

Hierarchy-aware Label Semantics을 활용한 문장 태깅 분류

팀 AIDA

이소연, 한예지, 이병훈, 신우석

Contents

1. 프로젝트 개요
2. 활용 데이터
3. 모델 개발 방법
4. 실험 및 평가
5. 활용 계획 및 기대효과
6. 시연

프로젝트 개요

▪ 팀원 소개



이소연

고려대학교
산업경영공학과



한예지

고려대학교
산업경영공학과



이병훈

고려대학교
산업경영공학과



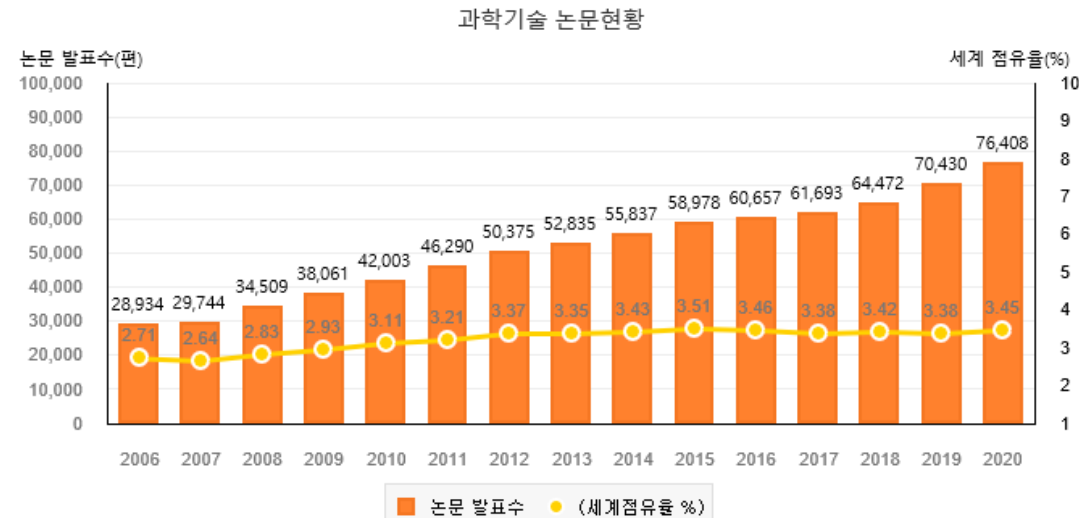
신우석

고려대학교
산업경영공학과

프로젝트 개요

■ 문제정의

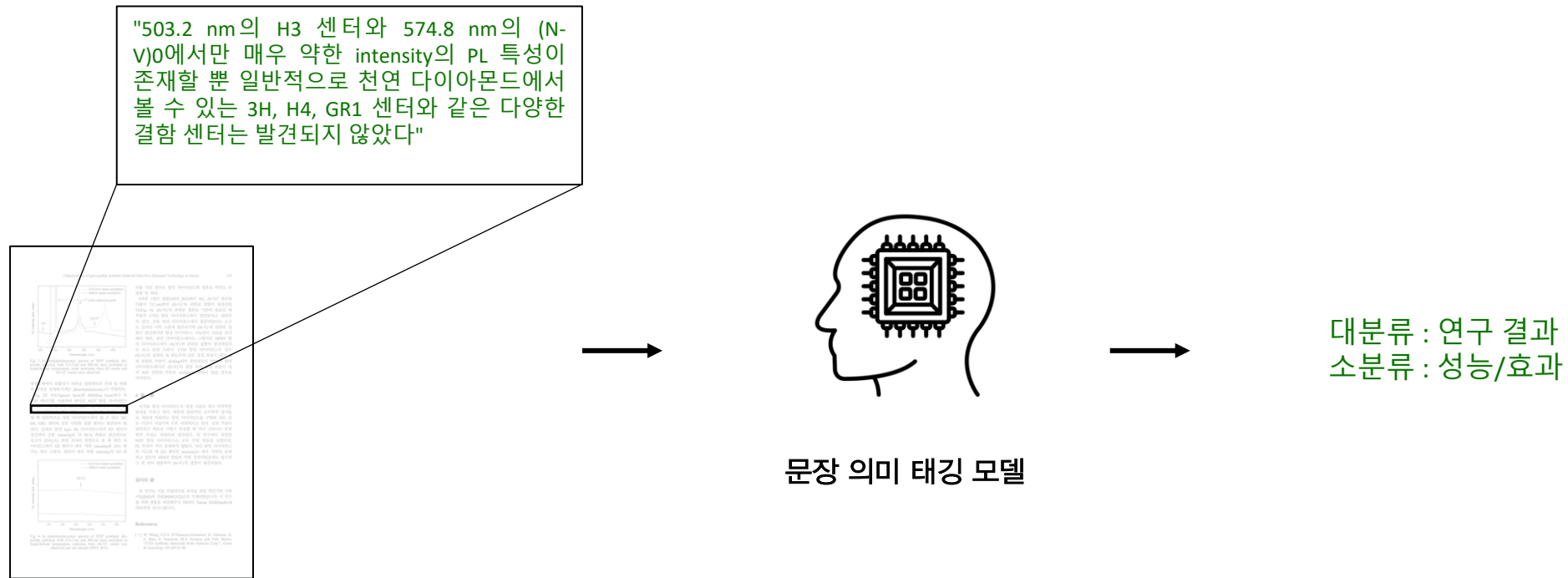
- 매년 다양한 연구분야에서 많은 양의 논문이 발표되고 있음.
- 특정 논문에서 원하는 정보를 추출하기 위해서는 시간과 비용이 많이 듦.
- 논문의 각 문장이 의미하는 바를 태그로 부착하여 자동화한다면 비용절감 효과를 볼 수 있음.



프로젝트 개요

개발목표

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



활용 데이터

■ 국내 논문 문장 의미 태깅 데이터셋

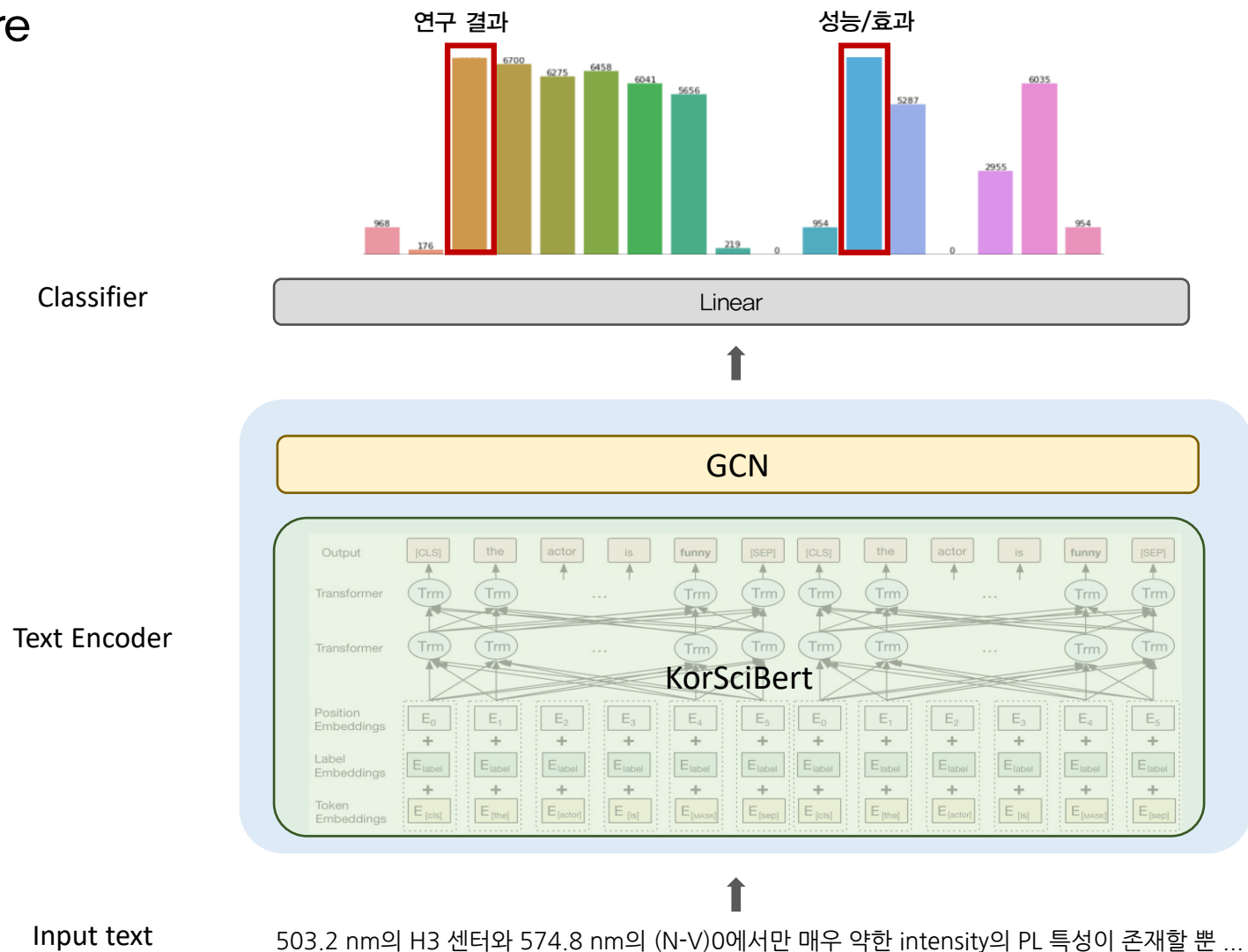
- 논문 자동 요약 및 논문의 목적, 방법, 결과, 결론 별 문서 분류를 위한 기계학습 데이터셋
- 총 155,740개의 논문 문장과 태그 쌍이 존재함.
- 의미 태그는 의미 구조 분류/세부 의미 분류로 계층적 구조를 이룸.

```
{
  "doc_id": "JAK0201534165174127",
  "sentence": "제시된 실험자료 및 예측 모델은 이들 물질을 취급하는 공정에서 방호 자료로 제공하고자 한다.",
  "tag": "문제 정의",
  "keysentence": "Y"
},
{
  "doc_id": "JAK0201534165174127",
  "sentence": "본 실험에서는 ASTM E659(Koehler사) 장치를 사용하여 톨루엔과 2-부탄올 계에 대해 AIT를 측정하였다. ",
  "tag": "제안 방법",
  "keysentence": "Y"
},
{
  "doc_id": "JAK0201534165174127",
  "sentence": "(4) 본 연구에서 제시한 톨루엔과 2-부탄올 계의 AIT 예측식은 다른 조성에서도 AIT의 예측이 가능해 졌다.
  ",
  "tag": "성능/효과",
  "keysentence": "Y"
},
{
  "doc_id": "JAK0201534165174127",
  "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

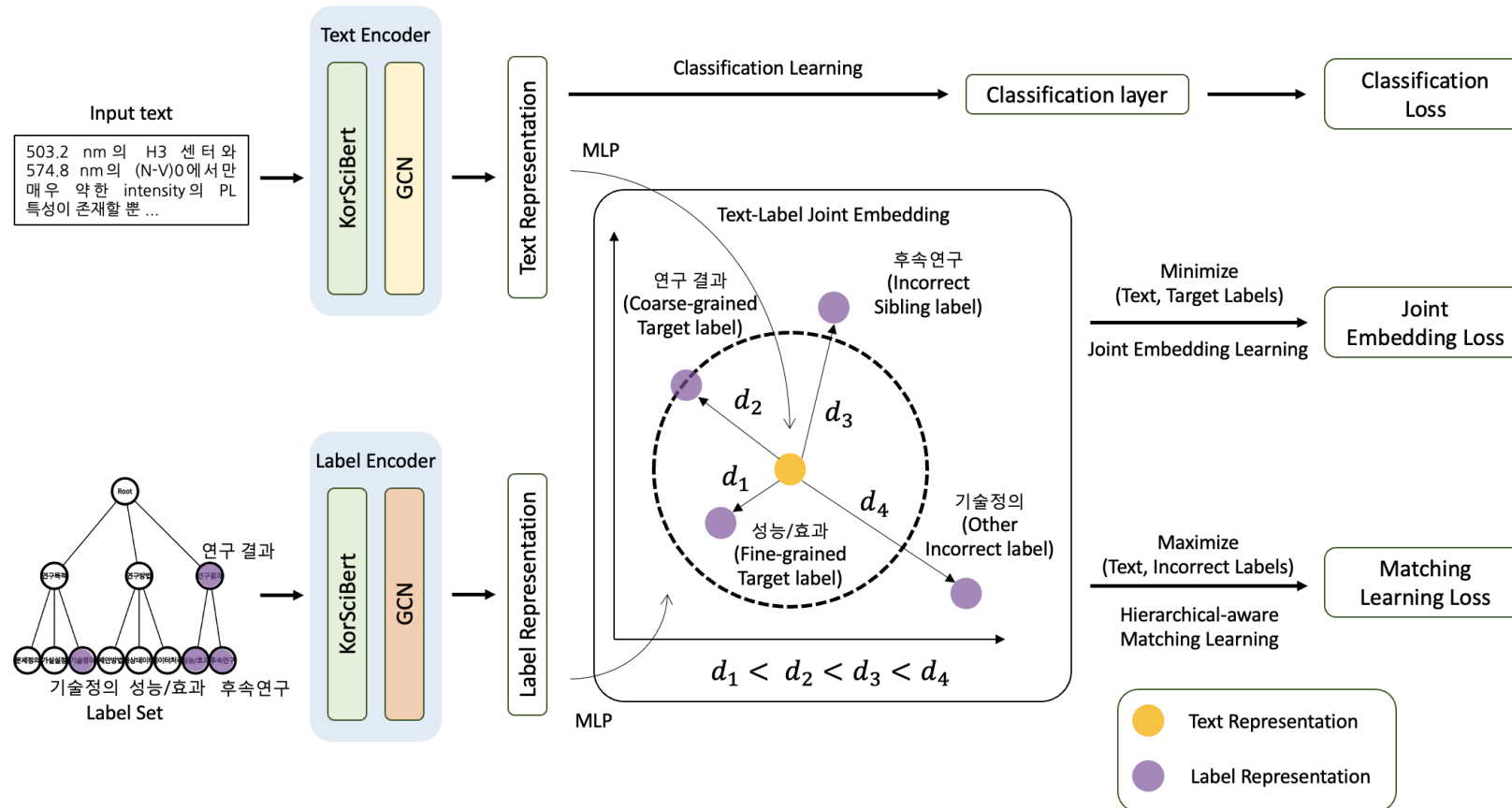
Model Architecture



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.

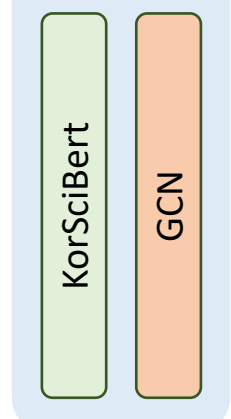


TF-IDF
Keyword 추출

	A	B
1	total_label	sentence
2	연구 목적	본 연구에서는 하고자 한다
3	연구 방법	이용하여 사용하였다
4	연구 결과	것으로 결과 나타냈다
5	문제 정의	본 연구에서는 제안한다 살펴본다 하고자 한다
6	가설 설정	유익한 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 나타냈다 있었다
13	후속연구	따라서 추후 연구에서는 활용될 수 있다 이후 필요할 것으로 것이다 필요하다

Label Description

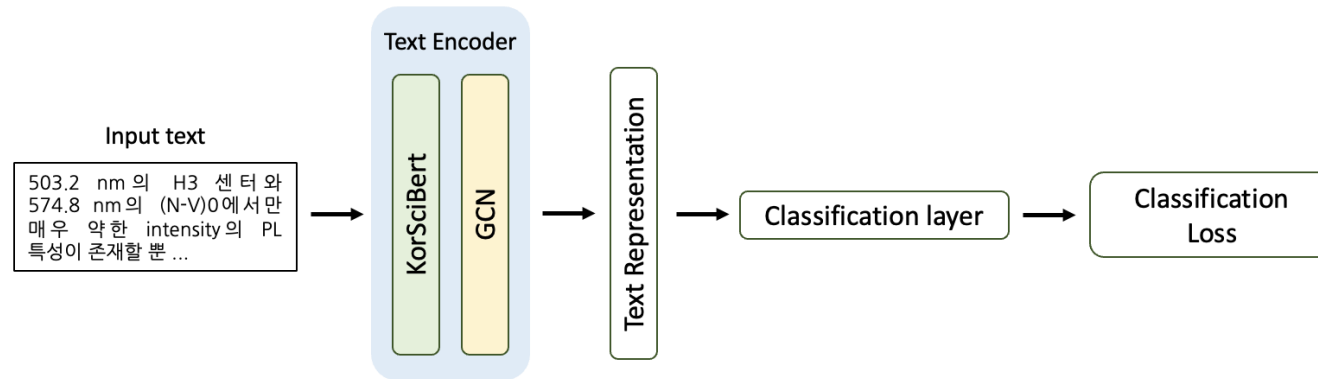
Label Encoder



모델 개발 방법

Classification Loss

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



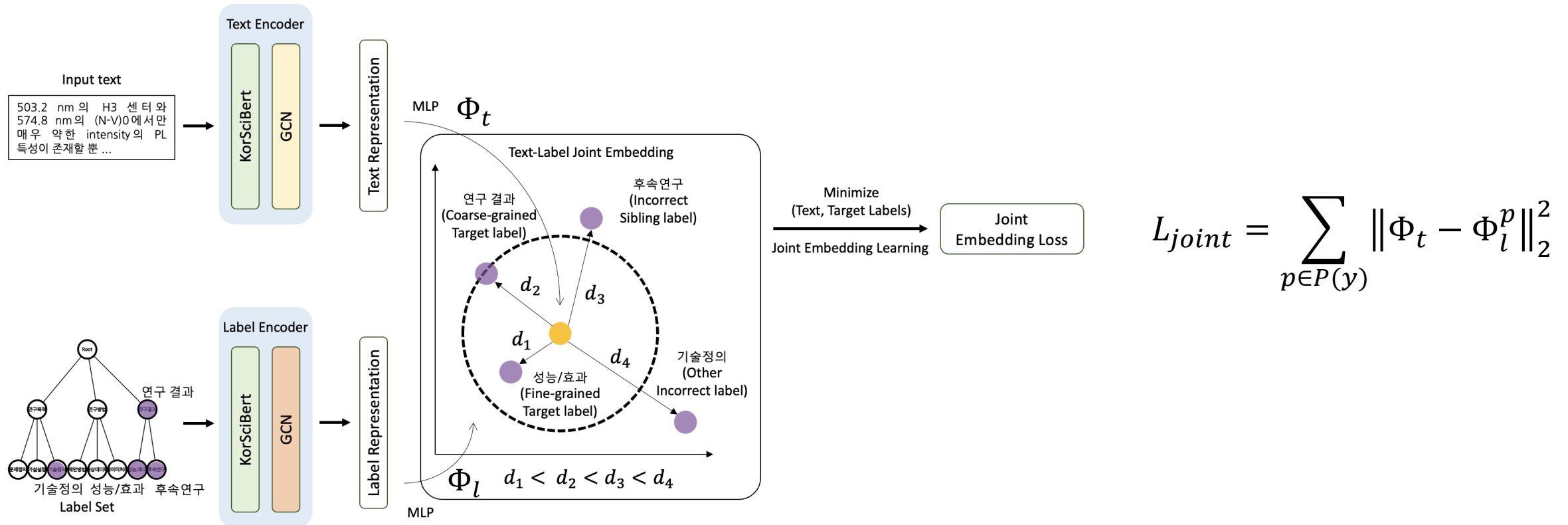
The loss can be described as:

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

모델 개발 방법

Joint Embedding Loss

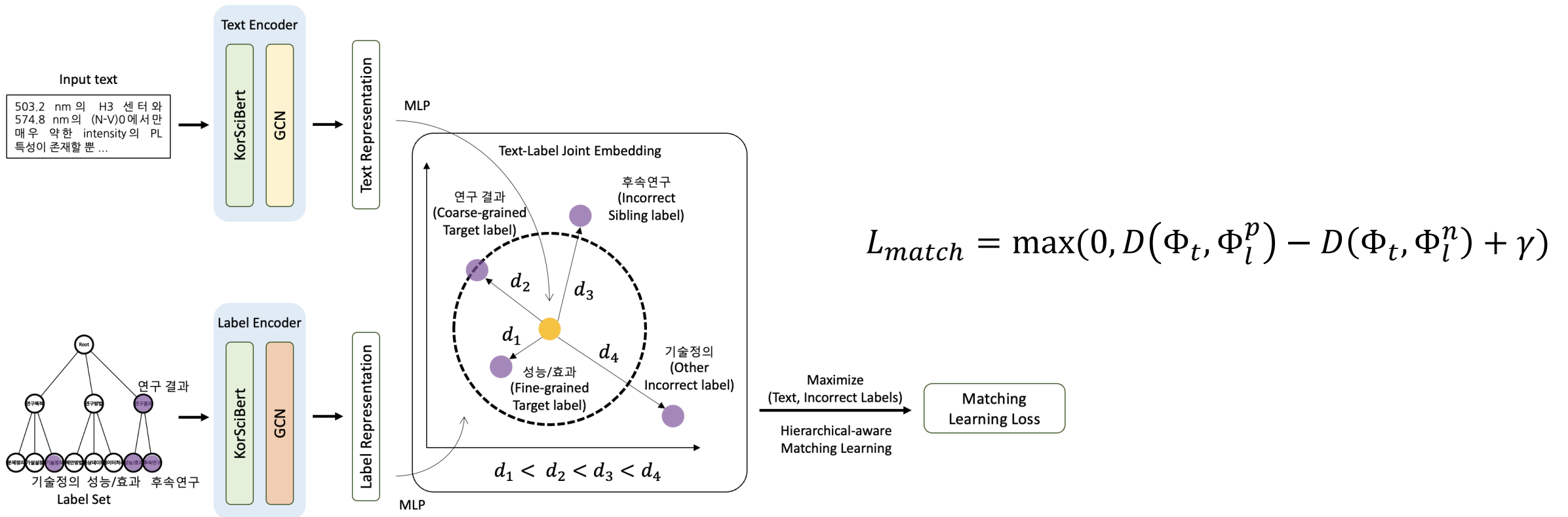
- 같은 latent space 안에서 text semantic과 target label semantic 간의 거리를 최소화 함.



모델 개발 방법

Matching Loss

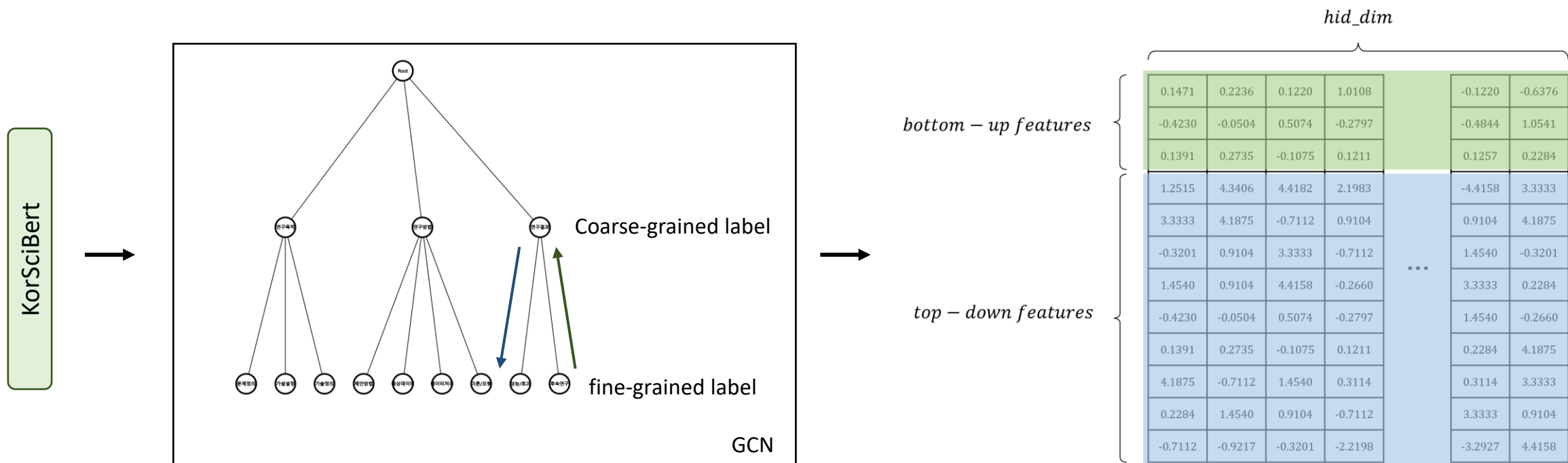
- text semantic과 incorrect label semantic간에 distance(margin γ)를 둠.
- Sibling label representation | Other Incorrect label representation보다 더 가깝게 embedding 되도록 함.



모델 개발 방법

Why GCN ?

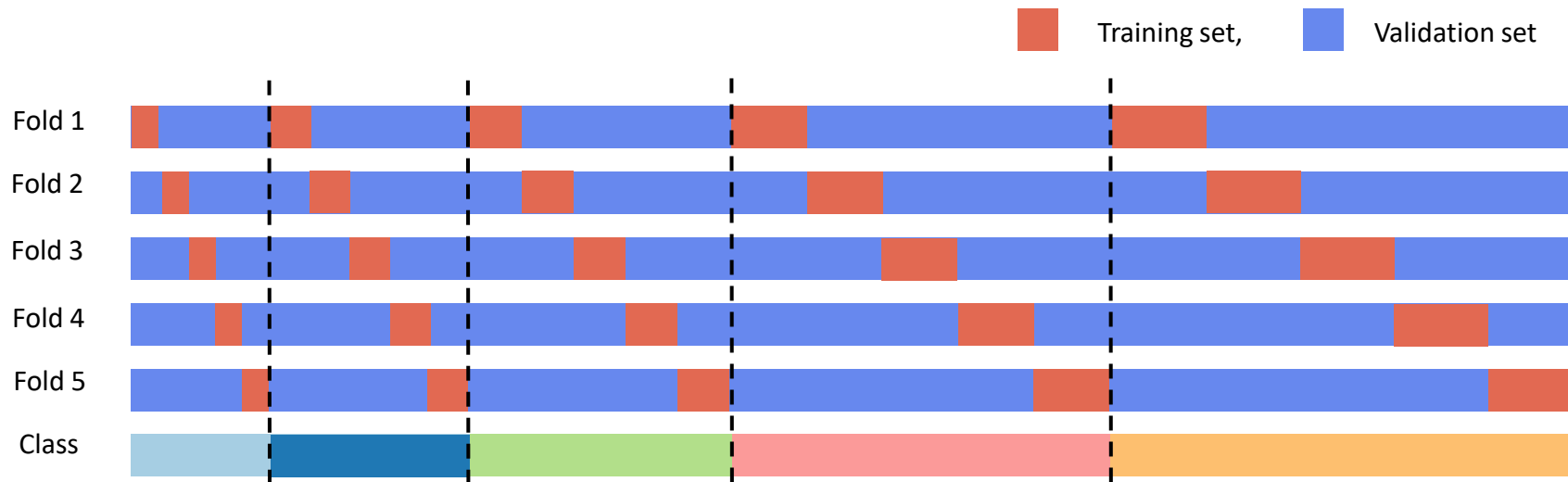
- target label의 계층적 속성을 반영하여 representation을 구성함.
- top-down & bottom-up 방식으로 feature를 조합함.



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.

$$\text{Macro F1 - Score} = \frac{1}{n} \sum_{i=1}^n \frac{(2 \times \text{precision}_i \times \text{recall}_i)}{\text{precision}_i + \text{recall}_i}$$

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

실험 및 평가

Experiments

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

	Train	Test
size	124592	31148

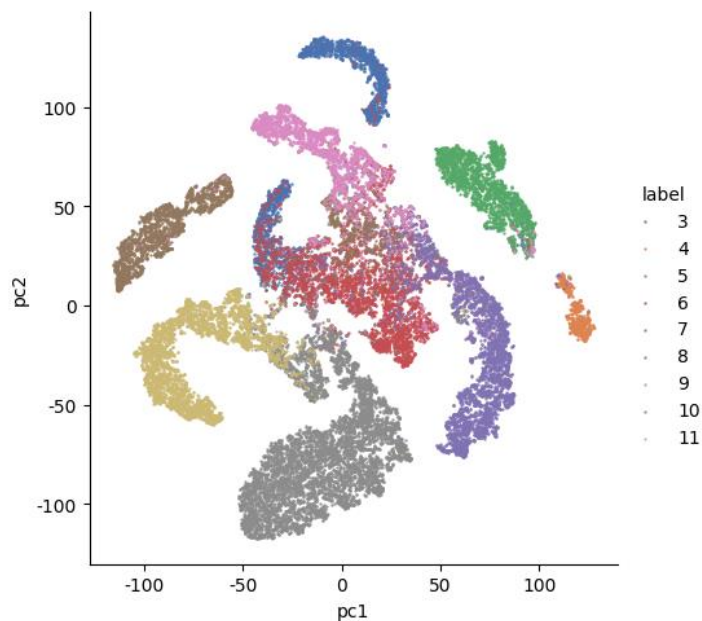
	Coarse-grained		Fine-grained		Total	
	micro Acc	macro F1	micro Acc	macro F1	micro Acc	macro F1
KorSciBERT	0.9609	0.9552	0.8943	0.9008	0.8892	0.9007
w/ GCN	0.9615	0.9561	0.8971	0.9013	0.8905	0.9011
w/ matching	0.9618	0.9565	0.8982	0.9023	0.8907	0.9022

Result Table

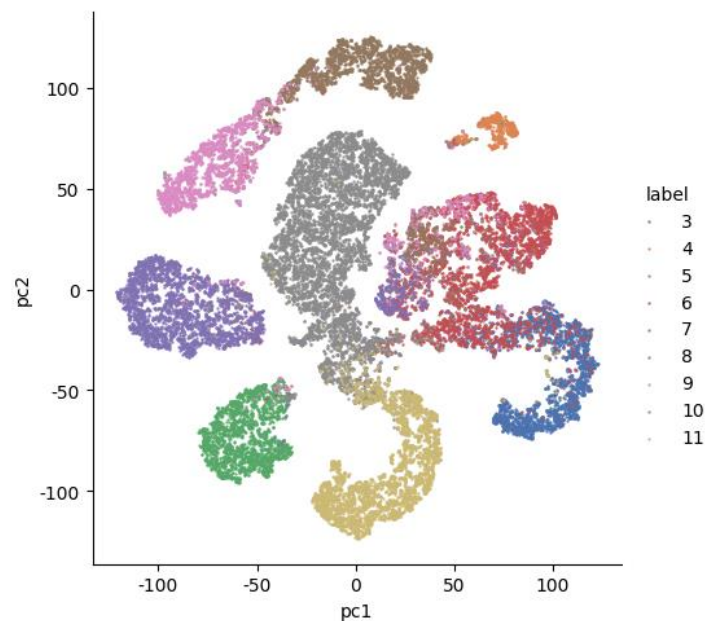
실험 및 평가

Experiments

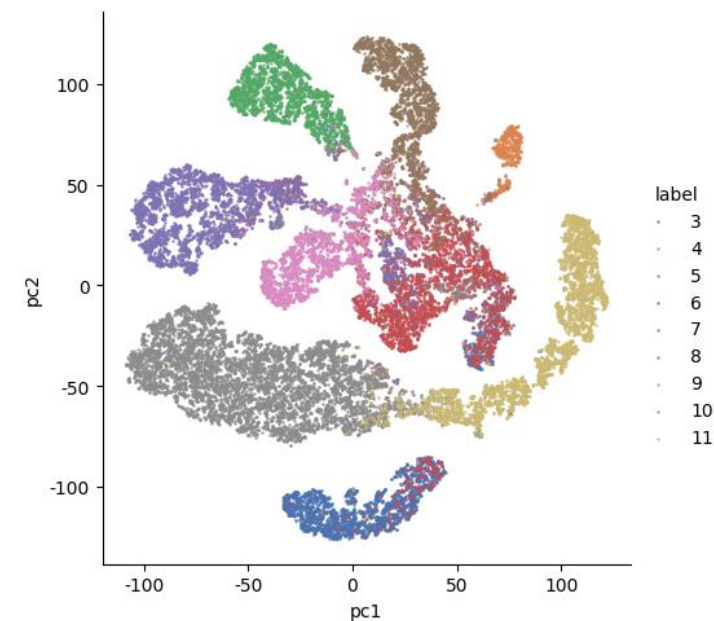
- PCA를 통해 text embedding 값을 비교함.
- KorSciBERT와 달리 coarse / fine-grained label의 정보가 잘 반영됨.



KorScibert



w/ GCN

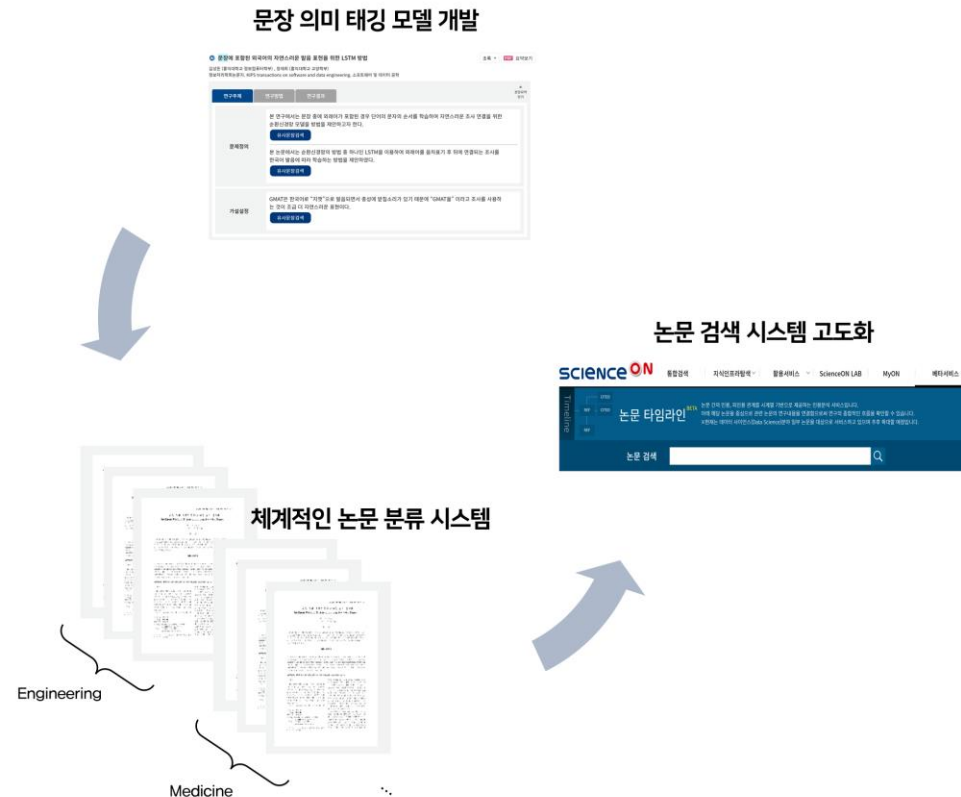


w/ match

활용 계획 및 기대효과

■ 체계적인 논문 분류 시스템을 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.

The screenshot shows the ScienceON website's AI Paper Summary service. The header includes navigation links for '통합검색', '지식인프라탐색', '활용서비스', 'ScienceON LAB', 'MyON', and '베타서비스'. The main banner features the text 'AI 논문 요약 BETA' and a description: '한국융합학회 2015~2020년 논문을 대상으로 AI기술(초록요약)을 활용하여 연구주제, 연구방법, 연구결과를 추출하고, 논문을 요약하는 베타서비스입니다. 서비스는 아래와 같이 3단계로 구성되어 있습니다.' Below the banner is a search bar with the text '한국융합학회 논문 검색' and a search icon. Below the search bar, it says '전체논문 940 건'. The main content area displays a search result for the article '국내 e스포츠산업의 비즈니스 모델에 관한 연구' by Yang Ji-hoon. The article details include the author's affiliation (Korea Convergence Society), the journal name, and a table with three rows: '연구주제' (Research Topic), '연구방법' (Research Method), and '연구결과' (Research Results).

ScienceON 통합검색 | 지식인프라탐색 | 활용서비스 | ScienceON LAB | MyON | 베타서비스

AI 논문 요약 BETA

한국융합학회 2015~2020년 논문을 대상으로 AI기술(초록요약)을 활용하여 연구주제, 연구방법, 연구결과를 추출하고, 논문을 요약하는 베타서비스입니다. 서비스는 아래와 같이 3단계로 구성되어 있습니다.

초록요약 ▶ 본문요약 ▶ PDF보기

한국융합학회 논문 검색

전체논문 940 건

국내 e스포츠산업의 비즈니스 모델에 관한 연구 PDF 다운로드 본문 요약하기

양지훈 (한국방송통신전파진흥원), 이인규 (KT 미래사업개발단), 이상호 (경성대학교 디지털미디어학부)
한국융합학회논문지 = Journal of the Korea Convergence Society

연구주제	본 연구는 국내 e스포츠 산업의 비즈니스 모델 실태를 진단 및 분석하고 시사점을 도출하는 것이다.
연구방법	e스포츠의 수익창출 경로, 현황 등을 파악하기 위해 e스포츠 구단, e스포츠 미디어, e스포츠 경기장, e스포츠 종목사로 나누고 각 주체별로 사업실 무자들을 대상으로 인터뷰를 실시했다.
연구결과	본 연구는 지금까지 거의 취급되지 못했던 e스포츠산업의 비즈니스 모델을 수익과 비용 중심으로 파악함으로써 업계와 산업 발전에 기여할 수 있을 것이라 기대된다.

결론

- GCN을 활용한 논문 문장 의미 태깅 모델 개발
 - top-down & bottom-up 방식으로 feature를 조합하여 계층적 속성을 반영한 Representation을 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 Classification의 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 😊