계층적 다중 레이블 임베딩을 이용한 논문 문장 수사학적 분류

인공지능 부문 – 국내 논문 문장 의미 태깅

주제 : 문장 태깅 모델

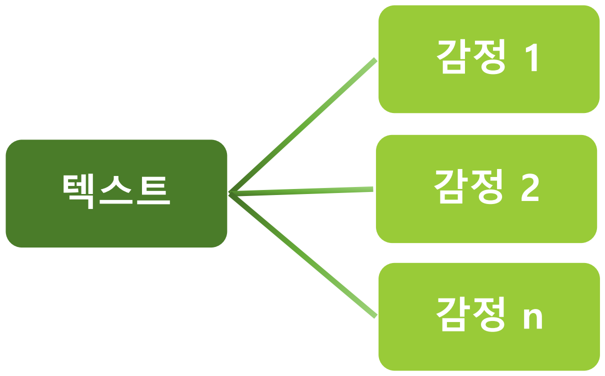
김슬기, 김담린, 이주상, 이우진

목차

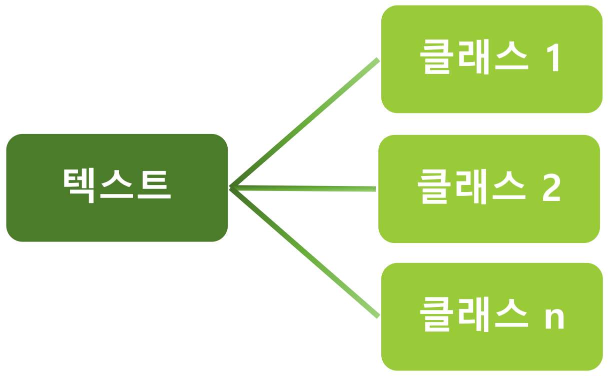
- 개요
  - 텍스트 분류
  - 대분류
  - 세부분류
- 모델
- 실험 및 결과
- 결론

텍스트 분류란 ?

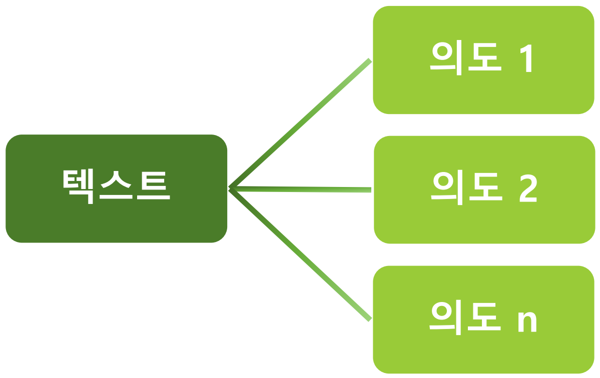
- 텍스트 분류
  – 단어, 문장 또는 지문 전체를 n개의 범주 중 하나로 분류하는 작업
  – 예시



< 감성 분석 >



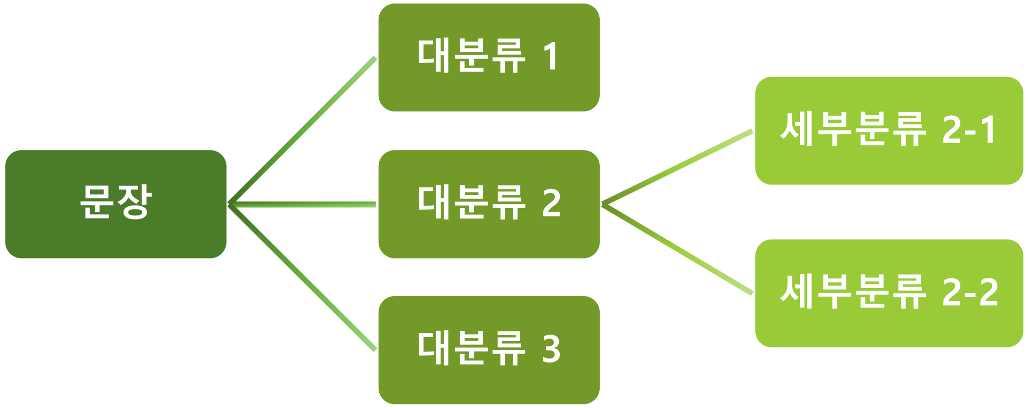
<클래스 분류>



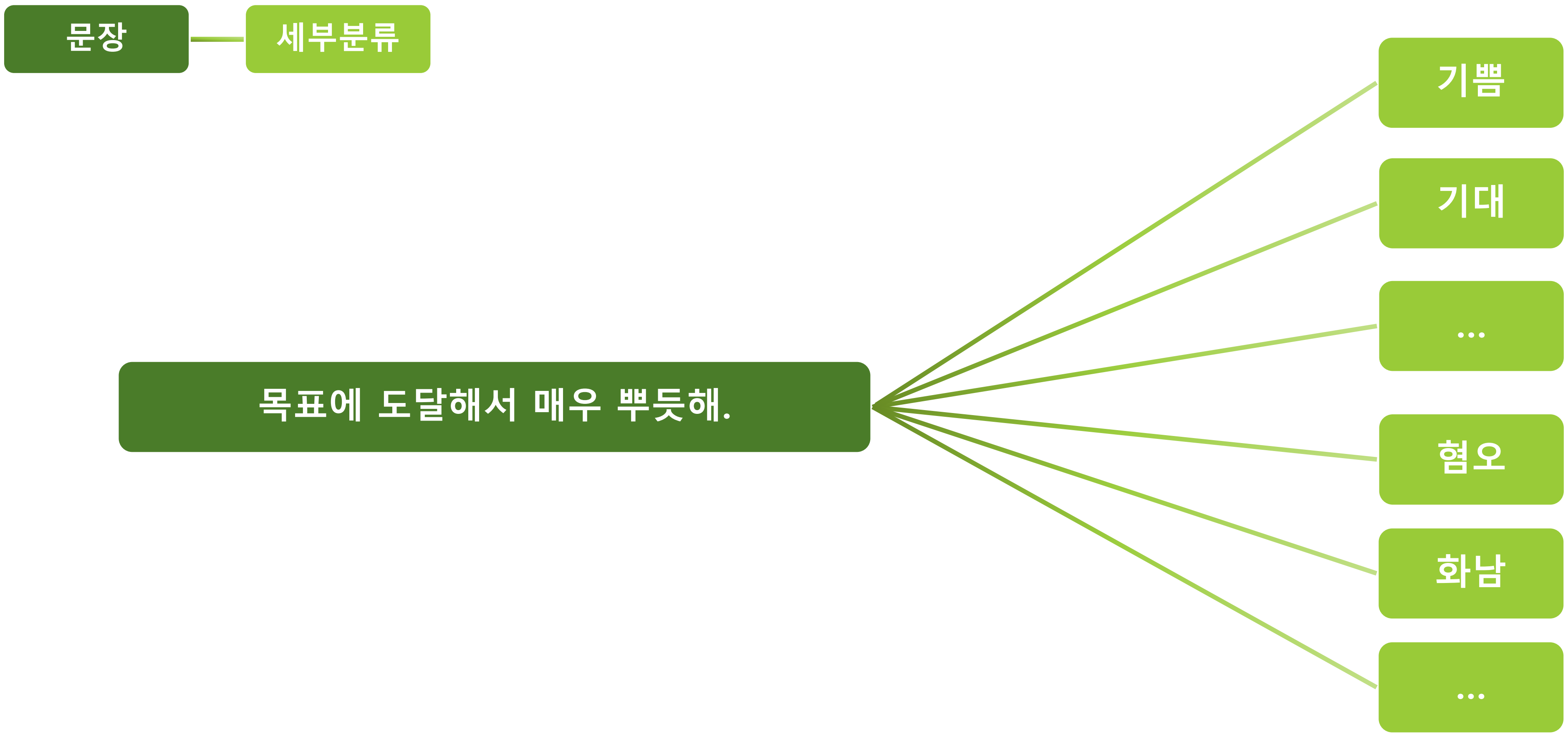
< 의도 분류 >

텍스트 분류 - 대분류 , 세부분류

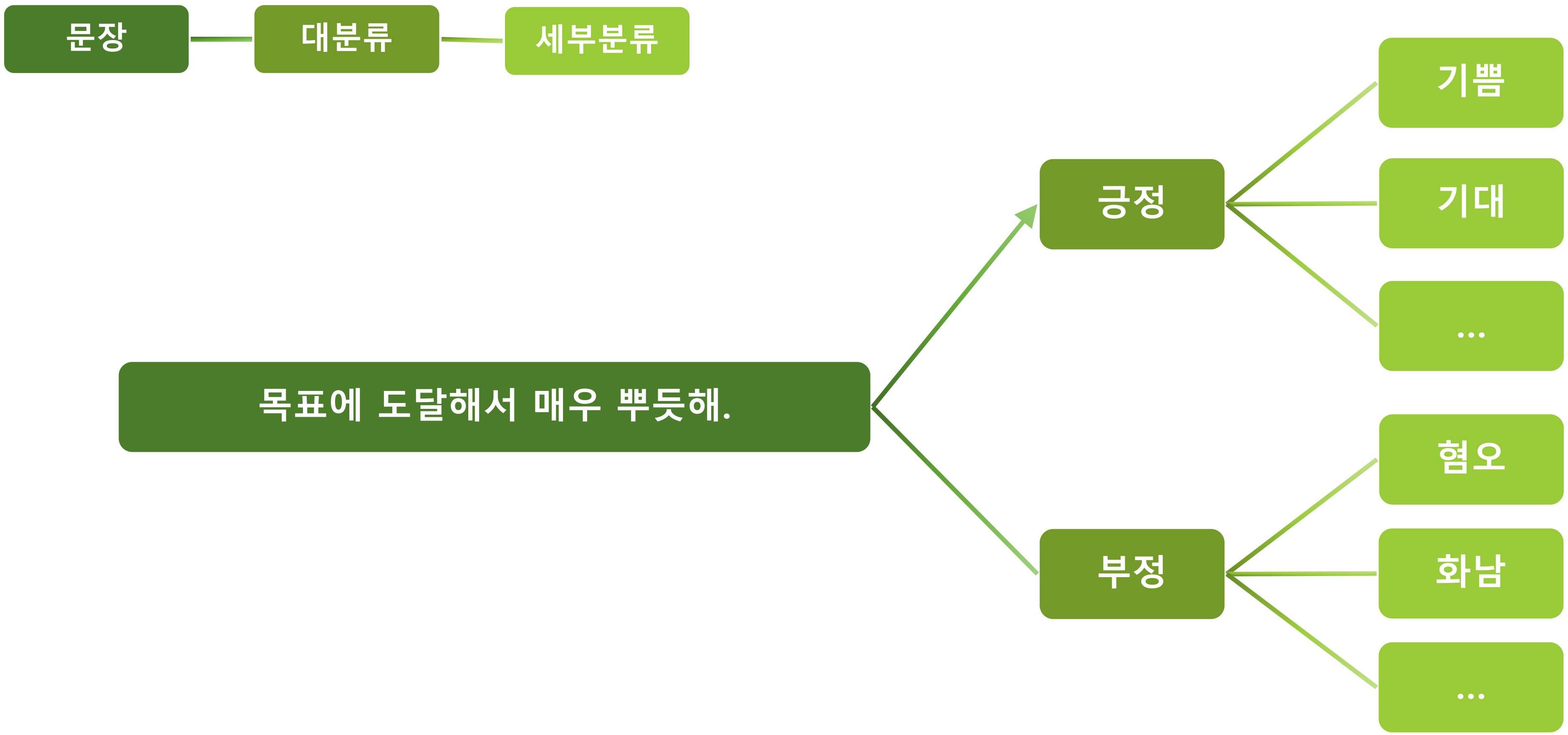
- 대분류 (Coarse-grained)
  - "크게 나누어 가름"
  - 세부분류들을 포함하는 범주형 데이터
- 세부분류 (Fine-grained)
  - "작게 나누어 가름"
  - 대분류에 속하는 범주형 데이터



대분류 , 세부분류 – 감정분석 적용 예시

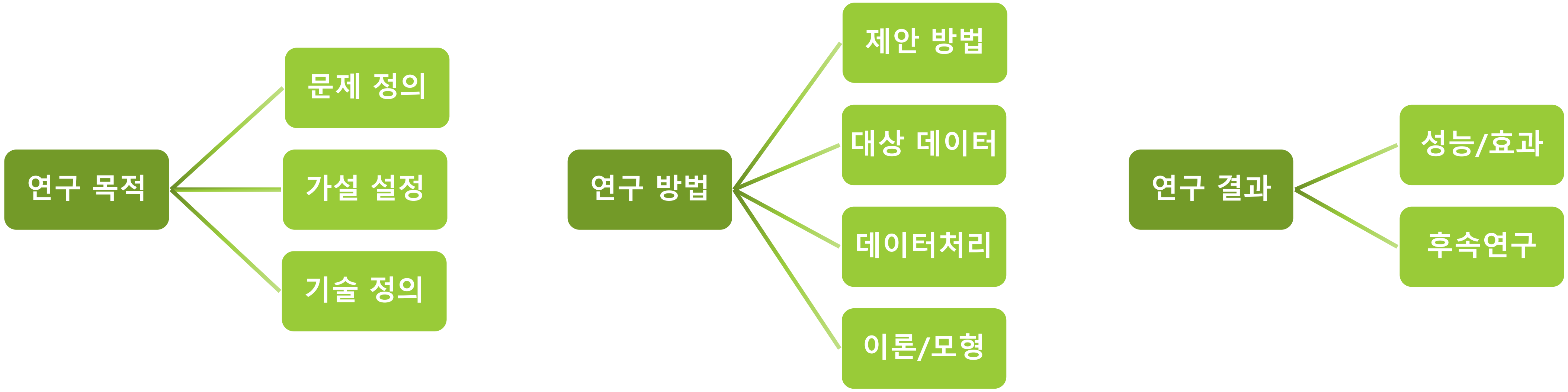


대분류 , 세부분류 – 감정분석 적용 예시



논문 문장 수사학적 분류 – 대분류, 세부분류

- 대분류
  – 의미 구조 분류
- 세부분류
  – 문장 의미 세부 분류



제안 모델

- 국내 논문 문장 의미 태깅
  - 계층적 다중 레이블 임베딩을 이용한 대분류, 세부분류 태그 예측

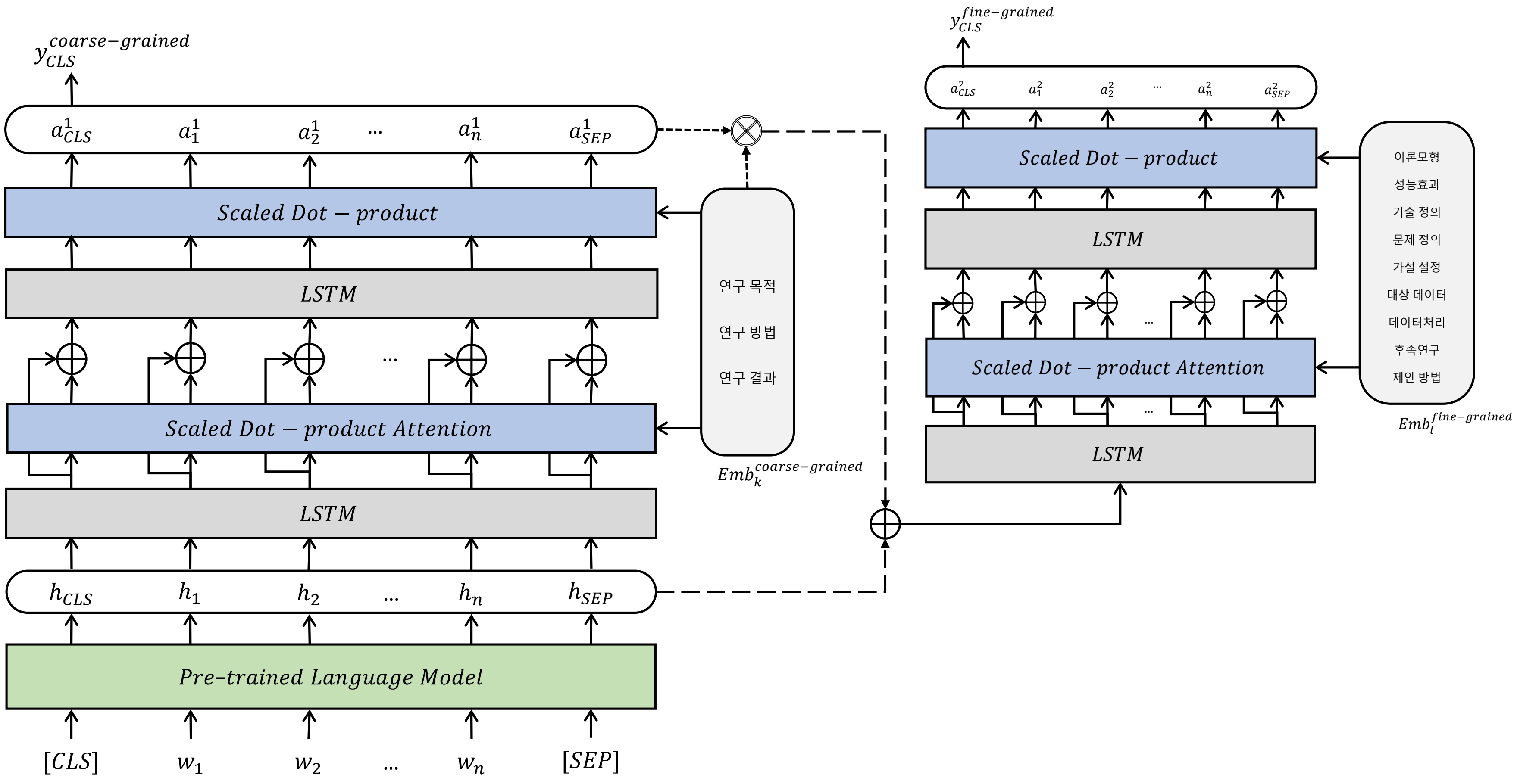
  - 대분류 태그에 비해 세부분류 태그의 학습데이터 개수 차이가 큼

    → 대분류 태그 임베딩 정보를 세부분류 태그 학습에 이용하여 불균형 문제 해결 가능

  - 대분류 태그 임베딩도 같이 학습

    → 세부분류 태그 뿐만 아니라 대분류 태그도 예측 가능

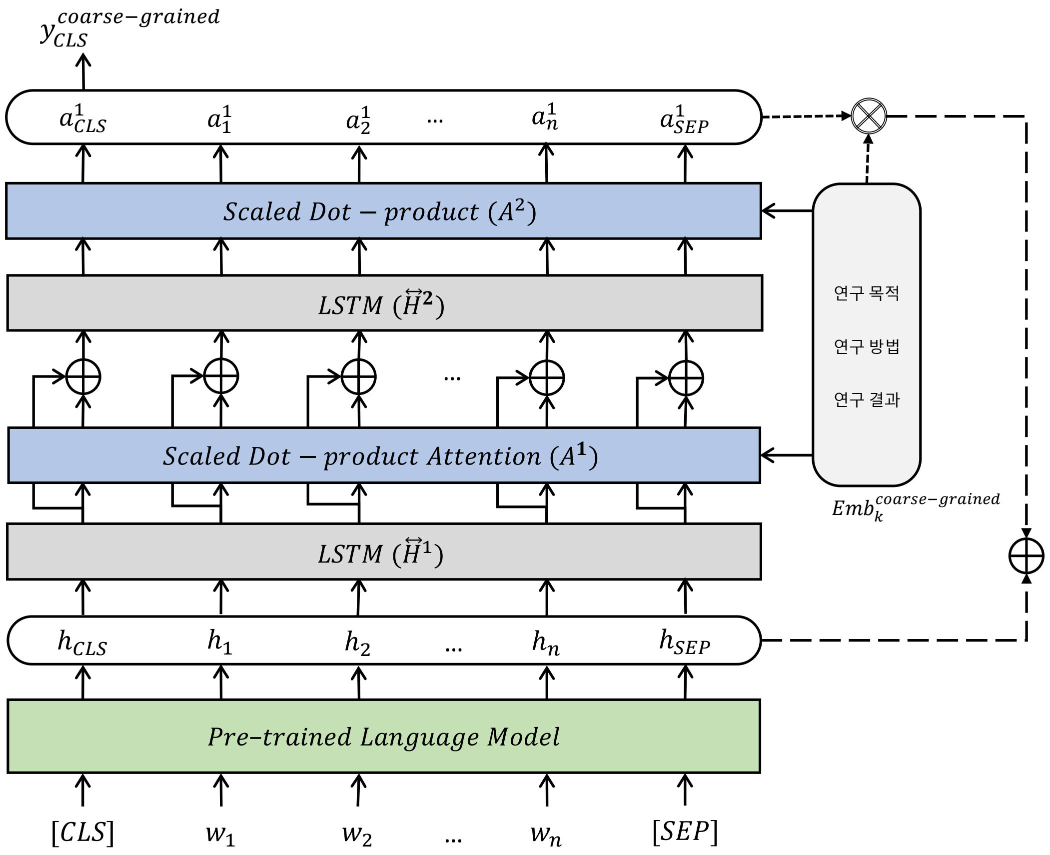
전체 모델 구조도



$y_{CLS}^{coarse-grained}$
a_{CLS}^1 a_1^1 a_2^1 ... a_n^1 a_{SEP}^1
Scaled Dot − product
LSTM
Scaled Dot − product Attention
LSTM
h_{CLS} h_1 h_2 ... h_n h_{SEP}
Pre-trained Language Model
[CLS] w_1 w_2 ... w_n [SEP]
연구 목적
연구 방법
연구 결과
$Emb_k^{coarse-grained}$
$y_{CLS}^{fine-grained}$
a_{CLS}^2 a_1^2 a_2^2 ... a_n^2 a_{SEP}^2
Scaled Dot − product
LSTM
Scaled Dot − product Attention
LSTM
이론모형
성능효과
기술 정의
문제 정의
가설 설정
대상 데이터
데이터처리
후속연구
제안 방법
$Emb_l^{fine-grained}$

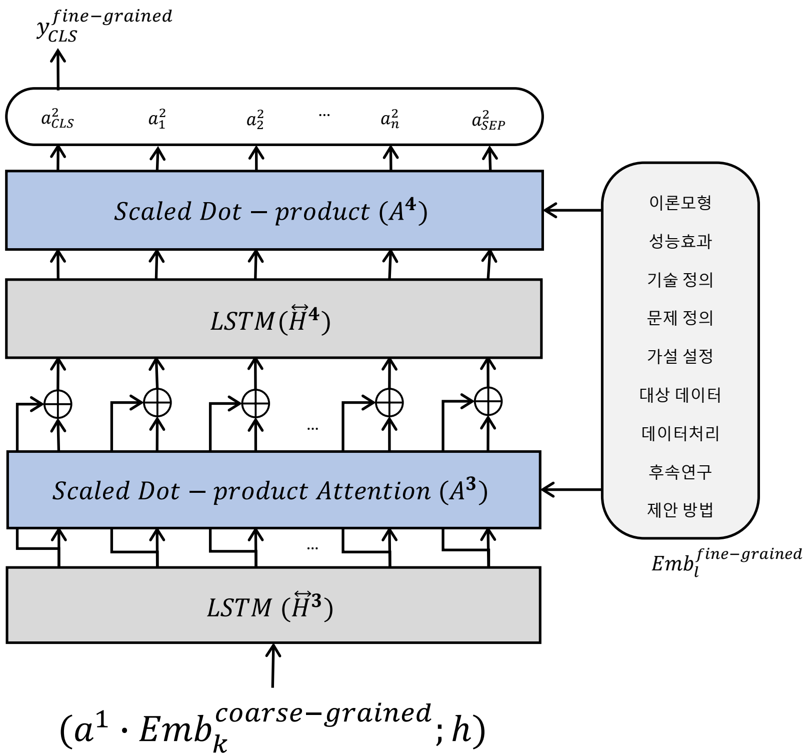
대분류 태그 예측

- $Scaled\ Dot - product\ Attention$
  - $A^1(C_i) = softmax\left(\frac{Q*K^T}{\sqrt{d_h}}\right) \cdot V$
  - $Q = \overleftrightarrow{H}^1, \quad K = V = Emb_k^{coarse-grained}$
- $Scaled\ Dot - product$
  - $A^2(C_i) = softmax\left(\frac{Q*K^T}{\sqrt{d_h}}\right)$
  - $Q = \overleftrightarrow{H}^2, \quad K = Emb_k^{coarse-grained}$



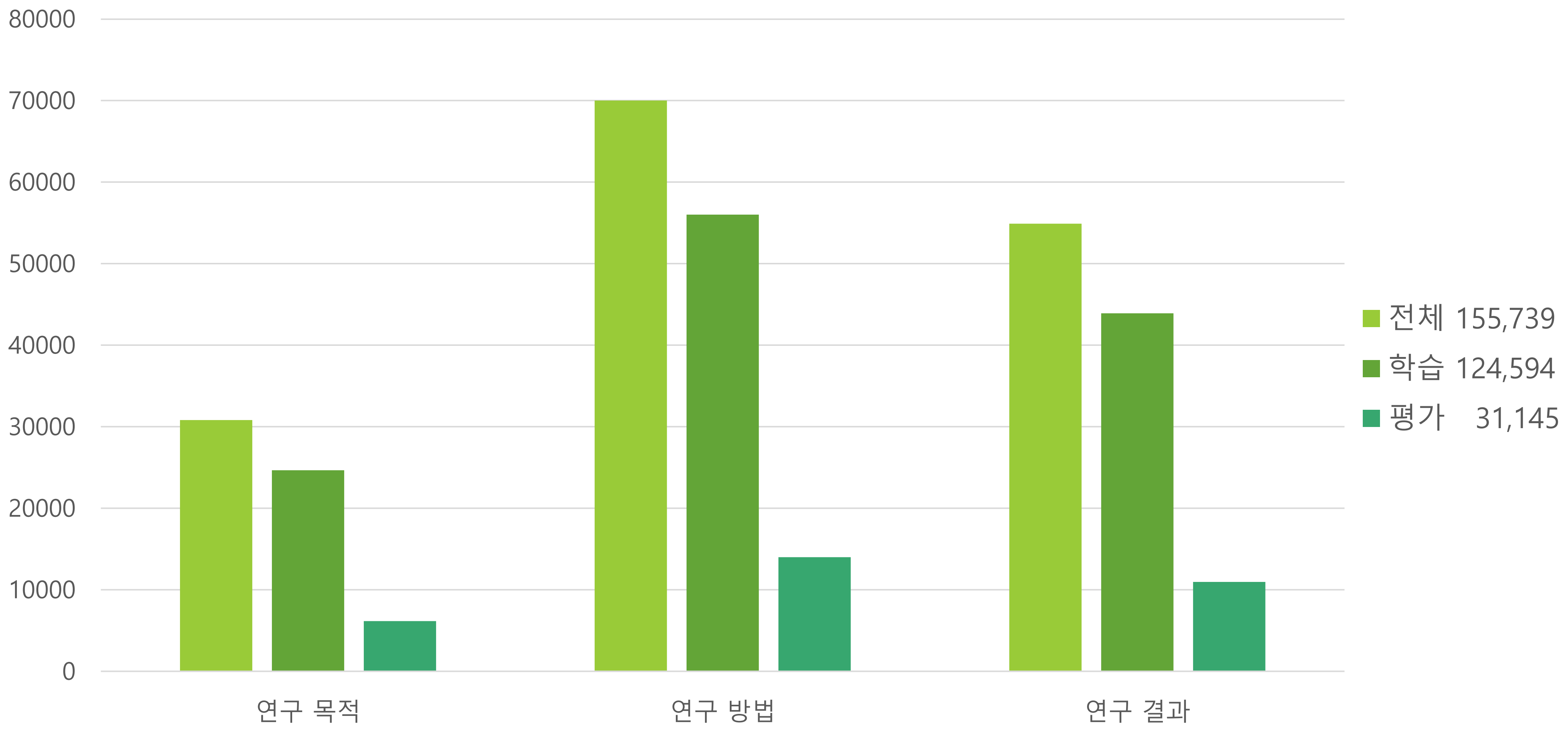
세부분류 태그 예측

- $Scaled\ Dot - product\ Attention$
  - $A^3(C_i) = softmax\left(\frac{Q * K^T}{\sqrt{d_h}}\right) \cdot V$
  - $Q = \overleftrightarrow{H}^3, \quad K = V = Emb_k^{fine-grained}$

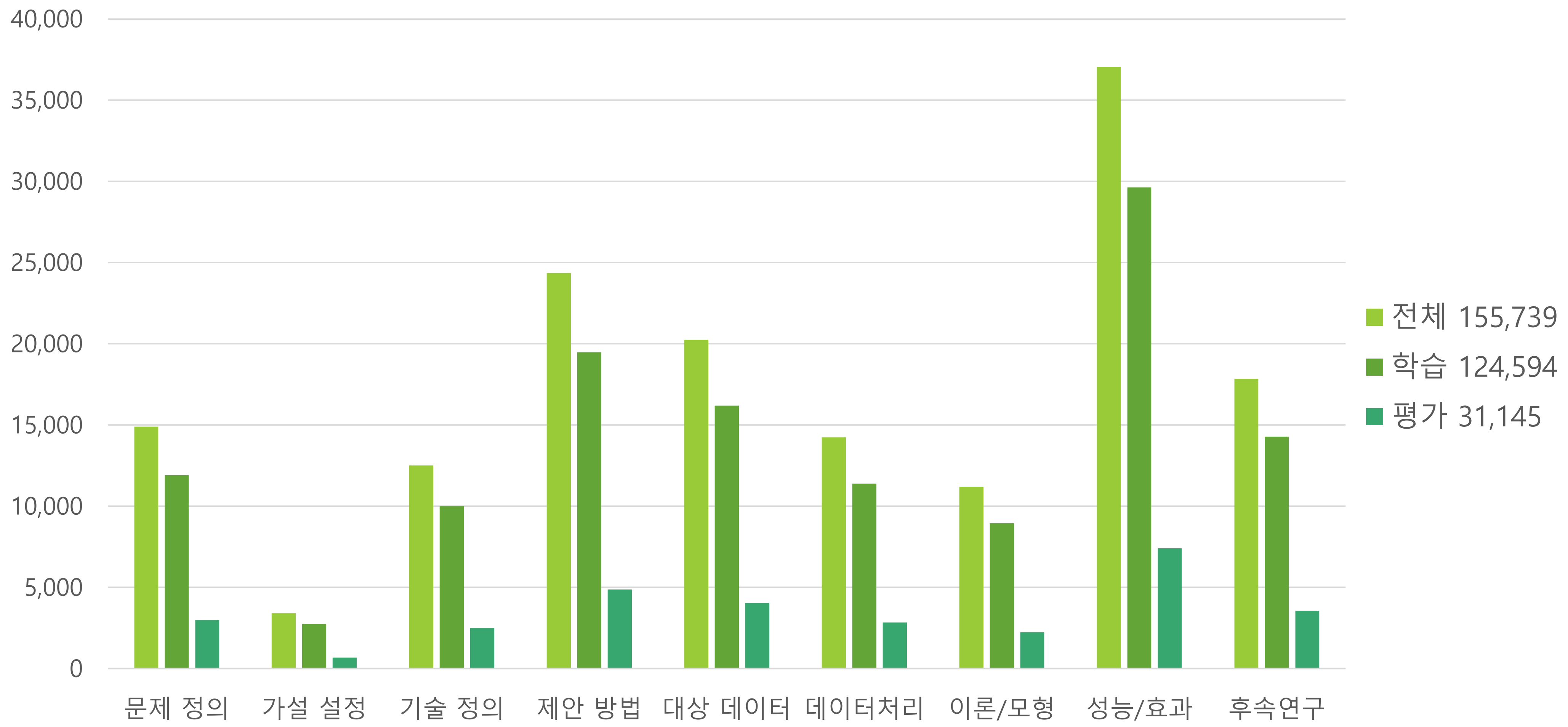


- $Scaled\ Dot - product$
  - $A^4(C_i) = softmax\left(\frac{Q * K^T}{\sqrt{d_h}}\right)$
  - $Q = \overleftrightarrow{H}^4, \quad K = Emb_k^{fine-grained}$

데이터셋 - 대분류


80000
70000
60000
50000
40000
30000
20000
10000
0
연구 목적
연구 방법
연구 결과
전체 155,739
학습 124,594
평가 31,145

데이터셋 – 세부분류


40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
문제 정의
가설 설정
기술 정의
제안 방법
대상 데이터
데이터처리
이론/모형
성능/효과
후속연구
전체 155,739
학습 124,594
평가 31,145

평가 방법

- $Macro\ F1-Score = \frac{1}{n}\sum_{i=1}^{n}\frac{(2\times precision_i\times recall_i)}{precision_i+recall_i}$

- $Micro\ Accuracy = \frac{\sum_{i=1}^{n}TP_i+\sum_{i=1}^{n}FN_i}{\sum_{i=1}^{n}TP_i+\sum_{i=1}^{n}TN_i+\sum_{i=1}^{n}FP_i+\sum_{i=1}^{n}FN_i}$

	Positive	Negative
Positive	**T**rue **P**ositive	**F**alse **P**ositive
Negative	**F**alse **N**egative	**T**rue **N**egative

실험 결과

번호	사전학습 언어모델	모델	대분류		세부분류		대분류+세부분류
			macro F1	micro accuracy	macro F1	micro accuracy	micro accuracy
1	KorSciBERT	LAN (세부분류)	-	-	89.89	89.65	-
2		LAN (대분류, 세부분류)	95.51	**96.85**	89.95	89.76	88.59
3	KLUE BERT-base	LAN (대분류, 세부분류)	95.54	96.43	89.77	89.58	88.55
4	KLUE roBERTa-base	LAN (대분류, 세부분류)	**95.64**	96.77	**90.00**	**89.85**	**88.79**

결론

- 계층적 다중 레이블 임베딩을 이용한 대분류, 세부분류 태그를 예측하여 대분류 95.64%, 세부분류 90%의 F1-score를 보임

- 세부분류 뿐만 아니라 대분류 태그도 한번에 예측함

- 문장 분류 뿐만 아니라 다른 태스크에도 적용하여 향후 연구 예정

QnA

감사합니다