

[Wrap-up Report] 데이터 제작 - KiYOUNG2

1. 프로젝트 개요

▼ 1.1 프로젝트 주제

- 관계 추출 데이터 제작

한국어 및 다른 언어에서의 자연어처리 데이터셋의 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 목표로 합니다.

▼ 1.2 프로젝트 일정

- 일정 : 2021.11.08 14:00 ~ 2021.11.19 12:00

▼ 1.3 평가 지표

- IAA - Fleiss' Kappa
- 가이드라인 및 relation set 정의의 정밀성에 대한 정성평가

▼ 1.4 작업 환경

- GPU - V100(32gb)
- Annotation tool 1 - tagtog

Projects / eliza-dukim/RE-NLP-14 / 2045.txt

Settings Documents Metrics Downloads



















+ Content  master 

pool

1st_complete

편집성 인격장애 환자는 지속적으로 **원인**을 품는다. (모욕, 상해, 경멸을 용서하지 않음)

- Annotation tool 2 - Google Sheet

[NLP-14]심리학 RE 데이터 제작 ☆ 							
파일 수정 보기 삽입 서식 데이터 도구 확장 프로그램 도움말 <small>인명의 링크님이 몇 초 전에 마지막으로 수정했습니다.</small>							
F149     100%  10  arial           							
1	A	B	C	D	F	H	I
	id	sentence	sub_tag	obj_tag	label_worker2	data_error	hate_bias_p
116	727	타르드는 군중과는 엄격히 구별한 _{공중}의 개념을 생각해내 <obj> THRY	THRY	이론-상위_이론			
118	960	_{인간 소여 치료}(human givens therapy)는 안정(security), 자 THRY	THRY	이론-상위_이론		1	
122	1346	_{교육 방식}는 간병인이 학교 시스템에 적극적으로 참여할 수 있는 THRY	THRY	이론-상위_이론			
123	1503	몸 자신은, <obj>과학이론</obj>으로서 신중하게 이론을 구성했지만, 그것은 < THRY	THRY	관계_없음			
127	2115	부정망상(不實妄想, delusion of infidelity 또는 delusional jealousy) 또는 <si THRY	THRY	이론-대체어			
149	268	<obj>인지행동치료</obj>는 _{노출치료}(exposure therapy), 스 THRY	THRY				
150	281	<obj>신경생물학 방법</obj>(Neurobiological methods)에 해당하는 연구 방 THRY	THRY				
153	316	4단계 _{성숙 병주}는 정서가 건강한 성인에게서 보인다. 이 <obj> THRY	THRY	관계_없음			
162	435	<obj>정신질환 진단 및 통계 편람</obj>(DSM)-IV에서는 3개의 부류로 나뉘는 THRY	THRY	이론-대체어			
167	530	<obj>성숙기</obj>(genital stage, 12세 이후)는 사춘기로서 성적 충동을 정상 THRY	THRY	이론-상위_이론			
168	535	<obj>통찰 치료</obj>에도 여러 유형이 있고 이들이 사용하는 치료 방법도 매우 THRY	THRY	이론-하위_이론			

2. 팀 구성 및 역할

▼ 2.1 팀원 역할

진명훈 : IAA 계산
 김대웅 : 가이드라인 작성, Tagtog 플랫폼에 문장 업로드 및 취합
 김태욱 : Relation map 작성
 허진규 : 모델 튜닝
 이하람 : Relation map 작성
 김채은 : 가이드라인 작성
 유영재 : 모델 튜닝

3. 프로젝트 계획

▼ 3.1 일정

- 1주차 : 데이터 파악, entity 결정 및 가이드라인 작성
- 2주차 : annotation 작업, IAA 계산 및 모델 학습

▼ 3.2 목표

- NLP 데이터 제작의 단계별 흐름 이해
- NLP task 별 데이터의 특성 이해
- 실제 RE 데이터 제작 실습을 통한 NLP 데이터 제작 체험

4. 프로젝트 수행 결과

▼ 4.1 Relation map 제작

no	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	이론:대체어	thry:alternate_names	(THRY, THRY)	이론(SUBJ)을 대체할 수 있는 명칭(OBJ)
3	이론:상위_이론	thry:theory_of	(THRY, THRY)	이론(SUBJ)을 포함하는 상위 이론(OBJ) (SUBJ < OBJ)
4	이론:하위_이론	thry:theory	(THRY, THRY)	이론(SUBJ)에 포함되는 하위 이론(OBJ) (SUBJ > OBJ)
5	이론:상위_학문분야	thry:studyfield	(THRY, STF)	이론(SUBJ)을 포함하는 학문분야(OBJ) (SUBJ < OBJ)
6	학문분야:하위_이론	stf:theory	(STF, THRY)	학문분야(SUBJ)에 포함되는 이론(OBJ) (SUBJ > OBJ)
7	인물:소속이론또는학문분야	per:affiliation	(PER, THRY/STF)	인물(SUBJ)이 관련된 이론(OBJ) 또는 학문분야(OBJ)
8	용어:치료기법	term:therapy	(TERM, THRY)	질환(SUBJ)에 적용되는 치료기법(OBJ)
9	용어:약	term:medicine	(TERM, TERM)	질환(SUBJ)에 적용되는 약(OBJ)
10	용어:증상또는질환	term:disorder	(TERM, TERM)	질환(SUBJ)에 수반되는 증상(OBJ), 질환(SUBJ)과 동반되는 질환(OBJ)
11	용어:대체어	term:alternate_names	(TERM, TERM)	질환(SUBJ)을 대체할 수 있는 명칭(OBJ)
	cf. 학문분야 : -학, -주의, -학파			
	cf. 이론 : 이론, 개념, 연구 방법론			
	cf. 용어 : 약물, 질병, 증상 한정			

▼ 4.2 데이터 제작 가이드라인 제작

• 데이터 제작 가이드라인

1. 심리학 분야의 관계 추출 태스크 알아보기

1-1 관계 추출 태스크 설명

1-2 관계의 소개

1-3 예시

2. Annotation 가이드라인

2-1 외부 지식의 활용

2-2 관계 없음

3. Annotation 환경

3-1. 관계 클래스 선택지

3-2. Data Error

3-3. Hate / Bias / Privacy

4. 예시 & FAQ

• 예시

1-3) 예시

관계 추출 태스크에 대한 이해를 돕기 위해 다음의 예시를 살펴보세요.
색상 : **Subject** / **Object**

Example #1:

인지주의 심리학의 가장으로는 **알버트 엘리스** 이론 역동이 거론된다.

위의 문장에서 Entity 쌍을 이루는 Subject Entity는 "**알버트 엘리스**"이며, Object Entity는 "**인지주의 심리학**"입니다. 의미적 관계는 "**인물:소속이론또는학문분야**" 클래스로 분류할 수 있습니다.

Example #2:

편집증(偏執症) 또는 **파라노이아**(영어: paranoia)는 심각한 걱정이나 두려움으로 자신이 주변으로부터 피해를 받을 것이라는 병리적인 의심을 고집하는 이상심리학적 상태를 일컫는다.

위의 문장에서 Entity 쌍을 이루는 Subject Entity는 "**편집증**"이며, Object Entity는 "**파라노이아**"입니다. 의미적 관계는 "**용어:대체어**" 클래스로 분류할 수 있습니다.

• FAQ

4. 자주 묻는 질문

Q1. 학파나 이론의 창시자 또는 저술의 자라도 "**인물:소속이론또는학문분야**"에 해당하나요?

· 네, 해당합니다. 학파의 창시자나 학파와 관련된 저술의 저자 또한 해당 학파의 일원으로 볼 수 있으므로 "**인물:소속이론또는학문분야**" 클래스로 분류합니다.

Example #6:

호프만(Hofmann)은 근대 **인지행동치료** 접근법을 간략 기술하였다

Q2. 특정 이론이 정신분석학과 같은 학문에 속할 경우 "**이론:상위_학문분야**"에 해당되는지 "**이론:상위_이론**"에 해당되는지 궁금합니다.

· "**이론:상위_학문분야**"에 해당합니다. "**이론:상위_이론**"의 경우, subject entity와 object entity 모두 이론에 해당하며, "**이론:상위_학문분야**"의 경우, subject entity는 이론이지만 object entity는 이론보다 상위 개념인 학문입니다.

Example #7:

1879년, 존의 "심리학의 아버지"라 불리는 볼리는 보토는 라이프치히 대학에 첫 심리학 연구소인 정신물리실험실을 개설하였다. 그는 **실험법**을 연구하는 방법론으로 **내성법**을 주장하였다.

▼ 4.3 데이터 제작 결과물

- Tagtag에서 subject 및 object entity, relation태깅한 것을 엑셀로 받아와 labeling작업 수행
 - 작업자마다 본인이 Tagtag에서 태깅한 것 이외의 데이터에 대해 labeling

	id	sentence	sub_tag	obj_tag	label
0	0	<subj>망상형 조현병</subj> 환자는 주로 자신의 망상과 관련된 내용의 환각...	TERM	TERM	용어:증상또는질환
1	1	<subj>합리화</subj>는 실망을 주는 현실에서 도피하기 위해 그럴듯한 구실을...	THRY	THRY	이론:상위_이론
2	3	<subj>자기애성 인격장애</subj> 환자는 자신의 중요성에 대해 지나치게 <o...	TERM	TERM	용어:증상또는질환
3	7	<subj>자기애성 인격장애</subj>가 심한 경우에는 자기를 향한 <obj>공격...	TERM	TERM	용어:증상또는질환
4	8	<subj>정서적 방치</subj>는 양육, 격려 및 지원 부족이 특징인 <obj>...	THRY	THRY	이론:상위_이론
...
901	2498	<subj>포기</subj>는 부모 또는 보호자가 베이비 시터 나 보호자 없이 오랫동안...	THRY	THRY	이론:상위_이론
902	2499	<subj>코헛</subj>이 설명하였듯, 자기대상이 자기에게 하는 일을 뜻하는 <...>	PER	THRY	인물:소속이론또는학문분야
903	2500	이러한 <subj>강박 장애</subj>는 현재로서는 강박사고의 침습적인 반복으로 ...	TERM	TERM	용어:증상또는질환
904	2501	당시 뱀장어의 생애 주기는 아직 알려지지 않은 상태였다. <subj>프로이트</su...	PER	STF	인물:소속이론또는학문분야
905	2503	<subj>정신분석학</subj>에서 말하는 것 같은 별개의 영역으로서의 무의식 개...	STF	THRY	학문분야:하위_이론

▼ 4.4 Fleiss' Kappa 결과

총계	7명 일치	6명 일치	5명 일치	4명 일치	3명 일치	2명 일치	1명 일치
906	558 (61.5%)	156 (17.2%)	86 (9.49%)	77 (8.49%)	27 (2.98%)	2 (0.22%)	0 (0%)

이번 과정에서 Fleiss Kappa Score가 70% 이상이 나오는 것이 목적이었습니다.

- 단순히 생각했을 때 sample별 전원 일치 비율이 70% 이상

심리학이라는 어려운 주제를 잡아서 낮은 평가 점수가 나올 것을 우려하였으나 우수한 가이드라인 작성과 팀원들의 작업 일치로 높은 평가 점수(0.8)를 기록하여 양질의 데이터를 구축하는 데 성공했습니다.

- 아래 사진은 fleiss kappa 계산에 입력으로 넣어준 정형 데이터 포맷

	label_worker1	label_worker2	label_worker3	label_worker4	label_worker5	label_worker6	label_worker7	sentence
0	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	일찍이 플라톤과 _j아리스토텔레스</sub>는 <obj>인식론</obj>에...
1	이론:상위_이론	이론:상위_이론	이론:상위_이론	이론:상위_이론	이론:대체어	이론:상위_이론	이론:대체어	프로이트 이후 직/간접적으로 그의 영향을 받은 수 많은 정신분석가들이 배출되었으며, ...
2	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	예를 들면, <obj>언어 습득 분야</obj>에서 _j 스티븐 핑커(Stev...
3	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	_j프로이트</sub>는 자기 분석을 계속하여 지 금까지 수집한 자료들을 모...
4	이론:상위_이론	이론:상위_이론	이론:상위_이론	이론:상위_이론	이론:상위_이론	이론:상위_이론	이론:상위_이론	브뤼케의 '생리학 강의'에서 그는 살아있는 유기체는 하나의 역학계이며 화학과 물리학...
...
901	이론:상위_이론	이론:상위_이론	이론:대체어	이론:상위_이론	이론:상위_이론	이론:상위_이론	이론:대체어	<obj>인지과학</obj>에서의 _j언어 처리 연 구</sub>는 언어학 ...
902	인물:소속이론 또는학문분야	인물:소속이론 또는학문분야	인물:소속이론 또는학문분야	인물:소속이론 또는학문분야	인물:소속이론 또는학문분야	인물:소속이론 또는학문분야	인물:소속이론 또는학문분야	빈약한 언어와 무감정, 사회적 활동의 <obj>위축 증 상</obj>은 다른 사람으로...
903	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	용어:증상또는 질환	⑥ _j주지화</sub>(<obj>지성화</obj>): 위 협적인 대상에 대...
904	인물:소속이론 또는학문분야	인물:소속이론 또는학문분야	관계 없음	인물:소속이론 또는학문분야	인물:소속이론 또는학문분야	인물:소속이론 또는학문분야	인물:소속이론 또는학문분야	Lewandowski와 Strohmets는 2009년에 심리학에서 사용하는 <obj>...
905	학문분야:하위 이론	이론:하위_이론	학문분야:하위 이론	학문분야:하위 이론	학문분야:하위 이론	학문분야:하위 이론	학문분야:하위 이론	_j포기</sub>는 부모 또는 보호자가 베이비 시터 나 보조자 없이 오랫동안...

- 아래 사진은 fleiss kappa 계산 결과

```

kappa = fleissKappa(transformed_result,len(result[0]))
[105] ✓ 0.3s

... #raters = 7 , #subjects = 906 , #categories = 11
PA = 0.8270787343635013
PE = 0.14815244110011885
Fleiss' Kappa = 0.797

```

▼ 4.5 모델 학습 결과

- train test split (8:2, stratify)

```
from sklearn.model_selection import train_test_split

X_train, X_dev, y_train, y_dev = train_test_split(
    result[[col for col in result.columns if col != "label"]],
    result.label,
    test_size=0.2,
    stratify=result.label
)

pd.concat([X_train, y_train], axis=1).to_csv("dataset/train/train.csv")
pd.concat([X_dev, y_dev], axis=1).to_csv("dataset/train/dev.csv")
```

- train: dev: test ratio

relation	train		dev		test	
	count	ratio	count	ratio	count	ratio
no_relation	107	14.77900552	13	14.28571429	14	15.38461538
thry:alternate_names	81	11.1878453	10	10.98901099	11	12.08791209
thry:theory_of	61	8.425414365	7	7.692307692	8	8.791208791
thry:theory	36	4.972375691	5	5.494505495	4	4.395604396
thry:studyfield	24	3.314917127	3	3.296703297	3	3.296703297
stf:theory	10	1.38121547	1	1.098901099	1	1.098901099
per:affiliation	123	16.98895028	16	17.58241758	15	16.48351648
term:therapy	39	5.386740331	5	5.494505495	5	5.494505495
term:medicine	18	2.486187845	2	2.197802198	2	2.197802198
term:disorder	188	25.96685083	24	26.37362637	23	25.27472527
term:alternate_names	37	5.110497238	5	5.494505495	5	5.494505495
	724	100	91	100	91	100

- dev set 예측 결과

모델의 예측 결과 또한 klue re 태스크의 bert-base 결과와 유사한 점수가 나왔습니다.

```
wandb: Run summary:
wandb:           eval/accuracy 0.64835
wandb:           eval/auprc 66.28294
wandb:           eval/loss 1.59196
wandb:           eval/micro f1 score 70.03155
wandb:           eval/runtime 0.6912
wandb:           eval/samples_per_second 263.295
wandb:           eval/steps_per_second 17.36
wandb:           train/epoch 20.0
wandb:           train/global_step 920
wandb:           train/learning_rate 0.0
wandb:           train/loss 0.0014
wandb:           train/total_flos 1547875852823040.0
wandb:           train/train_loss 0.46138
wandb:           train/train_runtime 277.8944
wandb:           train/train_samples_per_second 52.106
wandb:           train/train_steps_per_second 3.311
```

5. Conclusion

▼ 5.1 데이터 제작을 경험하고 각자 느낀 부분

- 김대웅
 - 각 개념간의 관계를 표현하기에 적절한 구조를 설계하는 일이 쉽지 않다는 것을 경험했습니다.
- 김채은
 - 초반에 데이터를 전체적으로 본 후 entity 및 class를 고민할 때 그 두 가지를 매칭해서 생각했는데, 추후 entity와 class를 분리하여 개념을 확립해야함을 배웠습니다. 가이드라인을 작성할 때, 우리가 고민했던 것을 작업자들에게 잘 전달하기 위해서는 예시 등을 사용하여 구체적으로 안내하는 것이 중요하다고 생각했습니다.
- 김태욱
 - 이미 제작된 데이터를 받아서 모델링을 진행할때는 소중함을 느끼지 못했는데 직접 제작을 해보니 명확히 구분되고 헷갈리지 않게 labeling을 하는 것이 정말 어렵다는 것을 느꼈습니다.
- 유영재
 - RE 데이터 제작 난이도가 높다는 것과 특정 주제에 대한 데이터 제작을 진행할 경우 도메인에 대한 이해가 중요하다는 것을 느꼈습니다.
- 이하람
 - 파일럿 태깅의 중요성을 깨달았습니다. 직접 태깅을 하면서 데이터의 예시를 살펴본 후에 작업 가이드라인을 맞춰가는 것이 제일 효율적이라는 것을 깨달았습니다.
- 진명훈
 - 누군가 만들어둔 데이터로 모델링을 하는 것은 상대적으로 굉장히 쉽다는 것을 깨달았습니다. 양질의 데이터를 확보하고 구축하는데 정말 많은 이들의 노력이 필요하다고 느꼈으며 향후 현업/연구직에 종사하며 직접 구축할 기회가 생기면 2주 동안의 경험을 살려서 체계적으로 가이드라인을 구축하고 예외 케이스를 잡아서 작업자들에게 배포해야겠습니다.
- 허진규
 - 막연하게 '데이터 제작을 하면 된다.' 라고 생각을 했었는데, 실제로 데이터 제작 과정을 진행해보니 굉장히 고려해야 될 부분도 많고 어려운 작업이라는 것을 깨달았습니다. 또한, 데이터의 특성을 이해하고 entity와 class를 어떻게 설정하는 지에 따라 성능이 크게 변한 다는 것을 실감 할 수 있었습니다.

▼ 5.2 프로젝트를 수행하며 중요했던 점 및 아쉬웠던 점

- 중요했던 점
 - 라벨러들 간 클래스에 대한 정의 합의
 - SUBJ / OBJ의 엔티티 태깅
 - 파일럿 태깅
 - 가이드라인
 - 예외 케이스 작성
 - 데이터가 너무 쉬워지지 않게끔 적절한 no_relation 추가

- 아쉬웠던 점
 - 클래스간에 완전히 독립되었다 라는 느낌을 받지 못했다.
 - 데이터의 한계로 모든 엔티티의 관계를 표현하지 못했다.
 - 더 많은 클래스를 만들어 보았다면 좋았을 것 같다.
 - 최종 프로젝트에 쓸 수 있는 데이터를 만들었으면 더 좋았을 것 같다.