대용량 표준 말뭉치 구축을 위한 다수 형태소 분석 결과 통합 방법론

Unified Methodology of Multiple POS Taggers for Large-scale Korean Linguistic GS Set Construction

저자
(Authors)

김태영, 류법모, 김한샘, 오효정
Tae-Young Kim, Pum-Mo Ryu, Hansaem Kim, Hyo-Jung Oh

출처
(Source)

정보과학회논문지 47(6), 2020.6, 596-602 (7 pages)
Journal of KIISE 47(6), 2020.6, 596-602 (7 pages)

발행처
(Publisher)

한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09353168

APA Style

김태영, 류법모, 김한샘, 오효정 (2020). 대용량 표준 말뭉치 구축을 위한 다수 형태소 분석 결과 통합 방법론. 정보과학회논문지, 47(6), 596-602.

ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 47, No. 6, pp. 596-602, 2020. 6
https://doi.org/10.5626/JOK.2020.47.6.596

대용량 표준 말뭉치 구축을 위한 다수 형태소 분석 결과 통합 방법론

(Unified Methodology of Multiple POS Taggers for Large-scale Korean Linguistic GS Set Construction)

김 태 영 † (Tae-Young Kim)　류 법 모 †† (Pum-Mo Ryu)　김 한 샘 ††† (Hansaem Kim)　오 효 정 †††† (Hyo-Jung Oh)

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 말뭉치 구축 사업의 일환으로, 현재 국내에서 개발된 다양한 한국어 언어분석 모듈을 활용하여 공통 정답셋 구축을 위한 방법론을 제안하고자 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 형태소 분석 모듈 적용 결과를 정규화하여 통합 포맷인 U-POS를 기반으로 대용량 한국어 언어분석 표준 말뭉치를 구축하였다. 본 연구를 통해 348,229 문장, 총 9,455,930 어절이 한국어 표준 말뭉치로 구축되었으며, 이는 차후에 한국어 정보처리를 위한 기초 학습자원으로 활용될 수 있다.

키워드: 한국어 코퍼스, 형태소 분석, 품사 태깅, 반자동 구축

Abstract In recent years, there has been national support for constructing, sharing, and spreading a large-scale Korean linguistic GS set for Korean information processing. As part of the corpus construction project, this study proposes the methodology for constructing the Korean linguistic GS set using various Korean language analysis modules developed in Korea. To build a large-scale training set, we referred to automatic tagged candidate answers from the N-modules. We then minimized manual effort by classifying the error types from the candidate responses and semi- automatically correcting the major error types. In this study, we normalized results of the morphological analysis and constructed a large-scale Korean linguistic GS set based on the unified format U-POS. As a result of this study, 348,229 sentences, a total of 9,455,930 words, were constructed as the Korean linguistic GS set. This can be practically applied later as a basic training resource for Korean information processing.

Keywords: Korean corpus, morphological analysis, POS tagging, semi-automatic construction

· 이 논문은 2020년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음
· 이 논문은 2017년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017M3C4A7068186)
· HCLT2019: 이 논문은 제31회 한글 및 한국어 정보처리 학술대회에서 '다수 형태소 분석 결과를 활용한 표준 말뭉치 반자동 구축'의 제목으로 발표된 논문을 확장한 것임

† 정 회 원 : 전북대학교 기록관리학과 박사
fnty127@hanmail.net
†† 정 회 원 : 부산외국어대학교 사이버경찰전공 교수
pmryu@bufs.ac.kr
††† 비 회 원 : 연세대학교 언어정보연구원 교수
khss@yonsei.ac.kr
†††† 비 회 원 : 전북대학교 문헌정보학과 교수(Jeonbuk Nat'l Univ.)
ohj@jbnu.ac.kr
(Corresponding author임)

논문접수 : 2020년 3월 6일
(Received 6 March 2020)
논문수정 : 2020년 4월 6일
(Revised 6 April 2020)
심사완료 : 2020년 4월 7일
(Accepted 7 April 2020)

정보과학회논문지 제47권 제6호(2020. 6)

1. 서 론

언어처리 기술은 범람하는 텍스트 정보를 분석하고 이해하기 위한 가장 기본적인 기반 기술로, 각 국가마다 자국어의 특성에 맞는 어휘 지식을 구축하기 위한 노력이 계속되고 있다. 국내에서는 1990년대 한국어 정보처리 분야에 제1의 전성시대를 맞이한 이래로, 많은 기관과 연구진들이 한국어 분석을 위한 말뭉치(corpus) 구축과 관련된 연구들을 진행하고 있다. 특히, 최근에는 딥러닝을 비롯한 인공지능 기술의 비약적인 발전으로 고품질의 학습데이터 구축에 대한 요구가 더욱 증대되고 있다.

자연어처리(NLP: Natural Language Processing) 기술은 인공지능 개발에 있어서 핵심적인 요소로, 대용량·고품질의 말뭉치를 기반으로 컴퓨터를 학습시킴으로써 인공지능의 정확성과 정밀성 향상이 가능하다. 따라서 제4차 산업혁명 시대에 국가 경쟁력 향상을 위해서는 대용량·고품질의 한국어 말뭉치 구축이 필수적이다. 이러한 요구에 부흥하여 정부에서도 한국어 분석을 통한 다양한 산업 활성화를 위해 표준 말뭉치(GS: Gold Standard Set)를 구축하고, 이를 공유·확산하기 위한 국가차원의 지원을 시작했다.

이에 한국어 인공지능 발전을 위해 문화체육관광부·국립국어원에서는 2018년부터 2022년까지 총 154억 7천만 어절의 말뭉치를 구축하는 국어 정보화사업을 계획하였다[1]. 또한 한국전자통신연구원(이하 ETRI)에서는 다양한 지식산업 환경에서 전문가 수준의 질의응답 서비스의 제공하기 위해 여러 연구기관과 협력하여 다양한 언어처리 학습데이터(엑소브레인 말뭉치 v4.0)를 제공하고 있다. 현재 공개된 언어처리 학습데이터 중에서도 엑소브레인 언어분석 말뭉치와 QA Datasets은 한국정보통신기술협회(TTA), 국가기술표준원 KS 표준안에 입각하여 구축되었다[2].

본 연구는 이러한 말뭉치 구축 사업의 일환으로, 현재 국내에서 개발된 다양한 한국어 언어분석 모듈을 활용하여 공통 정답셋 구축을 위한 방법론을 제안하고자 한다. 특히, 대량의 학습셋을 구축하기 위해 본 연구에서는 다수의 모듈(N-modules)로부터 제시된 후보 정답을 참조, 오류 형태를 분류하여 주요 유형을 반자동으로 보정함으로써 수작업을 최소화하였다. 본 연구의 최종 목적인 한국어 언어분석 표준 말뭉치 구축을 위해서는 통합 정답셋(CoNLL-U Format, [3])으로의 변환이 필수적이다. 이에 본 연구에서는 이를 위한 첫 단계로서 형태소 분석 모듈 적용 결과를 정규화하여 U-POS 기반으로 대용량 표준 말뭉치를 구축한 결과에 대해 논하고자 한다.

2. 관련 연구

2.1 대용량 말뭉치 구축

영어권의 경우 현재 iWeb, NOW, Wikipedia, COCA, COHA, GloWbE 등 대용량의 영어 말뭉치가 표 1과 같이 전 세계에 제공되고 있고[4], 전체 영어 텍스트 말뭉치는 관계형 데이터베이스 정보(textID, ID, wordID), PoS 정보(word/lemm/pos), words(paragraph format) 정보, 이렇게 세 가지 형식으로 제공되고 있다.

한편, 중국에서도 표 2와 같이 SIGHAN을 통해 중국어 어휘 분할(word segmentation) 등의 학습 말뭉치를 제공하고 있으며[5], 그 양이 한국어 공개 학습셋 대비 매우 크다. 따라서 고차원의 한국어 분석을 위해서는 대량의 검증된 표준 말뭉치 구축이 필수적이다.

이 외에 노르웨이의 경우에도 The Norwegian Dependency Treebank(NDT)를 UD(Universal Dependencies) 스키마로 변환하여 배포하고 있다. 노르웨이어 학습자의 에세이 말뭉치인 ASK corpus(Norsk andrespråkskorpus)는 표 3과 같이 텍스트 태그셋 원본과 오류 수정본으로 구성되어 있다[6].

현재 국내에서는 한국어의 특성을 반영하여 ETRI에서 개발한 KoBERT(Korean Bidirectional Encoder Representations from Transformers) 언어모델이 공개되어 활용되고 있으며, 학습을 위한 한국어 언어모델 말뭉치로 신문기사와 백과사전 등 대용량의 텍스트로부터 추출된 47억 개의 형태소가 활용되었다[2].

2.2 통합 포맷 : U-POS

U-POS(Universal POS Tagset)는 병렬 언어 처리를 위해 형태·통사적 특성을 찾는 것을 목표로 The Universal Dependencies 프로젝트(이하, UD 프로젝트)에서 사용하는 주석체계이다. U-POS는 Stanford Dependencies

표 1 대용량 영어 말뭉치
Table 1 Large-scale English Corpus

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

표 2 중국어 어휘 분할 말뭉치
Table 2 Chinese Word Segmentation Corpus

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

표 3 노르웨이어 ASK 말뭉치
Table 3 Norwegian ASK Corpus

Corpus	Language(s)	Size(words &punctuation)
ASK Hovedkorpus	Norwegian Bokmål (nob)	768,043
ASK Korrektkorpus	Norwegian Bokmål (nob)	785,451
ASK Tillegg	Norwegian Bokmål (nob)	44,529
ASK Hovedk./2015	Norwegian Bokmål (nob)	36,142

의 주석체계와 Google Universal Part-of-Speech Tags의 주석체계, Interset Interlingua for Morphosyntactic Tagsets의 주석체계를 기반으로 설계되었다[7].

U-POS 주석체계는 모든 언어에 공통적으로 적용될 수 있도록 고안된 체계이며, CoNLL-U Format과 같은 하나의 통일된 형식으로 변환됨으로써 범언어적인 언어 처리가 가능하다. 그리고 이러한 U-POS 주석체계는 열린 범주(ADJ, ADV, INTJ, NOUN, PROPN, VERB, NUM)와 닫힌 범주(ADP, AUX, CCONJ, DET, PART, PRON, SCONJ), 그리고 그 외(PUNCT, SYM, X)로서, 크게 세 가지 범주로 구분된다. 각 표지들은 대부분의 언어에서 높은 빈도로 나타나는 품사들을 묶은 것이지만, 모든 언어의 품사가 17개로 한정되지는 않는다. 다만, UD 프로젝트에 참여하는 언어들은 위의 공통된 체계를 기반으로 개별 언어의 품사 범주 체계를 U-POS에 대응시켜 활용하고 있다[7,8].

U-POS를 한국어에 적용시키기 위해서는 공백으로 분리되는 어절을 한국어 주석 단위로 정하고, 어절 내부의 형태 주석 표지 조합을 상정하여 그 기능에 따라 일대다의 매핑(Mapping)을 시도해야 한다. 즉, U-POS 태그를 한국어 주석체계인 '21세기 세종계획 형태 주석 표지'에 대응시킴으로써 한국어에 U-POS 적용이 가능하다[8].

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 1과 같은 방법론을 제시하며, 세부 수행 과정은 다음과 같다.

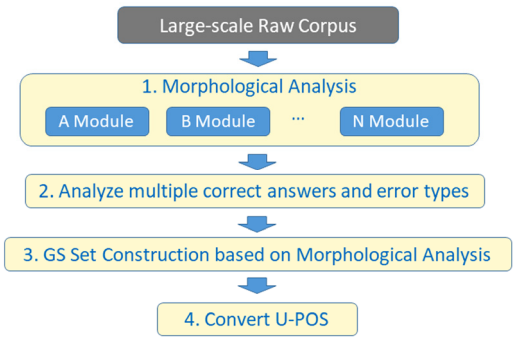

그림 1 한국어 표준 말뭉치 반자동 구축 흐름도
Fig. 1 Semi-automatic Construction Flowchart of Korean Linguistic GS Set

우선 첫 번째 단계에서는 대용량 원시 말뭉치에 공개된 다수 n개의 한국어 형태소 분석기를 적용하여 그 결과와 특성을 비교한다. 원시 말뭉치로는 GitHub[9]에 공개되어 있는 신문기사 74만 문장(10,081,411 어절)을 활용하였으며, 다수 n개의 형태소 분석기 결과를 비교한다. 두 번째 단계에서는 다수 n개 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-POS을 적용[8]하여 정답셋으로 변환하는 과정을 거치게 된다.

본 연구에서는 국내에 공개된 다수 n개의 형태소 분석기 중 ETRI[2], 울산대[10], 국민대[11] 형태소 분석기를 선정하였다. 선정 이유는 세 기관의 형태소 분석기 모두 현재 웹상에 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 정규화

본 절에서는 세 기관(A, B, C로 표기) 각각의 형태소 분석 결과를 아래 표 4와 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 정의하고, 일관성 유지를 위해 오류를 수정하였다.

3.2.1 변환 규칙 적용

본 장에서는 세 가지 형태의 분석기로 자동 주석하여 만들어진 말뭉치 결과를 비교·분석하여 변환 규칙을 제시하였다. 변환 규칙은 크게 세 가지로서 각각의 변환 규칙에 대한 설명과 일부 예시를 제시하고자 한다.

첫 번째 변환 규칙은 품사 변별에서의 차이에서 나타난 것으로. 그림 2의 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.

두 번째 변환 규칙은 고유명사 처리 변별에서의 차이에서 나타난 것으로, 그림 3의 예시를 살펴보면, B 분석기는 고유명사 '연화색(NNP)'을 미등록어로 인식하여 일반명사로 분석한 결과를 보여주었다. 이 같은 유형에서도 품사를 올바르게 분석하여 태깅한 A와 C 분석기의 결과로 B의 오류를 수정해야 한다.

세 번째 변환 규칙은 용언과 체언의 통합 및 분해에서의 차이에서 나타난 것으로, 그림 4의 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 대해A와 B의 형태소 분석기가 더 자세히 분석하였다. 이 같은 유형에는 보다 상세하게 분석된 결과물을 합치는 것이 활용에 용이하므로 최장 일치 규칙을 적용하여 오류를 수정해야 한다.

3.2.2 일관성 유지를 위한 오류 수정

첫 번째 수정 대상은 특정 토큰(token)의 일관성 오류로 '및, 혹은, 또는, 즉'의 품사를 MAG(일반부사)에서 MAJ(접속부사)로 일괄 변경하였다. 두 번째 수정 대상은 연결어미(EC), 종결어미(EF)의 일관성 오류로 문장 구조상 종결어미(EF)어야 함에도 불구하고 연결어미(EC)로 품사를 인식하여 태깅되는 오류가 발생하였기 때문에 이를 수정해야 한다.

표 4 기관별 형태소 분석 결과 비교

Table 4 Comparison of Morphological Analysis Results by Institution

No.	Analysis Type		Number of Words
1	Answer	A/B/C are the same	7,644,916 (Number of Common Sentences: 4,423)
2-1	Rule-based Conversion	B/C are the same, only A is different	550,679
2-2		A/C are the same, only B is different	686,368
2-3		A/B are the same, only C is different	888,361
3	Manual Verification	A/B/C are different	248,087

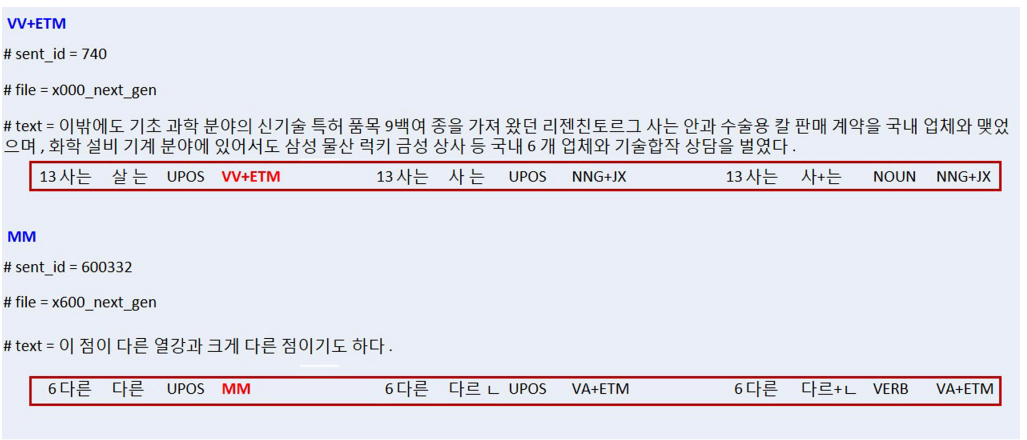

그림 2 첫 번째 변환 규칙 예시

Fig. 2 First Conversion Rule Example

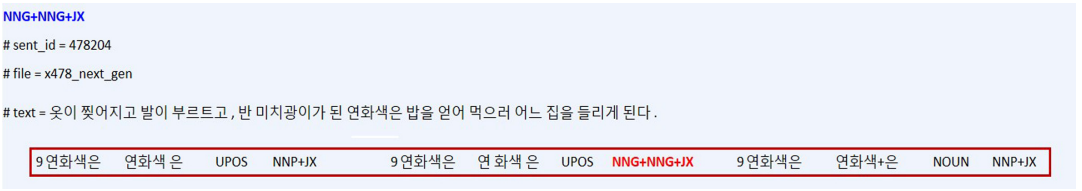

그림 3 두 번째 변환 규칙 예시
Fig. 3 Second Conversion Rule Example

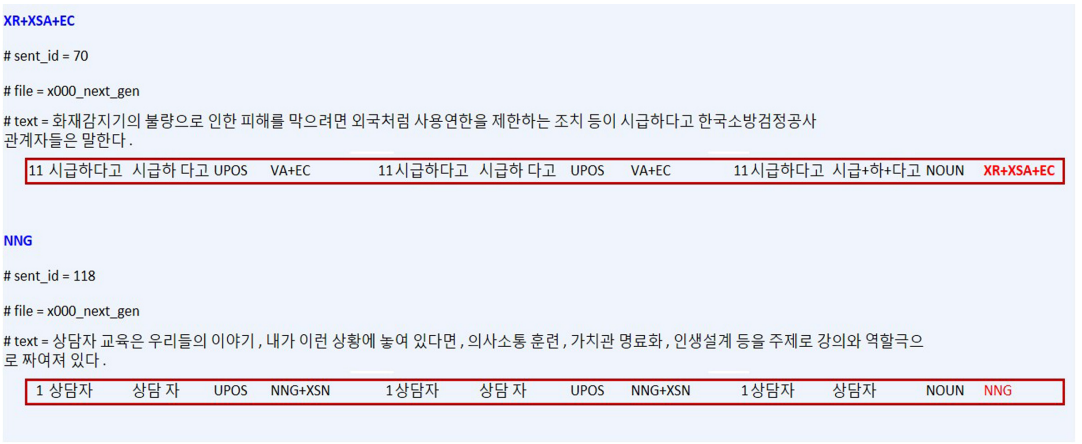

그림 4 세 번째 변환 규칙 예시
Fig. 4 Third Conversion Rule Example

수정 방법 방법은 크게 세 가지로 문장 중간 토큰이 종결어미(EF)인 경우 연결어미(EC)로 변환(단, 다음 토큰이 기호인 경우는 제외), 문장 마지막 토큰이 연결어미(EC)인 경우 종결어미(EF)로 변환(단, 다음 토큰이 마침표, 물음표, 느낌표(SF)인 경우만 적용), 마지막 토큰이 연결어미(EC)로 끝나는 경우 종결어미(EF)로 변환하였다.

표 5 형태소 분석 결과에 대한 표준 변환 규칙
Table 5 Standard Conversion Rule for Morphological Analysis Results

Conversion Rule Type	Rule Description
1	Batch adjustment with multiple morphological analyzer answer
2	Multiple morphological analyzer do not provide the answer (No batch conversion)
3-1	Verb Analysis-Integration Discrepancy → Apply longest rule
3-2	Noun Analysis-Integration Discrepancy → Apply longest rule

상기 각 변환 및 오류 유형에 해당하는 고빈도 100개 유형을 수작업으로 검증한 다음, 형태소 분석 결과에 대한 일괄 변환 규칙을 표 5와 같이 정의하였다.

이상의 변환 규칙을 적용하여 형태소 분석 결과를 정규화하는 방법은 아래 표 6과 같으며, 이는 각각의 변환 규칙 사례 중에서 우선순위 변환 대상이거나 오류가 많이 나타난 예시를 정리한 결과이다.

4. 대용량 한국어 표준 말뭉치 구축 결과

4.1 정규화 결과

앞서 정의한 변환 규칙을 적용하여 한 문장 내 정답만을 포함한 문장을 최종 추출한 결과, 표 7과 같이 348,229 문장, 총 9,455,930 어절을 한국어 표준 말뭉치로 구축하였으며 현재 GitHub에 공개를 준비 중이다.

표 7에서 나타나듯이 전체 정답 어절은 9백 45만 어절임에도 불구하고 문장 전체가 정답인 경우는 35만여 문장으로 매우 적게 취합되었다. 이를 보완하기 위해 다음 표 8과 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 규칙 기반 형태소 분석 결과 정규화 예시
Table 6 Normalize Rule-based Morphological Analysis Results Example

Conversion Rule Type	Normalization Result
1	**NNG+XSV → VV** 공부 + 하 → 공부하(VV)
	NNG+VV → VV 존중 + 받 → 존중받(VV)
	NNG+XSA → VA 적정 + 하 → 적정하(VA)
	NNG+VA → VA 독기 + 어리 → 독기어리(VA)
2	**NNP+NNP → NNP** 경기 + 포천 → 경기포천(NNP)
	NNP+NNG → NNP 양주 + 시청 → 양주시청(NNP)
	NNG+NNP → NNP 더본 + 코리아 → 더본코리아(NNP)
	XPN+NNP → NNP 반 + 스탈린주의 → 반스탈린주의(NNP)
	NNP+SN → NNP 갤럭시노트 + 7 → 갤럭시노트7(NNP)
3-1	**MAG+XSA → VA** 짜릿 + 하 → 짜릿하(VA)
	XR+XSA → VA 치밀 + 하 → 치밀하(VA)
	NNG+XSN → NNG 위원 + 장 → 위원장(NNG)
3-2	**NNG+NNG → NNG** 입당 + 원서 → 입당원서(NNG)
	XPN+NNG → NNG 초+강수(NNG)
	SN+NR → NR 500+만(NR)

표 7 기관별 형태소 분석에 대한 변환 규칙 적용 결과
Table 7 Results of Applying Conversion Rule for Morphological Analysis by Institution

No.	Analysis Type	Number of Words
1	A/B/C are the same	9,455,930 (Number of Sentences: 348,229)
2-1	B/C are the same, only A is different	74,807
2-2	A/C are the same, only B is different	64,856
2-3	A/B are the same, only C is different	174,731
3	A/B/C are different	248,087

표 8 오류 미수정 문장 추가 검증 계획
Table 8 Additional Verification Plan for Uncorrected Sentences

Type	Number of Words
Number of words in common sentences (348,229)	4,196,505
One word in the whole sentence is different	270,752
Two word in the whole sentence is different	87,570

표 9 U-POS와 세종계획 형태 주석 변환표
Table 9 U-POS and Sejong Project Tagsets Conversion Table

	U-POS Tags	U-POS Name	U-POS Translation	Sejong Project Tagsets (Word Unit)
1	ADJ	Adjective	형용사	MM(성상 관형사) VA+E VCN+E ([NNG, NNP, MAG, XR])+XSA+E ([N, MAG, SN])+VCP+E
2	ADP	Adposition	부치사	(JK, JX)
3	ADV	Adverb	부사	MAG
4	AUX	Auxiliary	조동사	VX+E
5	CCONJ	Coordinating Conjunction	등위 접속사	MAJ(및, 또는) JC
6	DET	Determiner	한정사	MM(수·성상 관형사를 제외한 관형사)
7	INTJ	Interjection	감탄사	IC
8	NOUN	Noun	명사	[NNG, NNB] (+[JK, JX])
9	NUM	Numeral	수사	[NR, SN](+[JK, JX]) MM(수 관형사)
10	PART	Particle	불변화사	(EP, EC, EF, ET, XP, XS)
11	PRON	Pronoun	대명사	NP(+[JK, JX])
12	PROPN	Proper Noun	고유명사	NNP(+[JK, JX])
13	PUNCT	Punctuation	구두점	SF, SP, SS, SE, SO
14	SCONJ	Subordinating Conjunction	종속 접속사	MAJ('및, 또는'을 제외한 모든 접속 부사)
15	SYM	Symbol	기호	SW
16	VERB	Verb	동사	VV+E ([NNG, NNP, MAG, XR])+XSV+E
17	X			SL, SH, NA, NF, NV

4.2 U-POS 변환

U-POS는 형태 단위가 아닌 공백에 따라 나누어지는 단위(한국어에서는 어절에 해당)에 따라 주석하는 것이 기본 원칙이지만, 개별 언어의 특성에 따라서 형태 단위

의 분석까지도 허용하고 있다. 따라서 본 연구에서는 한국어 주석 단위를 어절로 설정하여 앞의 표 9와 같이 구축된 U-POS와 '21세기 세종계획 형태 주석표지' 간의 대응 체계[7,8]를 참조하여 대용량 표준 말뭉치를 구축하였다. 이와 같이 U-POS 대응 체계를 기반으로 말뭉치를 구축해야 차후에 UD 정답셋 구축이 가능하다.

5. 결 론

본 연구에서는 한국어 분석을 위한 대용량 어휘자원 구축의 일환으로, 국내에서 개발된 다수의 형태소 분석기를 활용하여 표준 말뭉치를 구축하고자 하였다. 특히, 대용량의 정답셋 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. 또한 최종 결과를 통합 포맷인 U-POS로 변환함으로써 궁극적으로 CoNLL-U Format의 구문분석(UD)으로 확장하기 위한 기초자원을 마련하였다.

본 연구를 통해 348,229 문장, 총 9,455,930 어절이 한국어 표준 말뭉치로 구축되었으며, 현재 GitHub에 공개를 준비 중이다. 본 연구에서 구축한 표준 말뭉치는 차후 한국어 정보처리를 위한 기초 학습자원으로 활용될 수 있다. 향후 연구로는 오류 미수정 문장에 대한 추가 검증을 수행할 예정이며, 나아가 본 논문에서 제안한 방법론을 다수의 구문 분석 모듈에 적용함으로써 국내 최대 구문분석 표준 말뭉치를 구축하고자 한다.

References

[1] Constructed 15.5 Billion Words of Korean Linguistic Set for AI, [Online]. Available: https://www.yna.co.kr/view/AKR20171008048600005

[2] AI API · DATA, [Online]. Available: http://aiopen.etri.re.kr/index.php

[3] CoNLL-U Format, [Online]. Available: https://universaldependencies.org/format.html

[4] Full-text Corpus Data, [Online]. Available: https://www.corpusdata. org/corpora.asp

[5] Third International Chinese Language Processing Bakoff, [Online]. Available: http://sighan.cs.uchicago.edu/bakeoff2006/download.html

[6] ASK - Norsk andrespråkskorpus, [Online]. Available: http://clarino.uib.no/korpuskel/corpus-list?collection=ASK

[7] National Institute of the Korean Language, *Research and establishment of Korean corpus in 2018*, National Institute of the Korean Language, 2018.

[8] H. Park, T. Oh, and H. Kim, Universal POS Tagset for Korean, *Language and Information*, Vol. 22, No. 3, pp. 67-89, 2018.

[9] Korean Large-scale Raw Corpus, [Online]. Available: http://nlp.kookmin.ac. kr/kcc/

[10] UTagger, [Online]. Available: http://nlplab.ulsan.ac.kr/doku.php?id= utagger

[11] Korean Morphological Analyzer and Korean analysis module, [Online]. Available: http://nlp.kookmin.ac.kr/HAM/kor/index.html

김 태 영

2015년 전북대학교 기록관리학과(석사) 2020년 전북대학교 기록관리학과(박사) 2019년~현재 전북대학교 재난안전정보표준화사업단 연구원. 관심분야는 이용자 행태, 빅데이터 분석, 온톨로지, 네트워크 분석, 시소러스

류 법 모

1995년 경북대학교 컴퓨터공학(학사). 1997년 POSTECH 컴퓨터공학과(석사). 2009년 KAIST 전산학과(박사). 2009년~2015년 한국전자통신연구원 지식마이닝연구실 책임연구원. 2015년~현재 부산외국어대학교 사이버경찰전공 부교수. 관심분야는 정보검색, 자연어처리, 온톨로지, 질의응답기술 등

김 한 샘

1998년 연세대학교 국어국문학과 졸업(학사). 2000년 연세대학교 국어정보학 협동과정 졸업(석사). 2005년 연세대학교 언어정보학 협동과정 졸업(박사). 현재 연세대학교 언어정보연구원 부교수. 관심분야는 언어 자원 구축, 자연 언어 처리

오 효 정

2008년 한국과학기술원 컴퓨터공학과(박사). 2000년~2015년 한국전자통신연구원 지식마이닝연구실 책임연구원. 2015년~현재 전북대학교 문헌정보학과 부교수 관심분야는 정보검색, 텍스트마이닝, 빅데이터 정보처리