Wrap Up Report

- **NLP 2조 퍼스트펭귄**

[목차]

---

1. 프로젝트 개요

- 한국어 및 다른 언어에서의 자연어처리 데이터셋 유형 및 포맷이 어떠한지, 그리고 데이터셋을 구축하는 일반적인 프로세스가 무엇인지 알아보는 것을 목표로 한다.
- 위키피디아에서 얻은 **스마트폰** 관련 원시 말뭉치를 활용하여 직접 관계 추출 태스크에 쓰이는 주석 코퍼스를 만들어본다.

2. 프로젝트 팀 구성 및 역할

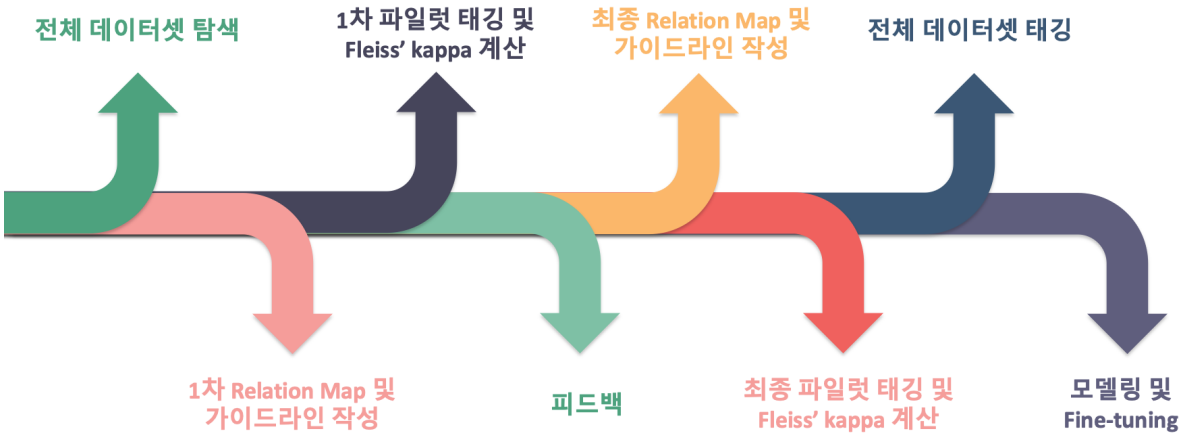
- **고우진** : relation map 제작, 가이드 라인 작성, annotation 작성, 데이터셋 제작 코드 작성
- **현승엽** : relation map 제작, 가이드 라인 작성, annotation 작성
- **이종윤** : relation map 제작, annotation 작성, Fleiss' kappa 계산
- **김상윤** : relation map 제작, annotation 작성, 모델 Fine tuning

3. 데이터셋 설명

- **스마트폰** 관련 키워드로 검색된 위키피디아 코퍼스
- 총 **35개**의 text파일, **1190개**의 문장으로 구성

4. 프로젝트 수행 절차 및 방법

4.1. Workflow



4.2. 세부사항

1. **전체 데이터셋 탐색**
   - 팀원별로 데이터를 4등분하여 전체 데이터셋을 탐색하였다.
   - 이 과정에서 팀원별로 **스마트폰** 주제와 관련된 entity 및 relation을 정의 하였다.

2. **1차 Relation Map 및 가이드라인 작성**
   - 각자 정의한 entity와 relation을 비교하여 총 **7개**의 **entity**와 **14개**의 **relation**을 정의하였고, 이를 기반으로 1차 Relation Map 및 가이드라인을 작성하였다.
   - **Entities**

     `PER(사람)` , `ORG(단체)` , `POH(고유명사)` , `NOT(수량)` , `DAT(날짜)` , `VIS(유형기술)` , `INV(무형기술)`
   - **Relations**

	class_name (ko)	class_name (en)	direction (sub, obj)	description
1	관계_없음	no_relation	(*,*)	관계를 유추할 수 없음. 정의된 클래스 중 하나로 분류할 수 없음
2	단체 : 제작	org : produce	(ORG, VIS/INV)	Object는 Subject가 제작/제공한 것
3	단체 : 구성원	org : employees	(ORG, PER)	Object는 Subject의 구성원
4	유형기술 : 출시/제작일	vis : release_date/production_date	(VIS, DAT)	Object는 Subject의 출시/제작일
5	무형기술 : 출시/제작일	inv : release_date/production_date	(INV, DAT)	Object는 Subject의 출시/제작일
6	무형기술 : 사용자_수	inv : number_of_users	(INV, NOH)	Object는 Subject의 사용자수
7	유형기술 : 별칭	vis : alternative_names	(VIS, VIS)	Object는 Subject의 별칭
8	무형기술 : 별칭	inv : alternative_names	(INV, INV)	Object는 Subject의 별칭
9	유형기술 : 기반기술	vis : foundation_technique	(VIS, VIS/INV)	Object는 Subject의 기반 기술
10	무형기술 : 기반기술	inv : foundation_technique	(INV, VIS/INV)	Object는 Subject의 기반 기술
11	유형기술 : 구성요소	vis : component	(VIS, VIS/INV)	Object는 Subject의 구성 요소
12	무형기술 : 구성요소	inv : component	(INV, VIS/INV)	Object는 Subject의 구성 요소
13	유형기술 : 기능	vis : function	(VIS, INV)	Object는 Subject의 기능
14	무형기술 : 기능	inv : function	(INV, INV)	Object는 Subject의 기능

3. **1차 파일럿 태깅 및 Fleiss' kappa 계산**

- 전체 데이터셋에서 랜덤으로 **100개**의 문장을 추출하였고, **1차 Relation Map 및 가이드라인**을 기반으로 파일럿 태깅을 진행하였다.
- 작업자간 일치도를 측정하기 위해 **Fleiss' kappa score**를 계산하였다.
  - **Fleiss' kappa score: 0.81**
  - (팀원 모두 no_relation이라고 태깅한 데이터 제외 후 계산시 **0.717**)

4. **피드백**

- 1차 Relation Map 및 가이드라인을 기반으로 태깅을 진행시 relation이 `no_relation`으로 태깅되는 데이터가 많았다.(59.43%)
- 전체 데이터셋에서 `no_relation`에 해당하는 데이터를 줄이고, entity 사이의 보다 구체적인 관계를 파악하기 위해 `유형기술 : 제원`, `유형기술 : 기타속성`, `무형기술 : 기타속성` 에 해당하는 3가지 relation을 추가하였다.

5. **최종 Relation Map 및 가이드라인 작성**

- 피드백 결과를 반영하여 총 **7개의 entity**와 **17개의 relation**을 정의하였고, 이를 기반으로 **최종 Relation Map 및 가이드라인**을 작성하였다.
- **Entities**

  `PER(사람)` , `ORG(단체)` , `POH(유형기술)` , `NOT(수량)` , `DAT(날짜)` , `VIS(유형기술)` , `INV(무형기술)`
- **Relations**

	class_name (ko)	class_name (en)	direction (sub, obj)	description
1	관계_없음	no_relation	(*,*)	관계를 유추할 수 없음. 정의된 클래스 중 하나로 분류할 수 없음
2	단체 : 제작	org : produce	(ORG, VIS/INV)	Object는 Subject가 제작/제공한 것
3	단체 : 구성원	org : employees	(ORG, PER)	Object는 Subject의 구성원
4	유형기술 : 출시/제작일	vis : release_date/production_date	(VIS, DAT)	Object는 Subject의 출시/제작일
5	무형기술 : 출시/제작일	inv : release_date/production_date	(INV, DAT)	Object는 Subject의 출시/제작일
6	무형기술 : 사용자_수	inv : number_of_users	(INV, NOH)	Object는 Subject의 사용자수
7	유형기술 : 별칭	vis : alternative_names	(VIS, VIS)	Object는 Subject의 별칭
8	무형기술 : 별칭	inv : alternative_names	(INV, INV)	Object는 Subject의 별칭
9	유형기술 : 기반기술	vis : foundation_technique	(VIS, VIS/INV)	Object는 Subject의 기반 기술
10	무형기술 : 기반기술	inv : foundation_technique	(INV, VIS/INV)	Object는 Subject의 기반 기술
11	유형기술 : 구성요소	vis : component	(VIS, VIS/INV)	Object는 Subject의 구성 요소
12	무형기술 : 구성요소	inv : component	(INV, VIS/INV)	Object는 Subject의 구성 요소
13	유형기술 : 기능	vis : function	(VIS, INV)	Object는 Subject의 기능
14	무형기술 : 기능	inv : function	(INV, INV)	Object는 Subject의 기능
15	유형기술 : 제원	vis : specification	(VIS, POH/NOH)	Object는 Subject의 속성(길이, 모양(비율), 무게, 색깔, 디자인)
16	유형기술 : 기타속성	vis : property	(VIS, POH/NOH)	Object는 Subject의 속성(가격 등)
17	무형기술 : 기타속성	inv : property	(INV, POH/NOH)	Object는 Subject의 속성(가격, 주파수 등)

6. **최종 파일럿 태깅 및 Fleiss' kappa 계산**

- 전체 데이터셋에서 랜덤으로 **100개**의 문장을 추출하였고, **최종 Relation Map 및 가이드라인**을 기반으로 파일럿 태깅을 진행하였다.
- 작업자간 일치도를 측정하기 위해 **Fleiss' kappa score**를 계산하였다.
  - **Fleiss' kappa score: 0.87** (1차 대비 0.06상승)
  - (팀원 모두 no_relation이라고 태깅한 데이터 제외 후 계산시 **0.833**(1차 대비 0.116 상승))

- 1차 파일럿 태깅 결과 대비 더 향상된 작업자간 일치도를 가지는 것을 확인하였다.

7. **전체 데이터셋 태깅**
   - 최종 파일럿 태깅에서 **Fleiss' kappa score**가 상승한 것을 확인했기 때문에, **최종 Relation Map 및 가이드라인**을 기준으로 전체 데이터셋 태깅을 진행하였다.
   - 지나치게 짧은 문장이나 entity를 태깅할 수 없는 문장들을 제외하였고, 한 문장에서 여러 관계가 발견된다면 모두 추가하였다. 그 결과 기존 1190개 문장의 데이터에서 총 **890**개 의 데이터를 얻었다.

8. **모델링 및 Fine-tuning**
   - 제작한 데이터셋의 label 별 비율을 고려하여 stratified 방식으로 train(0.8), test(0.2)로 나누었다.
     - train: 716개, test: 174개
   - 이전의 RE task 대회를 통해 전반적으로 가장 성능이 좋았던 `klue/roberta-large` 모델을 사전 학습 모델로 하여 fine-tuning을 진행하였다.
     - Model
       - `klue/roberta-large` , `klue/roberta-large + entity marker`
     - Hyperpameter
       - epochs = 40, lr = [1e-4, 1e-5, 5e-6, 1e-6, 5e-7, 1e-7], batch size = [4, 8, 16, 32, 64]
   - `klue/roberta-large + entity marker`, lr=1e-4, batch size=64 일 때, **test f1-score = 82.759**로 가장 높은 성능을 기록하였다.
   - 총 60가지의 모든 Model, lr, batch size 조합에 대한 결과는 아래의 **5.6. Fine-tuning 결과**에 정리해 두었다.

5. 프로젝트 수행 결과

5.1. 최종 Entities & Relations

- **최종 Relation Map** : [Relation Map]
- **최종 Entities**
  - KLUE-RE 데이터셋에서 사용된 기존의 6개의 entity중 `LOC(위치)` 를 제외한 5개를 사용했고, 주어진 데이터를 잘 표현하기 위해 회의를 통해 결정한 2가지 entity `VIS(유형기술)`, `INV(무형기술)` 를 추가하였다.

Entity	description
PER	사람
ORG	단체
POH	기타 명사
NOH	수량

Entity	description
DAT	날짜
VIS	유형기술
INV	무형기술

- **최종 Relations**
  - 회의를 통해 `단체 중심 관계` 2개, `기술 중심 관계` 14개, `관계_없음` 1개로 총 17개의 relation을 구성하였다.

class_name (ko)	class_name (en)	count	ratio(%)
관계_없음	no_relation	317	36.44
단체 : 제작	org : produce	49	5.63
단체 : 구성원	org : employees	47	5.40
유형기술 : 출시/제작일	vis : release_date/production_date	45	5.17
무형기술 : 출시/제작일	inv : release_date/production_date	52	5.98
무형기술 : 사용자_수	inv : number_of_users	12	1.38
유형기술 : 별칭	vis : alternative_names	17	1.95
무형기술 : 별칭	inv : alternative_names	35	4.02
유형기술 : 기반기술	vis : foundation_technique	16	1.84
무형기술 : 기반기술	inv : foundation_technique	28	3.22
유형기술 : 구성요소	vis : component	41	4.94
무형기술 : 구성요소	inv : component	28	3.22
유형기술 : 기능	vis : function	28	3.22
무형기술 : 기능	inv : function	59	6.78
유형기술 : 제원	vis : specification	8	0.92
유형기술 : 기타속성	vis : property	59	6.78
무형기술 : 기타속성	inv : property	29	3.33

5.2. 최종 가이드라인

최종 가이드라인:

> https://s3-us-west-2.amazonaws.com/secure.notion-static.com/282343b1-64d6-4310-9cae-b87713cacc72/NLP02_KLUE-RE_%EA%B4%80%EA%B3%84_%EC%B6%94%EC%B6%9C_%ED%83%9C%EC%8A%A4%ED%81%AC_%EA%B0%80%EC%9D%B4%EB%93%9C%EB%9D%BC%EC%9D%B8.pdf

5.3. 최종 Dataset

최종 Dataset:

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/ab2f23d6-026a-4f78-ab8f-eb8a1e6f7691/02_dataset.xlsx

5.4. 작업자간 일치도 계산(Fleiss' kappa)

	Case A	Case B
1차 Relation Map 및 가이드라인 기준	0.81	0.717
최종 Relation Map 및 가이드라인 기준	0.87	0.833

Case A : 모든 데이터에 대해 계산한 경우

Case B : 팀원 모두가 no_relation 으로 태깅한 데이터를 제외한 경우

5.5. Fine-tuning 결과

1. Model : **klue/roberta-large**
   - lr = [1e-4, 1e-5, 5e-6, 1e-6, 5e-7, 1e-7], batch size = [4, 8, 16, 32, 64], epochs = 40
   - 모든 lr, batch size의 조합에 대해 40 epochs 동안 학습을 진행하였고, 아래의 표는 각 경우에서의 최고의 test f1-score 를 기록한 것이다.

	batch size: 4	batch size: 8	batch size: 16	batch size: 32	batch size: 64
lr: 1e-4	0	0	0	0	`75.745`
lr: 1e-5	79.325	`78.788`	`78.414`	75.949	73.214
lr: 5e-6	`79.661`	76.923	74.459	`75.983`	73.913
lr: 1e-6	74.236	74.236	69.058	60.784	34.014
lr: 5e-7	69.369	61.611	50.588	8.621	7.018
lr: 1e-7	7.111	7.719	8.163	8.571	7.067

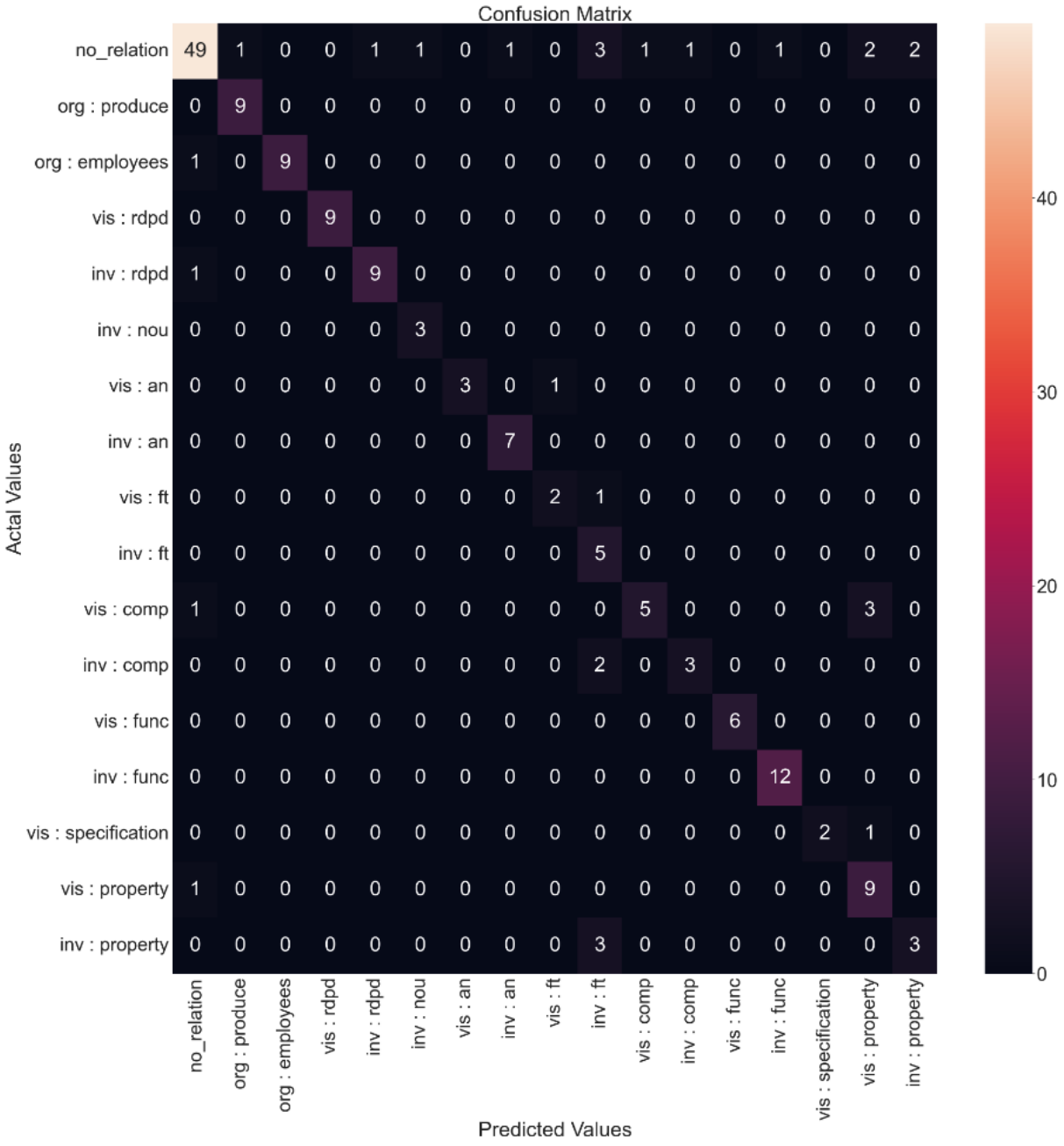
2. Model : **klue/roberta-large** + **entity marker**
   - lr = [1e-4, 1e-5, 5e-6, 1e-6, 5e-7, 1e-7], batch size = [4, 8, 16, 32, 64], epochs = 40
   - 모든 lr, batch size의 조합에 대해 40 epochs 동안 학습을 진행하였고, 아래의 표는 각 경우에서의 최고의 test f1-score 를 기록한 것이다.

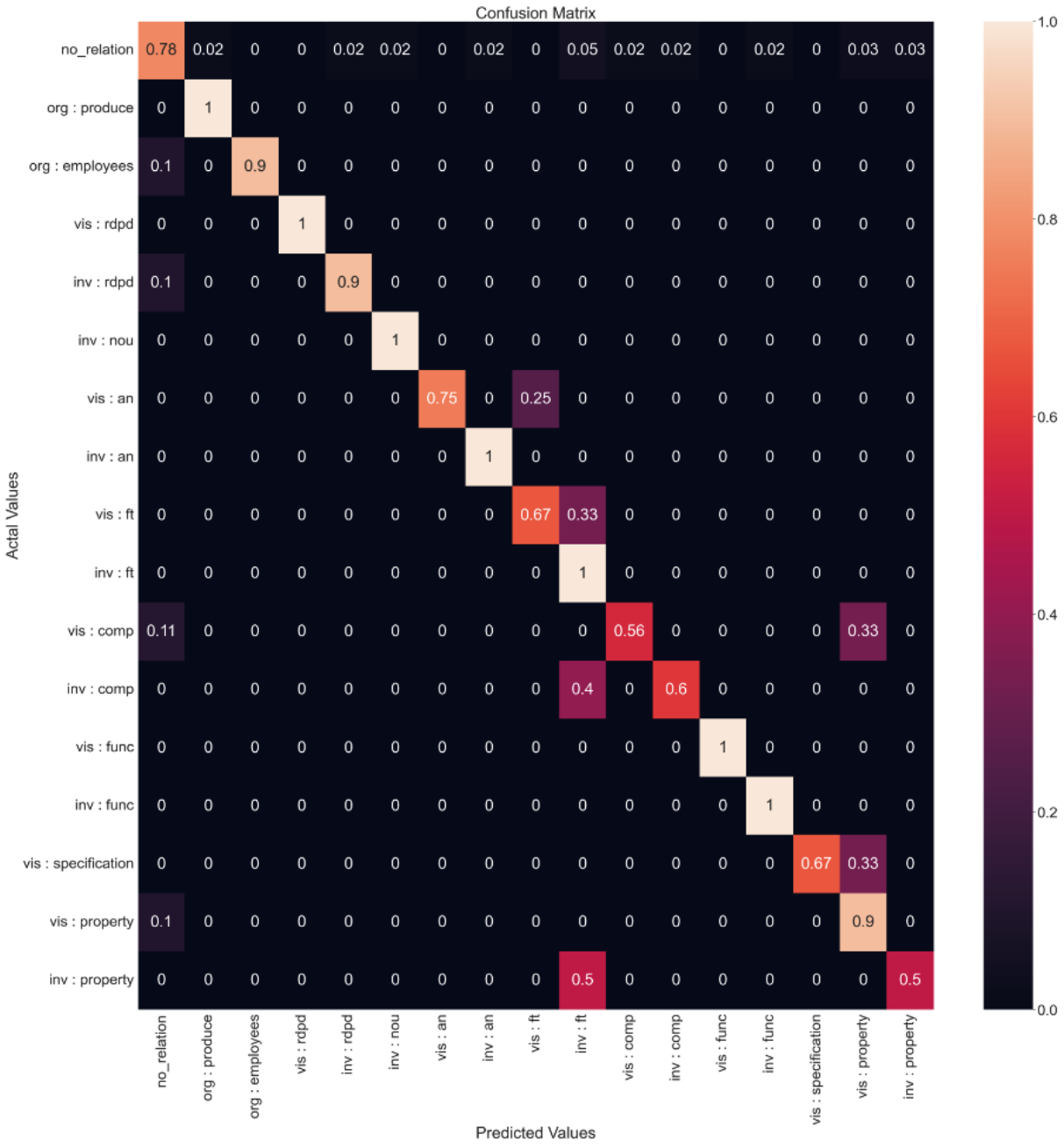
	batch size: 4	batch size: 8	batch size: 16	batch size: 32	batch size: 64
lr: 1e-4	0	0	0	0	`82.759`
lr: 1e-5	`80.349`	`78.448`	`76.19`	`76.19`	68.67
lr: 5e-6	78.414	76.624	68.936	70.042	64.135
lr: 1e-6	71.489	67.826	58.333	59.821	17.054
lr: 5e-7	58.824	53.659	29.176	39.024	8.421

	batch size: 4	batch size: 8	batch size: 16	batch size: 32	batch size: 64
lr: 1e-7	1.429	2.105	8.541	2.105	7.719

5.6. Confusion Matrix

- 가장 높은 test f1-score(=82.759)를 기록한 경우에 대해 confusion matrix를 작성하였다.
  → model=`klue/roberta-large + entity marker`, lr=1e-4, batch size=64, epoch=40





- 전체적으로 잘 예측하였음을 확인할 수 있다.
- `단체 : 제작`, `유형기술 : 출시/제작일`, `무형기술 : 사용자_수`, `무형기술 : 별칭`, `무형기술 : 기반기술`, `유형기술 : 기능`, `무형기술 : 기능` 의 경우 100% 정확히 예측을 하였다.
- `유형기술 : 기반기술` 의 경우 67%는 정확히 예측하였으나, 33%는 `무형기술 : 기반기술` 로 잘못 예측하는 것으로 보아 두 관계 사이의 약간의 모호성이 존재해 보인다.
- `유형기술:구성요소`의 경우 56%는 정확히 예측하였으나, 33%는 `유형기술 : 기타속성` 로 잘못 예측하는 것으로 보아 두 관계 사이의 모호성이 존재해 보인다.
- `무형기술 : 구성요소` 의 경우 60%는 정확히 예측하였으나, 40%는 `무형기술 : 기반기술`로 잘못 예측하는 것으로 보아 두 관계 사이의 약간의 모호성이 존재해 보인다.
- `유형기술 : 제원` 의 경우 67%는 정확히 예측하였으나, 33%는 `유형기술 : 기타속성`로 잘못 예측하는 것으로 보아 두 관계 사이의 약간의 모호성이 존재해 보인다.

- `무형기술 : 기타속성` 의 경우 50%는 정확히 예측하였으나, 50%는 `무형기술 : 기반기술`로 잘못 예측하는 것으로 보아 두 관계 사이의 모호성이 존재해 보인다.

6. 자체 평가 의견

6.1. 프로젝트 달성 요소

- 피드백을 통해 Relation Map 및 가이드라인을 개선하였고, Fleiss' kappa score를 0.81에서 0.87로 올렸다
- 태깅하는 사람마다 관점이 달라질 수 있는 relation에 대해, 명확한 태깅 기준이 되는 "함께 등장하는 단어"를 명시하는 등 가이드라인에 Syntactic한 규칙들을 최대한 자세하게 구축함으로써, rule-base 태깅이 가능하도록 하였다
- 제작한 Data로 Fine-tuning한 모델로 test한 결과 micro_f1_score 82.759점을 달성했다.

6.2. 프로젝트 자체 평가

- **잘한 점**
  - 사람마다 다르게 판단할 수 있는 entity, relation의 구분 기준에 대해(ex.기반기술, 구성요소, 기능), 팀원들이 지속적으로 사례를 들어가며 대화를 통해 더 명확한 태깅 rule을 만들었다.
  - relation(label) 종류를 정할 때 기술에 대한 정확한 정보가 반영되야 한다는 "data 활용의 목적성"을 고려하여 선정함. : 서비스 목적에 맞는 분류 기준 세웠다.
  - klue/roberta-large 사전 모델을 fine-tuning하여 결과를 확인하는데에 그치지 않고, entity를 명시하는 special token을 추가하는 등 모델 구조 및 hyper-parameter tuning을 통해 제작한 데이터에 대한 RE-task 성능을 끌어올렸다.
- **시도 했으나 잘 되지 않았던 것**
  - tagtog 플랫폼 활용 때, relation 추가하는 과정에서 플랫폼 활용이 익숙하지 않아 어려움이 있었다.
- **아쉬웠던 점**
  - 의미없는 no_relation을 더 줄이지 못했다.
  - 주어진 스마트폰 관련 위키피디아 코퍼스의 총 문장이 1190개로 적어, 큰 데이터셋을 만들지 못했다.
- **프로젝트를 통해 배운 점 또는 시사점**
  - 정제돼있지 않은 인터넷상의 데이터를 목적에 맞게 전처리하고 새롭게 제작하는 것은 답이 정해진 문제가 아니라서 팀원들 간 의견 조율이 정말 중요한 것 같다.
  - 팀 내에서 임의로 정한 기준을 klue/roberta-large가 이해해서 문장내의 relation을 구분하는 것이 가능하는 것을 알게 되었다.