International Journal of Internet, Broadcasting and Communication Vol.14 No.2 55-62 (2022)
http://dx.doi.org/10.7236/IJIBC.2022.14.2.55

IJIBC 22-2-9

Development of Tourism Information Named Entity Recognition Datasets for the Fine-tune KoBERT-CRF Model

Myeong-Cheol Jwa* and Jeong-Woo Jwa**

**Student, Korea University of Technology and Education, Cheonan-si, Korea*
***Professor, Department of Telecommunication Eng., Jeju National University, Jeju, Korea*
e-mail : jmc0407@koreatech.ac.kr*****, lcr02@jejunu.ac.kr******

Abstract

A smart tourism chatbot is needed as a user interface to efficiently provide smart tourism services such as recommended travel products, tourist information, my travel itinerary, and tour guide service to tourists. We have been developed a smart tourism app and a smart tourism information system that provide smart tourism services to tourists. We also developed a smart tourism chatbot service consisting of khaiii morpheme analyzer, rule-based intention classification, and tourism information knowledge base using Neo4j graph database. In this paper, we develop the Korean and English smart tourism Name Entity (NE) datasets required for the development of the NER model using the pre-trained language models (PLMs) for the smart tourism chatbot system. We create the tourism information NER datasets by collecting source data through smart tourism app, visitJeju web of Jeju Tourism Organization (JTO), and web search, and preprocessing it using Korean and English tourism information Name Entity dictionaries. We perform training on the KoBERT-CRF NER model using the developed Korean and English tourism information NER datasets. The weight-averaged precision, recall, and f1 scores are 0.94, 0.92 and 0.94 on Korean and English tourism information NER datasets.

Keywords: *Tourism Information NER, Smart Tourism Chatbot, KoBERT model, Conditional Random Fields (CRF), pre-trained language models (PLMs).*

1. Introduction

The smart tourism service uses 4th industrial revolution technologies such as the Internet of Things, communication infrastructure, big data, artificial intelligence, AR/VR/MR, metaverse, and drones to create a personalized context-aware itinerary and provide tour guide services. A smart tourism chatbot service is required to efficiently provide smart tourism services such as recommended travel products, tourist information, my travel itinerary, and tour guide service to tourists using a smart tourism information system. We developed the rule-based chatbot service using the Neo4J graph database to efficiently provide tourism information provided by the smart tourism app to tourists [1, 2]. The developed smart tourism chatbot system identifies user intent based on Khaiii (Kakao Hangul Analyzer III) morpheme analyzer [3] and predefined rules. When the user's query includes a tourist name consisting of a compound noun, an error occurred in the result of the

Manuscript Received: February. 24, 2022 / Revised: February. 27, 2022 / Accepted: March. 2, 2022
Corresponding Author: lcr02@jejunu.ac.kr
Tel: +81-64-754-3638, Fax: +81-64-754-3610
Professor, Department of Telecommunication Eng., Jeju National University, Korea

Khaiii morpheme analyzer, so we modified the error patch dictionary. The tourism knowledge base of the chatbot system was created using a Neo4J graph database [4] based on the tourism information built in the smart tourism information system. The tourism knowledge base includes service attribute data for searching for a tourist destination according to a user's query intention. When the intent of the question is identified through the analysis of the user's question, the Neo4J query is written to search the related information in the knowledge base, and the answer sentence is provided to the user.

The chatbot system uses speech recognition for natural language understanding (NLU) and speech synthesis technology for natural language generation (NLG) for the user interface. In addition, a NER model using PLMs is used to identify the intent of a user's query, and a DST model is used for dialog management (DM). Large-scale PLMs such as Embeddings from Language Models (ELMo), Bidirectional Encoder Representations from Transformer (BERT), and OpenAI Generative Pre-Training (GPT) can be used in a wide range of natural language processing (NLP) tasks [5-7]. There are there are ETRI KorBERT, SKT KoBERT, AWS-SKT KoGPT-2, TwoBlock AI HanBERT, and Samsung SDS KoreALBERT as the Korean pre-trained language models. BERT NER is a fine-tuned BERT model that is fine-tuned on the English version of the standard CoNLL-2003 Named Entity Recognition dataset [8]. The standard CoNLL-2003 NER dataset consists of entities such as location (LOC), organizations (ORG), person (PER) and Miscellaneous (MISC). There are the National Institute of the Korean Language NER data [9], the Natural Language Processing Lab NER data at Korea Maritime & Ocean University [10], and NAVER NLP Challenge 2018 NER data [11] in the Korean datasets.

To develop a competitive smart tourism chatbot system, it is necessary to develop the Named Entity Recognition (NER), Dialogue State Tracking (DST), and Question Answering (QA) models using pre-learned language models (PLMs). The chatbot system uses the named entity recognition result to understand the intent of the user's question. The tourism information name entity training dataset is needed to develop a NER model for a smart tourism chatbot system. In this paper, we develop NER data for tourism information in Korean and English using the tourism information data collected from the smart tourism app, Jeju Tourism Organization visitJeju, and web data. We perform training on the developed the Korean and English tourism information NER data using the KoBERT-CRF model.

2. The Korean and English Tourism Information NER Datasets

It is necessary to create the tourism information NER datasets to develop a NER model using the pre-trained language models of the smart tourism chatbot. In this paper, we create the Korean and English tourism information Name Entity datasets for the development of the NER models using the Korean pre-trained language models.

2.1 Source data of smart tourism NER datasets

The tourism information source data for generating tourism information NER datasets was constructed from the following three data:

(1) Tourism information data of smart tourism information system that provides smart tourism app and chatbot app service

(2) Tourism information data from the VisitJeju web of the Jeju Tourism Organization [12]

(3) Tourism information data from internet websites

Table 1 shows examples of source data for tourism information in Korean and English used to create the

tourism information NER dataset.

Table 1. Source data of the Korean and English smart tourism information NER datasets

(a) Korean tourist information data

index	Tourism Information Data
6574	최고봉, 민족의 영산인 한라산은 1966 년 10 월 12 일 천연기념물 제 182 호로 한라산 천연보호 구역으로 지정 보호되고 있으며, 2002 년 12 월에는 UNESCO 생물권 보전지역으로 등록되었다.
6575	제주올레 15 코스는 한수리를 지나자마자 길은 바다를 등진다. 마을 올레의 시작이다. 인적 드문 한적한 마을이 있고, 사시사철 푸른 밭이 있고, 그 밭에 물을 대는 작은 못들이 있고, 두 개의 오름이 있고, 감춰진 난대림 숲이 있다.
6576	거문오름 내부의 대표적인 수종으로는 식나무와 붓순나무를 들 수 있다. 층층나무과의 상록수인 식나무는 식물 중에서 산소정화능력이 가장 뛰어나다. 붓순나무는 예전에 주판알 재료로 쓰였을 뿐 아니라 장작을 때도 연기가 나지 않는다 하여 4.3 항쟁 당시 산사람들이 토벌군의 눈을 피해 밥을 지을 때 즐겨 썼다고 한다.
6577	수월봉은 분화구와 가까운 부분에는 무게 때문에 멀리가지 못한 큰 돌들이 박혀 있고, 거리가 점점 멀어지면서 완만한 경사를 보여주는 물결 모양의 구조가 나타난다.
6578	용눈이 오름은 다른 오름과 달리 세 개의 오름이 능선으로 이어져 있어서 전체적으로 부드럽다. 용눈이 오름에서 날씨가 좋은날에는 성산일출봉 城山日出峰과 우도까지 전망이 가능하다.

(b) English tourist information data

index	Tourism Information Data
6574	Hallasan, the highest peak and the spiritual mountain of the nation, was designated and protected as Natural Monument No. 182 on October 12, 1966 as a natural protected area of Mt. Halla, and in December 2002, it was registered as a UNESCO Biosphere Reserve.
6575	On Jeju Olle Course 15, the road faces the sea as soon as it passes Hansu-ri. It is the beginning of the village Olle. There is a quiet village with few people, there is a green field all year round, there are small ponds that water the field, there are two oreums, and there is a hidden temperate forest.
6576	The representative tree species inside Geomunoreum are cypress trees and cypress trees. As an evergreen tree of the Dogwood family, the cedar has the best oxygen purification ability among plants. It is said that brush-headed wood was used as a material for abacus in the past as well as smoke when burning.
6577	In Suwolbong Peak, there are large stones that cannot be reached due to its weight in the part close to the crater, and as the distance increases, a wavy structure showing a gentle slope appears.
6578	Unlike other volcanic cones, Yongnuni Oreum has three volcanic cones connected by a ridgeline, so it is soft overall. On a clear day from Yongnuni Oreum, you can see Seongsan Ilchulbong Peak and Udo Island.

2.2 Tourism information Name Entity BIO tagging dictionary

In the pre-processing process, we remove special characters from the source data of tourism information and perform BIO (begin, inside, outside) tagging for the tourism information name entity used by the smart tourism chatbot. Korean and English versions of the tourism information Name Entity (NE) BIO tagging dictionaries are used to create the NER datasets from tourism information source data. Table 2 shows the Korean and English tourism information Name Entity BIO tagging dictionaries.

We create the tourism information Name Entity BIO tagging dictionary considering the following:

(1) Tourism information data consisting of compound nouns

We create a Korean BIO tagging dictionary by considering various spaces for the tourism information Name Entity consisting of compound nouns registered in the error patch dictionary due to an analysis error in the khaiii morpheme analyzer.

(2) Tourism information expressed in various notations

We create a BIO tagging dictionary considering the tourism information name entity expressed in various notations.

(3) Tourism information with various English notations

We create an English BIO tagging dictionary considering the tourism information name entity with various English notations.

Table 2. Tourism information Name Entity BIO tagging dictionaries.

(a) Korean NE BIO tagging dictionary

index	Name Entity	Tag
2705	한라산	A010000
2706	한라산국립공원	A010000
2707	한라산 국립공원	A010000
2708	한라산 국립 공원	A010000
2709	협재해수욕장	A020000
2710	에코랜드 테마파크	A030000
2711	올레 1 코스	A050000

(b) English NE BIO tagging dictionary

index	Name Entity	Tag
2705	Hallasan Mountain	Hallasan A010000 Mountain A010000_I
2706	Mountain Hallasan	Mountain A010000 Hallasan A010000_I
2707	Mt Hallasan	Mt A010000 Hallasan A010000_I
2708	Hallasan National Park	Hallasan A010000 National A010000_I Park A010000_I
2709	Hyeopjae Beach	Hyeopjae A020000 Beach A020000_I
2710	Eco Land Theme Park	Eco A030000 Land A030000_I Theme A030000_I Park A030000_I
2711	Olle trail roue # 1	Olle A050000 trail A050000_I route A050000_I # A050000_I 1 A050000_I

We intend to provide BIO tagging information according to the tourism information classification of the smart tourism information system that provides smart tourism services to provide tourism information provided through various channels such as smart tourism apps, Instagram, and YouTube as a smart tourism chatbot system. Table 1(a) is a dictionary for Korean tourism information name entity BIO tagging, and the tagging follows the tourism information classification in the smart tourism app. The tourism information classification code consists of 1 English letter and 6 numbers, and the first letter of the classification code, A, indicates a tourist destination, B indicates accommodation, and C indicates restaurants and cafes. The six numbers are divided into two and classified into major, medium, and sub-category. For example, classification code A010101 indicates Mt. Hallasan (medium) and Seongpanak Trail (sub) in the classification of mountains and Oreums (major). BIO tagging is performed according to the tourism information classification system of the smart tourism information system, and tourist destinations are grouped into one Name Entity in the NER model learning process.

2.3 Pre-processing for tourism information NER data generation

We perform BIO tagging using the tourism information name entity dictionaries in the pre-processing process. Figure 1 shows the process of generating NER data through preprocessing using the Name Entity BIO tagging dictionaries for source data. The process of creating Korean and English NER datasets is performed differently depending on the BIO tagging dictionary. The Korean NER dataset is created by finding the tourism information Name Entity in the BIO tagging dictionary in the source data and tagging the first word as Begin and the following words as Inside. For example, if the source data has a Name entity of 한라산 국립공원(Hallasan National Park) in the Korean BIO tagging dictionary of Table 2(a), it is tagged as 한라산(Hallasan) A010000 국립공원(National Park) A010000_I. The BIO tagging accuracy of the Korean NER dataset is determined according to the BIO tagging dictionary, just as the morpheme analysis accuracy is determined by the error-patch dictionary in the khaiii morpheme analyzer. The English NER dataset is created by finding the tourism information Name Entity in the BIO tagging dictionary in the source data and replacing the tagging information in the dictionary with sentence replacement. For example, if the source data has the name entity of Hallasan National Park in the English BIO tagging dictionary of Table 2(b), it is tagged as Hallasan Mountain A010101 National A010101_I Park A010101_I according to the tagging information in the dictionary. The BIO tagging accuracy of the English NER dataset is also determined according to the tagging information of the English BIO tagging dictionary. Table 3 shows the Korean and English NER learning datasets generated by preprocessing the source data using the BIO tagging dictionary. There are 20,768 words in the Korean NER learning data and 12,024 words in the English NER learning data.

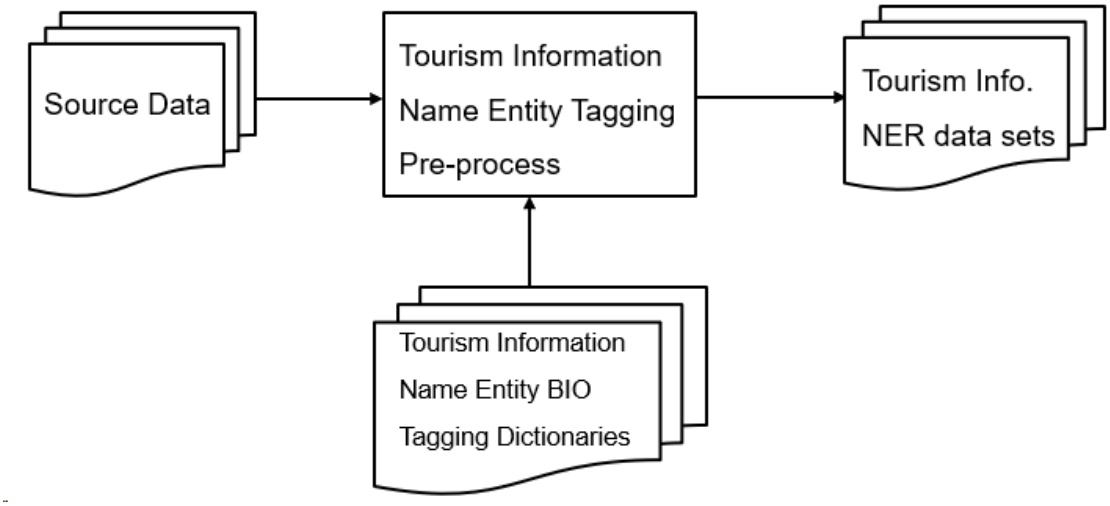
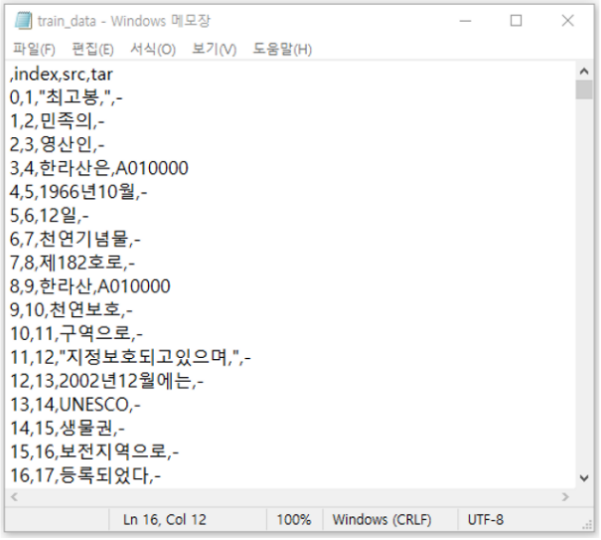


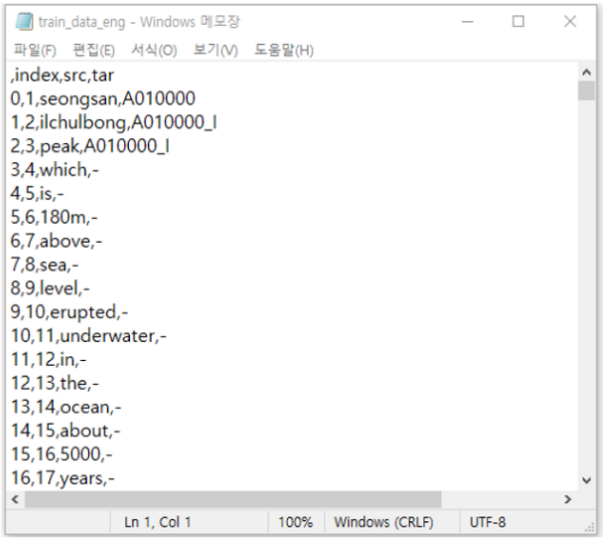
Figure 1. Preprocessing for NER dataset creation using Name Entity BIO tagging dictionaries.

Table 3. Korean and English tourism information NER datasets

(a) Korean NER dataset



(b) English NER dataset



3. Tourism Information NER Performance of the KoBERT-CRF NER model

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model developed by Google. The BERT model can be used for a wide variety of NLP tasks, including NER, DST, QA, and others. KoBERT is a Korean pre-trained language model developed by SK Telecom. KoBERT trained a large corpus composed of millions of Korean sentences collected from Wikipedia and news. The linear chain CRF (Conditional Random Fields) model has achieved significant improvement in NER tasks. In this paper, we use the KoBERT-CRF model [13] to train Korean and English tourism information NER datasets. Table 4 shows the tourism information NER performance the KoBERT-CRF model on the Korean and English Tourism Information NER datasets.

Table 4(a) shows the performance of the KoBERT-CRF NER model on the Korean tourism information

NER dataset as micro average, macro-average, and weighted average for F1, recall, and precision scores. To evaluate the tourism information NER performance of the KoBERT-CRF NER model, name entity tags according to tourism information classification are grouped into KorTour and KorTour_I. The number of tourism information entities tagged with KorTour and KorTour_I is 8,654 and 2095, respectively. The performance evaluation of the KoBERT-CRF NER model is performed only on the tourism information name entity tagged with KorTour and KorTour_I. The precision, recall, and F1 scores of the KorTour tag are 0.95, 0.96, and 0.95, and the precision, recall, and F1 scores of the KorTour _I tag are 0.88, 0.96, and 0.92. In the tourism information NER dataset, the number of KorTour _I tags is smaller than the KorTour tags, so the learning results seem to be poor. The micro-averaged precision, recall, and F1 of the KoBERT-CRF model on the Korean tourist information NER data are 0.93, 0.96 and 0.93. The macro-averaged precision, recall, and F1 scores are 0.91, 0.96 and 0.96. The weight-averaged precision, recall, and F1 scores are 0.93, 0.96 and 0.95 as shown in Table 4(a).

Table 4. NER performance of KoBERT-CRF on the Korean and English Tourism Information NER datasets

(a) Korean NER dataset

	Micro Avg	Macro Avg	Weighted Avg
Precision	0.93	0.91	0.93
Recall	0.96	0.96	0.96
F1-score	0.93	0.96	0.95

(b) English NER dataset

	Micro Avg	Macro Avg	Weighted Avg
Precision	0.94	0.91	0.94
Recall	0.89	0.88	0.89
F1-score	0.92	0.89	0.92

(C) Korean and English NER datasets

	Micro Avg	Macro Avg	Weighted Avg
Precision	0.93	0.90	0.93
Recall	0.95	0.95	0.95
F1-score	0.94	0.92	0.94

To evaluate the performance of the KoBERT-CRF NER model using English tourism information NER data, name entity tags according to tourism information classification are grouped into EngTour and EngTour_I. The total number of tourism information entities tagged with EngTour and EngTour_I are 11,52 4 and 2,892, respectively. The precision, recall, and F1 scores of the EngTour tag are 0.97, 0.91, and 0.93, and the precision, recall, and F1 scores of the EngTour _I tag are 0.85, 0.85, and 0.85. As in the Korean NER learning results, the number of EngTour_I tags is smaller than the EngTour tags, so the learning results seem to be poor. The micro-averaged precision, recall, and f1 of the KoBERT-CRF model on the English tourist information NER data are 0.94, 0.89 and 0.92. The macro-averaged precision, recall, and F1 scores are 0.91, 0.88 and 0.89. The weight-averaged precision, recall, and F1 scores are 0.94, 0.89 and 0.92 as shown in Table 4(b).

Table 4(c) shows the performance evaluation results of the KoBERT-CRF NER model using the NER datasets of Korean and English tourism information. The precision, recall, and F1 scores were 0.96, 0.95, 0.96

in KorTour tag, 0.89, 0.97, 0.93 in KorTour _I tag, 0.96, 0.93, 0.95 in EngTour tag, and 0.85, 0.87, 0.86 in EngTour _I tag. The micro-averaged precision, recall, and F1 of the KoBERT-CRF model on the Korean and English tourist information NER data are 0.93, 0.90 and 0.93. The macro-averaged precision, recall, and F1 scores are 0.95, 0.95 and 0.95. The weight-averaged precision, recall, and F1 scores are 0.94, 0.92 and 0.94, respectively.

4. Conclusions and Further Study

The smart tourism chatbot is a user interface that efficiently provides smart tourism services to tourists along with mobile apps, Instagram, YouTube, and blogs. We developed a smart tourism chatbot system using Khaiii morpheme analyzer, rule-based user's question intention identification, and Neo4J graph database as a knowledge base. To develop a competitive smart tourism chatbot system, it is necessary to develop NER, DST, and QA models based on PLMs. In this paper, we develop the Korean and English tourism information NER datasets for the development of NER model to understand the intention of users for tourism information questions in the previously developed chatbot system. We collect tourism information data and create Korean and English tourism information Name Entity tag dictionaries to create a tourism information NER dataset through the preprocessing process.

We perform tourism information NER performance evaluation using the Korean, English, Korean and English tourism information NER datasets using the KoBERT-CRF NER model. As a result of evaluating the tourism information NER performance, the performance of the Begin tag (KorTour, EngTour) is higher than that of the Inside tag (KorTour_I, EngTour_I) due to the difference in the number of Name entities in the training dataset. We plan to develop tourism domain DST and tourism information QA model using tourism information NER data set and tourism information name entity dictionaries.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2021R1A2C1093283).

References

[1] JeongWoo Jwa, “Development of Personalized Travel Products for Smart Tour Guidance Services”, International Journal of Engineering & Technology, 7 (3.33) 58-61, 2018.
DOI: DOI: 10.14419/ijet.v7i3.33.18524

[2] Dong-Hyun Kim, Hyeon-Su Im, Jong-Heon Hyeon, Jeong-Woo Jwa, “Development of the Rule-based Smart Tourism Chatbot using Neo4J graph database”, International Journal of Internet, Broadcasting and Communication, Vol.13, No.2, pp 179-186, 2021.
DOI: 10.7236/IJIBC.2021.13.2.179

[3] Kakao khaiii (Kakao Hangul Analyzer III), *https://tech.kakao.com/2018/12/13/khaiii/*

[4] Neo4j graph database, *https://neo4j.com/*

[5] Guendalina Caldarini, Sardar Jaf, Kenneth McGarry, ‘A Literature Survey of Recent Advances in Chatbots’, Information vol.13, no.1, 41, 2022.
DOI: 10.3390/info13010041

[6] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, Jason Weston, ‘Recipes for Building an Open-Domain Chatbot’, EACL 2021, pp. 300–325. 2021.

DOI: 10.18653/v1/2021.eacl-main.24,

[7] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, Xuanjing Huang, “Pre-trained Models for Natural Language Processing: A Survey”, Science China Technological Sciences 63(10), pp.1872-1897, 2020. DOI: 10.1007/s11431-020-1647-3

[8] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li, "A Survey on Deep Learning for Named Entity Recognition", IEEE Trans. on Knowledge and Data Eng., pp. 50-70, 2020. DOI:10.1109/TKDE.2020.2981314

[9] *https://github.com/kmounlp/NER*

[10] National Institute of the Korean Language NER data, *https://corpus.korean.go.kr/*

[11] NAVER NLP Challenge 2018 NER data, *https://github.com/naver/nlp-challenge/tree/master/missions/ner*

[12] Visit Jeju Website, *https://www.visitjeju.net/kr*

[13] SKT KoBERT, *https://github.com/SKTBrain/KoBERT*