Artficial Intelligence Laboratory

세분화된 한국어 형태소 규칙에 기반한 의존구문분석 모델 개발

부산대학교 정보컴퓨터공학과
201724443 김준기
201724465 박기훈
201724568 정대성

부산대학교
PUSAN NATIONAL UNIVERSITY

주요내용

- 현재 의존구문분석의 연구
- 기존 시스템의 한계 및 부산대 구문분석기
- 형태소 분석기 문제 & 대안
- 제안 시스템 구조
- 실험 결과
- 시스템 시연
- 한계점 및 향후 연구

현재 의존구문분석의 연구

의존구문분석이란

- 자연어 문장을 지배소-피지배소 의존 관계로 분석하는 구문 분석 방법론
- 문장의 구조적 중의성 해소 가능
- 어순이 고정적이지 않고 문장 성분의 생략이 빈번한 한국어에 적합
- 의존관계 레이블 : 구문태그_기능태그 형태로 태그를 결합하여 사용
  - 예) NP_SBJ(체언_주어), VP_MOD(용언_관형어)

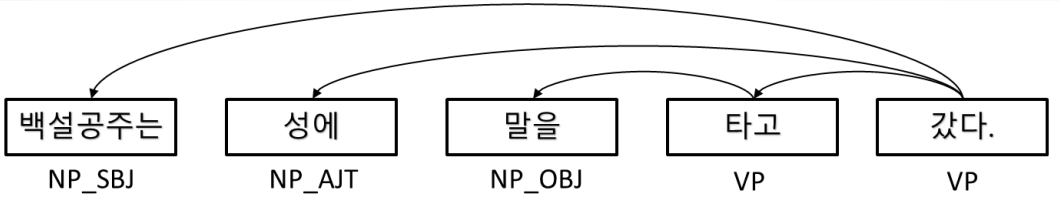


그림1. 의존구문분석 예시

현재 의존구문분석의 연구

- 전이학습을 이용한 그래프기반 딥러닝 시스템
  - 전이학습(Transfer Learning) : 사전학습된 시스템에 추가 출력층을 활용하여 응용시스템 구현
  - 그래프 기반 시스템 : 각 노드 간의 에지(Edge)를 Score 형태로 출력, Score가 가장 높은 것을 최종 의존관계로 설정

기존 시스템의 한계 및 부산대 구문분석기

기존 시스템의 한계

- 데이터셋에 의존적 : 과적합(Overfitting) 문제 발생

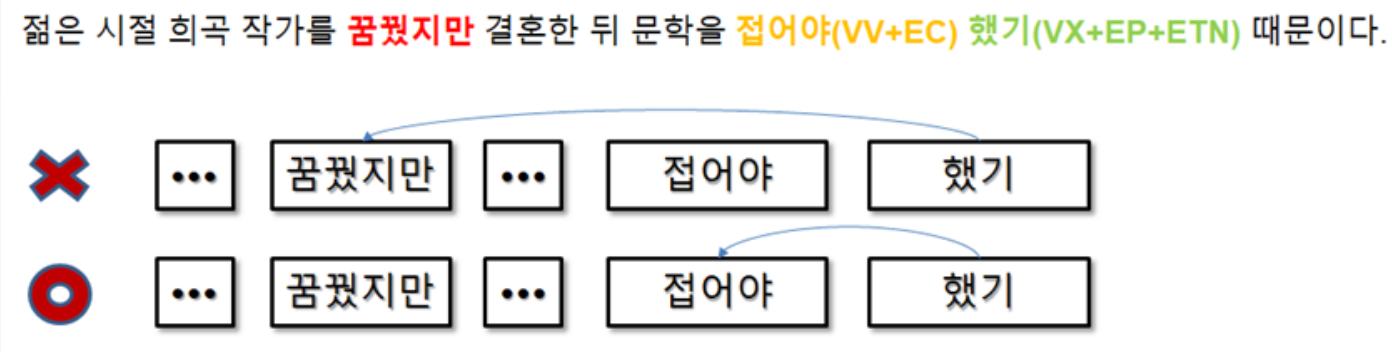


그림2. 과적합현상 예시

부산대학교 AILAB 심층학습 및 규칙 결합 구문분석기

최종 의존관계 Score(Attention score)를 0(위반규칙)과 100(정답규칙)으로 조정

- 위반규칙 : Edge[꿈꿨지만 – 했기] Score = 0
- 정답규칙 : Edge[접어야 – 했기] Score = 100

- 규칙을 통한 데이터셋 의존성(과적합) 해결
- 현재 한국어 의존구문분석의 SOTA(UAS 96.28, LAS 93.19%)

형태소 분석기 문제&대안

형태소 분석기 문제

Mecab 형태소 분석기

- 형태분석 성능 93% : 구문 분석 에러 전파 (UAS 96.28 -> 95.06%, 성능 1.22% 감소)

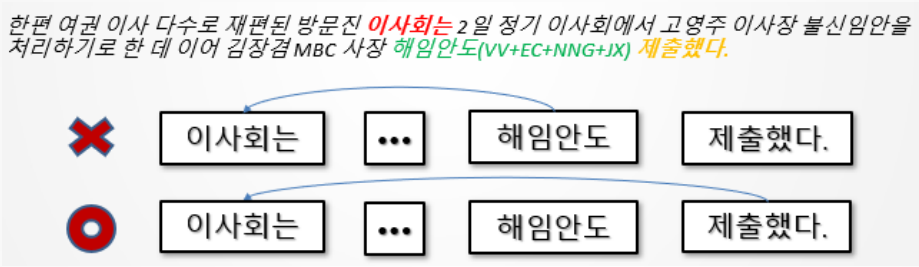


그림3. 형태소 에러-구문분석 에러 예시

Ex) 동사(VV) -> 자동사, 타동사 분류 X : 목적어와 보어 연결여부 알지 못함

형태소 분석기 대안

KLTagger 형태소 분석기

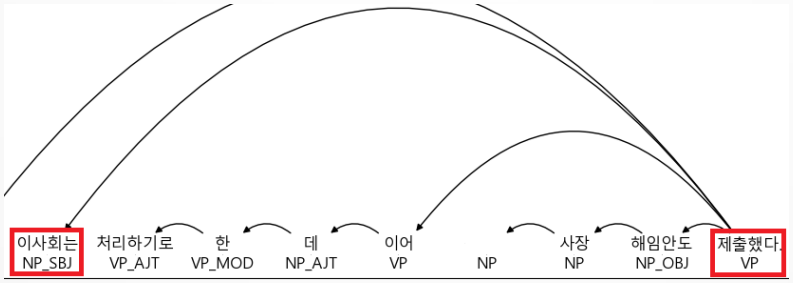


그림4. 형태소 구문분석 KLTagger 적용

⇒ **TTAS표준 46개 -> KLTagger 69개(동사, 일반명사, 의존명사 등 세부분류)**

제안 시스템 구조 – 구문분석 모델 전체 구조

구문분석 모델 구조

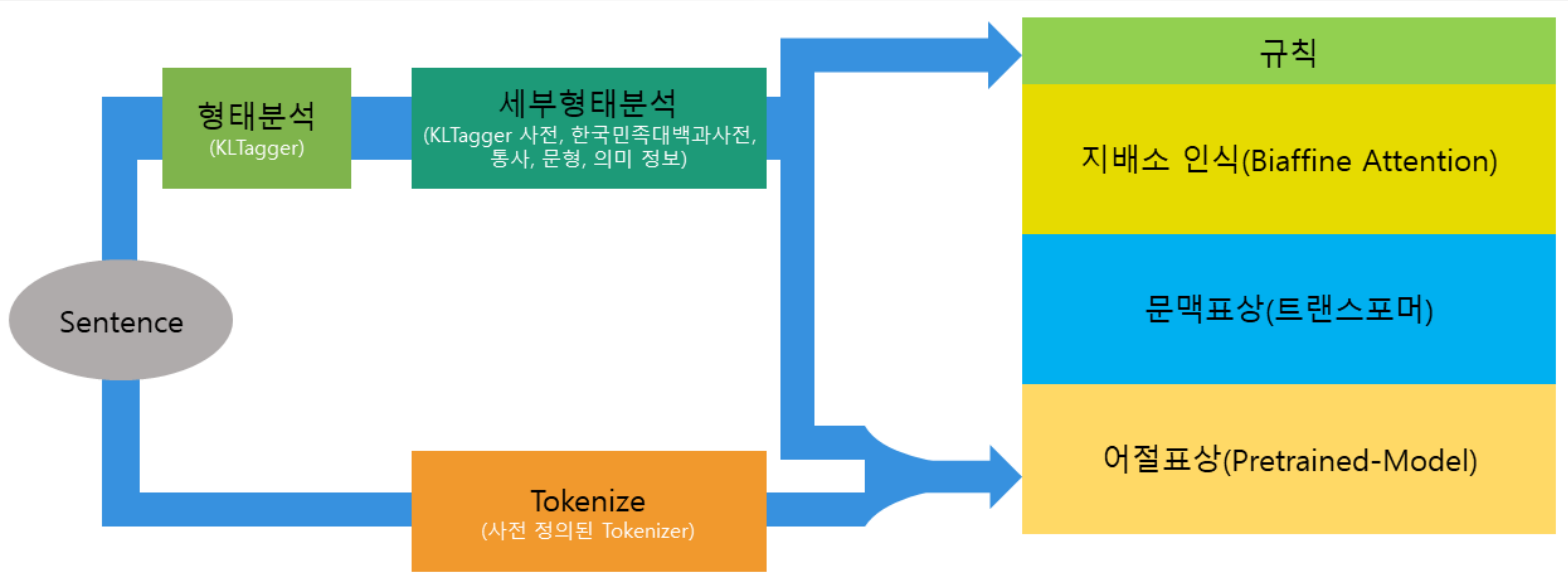


그림5. 구문분석모델 전체 구조도

Biaffine

$$f(Q,K) = QU^{T}K + W^{T}(Q+K) + b$$

Q = 트랜스포머 디코더 출력값

K, V = 트랜스포머 인코더의 출력값

실험 결과

실험 결과

- KLUE 오픈소스(Baseline)과 추가 모델링(형태소,트랜스포머), 규칙적용 모델 비교
- 데이터셋 : KLUE-DP(Train 10000, Dev & Test 2000)

평가메트릭 UAS/LAS macro 사용

Macro : 클래스별 정확도의 평균, KLUE DP의 평가기준

구분	학습데이터	평가데이터	Metric	UAS	LAS
(1) Baseline	KLUE-DP	KLUE-DP	Macro	93.48	87.82
(2) 추가형태소임베딩+트랜스포머	KLUE-DP	KLUE-DP	Macro	93.73	88.03
(3) KLTagger	KLUE-DP	KLUE-DP	Macro	**94.30**	**87.94**
(4) KLUE 형태소 규칙	KLUE-DP	KLUE-DP	Macro	94.22	88.39
(5) KLTagger 규칙	KLUE-DP	KLUE-DP	Macro	**94.88**	**88.48**

표1. Macro 매트릭 비교 결과

실험 결과 분석

- 한국어 형태소 기반 KLTagger 사용시 UAS 0.57% 향상(모델 2와 3 비교)
  -> 한국어 형태소 세부 분석에 Parser의 성능이 크게 향상

- 규칙 사용시 KLUE 형태소 0.74%(모델 1과 4 비교), KLTagger 1.4% 향상(모델 1과 5 비교)
  -> 형태소 분석에 규칙이 효과적

실험 결과

실험 결과

- 건국대학교모델과 추가 KLTagger 규칙적용 모델 비교
- 데이터셋 : KLUE-DP(Train 10000, Dev & Test 2000)

평가메트릭 UAS/LAS micro 사용

Micro : 전체 정확도, 국립국어원 DP의 평가기준

구분	학습데이터	평가데이터	Metric	UAS	LAS
(1) 건국대학교모델	KLUE-DP	KLUE-DP	Micro	95.14	92.68
{2}KLUE 규칙	KLUE-DP	KLUE-DP	Micro	95.36	**93.66**
(3) KLTagger 규칙	KLUE-DP	KLUE-DP	Micro	**95.51**	93.49

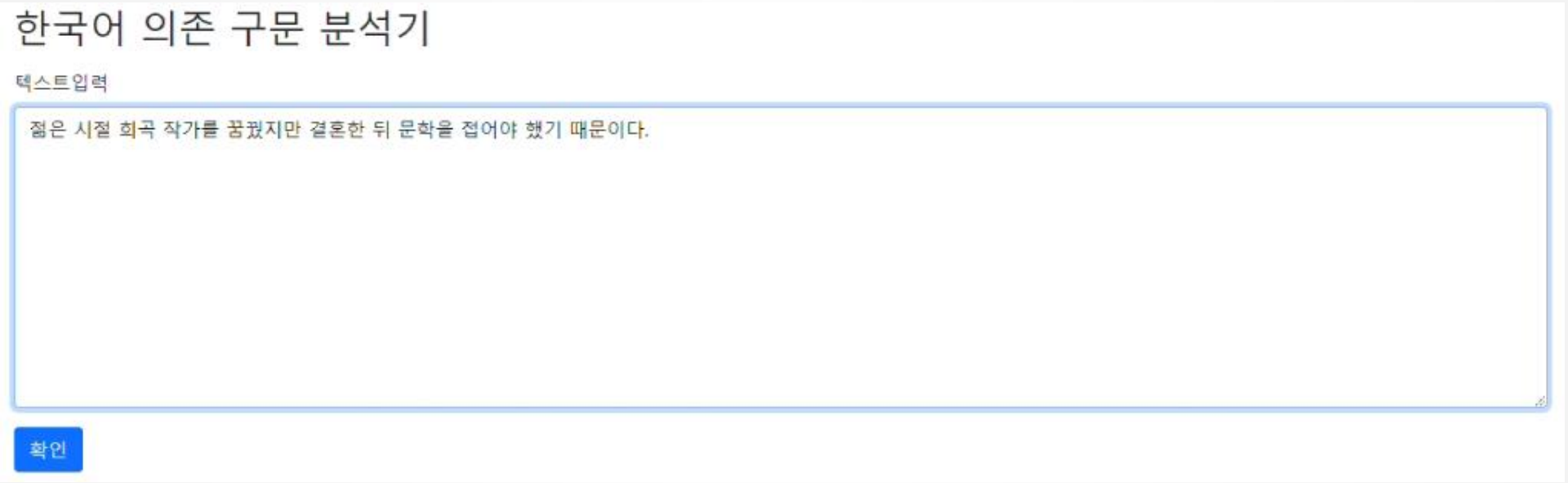
표2. Micro 매트릭 비교 결과

실험 결과 분석

- 같은 형태분석 : UAS 0.22%, LAS 0.98% 높은 성능(모델 1과 2 비교)
- KLTagger사용 : 선행연구보다 UAS 0.37%, LAS 0.81% 높은 성능(모델 1과 3 비교)
  -> 형태소 분석에 상관없이 규칙이 효과적

테스트셋 태깅에러 KLUE 형태소-117개, KLTagger 형태소- 133개 평가에서 제외*(상세 예시 appendix참조)

시스템 시연

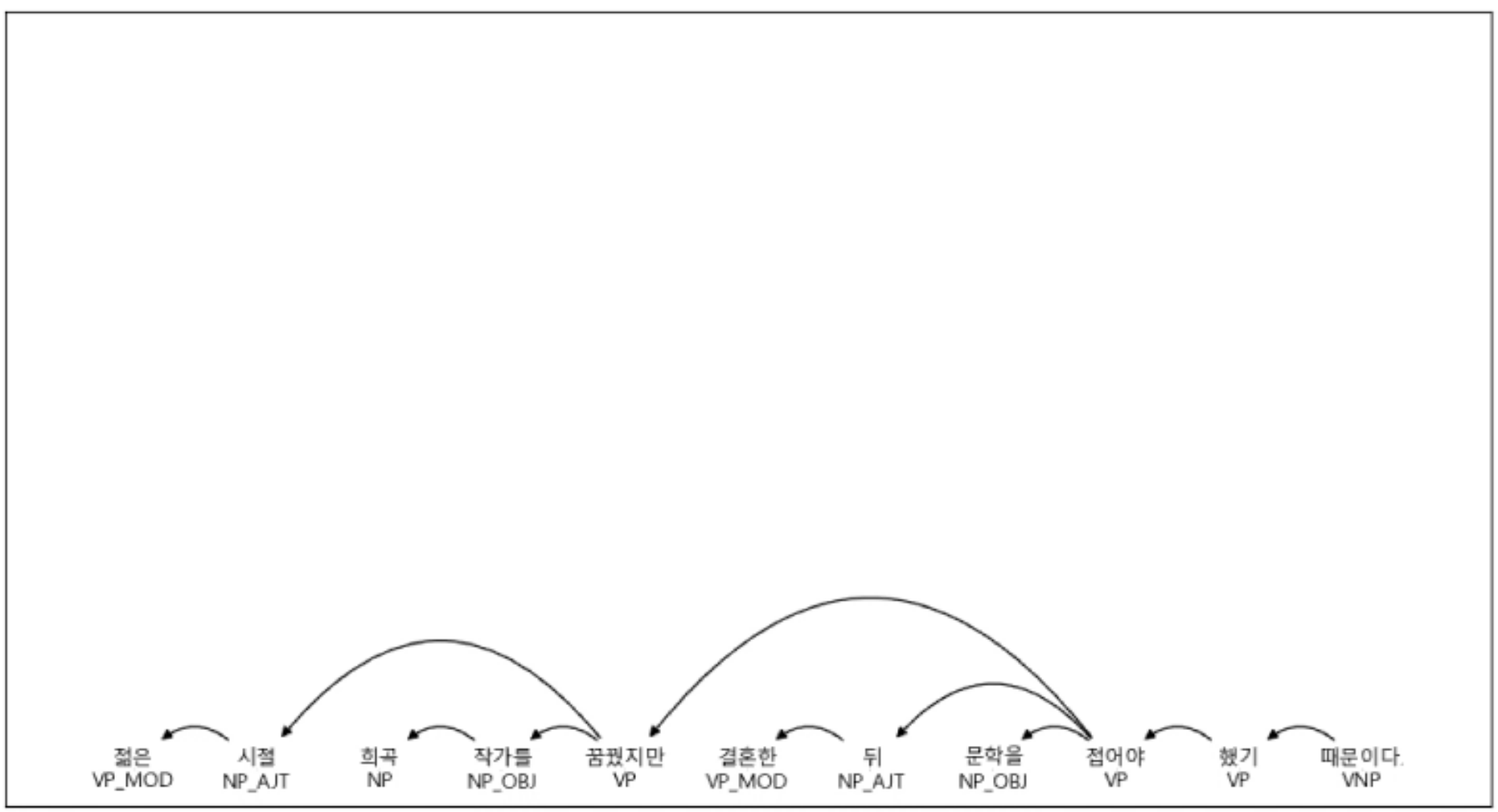


시스템 시연

입력문장 : 젊은 시절 희곡 작가를 꿈꿨지만 결혼한 뒤 문학을 접어야 했기 때문이다.

인덱스	단어형식	지배소인덱스	의존관계레이블
1	젊은	2	VP_MOD
2	시절	5	NP_AJT
3	희곡	4	NP
4	작가를	5	NP_OBJ
5	꿈꿨지만	9	VP
6	결혼한	7	VP_MOD
7	뒤	9	NP_AJT
8	문학을	9	NP_OBJ
9	접어야	10	VP
10	했기	11	VP
11	때문이다.	0	VNP

시스템 시연


젊은
VP_MOD
시절
NP_AJT
희곡
NP
작가들
NP_OBJ
꿈꿨지만
VP
결혼한
VP_MOD
뒤
NP_AJT
문학을
NP_OBJ
접어야
VP
했기
VP
때문이다.
VNP

한계점 및 향후 연구

모델

- KLUE 추가 에러케이스 분석 및 규칙작성
- 국립국어원 모두의 말뭉치 또한 KLTagger로 분석 후 실험 예정

모듈

- KLTagger결과를 cmd로 txt파일로 추출하는 과정에서 특수문자 및 알파벳이 유니코드로 변환 저장되어 텍스트 저장 불가

  -> binary값으로 비교

  • ("`","”")("·", "・")("~","~")("<","《")("十","+")

- Javascript 텍스트(http 및 www 포함된 문장)가 '<ELSS>'로 변경됨

  -> '<ELSS>'를 http등으로 치환 필요

참고문헌

[1] Michael A. Covington, "A dependency parser for variable-word-order languages," Research Reprot AI-1990-01, University of Georgia, 1990.

[2] J. Nivre, "An efficient algorithm for projective dependency parsing," Proc. Of IWPT, pp. 149-160, 2003.

[3] R. McDonald, K. Crammar, F. Pereira, "Online Large-margin Training of Dependency Parsers," Proc. Of ACL, pp. 91-98, 2005.

[4] Y.-H. Lee, J.-H. Lee, "Korean Parsing using Machine Learning Techniques," KIISE, Vol. 35, No. 1C, pp. 285-288, 2008. (in Korean)

[5] M. Choi, S. Jeong, H. Kim, "Dependency Structure Analysis and Dependency Label Annotation Using CRFs," Journal of KIISE, Vol. 41, No. 4, pp. 302-308, 2014. (in Korean)

[6] C. Lee, J. Kim, J. Kim, "Korean Dependency Parsing using Deep Learning," Proc. KIISE for HCLT, pp. 87-91, 2014. (in Korean)

[7] S.-H. Na, K. Kim, Y.-K. Kim, "Stack LSTMs for Transition-Based Korean Dependency Parsing," KCC 2016, pp. 732-734, 2016. (in Korean)

[8] S.-Y. Hong, S.-H. Na, J.-H. Shin, Y.-K. Kim, "BERT and ELMo for contextualized word embed□dings in Korean Dependency Parsing," KCC 2019, pp. 491-493, 2019. (in Korean)

[9] M. King, "Natural Language Parsing," pp. 58-87, Academic Press, 1983.

[10] H. Y. KIM, J. H. CHOI, S. J. LEE, "Improved Chart Parsing Algorithm based on Korean Syntactic Rules," KIISE, Vol. 17, No. 1, Apr. 1990. (in Korean)

[11] Y.-G. Hwang, H.-Y. Lee, Y.-S. Lee, "Using Syntactic Unit of Morpheme for Reducing Mor□phological and Syntactic Ambiguity," Journal of KIISE, Vol. 27, No. 7, pp. 784-793, 2000. (in Korean)

[12] M. Kim, S. Kang, J.-H. Lee, "Dependency Parsing by Chunks," KIISE, Vol. 27, No. 1B, pp. 327-329, Apr. 2000. (in Korean)

[13] S. K. Park, C. M. Jeong, J. M. Jo, S. J. Lee, "AnEffective Korean Syntatic Analyzer Using Longest Grouping Method," KIISE, Vol. 22, No. 1, pp. 961-964, Apr. 1995. (in Korean)

[14] H. Lee, "Korean Lexical Disambiguation using Tail-Head Co-occurrence Information," Journal for KIISE(B), Vol. 24, No. 1, pp. 82-89, 1997. (in Korean)

[15] Y.-M. Woo, Y.-I. Song, S.-Y. Park, H.-C. Rim, "Modification Distance Model for Korean Depen□dency Parsing Using Headible Path Context," Journal of KIISE, Vol. 34, No. 2, pp. 140-149, 2007. (in Korean)

[16] M.G. Jang, G.S. Yoon, and H.C kwon, "Korean Parsing System Based on Chart," KCC 1989.10, 571-574. (in Korean)

[17] J.-Ryu, "A rule-based Ambiguity resolution proposal for extensive Korean Parsing," Pusan National University Master's Thesis, 2018. (in Korean)

[18] A. Yoon, S. Hwang, E. Lee, H.-C. Kwon, "Construction of Korean Wordnet KorLex 1.5 ," Journal of KIISE, Vol. 31, No. 1, pp. 92-108, 2009. (in Korean)

[19] S. T. Kim, M. H. Kim, H. C. Kwon "Rules-based Korean Dependency Parsing Using Sentence Pattern Information,"Journal of KIISE, Vol. 47, No. 5, pp. 488-495, 2020.

[20] C. E. Park, et al., "Korean Dependency Parsing with Multi-layer Pointer Networks," Proc. of the 29th Annual Conference on Human & Cognitive Language Technology, 2017.

[21] S. H. Na, et al., "Deep Biaffine Attention for Korean Dependency Parsing," Proc. of the KIISE Korea Computer Congress 2017, pp. 584-586, 2017. (in Korean)

[22] J.-H. Lim and H. Kim, "Korean Dependency Parsing using the Self-Attention Head Recognition Model," Journal of KIISE, Vol. 46, No. 1, pp. 22-30, 2019.

[23] C. Park, C. Lee, J.-H. Lim, and H.-k. Kim, "Korean Dependency Parsing with BERT," Proc. of the KIISE Korea Computer Congress (KCC) 2019, pp. 530-532, 2019. (in Korean)

[24] J. H. Han, Y. J. Park, Y. H. Jeong, I. K. Lee, J. W. Han, S. J. Park, J. A. Kim, and J. Y. Seo, "Korean Dependency Parsing Using Sequential Parsing Method Based on Pointer Network," Proc. of the 31th Annual Conference on Human & Cognitive Lan□guage Technology, pp. 533-536, 2019. (in Korean)

[25] J.-H. Lim and H. Kim, "Korean Dependency Parsing using Token-Level Contextual Representation in Pre-trained Language Model," Journal of KIISE, Vol. 48, No. 1, pp. 27-34, 2021.

[26] J. H. Lim, Y. J. Bae, H. K. Kim, Y. J. Kim, and K.C. Lee, "Korean Dependency Guidelines for Depen□dency Parsing and Exo-Brain Language Analysis Corpus," Proc. of the 27th Annual Conference on Human & Cognitive Language Technology, pp. 234-239, 2015. (in Korean)

[27] 국립국어원, "구문 및 무형 대용어 복원 말뭉치 연구 분석", 2021

[28] Gawlikowski, J., Tassi, C.R., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A.M., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., & Zhu, X. (2021). A Survey of Uncertainty in Deep Neural Networks. ArXiv, abs/2107.03342.

[29] J. M. Shin, S. H. Cho, S. R, Park "Neural network-based dependency parsing with rules applied"한국컴퓨터종합학술대회 논문집, 2022

참고자료 : Appendix

테스트셋 태깅에러 KLUE 형태소-117개, KLTagger 형태소- 133개

지배소	자질	규칙	태깅에러 (KLTagger)	태깅에러 (KLUE)
동사	-동사 공통	:관형절은 용언에 연결되지 못함	27	26
	-문장 주동사	:문장부사는 문장 주동사와 연결	39	23
	-자동사	:목적격은 자동사와 연결되지 못함	2	
	-불완전동사	:바로 앞 어절을 제외하곤 연결되지 못함	45	45
	-동사+동사	:본용언이 연속적으로 나타날 경우, 주어를 앞에 위치한 서술어에 연결		
	-부사형 용언	:'~도록' 과 같은 형태의 부사격으로 쓰이는 용언은 내포문의 주어가 아닌 문장주어가 연결되지 못함		
보조용언	-보조용언	: 문장부사와 바로 앞 어절을 제외하곤 다른 문장성분은 연결되지 못함	7	6
		: 바로 앞 어절과 연결		
인용	-직접인용	: 문장주어는 인용절 끝에 연결되지 못한다		
	-간접인용	: 직접인용과 같은 원칙, 문장주어는 인용절 끝에 연결되지 못한다	5	3
의존명사	-의존명사 공통	:바로 앞에 지시관형사가 올 경우 연결		
		:바로 앞에 관형사가 올 경우 연결		
		:바로 앞에 명사가 올 경우 연결		
		:바로 앞에 명사파생접미사가 올 경우 연결		

지배소	자질	규칙	태깅에러 (KLTagger)	태깅에러 (KLUE)
의존명사	-의존명사 공통	:바로 앞에 지시관형사가 올 경우 연결		
		:바로 앞에 관형사가 올 경우 연결		
		:바로 앞에 명사가 올 경우 연결		
		:바로 앞에 명사파생접미사가 올 경우 연결		
	-일반의존명사	:바로 앞에 대명사가 올 경우 연결		
		:바로 앞에 명사가 올 경우 연결		
	-단위의존명사	:바로 앞에 수 관형사가 올 경우 연결		
		:바로 앞에 수사가 올 경우 연결		
	-의사보조용언	:문장부사와 바로 앞 어절을 제외하고 문장 성분이 연결되지 못함	8	14
		:바로 앞 어절(관형형 ㄴ/ㄹ, 동작성명사)과 연결		
일반명사	일반명사	:연결어미는 목적격 조사를 가지는 명사절에는 연결되지 못함		

감사합니다.