Hierarchy-aware Label Semantics을 활용한 문장 태깅 분류

팀 AIDA
이소연, 한예지, 이병훈, 신우석

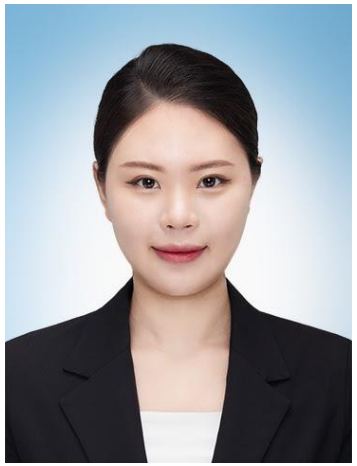
Contents

프로젝트 개요

- 팀원 소개

이소연

고려대학교
산업경영공학과



한예지

고려대학교
산업경영공학과

이병훈

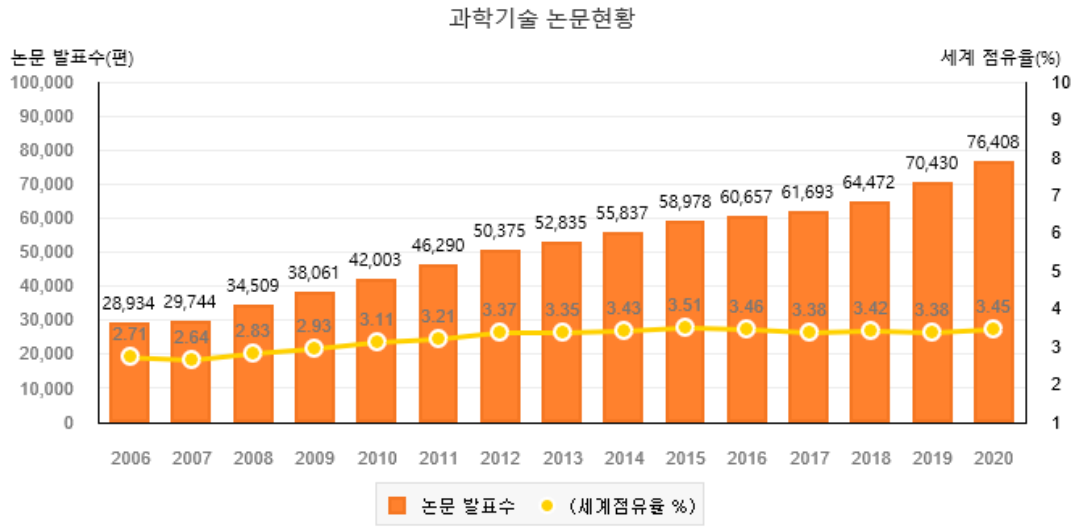
고려대학교
산업경영공학과

신우석

고려대학교
산업경영공학과

프로젝트 개요

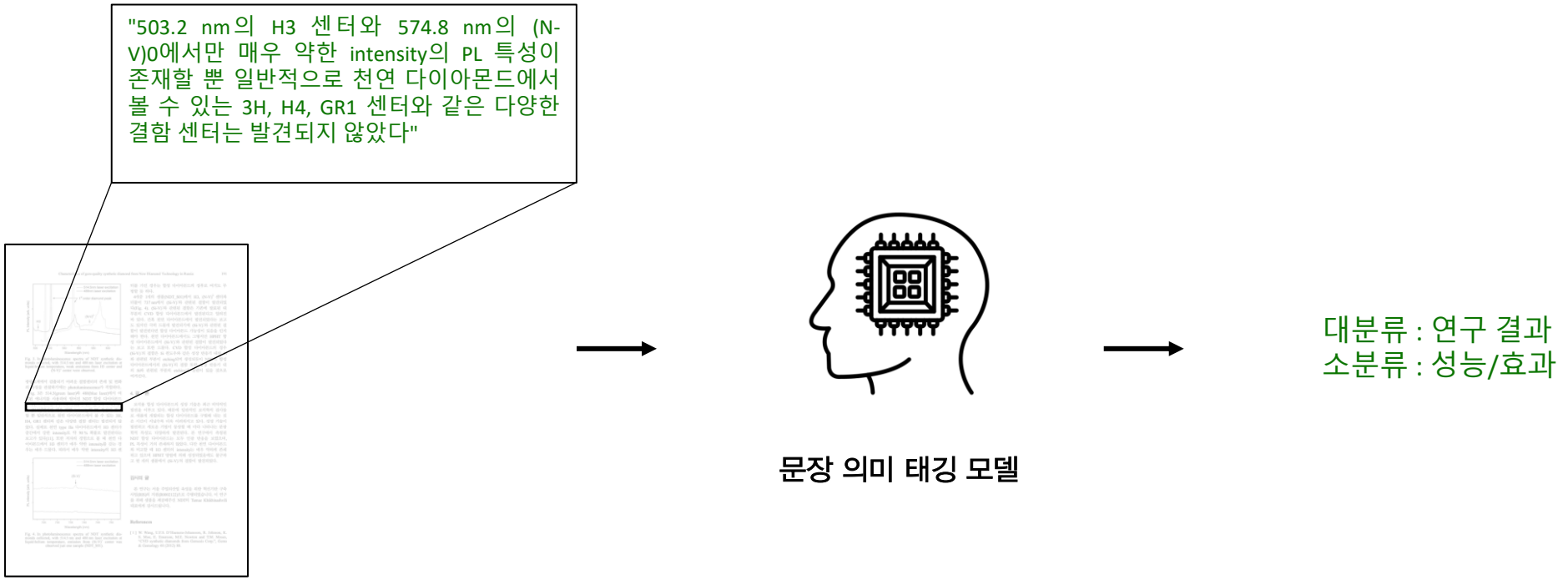
- 문제정의
  - 매년 다양한 연구분야에서 많은 양의 논문이 발표되고 있음.
  - 특정 논문에서 원하는 정보를 추출하기 위해서는 시간과 비용이 많이 듬.
  - 논문의 각 문장이 의미하는 바를 태그로 부착하여 자동화한다면 비용절감 효과를 볼 수 있음.



출처 : 과학기술정보통신부/한국과학기술기획평가원에서 발간한 "과학기술 논문성과 분석연구" 의 각 연도별 보고서

프로젝트 개요

- **개발목표**
  - **논증적 의미 구조 요소**에 따라 논문 문장에 의미 태그를 부착하는 모델을 개발하고자 함.
  - 문장 의미 태깅은 논문의 주요 문장 추출을 가능하게 하고, 트랜드 분석/자동요약 등에 활용될 수 있음.



활용 데이터

- **국내 논문 문장 의미 태깅 데이터셋**
  - 논문 자동 요약 및 논문의 목적, 방법, 결과, 결론 별 문서 분류를 위한 기계학습 데이터셋
  - 총 155,740개의 논문 문장과 태그 쌍이 존재함.
  - 의미 태그는 의미 구조 분류/세부 의미 분류로 계층적 구조를 이룸.

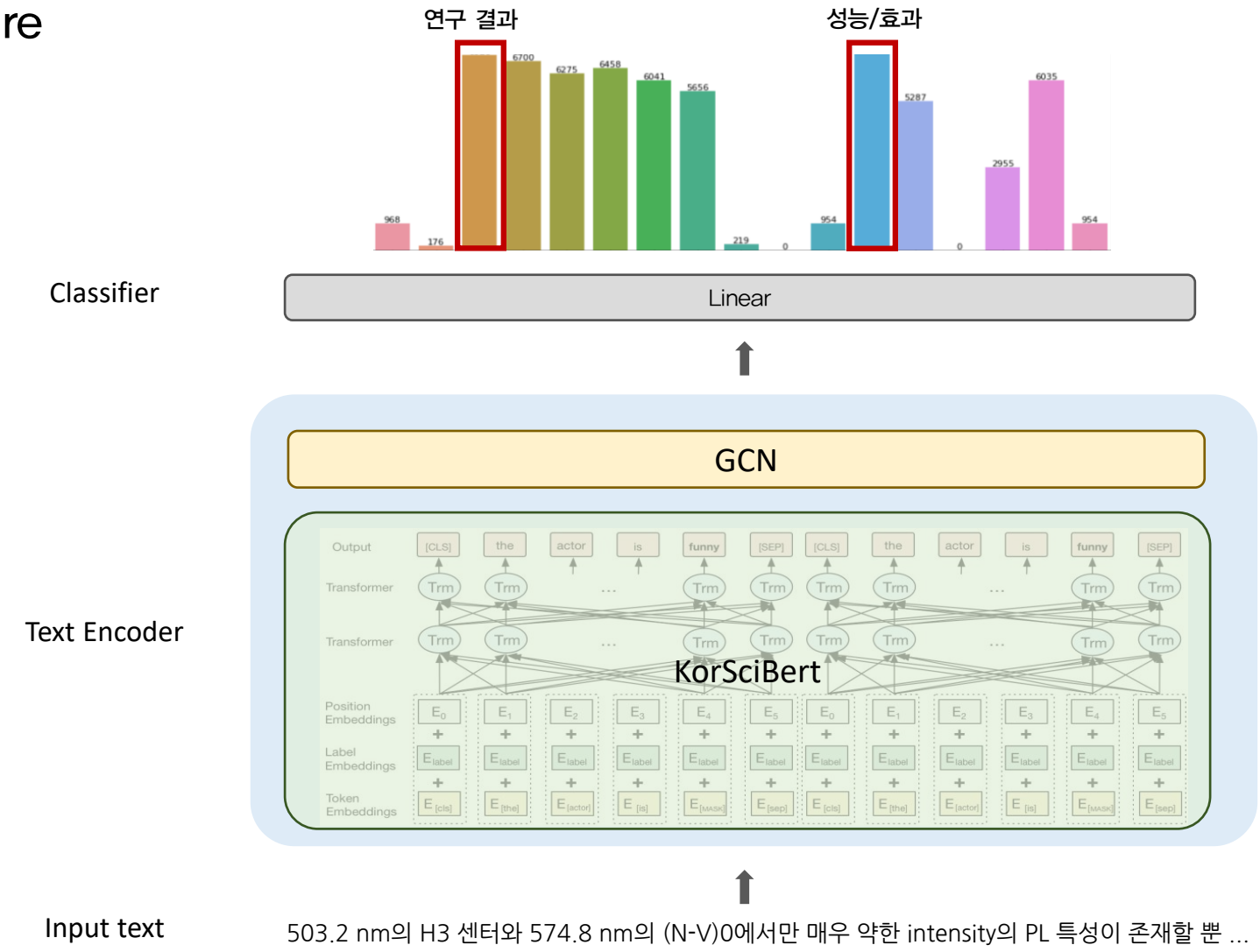

```
{
    "doc_id": "JAKO201534165174127",
    "sentence": "제시된 실험자료 및 예측 모델은 이들 물질을 취급하는 공정에서 방호 자료로 제공하고자 한다.",
    "tag": "문제 정의",
    "keysentence": "Y"
},
{
    "doc_id": "JAKO201534165174127",
    "sentence": "본 실험에서는 ASTM E659(Koehler사) 장치를 사용하여 톨루엔과 2-부탄올 계에 대해 AIT를 측정하였다. ",
    "tag": "제안 방법",
    "keysentence": "Y"
},
{
    "doc_id": "JAKO201534165174127",
    "sentence": "4) 본 연구에서 제시한 톨루엔과 2-부탄올 계의 AIT 예측식은 다른 조성에서도 AIT의 예측이 가능해 졌다.

    ",
    "tag": "성능/효과",
    "keysentence": "Y"
},
{
    "doc_id": "JAKO201534165174127",
    "sentence": "그리고 이성분계를 구성하는 순수성분인 톨루엔과 2-부탄올의 자연발화온도와 발화지연시간 관계를 측정하였다. ",
    "tag": "제안 방법",
    "keysentence": "Y"
},
```


의미 구조 분류	문장 의미 세부 분류
연구 목적	문제 정의
	가설 설정
	기술 정의
연구 방법	제안 방법
	대상 데이터
	데이터처리
	이론/모형
연구 결과	성능/효과
	후속연구

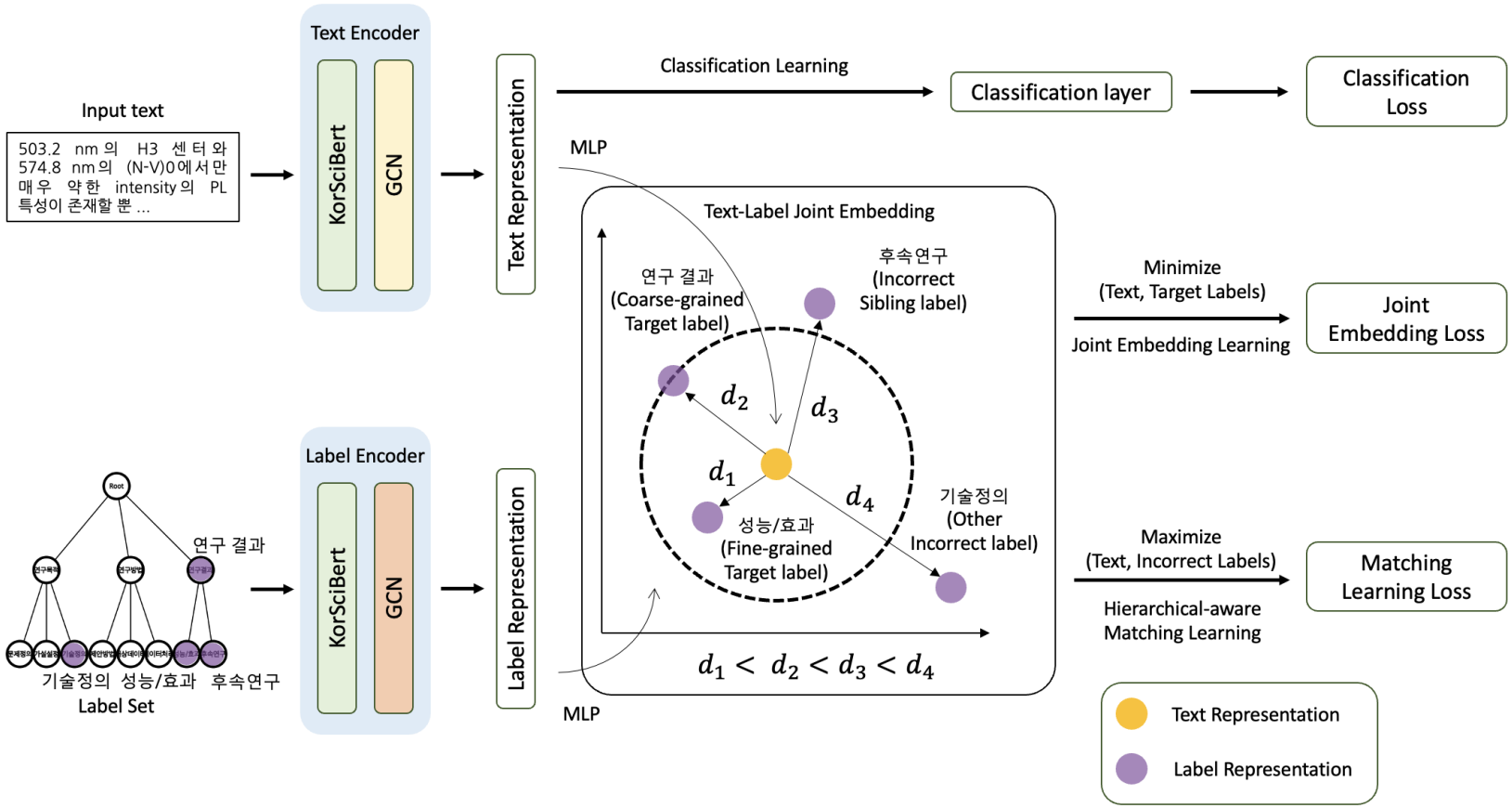
모델 개발 방법

- Model Architecture



모델 개발 방법

- Training Scheme
  - 서로 다른 3가지의 loss function을 활용하여 구성함.



모델 개발 방법

- Label Encoder
  - TF-IDF를 활용하여 각 label을 나타낼 수 있는 keyword를 추출함.

TF-IDF

Keyword 추출

	A	B
1	total_label	sentence
2	연구 목적	본 연구에서는 하고자 한다
3	연구 방법	이용하여 사용하였다
4	연구 결과	것으로 결과 나타났다
5	문제 정의	본 연구에서는 제안한다 살펴본다 하고자 한다
6	가설 설정	유의한 가설 가정하자 영향을 미칠 것이다
7	기술 정의	이다 말한다 알려져 있다 라고 한다
8	제안 방법	제안한다 제안하였다 확인하였다
9	대상 데이터	본 연구에서 사용한 본 실험에 사용된 대상으로
10	데이터처리	test를 이용하여 spss 실시하였다 대상자의 anova test 분석하였다
11	이론/모형	척도를 사용하였다 위해 개발한 도구를 측정하였다 이용하여
12	성능/효과	예측된다 판단된다 알 수 있다 결과 나타났다 있었다
13	후속연구	따라서 추후 연구에서는 활용될 수 있다 향후 필요할 것으로 것이다 필요하다

Label Description

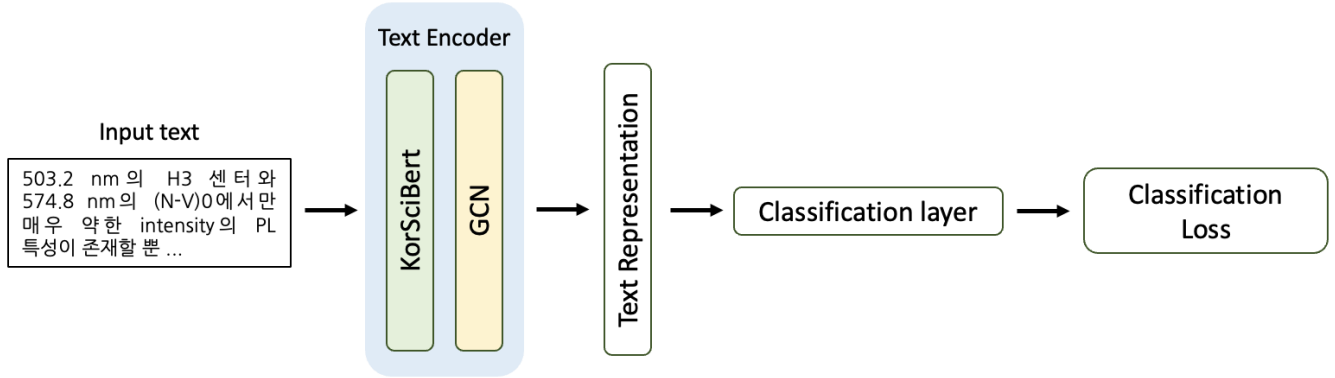
Label Encoder

KorSciBert

GCN

모델 개발 방법

- Classification Loss
  - Text Representation 만을 활용한 학습임.
  - 문장에 대해 대분류와 소분류를 예측하는 Multi-label classification task임.

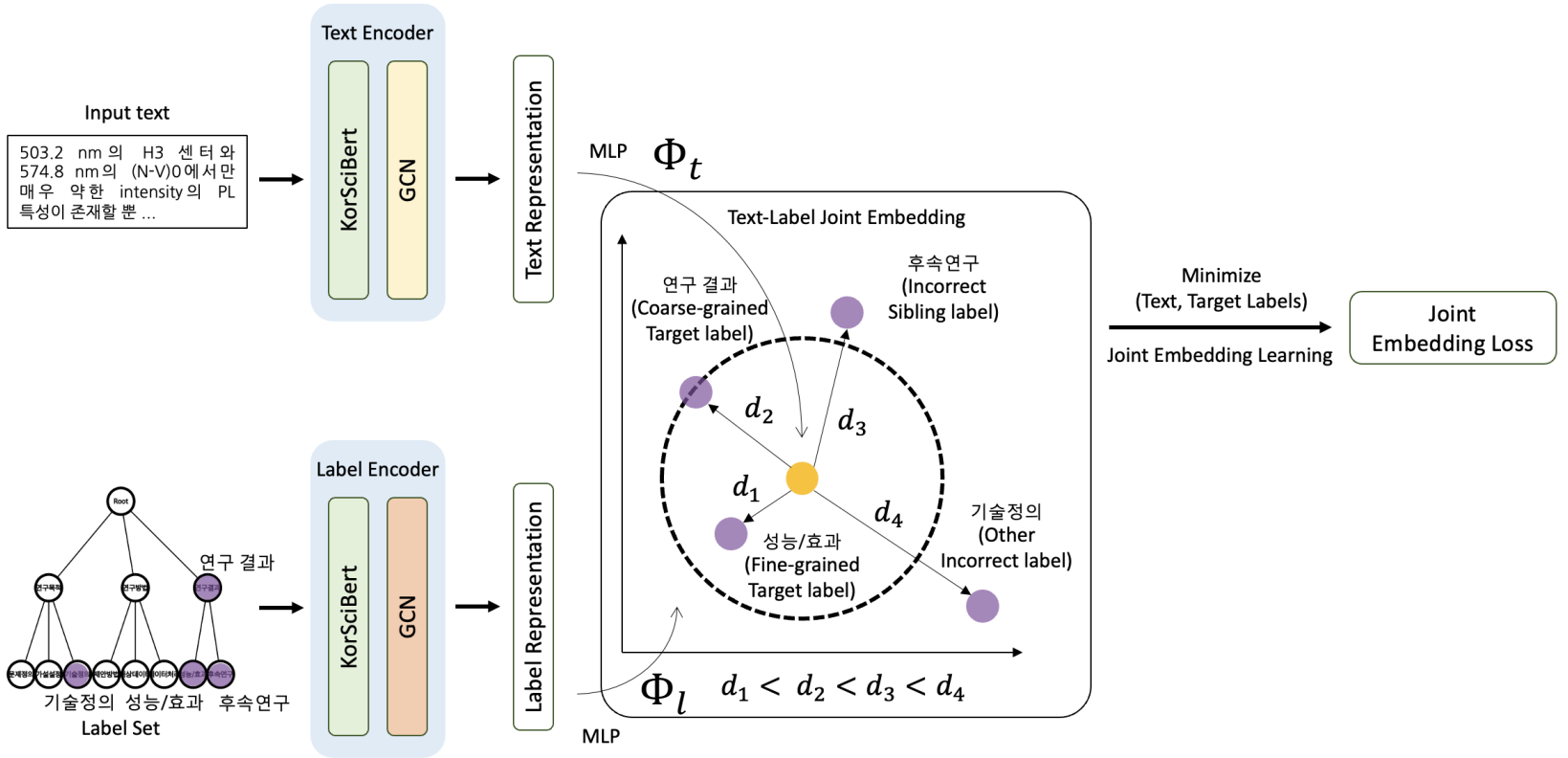


The loss can be described as:

$$\ell(x, y) = L = \{l_1, \ldots, l_N\}^\top, \quad l_n = -w_n \left[t_n \cdot \log \sigma(x_n) + (1 - t_n) \cdot \log(1 - \sigma(x_n)) \right]$$

모델 개발 방법

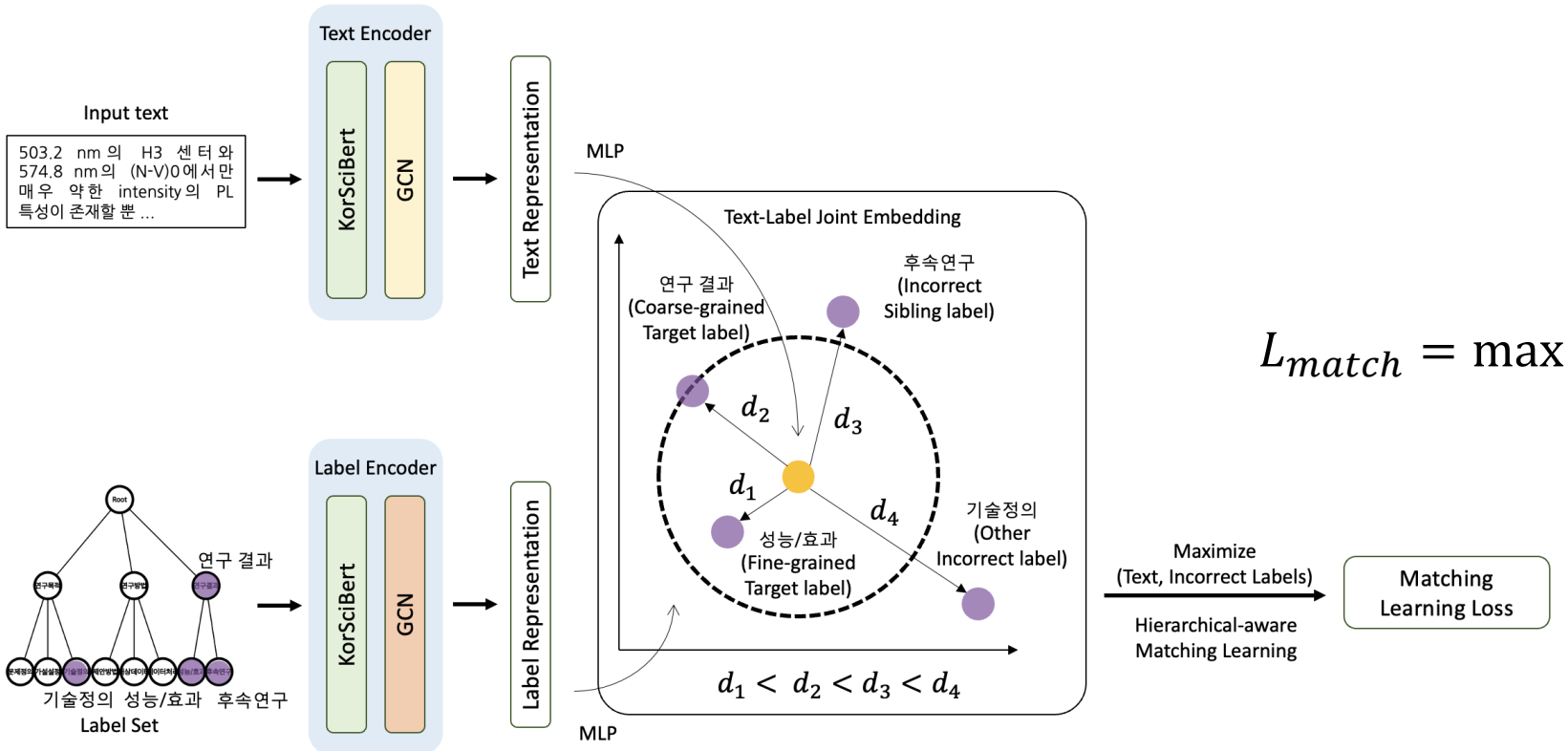
- Joint Embedding Loss
  - 같은 latent space 안에서 text semantic과 target label semantic 간의 거리를 최소화 함.



$$L_{joint} = \sum_{p \in P(y)} \left\| \Phi_t - \Phi_l^p \right\|_2^2$$

모델 개발 방법

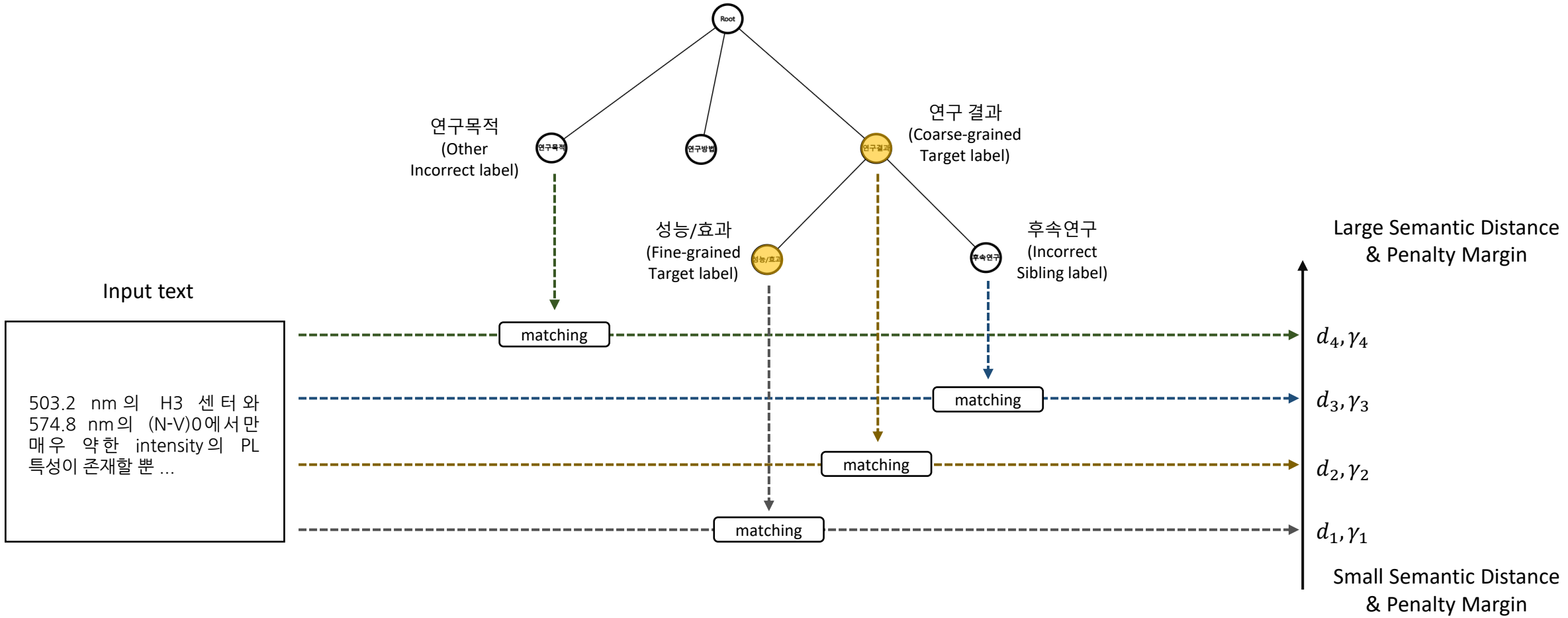
- Matching Loss
  - text semantic과 incorrect label semantic간에 distance(margin γ)를 둠.
  - Sibling label representation이 Other Incorrect label representation보다 더 가깝게 embedding 되도록 함.



$$L_{match} = \max(0, D(\Phi_t, \Phi_l^p) - D(\Phi_t, \Phi_l^n) + \gamma)$$

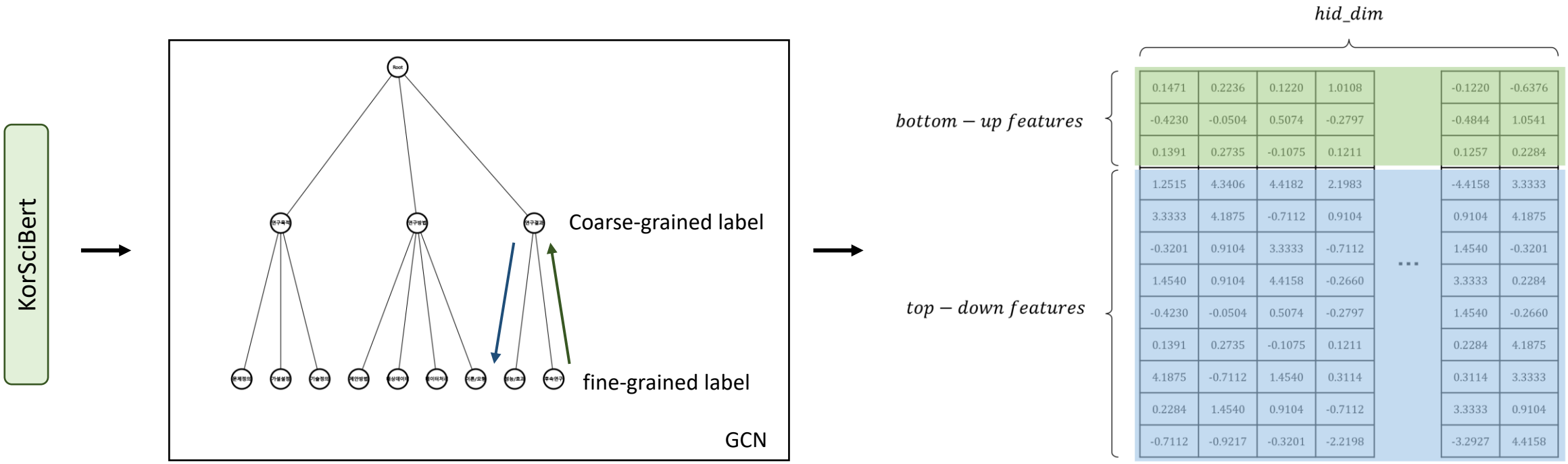
모델 개발 방법

- Hierarchy-aware margin
  - 각 label의 계층적 속성을 반영하여 margin이 부여됨.



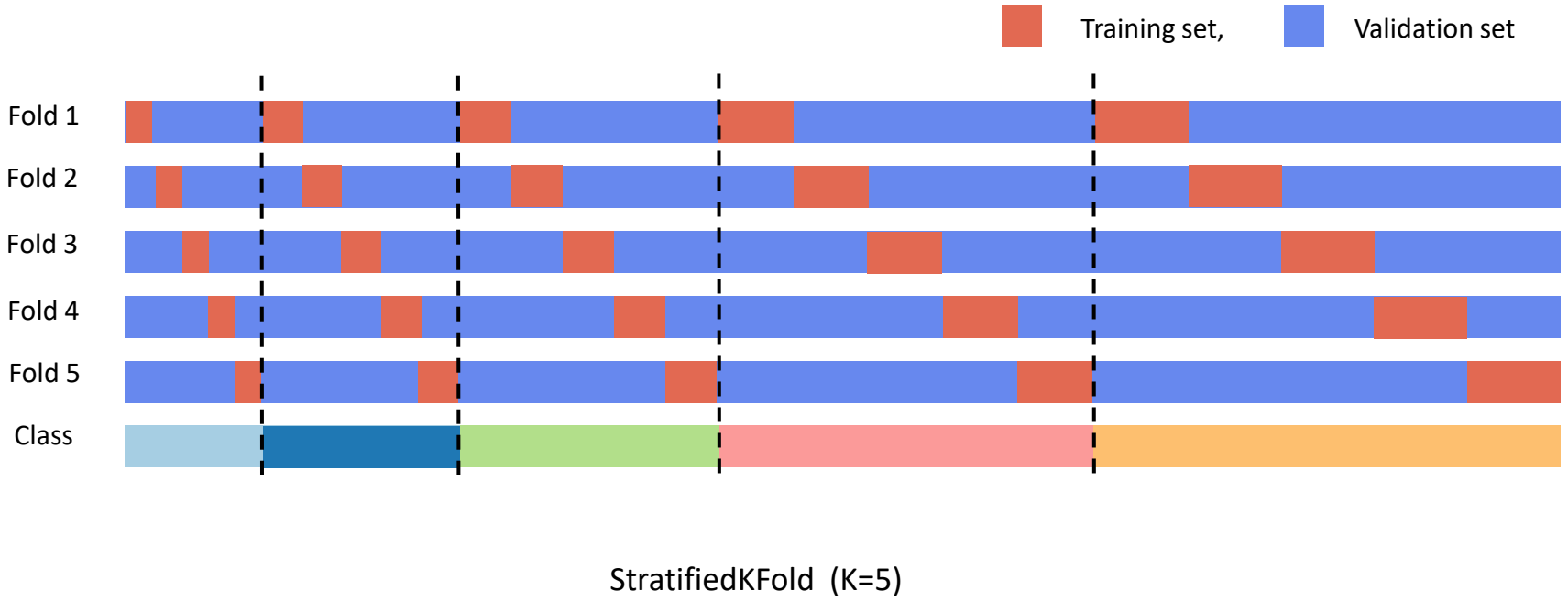
모델 개발 방법

- Why GCN ?
  - target label의 계층적 속성을 반영하여 representation을 구성함.
  - top–down & bottom–up 방식으로 feature를 조합함.



실험 및 평가

- Data Split
  - 데이터의 클래스 불균형을 고려하여 데이터를 나누는 StratifiedKfold를 사용함.
  - 5 Fold로 학습과 검증데이터를 나누고 교차검증을 진행함.



StratifiedKFold (K=5)

실험 및 평가

- Metric
  - 본 연구에 사용된 데이터는 클래스마다 데이터 수가 상이함.
  - 따라서 클래스 불균형을 고려한 Macro F1–Score와 Micro Accuracy를 평가지표로서 활용함.

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

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

실험 및 평가

- Experiments
  - 5 fold Cross-validation을 적용하여 평균낸 결과임.
  - Trainset 124592 / Testset 31148

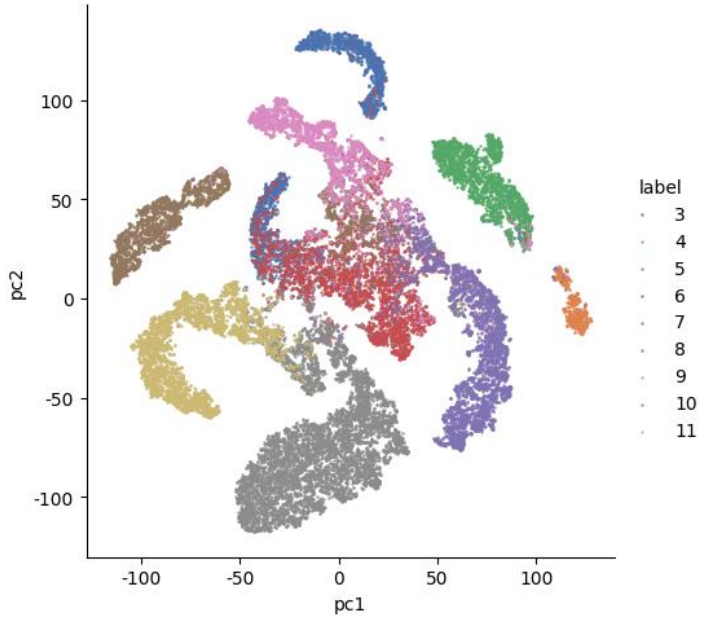
	Train	Test
size	124592	31148

	Coarse-grained		Fine-grained		Total	
	micro Acc	macro F1	micro Acc	macro F1	micro Acc	macro F1
KorSciBERT	0.9609	0.9552	0.8943	0.9008	0.8892	0.9007
w/ GCN	0.9615	0.9561	0.8971	0.9013	0.8905	0.9011
w/ matching	**0.9618**	**0.9565**	**0.8982**	**0.9023**	**0.8907**	**0.9022**

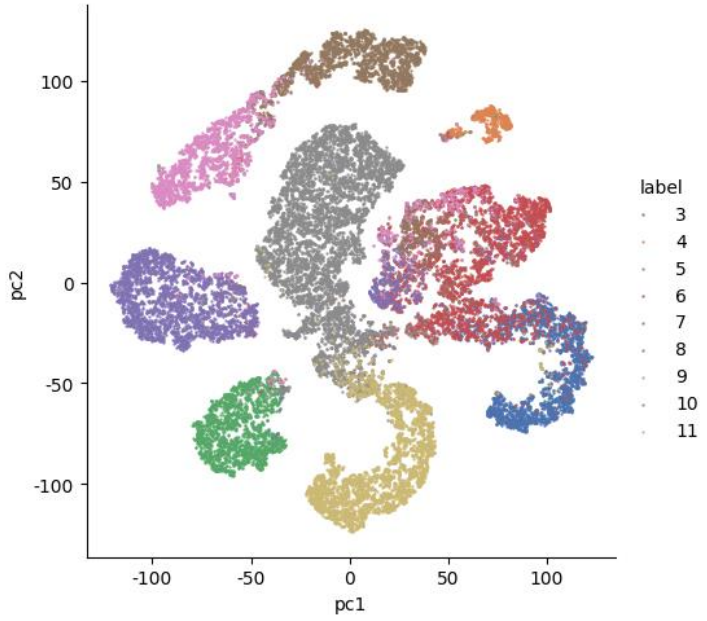
Result Table

실험 및 평가

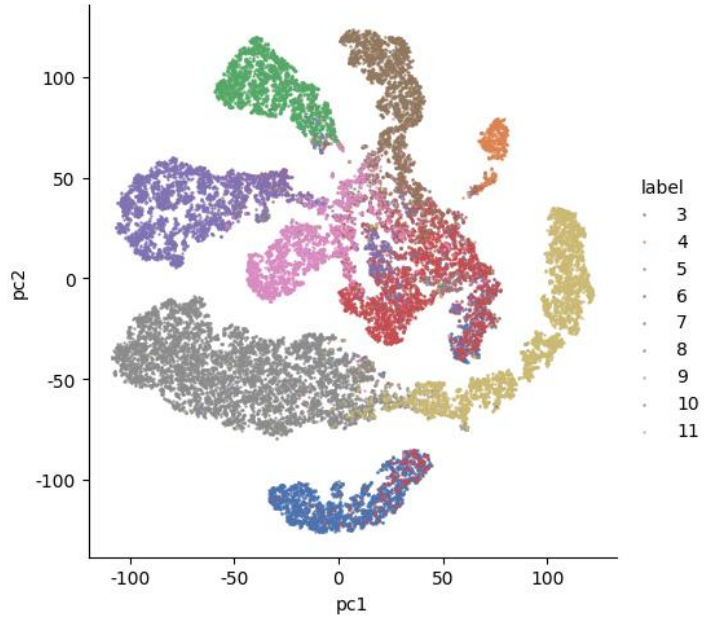
- Experiments
  - PCA를 통해 text embedding 값을 비교함.
  - KorSciBERT와 달리 coarse / fine-grained label의 정보가 잘 반영됨.



KorScibert



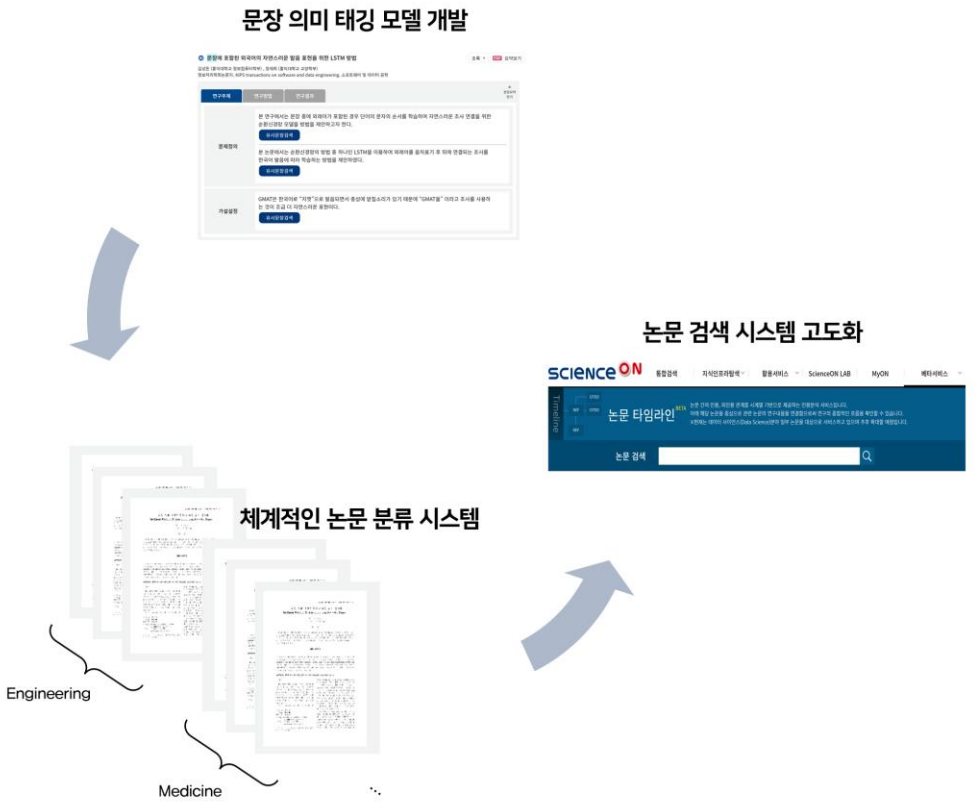
w/ GCN



w/ match

활용 계획 및 기대효과

- **체계적인 논문 분류 시스템 구축을 통한 검색 시스템 고도화**
  - 수 많은 논문 중 필요한 논문을 찾기 위해 빠르고 용이한 검색 시스템이 필요함.
  - 논문 문장 의미 태깅 모델은 논문 분류 시스템의 근간이 될 수 있으며, 이는 논문 검색 시스템 고도화에 기여할 수 있음.



활용 계획 및 기대효과

- **논문 자동 요약 시스템 고도화**
  - KISTI는 과학기술 지식 인프라 서비스 ScienceON에서 과학기술정보 활용을 용이하게 하는 AI 기반논문 요약 서비스를 제공하고 있음.

결론

- GCN을 활용한 논문 문장 의미 태깅 모델 개발
  - top-down & bottom-up 방식으로 feature를 조합하여 계층적 속성을 반영한 Representaion을 구성할 수 있었음.
  - Coarse-grained 95.65%, Fined-grained 90.23%의 macro F1 score를 보임.

- Label semantic distance를 고려한 다중 손실함수 사용
  - Hierarchical Matching Learning을 통해 계층적 속성을 반영함.
  - 다중 손실함수를 사용하여 보다 정교한 학습을 진행함.

향후 계획

- **데이터 전처리 및 후처리**
  - 현재는 데이터의 전처리와 후처리가 진행되지 않았음.

- **이전 문장 정보를 활용한 모델 고도화 및 성능 향상**
  - 현재 문장의 문맥을 파악하는데 auxiliary information 역할을 할 수 있는 이전 문장 정보를 활용하면 성능 향상에 도움이 될 것임.

- **Hierarchical Text Classfication의 여러 최신 모델 적용**
  - Hierarchical Text Classification의 최신 모델 비교실험을 통해 현재 국내 논문 문장 태깅 데이터셋에 가장 적합한 모델을 구축하고자 함.

Q&A

감사합니다 ☺