Data Annotation Project

8조 MIML

[1] 프로젝트 주제

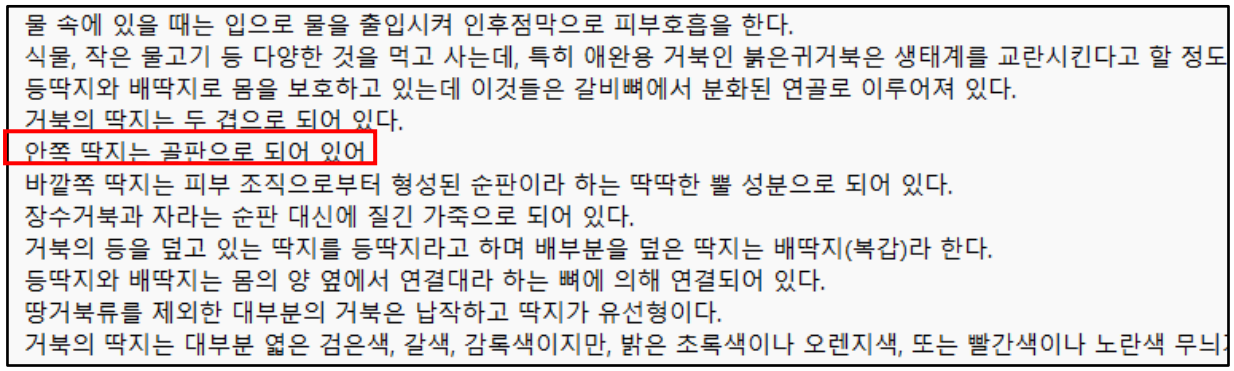
한국어 위키피디아의 반려동물과 관련된 원시 데이터를 활용해 관계 추출 태스크에서 사용할 수 있는 주석된 코퍼스 제작

[2] 프로젝트 팀 구성 및 역할

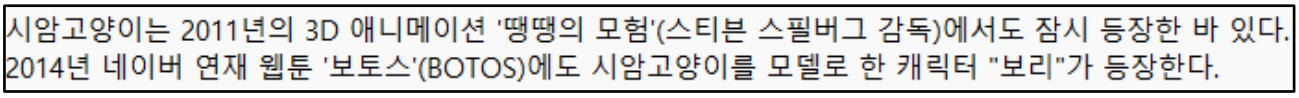
- 임동진 : Data Annotation, IAA 계산
- 정재윤 : Data Annotation, Relation Map 제작
- 조설아 : Data Annotation, Guideline 작성
- 허치영 : Data Annotation, Modeling
- 이보림 : Data Annotation, Guideline 작성

[3] 프로젝트 수행 절차 및 방법

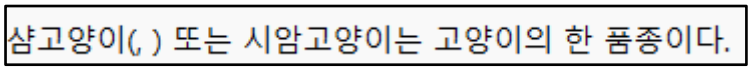
1. 원시 데이터 정제


물 속에 있을 때는 입으로 물을 출입시켜 인후점막으로 피부호흡을 한다.
식물, 작은 물고기 등 다양한 것을 먹고 사는데, 특히 애완용 거북인 붉은귀거북은 생태계를 교란시킨다고 할 정도
등딱지와 배딱지로 몸을 보호하고 있는데 이것들은 갈비뼈에서 분화된 연골로 이루어져 있다.
거북의 딱지는 두 겹으로 되어 있다.
안쪽 딱지는 골판으로 되어 있어
바깥쪽 딱지는 피부 조직으로부터 형성된 순판이라 하는 딱딱한 뿔 성분으로 되어 있다.
장수거북과 자라는 순판 대신에 질긴 가죽으로 되어 있다.
거북의 등을 덮고 있는 딱지를 등딱지라고 하며 배부분을 덮은 딱지는 배딱지(복갑)라 한다.
등딱지와 배딱지는 몸의 양 옆에서 연결대라 하는 뼈에 의해 연결되어 있다.
땅거북류를 제외한 대부분의 거북은 납작하고 딱지가 유선형이다.
거북의 딱지는 대부분 엷은 검은색, 갈색, 감록색이지만, 밝은 초록색이나 오렌지색, 또는 빨간색이나 노란색 무늬

문장이 너무 짧아 entity 선정이 불가능한 경우 및 불완전한 문장은 하나의 문장으로 합쳤습니다.


시암고양이는 2011년의 3D 애니메이션 '땡땡의 모험'(스티븐 스필버그 감독)에서도 잠시 등장한 바 있다.
2014년 네이버 연재 웹툰 '보토스'(BOTOS)에도 시암고양이를 모델로 한 캐릭터 "보리"가 등장한다.

프로젝트 주제인 "반려동물"과 연관이 없는 문장의 경우 문장을 제외하거나 entity로 선정하지 않았습니다.

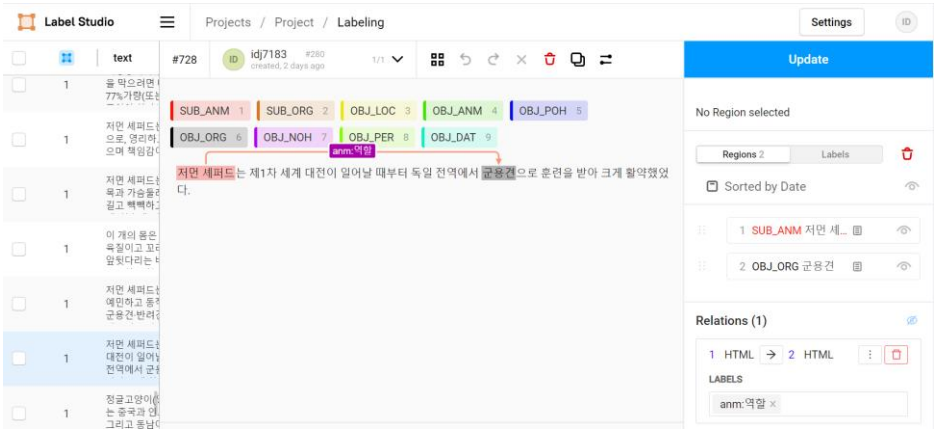

샴고양이(,) 또는 시암고양이는 고양이의 한 품종이다.

학명 또는 외국어의 경우 내용이 누락되고 괄호만 남는 경우가 존재하여 확인 후 제거했습니다.

데이터 정제 결과 총 2193개 문장에서 **1755개**의 문장으로 데이터 양이 줄었습니다.

2. Annotation Tool

- Label Studio



- tagtog

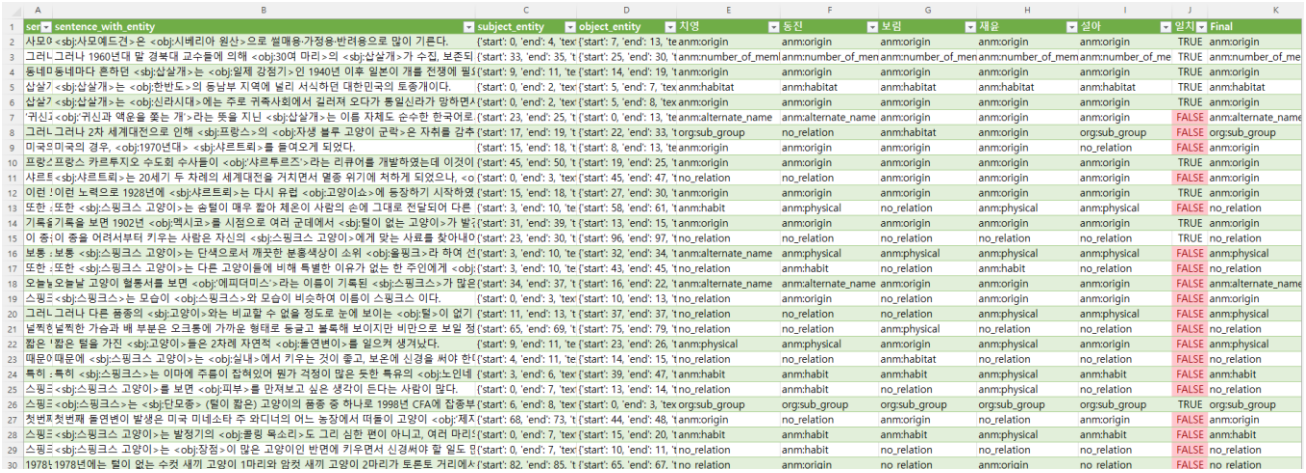
총 두 가지 Annotation Tool로 정제된 데이터에 대해 엔티티와 관계를 지정했습니다.

3. relation map 초안 작성

총 12개의 relation을 선정하여 초안을 작성했습니다.

[동물:서식지] [동물:대체표현] [동물:신체적 특징] [동물:질병] [동물:비신체적 특징] [동물:사냥감] [동물:역할] [동물:유래] [동물:개체수] [집단:상위집단] [집단:하위집단] [관계없음]

4. Pilot tagging



ser	sentence_with_entity	subject_entity	object_entity	지영	동진	보람	재윤	셀아	일치	Final
2	사모<sbj:사모예드견>은 <obj:시베리아 원산>으로 썰매용·가정용·반려용으로 많이 기른다.	{'start': 0, 'end': 4, 'tex	{'start': 7, 'end': 13, 'te	anm:origin	anm:origin	anm:origin	anm:origin	anm:origin	TRUE	anm:origin
3	그러나 그러나 1960년대 말 경복대 교수들에 의해 <obj:30여 마리>의 <sbj:삽살개>가 수집, 보존되	{'start': 33, 'end': 35, 't	{'start': 25, 'end': 30, 't	anm:number_of_mem	anm:number_of_mem	anm:number_of_mem	anm:number_of_mem	anm:number_of_me	TRUE	anm:number_of_me
4	동네마동네마다 흔하던 <sbj:삽살개>는 <obj:일제 강점기>인 1940년 이후 일본이 개를 전쟁에 필	{'start': 9, 'end': 11, 'te	{'start': 14, 'end': 19, 't	anm:origin	anm:origin	anm:origin	anm:origin	anm:origin	TRUE	anm:origin
5	삽살개 <sbj:삽살개>는 <obj:한반도>의 동남부 지역에 널리 서식하던 대한민국의 토종개이다.	{'start': 0, 'end': 2, 'tex	{'start': 5, 'end': 7, 'tex	anm:habitat	anm:habitat	anm:habitat	anm:habitat	anm:habitat	TRUE	anm:habitat
6	삽살개 <sbj:삽살개>는 <obj:신라시대>에는 주로 귀족사회에서 길러져 오다가 통일신라가 망하면서	{'start': 0, 'end': 2, 'tex	{'start': 5, 'end': 8, 'tex	anm:origin	anm:origin	anm:origin	anm:origin	anm:origin	TRUE	anm:origin
7	'귀신과 액운을 쫓는 개'라는 뜻을 지닌 <sbj:삽살개>는 이름 자체도 순수한 한국어로	{'start': 23, 'end': 25, 't	{'start': 0, 'end': 13, 'te	anm:alternate_name	anm:alternate_name	anm:origin	anm:origin	anm:origin	FALSE	anm:alternate_name
8	그러나 그러나 2차 세계대전으로 인해 <sbj:프랑스>의 <obj:자생 볼루 고양이 군락>은 자취를 감추	{'start': 17, 'end': 19, 't	{'start': 22, 'end': 33, 't	org:sub_group	no_relation	anm:habitat	anm:origin	org:sub_group	FALSE	org:sub_group
9	미국의미국의 경우, <obj:1970년대> <sbj:샤르트뢰>를 들여오게 되었다.	{'start': 15, 'end': 18, 't	{'start': 8, 'end': 13, 'te	anm:origin	anm:origin	anm:origin	anm:origin	no_relation	FALSE	anm:origin
10	프랑스프랑스 카르투지오 수도회 수사들이 <obj:'샤르투르즈'>라는 리큐어를 개발하였는데 이것이	{'start': 45, 'end': 50, 't	{'start': 19, 'end': 25, 't	anm:origin	anm:origin	anm:origin	anm:origin	anm:origin	TRUE	anm:origin
11	샤르트<sbj:샤르트뢰>는 20세기 두 차례의 세계대전을 거치면서 멸종 위기에 처하게 되었으나, <o	{'start': 0, 'end': 3, 'tex	{'start': 45, 'end': 47, 't	no_relation	no_relation	anm:origin	anm:origin	anm:origin	FALSE	anm:origin
12	이런 !이런 노력으로 1928년에 <sbj:샤르트뢰>는 다시 유럽 <obj:고양이쇼>에 등장하기 시작하였	{'start': 15, 'end': 18, 't	{'start': 27, 'end': 30, 't	anm:origin	anm:origin	anm:origin	anm:origin	anm:origin	TRUE	anm:origin
13	또한 ;또한 <sbj:스핑크스 고양이>는 솜털이 매우 짧아 체온이 사람의 손에 그대로 전달되어 다른	{'start': 3, 'end': 10, 'te	{'start': 58, 'end': 61, 't	anm:habit	anm:physical	no_relation	anm:physical	anm:physical	FALSE	no_relation
14	기록§기록을 보면 1902년 <obj:멕시코>를 시점으로 여러 군데에서 <sbj:털이 없는 고양이>가 발견	{'start': 31, 'end': 39, 't	{'start': 13, 'end': 15, 't	anm:origin	anm:origin	anm:origin	anm:origin	anm:origin	TRUE	anm:origin
15	이 종¦이 종을 여러서부터 키우는 사람은 자신의 <sbj:스핑크스 고양이>에게 맞는 사료를 찾아내어	{'start': 23, 'end': 30, 't	{'start': 96, 'end': 97, 't	no_relation	no_relation	no_relation	no_relation	no_relation	TRUE	no_relation
16	보통 ;보통 <sbj:스핑크스 고양이>는 단색으로서 깨끗한 분홍색상이 소위 <obj:올핑크>라 하여 선	{'start': 3, 'end': 10, 'te	{'start': 32, 'end': 34, 't	anm:alternate_name	anm:physical	anm:physical	anm:physical	anm:physical	FALSE	anm:physical
17	또한 ;또한 <sbj:스핑크스 고양이>는 다른 고양이들에 비해 특별한 이유가 없는 한 주인에게 <obj:(	{'start': 3, 'end': 10, 'te	{'start': 43, 'end': 45, 't	no_relation	anm:habit	no_relation	anm:habit	no_relation	FALSE	no_relation
18	오늘늘오늘날 고양이 혈통서를 보면 <obj:'에피더미스'>라는 이름이 기록된 <sbj:스핑크스>가 많은	{'start': 34, 'end': 37, 't	{'start': 16, 'end': 22, 't	anm:alternate_name	anm:alternate_name	anm:origin	anm:origin	anm:origin	FALSE	anm:alternate_name
19	스핑ㅌ<sbj:스핑크스>는 모습이 <obj:스핑크스>와 모습이 비슷하여 이름이 스핑크스 이다.	{'start': 0, 'end': 3, 'tex	{'start': 10, 'end': 13, 't	no_relation	anm:origin	no_relation	anm:origin	anm:origin	FALSE	anm:origin
20	그러나 그러나 다른 품종의 <sbj:고양이>와는 비교할 수 없을 정도로 눈에 보이는 <obj:털>이 없기	{'start': 11, 'end': 13, 't	{'start': 37, 'end': 37, 't	no_relation	no_relation	anm:physical	anm:physical	anm:physical	FALSE	anm:physical
21	넓찍넓찍한 가슴과 배 부분은 오크통에 가까운 형태로 등글고 볼록해 보이지만 비만으로 보일 정	{'start': 65, 'end': 69, 't	{'start': 75, 'end': 79, 't	no_relation	no_relation	anm:physical	no_relation	no_relation	FALSE	no_relation
22	짧은 '짧은 털을 가진 <sbj:고양이>들은 2차례 자연적 <obj:돌연변이>를 일으켜 생겨났다.	{'start': 9, 'end': 11, 'te	{'start': 23, 'end': 26, 't	anm:physical	anm:physical	anm:origin	anm:physical	anm:origin	FALSE	anm:physical
23	때문이때문에 <sbj:스핑크스 고양이>는 <obj:실내>에서 키우는 것이 좋고, 보온에 신경을 써야 한다	{'start': 4, 'end': 11, 'te	{'start': 14, 'end': 15, 't	no_relation	anm:habitat	no_relation	no_relation	no_relation	FALSE	no_relation
24	특히 ;특히 <sbj:스핑크스>는 이마에 주름이 잡혀있어 뭔가 걱정이 많은 듯한 특유의 <obj:노인네	{'start': 3, 'end': 6, 'tex	{'start': 39, 'end': 47, 't	anm:habit	anm:habit	anm:habit	anm:physical	anm:habit	FALSE	anm:habit
25	스핑ㅌ<sbj:스핑크스 고양이>를 보면 <obj:피부>를 만져보고 싶은 생각이 든다는 사람이 많다.	{'start': 0, 'end': 7, 'tex	{'start': 13, 'end': 14, 't	no_relation	anm:habit	no_relation	anm:physical	no_relation	FALSE	no_relation
26	스핑ㅌ<obj:스핑크스>는 <sbj:단모종> (털이 짧은) 고양이의 품종 중 하나로 1998년 CFA에 잡종부	{'start': 6, 'end': 8, 'tex	{'start': 0, 'end': 3, 'tex	org:sub_group	org:sub_group	org:sub_group	org:sub_group	org:sub_group	TRUE	org:sub_group
27	첫번ㅈ첫번째 돌연변이 발생은 미국 미네소타 주 와디너의 어느 농장에서 떠돌이 고양이 <obj:제주	{'start': 68, 'end': 73, 't	{'start': 44, 'end': 48, 't	anm:origin	no_relation	anm:origin	no_relation	anm:origin	FALSE	anm:origin
28	스핑ㅌ<sbj:스핑크스 고양이>는 발정기의 <obj:롤링 목소리>도 그리 심한 편이 아니고, 여러 마리	{'start': 0, 'end': 7, 'tex	{'start': 15, 'end': 20, 't	anm:habit	anm:habit	anm:habit	anm:physical	anm:habit	FALSE	anm:habit
29	스핑ㅌ<sbj:스핑크스 고양이>는 <obj:장점>이 많은 고양이인 반면에 키우면서 신경써야 할 일도 많	{'start': 0, 'end': 7, 'tex	{'start': 10, 'end': 11, 't	no_relation	anm:habit	no_relation	no_relation	no_relation	FALSE	no_relation
30	1978년 1978년에는 털이 없는 수컷 새끼 고양이 1마리와 암컷 새끼 고양이 2마리가 토론토 거리에서	{'start': 82, 'end': 85, 't	{'start': 65, 'end': 67, 't	no_relation	anm:origin	no_relation	anm:origin	no_relation	FALSE	no_relation

전체 데이터 중 약 10%를 Pilot tagging을 위해 사용했습니다. 각자 임의로 tagging한 후 취합하여 상이한 결과를 나타내는 데이터에 대해서 논의를 진행한 후 가이드라인을 수정했습니다.

5. 최종 Entity Type, Relation Map 도출

1) Entity Type (총 7개)

[ANM]: 동물 **[ORG]**: 집단 **[POH]**: 기타 **[NOH]**: 수치

[PER]: 인물 **[LOC]**: 지역 **[DAT]**: 날짜

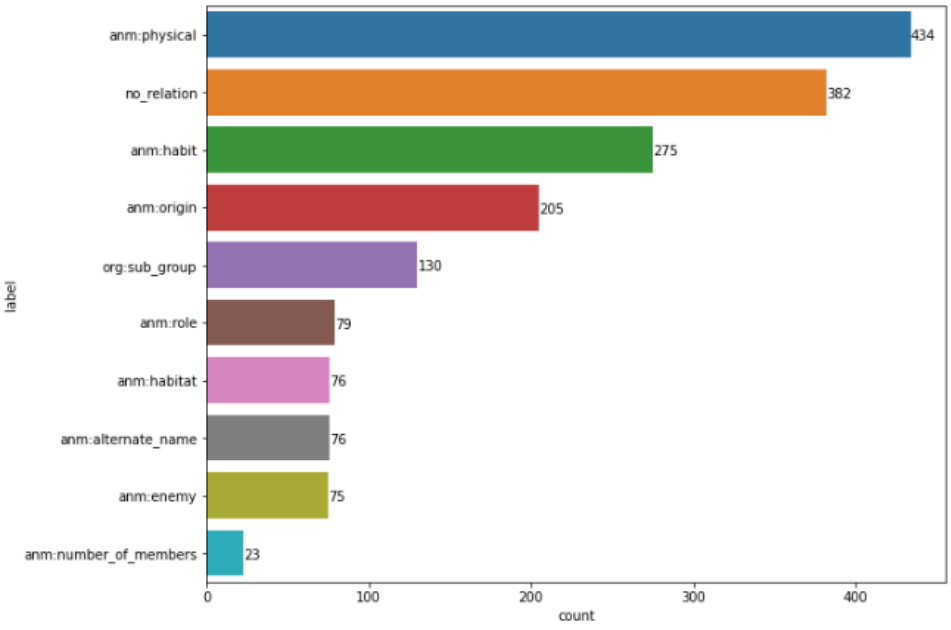
+) DAT : 수량 표현(NOH)과 뚜렷하게 구분되는 특성을 갖고 있기에 분리해서 사용했습니다.

2) Relation map (총 10개)

- **동물:서식지** (anm: habitat) → (ANM, LOC)
- **동물:대체표현** (anm: alternate_name) → (ANM, ANM/POH/ORG)
- **동물:신체적 특징** (anm: physical) → (ANM, POH/NOH)
- **동물:비신체적 특징** (anm: habit) → (ANM, POH/NOH/DAT)
- **동물:사냥감** (anm: enemy) → (ANM, ANM/POH)
- **동물:역할** (anm: role) → (ANM, ORG/POH)
- **동물:유래** (anm: origin) → (ANM, ANM/ORG/POH/PER/LOC)
- **동물:개체수** (anm: number_of_members) → (ANM, NOH)
- **집단:하위집단** (org: sub_group) → (ORG, ANM/ORG)
- **관계 없음** (no_relation) → (ANM/ORG, *)

6. 모델 성능, Fleiss' Kappa

1) 데이터셋 분포



< Data Distribution >

2) Inter Annotator Agreement (IAA) Score

Fleiss' Kappa Score (Pilot Tagging) : 0.72

Fleiss' Kappa Score (Final Tagging) : **0.79**

3) 모델 학습 성능 (KLUE/BERT-base)

	Only pre-training	With fine-tuning	KLUE Benchmark
Micro F1	12.7119	57.2581	66.44
AUPRC	17.1305	60.4493	66.17

Hyperparameters

Epochs : 3 Batch Size : 2 Learning Rate : 2e-5 Evaluation Step : 60

[4] 자체 평가 의견

Relation Map 설정 시 anm:origin에 해당하는 관계의 기준을 다소 모호하게 잡았던 것과 너무 많은 범위를 한 관계에 포함시키려 했던 점이 아쉬웠습니다. 또한 anm:physical의 관계를 갖는 데이터가 no_relation보다 많이 등장한 것이 아쉬웠습니다. anm:disease와 나누어 사용했다면 더 좋은 분포를 얻을 수 있었을 것이라는 생각이 들었습니다.

데이터 클래스 간의 불균형을 해소하고자 했으나 완전히 해소하지 못한 것이 아쉬웠습니다. 이러한 데이터 불균형 또한 데이터의 특징 중 하나로 간주하기로 생각했습니다.

KLUE/BERT-base의 리더보드 성능과 비교했을 때, 데이터셋의 크기가 약 20배 차이남에도 불구하고 큰 성능 차이를 보이지 않았음에 쓸 만한 데이터를 제작했다고 생각됩니다. 데이터셋의 크기를 비슷한 수준까지 끌어올린다면 KLUE 벤치마크 점수와 비슷한 수준의 성능을 기대할 수 있다고 생각됩니다.