

청크 표현 방법과 의존 구문 분석을 활용한 논문 문장 이중 분류

김슬기(cloudyju11@konkuk.ac.kr), 김홍진(jin3430@gmail.com), 김학수(nlpdrkim@konkuk.ac.kr)
건국대학교 컴퓨터공학과, 건국대학교 인공지능학과

모델을 제안하는 이유

논문 문장 분류

- 연구 동향을 파악하기 위해 논문 내 주요 문장을 찾아 이를 자동으로 분류하기 위한 작업

논문 문장 이중 분류

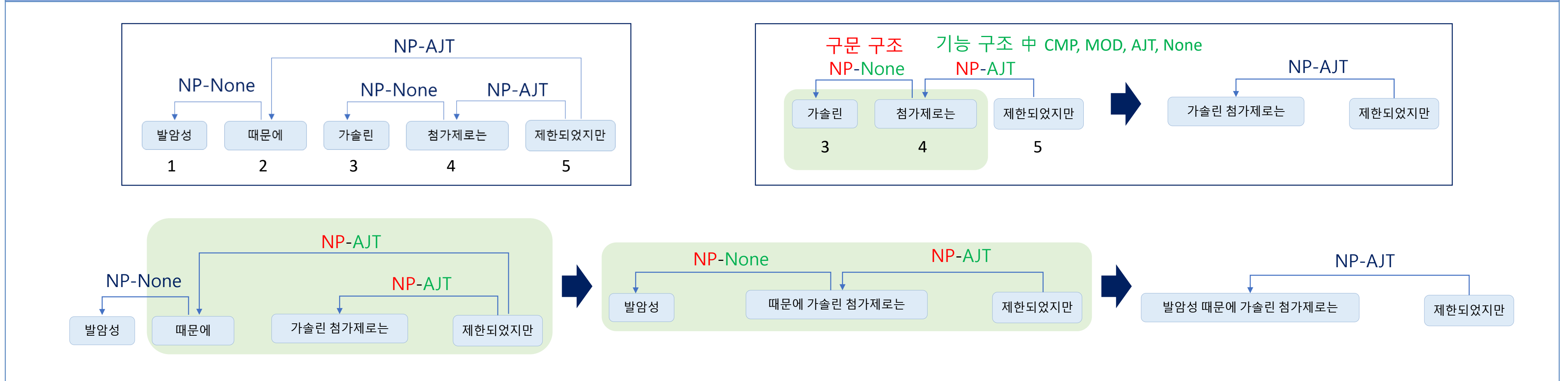
- 각 분류 태그별 데이터 개수 불균형 문제를 완화하기 위해 기존의 분류 태그를 세부 분류 태그로 보고 일정 기준으로 세부 분류 태그를 대분류 태그로 나누어 대분류와 세부분류를 하는 작업

청크 표현 방법

- 기존 문장에서 어절 단위로 만들어진 의존 구문 구조를 의미적 기준으로 만들어진 청크 단위 의존 구문 구조를 만들기 위해 사용하는 방법

대분류	세부 분류	데이터 개수
연구 목적	문제 정의	14,898
	가설 설정	3,413
	기술 정의	12,358
연구 방법	제안 방법	24,358
	대상 데이터	20,249
	데이터처리	14,239
연구 결과	이론/모형	11,190
	성능/효과	37,053
	후속 연구	17,858

청크 표현 방법



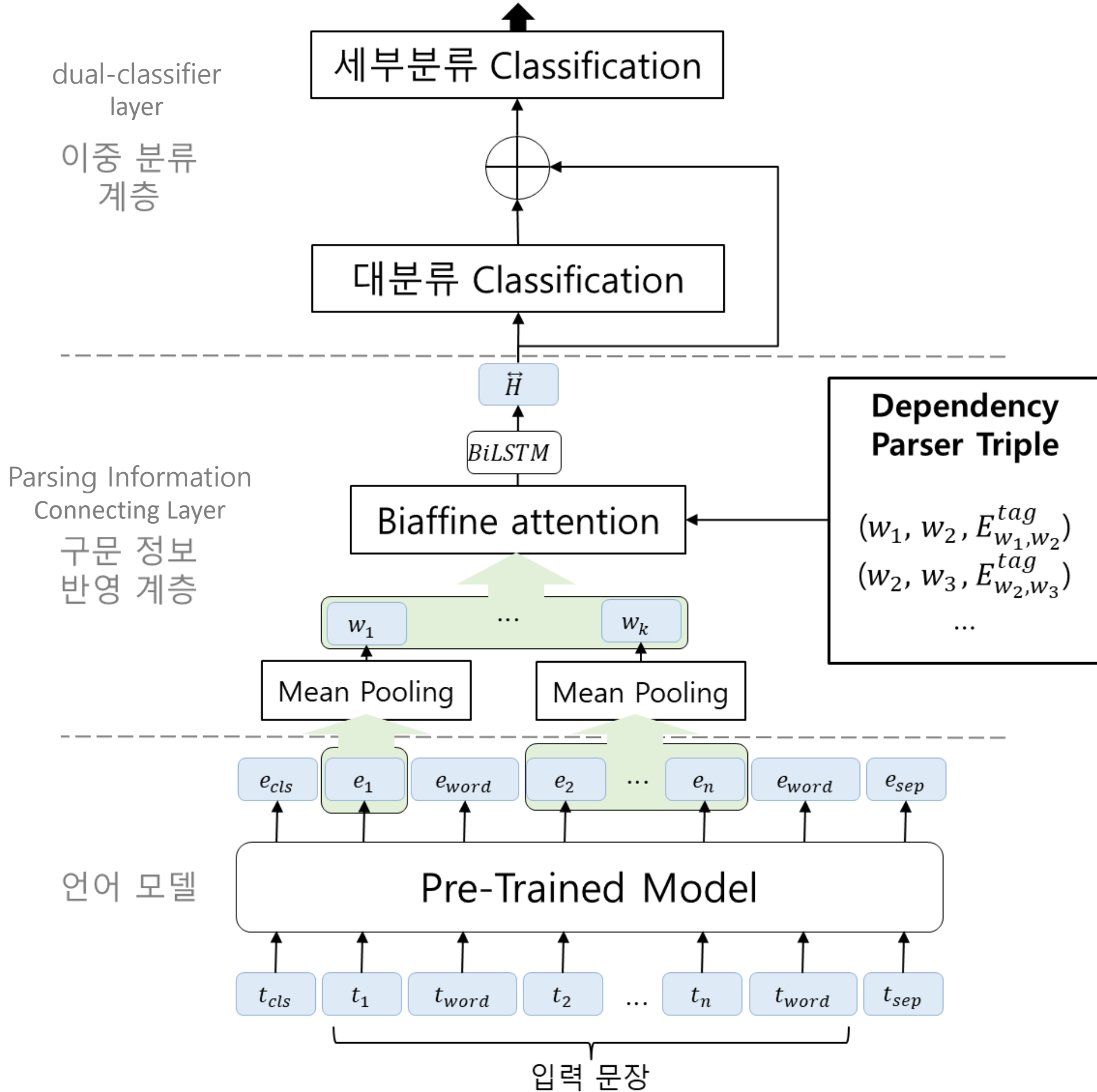
문제 정의 or 가설 설정 or 기술 정의 or 제안 방법 or 대상 데이터 or 데이터처리 or 이론/모형 or 성능/효과 or 후속연구

제안 모델에 대한 성능 평가

Model		Accuracy	F1-Score
KLUE/ BERT- base	Baseline (1)	89.66	89.90
	Chunk+Biaffine (2)	89.66	89.91(+0.01)
	Chunk+Biaffine+dual-classifier (3)	89.75(+0.09)	89.99(+0.09)

세부분류 태그별 성능 평가

세부 분류	Accuracy		
	(1)	(2)	(3)
문제 정의	91.80	90.30 (-1.50)	90.03 (-1.77)
가설 설정	94.72	95.60 (+0.88)	95.30 (+0.58)
기술 정의	97.72	97.72 (+0.00)	97.36 (-0.36)
제안 방법	75.50	74.83 (-0.67)	76.88 (+1.38)
대상 데이터	88.72	89.31 (+0.59)	88.96 (+0.24)
데이터처리	88.76	87.42 (-1.34)	88.58 (-0.18)
이론/모형	81.57	84.45 (+2.88)	82.78 (+1.21)
성능/효과	95.15	95.32 (+0.17)	95.25 (+0.10)
후속 연구	94.80	94.48 (-0.32)	95.41 (+0.61)



대분류를 “연구 목적”으로 예측했을 때
세부분류의 정답과 예측
(acc: 92.35%p)

문제 정의	2750	1	0
가설 설정	3	653	6
기술 정의	3	0	2436
제안 방법	65	2	0
대상 데이터	1	0	1
데이터처리	0	0	0
이론/모형	0	0	1
성능/효과	0	0	1
후속연구	0	0	2

대분류를 “연구 방법”으로 예측했을 때
세부분류의 정답과 예측
(acc: 95.94%p)

문제 정의	2	0	0	0
가설 설정	13	6	1	6
기술 정의	4	1	1	4
제안 방법	3469	258	286	367
대상 데이터	338	3659	14	43
데이터처리	118	11	2444	75
이론/모형	131	28	77	1688
성능/효과	6	4	1	2
후속연구	0	0	0	0

대분류를 “연구 결과”으로 예측했을 때
세부분류의 정답과 예측
(acc: 98.18%p)

문제 정의	2	2
가설 설정	1	0
기술 정의	2	0
제안 방법	15	1
대상 데이터	0	0
데이터처리	0	0
이론/모형	0	0
성능/효과	7070	261
후속연구	120	3282