KoCHET : a Korean Cultural Heritage corpus for Entity-related Tasks

Gyeongmin Kim[*], **Jinsung Kim**[*], **Junyoung Son**[*], **Heuiseok Lim**[†]
Korea University, Korea
{totoro4007, jin62304, s0ny, limhseok}@korea.ac.kr

Abstract

As digitized traditional cultural heritage documents have rapidly increased, resulting in an increased need for preservation and management, practical recognition of entities and typification of their classes has become essential. To achieve this, we propose **KoCHET** - a Korean cultural heritage corpus for the typical entity-related tasks, i.e., named entity recognition (NER), relation extraction (RE), and entity typing (ET). Advised by cultural heritage experts based on the data construction guidelines of government-affiliated organizations, **KoCHET** consists of respectively 112,362, 38,765, 113,198 examples for NER, RE, and ET tasks, covering all entity types related to Korean cultural heritage. Moreover, unlike the existing public corpora, modified redistribution can be allowed both domestic and foreign researchers. Our experimental results make the practical usability of **KoCHET** more valuable in terms of cultural heritage. We also provide practical insights of **KoCHET** in terms of statistical and linguistic analysis. Our corpus is freely available at https://github.com/Gyeongmin47/KoCHET.

1 Introduction

Recently there has been an increasing interest in the preservation of national historical artifacts and traditional cultural heritage, and also grows up the importance of effective management of them through digitization and archival. As the amount of digitized information materials increases rapidly, information extraction (IE) tasks in natural language processing (NLP), such as named entity recognition (NER), relation extraction (RE), and entity typing (ET), have become an essential and fundamental step in the field of historical document analysis.

Despite the necessity of a well-refined entity-centric corpus specialized in domestic cultural heritage, unfortunately, there no exists any cultural heritage domain-specialized corpus in Korean. Moreover, conventional entity-related systems deal only with a coarse set of entity types such as person, location, and organization which is significantly limited in terms of application (Kim et al., 2020). This absence of cultural heritage domain-specialized corpus and narrow coverage of entity types hinders the effective digitization of domestic historical documents because training the model with general corpus for entity-related tasks cannot afford to learn enough significant entity types such as pagodas, historical sites and intangible heritage, and their relations. Furthermore, not in the cultural heritage domain, the existing entity-related datasets supervised by the public institutions have a complicated procedure for data acquisition, and they are also restricted from modification and redistribution. These cumbersome procedures and restrictions have been stumbling blocks for researchers against the rapid increase in digitized cultural heritage materials over the past few decades.

To address these difficulties against the conservation of Korean cultural heritage, we introduce a new dataset collection called **KoCHET** - **Ko**rean **C**ultural **H**eritage corpus for **E**ntity-related **T**asks, a high-quality Korean cultural heritage domain-specialized dataset for NER, RE, and ET tasks. For corpus construction, we crawled the e-museum digitized data of the National Museum of Korea[1] (including data from all 50 museums) as the source text which is for the interested public. We selectively used resources from the museums in which the details of artifacts were registered; moreover, for the completeness of the attribute data, we limited the chronological range of the data from the prehistoric era to the Korean Empire era, excluding the Japanese colonial period. For the annotation, the categorization for classes and attributes appropriate was defined and developed following

[*] These authors have equally contributed to this work
[†] Corresponding author

[1] https://www.emuseum.go.kr/

Proceedings of the 29th International Conference on Computational Linguistics, pages 3496–3505
October 12–17, 2022.

the *2020 Named Entity Corpus Research Analysis*[2] which was published under the guidelines as institutional organizations.

As our corpus focuses on the entity features, it has more detailed and abundant entity types including diverse cultural heritage artifacts, compared to the existing accessible datasets that aim to deal with several downstream tasks in addition to entity-related tasks. Furthermore, the ET of **KoCHET** is the first freely available corpus for the ET task in Korea. In addition to providing these values, this paper provides detailed statistics and linguistic analysis of **KoCHET** for each entity-related task to demonstrate their applicability and enhance understanding of the data, along with baseline experiments with language models.

Our contributions are summarized as follows:

- We introduce **KoCHET** designed for entity-related tasks. This guarantees a high-quality corpus without restrictions regarding modification and redistribution. Moreover, to the best of our knowledge, the ET corpus is the first proposed corpus in Korean.
- We categorized the detailed entity types specialized in the cultural heritage domain, which is essential for preserving our cultural and historical artifacts, thereby contributing as an alternative to the increased demand for the digitalized archiving of cultural heritage documents.
- We prove the applicability of our entity-abundant corpus in each task by providing statistics and linguistic analysis, along with the experiments with pre-trained language models.

2 Related Works

As domains that require expertise, such as the cultural heritage, contain entities or relationships that rarely appear in general domains, the necessity of a corpus specialized in the domain is obvious. Despite such demand, Korean does not yet have a corpus specialized in the cultural heritage area, unlike other languages.

2.1 General cultural heritage corpora

There have been the disclosures of corpora in an effort to preserve traditional culture including the cultural heritage, composing data from the perspective of the entity-related tasks that we deal with. For example, these include a Czech NER corpus constructed based on public optical character recognition data of Czech historical newspapers (Hubková et al., 2020), a Chinese corpus suitable for the computational analysis of historical lexicon and semantic change (Zinin and Xu, 2020), and an English corpus that is one of the most commonly used large corpora in diachronic studies in English (Alatrash et al., 2020).

2.2 Korean public corpora

The National Institute of Korean Language , which is an institution that has established the norms for Korean linguistics, constructed a large-scale dataset[3] for the study of new computational linguistics of Korean (Kim, 2006).

AI HUB is a massive dataset integration platform[4] hosted by the National Information Society Agency (NIA)[5], a government-affiliated organization. To support the development of the Korean artificial intelligence industry for the NLP field, the NIA disclosed domain-specific corpora and 27 datasets have been released or are being prepared.

Electronics and Telecommunications Research Institute , as part of the Exo-brain project[6], provides corpora for NLP tasks such as morphological analysis, entity recognition, dependency parsing, and question answering, and guidelines for building such high-quality corpora[7]. In addition to public datasets opened by public institutions, there is a Korean dataset publicly available for free without the requirement for an access request.

Korean Language Understanding Evaluation (KLUE) dataset was recently released to evaluate the ability of Korean models to understand natural languages with eight diverse and typical tasks (Park et al., 2021b). The tasks include natural language inference, semantic textual similarity, dependency parsing, NER, and RE.

3 KoCHET

Following the guidelines of Korean institutional organizations, **KoCHET** is a domain specialized

[2] https://www.korean.go.kr

[3] https://stdict.korean.go.kr/

[4] https://aihub.or.kr/

[5] https://www.nia.or.kr/

[6] http://exobrain.kr/pages/ko/result/outputs.jsp

[7] https://www.etri.re.kr/

corpus for cultural heritage, which ensures quality and can be freely accessed. In this section, we report the annotation process and guidelines in detail.

3.1 Annotation Process

To improve the quality of annotations on our entity-rich corpus related to cultural heritage, we conducted the annotation process based on expertise in the cultural heritage domain.

Annotation Guidelines The raw corpus annotated by each annotator is equally divided by the category. The annotators were instructed to follow two types of rules by the aforementioned entity guidelines in Section 1; one is related to tagging units and categories, and the other is the principle of unique tagging. The minimum unit is based on one word for the tagging units and categories. In addition, it is applied only to cases written in Korean, where the notation is possible. It is not tagged in the case of Chinese characters and English, but if it is read in Korean, it is included in the tagging range. For the principle of unique tagging, there are cases of duplication in entities that belong to two or more semantic regions. This guideline grants a single tag to a semantically suitable word and refers to assigning only one tag by prioritizing it accordingly. There are two cases in which this principle should be applied. The first case is where the entity belongs to two semantic categories regardless of the context. The second refers to the case where it may vary depending on the context. In both cases, tagging is determined according to the pre-defined priority.

Annotator Training and Cross-Checking We recruited 34 college and graduate annotators who have been professionally educated on the cultural heritage domain in Korea to participate in the annotation process. All annotators were trained for a week, and each of them was familiarized with the annotation guideline and conducted practice annotation on test samples. The annotation team met once every week to review and discuss each member's work during the annotation process. All entity types and relations were reviewed by four cross-checking annotators, afterward, were additionally checked by two expert supervisors. The discrepancy between annotators on the annotated entity types and relations is also discussed and agreed upon in the period. These procedures allowed the reliability and validity of **KOCHET** on the cultural heritage objects to be improved.

3.2 Schema for Task Annotation

3.2.1 Named Entity Recognition

Label	Train	Dev	Test
	Counts (%)		
Artifacts (AF)	91,453 (35.57)	11,374 (35.54)	11,366 (35.35)
Person (PS)	51,758 (20.13)	6,455 (20.17)	6,744 (20.97)
Term (TM)	25,781 (10.02)	3,175 (9.92)	3,159 (9.82)
Date (DT)	23,636 (9.19)	2,943 (9.20)	3,078 (9.57)
Political location (LCP)	20,076 (7.80)	2,375 (7.42)	2,384 (7.41)
Civilization (CV)	15,404 (5.99)	1,929 (6.03)	1,835 (5.71)
Material (MT)	8,893 (3.45)	1,160 (3.62)	1,046 (3.25)
Location (LC)	6,881 (2.67)	857 (2.68)	881 (2.74)
Animal (AM)	4,376 (1.70)	578 (1.81)	566 (1.76)
Plant (PT)	3,952 (1.53)	549 (1.72)	498 (1.55)
Geographical location (LCG)	2,821 (1.09)	354 (1.11)	348 (1.08)
Event (EV)	2,045 (0.79)	254 (0.79)	248 (0.77)

Table 1: The counts of entities and their distributions (%) in our NER data.

As described in Table 1, we defined 12 entity types. They were tagged with the character-level beginning-inside-outside (BIO) tagging scheme, which is the generally adopted method for sequence labeling problems. For example, "아시아 (Asia): Geographical Location (LCG)" is tagged as "아: B-LCG," "시: I-LCG," "아: I-LCG." Therefore, we evaluated the model not only with entity-level F1 score but also with character-level F1 score (Park et al., 2021b).

Label Description

- **Artifacts (AF)** generally refer to objects created by humans corresponding to common and proper nouns and also include cultural properties. Therefore, artificial materials such as buildings, civil engineering constructions, playground names, apartments, and bridges fall under this category.
- **Person (PS)** is a category for content related to people, including real persons, mythical figures, fictional characters in games/novels, occupations, and human relationships.
- **Term (TM)** includes the color, direction, shape, or form that describes an artifact. Patterns and drawings are classified as TM, owing to the characteristics of movable cultural properties.
- **Civilization (CV)** is defined as terms related to civilization/culture. It targets words classified by detailed civilizations/cultures, such as clothing and food.

- **Date (DT)** includes all entities related to date and time, such as date, period, specific day, or season, month, year, era/dynasty. However, in the case of an unclear period that cannot be tagged with a separate entity, tagging is not performed.

- **Material (MT)** includes a substance used as a material or an expression for the substance. In other words, it indicates the entity corresponding to the detailed classification of a substance (metal, rock, wood, etc.). When an entity can be tagged as both natural objects (AM, PT) and MT, tagging as MT takes precedence.

- **Geographical location (LCG)**, **Political location (LCP)**, and **Location (LC)** are defined as geographical names, administrative districts, and other places, respectively.

- **Animal (AM)** and **Plant (PT)** are defined as animals and plants, respectively, excluding humans. If it is applied as a subject of a picture, it is also included in the category of animals and plants.

- **Event (EV)** contains entities for a specific event/accident. In principle, social movements and declarations, wars, revolutions, events, festivals, etc., fall under this category and should be classified only if they exist as a separate entity.

3.2.2 Relation Extraction

Unlike the other existing corpora, our corpus has the advantage of capturing various relationships between multiple entities that are included in a sentence because more than one relation can exist per raw sentence. We consider the relations between annotated entities in the NER annotation procedure. In the case of certain tokens, it can be a subject or an object depending on the relationship with other tokens. A relationship in the form of a self-relationship between identical tokens does not exist.

As shown in Table 2, our RE corpus consists of 14 labels, and these were defined based on the Encyves ontology research of the National Culture Research Institute[8].

Label Description

- "**A depicts B**" implies the relationship between an object and its color, shape or pattern, etc. For example, "Green Door" corresponds to this relationship. It can also represent a descriptive relationship such as "Picture of a place-the place where it was taken" or "Picture of a person-the person who is the object of the painting."

- "**A documents B**" implies "~ records -." ;a relationship such as "Record-The person who records it" can be represented by this. It also indicates the relationship like a record written on an object such as "Postcard-Explanation" or a specific language written on a document such as "Record-Chinese characters."

- "**A hasSection B**" indicates "~ is located at -." It represents the relationship between a statue, building, or specific attraction and a location, such as a certain city and place.

- "**A servedAs B**" implies "~ is the role of -," which corresponds to the relationship between a person, and his/her position or occupation, etc.

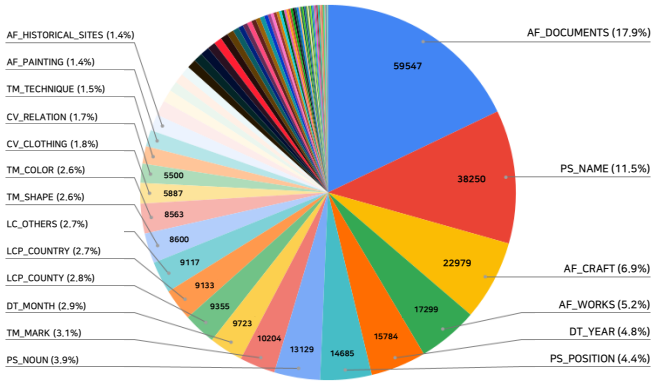
- "**A hasCreated B**" demonstrates, for example, "Person-Documents" or "Person-Painting," which refers to the relationship between a person and a document such as a book, map, or drawing, or his/her activities to record works.

- "**A OriginatedIn B**" means "~ is discovered at –" or "~ is produced at -(time)." It indicates that cultural property is produced at a specific time such as "Craft-Year" or is discovered at a particular place such as "Object-Place," or is produced at a certain site such as "Document-Place." For example, the relation

Label	Train	Dev	Test
	Counts (%)		
A depicts B	14,157 (22.09)	1,803 (22.45)	1,711 (21.85)
A documents B	10,214 (15.94)	1,244 (15.49)	1,220 (15.58)
A hasSection B	6,542 (10.21)	818 (10.19)	776 (9.91)
A servedAs B	6,546 (10.22)	780 (9.71)	740 (9.45)
A hasCreated B	6,136 (9.58)	759 (9.45)	744 (9.50)
A OriginatedIn B	5,456 (8.51)	679 (8.45)	663 (8.47)
A consistsOf B	4,331 (6.76)	569 (7.09)	586 (7.48)
A isConnectedWith B	3,489 (5.44)	501 (6.24)	461 (5.89)
A fallsWithin B	3,454 (5.39)	415 (5.17)	483 (6.17)
A isUsedIn B	1,906 (2.97)	238 (2.96)	244 (3.12)
A hasTime B	934 (1.46)	111 (1.38)	95 (1.21)
A wears B	798 (1.25)	97 (1.21)	86 (1.10)
A hasCarriedOut B	112 (0.17)	15 (0.19)	19 (0.24)
A hasDestroyed B	5 (0.01)	2 (0.02)	3 (0.04)

Table 2: Relation counts and distributions (%) for our RE corpus.

[8] http://dh.aks.ac.kr/Encyves/wiki



ARTIFACT	MATERIAL	PLANT	DATE	CIVILIZATION
DOCUMENTS CRAFT WORKS PAINTING HISTORICAL_SITES MONUMENT INTANGIBLE_HERITAGE PAGODA TRANSPORT WEAPON BUILDING MUSICAL_INSTRUMENT ROAD RELIC	SOIL PAPER FIBER ROCK METAL WOOD BONE JEWELRY SKIN GRASS SYNTHETIC SEED RUBBER OTHERS	PART FLOWER TREE TYPE GRASS FRUIT OTHERS	YEAR MONTH DURATION DYNASTY DAY SEASON GEOAGE	CLOTHING REALTION LANGUAGE BUILDING_TYPE FOOD CURRENCY TRIBE LAW DRINK POLICY TAX SPORTS SPORTS_INST NAME FOOD_STYLE
		LCP	ANIMAL	
		COUNTY COUNTRY CITY PROVINCE CAPITALCITY	PART BIRD MAMMALIA TYPE FISH INSECT REPTILIA AMPIHBIA OTHERS	
EVENT	LCG	TERM	LC	PERSON
FESTIVAL ACTIVITY WAR_REVOLUTION SPORTS OTHERS	MOUNTAIN RIVER ISLAND CONTINENT BAY	MARK SHAPE COLOR TECHNIQUE	SPACE OTHERS	NAME POSITION NOUN IMAGINARY

Figure 1: Visualization of all the labels that cover 84% of the entity types is shown on the left side, and 106 general and fine-grained entities with their distributions (%) are shown on the right side.

between earrings and tombs or a newspaper and the company of the newspaper fall into this.

- "**A consistsOf B**" refers to the relation between an object and its raw ingredients, such as soil, iron, and wood that constitute an object.
- "**A isConnectedWith B**" represents a person-to-person association. The relationships between two positions or a person and the position he or she holds do not fall into this.
- "**A fallsWithin B**" implies "~ is denominated as -." It indicates the relationship of alternate names such as "Person-Specific name," or between a name and designation in front of the name, or between words that refer to synonymous concepts such as "Verse-Poetry."
- "**A isUsedIn B**" indicates "~ is used for the purpose of -" or literally "~ is used in -." For example, it can also indicate the material used for a certain object, such as "Raw material-Clothes." The relationship between an object and the place where the object is used, such as a signboard and a palace, or the relationship between certain means of performing a function and an object such as "Bowl-Rice cake" can correspond to this category.
- "**A hasTime B**" implies "~ has happened at -." For example, it can indicate the relationship between a particular event and a specific date, such as "Presidential election-1928." The relation between a specific date and a certain work, such as the year of production of a work and the year of construction of a building, can fall under this category, for example, "Year-Craftwork."
- "**A wears B**" implies "~ puts - on." For instance, not only clothes such as school uniforms but also crafts, etc. may correspond to the object argument.
- "**A hasCarriedOut B**" indicates "- is caused by ~." It can represent a relationship between a specific organization or group and an event conducted by it, such as a festival or social movement.
- "**A hasDestroyed B**" implies the event that caused destruction such as "War-Destroyed place," or the collapse of a country in a specific year such as "Country-Year," or the relationship in which a building, structure, monument, etc. is destroyed at a particular period.

3.2.3 Fine-grained Entity Typing

Given a sentence and entity mention within it, the ET task predicts a set of noun phrases that describe the mention type. For example, in "김홍도는 조선 후기의 화가이다. (*Kim Hong-do was a painter of the Joseon era of Korea.*)," *Joseon* should be typed as "dynasty/Date" and not "country/Location." This typification is crucial for context-sensitive tasks such as RE, coreference resolution, and question answering (e.g., "In which era was Kim Hong-do, an artist?"). Unlike high resource languages, we found that the Korean corpus for the ET task has not been released. In dealing with this data scarcity problem and promoting universal studies, we release a Korean ET task corpus for the first time, to the best of our knowledge.

Sentence with Entity Mention	Entity Types
조선시대에는 전통 관습을 잇기 위한 많은 향로가 제작되었다. (In the **Joseon dynasty**, many fragrance burners were created for traditional customs.)	**DT_DYNASTY**, DT_DURATION LCP_COUNTRY, LCP_CITY, LCP_COUNTY LC_OTHERS, AF_DOCUMENTS
노란 바탕의 **모란**이 양쪽에 그려져 있다. (The yellow background **peony** is drawn on both sides.)	**PT_FLOWER**, PT_TYPE, PT_OTHERS, TM_SHAPE
19세기 후반 청주의 재정을 파악할 수 있는 자료가 있다. (There are data to comprehend the finances of Cheongju in the **late 19th century**.)	**DT_YEAR**, DT_DYNASTY, DT_DURATION

Table 3: Examples including entity mentions and their fine-grained entity types. Entity mentions and the correct types in the given context are bold. All fine-grained entity types are shown in Figure 1.

The schema for the ET task was designed with reference to the data construction process of the Fine-Grained Entity Recognition dataset (Ling and Weld, 2012). Considering the properties of the cultural heritage domain, we categorized the 12 general entity types aforementioned in the NER task (Section 3.2.1) into a fine-grained set of 94 types with detailed meanings. Particularly, the cultural taxonomy defined in the *Cultural Properties Protection Law*[9] was applied to AF, and the 2004 Cavalier-Smith's classification system (Cavalier-Smith, 2004) was applied to the biological scope of PT and AM. All fine-grained entity types are detailed in Figure 1.

The fine-grained entities for entity-related downstream tasks in the cultural heritage domain enable a more detailed contextualized representation for each entity mention than the previous typing schemas, which only predict relatively coarse types of entities. Table 3 lists three example sentences with entity mention that can represent several fine-grained types. Given a sentence with an entity mention, the appropriate type that describes the role of the entity span in the sentence should be predicted. Our fine-grained entity types can embrace all the existing general types and categorize them in greater detail. Accordingly, they can let models understand richly the noun phrases including entity, compared to when the models are trained to predict only relatively coarse types. For Figure 1, the circle on the left shows the visualization of fine-grained entity types that possess approximately 84% among all labels in the corpus, and the set on the right shows the detailed distributions of all fine-grained types. Each example includes 2.94 fine-grained entities on average; there are up to nine several fine-grained entity types per entity. The category to which the most entities belong is "AF_DOCUMENTS," which possesses 17.9%, and that on the second place is "PS_NAME," having 16.7%.

Label Description

- **12 general** types: PS, AF, AM, CV, DT, EV, PT, MT, TM, LC, LCG, LCP
- **94 fine-grained** types, which were mapped to the cultural heritage-specialized fine-grained entity labels, were inspired by prior works (Ling and Weld, 2012; Gillick et al., 2014; Choi et al., 2018).

3.3 Analysis on KoCHET

3.3.1 Diachronic and Linguistic Analysis

There are mainly two differences between the entities in the proposed corpus and those commonly used.

First, archaic expressions that are not used in modern times are frequently shown in our corpus. Specifically, such expressions continually appear when ancient documents or historical artifacts are quoted. Let us consider the phrase "한번사신레꼬-드는승질상밧고거-나믈느지는안슴니다" in sentence 1 in Table 4. Although it is written using syllables of modern Korean, the grammar and the vocabulary are fairly dissimilar from those of contemporary Korean, such as word spacing and syllabification, i.e., separation rule between the units of the word. When translating the sentence with quotation marks into modern Korean, it can be expressed as "한번 사신 레코드는 성질상 바꾸거나 무르지는 않습니다 (Once a record is purchased, it cannot be exchanged or refunded due to its characteristics)."

[9] www.cha.go.kr

Index	Example sentences
1	앞면 좌측 하단에 '한번사신레꼬-드는승질상밧고거-나믈느지는안슴니다' 문구가 있음. There is a phrase '한번사신레꼬-드는승질상밧고거-나믈느지는안슴니다'(archaic Korean) on the left corner of the front side.
2	1면에는 안창호씨(安昌浩氏)의 연설, 편집실 여언(餘言) 등의 기사가, · · ·, 인쇄됨. On the first page, articles such as Mr. Changho Ahn(安昌浩氏)(Chinese character)'s speech and editorial comments(餘言)(Chinese character), · · ·, were printed.
3	'戰爭の訓示', · · ·, 등의 기사와 일본 언어학자 가나자와 쇼자부로(金澤庄三郎, 1872∼1967)의 현대 국어 음운에 대한 연구물인「朝鮮語發音篇」의 일부를 게재함. '戰爭の訓示(Japanese)', · · ·, the articles and「朝鮮語發音篇」(Chinese character), the part of a study on the modern Korean phonology of Japanese linguist Kanazawa Shouzaburou(金澤庄三郎 (Chinese character), 1872∼1967) were published.

Table 4: Example sentences contained in our corpus. These examples include not only Korean but also Japanese and Chinese characters. Also, they contain archaic expressions that are not used in modern times. These characteristics make it more suitable for the learning of cultural heritage domain. Note that we omitted some of the words in the sentence for brevity.

Task		Train	Dev	Test
NER	# of examples	89,884	11,245	11,233
	# of entities	393,076	32,003	32,153
RE	# of examples	31,012	3,876	3,877
	# of relations	64,080	8,031	7,831
ET	# of examples	90,558	11,320	11,320
	# of mentions	266,209	33,226	33,395

Table 5: Statistics of **KoCHET** for each task.

Second, several entities contained in **KoCHET** written in Korean are followed by the descriptions written in either Chinese or Japanese characters. For example, as shown in sentence 2 in Table 4, the description with Chinese characters in parentheses follows the entity "안창호씨," and is usually written such as "안창호씨(安昌浩氏)." Further, Japanese characters are also present throughout the corpus, enhancing the polyglot property of the corpus, as shown in sentence 3. Therefore, to fully understand such expression types in our corpus, multilingual factors of language models should be considered; particularly in the case of token classification tasks, in which the meaning of each token directly affects the model performance.

3.3.2 Statistics

The overall statistics of **KoCHET** are showed in Table 5. For the NER corpus, 457,232 entities from 112,362 examples in total. For the RE corpus, 79,942 relations from 38,765 examples were annotated in total. For the ET corpus, 332,830 entity mentions from 113,198 examples were annotated in total. The annotated corpus was divided into three subsets for each task, i.e., a ratio of 8:1:1 for training, development, and testing, respectively. In this section, we describe our corpus statistically in the order of NER, RE, and ET.

First, as shown in Table 1, we used 12 entity types for our cultural heritage NER corpus. Due to the properties of the cultural heritage domain, the three primary entity types, i.e., artifacts (AF), person (PS), and term (TM), account for the majority of the total entity population. AF, PS, and TM entities possess approximately 36%, 20%, and 10%, respectively, which are used as crucial information in the cultural heritage domain. The AF type includes cultural assets and historical landmarks, the TM type includes patterns or traces engraved on certain cultural assets, and the PS type particularly includes not only general people but also particular types of persons such as mythical figures. On the other hand, the EV type occupies the most minor proportion, approximately 0.8%, because our corpus especially aims to concentrate on the cultural heritage.

Second, Table 2 demonstrates the distribution of 14 RE labels. In the case of "A depicts B" and "A documents B," cultural assets left in a specific form such as records, drawings, and photographs are included, whereas "A hasSection B" contains cultural heritage or historical landmarks located at a specific place. Among them, "A depicts B," "A documents B," and "A hasSection B" are the most relationship labels with approximately 22%, 16%, and 10% of the total, respectively. "A depicts B" and "A documents B" include cultural assets left in a specific form such as records, drawings, and

Model	NER		RE	ET
	Entity F1 (σ)	Character F1 (σ)	F1 (σ)	F1 (σ)
Multilingual fine-tuned Models				
Multilingual BERT	59.81 (0.09)	71.80 (0.12)	80.85 (0.39)	91.64 (0.10)
XLM-RoBERTa-base	**76.57** (0.13)	**82.69** (0.09)	80.29 (0.53)	91.13 (0.16)
Korean fine-tuned Models				
KLUE-BERT-base	39.31 (0.10)	55.63 (0.15)	**82.44** (0.18)	**93.08** (0.27)
KLUE-RoBERTa-base	38.92 (0.28)	55.47 (0.21)	82.42 (0.57)	92.80 (0.17)

Table 6: Experiments results on the NER, RE, and ET tasks. F1 score (%) is used for the evaluation metric with σ which shows the standard deviation of the score. We divide the baseline models into two parts: the Multilingual models and the Korean models, marking the highest performances with bold text.

photographs, whereas “A hasSection B" contains cultural heritage or historical landmarks located at a particular place. “A hasDestroyed B" has the smallest proportion with ten relations in total because, in actual history, significant events such as the collapse of a nation or the loss of cultural properties are not as diverse as the types of general cultural assets.

Finally, among the fine-grained entity types, the “AF_DOCUMENTS" type, such as historical documents, occupies the largest part with 17.9%, and “PS_NAME" including the names of historical figures, takes second place by occupying 11.5%. On the other hand, the entity types to which belong to the AM, PT, MT, and EV almost account for under 1.0%.

4 Experiment

The detailed experimental settings are in Appendix A.

Experimental results According to Table 6, two tendencies are observed. One is that in the NER task, the multilingual models, i.e., multilingual BERT and xlm-RoBERTa-base, showed better performance by more than 30% difference in both Entity F1 and Character F1 scores compared to the Korean models, i.e., KLUE-BERT-base and KLUE-RoBERTa-base. The other is that in the RE and ET tasks, the performances of the Korean models were at least 1.1% higher than those of the multilingual models.

Experimental Analysis As the token classification tasks are directly affected by segmentation (Kim et al., 2021; Park et al., 2021a), models with linguistic knowledge of Chinese and Japanese overperform in such tasks (Pires et al., 2019). In other words, the multilingual models are considered to segment better each token composed of various languages, especially in the NER corpus. In addition, in Table 7, the Korean models, i.e., KLUE-BERT-base and KLUE-RoBERTa-base show a significantly higher ratio of unknown tokens than the multilingual language models. It is attributed that the NER task requires more polyglot features of the model compared to the other tasks, i.e., RE and ET, which has the properties of sentence classification tasks. On the other hand, as the RE or ET task does not classify all tokens in a sentence, the correct answer can be satisfactorily inferred from only the given Korean words; thereby, the language models pre-trained in Korean show better performance in the two tasks compared to the multilingual model.

Model	UNK_dev (%)	UNK_test (%)
Multilingual BERT	0.8156%	0.7684%
XLM-RoBERTa-base	0.1952%	0.1810%
KLUE-BERT-base	5.8670%	5.9677%
KLUE-RoBERTa-base	5.8670%	5.9677%

Table 7: Unknown (UNK) token ratio (%) of each model for development and testing set in the corpus. Baseline models pre-trained in Korean show the same proportions because they use identical vocabulary and tokenizers.

5 Conclusion

In this paper, we introduced **KoCHET** - a Korean cultural heritage corpus for three typical entity-related tasks, i.e., NER, RE, and ET. Unlike the existing public Korean datasets with additional restrictions, **KoCHET** obviated the cumbersome prerequisite and can be freely modified and redistributed. Furthermore, we proved the applicability

of our entity-abundant corpus with the experiments employing the various pre-trained language models and provided practical insights regarding the statistical, diachronic, and linguistic analysis. Above all, the most significant contributing point is that the disclosure of our corpus is expected to serve as a cornerstone for the development of IE tasks for a traditional cultural heritage. We hope that the continuous effort to preserve cultural heritage with the effective management of digitized documents containing cultural artifacts is encouraged by this research.

Acknowledgements

This research is supported by Ministry of Culture, Sports and Tourism and Korea Creative Content Agency(Project Number: R2020040045), MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), and Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques).

References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. Ccoha: Clean corpus of historical american english. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6958–6966.

Thomas Cavalier-Smith. 2004. Only six kingdoms of life. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1545):1251–1262.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dan Gillick, Nevena Lazic, Kuzman Ganchev, Jesse Kirchner, and David Huynh. 2014. Context-dependent fine-grained entity type tagging. *arXiv preprint arXiv:1412.1820*.

Helena Hubková, Pavel Král, and Eva Pettersson. 2020. Czech historical named entity corpus v 1.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4458–4465.

Gyeongmin Kim, Chanhee Lee, Jaechoon Jo, and Heuiseok Lim. 2020. Automatic extraction of named entities of cyber threats using a deep bi-lstm-crf network. *International Journal of Machine Learning and Cybernetics*, 11(10):2341–2355.

Gyeongmin Kim, Junyoung Son, Jinsung Kim, Hyunhee Lee, and Heuiseok Lim. 2021. Enhancing korean named entity recognition with linguistic tokenization strategies. *IEEE Access*, 9:151814–151823.

Hansaem Kim. 2006. Korean national corpus in the 21st century sejong project. In *Proceedings of the 13th NIJL International Symposium*, pages 49–54. National Institute for Japanese Language Tokyo.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Chanjun Park, Sugyeong Eo, Hyeonseok Moon, and Heuiseok Lim. 2021a. Should we find another model?: Improving neural machine translation performance with ONE-piece tokenization method without model modification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 97–104, Online. Association for Computational Linguistics.

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

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

Sergey Zinin and Yang Xu. 2020. Corpus of Chinese dynastic histories: Gender analysis over two millennia. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 785–793, Marseille, France. European Language Resources Association.

A Experimental Setup

As the baseline models, we employed two global language models: multilingual bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019) and a cross-lingual language model XLM-RoBERTa-base (Conneau et al., 2020) containing the Korean language, and two KLUE language models: KLUE-BERT-base, KLUE-RoBERTa-base, which were recently published covering various Korean downstream tasks. In all the model experiments, the performance of each model was measured five times, and the average of each result was evaluated as the final result. Further, we set our environment for the experiment with four A6000 GPUs and 384 GB memory. The hyperparameters in the fine-tuning step were set as follows. The learning rate and weight decay were consistently set at 5e-5 and 0.01 across all three tasks. The number of training epochs was set to 10 in NER, RE and 3 in ET. The batch size in training and testing procedures was set to 128 in NER, RE and 256 in ET. In the case of max sequence length, the lengths of 256 and 128 were used for each task.

We evaluated our system by employing F1 score, which is standard metric for classification tasks. Specifically, the evaluation metrics for NER task were Entity F1 and Character F1 based on previous research (Park et al., 2021b). Entity F1 is a metric that is recognized as a correct answer only when all types included in an entity are matched accurately. Conversely, Character F1 is a metric that evaluates each type of syllable in a sentence individually. The evaluation metrics for the RE task were F1 score in the Scikit-learn library (Pedregosa et al., 2011). As for ET, we adopted the evaluation metrics of loose F1 score following the same evaluation criteria used in previous works (Ling and Weld, 2012; Wang et al., 2020).