	어절 단위 문맥 규칙 사전
	기호 사전

1) 'CET'라는 명칭은 세종 분석 말뭉치를 구축하고 있는 고려대학교 민족문화연구원 산하의 전자텍스트연구소의 영문명인 'Center for Electronic Text'에서 따온 것이다. 이와 아울러 세종 분석 말뭉치 구축을 위해 사용된 태거는 '세종 태거'라 부르기로 한다.

태깅 작업 테이블	어절 태깅 결과 저장
	어절 문맥 규칙 처리
	형태소 문맥 규칙 처리
	미등록 어절 처리
	기호 처리

〔표-1〕 : CETtagger 시스템의 데이터베이스 구성

[그림-1]은 프로그램과 DB 사이의 정보 교환을 나타낸 것이다.



〔그림-1〕 프로그램과 테이블 사이의 정보 교환

4.2 형태소 기분석 사전의 구성

150만 어절의 세종 분석 말뭉치를 이용하여 CETtagger용 형태소 기분석 사전을 구축하였다. 형태소 기분석 사전은 자동 태거의 성능에 큰 영향을 미치는 기본적인 데이터이기 때문에 몇 가지 고려해야 할 요소가 있다. 첫째, 기분석 사전을 구성하는 분석 말뭉치가 질적으로 우수해야 한다. CETtagger는 150만 분석 말뭉치에서 확실하게 1개 유형으로 태깅된 어절에 대해서는 분석 결과를 그대로 가져와 새로운 어절을 태깅하는 방법을 사용한다. 형태 분석 말뭉치의 분석 결과에 오류가 많으면 이를 기반으로 태깅 작업을 수행하는 태거의 성능을 향상시킬 수 없다.

둘째, 분석 말뭉치는 일관성이 확보되어야 한다. 일관성이 없는 분석은 같은 어절에 대해 다른 분석 결과를 생성함으로써 형태소 기분석 사전에 중의성이 있는 어절 유형으로 등재되기 때문이다.

셋째, 분석 말뭉치의 규모에 대한 고려가 필요하다. 분석 말뭉치의 규모에 따라 형태소 기분석 사전의 크기가 결정된다. 형태소 기분석 사전의 크기가 커질수록 더 많은 어절 유형을 포함한다는 것을 의미하기 때문에 자동 태깅의 분석 정확도를 높일 수 있을 것으로 기대된다.

넷째, 형태소 기분석 사전은 어절 유형에 대한 빈도 정보를 지니고 있는 것이 좋다. 빈도 정보를 포함하고 있는 형태소 기분석 사전은 본 논문에서 제안한 자동 태거뿐 아니라 다양한 방면에서 활용될 수 있을 것이다.

이러한 고려를 바탕으로 CETtagger용 형태소 기분석 사전이 테이블 형식으로 구축되었다. 분석 말뭉치의 질을 향상시키기 위해 모든 어절에 대한 확인 작업을 수작업으로 수행하였고, 분석 말뭉치 내에서 분석 유형의 빈도를 계산하여 데이터베이스에 입력하였다.

4.3 단일 분석 어절의 선택

4.3.1 단일 분석 어절의 선택 방법

본 논문에서 제안하는 반자동 태거는 확실하게 하나로 분석되는 어절 유형은 형태소 분석 없이 그대로 태깅을 수행하고, 다중 분석 어절에 대해서는 적절한 처리를 통해 단일 분석을 선택하며, 단일 분석의 선택이 어려운 어절에 대해서는 적절한 범주를 설정하여 출력함으로써 후처리의 수작업을 용이하게 하는데 그 목적을 두고 있다. 따라서 단일 분석에 대한 기준을 설정하는 것이 무엇보다 중요하다.

150만 형태소 분석 말뭉치에서 하나로 분석되었다고 해서 그것을 그대로 채택하는 것은 무리가 있다. 예를 들어 '의사가'라는 어절이 사전 '명사 + 주격조사'의 형태로만 등재되어 있다고 해서 입력되는 '의사가'를 그대로 태깅할 수 없다. 왜냐하면 '명사+보격조사'에 대한 태깅 가능성을 놓치게 되어 오류를 발생시킬 수 있기 때문이다. 기분석 사전으로 구축된 350,639개의 어절 유형 중 분석 유형이 1개인 어절은 341,531개로 약 97.4%를 차지한다. 또한 총 1,504,008개의 어절 중 한가지 유형으로만 태깅된 어절은 1,147,803개이다.

150만 말뭉치에 나타난 어절 유형을 분석 개수에 의해 구분한 후 각각의 타입(type)과 토큰(token)을 살펴보면 다음과 같다.

어절 분석개수	타입(Type)	토큰(Token)
1	341,531 (97.4%)	1,114,783 (76.3%)
2	8,517 (24.3%)	268,630 (17.3%)
3	506 (0.14%)	52,126 (3.5%)
4	59 (0.017%)	11,184 (0.79%)
5	19 (0.005%)	8,307 (0.06%)
6	6 (0.002%)	15,848 (4.5.3%)
7	1 (0%)	137 (0.009%)

〔표-2〕 : 150만 말뭉치의 어절 분석 개수

이러한 수치는 단일 분석으로 확정할 제어 장치가 없다면 오류 발생 가능성이 높게 있음을 알려주는 것으로 어떤 방식으로든 단일 분석에 대한 선택 기준 설정이 필요함을 나타낸다. 즉, 중의성이 존재할 수 있지만 150만 형태소 분석 말뭉치에서는 단일 분석 유형으로 태깅된 어절을 그대로 태거에 적용하지 않고 일정한 기준에 의해 선별하여 태깅하는 것이 필요하다.

약 5만 어절에 대한 샘플 테스트 결과 단일 분석 결정에 대한 제어 없이 기분석 사전에 한가지로만 분석되어 있는 유형을 그대로 적용한 결과 약 60%가 단일 분석으로 태깅되었다. 이를 토대로 단일 분석에 대한 정확도를 조사하였다. 그 결과 빈도에 상관없이 중의성을 발생시킬 수 있는 요소는 주로 중의성으로 분석될 가능성이 있는 일부 조사, 어미 등의 문법 형태소와 본동

사, 보조동사의 중의성이 대부분이었고, 일부 품사 중의성이 발견되었다.

주격, 보격조사나 접속, 부사격 조사 등의 문법 형태소나 본동사, 보조동사의 경우 태깅의 과정에서 어느 정도 예측이 가능하기 때문에 적절하게 처리를 해 줄 수 있지만, 품사 중의성의 경우에는 그 수가 많고 다양하기 때문에 기분석 사전에서 단일 분석으로 제시한다 하더라도 그것을 확실하게 인정하고 태깅을 수행하기가 상당히 어렵다. 따라서 본 논문에서는 이러한 문제를 해결하기 위해 기분석 사전의 단일분석 항목을 형태소 분석기로 분석한 뒤 이를 데이터베이스로 구축하여 형태소 분석의 결과와 기분석 사전에서 모두 단일분석으로 제시한 항목만을 확실한 단일분석으로 인정하는 방식을 택하기로 한다. 단, 형태소 기분석 사전의 검토 결과 이러한 중의성의 가능성을 내포하고 있는 어절이 대부분 빈도가 30개 미만인 경우에서 발생하고 있기 때문에 빈도가 30개 이하이고 어절의 분석이 1개로 제시된 어절은 단일분석 선택 대상에서 제외하여 '추정([ASSUMP])'의 표지를 부착하도록 했다. 또한 30개 이상 중에서도 위에서 언급한 문법 형태소가 존재하는 어절에 대해서는 '추정'의 표지를 부착하여 출력한다. 만일 말뭉치에서의 빈도가 30개 이상이고 중의성을 유발시킬 수 있는 문법 형태소가 결합되지 않은 어절 유형에 대해서는 형태소 분석기에 의한 태깅 결과와 상관없이 확정된 단일분석으로 채택하기로 한다.

4.3.2 중의성을 지닌 다중 분석 어절의 처리 방법

형태소 기분석 사전에 총 어절 유형 빈도(total_type_freq)가 2 이상인 어절과 단일 후보 선택에서 제외된 어절, 미등록 어절의 처리에서 다중 분석으로 판명된 어절은 중의성을 지닌 어절로 간주되어 규칙을 적용받는 후보가 된다. 예를 들어 태깅 대상 어절이 '의사가'일 때 이를 형태소 기분석 사전에서 검색하면 [의사/NNG + 가/JKS], [의사/NNG + 가/JKC]의 두 가지 분석 결과를 제시하게 된다. 이 때 자동 태거가 하나의 분석을 선택할 수 있도록 기준을 제공해 주는 장치가 문맥 규칙(context rule)이다. 문맥 규칙은 해당 어절을 중심으로 앞뒤 한 어절, 또는 그 이상(최대 세 어절) 확장하여 중심 어절이 주위 어절들과 맺고 있는 관계(relation)를 문맥정보로 활용하는 방법이 주로 사용된다. CETtagger에서는 해당 어절을 중심으로 최대 앞뒤 세 어절을 볼 수 있도록 구축하였다.

CETtagger 시스템에서 사용하는 문맥 규칙은 앞뒤 어절의 형태소 단위 정보를 이용하여 해당 어절의 형태소 유형을 확정하는 것이다. 비교 대상은 분석된 형태소나 분석 태그, 태그가 부착된 형태소 모두 가능하다. 즉, 태깅 대상 어절을 중심어로 하여 앞이나 뒤의 형태소 분석 유형을 비교하여 분석 후보로부터 형태소 단위로 올바른 태그를 부여할 수 있도록 한다. 예를 들어 '-이/가' 격조사가 나온 다음 어절의 첫 형태소가 '되-(본동사)', 아니-(부정 지정사)'로 시작되면 '보격조사 태그(JKC)'를 부착하는 규칙을 설정하고 이에 따라 주격조사와 보격조사의 중의성을 해결하는 방식의 문맥 규칙이다.

지금까지 제시한 문맥 규칙은 결정적 규칙이다. 이상주 외(1998)에서는 결정적 규칙 부여의 단점으로 규칙의 수가 굉장히 많아지고 규칙을 얻기 위한 수작업의 비중이 커져 오류의 확률이 높을 수 있음을 지적하였다. 또한 모든 용례를 검토한 후 중의성을 발생시킬 수 없는 경우에만 규칙화할 수 있기 때문에 다양한 유형의 중의성을 해결하는데 적합하지 못함을 단점으로 지적하면서 이에 대한 대안으로 자동 습득이 가능한 통계기반 어휘규칙을 제안하였다. 하지만 본 논문에서 지향하는 자동 태거는 분석 말뭉치 구축을 위한 후처리 작업을 가정하고 있다. 이러한 가정에 부합되는 태깅 방식은 확률적, 통계적 분석 결과의 선택이 아니라 확정적인 분석 결과는 선택하여 제시하고 그럴 수 없는 어절에 대해서는 수작업의 처리로 오류를 최소화하는 것이다. 따라서 CETtagger 시스템에서는 결정적 규칙을 통한 중의성 해결 방식을 채택할 필요성이 있다.

4.3.2.1 문맥 규칙의 구성과 적용 방식

앞뒤 어절의 형태소를 이용한 문맥 규칙을 적용하기 위해서는 분석된 형태소를 이용하거나 형태소에 부착된 태그 정보를 함께 이용할 수 있다. 앞에서 언급했던 주·보격조사의 처리는 뒤 어절의 첫 형태소와 비교하는 방법이다. 예를 들어 75만 어절의 분석 말뭉치를 형태소 기분석 사전으로 구축하고 문맥 규칙을 전혀 주지 않은 상태에서 총 233,058개의 어절에 대해 자동 태깅을 시도한 결과, 44,451개의 중의성이 있는 어절 중 6,772개의 어절(약 15%)이 주격·보격조사의 중의성을 지닌 것으로 나타났다. 주격조사와 보격조사는 문맥 규칙을 적용할 수 있는 환경이 비교적 분명하기 때문에 이에 대한 적절한 처리로 태거의 성능을 향상시킬 수 있을 것으로 예측된다.

형태소와 태그를 함께 사용하여 중의성을 해결하는 예로 '보조용언'을 들 수 있다. 대부분의 보조용언은 앞 어절에 일정한 형태의 어미와 연결되는데, '21세기 세종 계획 기초 언어자료기반 분과' 세부연구과제의 최종 연구 보고서로 제출된 '한국어 정보처리를 위한 어절 분석 표지의 표준화 연구'(임홍빈, 송철의 1998, 이하 표준화 연구)에서는 전산적인 처리의 목적으로 보조용언 앞에 오는 어미에 대한 목록화를 시도하였다. 본 논문에서는 '표준화 연구'에서 제시한 보조용언에 연결되는 어미 목록을 참고하여 세종 분석 말뭉치의 보조용언에 대해 검토하여 반례가 나오는 어미 유형에 대해서는 규칙 설정에서 제외하였다. 예를 들어 '표준화 연구'에서는 동사 '만들-'는 어미 '-게' 다음에서만 보조동사로 사용됨을 제시하고 있다.

기어코 친구를 가지 못하**게 만들었다.**
보조용언

하지만 분석 말뭉치와 세종 말뭉치의 용례를 검토한 결과 다음에서 볼 수 있듯이 '-게' 다음에 본동사로 사용되는 반례가 발견되었다.

1. 이**렇게 [만들어진]** 것이 아스피린이다.
2. 모방한 형으로 매우 정교하**게 [만들어졌다.]**
3. 여러 가지 틀에 따라 다양하**게 [만드는]** 것을 성형이라
4. 다리는 잘 만들어져야 하며, 멋있**게 [만들어져야]** 하고,

이러한 반례는 주로 '형용사+게'의 형태에서 나타나는데, 이를 위해 '형용사+게'를 기준으로 규칙화할 수 있는지 살펴보았다. 하지만 '경제를 어렵게 만든 것은…'과 같이 '형용사+게' 다음에 보조용언으로 사용된 예들이 존재하기 때문에 어미 '-게' 다음의 '만들-'는 본용언과 보조용언을 구분하는 규칙에서 제외되어 수작업에 의한 후처리 단계에서 선택하도록 하였다. 이와 반대로

'만들-' 앞의 유형이 '-게'가 아니면 모두 본동사로 처리될 수 있도록 규칙화하였다. CETtagger에서 규칙으로 처리하기 위해 선택된 중의성이 있는 용언에는 '가지-', '나가-', '놓-', '대-', '되-', '두-', '드리-', '들-', '말-', '만들-', '먹-', '버리-', '보이-', '빠지-', '생기-', '쌓-', '아니하-', '않-', '오-', '있-', '주-', '지-', '하-' 등이 있다. (형태소 단위 문맥 규칙의 설정을 위해 150만 세종 분석 말뭉치, 2000년도에 구축한 200만 분석 말뭉치, 1천만 세종 원시 말뭉치, 국립국어연구원의 「표준국어대사전」,금성출판사의 「국어대사전」을 참조하였다. 이하 사전에 대한 명칭은 「표준」과 「금성판」으로 통일하도록 한다.)

특히 논란의 여지가 있었던 중의적 용언은 '되-'이다. 세종 분석 말뭉치에서는 어미 '-게', '-아야/어야' 다음에서 이를 보조용언으로 분석했는데, 「표준」에서는 모든 유형을 본용언으로, 「금성판」에서는 '-게' 다음에만 보조용언으로 다르게 처리하고 있다. 또한 「금성판」에서는 '-게' 다음에서 본용언으로 사용되는 경우가 있음을 밝히고 있다. 다만 「금성판」에서 제시한 예에서는 본동사와 보조용언의 의미적 차이를 발견하기 어려웠고 이는 사전 기술상의 문제가 아닌가 판단된다. 이처럼 사전에서도 다르게 구분하고 있는 동사 '되-'에 대해 분석 말뭉치 지침의 보조용언 목록에서 삭제할 것인가에 대한 논의가 이루어졌는데, 다른 용언에 연결되어 그 의미를 도와 문장을 완결하는 역할을 수행한다는 점과 '되-'가 주로 '하-'와 비슷한 방식으로 문장에서 발현되고 의미적 차이만을 발생시킨다는 점을 중시하여 보조용언 목록에 포함하기로 했다. 이에 따라 말뭉치에서 용례를 추출한 결과 '이렇게', '그렇게', '저렇게' 다음에서 사용되는 경우를 제외하고는 보조용언으로 사용되고, '-아야/어야'는 모든 환경에서 보조용언으로 사용되었다. 따라서 세종 분석 말뭉치의 방식대로 '-게'와 '-아야/어야' 다음에 나오는 '-되'만 보조용언으로 인정하기로 했다. 다만 '이렇게', '그렇게', '저렇게' 다음에 나오는 '되-'는 어절 단위 문맥 규칙에서 처리함으로써 '되-'에 의해 발생되는 동사와 보조용언에 대한 중의성을 해결하고자 하였다. 형용사로 사용되는 '되-'의 경우에는 문제가 있는데, 이에 대한 논의는 다음 절에서 하도록 하겠다.

태그가 부여된 형태소를 이용하는 문맥 규칙 적용 방식은 몇 가지 한계를 가지고 있다. 첫째, 선행 어절의 형태소 분석 결과를 참조하는 어절의 경우 분석 결과가 단일 분석이 아니라면(중의성이 있는 어절이라면) 문맥 규칙의 적용이 어려워진다.

둘째 일단 규칙으로 설정된 형태소는 어떠한 어절 유형을 형성하건 모든 곳에서 적용되기 때문에 완벽하지 못한 규칙 설정은 심각한 오분석을 일으킬 가능성을 지닌다.

셋째 두 어절이 상호 참조하는 경우를 가정할 수 있다. 이러한 문제가 발생하게 되면 프로그램은 반복적인 루프(loop)에 빠져들 가능성을 배제할 수 없다.

이와 같은 문제점을 고려하여 형태소 단위 문맥 규칙을 설정하였다. 규칙은 확정적인 분석 결과를 얻을 수 있도록 최소화하고자 하였고, 주로 주격조사과 보격조사, 본용언과 보조용언을 구분하는데 국한하였다.

4.3.2.2 문맥 규칙의 설정에 있어서의 문제점

규칙 설정 과정에서 가장 어려운 문제는 적용 기준을 어느 수준에서 결정해야 하는가 하는 점이다. 어떤 어절이 전체 용례에서 99%는 A로, 1%가 B로 분석된다면 전체에서 1%만 나타나는 분석 B를 위해 분석 A로의 태깅 가능성을 포기할 것인가, 아니면 규칙으로 설정하여 1%의 오류 가능성을 무시할 것인가 선택을 해야한다. 또한 개념적으로는 가능하지만 실제 용례에서 나타나지 않는 어절 형태에 대해 어떤 결정을 내려야만 하는가의 문제도 제기된다. 만일 해당 어절이나 형태소가 텍스트 상에서 사용되는 빈도가 매우 많다면 이 문제는 태거의 성능과 직접적으로 연관이 되기 때문에 무시할 수 없는 요소이다. 이에 대한 가장 큰 예로 지적할 수 있는 것이 바로 주격조사와 보격조사의 문제이다.

앞에서 언급하였듯이 주격조사와 보격조사, 특히 주격조사는 일반적으로 텍스트에서 가장 많이 사용되는 조사중의 하나이다. 이를 구분해 줄 수 있는 기준이 바로 동사 '되-'이다. 즉 다음 어절에 나오는 형태소가 본동사 '되-'면 보격조사로, 그렇지 않으면 주격조사로 분석할 수 있다. 그러나 '되-'는 동사, 형용사, 보조동사로 분석될 수 있는 중의성을 갖는 형태소이다. 만일 빈도가 많은 주격조사의 선택을 위해 [다음 어절이 '되/V-'로 시작하면 보격조사, 그렇지 않으면 주격조사로 태깅한다]라는 규칙을 설정한다면 다음 어절이 형용사 '되-'인 주격조사 포함 어절은 모두 보격조사로 분석될 것이다. (세종 분석 말뭉치 구축 지침에서 용언에 속하는 품사 범주로는 'VV-동사', 'VA-형용사', 'VX-보조용언' 등이 있다.)

밥이 너무 되서 먹을 수 없었다.
주격조사 형용사
→ 밥이 너무 되서 먹을 수 없었다.
보격조사 형용사

그렇다고 해서 주격·보격조사에 대한 분석을 포기하기에는 후처리의 수작업에 큰 부담으로 작용할 수 있다. 왜냐하면 '되-'가 형용사로 사용되는 환경이 극히 제한적이고 그 빈도가 많지 않기 때문이다.

이와 같은 문제의 처리를 위해 CETtagger에서는 빈도가 현저하게 차이가 나는 어절이나 형태소에 대해서는 분석 오류의 가능성을 안고 그대로 규칙을 적용하기로 한다. 이러한 처리 방식은 확정적인 분석 처리를 위한 단일 분석의 선택 기준을 위배하는 측면이 존재하지만 태거 성능의 개선을 위한 잠정적인 처리 방침이다. 그리고 향후 지속적인 태거의 보완을 통해 해결해 나가야 하는 문제이다. 다만 CETtagger의 가장 큰 목적인 후처리 수작업의 편의를 위해 '되-'가 형용사로 사용된 용례를 추출하여 논항으로 사용된 어절의 유형을 목록화하고 태깅시 이를 마크업(markup)의 형태로 표시해 줌으로써 어느 정도 이 문제를 해결하고자 하였다. 이러한 예로는 '밥이', '죽이', '일이', '반죽이' 등이 있다. 일단 모두 본동사로 분석한 다음에 주어로 사용된 앞 어절이 마크업된 어절이라면 수작업으로 이를 판단하여 수정해 줄 수 있을 것이다.

개념적, 이론적으로는 가능하지만 실제 텍스트에서 나타나지 않는 유형은 더 큰 문제를 안겨준다. 적은 빈도라도 다양한 텍스트에서 출현한 경우에는 그 빈도와 유형을 검토하여 처리 방안을 논할 수 있지만, 이 경우는 규칙 설정에 있어 어려움을 준다. CETtagger에서는 이러한 어절 유형이나 형태소의 문제는 위에서 처리한 방식대로 경험적인 가능성을 우선하기로 한다. 즉 규칙을 만드는데 참조하는 350만 분석 말뭉치나 세종 원시 말뭉치에서 출현하지 않는 유형에 대해서는 규칙 설정에 있어 출현 가

능성을 배제하여 처리하도록 하였다.

4.4 미등록 어절의 처리 방법

미등록 어절은 형태소 기분석 사전에 분석 대상 어절에 대한 분석 정보가 들어있지 않아 어떤 형태로든 분석을 수행하지 못하는 어절을 말한다. 미등록 어절의 생성은 형태소 기분석 사전의 크기, 기호에 의한 어절의 형태 변화, 분석 말뭉치의 장르별 균형 여부에 따라 영향을 받게 된다.

본 논문에서는 형태소 분석 결과의 후보를 얻기 위해 기존의 형태소 분석기를 이용하는 방법을 사용한다. CETtagger는 이미 생성된 형태소 분석 결과를 가지고 처리하는 방식이기 때문에 전처리 단계에서 별도의 형태소 분석을 하지 않는다. 따라서 미등록 어절에 대한 형태소 분석 후보를 얻는 작업은 CETtagger 시스템에서 처리할 수 없는 작업이므로 형태소 분석기를 사용할 것이다. 이 때 생성된 형태소 분석 후보들 중 의미있는 분석 후보를 선택하기 위해 앞에서 언급했듯이 어절별 태그 유형으로 판단하여 분석 후보를 줄이도록 했다. 이렇게 선택된 분석 후보를 다시 태거에 넣고 문맥 규칙을 적용하여 최대한 단일 분석을 선택하고, 그렇지 못한 어절에 대해서는 분석 후보를 적절한 출력형태를 지정하여 파일로 출력한다.

이와 같은 미등록 어절의 처리를 위해서는 형태소 분석기가 태거에 포함되어 있어야 한다. 하지만 기존의 분석기를 그대로 사용하기 때문에 이를 태거에 내장하기란 쉬운 작업이 아니다. 일단 소스의 공개 문제부터 다른 운영체제와 프로그램 언어로 되어 있는 형태소 분석기를 CETtagger의 환경에 맞게 변환시키기는 문제 등이 발생한다. 따라서 CETtagger에서는 세종 분석 말뭉치 구축에 사용된 형태소 분석기를 사용하였으며, 프로그램을 둘로 나누어 1차 태깅 단계에서 미등록 어절만을 파일로 출력하고, 미등록 어절 목록을 형태소 분석기로 분석하여 이를 다시 입력하여 2차 태깅을 실시하였다.

형태소 분석기로 처리한 미등록 어절을 다시 태거에 입력하기 위해서는 분석 결과의 수를 제한할 필요가 있다. 예를 들어 미등록 어절에 대한 형태소 분석 결과 분석 유형이 63개인 어절이 발견되었다. 이 중에서 맞는 한 개의 분석을 선택하기 위해 모든 유형을 검사하기란 쉬운 작업이 아니다. 따라서 CETtagger에서는 앞에서 언급한 대로 150만 분석 말뭉치로부터 태그 결합 유형을 추출하여 형태소 분석 결과 중 태그 결합 유형에 있는 분석만을 채택하였다.

5. 실험 및 평가

5.1 실험 결과

실험은 2001년 형태 분석 말뭉치 구축 대상인 텍스트 중 장르가 다른 네 가지 텍스트, 총 88,879 어절을 선정하여 실시하였다.

텍스트명	텍스트 출처	장르	총어절수
Text1	문화의 시대(김성우)	책/예술론	37,890
Text2	학생백과사전(계몽사)	책/총류일반	23,005
Text3	작은 그림책(송재찬)	아동/상상적 산문	14,013
Text4	신문 칼럼(조선일보)	신문/문화	13,971

〔표-3〕 실험 대상 텍스트

이들 텍스트를 대상으로 형태소 분석을 통한 미등록 어절의 처리를 완료한 후 태깅을 수행한 결과는 **〔표-4〕, 〔표-5〕**과 같다.

텍스트	총어절수	[1TYPE]	[CR]	[2TYPE]	[ASSUMP]	[NOT]
Text1	37,890	16,218	2,779	10,788	7,852	253
Text2	23,005	10,876	1,303	5,592	5,152	82
Text3	14,013	6,321	1,087	4,147	2,393	65
Text4	13,971	6,832	721	3,435	4,847	136
계	88,879	40,247	5,890	23,962	18,244	536

〔표-4〕 CETtagger를 이용한 태깅 결과
(출력 표지별 빈도수)

텍스트	[1TYPE]	[CR]	[1TYPE]+[CR]	[2TYPE]	[ASSUMP]	[NOT]
Text1	42.8	7.3	50.1	28.5	20.7	0.7
Text2	47.3	5.7	52.9	24.3	22.4	0.4
Text3	45.1	7.8	52.9	29.6	17.1	0.5
Text4	48.9	5.2	54.1	24.6	20.4	1.0
계	45.3	6.6	51.9	27.0	20.5	0.6

〔표-5〕 CETtagger를 이용한 태깅 결과
(출력 표지별 비율-단위:%)

1.[1TYPE] : 형태소 기분석 사전에 의해 단일 분석으로 확정된 어절
2.[2TYPE] : 다중 분석 어절
4.[CR] : 다중 분석 어절 중 규칙이 적용된 어절
5.[ASSUMP]: 단일 분석 추정 어절
6.[NOT] : 분석 실패 어절

실험 결과는 예상과 다르게 네 가지 장르의 텍스트에서 비슷한 수준의 결과를 얻을 수 있었다. 네 가지 장르에 대한 실험에서는 장르별 편차가 크지 않고 단일 분석으로 확정된 어절도 52%정도로 낮은 확률을 보였다. 텍스트별 단일 분석 생성의 차이가 5% 미만으로 고른 것은 각 텍스트의 특징을 유발시키는 전문어, 고유 명사 등의 특징적 어휘나 어절 형태가 텍스트 전체에서 나타나는 것이 아니라 어느 정도(50%)는 기본적으로 유사한 어절 형태가 사용되었기 때문이라 추정해 볼 수 있다.

이 중에서 Text2와 Text4의 경우 기분석 사전에 의한 단일 분석 확정 어절이 다른 텍스트들보다 높게 나타났다. 이것은 이들 텍스트가 '사전'과 '신문'이라는 장르의 성격을 지니고 있어 고유명사와 전문어, 숫자, 기호의 사용이 많은 반면 이를 제외한 어절유형에서는 일반적으로 평이하고 쉬운 어휘들이 사용되었기 때문으로 보인다.

또한 비교적 태깅이 쉬운 장르라고 판단되었던 Text3에서는 가장 많은 다중 분석 어절이 산출되었다. 이것은 소설이기 때문에 대화체 형식이 많아 '나는', '난', '나를', '내', '그' 등 중의성이 있는 대명사 포함 어절 유형들의 사용이 많고, 비교적 문장이 간결하여 규칙처리가 어려운 '가다'와 같은 중의적 용언, '그리고', '그래서' 등의 접속사 등이 많이 사용되었기 때문으로 판단된다. 이와는 다르게 단일 분석 중 확정적인 결정이 어려운 어절 유형이 상대적으로 적음을 [ASSUMP]의 표지를 통해 확인할 수 있다(17%). 이는 Tex2나 Text4와 마찬가지로 비교적 쉽고 많이 통용되는 어휘들이 사용된 결과로 여겨진다.

이와 같이 전체적으로 낮은 비율의 단일 분석 결과는 우선 150만 어절 규모의 형태소 분석 말뭉치가 다양하게 발현되는 한국어의 어절 형태를 충분히 반영할 수 없는 크기라는 점과 다중

분석 어절, 즉 중의성의 해결을 필요로 하는 어절 형태가 형태적, 통사적으로 복잡하게 나타나는 한국어의 특징이 작용하였기 때문으로 판단된다.

5.2 단일 분석 어절의 정확도 검사

오류 검증에 대한 기준은 0.5%로 설정하였는데, 이는 세종 분석 말뭉치 구축에 있어 목표로 하고 있는 오류률이 0.5% 이내이기 때문이다. 따라서 단일 분석 어절에 대한 오류 분석을 실시하여 CETtagger 방식의 타당성을 검증하고자 하였다.

5.2.1 [1TYPE] 유형에서의 오류

텍스트	[1TYPE] 어절수	[1TYPE] 오류수	오류율(%)
Text1	16,218	74	0.46
Text2	10,876	29	0.27
Text3	6,321	25	0.40
Text4	6,832	32	0.47
계	40,247	290	0.40

[표-6] 텍스트 교정후 [1TYPE] 유형에서의 오류

각 텍스트들에서 나타난 오류 유형 중 대부분을 차지하고 있는 것은 공통적으로 일반명사와 고유명사의 구분에 대한 오류였다. 이에 대한 문제는 본 논문에서 제안하는 CETtagger에서 뿐만 아니라 일반적인 태거에서도 심각하게 제기되는 문제이다. 결과는 모든 텍스트에서 [1TYPE] 오류의 비율이 0.5% 미만으로 세종 분석 말뭉치의 구축에 있어 목표로 삼고 있는 오류율의 범위 이내인 것으로 나타났다.

5.2.2 [CR] 유형에서의 오류

[표-7]는 규칙 적용에 있어 오류가 발생한 어절에 대한 통계를 제시한 것이다.

텍스트	[CR] 어절수	[CR] 오류수	오류율(%)	주/보격조사 오류	용언 규칙 오류
Text1	2,779	16	0.58	8	8
Text2	1,303	2	0.15	0	2
Text3	1,087	2	0.18	0	2
Text4	721	5	0.69	2	3
계	5890	25	0.42	10	15

[표-7] [CR] 유형에서의 오류

본 논문에서 사용한 문맥 규칙은 주격 조사와 보격 조사, 본용언과 보조용언을 구분하는 규칙이 주를 이룬다. 오류 유형도 크게 이 두 부분에서 각각 존재하였다. 전자의 경우 주격 조사와 보격 조사가 결합된 어절에서 해당 규칙의 적용은 성공하였으나, 모두 조사와 결합되는 명사의 분석에서 오류가 발생하였다. 주격 조사와 보격 조사의 구분에서 발생된 오류를 제외한 나머지 오류는 동사 '하다'에 대한 규칙이 적용된 어절 '한'에서 모두 발생하였다.

첫　　첫/MM
장을　　장/NNB+을/JKO　장/NNG+을/JKO　장/NNP+을/JKO
한　　**하/VV+ㄴ/ETM**
학생더러　학생/NNG+더러/JKB

이러한 오류에도 불구하고 규칙 적용이 이루어진 어절에 대한 전체 오류율은 0.42%로 기준 오류율을 밑돈다. 이에 따라 [1TYPE]유형과 [CR]유형을 결합한 태깅 오류율은 다음과 같다.

텍스트	[1TYPE]+[CR] 어절수	[1TYPE]+ [CR] 오류수	오류율(%)
Text1	19,110	90	0.47
Text2	12,233	31	0.25
Text3	7,476	27	0.36
Text4	7,595	37	0.49
계	46,404	185	0.40

[표-8] [1TYPE]과 [CR] 유형을 합산한 오류율

결과적으로 단일 분석으로 확정된 어절의 총 오류율은 0.40%로 기준 오류율 0.5% 미만이기 때문에 이에 해당하는 어절들에 대해서는 수작업에 의한 교정을 필요로 하지 않을 것으로 판단된다. 그러나 약 9만 어절에 대한 실험은 결과의 타당성을 검증하기에는 적은 양으로 판단되기 때문에 좀더 다양하고 많은 양에 대한 실험을 통하여 지속적인 신뢰 검증이 필요할 것으로 여겨진다.

5.3 세종 태거에 의한 분석 결과와 정확도 검사

CETtagger에 의해 실험을 실시한 텍스트를 대상으로 세종 태거를 이용하여 자동 태깅을 시도하였다. 각 텍스트에 대한 태깅 결과와 정확도 검사를 실시하여 발생된 오류율은 다음과 같다.

텍스트	총어절수	태깅성공	비율(%)	추정	비율(%)	분석실패	비율(%)
Text1	37,890	37,789	99.73	101	0.27	0	0
Text2	23,005	22,818	99.19	187	0.81	0	0
Text3	14,013	13,584	96.93	429	3.06	0	0
Text4	13,971	13,817	98.90	154	1010	0	0
계	88,879	88,217	99.26	662	0.74	0	0

[표-9] 세종 태거를 이용한 태깅 결과

텍스트	태깅성공 어절수	오류 어절수	오류율(%)
Text1	37,789	4,858	12.86
Text2	22,818	2,470	10.82
Text3	13,584	1,042	7.67
Text4	13,971	2,080	14.89
계	88,879	10,450	11.76

[표-10] 세종 태거의 오류율

전체적으로 88% 이상의 태깅 정확도를 보여 주었는데, 텍스트별로 오류율에 대한 편차가 매우 크다는 사실을 알 수 있다. 이러한 오류 편차에 대한 명확한 이유를 파악하기가 쉽지 않다. 다만 오류 유형이 비교적 적은 Text3보다는 다른 텍스트들에서 오류 어절의 반복 사용이 훨씬 많음을 볼 수 있다.

이렇듯 오류 분석의 차이를 파악하기 힘든 것은 네 가지 텍스트에서 나타나는 오류의 유형이 매우 다양하기 때문인 것으로 여겨진다. 고유명사와 일반명사의 구분 오류, 품사 중의성 해결 실패, 합성명사의 분석 문제, 본용언과 보조 용언의 구분, 지침 수정으로 인한 오류 등 앞에서 살펴보았던 대부분의 오류 유형이 발견되고 있다. 게다가 기호가 부착된 어절의 분석 결과에서 기호가 사라지거나 원어절과 다른 분석 결과가 제시되는 문제점

도 발견되었다.

이러한 오류는 분석 말뭉치 구축에 있어 후처리 작업을 힘들게 하는 요소로 작용한다. 오류의 유형이 다양하기 때문에 후처리 작업에 있어 오류 가능 어절에 대한 예측이 거의 불가능하다. 예를 들어 세종 태거는 같은 어휘가 결합된 '귀하는'과 '귀하에게'이라는 어절에 대해 '귀하/의존명사+는/보조사'와 '귀하/대명사+에게/부사격조사'로 다르게 태깅하였다. 만일 다양한 어휘들에 대해 이와 같은 분석 결과를 생성해 낸다면 그 오류의 양상을 예측할 수 없어 효율적이고 체계적인 작업을 어렵게 만드는 한편 말뭉치의 일관성을 떨어뜨리는 결과를 가져온다. 따라서 약 10% 이상 발생하는 오류를 바로 잡기 위하여 모든 태깅 결과를 일일이 확인해야 하는 작업이 반복되어 일어나게 된다.

5. 결 론

앞에서 살펴보았듯이 CETtagger는 좀더 체계적인 분석 결과를 생성한다. 비록 전체적인 태깅 성공률과 정확도는 다른 태거에 비해 떨어지지만 비교적 정확한 단일 분석 결과를 큰 편차없이 50% 이상으로 산출하고, 해결이 어려운 어절 유형에 대해서 완전히 작업자의 판단에 맡김으로써 오류의 가능성을 줄인다. 또한 분석 어절에 대해 적절한 표지를 부착함으로써 체계적이고 단계적인 후처리 작업이 가능하도록 한다.

지금까지의 논의를 종합해 살펴보면 분석 말뭉치의 구축을 위한 태거는 일반적인 언어 처리를 위한 태거와는 다르다는 점을 확인할 수 있다. 확률적 선택에 의한 자동 태깅 방식보다는 확정적인 어절만 자동 태깅하고 나머지는 수작업으로 처리할 수 있도록 편의를 제공하는 태깅 방식이 필요한 것이다. 따라서 본 논문에서 제안한 반자동 태깅 방식이 말뭉치 구축에 있어 보다 효율적으로 후처리 작업을 수행할 수 있는 안정적 기반을 제공할 수 있을 것으로 판단된다.

본 논문에서 제안하고 있는 반자동 태깅 방식은 여러 면에서 개선할 여지를 안고 있다. 첫째, 기존 태거와 마찬가지로 형태소 분석의 오분석이 그대로 태거에 반영될 가능성을 안고 있다. 이러한 문제는 주로 미등록 어절의 처리 과정에서 나타나는데, CETtagger는 미등록 어절에 대한 분석의 타당성을 검증할 장치가 미비하기 때문에 발생될 수 있는 문제이다. CETtagger는 150만 분석 말뭉치의 태그 결합 유형으로 형태소 분석 결과를 제한하는 방식을 사용한다. 분석 말뭉치의 태그 결합 유형은 12,965개 정도인데, 기호 등에 의한 어절 유형 변화가 다양한 한국어의 특징에서 이러한 결합 정보에 해당하지 않는 유형들이 나타난다. 따라서 보다 정확하게 형태소 분석 결과를 제한할 수 있는 방법론을 개선하는 것이 필요하다.

둘째, 150만 형태소 분석 말뭉치에 의한 기분석 사전은 그 규모가 작기 때문에 문법 형태소에 의해 다양하게 나타나는 한국어의 어절 유형을 모두 반영할 수 없다. 따라서 이후 구축되거나 구축될 분석 말뭉치를 추가하여 방대한 규모의 기분석 사전을 구축한다면 보다 효과적이고 정확한 태깅을 수행할 수 있을 것이다.

셋째, 수작업에 의한 처리의 문제점을 들 수 있다. CETtagger에서 사용하는 방식은 수작업에 의한 처리를 인정하고 분석의 많은 부분을 후처리 단계에서 처리하도록 하고 있다. 그런데 후처리 작업을 수행하는 작업자의 개인적, 학문적인 차이가 결국 전체적인 일관성 결여로 연결되어 분석 말뭉치의 질을 현저히 떨어뜨릴 가능성을 내포하고 있다. 따라서 작업자에 대한 철저한 사전 교육이 필요하고 후처리 작업 각 단계에 따른 일관성 확인 여부 과정을 충분히 수행해야 할 것이다.

세종 분석 말뭉치의 구축에 사용된 세종 태거의 경우 태깅 성공률이 약 88%로 비교적 우수한 성능을 지니고 있다. 따라서 일반적인 언어처리용으로 사용되는데 큰 무리가 없을 것으로 보인다. 하지만 분석 말뭉치의 구축을 위해서 사용되는 과정에서 앞서 살펴보았던 대로 후처리 작업에 어려움을 일으키는 요소들이 존재한다. 말뭉치 구축을 위해서는 궁극적으로 자동화율을 높이고 인간의 수작업을 최소한으로 줄이는 자동 태깅 방식으로 나아가야 하지만 현실적으로 어려움이 존재하기 때문에 본 논문에서는 이러한 어려움을 줄이고, 말뭉치 구축을 좀더 효율적으로 수행할 수 있는 반자동 방법론을 제시하고자 하였다. 이와 같은 방법론을 통해 훌륭한 품질을 지닌 형태 분석 말뭉치가 지속적으로 구축되고 활용되는데 조금이나마 기여를 할 수 있을 것으로 기대해 본다.

[참고문헌]

[1] 김진동, 이상주, 임해창. 1998. "어절 띄어쓰기를 고려한 형태소 단위 품사 태깅 모델", 제10회 한글 및 한국어 정보처리 학술대회 '인간과 기계와 언어' 논문집, 「한글 및 한국어 정보처리」 3-8.

[2] 김진동. 1996. "어절 문맥을 고려하는 형태소 단위의 한국어 품사 태깅 모델", 고려대학교 석사학위 논문.

[3] 김홍규 외. 2000. 「21세기 세종계획 국어 기초자료 구축 연구 보고서」, 문화관광부.

[4] 김홍규, 강범모. 2000. 「한국어 형태소 및 어휘 사용 빈도의 분석 1」, 고려대학교 민족문화연구원.

[5] 이상주, 류원호., 김진동, 임해창. 1998. "품사태깅을 위한 어휘규칙의 자동 획득", 제10회 한글 및 한국어 정보처리 학술대회 '인간과 기계와 언어' 논문집, 「한글 및 한국어 정보처리」 20-27.

[6] 임홍빈. 1998. "한국어 정보 처리를 위한 어절 분석 표지의 표준화 연구", 정광, 이기용, 김홍규,

[7] 정광, 이기용, 김홍규, 임해창, 강범모. 1995. 「한국어 데이터베이스의 설계 및 응용을 위한 기초 연구」, 서울 : 민음사

[8] 한영균 1996. "전산기에 의한 형태분석과 사전정보", 「국어학 27」 251-276.

[9] 한영균 1998. "문어 코퍼스의 형태 정보 주석에서 선결되어야 할 몇 문제", 「한국어 전산학 2집」, 293-307.

[10] Biber, Douglas. Susan, Conard. and Randi, Reppen. 1998. *Corpus Linguistics : Investigating Language Structure and Use (Cambridge Approaches to Linguistics)*, Cambridge University Press.

[11] Rob, Peter. and Carlos, Coronel. 1999. *Database systems : design, implementation, and management(Fourth Edition)*, Belmont, California : Wadsworth Pub. Co.