Small-Scale Korean Corpus for Relation Extraction Task on Nature and Environments

Soyeon Kim **Sangryul Kim** **Eunki Kim**

Seyon Park **Sujeong Im**

Abstract

Relation Extraction is the task to predict the relationship between two entities in a single sentence. Extracting such triplets $< e_{subj}, rel, e_{obj} >$ is a building block for constructing relation knowledge graphs used for structured search and question answering. Considering the aforementioned applications, gathering and refining dataset from a wide range of topics to specific topic are essential. In this work, we first present 2K manually annotated data for relation extraction specifically focused on natures and environments collected from Korean wiki. We provide the detailed explanation for annotating process, trial and errors with final relation maps, the guideline and baseline test results alongside the dataset. The final dataset's Fleiss kappa score is 0.546 and f1 score using klue/roberta-large and bert-base-multilingual-cased is 68.50, 65.41 respectively.

1 Introduction

We, as the internet-friendly generation, are familiar to finding information simply by searching through internet. From the old days, the number of works on retrieving the answers from the vast amount of knowledge have proposed. However, considering the property of knowledge that it is accumulating while the human live long and prosper, *How can we construct the vast knowledge of world we are living in?* is still an interesting question. One of the branches on constructing such knowledge is to embed them in knowledge graph in the form of ontological methods.

Relation Extraction(RE) is the task to predict the relationship between two entities in a single sentence. With RE, any sentence simply summarizes into triplets including subject entity, object entity and relation. Therefore, it is beneficial when constructing knowledge base(Lyu and Chen, 2021).

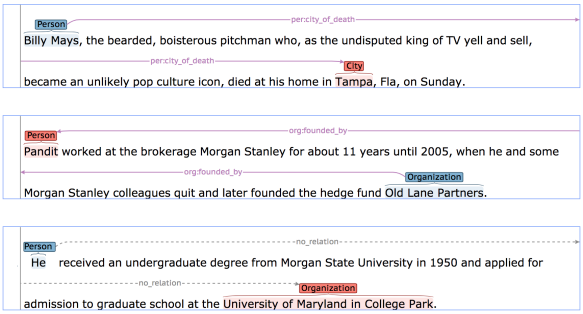

Figure 1: Relation extraction dataset composition (Zhang et al., 2017)

As the recent works put importance on data-driven approach utilizing emerging Natural Language Processing technologies, constructing the corresponding dataset becomes ever important. The wide extent of NLP datasets are already presented heavily focused on most-spoken languages. However, compared to the computer vision dataset, the universal NLP dataset is insufficient to cover all the language cases since each language has its own syntactic rules and attributes.

Recently, Korean Language Understanding Evaluation(KLUE) benchmark (Park et al., 2021) released 8 Korean natural language understanding (NLU) tasks including RE. They mainly collect data from WIKIPEDIA, WIKITREE, POLICY corpora and extend the pool from Namuwiki. The final dataset for RE is comprised of 6 types of entities- `PER`(Person), `ORG`(Organization), `LOC`(Location), `DAT`(Date and time), `POH`(Other proper nouns) and `NOH`(Other numerals) and 30 relations which 18 of them are person-related, 11 of them are organization-related and no relations. from news, review having comparably long sentences with the names of public figures and their relationships to organizations.

Nevertheless, it aims to cover the broad topics in Korea especially on the figures, very few of them is

related to scientific or environmental facts. Considering the performance of data-driven method such as deep learning heavily depends on which data the model has seen while training, RE dataset for Korean in the specific area may require for further industrial purposes. Motivated by open benchmarks for specific domain such as disease (Doğan et al., 2014) or scientific abstract(Luan et al., 2018) in english, we construct the RE dataset focusing on the natures and environments.

2 Adventures to Tag!

In this section, we explain the overall procedures to construct the dataset. Note that this is not about the perfectness, but to share several trial and errors we have gone through.

2.1 Collections

The dataset was collected from Korean wikipedia by the teaching assistant in our boostcamp teaching assistance. The number of sub-topic in the main topic-Natures and Environments- is 62.

2.2 Set Rules for Entities and Relations

We assign subtopics and glance each content to catch the possible distributions of entities and relations. After gathering all the candidates, we summarize 7 entities - `UL(lifeless)`, `LIV`(living things including human), `DAT`(periods, dates), `UNT`(units), `PHE`(phenomenon), `POH`, `LOC`, `NOH`- and 15 relations including `no relation` for the first draft. However, we find our relation map is too specified compared to other teams having 6 relations.

In addition, several instances lie in the border line such as bacteria or RNA. Therefore, we reorganize relations and entities with 7 classes and 7 types respectively based on the feedback. The final entities are `DAT`,`IDV`(individuals which are the very basic living creatures), `PHE`(phenomena), `RES`(resources including environments), `LOC`, `POH`, `NOH`. The first four entities can be used both subject and object but last three entities only for the object case.

Surprisingly, establishing the type of entities and relations is not the end. Following KLUE dataset, the class of relationship should be more defined with the combination of possible subject and relationship. If every 4 subject entity is paired with the 7 classes, the number of classes become complicated and a few classes might have only few samples. Thus, we discard several combinations. Please refer to Table 3 for our final relation map.

2.3 Tagging

We use Tagtog[1] platform for co-work. Even the number of class for relation is 20, we need to set all the possible direction pairs. For instance, if the class is `DAT:influence`, we need to add all possible entities for object, `(DAT, IDV)`, `(DAT, PHE)`, `(DAT, RES)`, `(DAT, POH)`, `(DAT, LOC)`, `(DAT, POH)`, `(DAT, NOH)`. The problem is that it has limit number for adding pairs. Ignorant of it, we again need to discard several possible objects for directions. If all the works are finished, it provides automatic results in json format. The caveat here is that it the json file is given by each text, not with line by line in the text. Since we did not annotate relations but only entities for `no relation` cases, we need to postprocess the json files. Please check the final result format before starting tagging.

2.4 Inter-Annotator Agreement(IAA)

To validate our assessment, we use Fleiss's Kappa score (Fleiss, 1971)[2]. It is the metric to evaluate the reliability of agreement between a fixed number of raters when allocating categorical ratings to a number of items or classifying items. We first pick 5 sentences randomly from all the classes. We label the type of classes given that the corresponding entities for each sentence are marked. The final score is 0.546 regarded as *moderate agreement*.

2.5 Missing Points in Our Works

Actually, the project goal is to achieve more than 0.7 Fleiss's Kappa score. After finishing the works, we compare to other team's procedure whose score is more than 0.9. In this section, we review the missing parts we were ignorant of.

Set the ultimate goal of dataset and narrow down to specific topic Although the topic is *Natures and Environments*, still may sub-topics exist. For instance, a few topics are more related to the histories of geological, biological chronology of earth, a few topics cover the facts of species , climates and so on. Therefore, we had better to fix the goal whether we focus on the historical sequence or the definition of the creatures and phenomenon.

[1] https://www.tagtog.net/
[2] https://github.com/Shamya/FleissKappa

It results in comparably large number of entities and relationships compared to other teams.

Consider to discard examples if it does not fit into the goal At the beginning of labeling process, we are confused to annotate any relation avoiding no relation cases even if it does not strictly fit. Moreover, no standard has been discussed on whether we label `no relation` if entities are meaningful but do not belong to any predefined classes or mark any entity and `no relation` if the sentence does not have any proper relations. Thus, when briefly check the `no relation` case, it includes too wide extent of examples. We think this results in the poor accuracy on `no relation`.

Be aware of human biases when assigning the load We were too oriented to efficient working! Because we assign the sub-topics according to the volume, we had no change to ponder on others' sub topics. We get to care on our own assigned topics. Even we did bring and discuss on confusing examples, we have not done the strict pilot tagging process. Therefore, we had no chance to consider each teammate's own decision standard and set our ground standard concretely. We think it would be better to shuffle all the sentences, allocate and do pilot tag so that every teammate can face and get chance to think over more topics, listen to other opinions and settle the standard within the group.

3 Corpus and Guidelines

From xx sentences, we only select xx sentences as examples. Any single sentence can have more than two triplets especially when the sentence has `alternate name` or listing examples. We split total dataset into train and test set simply stratifying the class. Table 1 summarizes the class distribution of the total dataset. For the access of dataset, relation map files, guidelines, check our github repository.[3]

4 Benchmark Experiment

4.1 Results

We test dataset with BERT(Devlin et al., 2018), Roberta(Liu et al., 2019) pretrained with KLUE dataset and bert-base- multilingual model using huggingface. Following KLUE competitions, we present f1 score, AUPRC and accuracy in Table 2. The ratio of training set and test setis 0.8, 0.2 respectively.

[3] https://github.com/boostcampaitech3/level2-data-annotation_nlp-level2-nlp-03/tree/main

Class names	Count	Ratio
no_relation	761	35.18%
phe:influence	259	11.97%
res:influence	162	7.49%
res:parent_con	126	5.83%
res:feature	116	5.36%
phe:alter_name	78	3.61%
idv:parent_con	77	3.56%
idv:influence	72	3.33%
idv:alter_name	67	3.10%
idv:feature	64	2.96%
res:location	59	2.73%
res:alter_name	58	2.68%
phe:parent_con	58	2.68%
phe:location	52	2.40%
dat:feature	31	1.43%
idv:location	31	1.43%
phe:feature	30	1.39%
dat:alter_name	28	1.29%
res:outbreak_date	27	1.25%
dat:influence	7	0.32%

Table 1: The detail class distribution of the total dataset.

4.2 Analysis

Zooming the confusion matrix, we observe strong ambiguity correlation of the model and human. For instance, `idv:feature`, `idv:parent_con` and `idv:influence` were the most confusing classes when we conducted IAA test. The model accuracy drops significantly on those classes. As mentioned before, `no_relation` score is also not high considering the overwhelming portion of the whole dataset. In terms of the smallest number of classes, it is interesting that `dat:outbreak_date` is nearly 100% but `dat:influence` is 0 %. Where there exist the general trends recording lower scores for the small number of classes, the existence of confusing classes for the small number classes accelerates the degradation.

5 Discussion and Conclusion

In this work, we present small-scale but specific domain, *Natures and Environments*, Korean dataset for Relation Extraction task. In addition to providing dataset, we explain the process to construct and

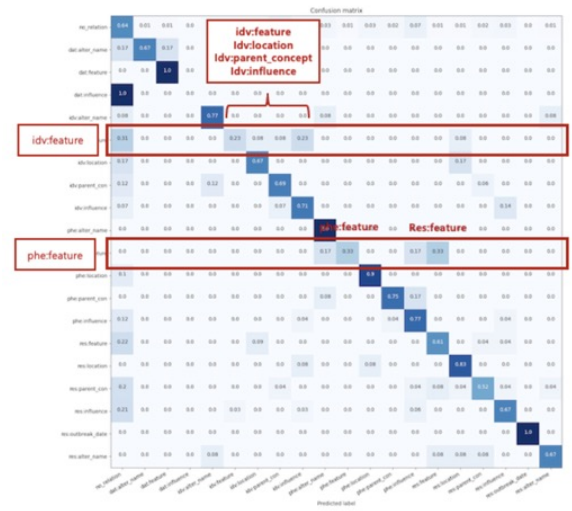

Figure 2: Confusion matrix when trained with klue/roberta-large

validate the dataset in details including the final relation map and guidelines in Korean. Although the dataset has noisy labels, class imbalance and not comparably high consensus scores, we believe our trial and error explanations can help those who require to construct the dataset for their own purposes.

6 Concluding Remarks

Hello! we would like to give short introductions about us informally! We get to construct this dataset as one of projects in Boostcamp AI Tech [4] supported by Naver Connect foundation. Among many provided topics such as animals, wars, COVID-19 and so on, we select *Natures and Environments* as we had a thought on *Maybe we can observe several insights on environmental destruction as time goes with this topic*. Nah,while settling the standards for labeleling and checking the consensus, we honestly have gone hard time and felt how **an annotating is truly human-cost process literally** and why self-supervised training method really requires. Though, we come to understand why the dataset inevitably includes noisy labels and inherent ambiguities, which we did not understand before. Finally, it is highly recommended to set the goal of what this dataset is used for before collecting, refining them. Just collecting and labeling will be highly led to obsolete ones.

[4] https://boostcamp.connect.or.kr/program_ai.html

Model	f1	AURPC	Acc
klue/bert-base	65.66	66.05	67.04
klue/roberta-small	60.72	62.59	60.68
klue/roberta-base	63.25	63.74	62.99
klue/roberta-large	68.5	68.82	69.43
bert-base-mult.-uncased	60.63	61.2	58.57
bert-base-mult.-cased	65.41	65.82	64.55

Table 2: The baseline performance measured by f1 score, AUPRC, accuracy

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*.

Shengfei Lyu and Huanhuan Chen. 2021. Relation classification with entity type restriction. *arXiv preprint arXiv:2105.08393*.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on Empirical Methods in Natural Language Processing*.

Index	Class name(KOR)	Class name(ENG)	Direction(subj, obj)	Description
1	관계없음	no_relation	(*, *)	관계를 유추할 수 없음. 정의된 클래스 중 하나로 분류할 수 없음
2	시기:대체어	DAT:alternate_name	(DAT, DAT / POH)	Object는 Subject의 다른 명칭
3	시기:특징	DAT:feature	(DAT, *)	Object는 Subject의 특징
4	시기:영향	DAT:influence	(DAT,*)	Object는 Subject에 영향을 미침
5	개체:대체어	IDV:alternate_name	(IDV, IDV / POH)	Object는 Subject의 다른 명칭
6	개체:특징	IDV:feature	(IDV, IDV / POH)	Object는 Subject의 특징
7	개체:위치	IDV:location	(IDV, LOC)	Object는 Subject의 위치
8	개체:상위관계	IDV:parent_concept	(IDV, IDV/POH)	Object는 Subject의 상위 개념
9	개체:영향	IDV:influence	(IDV, *)	Object는 Subject에 영향을 미침
10	현상:대체어	PHE:alternate_name	(PHE,PHE/POH)	Object는 Subject의 다른 명칭
11	현상:특징	PHE:feature	(PHE, *)	Object는 Subject의 특징
12	현상:위치	PHE:location	(PHE, LOC)	Object는 Subject의 위치
13	현상:상위관계	PHE:parent_concept	(PHE, PHE)	Object는 Subject의 상위 개념
14	현상:영향	PHE:influence	(PHE,*)	Object는 Subject에 영향을 미침
15	환경/자원:특징	RES:feature	(RES, *)	Object는 Subject의 특징
16	환경/자원:위치	RES:location	(RES, LOC)	Object는 Subject의 위치
17	환경/자원:상위관계	RES:parent_concept	(RES, RES / POH)	Object는 Subject의 상위 개념
18	환경/자원:영향	RES:influence	(RES, *)	Object는 Subject에 영향을 미침
19	환경/자원:발생(발견)날짜	RES:outbreak_date	(RES, DAT)	Object는 Subject의 발생(발견)날짜
20	환경/자원:대체어	RES:alternate_name	(RES, RES)	Object는 Subject의 다른 명칭 height

Table 3: Final relation maps