ChatGPT와 KoBERT을 활용한 지리 정보 처리:
개체명 인식 성능 비교 연구

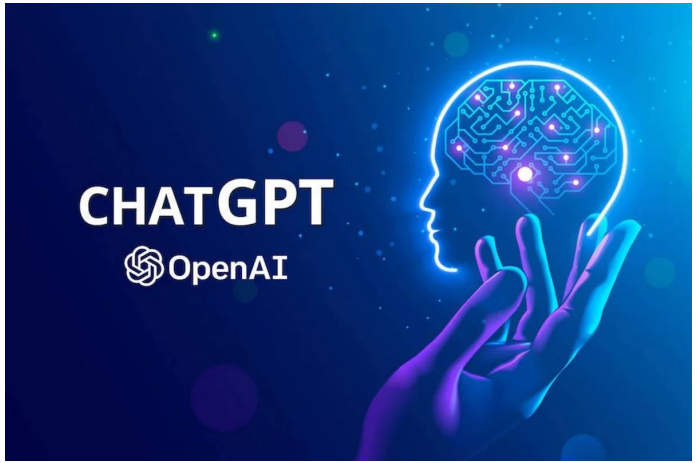
"Comparative Study of Geographical Information Processing with ChatGPT and KoBERT: Performance Comparison of Named Entity Recognition"

최대웅
Seoul National University
GIS/LBS Lab

Motivation

Motivation:

- 최근 **ChatGPT**의 개발로 **자연어 처리 분야**에서 **딥러닝 모델**은 많은 관심을 받고 있습니다.
- 하지만 **지리 정보 처리**에 딥러닝 모델을 적용하는 선행연구는 **많지 않습니다**.
- 지리 정보 처리에 자연어 딥러닝 모델을 적용함으로써 개체명 인식(Named Entity Recognition), 쿼리 생성과 같은 작업으로 지식 기반 질의응답(KBQA) 시스템 구축이 가능합니다.



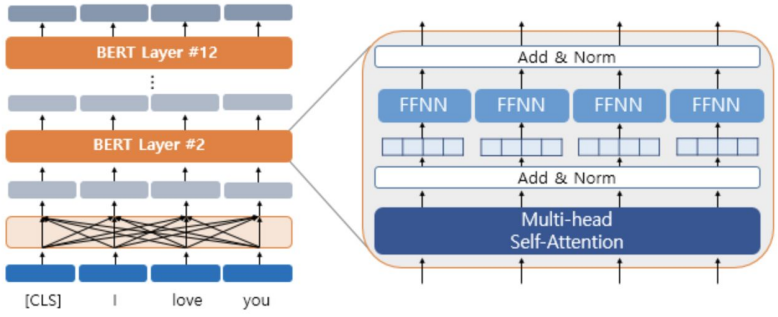
Background

Terminology of Area:

- **Named Entity Recognition (NER)**: 자연어 처리 분야에서 개체명 인식 작업을 의미
  a. 주어진 텍스트에서 중요한 개체(사람, 장소, 날짜, 조직 등)의 인식과 분류를 수행하는 작업
- **Entity Linking**: 개체명 해소 작업으로, 텍스트에서 인식된 개체를 지식 베이스나 데이터베이스에 연결하여 정확한 의미를 부여하는 과정
- **koBERT**: 한국어 자연어 처리를 위해 사전 학습된 BERT (Bidirectional Encoder Representations from Transformers) 모델
- **Fine-tuning**: 사전 학습된 모델을 특정 작업에 맞게 세부 조정하는 과정
- **Knowledge-Based Question Answering (KBQA)**: 지식 베이스를 기반으로 한 질문에 대한 답변을 제공하는 시스템이나 작업

Problem Definition

1. 이 연구의 목표는 공간 질의 처리 분야에서 **Named Entity Recognition(NER)**을 효과적으로 수행하는 딥러닝 모델을 개발을 목표로 합니다.
2. 현재까지 지리 정보 처리에 딥러닝 모델을 적용하는 연구는 많지 않으며, 특히 공간 질의에 대한 NER 선행 연구가 많지 않습니다.
3. **ChatGPT**와 **koBERT** 모델을 활용하여 개체명 인식 작업을 수행합니다.
4. 지리 정보 처리에 자연어 딥러닝 모델을 적용함으로써 개체명 인식(Named Entity Recognition), 쿼리 생성과 같은 작업으로 **지식 기반 질의응답(KBQA)** 시스템 구축을 위한 기반 마련을 목표로 합니다.



SKTBrain/**KoBERT**

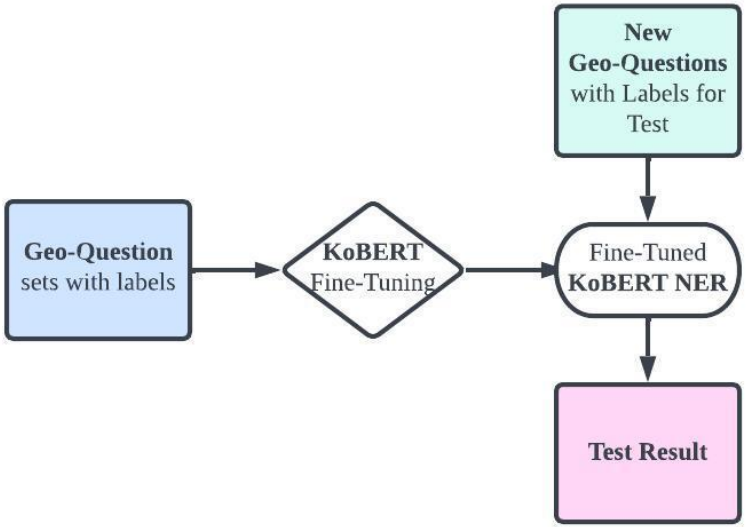
Korean BERT pre-trained cased (KoBERT)

5 Contributors | 3 Issues | 627 Stars | 159 Forks

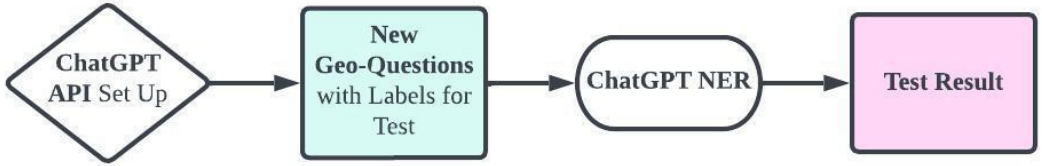
Solving Approach

Overall Pipeline

KoBERT



ChatGPT



Solving Approach

1. 지리 질문 데이터 세트를 **koBERT**의 **Training** 데이터 형식에 맞게 재생성합니다.(7500개)
   - koBERT Training-Data 형식
     i. Ex) **"질문 문장 \t entity_label(없으면 0)"**
     ii. 질문과 entity는 **tab**으로 구분 나머지는 스페이스

entity_list	NL_substituted
신반포21차아파트	신반포21차아파트의 인구는 몇명이야?
방배래미안아파트	방배래미안아파트의 인구는 몇명이야?
아크로리버파크아파트	아크로리버파크아파트의 인구는 몇명이야?
방배자이아파트	방배자이아파트의 인구는 몇명이야?
마제스타시티힐스테이트서리풀아파트	마제스타시티힐스테이트서리풀아파트의 인구는 몇명이야?
신화아파트	신화아파트의 인구는 몇명이야?
서초자이아파트	서초자이아파트의 인구는 몇명이야?
서초네이처힐3단지아파트	서초네이처힐3단지아파트의 인구는 몇명이야?
한라비발디스튜디오193(도시형)아파트	한라비발디스튜디오193(도시형)아파트의 인구는 몇명이야?
반포르엘아파트	반포르엘아파트의 인구는 몇명이야?
신반포4차아파트	신반포4차아파트의 인구는 몇명이야?
방배e-편한세상1차아파트	방배e-편한세상1차아파트의 인구는 몇명이야?
반포미도2차아파트	반포미도2차아파트의 인구는 몇명이야?
힐스테이트서초젠트리스아파트	힐스테이트서초젠트리스아파트의 인구는 몇명이야?
현대성우아파트	현대성우아파트의 인구는 몇명이야?
디에이치라클라스아파트	디에이치라클라스아파트의 인구는 몇명이야?
래미안신반포팰리스아파트	래미안신반포팰리스아파트의 인구는 몇명이야?
방배그랑자이아파트	방배그랑자이아파트의 인구는 몇명이야?
래미안퍼스티지아파트	래미안퍼스티지아파트의 인구는 몇명이야?
서초롯데캐슬프레지던트아파트	서초롯데캐슬프레지던트아파트의 인구는 몇명이야?
신반포21차아파트	신반포21차아파트의 인구는 몇명이야?

koBERT > KoBERT-NER > data > train.tsv

```
1  신반포21차아파트 인구는 몇명이야?  LOC-B 0 0
2  방배래미안아파트 인구는 몇명이야?  LOC-B 0 0
3  아크로리버파크아파트 인구는 몇명이야?  LOC-B 0 0
4  방배자이아파트 인구는 몇명이야?  LOC-B 0 0
5  마제스타시티힐스테이트서리풀아파트 인구는 몇명이야?  LOC-B 0 0
6  신화아파트 인구는 몇명이야?  LOC-B 0 0
7  서초자이아파트 인구는 몇명이야?  LOC-B 0 0
8  서초네이처힐3단지아파트 인구는 몇명이야?  LOC-B 0 0
9  한라비발디스튜디오193(도시형)아파트 인구는 몇명이야?  LOC-B 0 0
10 반포르엘아파트 인구는 몇명이야?  LOC-B 0 0
11 신반포4차아파트 인구는 몇명이야?  LOC-B 0 0
12 방배e-편한세상1차아파트 인구는 몇명이야?  LOC-B 0 0
13 반포미도2차아파트 인구는 몇명이야?  LOC-B 0 0
14 힐스테이트서초젠트리스아파트 인구는 몇명이야?  LOC-B 0 0
15 현대성우아파트 인구는 몇명이야?  LOC-B 0 0
16 디에이치라클라스아파트 인구는 몇명이야?  LOC-B 0 0
17 래미안신반포팰리스아파트 인구는 몇명이야?  LOC-B 0 0
18 방배그랑자이아파트 인구는 몇명이야?  LOC-B 0 0
19 래미안퍼스티지아파트 인구는 몇명이야?  LOC-B 0 0
20 서초롯데캐슬프레지던트아파트 인구는 몇명이야?  LOC-B 0 0
```

Solving Approach

2. **재생성된** 지리 질문 세트를 이용하여 **koBERT** 모델을 **fine-tuning** 합니다.

3. Fine-tuned 된 모델을 사용하여 **새롭게 생성한(Test)** 지리 질문 데이터 세트에 대해 **Named Entity Recognition (NER)** 작업을 수행합니다.

- F1 Score: **0.96**

```
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
06/01/2023 14:35:51 - INFO - trainer -   ***** Running training *****
06/01/2023 14:35:51 - INFO - trainer -     Num examples = 5100
06/01/2023 14:35:51 - INFO - trainer -     Num Epochs = 20
06/01/2023 14:35:51 - INFO - trainer -     Total train batch size = 32
06/01/2023 14:35:51 - INFO - trainer -     Gradient Accumulation steps = 1
06/01/2023 14:35:51 - INFO - trainer -     Total optimization steps = 3200
06/01/2023 14:35:51 - INFO - trainer -     Logging steps = 1000
06/01/2023 14:35:51 - INFO - trainer -     Save steps = 1000
Iteration: 100%| | 160/160 [00:25<00:00,  6.16it/s]
Iteration: 100%| | 160/160 [00:26<00:00,  6.12it/s]
Iteration: 100%| | 160/160 [00:26<00:00,  6.09it/s]
Iteration: 100%| | 160/160 [00:26<00:00,  6.07it/s]
Iteration: 100%| | 160/160 [00:26<00:00,  6.07it/s]
Iteration: 100%| | 160/160 [00:26<00:00,  6.06it/s]
Epoch:  30%|  [02:37<06:08, 26.33s/it0
  warnings.warn('{} seems not to be NE tag.'.format(chunk))
06/01/2023 15:12:35 - INFO - trainer -   ***** Eval results *****
06/01/2023 15:12:35 - INFO - trainer -     f1 = 0.955223880597015
06/01/2023 15:12:35 - INFO - trainer -     loss = 0.12490617483854294
06/01/2023 15:12:35 - INFO - trainer -     precision = 0.9696969696969697
06/01/2023 15:12:35 - INFO - trainer -     recall = 0.9411764705882353
06/01/2023 15:12:35 - INFO - trainer -
              precision    recall  f1-score   support

         LOC       1.00      0.94      0.97        17
          UN       0.94      0.94      0.94        17

   micro avg       0.97      0.94      0.96        34
   macro avg       0.97      0.94      0.96        34
weighted avg       0.97      0.94      0.96        34
```

```
44  계단식이야? UNK UNK
45
46  안성아파트는 LOC-B LOC-B
47  개별난방이야 UNK UNK
48  중앙난방이야? UNK UNK
49
50  홍대아파트는 LOC-B LOC-B
51  지하주차장이 UNK UNK
52  있어? UNK UNK
53
54  인사동의 LOC-B LOC-B
55  월세가 UNK UNK
56  가장 UNK UNK
57  저렴한 UNK UNK
58  아파트는 UNK UNK
59  어디야? UNK UNK
60
61  20대가 UNK UNK
62  많이 UNK UNK
63  거주하는 UNK UNK
64  강남구는? LOC-B UNK
```

Solving Approach

1. **ChatGPT API**를 사용하여 지리 **Test 질문 세트**에 대해 **Named Entity Recognition (NER)** 작업을 수행합니다.

```
response = openai.ChatCompletion.create(
    model="gpt-3.5-turbo",
    messages=[
        {"role": "user", "content": f'"{question}" 문장만 Named Entity Recognition 해줘'}
```

```
import openai
import csv
import pandas as pd
from fuzzywuzzy import fuzz
import time

API_KEY = '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'
openai.api_key = API_KEY

# 입력 파일 경로
input_file_path = "/Users/daewoong/Documents/서울대학교/NLP/연구 및 프로젝트/pororo_ner/question_r300.csv"
# 출력 CSV 파일 경로
output_file_path = "/Users/daewoong/Documents/서울대학교/NLP/2023 춘계 학술대회/chatGPT_output.csv"

df = pd.read_csv(input_file_path)

with open(output_file_path, "w", newline="") as output_file:
    writer = csv.writer(output_file)
    writer.writerow(["Question", "Predicted Entity", "Actual Entity", "Fuzzy Score"])

    for index, row in df.iterrows():
        question = row["NL_substituted"]
        entity = row["entity_list"]

        # NER 수행
        response = openai.ChatCompletion.create(
            model="gpt-3.5-turbo",
            messages=[
                {"role": "user", "content": f'"{question}" 문장만 Named Entity Recognition 해줘 PER, ORG, LOC 를 사용해줘'}
            ]
        )
        predicted_entities = response['choices'][0]['message']['content'].split("\n")
        predicted_entities_dict = {}
        for entity_info in predicted_entities:
            if ":" in entity_info:
                entity_type, entity_value = entity_info.split(":")
                entity_type = entity_type.strip()
                entity_value = entity_value.strip()
                predicted_entities_dict[entity_type] = entity_value

        fuzzy_scores = {}
        for entity_type in ["PER", "ORG", "LOC"]:
            predicted_entity = predicted_entities_dict.get(entity_type, "없음")
```

chatGPT_output

Question	Predicted Entity
인사동의 인구는 몇명이야?	{'ORG': '없음', 'LOC': '인사동'}
동탄파라곤아파트의 평균 소득은 얼마야?	{'ORG': '동탄파라곤아파트', 'LOC': '없음'}
수원대림아파트의 준공년도는 언제야?	{'ORG': '수원대림아파트', 'LOC': '수원'}
롯데타워의 가구당 주차대수는 몇 대야?	{'ORG': '롯데타워', 'LOC': '없음'}
63빌딩의 가장 높은 층은 몇 층이야?	{'ORG': '63빌딩', 'LOC': '없음'}
서울대학교를 만든 건설사는 뭐야?	{'ORG': '서울대학교', 'LOC': '없음'}
낙성대아파트 용적률은 얼마야?	{'ORG': '낙성대아파트', 'LOC': '없음'}
반포자이아파트의 제공하는 편의시설에는 어떤게 있어?	{'ORG': '반포자이아파트', 'LOC': '없음'}
전주아파트는 복도식이야 계단식이야?	{'ORG': '전주아파트', 'LOC': '없음'}
안성아파트는 개별난방이야 중앙난방이야?	{'ORG': '안성아파트', 'LOC': '없음'}
홍대아파트는 지하주차장이 있어?	{'ORG': '홍대아파트', 'LOC': '지하주차장'}
인사동의 월세가 가장 저렴한 아파트는 어디야?	{'ORG': '없음', 'LOC': '인사동'}
20대가 많이 거주하는 강남구는?	{'ORG': '강남구', 'LOC': '없음'}
교통사고가 많이 나는 동탄은?	{'ORG': '없음', 'LOC': '동탄'}
범죄율이 높은 관악구는?	{'ORG': '없음', 'LOC': '관악구'}
화성롯데캐슬아파트의 가까운 병원는?	{'ORG': '화성롯데캐슬아파트', 'LOC': '없음'}
부산해운대의 가까운 공원는?	{'ORG': '없음', 'LOC': '부산, 해운대, 공원'}

Evaluation

Test Dataset에 대한 성능 평가 지표

	KoBERT	ChatGPT
F1 Score	0.955524	0.576
Precision	0.969697	0.487
Recall	0.941176	0.412

- **정밀도(Precision)**: 정밀도는 모델이 양성(Positive)으로 예측한 샘플 중 실제로 양성인 샘플의 비율을 의미.
  정밀도 = TP / (TP + FP)
- **재현율(Recall)**: 재현율은 실제로 양성인 샘플 중에서 모델이 정확하게 양성으로 예측한 샘플의 비율을 의미.
  재현율 = TP / (TP + FN)
- **F1 스코어 = 2 * (정밀도 * 재현율) / (정밀도 + 재현율)**

Conclusion

1. **KoBERT**를 지리 정보 데이터셋으로 **Fine-Tuning**한 모델의 정확도가 **ChatGPT**보다 높았습니다.
2. ChatGPT는 많은 질문을 **request**하면 **overload**가 걸려 많은 수의 데이터를 처리하는데 **어려움**이 있습니다.
3. 지리 정보를 개체명 인식해 지리 **KBQA 시스템** 발전의 **기반**을 마련했습니다.
4. **공간정보**를 기반으로 한 **자연어 처리** 작업에서의 **NER**의 **중요성**을 강조하며, 공간 질의에 대한 정확한 답변을 제공하는 시스템의 필요성을 **제시** 합니다.

감사합니다