

# NLP-07조\_프로젝트 WrapUp

## 프로젝트 개요


데이터 제작 NLP에서는 **한국어 및 다른 언어에서의 자연어처리 데이터셋**의 유형 및 포맷이 어떠한지, 그리고 데이터셋을 구축하는 일반적인 프로세스가 무엇인지 학습하는 것을 목표로 합니다. 강좌에서 배운 내용을 바탕으로, 위키피디아 원시 말뭉치를 활용하여 직접 **관계 추출 태스크에 쓰이는 주석 코퍼스를 만들어보는 것**을 스페셜 미션으로 합니다.


본 대회에서 제공받은 데이터는 대학 데이터셋입니다. **이번 대회에 더 몰입하기 위해** 제공 받은 데이터와 추가로 크롤링한 데이터를 바탕으로 **대학에서 배울 수 있는 학문** 데이터 셋을 구축했습니다.


**자연어처리를 통해 대학 학문 관계 그래프를 만드는 이유**는 대학에서 배우는 여러 학문에 대한 구분이 명확하지 않다고 판단했기 때문입니다. 기존 분류에 대한 의견 및 이론 정립은 자주 나왔음에도 불구하고 많은 학생들이 여전히 명확한 구분을 하지 못하고 있습니다. 대량의 정보량을 토대로 자연어 학습을 하게 된다면 **‘실제로 쓰이는 바’를 토대로 지식그래프를 구축**할 수 있게 됩니다.


**대학에서 배울 수 있는 학문**들의 관계를 추출하고자 하는 연구자 분들에게 도움이 될 것이고 이외 **학문에 대한 지식 그래프 및 유관 분야 소개** 등으로의 확장 또한 가능할 것입니다.

## 프로젝트 팀 구성 및 역할

 김한성: 가이드라인 작성 + 데이터 증강 / LM Finetuning

 이재욱: 가이드라인 작성 + 관계 점검 및 수정

 최동민: 가이드라인 작성 + 관계 점검 및 수정

 염성현: 가이드라인 검수 + 어노테이션 관리

 홍인희: Relation map 작성 + IAA 측정

## 프로젝트 수행 절차 및 방법

### 데이터 확인 및 1차 관계 설정

#### 데이터 확인 - 대학 데이터셋

- 총 41개의 txt 파일, 총 1692개의 문장으로 구성, 아래는 각 txt 파일의 이름과 문장의 개수

이름	개수	이름	개수	이름	개수
영어	184	한문	38	생명과학	12
물리학	128	예비고사	37	현우진	12
교육과정	128	베트남어	34	재수	11
화학	127	중국어	33	교재	6
일본어	100	경제	32	최태성	5
아랍어	93	국어	28	고등학교	5
스페인어	89	학교생활기록부	23	성적표	5
교사	83	입학	23	지구과학	4
프랑스어	81	학력고사	20	문제	2

이름	개수	이름	개수	이름	개수
수학	66	한국교육과정평가원	19	강사	2
러시아어	53	졸업	19	정시	1
대학	52	학생	15	내신	1
독일어	51	시험	15	OMR	1
문학	40	성적	14		

- 주로 언어, 학문, 교육에 대한 데이터들로 데이터셋이 구성
- Entity를 찾기 어려운 문장, 단락이 끊긴 문장들이 다수 존재
- 데이터셋의 목적이 필요하다고 느껴 **대학에서 배울 수 있는 학문 데이터셋 구축**을 목적으로 하고 작업을 진행

## 1차 관계 설정

- 파일럿 태깅 이후 아래와 같이 임시 Relation map을 작성하였음

Id	Relation	SUBJ	OBJ	Description
1	no_relation	ALL	ALL	관련이 없음, 관계 유추 불가
2	학문:하위학문	STU	STU	Object는 Subject의 하위 학문
3	학문:상위학문	STU	STU	Object는 Subject의 상위 학문
4	학문:별칭	STU	STU	Object는 Subject의 별칭
5	학문:인물	STU	PER	Object는 Subject를 주장한 인물
6	학문:시대	STU	DAT	Object는 Subject가 등장한 시대
7	언어:하위언어	LAN	LAN	Object는 Subject의 하위 언어이다.
8	언어:상위언어	LAN	LAN	Object는 Subject의 상위 언어이다.
9	언어:과목	LAN	STU	Object는 Subject에 대한 과목
10	언어:사용지역	LAN	LOC	Object는 Subject 사용 기관/지역
11	언어:별칭	LAN	LAN	Object는 Subject의 또 다른 이름
12	언어:사용민족	LAN	ORG	Object는 Subject를 쓰는 집단

- 임시 Relation map을 기반으로 1차 분담 태깅 작업을 진행

## Case Study

- 1차 분담 태깅 작업을 진행하며, Entity 설정 및 태깅 관련 애로 사항들에 대해 토의
- Entity 설정
  - STU Entity
    - 학문 또는 이론으로 표현된 STU entity의 범위 설정에 대한 각각의 의견이 달랐고, 혼란을 방지하기 위해 **OO학, OO론, OO법칙 등의 예시를 두어 대상 범위를 한정**
    - 상위/하위 관계로 학문 간의 관계를 모두 표현하기 어렵다는 의견이 있어, **학문:영향 관계를 추가하여 데이터셋의 표현력 제고**
  - LAN Entity
    - 공용어, 외래어, 굴절어, 고립어 등 나라 및 집단에 종속되지 않은 언어에 대한 표현이나 언어학적 표현의 분류에 대한 논의가 있었음

## 언어학의 기초분야에 해당하는 개념들까지 포함한 것을 LAN Entity로 설정

- 태깅 관련 애로 사항
  - 태깅의 대상 : 고유 명사 vs 보통 명사
    - 어떤 종류의 Token까지 태깅의 대상으로 볼 것인가에 대한 논의가 있었고, **뜻이 한정되지 않거나 대상이 구체화되지 않는 Token들은 태깅하지 않기로 한정**
  - Data Error
    - 관계 없는 문장을 합쳐 관계 문장을 만들거나 문장을 정제하는 것에 대한 논의가 있었음, KLUE 가이드라인을 참고, **기준에 부합하는 문장 만으로 관계 문장을 만들기로 하였으며 문장들을 정제한 후 태깅 작업을 진행**
- 자주 묻는 질문에 그 외 Case Study 관련 내용들을 추가하여 혼란을 미연에 방지하기로 함

## 1차 산출물 제출 및 피드백

### Relation map

- 1차 분담 태깅 진행 후, Relation map을 아래와 같이 수정

	class_name (ko)	class_name (en)	direction (sub, obj)	description
1	관계_없음	no_relation	(*, *)	관계를 유추할 수 없음. 정의된 클래스 중 하나도
2	학문:하위_학문	stu:sub_study	(STU, STU)	Object는 Subject의 하위 학문
3	학문:상위_학문	stu:high_study	(STU, STU)	Object는 Subject의 상위 학문
4	학문:별칭	stu:alternate_names	(STU, STU/POH/ORG)	Object는 Subject의 또 다른 이름
5	학문:기여자	stu:contributor	(STU, PER)	Object는 Subject에 기여한 인물
6	학문:시대	stu:area	(STU, DAT)	Object는 Subject가 등장한 시대
7	학문:연구_집단	stu:research_group	(STU, ORG)	Object는 Subject를 다루는 기관 혹은 연구집단
8	학문:영향	stu:cause	(STU, STU/POH)	Object는 Subject의 영향을 받은 것
9	언어:하위_언어	lan:sub_language	(LAN, LAN)	Object는 Subject의 하위 언어
10	언어:상위_언어	lan:high_language	(LAN, LAN)	Object는 Subject의 상위 언어
11	언어:파생물	lan:product	(LAN, POH)	Object는 Subject 언어의 파생물
12	언어:사용_지역	lan:use_area	(LAN, LOC)	Object는 Subject 사용 지역
13	언어:별칭	lan:alternate_names	(LAN, LAN/POH/ORG)	Object는 Subject의 또 다른 이름
14	언어:사용_집단	lan:group_of_people	(LAN, ORG)	Object는 Subject를 사용 집단

- 수정 사항은 다음과 같음
  - 학문:인물을 학문:기여자로 수정
    - 학문에 대한 직접적인 기여가 있는 자로 의미를 제한
  - 학문:연구\_집단, 학문:영향 추가
    - 학문에 대한 연구를 진행하는 기관 혹은 집단을 표현
    - 학문이 영향을 끼친 대상을 추가적으로 표현
  - 언어:과목을 언어:파생물로 수정
    - 과목이 주로 교과에 한정되었기 때문에 특정한 언어로 쓰인 저서 및 파생물로 의미를 확장
  - 언어:사용민족을 언어:사용\_집단으로 수정
    - 민족 외에 기관 또는 기구 등 더 넓은 범위를 포괄

### 가이드라인

- 관계 및 Entity 가이드라인, Annotation 환경 가이드라인, 자주 묻는 질문을 나눠 작성 후 검수함
- 관계 및 Entity 가이드라인



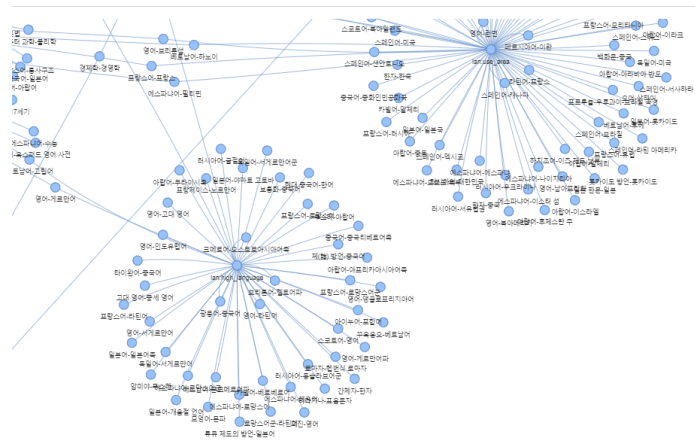
	class	class	동민	성현	인희	재욱	한성
lan:관계_없음	206	14.71%	lan:하위_언어	lan:별칭	lan:별칭	lan:하위_언어	lan:관계_없음
stu:관계_없음	161	11.5%	lan:사용_지역	lan:사용_지역	lan:사용_지역	lan:사용_지역	lan:사용_지역
stu:요소	136	9.71%	lan:하위_언어	lan:하위_언어	lan:하위_언어	lan:하위_언어	lan:하위_언어
stu:기여자	118	8.43%	lan:관계_없음	lan:관계_없음	lan:관계_없음	lan:관계_없음	lan:관계_없음
stu:하위_학문	107	7.64%	lan:사용_집단	lan:사용_집단	lan:사용_집단	lan:사용_집단	lan:사용_집단
lan:사용_지역	98	7.0%	lan:사용_지역	lan:사용_지역	lan:사용_지역	lan:사용_지역	lan:사용_지역
stu:상위_학문	97	6.93%	lan:파생물	lan:관계_없음	lan:별칭	lan:파생물	lan:파생물
stu:영향	79	5.64%	lan:관계_없음	lan:관계_없음	lan:관계_없음	lan:관계_없음	lan:관계_없음
lan:사용_집단	72	5.14%	lan:사용_지역	lan:사용_지역	lan:관계_없음	lan:사용_지역	lan:사용_지역
lan:하위_언어	68	4.86%	lan:하위_언어	lan:하위_언어	lan:하위_언어	lan:하위_언어	lan:하위_언어
lan:상위_언어	62	4.43%	lan:하위_언어	lan:하위_언어	lan:하위_언어	lan:하위_언어	lan:하위_언어
stu:연구_집단	48	3.43%	lan:별칭	lan:별칭	lan:별칭	stu:별칭	lan:별칭
lan:별칭	43	3.07%	lan:하위_언어	lan:하위_언어	lan:하위_언어	lan:하위_언어	lan:관계_없음
stu:시대	42	3.0%	lan:관계_없음	lan:관계_없음	lan:관계_없음	lan:관계_없음	lan:관계_없음
stu:별칭	35	2.5%	lan:사용_지역	lan:사용_지역	lan:사용_지역	lan:사용_지역	lan:사용_지역
lan:파생물	28	2.0%					

- 왼쪽 그림은 개별 어노테이션 이후의 관계 분포, 오른쪽 그림은 종합 어노테이션 시트 예시
- 관계 없음의 비중이 KLUUE와 같이 크게 높지 않았으나 가장 높은 축에 속하였음
- 학문에 대한 관계들이 많았고 **요소 및 기여자, 하위/상위가 그 중 많았음**
- 언어는 **사용 지역과 사용 집단** 관계가 가장 많았음
- **별칭** 관계가 학문, 언어에 대해 모두 적었고 **파생물** 관계가 가장 적었음

## 관계 네트워크 시각화



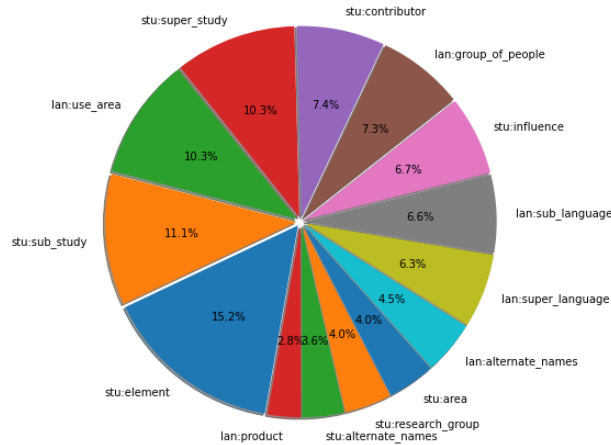
- 왼쪽은 Type별 Entity 쌍 네트워크 시각화 자료, 오른쪽은 단어별 Entity 쌍 네트워크 시각화 자료



- 단어별 Entity 쌍 네트워크 시각화 자료 확대 자료

- 관계들을 중심으로 각각의 Type 및 단어 쌍들이 연결된 모습을 관찰할 수 있음
- 더 큰 데이터셋을 활용한다면 단어 간의 관계에 대한 더욱 유의미한 관찰이 가능할 것으로 보임

## IAA, Fleiss Kappa 산출



```
#raters = 5 , #subjects = 1400 , #categories = 15
PA = 0.8752142857142855
PE = 0.09166873469387755
Fleiss' Kappa = 0.863
```

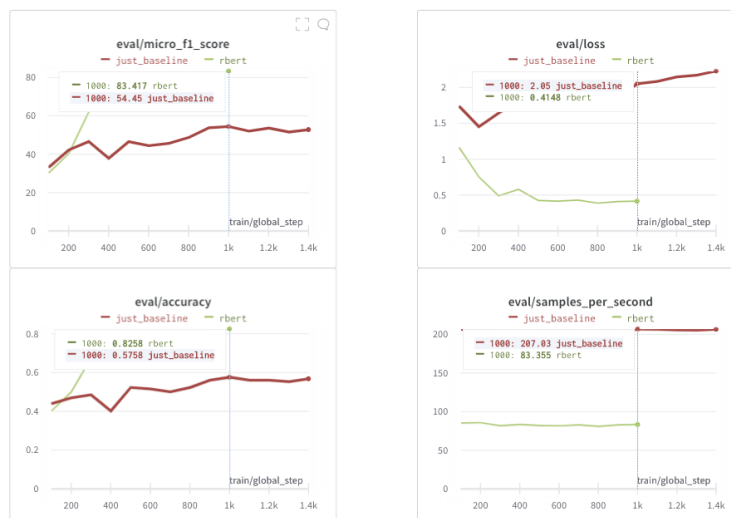
- 1차 태깅과 Case Study를 병행하였기 때문에 Fleiss Kappa가 높게 나왔음
  - Entity 정의와 범위 설정에 대해 각각의 Case를 보고 합의 과정을 거쳤기 때문으로 추정

## LM Finetuning

최종 학습 데이터 총 정리

- total /train /dev dataset : 1313/ 1181/ 132

### ▾ 베이스라인 및 RBERT 비교



제공 받은 베이스라인으로 모델을 학습, 또한 further reading을 위해 대회 당시 SOTA(State Of The Art)모델로 학습 후 비교진행

- f1-score 비교 결과
  - 제공된 baseline code : 54.45
  - rBERT(RE대회 SOTA) : **83.417**

[Wandb Link](#)

## 관계 점검 및 최종 산출

### 관계 점검

- STU:시대 정의가 모호
  - 시대에 대한 태깅 작업이 상이, 등장 시대가 아닌 연관된 시대로 변환
- LAN:파생물 정의가 모호
  - 언어와 관련 있는 POH가 아닌 언어로 쓰인 저작물, 작품 등 파생관계가 명확한 것으로 제한
- STU:요소 정의가 모호
  - STU:영향과 혼동, STU:요소는 entity들 간의 포함관계가 명확하게 드러나 있는 경우로 제한

### 최종 Relation map

	class_name (ko)	class_name (en)	direction (sub, obj)	description
1	관계 없음	no_relation	(*, *)	관계를 유추할 수 없음 / 정의된 클래스 중 하나로 분류
2	학문:하위_학문	stu:sub_study	(STU, STU)	Object는 Subject의 하위 학문
3	학문:상위_학문	stu:super_study	(STU, STU)	Object는 Subject의 상위 학문
4	학문:별칭	stu:alternate_names	(STU, STU/POH/ORG)	Object는 Subject의 또 다른 이름
5	학문:기여자	stu:contributor	(STU, PER)	Object는 Subject에 기여한 인물
6	학문:시대	stu:era	(STU, DAT)	Object는 Subject와 연관된 시대
7	학문:연구_집단	stu:research_group	(STU, ORG)	Object는 Subject를 다루는 기관 혹은 집단
8	학문:영향	stu:influence	(STU, STU/POH)	Object는 Subject의 영향을 받은 것
9	학문:요소	stu:element	(STU, POH)	Object는 Subject에 포함되는 요소
10	언어:하위_언어	lan:sub_language	(LAN, LAN)	Object는 Subject의 하위 언어
11	언어:상위_언어	lan:super_language	(LAN, LAN)	Object는 Subject의 상위 언어
12	언어:파생물	lan:product	(LAN, POH)	Object는 Subject 언어의 파생물
13	언어:사용_지역	lan:area_of_use	(LAN, LOC)	Object는 Subject 사용 지역
14	언어:별칭	lan:alternate_names	(LAN, LAN/POH/ORG)	Object는 Subject의 또 다른 이름
15	언어:사용_집단	lan:group_of_users	(LAN, ORG)	Object는 Subject를 사용 집단

### 최종 IAA, Fleiss Kappa 산출

```
#raters = 5 , #subjects = 1314 , #categories = 15
PA = 0.8951293759512942
PE = 0.12292918551878956
Fleiss' Kappa = 0.88
```

- Fleiss Kappa 상승 : 0.863 → 0.88

### 관계 비교 및 최종 산출물 제출

- 어노테이션 작업물을 비교하여 관계 데이터셋 최종본을 도출
- Relation map 및 가이드라인을 확정 후 제출

## 프로젝트 회고 및 향후 과제

### 데이터 및 모델 신뢰도에 대한 회고

#### 데이터 어노테이션에 대한 신뢰 → IAA 체크 → 가이드라인 업데이트

1. 1차적으로 가이드라인을 잘 작성하는 것이 중요함 하지만 주석자 간의 이해가 다르기 때문에 태깅 및 관계에 대한 이견이 생기는 것은 불가피.
2. 작업 중간에 의견을 합치하는 것보다 작업 후에 Fleiss Kappa를 산출하고 그를 바탕으로 가이드라인을 업데이트하는 것이 중요.
3. 블라인드 테스트 등으로 가이드라인과 어노테이션에 대한 실험을 한다면 그 정합성을 더 높일 수 있을 것이라고 보임.

#### 모델 학습에 대한 신뢰

1. Fleiss Kappa : 0.88, 총 데이터 수 : 1313
2. Fleiss Kappa 점수가 높다는 것이 관계가 비교적 명확하게 정의되었다는 것을 뜻하지 않을까 하여 모델 또한 학습을 잘하지 않을까 기대하였음
3. 하지만 동시에 데이터셋의 크기가 작기 때문에 학습이 잘 진행될 수 있을까 의문이 들기도 하였음
4. 베이스라인의 성능, f1-score가 54.45 였고 데이터셋이 작은 것이 어느 정도 문제가 된다 판단
5. 하지만, 자체 구축한 rBERT의 성능이 **83.417** 으로 높게 나와 모델링으로 극복 가능하다는 것을 확인하였고 데이터 설계 또한 잘 진행되었음을 이를 통해 가늠해볼 수 있었음

### 향후 과제

- 별도의 test\_data(unseen\_data)를 구축한다면 일반화 성능까지 확인해볼 수 있을 듯함