2020 졸업작품 프로젝트

1st Implementation Demo

김지은 | 채민형 | 최병규 | 최지원

목차
INDEX

01 개 요

01 과정

초기 : 개체명인식을 통한 한국어 문장, 문단의 토픽 유추

중기 : 개체명인식으로 문단의 키워드 체크 및 문단 요약

<추가 수정 사항>

1) 구현할 언어 모델을 구체화 : Seq2Seq 모델

2) “문단 요약 " 이 아닌, “의도 분석 " 으로 기능 수정

3) 개체명인식과 **의도분석**을 통합해 하나의 모델로 구현

02 최종 주제

: **Seq2Seq 모델**을 이용한 한국어 **개체명인식기** 및 **의도분석기** 구현

03 최종 산출물

: 입력된 문장의 개체명인식 결과와 의도분석 결과를 반환하는 웹사이트

04 프로젝트 구체화

1) 모델 : Seq2Seq

: Seq2Seq, 즉 sequence-to-sequence는 임의 길이의 한 시퀀스를
다른 종류의 시퀀스로 변환하는 확률모델

2) 기능 : 개체명인식 & 의도분석

: 개체명인식(NER : Named Entity Recognition)은 태깅 활용해 시퀀스 생성

: 의도분석(Intent Classification)은 다중 분류의 일종이나 시퀀스변환(번역)으로 접근

3) 방법

: 하나의 **Encdoer**로 문장을 입력하고,
개체명인식과 의도분석 각각의 결과를 출력하는 두 개의 **Decoder**로 구성
(Encoder : 입력 문장 / Decoder 1 : 개체명인식 결과, Decoder 2 : 의도분석 결과)

04 프로젝트 구체화

4) 데이터셋

: 건국 NLP Lab에서 제공받은 음식주문관련 데이터를 클리닝, 가공해 데이터셋 구성

5) 기존 연구와의 차별성

: 하나의 모델로 두 가지의 결과를 도출한다.

: 태깅, 분류를 시퀀스 생성이라는 새로운 시각으로 바라본다.

02 프로젝트 요구사항

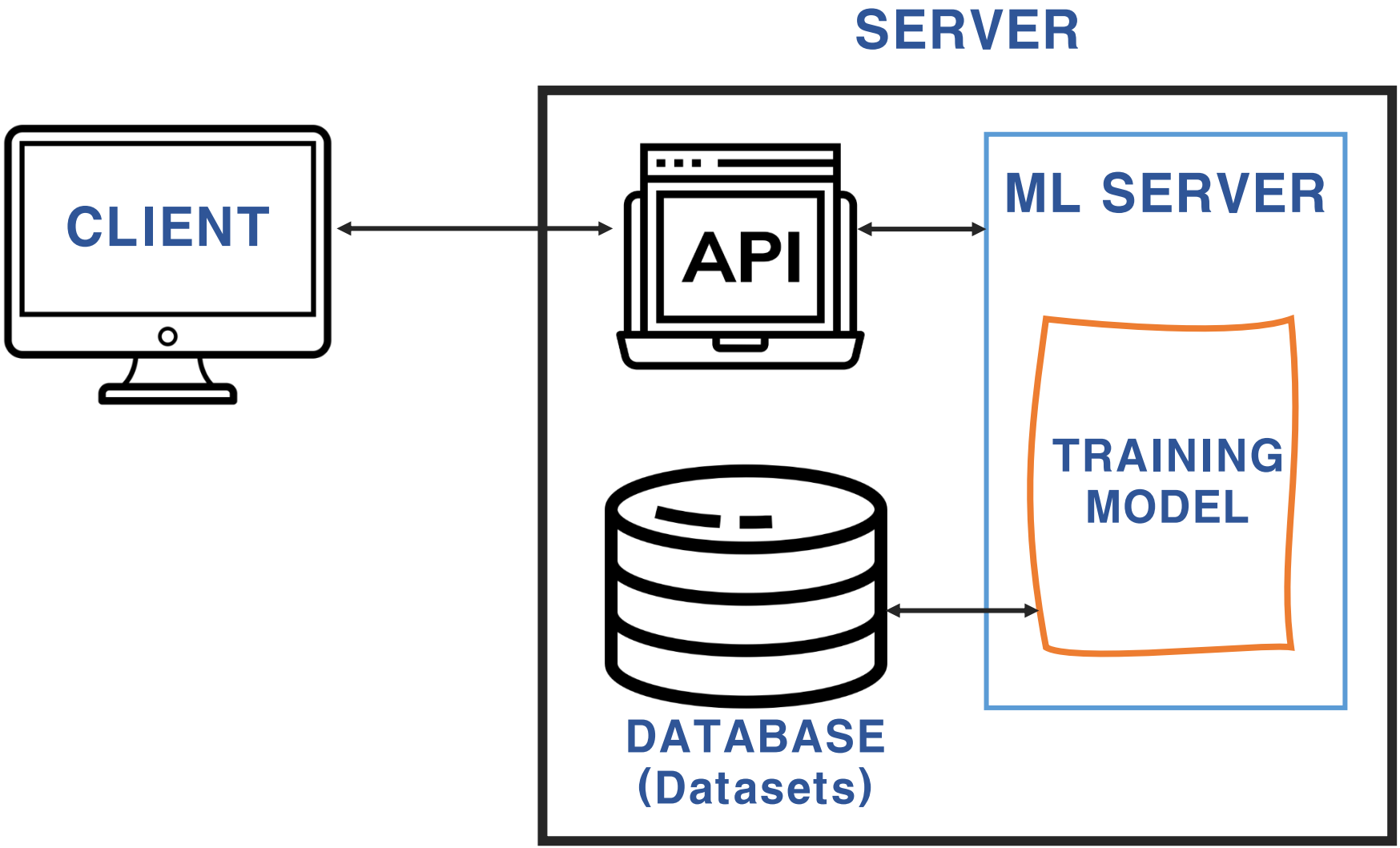
01 기능적 요구사항

1) 사용자가 원하는 글을 원하는 방식으로 입력할 수 있다.
2) 결과물을 디스플레이할 수 있다.

02 비기능적 요구사항

1) 응답시간 5초이내로 주어진 글에 대한 개체명인식, 의도분석이 완료되어야 한다.
2) 개체명인식, 의도분석의 정확도가 각각 70%이상이 되어야 한다.

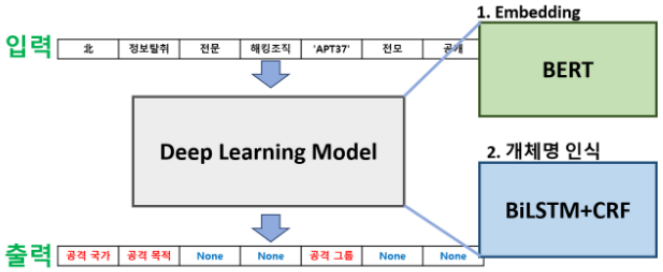
03 전체 시스템 아키텍처



04 NER : 개체명 인식기

01 기존의 개체명인식

: BERT를 이용한 Embedding, BiLSTM+CRF 구조를 이용한 개체명인식 모델이 가장 많이 사용된다.

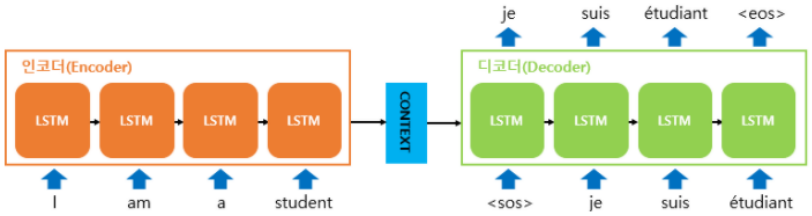


NER 순서, 임베딩 -> NER

02 기존의 Seq2Seq 모델

: LSTM, GRU 등 RNN cell을 길게 쌓아 방대한 양의 sequenc를 처리하는데 특화되었으며,
기계 번역에 탁월한 성과를 보여준다.

: Encoder와 Decoder로 구성되며, Encoder에서 input을 고정된 크기의 context vector로 만들고,
Decoder는 context vector로 output을 만든다.



04 NER : 개체명 인식기

03 우리 프로젝트에서의 변형과 활용

1) Seq2Seq 모델 기반의 개체명인식

: 일반적인 NLP문제는 Nto1(ex. 긍정or부정 분류) 문제이지만, 개체명인식 문제는 NtoN 문제(Sequence labeling)

-> 한 종류의 문장을 다른 종류의 문장으로 바꿔주는 **Seq2Seq 모델** 활용

2) Encoder : RNN이 아닌, **CNN** 기반의 자질 추출

- computer vision 분야에서 주로 사용하는 CNN의 적용
- 이미지 전체 영역에 대해 필터를 이용해 패턴을 스스로 학습하는 CNN의 특징 : 빠른 성능

3) Decoder : **BiLSTM** 이용해 대응하는 개체명을 예측

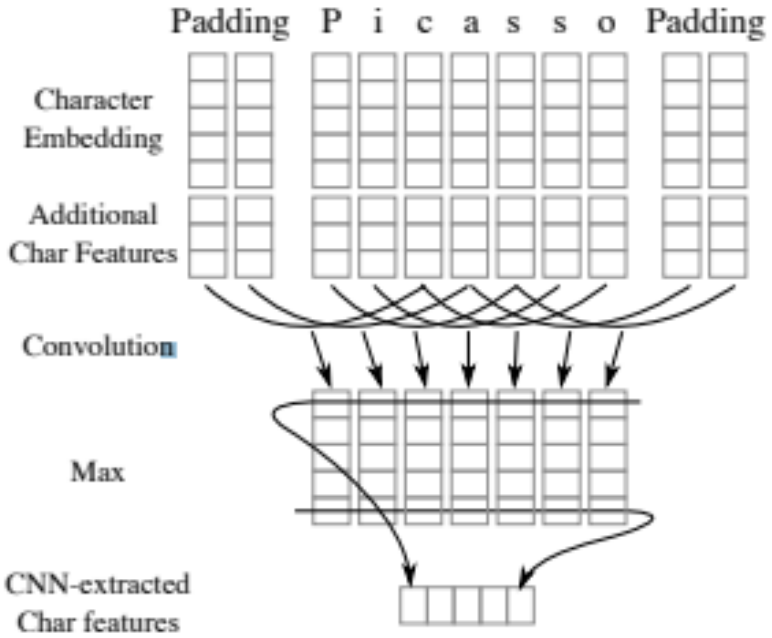
- 학습데이터를 사용해 Encoder에서 도출된 각각의 자질들을 연결하고, BiLSTM의 입력으로 사용
- 최종적으로 형태소에 대응하는 개체명을 예측

" CNN을 이용한 빠른 학습과 BiLSTM을 이용한 정확한 예측 기대 가능 "

04 NER : 개체명 인식기

04 모델 구조

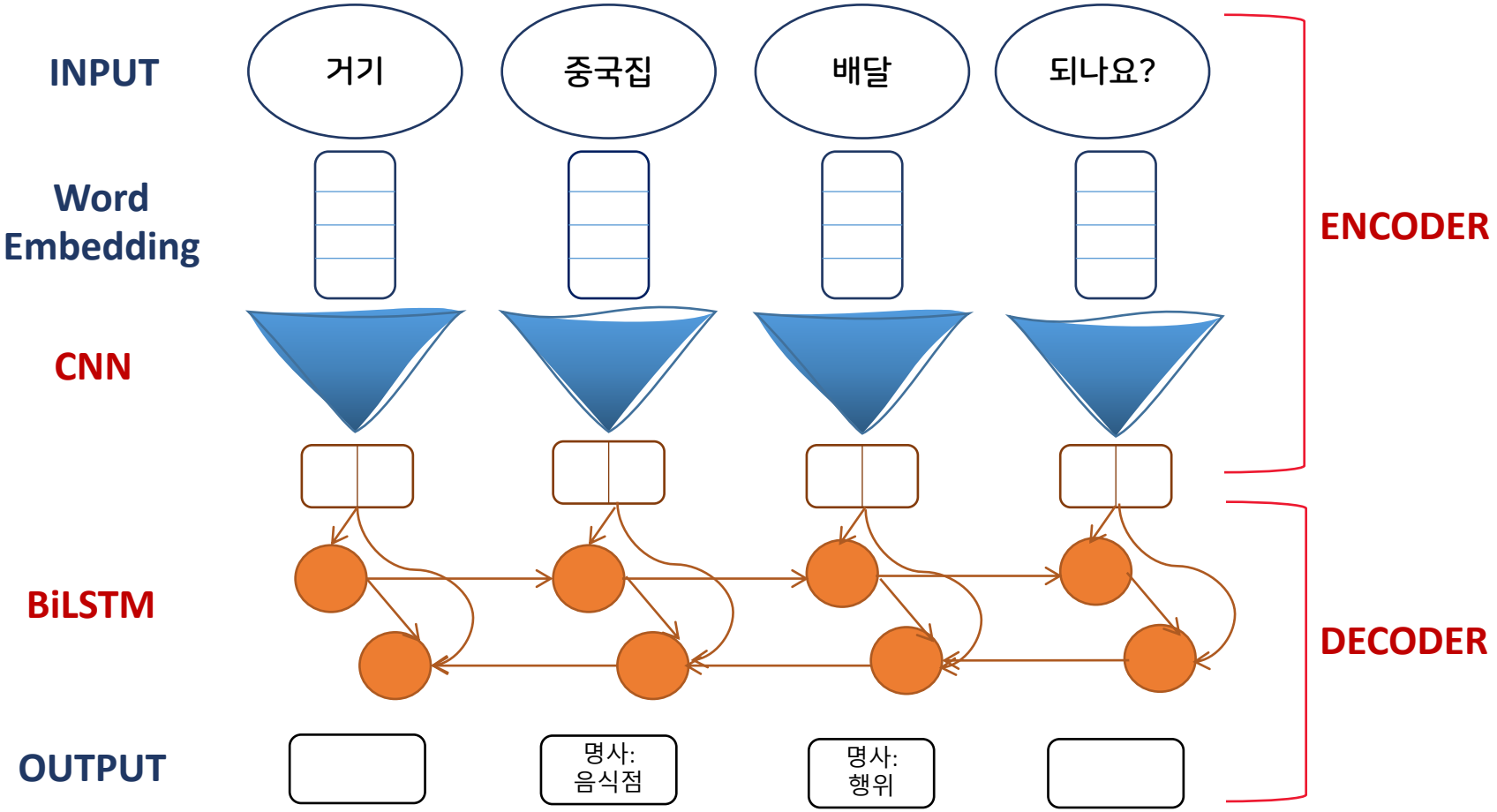
1) CNN 기반의 자질 추출 세부 구조



04 NER : 개체명 인식기

04 모델 구조

2) CNN과 BiLSTM 기반의 Seq2Seq 모델



04 NER : 개체명 인식기

05 모델 구현 결과

```
Seq2Seq(
  (encoder): Encoder(
    (embed): Embedding(28386, 100, padding_idx=0)
    (trainable_embed): Embedding(28386, 100, padding_idx=0)
    (char_embed): Embedding(2284, 100, padding_idx=0)
    (pos_embed): Embedding(232, 100, padding_idx=0)
    (convs1): ModuleList(
      (0): Conv2d(1, 128, kernel_size=(2, 100), stride=(1, 1))
      (1): Conv2d(1, 128, kernel_size=(3, 100), stride=(1, 1))
      (2): Conv2d(1, 128, kernel_size=(4, 100), stride=(1, 1))
      (3): Conv2d(1, 128, kernel_size=(5, 100), stride=(1, 1))
    )
    (dropout): Dropout(p=0.5, inplace=False)
    (fc1): Linear(in_features=400, out_features=10, bias=True)
  )
  (decoder): Decoder(
    (lstm): LSTM(818, 200, num_layers=2, batch_first=True, dropout=0.6, bidirectional=True)
  )
)
```

04 NER : 개체명 인식기

06 데이터셋 상세 내용

1) 개체명 태그 정보

: 전처리 데이터의 개체명 분석 후 사용 빈도가 많은 것을 선정

TAG	정의	세부사항
MN	메뉴	음식 메뉴로 완제품을 의미한다. 반찬을 포함한다.
IN	재료	완제품이 아닌 상태의 식재료를 의미한다. 소스를 포함한다.
DR	음료	커피, 콜라 등 액체류의 식음료를 의미한다.
CN	수량	물건에 대한 수량을 의미한다.
SI	사이즈	음식 및 물건에 대한 사이즈를 의미한다. 액체의 경우, 리터 단위, 메뉴의 경우 대/중/소 등을 포함한다.
PR	가격	물건의 가격을 의미한다.
PM	결제 수단	현금, 카드 등 결제 수단을 의미한다.
PC	인원수	사람의 수를 의미한다.
PX	식기	숟가락, 젓가락 등 식기를 의미한다.
DA	날짜	숫자 형태의 날짜, 혹은 오늘/내일 등을 포함한다.
TI	시간	숫자 형태의 시간, 오전/오후/저녁 등을 포함한다.
LO	장소	장소를 의미한다. 시/군/도 단위와 화장실 등 작은 단위의 장소를 모두 포함한다.

04 NER : 개체명 인식기

06 데이터셋 상세 내용

2) 현재 진행 상황

: 건국 NLP Lab 에서 제공한 전처리 데이터를 바탕으로, 말뭉치를 구축했다.

: 각각의 자질을 PKL 파일로 저장하여 사용한다.

파일명	자질	설명
vocab	형태소	non-static word2vec, static word2vec (mecab 사용, gensim으로 word2vec)
char_vocab	음절단위	character cnn
pos_vocab	POS	형태소에 따른 품사. mecab 사용
gazette	사전정보	사전에 처리한 형태소와, 그에 따른 Tag 쌍
lex_dict	형태소와 그에 따른 Tag	gazette을 vocabulary object로 변환

04 NER : 개체명 인식기

06 데이터셋 상세 내용

3) 현재 상황 척도

: 8376문장 중 현재 가공된 188문장을 학습시키고, 임의로 추가한 46문장의 테스트 결과.

-> F1 Score : 0.1947 (만점 1 기준) / Accuracy : 0.70 (만점 1기준)

```
Test:
Epoch [18/30], Step [30/63], Loss: 0.0681, accuracy: 0.6996, F1 Score: 0.1947, Max F1 Score: 0.1947lassification_report:
              precision    recall  f1-score   support

        B_CN       0.00      0.00      0.00        12
        B_DA       0.00      0.00      0.00         3
        B_DR       0.17      0.11      0.13         9
        B_FX       0.00      0.00      0.00         6
        B_IN       0.15      0.33      0.21         6
        B_LO       0.00      0.00      0.00         4
        B_MN       0.50      0.25      0.33        16
        B_PC       0.20      0.33      0.25         3
        B_PM       0.20      0.50      0.29         2
        B_PR       0.00      0.00      0.00         3
        B_SI       0.00      0.00      0.00         5
        B_TI       0.00      0.00      0.00         2
           I       0.58      0.68      0.62        28
           O       0.82      0.97      0.89       154

    accuracy                           0.70       253
   macro avg       0.19      0.23      0.19       253
weighted avg       0.61      0.70      0.65       253
```

05 Intent Classification : 의도 분석기

01 Seq2Seq 모델 기반의 의도분석

: 의도분석은 다중분류 문제의 일종이나, 기계번역 방식으로 접근하여 Seq2Seq 모델을 활용했다.

예시) Input : 혼자 먹을 수 있는 게 어떤 것이 있나요?

Output : 1인 메뉴 문의

05 Intent Classification : 의도 분석기

02 데이터셋

: 문장-의도 형태의 총 8347개의 데이터가 존재한다.

	SENTENCE	MAIN
4589	김밥 한 줄 여기서 먹을게요	식사주문
3743	온누리상품권 있는데 이걸로 결제가능해요?	상품권결제
6736	이거 저희가 주문한 거 아닌데요	주문과다른메뉴가전달된상황
1006	얼마나 기다려야 하나요?	대기시간문의
4205	치킨 시키려고 하는데요	식사배달요청
...	...	...
5351	혼 닭은 혼자 먹는 거다 그죠?	양에대한질문
7004	주차비는요?	주차비문의
2182	그럼 단품만 주문하는게 가능한가요?	메뉴주문문의
2882	밥도 같이 주나요?	밥포함여부문의
4480	아니요 안 맵게 해서요 짬뽕 하나 해주세요	식사주문

8347 rows × 2 columns

05 Intent Classification : 의도 분석기

03 진행 상황

: 총 8347개의 데이터 중 6498개는 학습에, 1892개는 검증에 사용했다.

: 총 500회 학습 후 검증한 결과 아래와 같은 정확도를 보여줬다.

Test accuracy: 76.80%, F1-Score: 82.03%

	precision	recall	f1-score	support
1인분배달문의	1.00	1.00	1.00	2
1인식사자리요청	0.67	0.67	0.67	6
가격문의	0.76	0.92	0.83	38
가격변경문의	1.00	1.00	1.00	1
결제문의	0.00	0.00	0.00	1
결제방식선택	0.83	0.89	0.86	28
계산서요청	0.89	0.80	0.84	10
계절메뉴문의	1.00	0.67	0.80	3
계좌번호요청	1.00	1.00	1.00	1
국물요청	1.00	0.50	0.67	2
국물제공문의	0.67	1.00	0.80	2
기본반찬요구	1.00	1.00	1.00	3
기본반찬종류문의	1.00	1.00	1.00	1
남은음식포장요구	0.86	0.86	0.86	7
냅킨요구	1.00	1.00	1.00	1
냉난방시설문의	1.00	1.00	1.00	2
넵킨물수건요청	0.80	0.80	0.80	5
단체주문문의	1.00	1.00	1.00	1
대기시간문의	0.80	1.00	0.89	4
두메뉴의차이에대한질문	1.00	0.80	0.89	5
런치메뉴주문문의	1.00	1.00	1.00	1
런치타임문의	1.00	1.00	1.00	1
룸가능인원문의	0.00	0.00	0.00	1
룸요구	1.00	1.00	1.00	1
리필문의	1.00	1.00	1.00	1
맛에대한질문	0.76	0.72	0.74	18
맛조절요청	1.00	0.33	0.50	6
매일달라지는'오늘의메뉴'문의	1.00	0.50	0.67	2
토핑문의	1.00	1.00	1.00	1
특정메뉴우선요청	0.00	0.00	0.00	1
특정메뉴종류문의	1.00	0.50	0.67	2
특정재료가포함되는지문의	0.75	0.75	0.75	4
특정재료첨삭요구	1.00	1.00	1.00	1
특정재료첨삭요청	1.00	1.00	1.00	1
판매하는술종류브랜드문의	1.00	1.00	1.00	2
포인트적립문의	0.75	1.00	0.86	3
포장메뉴문의	1.00	1.00	1.00	1
포장문의	0.75	0.60	0.67	10
포장비문의	1.00	1.00	1.00	1
포장용기문의	0.50	1.00	0.67	1
포장주문	0.67	0.83	0.74	12
포장할인문의	0.00	0.00	0.00	2
할인메뉴문의	1.00	1.00	1.00	2
할인문의	0.82	0.69	0.75	13
행사이름문의	1.00	1.00	1.00	1
현금영수증발행요청	1.00	0.83	0.91	6
현금할인문의	1.00	1.00	1.00	1
홀배달가격차이문의	1.00	1.00	1.00	2
화장실문의	1.00	1.00	1.00	15
후식문의	1.00	1.00	1.00	2
휴일문의	0.85	0.92	0.88	12
accuracy			0.77	1849
macro avg	0.85	0.82	0.82	1849
weighted avg	0.77	0.77	0.76	1849

06 향후 계획

01 [개체명인식] 현재 제공된 전처리데이터 8376문장 중
가공된 188문장을 제외한 나머지 문장 가공 및 학습 데이터 추가 후 성능향상

02 Cnn+Bilstm을 Seq2Seq 모델에 맞게 변경,
개체명인식 Decoder와 의도분석 Decoder를 하나의 공통된 Encoder로 연결

03 최적의 Batch Size와 Learning Rate을 찾고 성능 향상에 주력

04 가장 성능이 좋은 모델로 웹사이트 구축

감 사 합 니 다