20240424 Meeting

312707003 黃鈺婷

數據預處理

- 將中文文本進行預處理

```
with open("data/《斗破苍穹》.txt", "r", encoding="utf-8") as fp:
    #使用strip()去除每行文本的開頭和結尾的空白字符
    #再使用split("\n")將文本按行拆分成列表
    data = fp.read().strip().split("\n")
sentences = []

for d in data:
    d = d.strip()
    if "===" in d or len(d) == 0 or d == "《斗破苍穹》来自:":
        continue
    sentences.append(d)

with open("data/corpus.txt", "w", encoding="utf-8") as fp:
    fp.write("\n".join(sentences))
```

訓練中文詞表

- 使用sentencepiece來訓練中文詞庫

```
import sentencepiece as spm
spm.SentencePieceTrainer.train(
    input='data/corpus.txt',
    model_prefix='tokenizer',
    vocab_size=50000,
    user_defined_symbols=['foo', 'bar'],
    character_coverage=1.0,
    model_type="bpe",
)
```

- 得到tokenizer.model和tokenizer.vocab

使用transformers加載sentencepiece模型

```
chinese_sp_model_file = "tokenizer.model"

# load
chinese_sp_model = spm.SentencePieceProcessor()
chinese_sp_model.Load(chinese_sp_model_file)

chinese_spm = sp_pb2_model.ModelProto()
chinese_spm.ParseFromString(chinese_sp_model.serialized_model_proto())

## Save
output_dir = './transformers_tokenizer/chinese/'
os.makedirs(output_dir, exist_ok=True)
with open(output_dir + 'chinese.model', 'wb') as f:
    f.write(chinese_spm.SerializeToString())
tokenizer = ChineseTokenizer(vocab_file=output_dir + 'chinese.model')

tokenizer.save_pretrained(output_dir)
```

使用transformers加載sentencepiece模型

```
# Test
#從指定的目錄中加載已保存的tokenizer模型
chinese_tokenizer = ChineseTokenizer.from_pretrained(output_dir)
```

```
Test text:
 白日依山尽，黄河入海流。欲穷千里目，更上一层楼。
The primary use of LLaMA is research on large language models,
including
```

```
Tokenized by Chinese tokenizer:['▁', '白日', '依', '山', '尽', ',',
'黄', '河', '入', '海', '流', '。', '欲', '穷', '千里', '目', ',', '更',
'上一层', '楼', '。', '▁', 'T', 'h', 'e', '▁', 'p', 'r', 'i', 'm', 'a',
'r', 'y', '▁', 'u', 's', 'e', '▁', 'o', 'f', '▁', 'LL', 'a', 'MA',
'▁i', 's', '▁', 'r', 'e', 's', 'e', 'a', 'r', 'ch', '▁', 'o', 'n',
'▁', 'l', 'a', 'r', 'g', 'e', '▁', 'l', 'an', 'g', 'u', 'a', 'g',
'e', '▁', 'm', 'o', 'd', 'e', 'l', 's', ',', '▁i', 'n', 'c', 'lu',
'd', 'i', 'ng']
```

合併中文詞表到llama中

```
llama_tokenizer_dir = "transformers_tokenizer/llama/tokenizer.model"
chinese_sp_model_file = "tokenizer.model"

# load
llama_tokenizer = LlamaTokenizer.from_pretrained(llama_tokenizer_dir)
chinese_sp_model = spm.SentencePieceProcessor()
chinese_sp_model.Load(chinese_sp_model_file)

llama_spm = sp_pb2_model.ModelProto()
llama_spm.ParseFromString(llama_tokenizer.sp_model.serialized_model_proto())
chinese_spm = sp_pb2_model.ModelProto()
chinese_spm.ParseFromString(chinese_sp_model.serialized_model_proto())


print(len(llama_tokenizer), len(chinese_sp_model))
32000 50000
```

合併中文詞表到llama中

```
# Add Chinese tokens to LLaMA tokenizer
llama_spm_tokens_set = set(p.piece for p in llama_spm.pieces)
print(len(llama_spm_tokens_set))
print(f"Before:{len(llama_spm_tokens_set)}")
for p in chinese_spm.pieces:
    piece = p.piece
    if piece not in llama_spm_tokens_set:
        new_p = sp_pb2_model.ModelProto().SentencePiece()
        new_p.piece = piece
        new_p.score = 0
        llama_spm.pieces.append(new_p)
print(f"New model pieces: {len(llama_spm.pieces)}")
```

```
Before:32000
New model pieces: 81163
```

合併中文詞表到llama中

```
# Test
llama_tokenizer = LlamaTokenizer.from_pretrained(llama_tokenizer_dir)
chinese_llama_tokenizer = LlamaTokenizer.from_pretrained(output_hf_dir)
```

```
Tokenized by LLaMA tokenizer:['_', '白', '日', '<0xE4>', '<0xBE>',
'<0x9D>', '山', '<0xE5>', '<0xB0>', '<0xBD>', '，', '黃', '河', '入',
'海', '流', '。', '<0xE6>', '<0xAC>', '<0xB2>', '<0xE7>', '<0xA9>',
'<0xB7>', '千', '里', '目', '，', '更', '上', '一', '<0xE5>', '<0xB1>',
'<0x82>', '<0xE6>', '<0xA5>', '<0xBC>', '。', '<0x0A>', 'The',
'_primary', '_use', '_of', '_L', 'La', 'MA', '_is', '_research',
'_on', '_large', '_language', '_models', ',', '_including']
```

```
Tokenized by Chinese-LLaMA tokenizer:['_白', '日', '依', '山', '尽',
'，', '黄', '河', '入', '海', '流', '。', '欲', '穷', '千里', '目', '，',
'更', '上一层', '楼', '。', '<0x0A>', 'The', '_primary', '_use', '_of',
'_L', 'La', 'MA', '_is', '_research', '_on', '_large', '_language',
'_models', ',', '_including']
```

三種標記器比較

```
Tokenized by Chinese tokenizer:['_', '白日', '依', '山', '尽', ',',
'黄', '河', '入', '海', '流', '。', '欲', '穷', '千里', '目', ',', '更',
'上一层', '楼', '。', '_', 'T', 'h', 'e', '_', 'p', 'r', 'i', 'm', 'a',
'r', 'y', '_', 'u', 's', 'e', '_', 'o', 'f', '_', 'LL', 'a', 'MA',
'_i', 's', '_', 'r', 'e', 's', 'e', 'a', 'r', 'ch', '_', 'o', 'n',
'_', 'l', 'a', 'r', 'g', 'e', '_', 'l', 'an', 'g', 'u', 'a', 'g',
'e', '_', 'm', 'o', 'd', 'e', 'l', 's', ',', '_i', 'n', 'c', 'lu',
'd', 'i', 'ng']

Tokenized by LLaMA tokenizer:['_', '白', '日', '<0xE4>', '<0xBE>',
'<0x9D>', '山', '<0xE5>', '<0xB0>', '<0xBD>', '，', '黄', '河', '入',
'海', '流', '。', '<0xE6>', '<0xAC>', '<0xB2>', '<0xE7>', '<0xA9>',
'<0xB7>', '千', '里', '目', '，', '更', '上', '一', '<0xE5>', '<0xB1>',
'<0x82>', '<0xE6>', '<0xA5>', '<0xBC>', '。', '<0x0A>', 'The',
'_primary', '_use', '_of', '_L', 'La', 'MA', '_is', '_research',
'_on', '_large', '_language', '_models', ',', '_including']

Tokenized by Chinese-LLaMA tokenizer:['_白', '日', '依', '山', '尽',
'，', '黄', '河', '入', '海', '流', '。', '欲', '穷', '千里', '目', '，',
'更', '上一层', '楼', '。', '<0x0A>', 'The', '_primary', '_use', '_of',
'_L', 'La', 'MA', '_is', '_research', '_on', '_large', '_language',
'_models', ',', '_including']
```