# 다수 형태소 분석 결과를 활용한 표준 말뭉치 반자동 구축

김태영[O†], 류법모[‡], 김한샘[§], 오효정[†]

전북대학교[†], 부산외국어대학교[‡], 연세대학교[§]
fnty127@hanmail.net[O†], pmryu@bufs.ac.kr[‡], khss@yonsei.ac.kr[§], ohj@jbnu.ac.kr[†]

# Korean Linguistic GS Set Semi-Automatic Construction using Multiple POS taggers

Tae-Young Kim[O†], Pum-Mo Ryu[‡], Han-Saem Kim[§], Hyo Jung Oh[†]
Jeonbuk National University[†], Busan University of Foreign Studies[‡], Yonsei University[§]

**요 약**

최근 한국어 정보처리를 위한 대용량 언어분석 표준 말뭉치(GS:Gold Standard Set)를 구축하고 이를 공유·확산하기 위한 국가차원의 지원이 이뤄지고 있다. 본 연구는 이러한 사업의 일환으로, 현재 국내에서 개발된 다양한 한국어 언어분석 모듈을 활용하여 공통 정답셋을 구축하기 위한 방법론을 제시하고자 한다. 특히, 대량의 학습셋을 구축하기 위해 다수의 모듈(N-modules)로부터 제시된 후보 정답을 참조, 오류 형태를 분류하여 주요 유형을 반자동으로 보정함으로써 수작업을 최소화하였다. 본 연구에서는 우선 첫 단계인 형태소 분석 모듈 적용 결과를 토대로 표준 말뭉치를 구축한 결과에 대해 논하고자 한다.

주제어: 한국어 코퍼스, 반자동 구축, 형태소 분석

## 1. 서론

1990년대 한국어 정보처리 분야에 제1의 전성시대를 맞이한 이래로 많은 기관과 연구진에서 한국어 분석을 위한 코퍼스(corpus) 구축과 관련된 노력이 계속되고 있다. 최근 딥러닝을 비롯한 인공지능 기술의 비약적인 발전으로 고품질 학습데이터 구축에 대한 요구가 더욱 증대되고 있다. 이러한 요구에 부흥하여 정부에서도 한국어 분석을 통한 다양한 산업 활성화를 위하여 표준 말뭉치(GS:Gold Standard Set)를 구축하고 이를 공유·확산하기 위한 국가차원의 지원을 시작했다. 본 연구는 이러한 사업의 일환으로, 현재 국내에서 개발된 다양한 한국어 언어분석 모듈을 활용하여 공통 정답셋을 구축하기 위한 방법론을 제시하고자 한다. 특히 대량의 학습셋을 구축하기 위해 다수의 모듈(N-modules)로부터 제시된 후보 정답을 참조, 오류 형태를 분류하여 주요 유형을 반자동으로 보정함으로써 수작업을 최소화하였다. 이번 논문에서는 본 연구의 최종 목적인 한국어 언어분석 표준 말뭉치 구축을 위해 구문 분석을 적용하여 통합 정답셋(CoNLL-U Format, [1])까지 변환하여야 하지만, 우선 첫 단계인 형태소 분석 모듈 적용 결과를 토대로 표준 말뭉치를 구축한 결과에 대해 논하고자 한다.

## 2. 관련 연구

영어권의 경우 현재 iWeb, NOW, Wikipedia, COCA, COHA, GloWbE 등 대용량의 영어 말뭉치가 <표 1>과 같이 전 세계에 제공되고 있고[2], 전체 영어 텍스트 말뭉치는 관계형 데이터베이스 정보(textID, ID, wordID), word/lemma/PoS 정보, words(paragraph format) 정보, 이렇게 세 가지 형식으로 제공되고 있다.

<표 1> 대용량 영어 말뭉치

| Corpus | Texts (95% available in full-text data) |
|---|---|
| iWeb (The Intelligent Web Corpus) | • 14 billion words<br>• 22 million web pages |
| NOW (News on the Web) | • 6.04 billion words<br>• 6.0+ million texts |
| GloWbE (Global Web-based English) | • 1.9 billion words<br>• 1.8 million texts |
| Wikipedia Corpus | • 1.9 billion words<br>• 4.4 million texts |
| COCA (Corpus of Contemporary American English) | • 560 million words<br>• 220,000 texts |
| COHA (Corpus of Historical American English) | • 400 million words<br>• 107,000 texts |

이 외에도 중국의 경우에도 SIGHAN을 통해 중국어 어휘 분할(word segmentation) 등의 학습 말뭉치를 제공하고 있으며[3], 그 양이 한국어 공개 학습셋 대비 매우 크다. 따라서 고차원의 한국어 분석을 위해서는 대량의 검증된 표준 말뭉치 구축이 필수적이다

<표 2> 중국어 어휘 분할 말뭉치

| Corpus | | Word Types | Words | Character Types | Characters |
|---|---|---|---|---|---|
| Traditional | Academia Sinica | 141,340 | 5,449,698 | 6,117 | 8,368,050 |

| | | | | | |
|---|---|---|---|---|---|
| Chinese | City Univ.of Hong Kong | 69,085 | 1,455,629 | 4,923 | 2,403,355 |
| Simplified Chinese | Peking University | 55,303 | 1,109,947 | 4,698 | 1,826,448 |
| | Microsoft Research | 88,119 | 2,368,391 | 5,167 | 4,050,46 |

## 3. 다수의 형태소 분석기를 이용한 반자동 구축

### 3.1 방법론

본 연구에서 제안하는 방법이 지향하는 궁극의 목적은 가능한 수작업을 최소화하는 동시에 표준화된 대량의 정답셋을 구축하는 것이다. 이를 위해 본 연구에서는 다음과 <그림 1>과 같은 방법론을 제시하며, 세부 수행 과정은 다음과 같다.

우선 첫 번째 단계에서는 대용량 원시 말뭉치에 공개된 다수의 한국어 형태소 분석기를 적용하여 그 결과와 특성을 비교한다. 원시 말뭉치로는 GitHub[4]에 공개되어 있는 신문기사 74만 문장(10,081,411 어절)을 활용하였으며, 다수의 형태소 분석기 결과를 비교한다. 두 번째 단계에서는 다수 형태소 분석기의 품사태깅 결과에 대한 정답 및 오류 유형을 분석하여, 형태소 분석 결과에 대한 표준 변환 규칙 유형을 분류한다. 표준 변환 규칙 내용을 토대로 가이드라인을 생성하여 일괄적으로 자동수정을 한 후, 수작업 검증을 병행한다. 마지막으로는 범언어적인 언어 처리가 용이하도록 U-POS(Universal POS Tagset)을 적용[5]하여 정답셋으로 변환하는 과정을 거치게 된다.


<그림 1> 한국어 표준 말뭉치 반자동 구축 흐름도

본 연구에서는 국내에 공개된 다수의 형태소 분석기 중 한국전자통신연구원(ETRI)[6], 울산대[7], 국민대[8] 형태소 분석기를 선정하였다. 선정 이유는 세 기관의 형태소 분석기 모두 현재 웹 상에 공개되어 있고, 21세기 세종계획 형태 태그셋을 활용하기 때문에 공통된 태그셋을 기반으로 표준 말뭉치를 구축하기에 적합하다고 판단했기 때문이다. 또한 세 형태소 분석기 모두 기본 성능이 95%인 점을 감안하여, 세 기관의 결과가 모두 같은 경우에는 이를 정답으로 간주하였다.

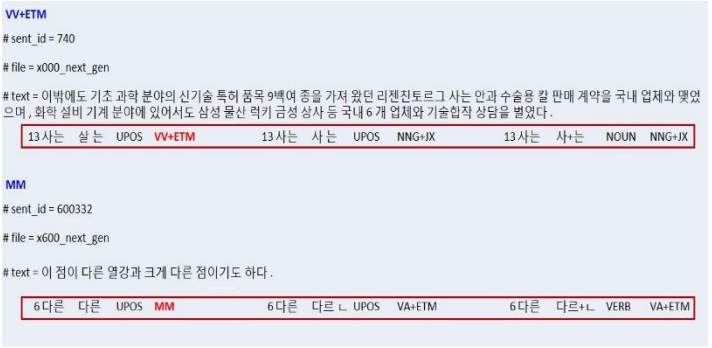
### 3.2 다수의 형태소 분석 결과에 대한 표준 변환 규칙

본 절에서는 세 기관(A, B, C로 표기) 각각의 형태소 분석 결과를 아래 <표 3>과 같은 유형으로 나누어 비교·분석하였다. 본 연구에서 선정한 세 기관의 형태소 분석기 기본 성능이 모두 95% 이상인 점을 감안하여 [유형 1]은 정답으로 간주하고, [유형 2]에 대한 보정작업을 수행하였다. 이를 위해 각 유형별 특성을 분석하여 일괄 변환 규칙을 정의하였다.

<표 3> 기관별 형태소 분석 결과 비교

| 구분 | 비고 | 유형 | 어절 수 |
|---|---|---|---|
| 1 | 정답으로 간주 | 세 기관 모두 같음 | 7,644,916 (공통문장수: 4,423) |
| 2-1 | 변환규칙 및 가이드라인 적용 대상 | B/C 같고 A만 다름 | 550,679 |
| 2-2 | | A/C같고 B만 다름 | 686,368 |
| 2-3 | | A/B 같고 C만 다름 | 888,361 |
| 3 | 수작업 검증 | 세 기관 모두 다름 | 248,087 |

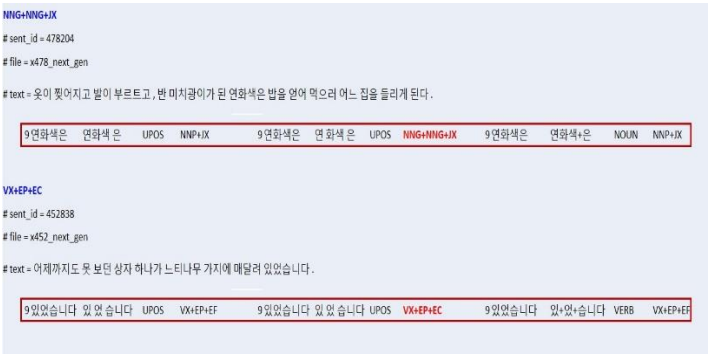
첫 번째 변환 규칙은 품사 변별에서의 차이에서 나타난 것으로 <그림 2> 예시를 살펴보면, '사는' 은 용언 '살다(VV)' 의 활용형이고 '다른(MM)' 은 관형사임에도 불구하고 B와 C 분석기는 올바른 품사를 판별해내지 못했다. 이 같은 유형은 품사를 올바르게 분석하여 태깅한 A 분석기 결과로 B와 C 분석기의 오류를 수정해야 한다.




<그림 2> 변환 규칙 (1) 예시

두 번째 변환 규칙은 고유명사 처리, 어미 세부 유형 변별에서의 차이에서 나타난 것으로 <그림 3> 예시를 살펴보면, B 분석기는 고유명사 '연화색(NNP)' 을 미등록

어로 인식하여 일반명사로 분석한 결과를 보여주었다. 더불어 문장 구조상 종결어미(EF)어야 함에도 불구하고 연결어미(EC)로 품사를 인식하여 태깅하는 오류를 발생시켰다. 이 같은 경우에도 올바르게 분석하여 태깅한 A와 C 분석기 결과로 B 분석기의 오류를 수정해야 한다.




<그림 3> 변환 규칙 (2) 예시

세 번째 변환 규칙은 용언과 체언의 통합 및 분해에서의 차이에서 나타난 것으로 <그림 4> 예시를 살펴보면, 용언의 경우 A와 B 형태소 분석기는 '시급하-'를 하나의 용언으로 보고 통합형으로 태깅하였으나, C 형태소 분석기는 어근(XR)과 형용사파생접미사(XSA)를 따로 분할하여 태깅하였다. 체언의 경우는 미등록어 '상담자'에 대해 A와 B의 형태소 분석기가 더 자세히 분석하였다. 이 같은 유형에는 보다 상세하게 분석된 결과물을 합치는 것이 활용에 용이하므로 최장 일치 규칙을 적용하여 오류를 수정해야 한다.


<그림 4> 변환 규칙 (3) 예시

상기 각 유형에 해당하는 고빈도 100개 유형을 수작업으로 검증, 일괄 변환 규칙을 다음 <표 4>와 같이 정의하였다.

<표 4> 형태소 분석 결과에 대한 표준 변환 규칙

| 규칙유형 | 설명 |
|---|---|
| 1 | 다수의 형태소 분석기 정답 쪽으로 일괄 조정 |
| 2 | 다수의 형태소 분석기 어느 곳에서도 정답을 내지 못함 (일괄변환 불가) |
| 3 | 용언 분석-통합 불일치 -> 최장 규칙 적용 |
| 4 | 체언 분석-통합 불일치 -> 최장 규칙 적용 |

### 3.3 대용량 한국어 표준 말뭉치 구축 결과

앞서 정의한 변환 규칙을 적용하여 한 문장내 정답만을 포함한 문장을 최종 추출한 결과, <표 5>와 같이 348,229 문장, 총 9,455,930 어절을 한국어 표준 말뭉치로 구축하였으며 현재 GitHub에 공개를 준비 중이다.

<표 5> 기관별 형태소 분석에 대한 변환 규칙 적용 결과

| 구분 | 유형 | 어절 수 |
|---|---|---|
| 1 | 세 기관 모두 같음 | 9,455,930 (문장수: 348,229) |
| 2-1 | B/C 같고 A만 다름 | 74,807 |
| 2-2 | A/C같고 B만 다름 | 64,856 |
| 2-3 | A/B 같고 C만 다름 | 174,731 |
| 3 | 세 기관 모두 다름 | 248,087 |

<표 5>에서 나타나듯이 전체 정답 어절은 9백 45만 어절임에도 불구하고 문장 전체가 정답인 경우는 35만여 문장으로 매우 적게 취합되었다. 이를 보완하기 위해 <표 6>과 같이 한 문장 전체가 정답인 4,196,505 어절을 제외한 나머지 5,259,425 어절 중 전체 문장에서 다른 어절이 1-2개 내외인 문장을 대상으로 수작업 검증할 예정이다.

<표 6> 오류 미수정 문장 추가 검증 계획

| 구분 | 어절 수 |
|---|---|
| 공통문장(348,229) 어절 수 | 4,196,505 |
| 전체 문장 중 다른 어절이 1개 | 270,752 |
| 전체 문장 중 다른 어절이 2개 | 87,570 |

## 4. 결 론

본 연구에서는 한국어 분석을 위한 대용량 코퍼스 구축을 위해 국내에서 개발된 다수의 형태소 분석기를 활용하여 표준 말뭉치를 구축하고자 하였다. 특히, 대량의 학습셋 구축을 위해 다수의 형태소 분석기로부터 제시된 후보 정답을 참조, 오류 형태를 분류하여 자동 변환 규칙을 생성하고 가이드라인을 구축하였다. 이후 일괄적으로 자동 수정한 후, 수작업 검증을 병행하여 국내 최대의 한국어 언어분석 표준 말뭉치를 구축하였다. 본 연구를 통해 구축된 표준 말뭉치은 차후 한국어 정보처리를 위한 기초 학습자원으로 활용될 수 있다. 향후 연구로는 오류 미수정 문장에 대한 추가 검증을 수행할 예정이며, 추가적으로 완전한 한국어 언어분석 표준 말뭉치 구축을 위해 구문 분석을 적용한 U-POS 정답셋 변환을 수행할 예정이다.

## 감사의 글

이 논문은 2017년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017M3C4A7068186).

## 참고문헌

[1] CoNLL-U Format, https://universaldependencies.org/format.html
[2] Full-text corpus data, https://www.corpusdata.org/corpora.asp
[3] Third International Chinese Language Processing Bakeoff, http://sighan.cs.uchicago.edu/bakeoff2006/download.html
[4] 한국어 대용량 원시말뭉치, http://nlp.kookmin.ac.kr/kcc/
[5] 박혜진, 오태환, 김한샘, “Universal Dependency를 위한 한국어 형태 주석 체계 연구”, 언어와 정보, 22권, 3호, pp.67-89, 2018.
[6] 공공 인공지능 오픈 API·DATA 서비스 포털, http://aiopen.etri.re.kr/service_api.php
[7] UTagger, http://nlplab.ulsan.ac.kr/doku.php?id=utagger
[8] 한국어 형태소 분석기와 한국어 분석 모듈, http://nlp.kookmin.ac.kr/HAM/kor/index.html