
Mental-2024

Preprocessing with LLaMA

모바일시스템공학과
이승재

0. facebook/m2m100_418M??

>
전부터 기억력이 떨어져서 작년에도 검사를 받아왔다 (2019년 MMSE: 26)
1 쯤부터 기억력이 더 떨어졌다
고 못찾다 나중에는 찾음(주1회)
고 엉뚱한 곳에서 내렸다 (2021.01)- 오늘은 버스타고 한정거장 전에 내려서 걸어왔다
맛이 없다(당뇨로 치아가 좋지 않아 틀니를 했음)
치매여서 걱정이 된다

<병력>

고혈압(+) 당뇨(+) 고지혈증(+)
복용약물 : 고혈압약, 당뇨약, 고지혈증약
수술 : -
뇌졸중 : 2020.12. 2021.01 저혈당으로 쓰러져
수가 없어서 8일 입원했었다 -지금은 괜찮음) :
즈사?

1
2 NC_Kim Yonghye 520215
3
4 Name: Kim Yong
5 The gender:
6 Age: 69 (52 years old)
7 Degree: 6 years
8 Read, Write, and Number: Possible
9 Resident family: Single (divorce 2020.01) -010-2466-5812
10
11
12 The Case Conference
13 Normal
14 The family needs regular tracking tests.
15
16 ??
17 The guard.
18 Two or three years ago the memory was down and last year was tested (MMSE: 26)
19 Since 2021.01, the memory has fallen further.
20 I can't find it afterwards.
21 The bus went down from the wrong place (2021.01) - Today I went down before the bus station and walked down.
22 There is no taste of food.
23
24 You are worried about dementia.
25
26 The Army
27 High blood pressure(+) Diabetes(+) High blood pressure(+)
28 Drugs: High Blood Pressure, Diabetes, High Blood Pressure
29 The operation:
30 A stroke: 2020.12.2021.01 falls down to low blood glucose and MRI screenshot (when I was in 8 days because I c
31 The symptoms?
32 The Special Army?
33 The sight: - Hearing
34 Eating: - If you sleep:
35 Drink : - tobacco : -
36 Family: your sister dementia

1. read_data.py – 1

■현재 복용 중인 약을
기관지약(폐): 10년 전부터 하루 세 번

- 두부 외상/ 뇌출증: 없음
- 알레르기 : 2006년부터 이마 쪽 피부 알레르기가 있는데 약을 발라도 낫지 않음.
- 수술: 없음.
- 시력: 문제없음. / 청력: 거의 안 들림.

■생활습관

- 수면: 허리가 많이 굽어 있어서 잘 때 어려움. 자기 보다는 거의 깨어 있다시피 함. 낮잠
- 식사 : 원래는 잘 먹었는데 2015년 5월 초부터 먹는 것을 이유 없이 갑자기 거부해서 입
- 술/ 담배: 안함.

translate.py



```
def convert_text(text):
    """
    텍스트 내에서 괄호 안의 ('+', '-', '0', 'x')를 ('1', '0', '1', '0')으로 치환하고,
    'Y'는 1, 'N'은 0으로 변환합니다.
    또한 특수문자 '■', '★', '●'을 공백으로 변경합니다.

    Parameters:
    |   text (str): 변환할 텍스트.

    Returns:
    |   str: 변환된 텍스트.
    """
    if not isinstance(text, str):
        return text # 문자열이 아닌 경우 변환 없이 그대로 반환

    # ✅ 특수문자 변환
    text = text.replace("■", " ")
    text = text.replace("★", " ")
    text = text.replace("●", " ")
    text = text.replace("→", " ")
```

"If there are special characters in the original text,
some parts cannot be translated in the
“translate.py” so they are left blank."

1. read_data.py - 2

```
# ✓ 괄호 안의 값 변환
text = re.sub(
    r"\(([^\)]+)\)",
    lambda m: "(" + "".join(["1" if char in ['+', 'O'] else "0" if char in ['-','X'] else char for char in m.group(1)]) + ")",
    text
)

# ✓ "Y", "N" 치환
text = re.sub(r"\bY\b", "1", text)
text = re.sub(r"\bN\b", "0", text)
```

Standardized

"+", "O", "Y"	→ int 1
"-", "X", "N"	→ int 0

혼자 화장실 사용? (O)

절반이상(+),

단순한 명령 이해? Y

"Integer for standardization"

1. read_data.py - 3

```
import json
```

```
if output_json:  
    # JSON 변환 후 반환  
    json_data = df.to_dict(orient="records") # 각 행을 JSON 객체로 변환  
    return json.dumps(json_data, ensure_ascii=False, indent=4)
```

"Convert csv file to JSON file in read_data.py to standardize data."

CSV 파일 상위 20줄 내용:
변환된 JSON 데이터:

```
[  
  {  
    "ID": "D1",  
    "SEX": "여",  
    "D_TEST": "9/23/14",  
    "AGE": 97,  
    "CGA2": "이정임\n\n딸 함께 내방함\\n\\n무학(ey)\\n한글 읽고 쓰기 미숙 / 숫자 가능\\n사별 / 아들 내외 함께 거주\\n<현병력>\\n혈압(1) 당뇨(0) 고지혈증(0)\\n약물복용 ; 혈압약 20년 정도 복용 중\\n수술 ; 없음\\n기타 : 없음\\n시력 ; 백내장 수술(양쪽) 후 잘 보이십\\n청력 ; 잘 못 들으신다. 보청기 안 끼심, 대화 불가\\n식사 ; 조금 드신다. 잘 못 드심\\n수면 ; 잘 주무신다.\\n술, 담배 : 안 함\\n\\n다리 허리가 아파서 잘 못걷는다.\\n\\n보호자 코멘트>\\n시간 관념 없어진 것 같다.\\n새벽 3시, 5시에 아들 깨우면서 일어나야 되지 않냐고 하고, \\n새벽에 샤워하려고 해서 이 간에 씻으시느냐고 하니까 언제 씻든 무슨 상관이냐 하셨다\\n바우처 했었는데 요양보호사가 할머니 안 입는 옷 있으면 달라고 했는지 쫒는지 모르겠는데 그 후로 옷 없어지면 없어졌다고요 요양보호사한테 가져갔다고 해서 매번인 적 있다.\\nTV 같은 거 보면 이상한 소리 한다. TV 앞에서 옷 갈아 입으면 저기 사람 있는데 옷 갈아입는다고 뭐라고 하고, 과일 깎으면 TV앞에다 과일 내민다.\\nTv에서 나쳐다본다고 한다. TV에 나오는 사람을 현실과 헷갈린다.\\n<기억력>\\n기억력 저하 호소함. 증상 4-5년 정도, 요즘 더 심해졌다.\\n서서히 시작, 점진적 진행, 현재 일상생활 지장 (없음)\\n수도를 켜놓고 잊어버린다.\\n물건 잊어버리는 경우 중요한 물건 도장 둘째등. 아들 잠자는 방에 가서 계속 뒤진다.\\n아침 식사 : 밥, 김치, 미역국, 어제 저녁 식사 : 알 수 없다.\\n흔히 쓰는 물건 ; 밖으로 훌자 안 나가서 알 수 없다.\\n최근 뉴스, 연속극 내용 : 내용 기억 못 하신, 귀가 안 들려서 하나도 안 들려서 그림만 본다.\\n냄비 태우는 경우 : 요리 안 함\\n매일 먹는 약 없음 혈압이 안 올라서 안 드신다고 말 하심(재자 않고 생각하시기에 그런 것 같다)\\n자식 이름 기억 : 2남1녀(김영현, 성현, 정분)\\n손자 손녀 이름 기억 : 얼굴보면 다 기억함\\n\\n<언어>\\n유창성 문제없음. WFD(0), Naming diff(0) 이해력 (0)\\nNEARING DIFF(1)\\n\\n<지남력> 2\\n연월일요일 DK/10/DK/DK/가을 (0,0,0,0,1)\\n기념일 생일 원래 잘 못챙긴다.\\n장소: 익숙한 장소 문제 없다. 익숙하지 않은 장소 문제 없다.\\n      => 알 수 없음, 항상 모시고 다니심\\n사람: 잘 알아봄.\\n\\n<판단 및 문제해결> 2\\n이해력 저하 있음 . (0) 사회적 판단력(0) 예의범절 (0)\\n\\n<성격, 행동>\\n무울(1) 흥미소실(0), 식욕저하(0), 체중변화(0), 수면장애(1), 지체초조(1), 피로감(0),\\n무가치감 또는 죄책감(1), 집중력저하(0), 자살보고(1)\\n*매일 훈자 있으셔서 항상 무울해 하신다. 74년도 아들 일음, 항상 무울해하심, 아들 얘기만 하면 우신다.\\n 한번은 옥상에 올라가서 죽으려고 했었다 함, 준비를 해놨단 말도 하십\\n\\n<사회활동> 2\\n모임 : 없다\\n산책 : 안 함\\n종교활동 : 없음\\n운동 : 안 함\\n\\n<가정생활> 2\\n취미활동 : 바느질 하셨는데 지금은 못 하신다.\\n음식 만들기, 음식맛 : 요리 안 함\\n가전제품 사용 문제 (1) 리모콘 사용 문제(1) 전화 걸기 문제 (1), 전화 받기 문제(0)\\n돈계산(0), 용돈관리(0), 통장 관리(0), 은행업무(0) : 직접 관리 불가능\\n\\n<개인일상생활> 2\\n옷입기, 식사하기, 세수, 목욕, 옷갈아입기 - 독립 수행 가능하나 미숙하다. 샤워 등은 도와드려야 한다.",  
    "FINAL_DX": "PSD"  
  },
```

2. translate.py - 1

```
def chunk_text(text, chunk_size=500):
    """ 긴 텍스트를 chunk_size 크기로 나누어 리스트로 반환합니다. """
    return [text[i:i + chunk_size] for i in range(0, len(text), chunk_size)]
```

"We set chunk_size to 500 to ensure natural translation and context while maintaining high accuracy."

1 What if chunk_size is too large?

Pros : Translation models can translate by considering a wider context
→ improved connectivity between sentences

Cons : Long response time
→ Because there is a lot of data to process at once

2 What if chunk_size is too short?

Pros : Translation speed is fast because less data is processed at once

Cons : Loss of contextual information
→ Information from previous sentences may not be remembered and may be translated separately

2. translate.py - 2

```
for chunk in text_chunks:  
    user_prompt = (  
        "Translate the following text from Korean to English."  
        "Ensure that all Korean text, including any sections inside angle brackets ('< >'), is fully translated."  
        "Translate everything exactly as it is, without summarizing, shortening, or removing any details."  
        "Provide only the translated text without extra explanations."  
        "Do not omit any part of the text.\n\n"  
        f"Text:{\n{chunk}\n\n"  
        "Translated Text:"  
    )
```



```
2601 Please note that translating medical records can be a delicate task, as some information may be sensitive or personal.  
2602 This translation is provided to facilitate understanding and should not be used for any purpose that could compromise the individual's privacy or confidentiality.
```

"When I translated using LLaMA, I got a problem that didn't occur when I translated using the facebook/m2m100_418M model, so I gave the following command in the prompt"

2. translate.py - 3

```
#  번역이 원본과 너무 비슷하면 재번역 수행
if translated_text == chunk:
    print("⚠ 번역이 누락된 부분 감지됨. 재번역 시도...")
    response = client.generate(
        model=model_name,
        prompt=user_prompt
    )
    translated_text = response.response.strip()

translated_chunks.append(translated_text)
```

GDS-K: Depression (0), Anxiety (0), Irritability (0)
The person is experiencing anxiety and irritability, which affects their daily life.

<Daily Life>
I experience dizziness when standing up quickly or after physical exertion.
In cold weather, I wear winter clothing less frequently than usual.", I can't provide a translation of an empty text
"I can't provide a translation for the given text as it appears to be incomplete or missing content. If you could p

The original request asked me to translate a given Korean text into English, however, the input provided contains r
아들, 자부 함께 내방함
무학(0y)
한글 읽고 쓰기 불가능 / 숫자 불가
독거(사별, 2000년)

<현병력>
혈압(1) 당뇨(0) 고지혈(0)
약 ; 약 먹는 것 좋아하지 않으심(보건소에서 혈압약 주시는데 안 먹는다)
수술 ; 수술 없음
기타 ; 없음
시력 ; 침침하다, 잘 보이는 편 ■ 청력 ; 잘 들리는 편 ■
식사 ; 잘 먹는 편 ■ 수면 ; 잘 자는 편이다 ■
술, 담배 ; 안 함

<기억력>
기억력저하 호소 함. 요즘 자고 일어나면 때를 모른신다 ■ 급격히 그러셨다. 몇 달 안된다. 정신 놓으실 때 있다.■ 시간 지나면 괜찮
현실감각 없은 꿈얘기를 주로 많이 한다 ■ 갑자기 시작, 점진적 진행, 현재 일상생활 지장 (있음)■ 물건 잊어버리는 경우 중요한 틀
아침식사 ; 밥, 아육회, 김치 (0, 콩나물 시금치 된장국) ■ 어제 저녁 식사 ; 기억 안 난다 ■ 흘미소설(0), 식욕저하(0), 체중변화
최근 뉴스 ; TV를 보시길 하는데 그냥 보고만 있다. ■ 희노애락이 없어졌다 ■ 연속극 ; 안 본다 ■ 냄비 태우는 경우 ; 요리 못 함■

As shown in the picture, there are cases where translation.py does not translate certain text and returns it as is, so I added code to retranslate it.

2. translate.py - 4

```
def translate_json_cga2(json_file_path):
    """
    JSON 파일에서 "CGA2" 항목만 번역합니다.
    """

    try:
        print(f"Processing JSON file: {json_file_path}")

        # 📁 JSON 파일 읽기
        with open(json_file_path, "r", encoding="utf-8") as json_file:
            json_data = json.load(json_file)

        print(f" TRANSLATING 'CGA2' field in JSON data...")

        translated_json = []
        for entry in json_data:
            translated_entry = entry.copy() # 원본 유지

            if "CGA2" in translated_entry and isinstance(translated_entry["CGA2"], str):
                text_to_translate = translated_entry["CGA2"].strip()
```

"Most fields consist of numbers, codes, abbreviations, etc., which may result in performance degradation due to unnecessary data processing, but CGA2 is a field that contains descriptive Korean sentences about the patient's health and daily life status, so translation is required."

"ID": "D1",
"SEX": "여",
"D_TEST": "9/23/14",
"AGE": 97,

"CGA2": "Eunjeong\n\nDaughter is staying with me\nNo education (0 years)\nReading and writing in Hangul are not proficient / numbers are possible\nSeparation /
"FINAL_DX": "PSD"

3. result

```
∨ data
  {} 0_AI_LKH_AD_converted.json
  [] 0_AI_LKH_AD_translated_1rows(LLaMA model).csv
  [] 0_AI_LKH_AD_translated_1rows(LLaMA model)2.csv
  [] 0_AI_LKH_AD_translated_2rows(facebook model).csv
  [] 0_AI_LKH_AD_translated_2rows(LLaMA model).csv
  [] 0_AI_LKH_AD_translated_3rows(LLaMA model)3.csv
  [] 0_AI_LKH_AD_translated_3rows(LLaMA model)4.csv
  [] 0_AI_LKH_AD_translated_3rows(LLaMA model)5.csv
  {} 0_AI_LKH_AD_translated_CGA2.json
  {} 0_AI_LKH_AD_translated.json
  [] 0_AI_LKH_AD_translated(LLaMA).csv
  [] 0_AI_LKH_AD_translated(LLaMA)(20_data_set_using_chunk).csv
  [] 0_AI_LKH_AD_translated(LLaMA)(20_data_set_using_chunk2).csv
  [] 0_AI_LKH_AD_translated(LLaMA)(20_data_set_using_chunk3).csv
  [] 0_AI_LKH_AD_translated(LLaMA)(20_data_set).csv
  [] 0_AI_LKH_AD_translated(with GPT).csv
  [] 0_AI_LKH_AD_translated(without GPT).csv
```

```
data > {} 0_AI_LKH_AD_translated_CGA2.json > ...
1  [
2   {
3     "ID": "D1",
4     "SEX": "남",
5     "D_TEST": "9/23/14",
6     "AGE": 97,
7     "CGA2": "Eunjeong\n\nDaughter is staying with me\nNo education (0 years)\nReading and writing in Hangul are not proficient / numbers are possible\nSeparation / Son and his family live together\nAvailable Forces>\nBlood pressure (1)\nDiabetic (1)\nHigh Cholesterol (0)\n\nTaking blood pressure medicine\n\nCurrent Military Status\n\nResidency status: Living alone, living with his second son in front of An-dong"
8   },
9   {
10    "ID": "D2",
11    "SEX": "남",
12    "D_TEST": "6/27/16",
13    "AGE": 97,
14    "CGA2": "Name: Kim Hyun Jae      Gender: Male      Age: 97 years      Insurance: Life/Health\n\nEducation: 9 years      Reading and writing in Korean: Possible      Numbers: Possible\n\nResidency status: Living alone, living with his second son in front of An-dong"
15    "FINAL_DX": "PSD"
16  },
17  {
18    "ID": "D3",
19    "SEX": "남",
20    "D_TEST": "5/20/14",
21    "AGE": 95,
22    "CGA2": "Park Gyu-eum\n\nSon, I will go to my separate place.\n\n\nIlliterate (0 years)\nReading and writing in Hangul are impossible / no numbers allowed\nSolo (separated, 2000)\n\n\nCurrent Military Status\n\nBlood pressure (1)\nDiabetic (1)\nHigh Cholesterol (0)\n\nTaking blood pressure medicine\n\nCurrent Military Status\n\nResidency status: Living alone, living with his second son in front of An-dong"
23    "FINAL_DX": "PRD"
24  },
25  {
26    "ID": "D4",
27    "SEX": "남",
28    "D_TEST": "2003-04-14",
29    "AGE": 94,
30    "CGA2": "Recently\nTook him with my son\nLack of education (0y)\nCannot read or write in Hanji / unable to use numbers\n\n\nCurrent Health Status\n\nBlood pressure (1)\nDiabetic (1)\nHigh Cholesterol (0)\n\nTaking blood pressure medicine\n\nCurrent Military Status\n\nResidency status: Living alone, living with his second son in front of An-dong"
31    "FINAL_DX": "PRD"
32  },
33  {
34    "ID": "D5",
35    "SEX": "남",
36    "D_TEST": "6/20/14",
37    "AGE": 92,
38    "CGA2": "Kang Il-seon\n\nEducation: Illiterate\nReading and Writing: Poor\nNumbers: None\n\nVision: Blurred\nHearing: Difficulty hearing on both sides\nAlcohol, Tobacco: Abstinence\nDiet: Thin\nSleep: Insomnia\n\nWaking up at night\n\n\nCurrent Military Status\n\nResidency status: Living alone, living with his second son in front of An-dong"
39    "FINAL_DX": "PRD"
40  },
41  {
42    "ID": "D6",
43    "SEX": "남",
44    "D_TEST": "2005-11-16",
45    "AGE": 94,
46    "CGA2": "Name: Kim Seol-yeon      Gender: Female      Age: 94\n\nEducation: 2 (y)\n\nReading and writing: Incomplete\n\nNumbers: Incomplete\n\nResidency status: Living alone, living with his second son in front of An-dong"
47    "FINAL_DX": "PSD"
48  }
```

4. Conclusion

<이상희,F,85><보호자: 손자 010-4425-3208>
손자와 내방
국졸(6y)
한글 읽고 쓰기 가능 / 수자 가능

1-3. 실제생년월일
1929년 월 일 (시 분)
1-4. 연령
만 92 세
1-5. 읽기
 불가능 ① 미숙 ② 가능
1-6. 성별
 여 ① 남
1-7. 쓰기
 불가능 ① 미숙 ② 가능
1-8. 교육
0 년
1-9. 결혼상태
 결혼 ① 사별 ② 별거/이혼 ③ 미혼 ④ 기타(기술 :)
1-10. 독신기간
40년



1. "The remaining task is to explore the remaining data and find and process non-formal data like the above to reduce exceptions."
2. "The task of investigating and extracting specific items other than those included in most patients' medical records"

<현병력>	<기억력>
<언어>	<지남력>
<판단및문제해결>	<성격, 행동>
<사회활동>	<가정생활>
<가정및취미활동>	<개인일상생활>
<주호소>	

+

<보호자 코멘트>
<과거병력>
<약물복용>
<두부외상>
<뇌졸중>
<기타>
<이해력, 판단력>
<증상>
<가족부담및사회적 환경>
...