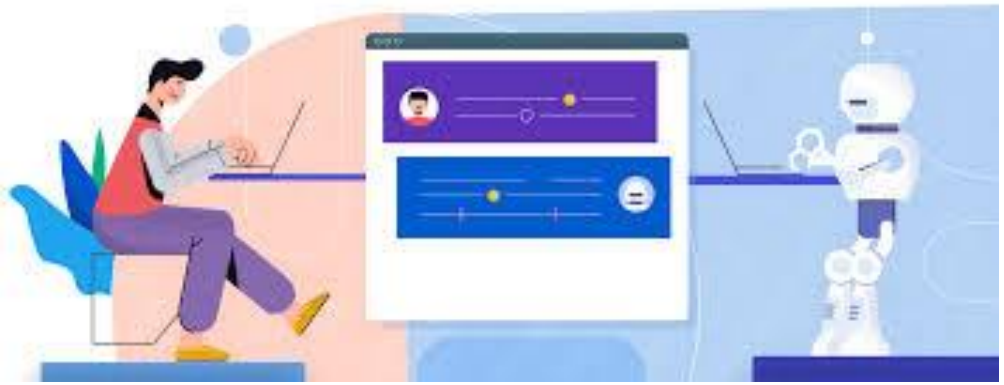
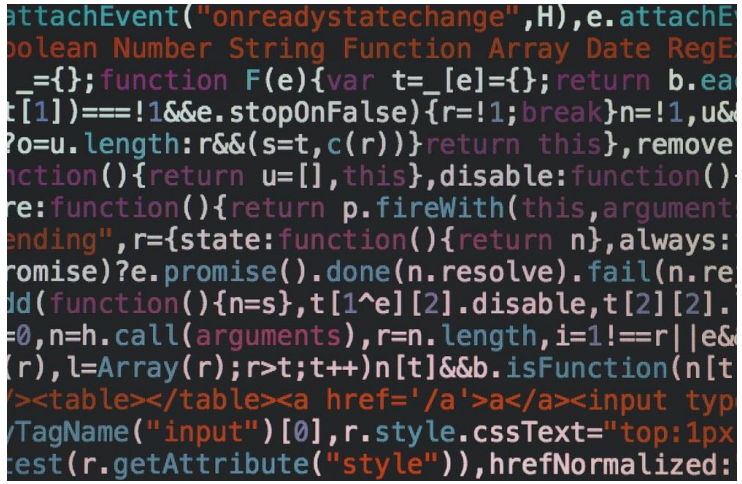
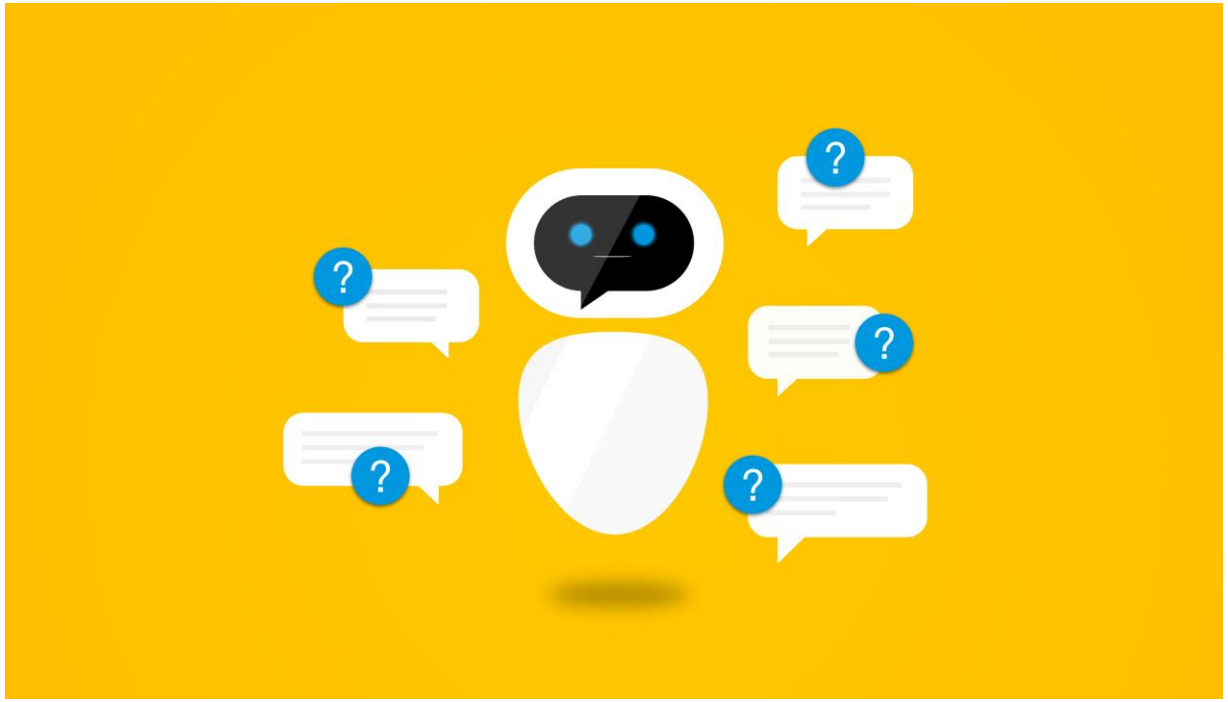




텍스트마이닝의 화룡점정 '챗봇' 만드는 여정

석사과정 정이태

챗봇



* 텍스트 생성(text generation)

딥러닝

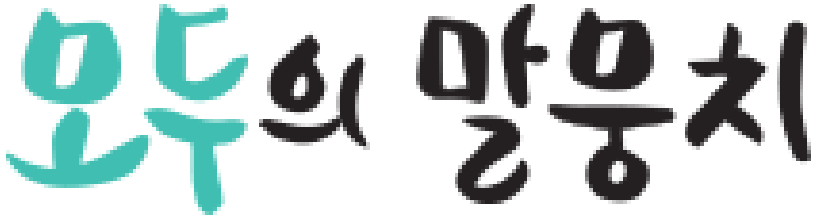
규칙기반

유사도 기반

하이브리드 기반
* 규칙+유사도

특정 시나리오 기반

데이터



음식점

16,000 sentences

의류

학원

소매점

생활서비스

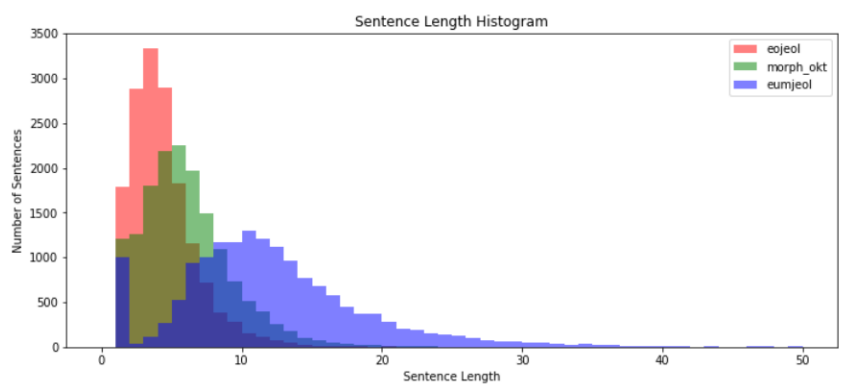
카페

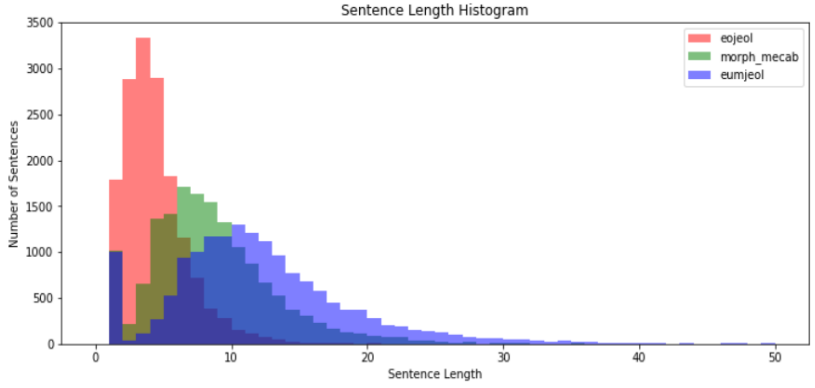
숙박업

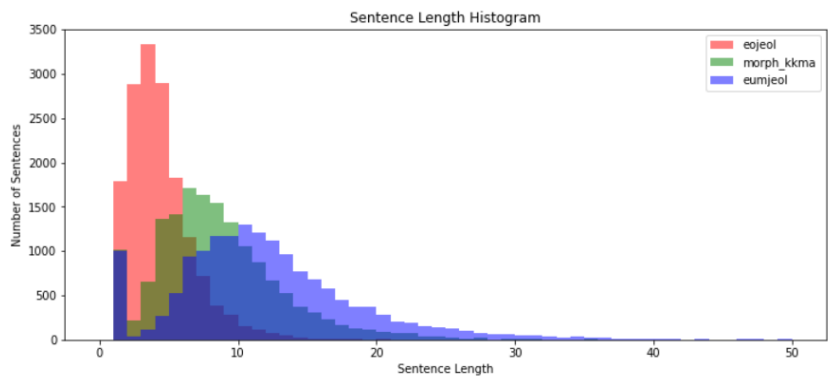
관광/여가/오락

민원

Exploratory Data Analysis

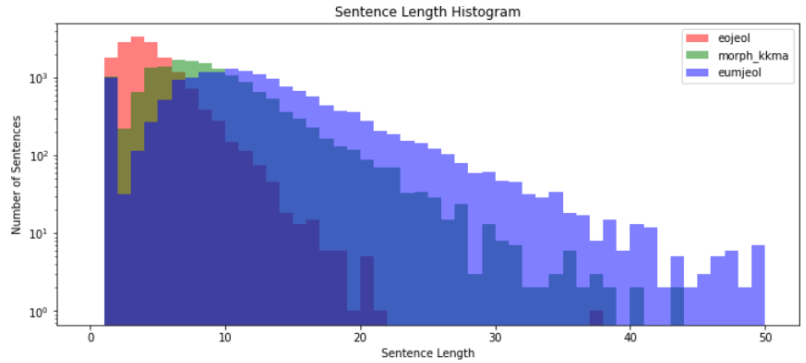


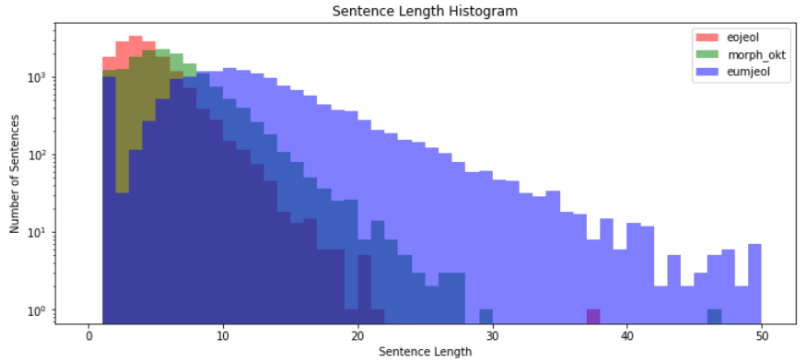


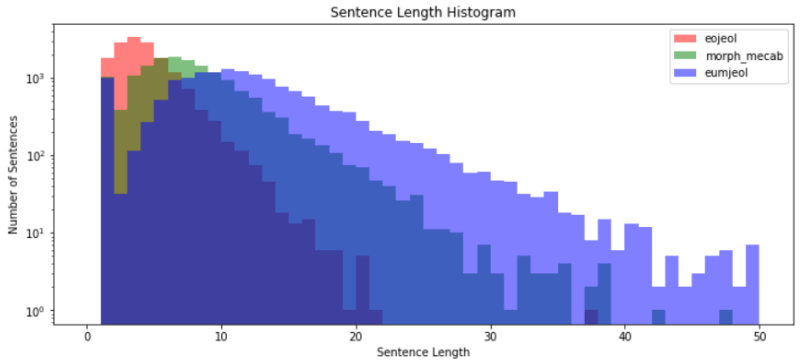


Sentence Length : 음절 > 형태소 > 어절

Number of Sentences : 어절 > 형태소 > 음절

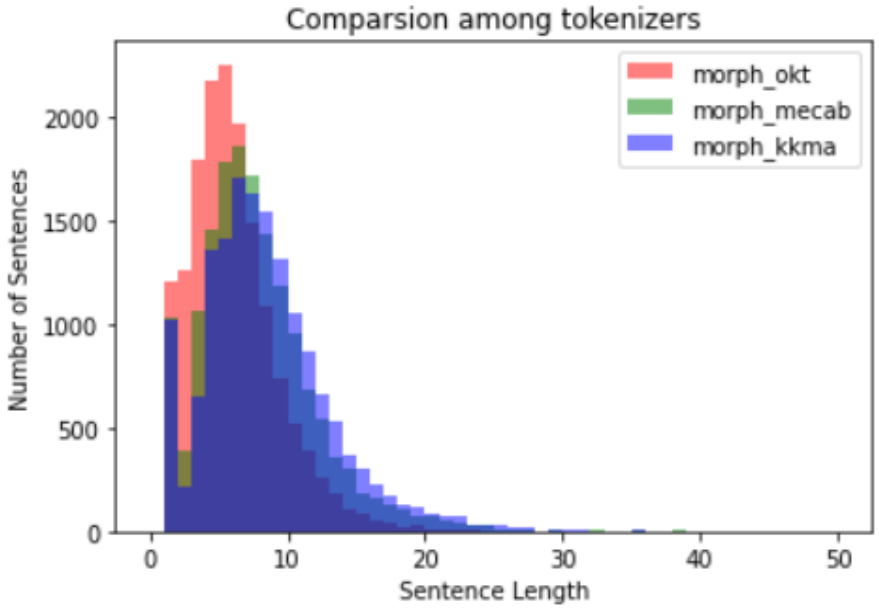


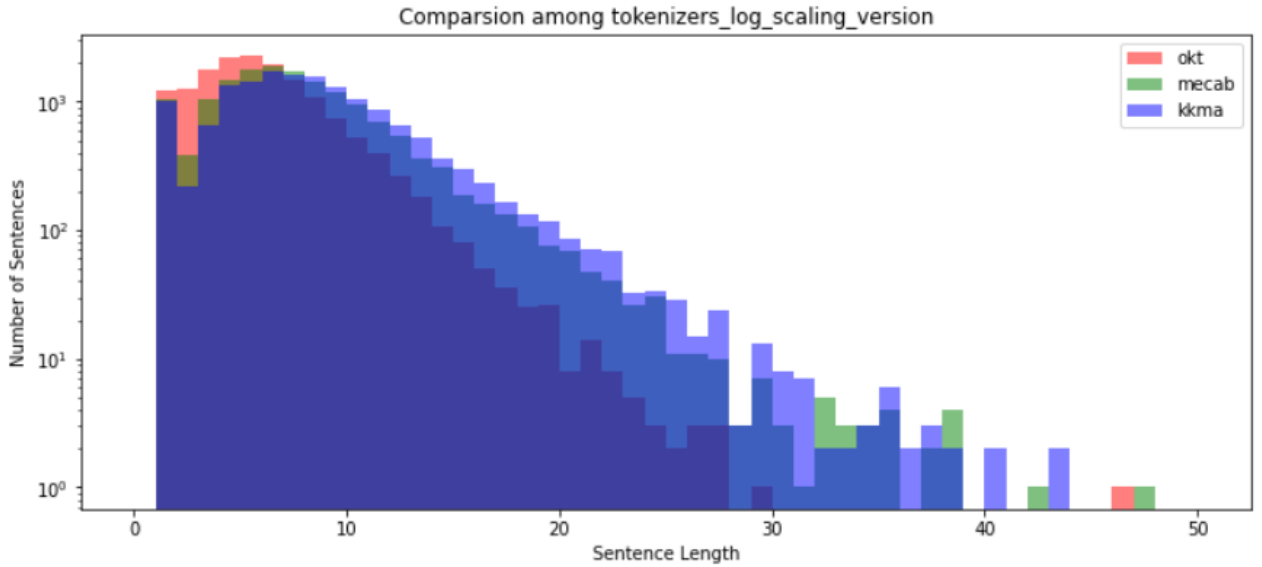




Y값의 스케일을 조정함으로써 차이가 큰 데이터에 대해서도 함께 비교할 수 있게끔 시각화 함.

-> 이전에 보이지 않았던 분포의 꼬리부분이 어떻게 분포돼 있는지 보기 쉽게 나옴.

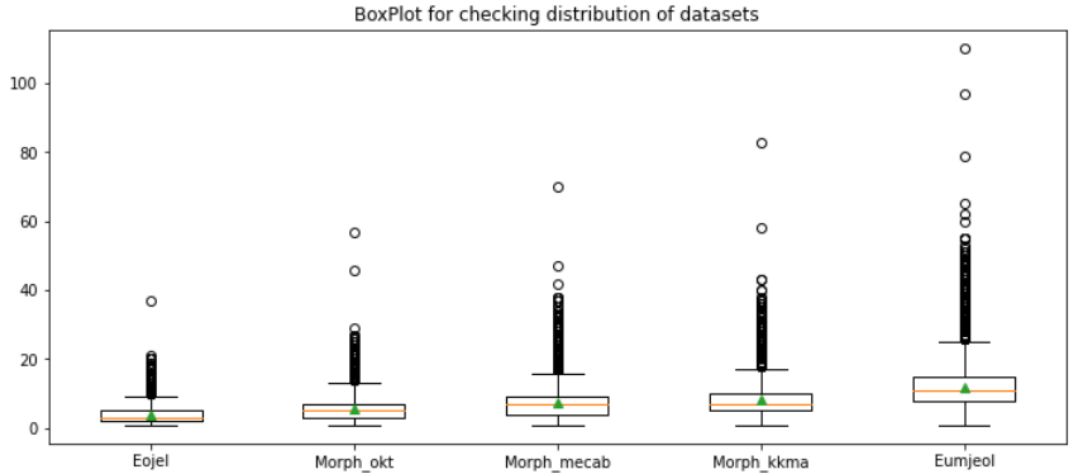




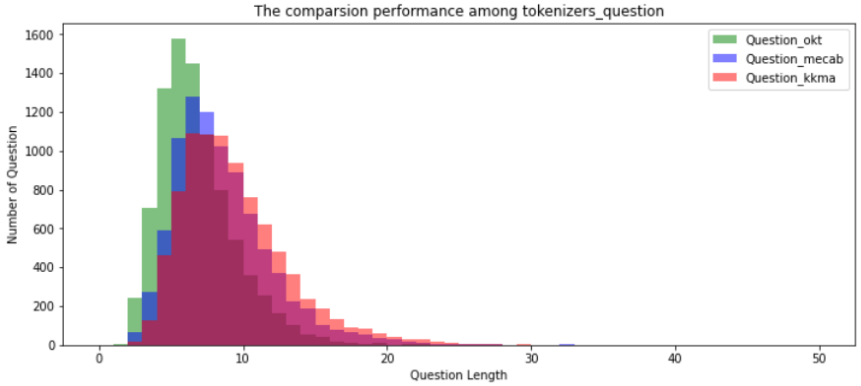
Sentence Length : kkma > mecab > okt

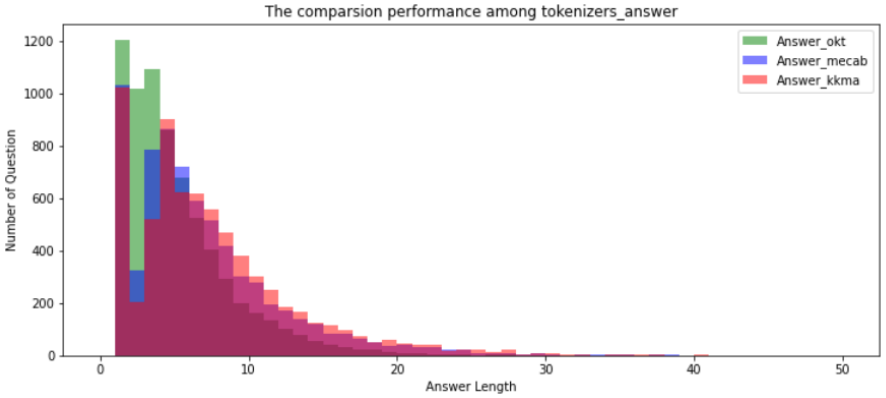
Number of Sentences : okt > mecab > kkma

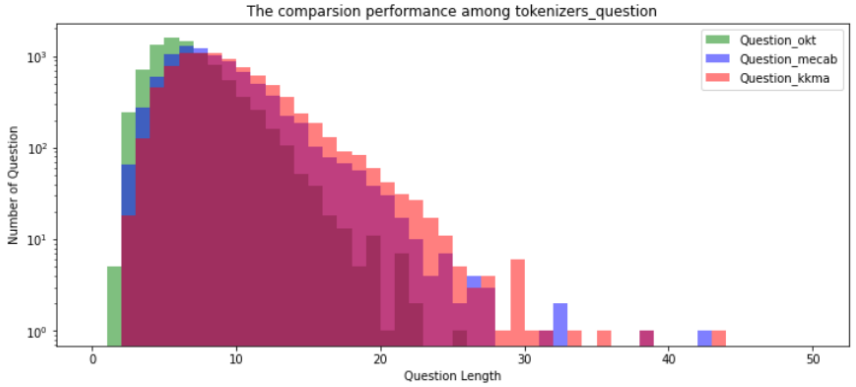
	최대	최소	평균	표준편차	중간	1사분위	3사분위
어절	37	1	3.89	2.35	3.0	2.0	5.0
형태소_okt	57	1	5.60	3.31	5.0	3.0	7.0
형태소_mecab	70	1	7.34	4.37	7.0	4.0	9.0
형태소_kkma	83	1	8.12	4.76	7.0	5.0	10.0
음절	110	1	11.79	6.83	11.0	8.0	15.0

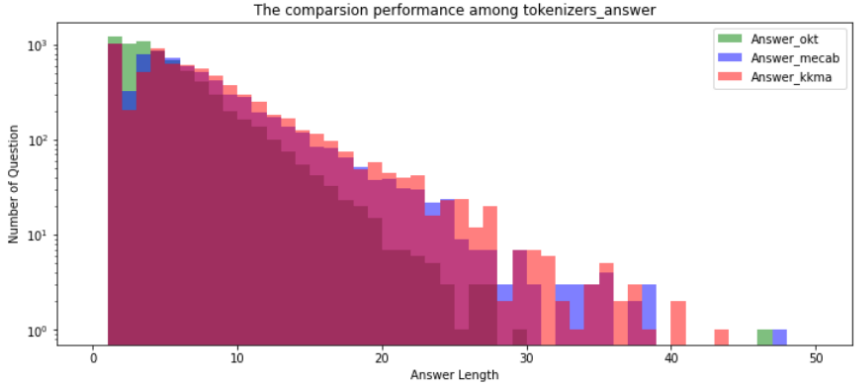


- 박스 플롯을 보아도 우측으로 꼬리가 긴 형태로 분포돼 있음을 확인 할 수 있었음.
- 대체로 문장의 길이는 3~12 의 길이를 중심으로 분포를 이루고 있음.
- 음절의 경우 길이 분포가 어절과 형태소에 비해 훨씬 더 크다는 점을 알 수 있었음.









Sentence Length(question) : kkma > mecab > okt

Sentence Length(answer) : kkma > mecab > okt

Number of Sentences(question) : okt > mecab > kkma

Number of Sentences(answer) : okt > kkma > mecab

* **Insight** ;

특정 서비스 평가 혹은 전사적 차원

customer interaction center (CIC) 서비스 품질 혁신전략 시

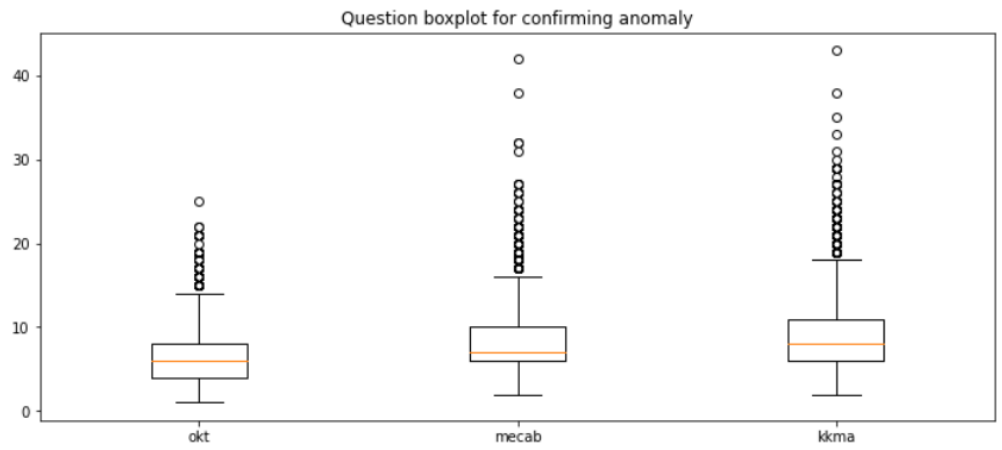
answer sentence-length 를 평가 척도 중 하나로 확인할 수 있음.

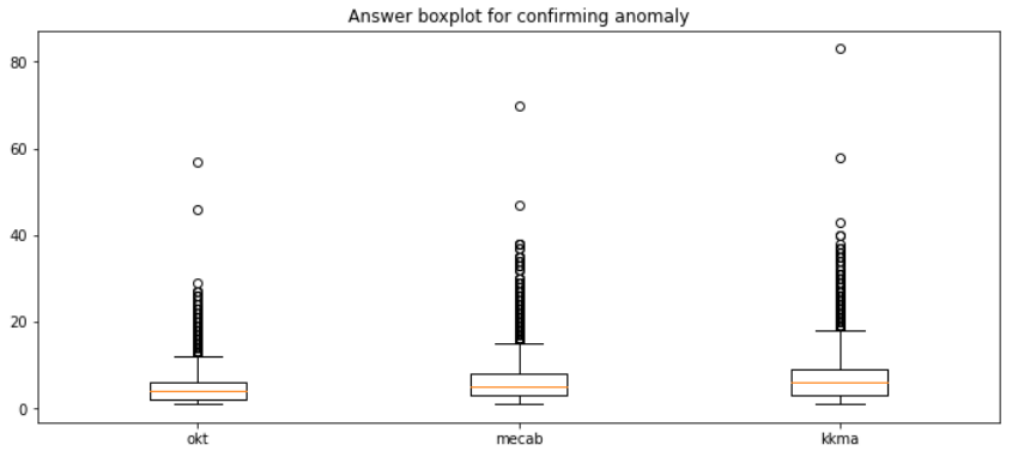
i.e.)

Q) 안녕하세요 제가 가방을 잃어버렸는데요.

A1) 네 고객님 마음 고생이 심하셨겠습니다. 먼저 …

A2) 네 담당부서로 연결해드리겠습니다.





- 박스 플롯을 보아도 우측으로 꼬리가 긴 형태로 분포돼 있음을 확인 할 수 있었음.

- Question 의 문장길이가 Answer에 비해 1.5배 가량 길다.

Question) 분포 토대로 okt 25 , mecab 30 , kkma 30 으로 max_sequence 선정 기준에 대한 인사이트 추출.

Answer) 분포 토대로 okt는 30 , mecab 40 , kkma 40 으로 max_sequence 선정 기준에 대한 인사이트 추출.

question_okt

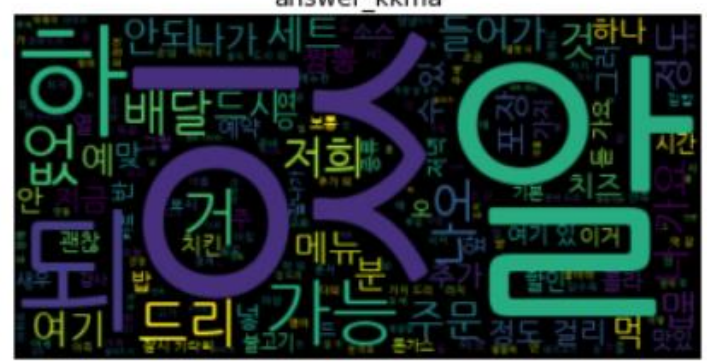
answer_okt

question_mecab

answer_mecab

question_kkma

answer_kkma



Max_Sequence = 5

김남규 교수님의 텍스트 마이닝 수업은 재밌다. 강추합니다. → 김남규 교수님의 텍스트 마이닝 수업은 | ?? ?? ??

Max_Sequence = 7

김남규 교수님의 텍스트 마이닝 수업은 재밌다. 강추합니다. → 김남규 교수님의 텍스트 마이닝 수업은 재밌다. 강추합니다.

Max_Sequence = 10

김남규 교수님의 텍스트 마이닝 수업은 재밌다. 강추합니다. → 김남규 교수님의 텍스트 마이닝 수업은 재밌다. 강추합니다. <PAD> <PAD> <PAD>

Max_Sequence = 20

김남규 교수님의 텍스트 마이닝 수업은 재밌다. 강추합니다. → 김남규 교수님의 텍스트 마이닝 수업은 재밌다. 강추합니다. <PAD> …

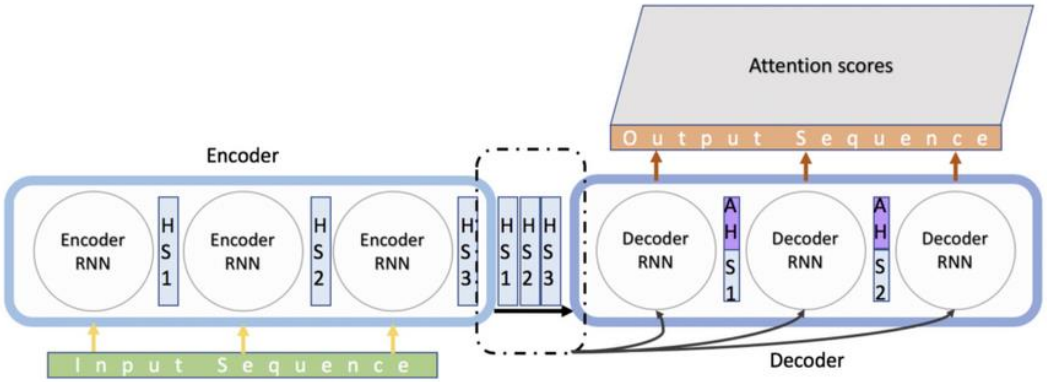
*어절단위

EDA 의 중요성

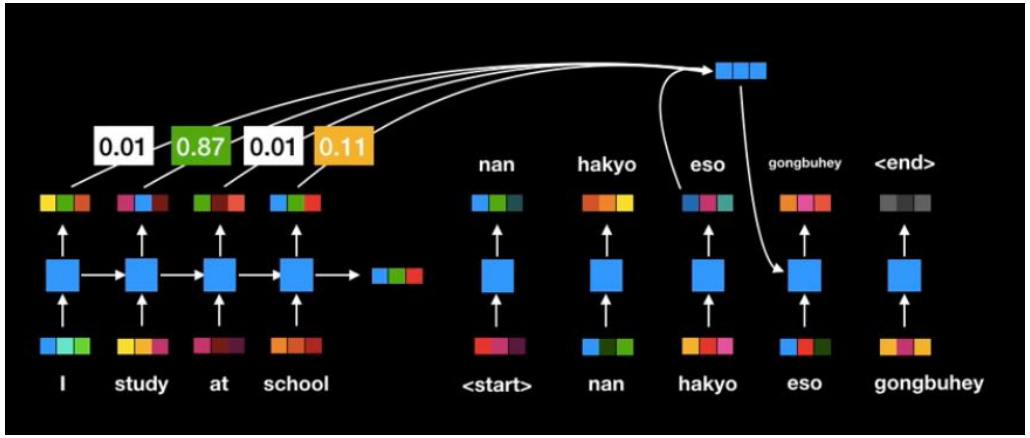
- Embedding vector로 나타내는 과정에 해당 문맥에 대한 고유한 정보를 놓칠 수 있음.
- Trade off ; 최대길이 ∝ 학습 속도
- Heuristic , Hyperparameter

정답은 없음! 다양한 시도를 통해서 낸 최적의 결과를 찾아야 함. 그 최적의 결과를 찾음에 있어 EDA는 중요한 과정!


MODELING



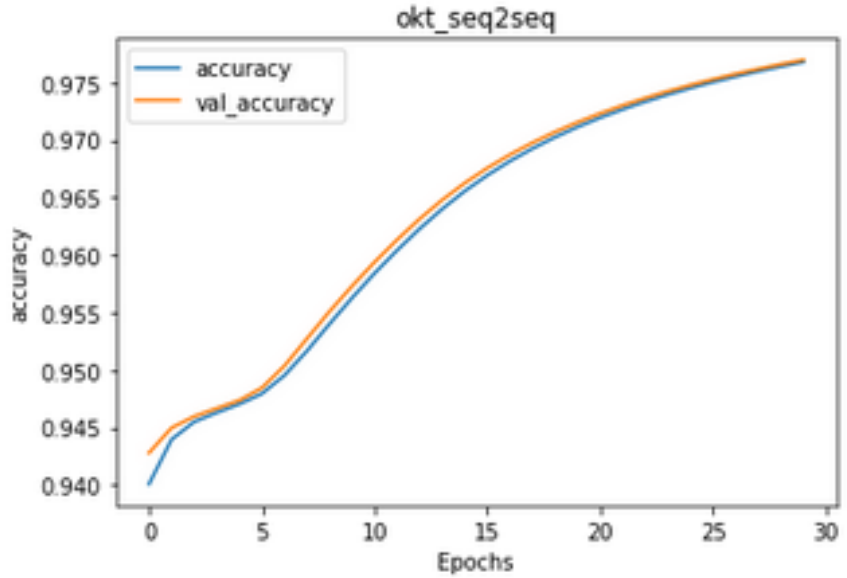
Seq2seq ; GRU + Attention

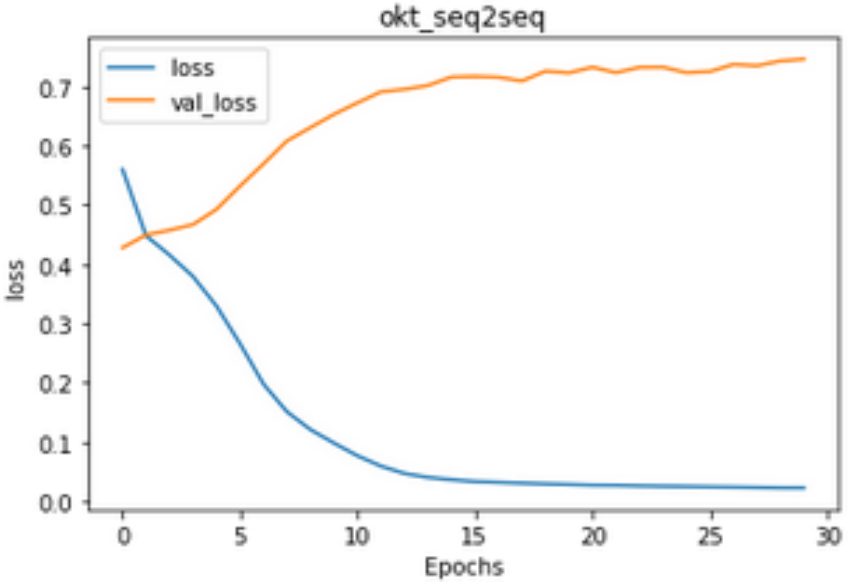


Attention

GRU ; RNN에서 발생하는 장기의존도 문제를 'gate'를 이용하여 정보의 양을 조절하고자 함.
'reset gate','update gate'를 추가하여 해결하고자 개발된 모델.
(LSTM 은 3개의 게이트를 가진 반면 GRU는 2개임. 상대적으로 학습속도가 빠르다고 알려짐, 허나 하이퍼 파라미터에 따라 성능의 차이가 있기에 case by case임.)

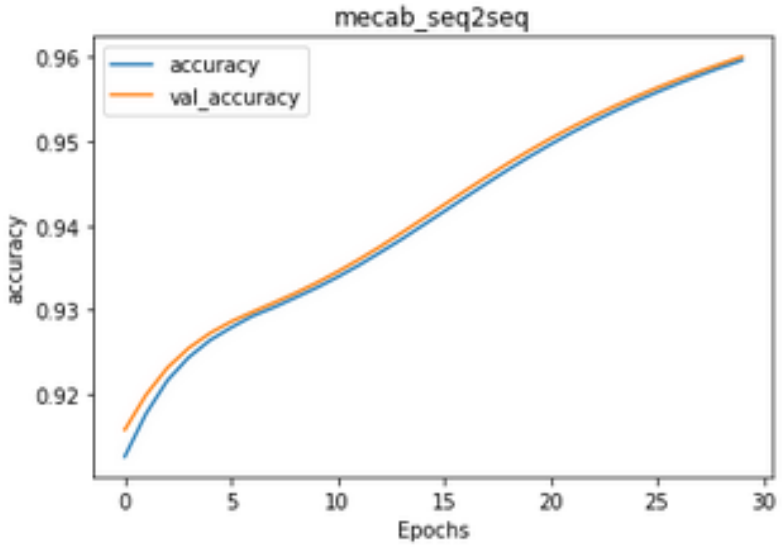
Attention ; 문장이 길어질수록 더 많은 정보를 고정된 길이에 담아야 하므로 정보의 손실이 있다는 점이 큰 문제로 지적됨.
또한 재귀순환망 특유 문제인 장기 의존성 문제를 보완하고자 개발된 알고리즘.
은닉 상태의 값을 통해 어텐션을 계산하고 디코더의 각 시퀀스 스텝마다 계산된 어텐션을 입력으로 넣음으로써 각 시퀀스 스텝마다 어텐션의 가중치가 다르게 적용됨.

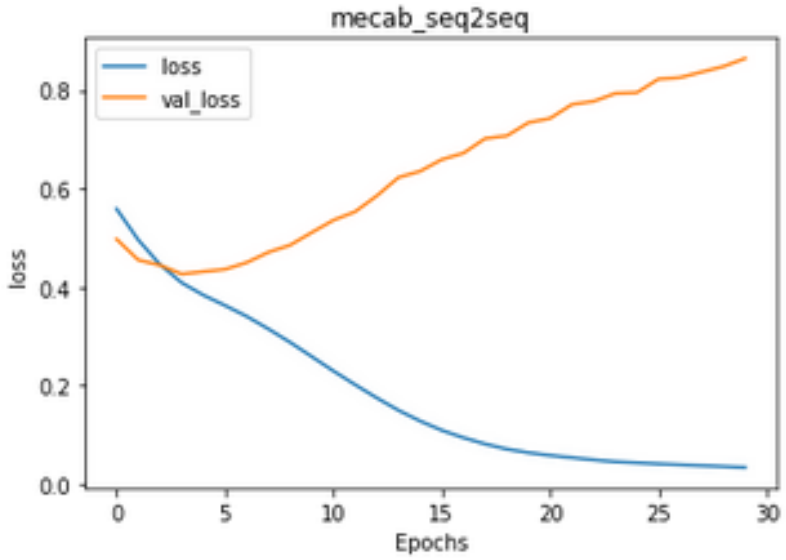




Question	Answer
짜장면 배달 되나요?	받죠 사이즈로 있어요
지금 배달 되나요?	아뇨
그럼 언제 배달 되나요?	네
홀 장사도 해요?	아 네
몇 시 마감이에요?	차슈가 여기 있습니다
짜장 1그릇 짬뽕 2그릇 요	전 대형 돈가스에요
정이태 잘생겼죠	룸은 조금 기다리셔야 합니다.

룸...? 진실의방...?





Question	Answer
짜장면 배달 되나요?	치킨 덮밥 이 에요
지금 배달 되나요?	전골 괜찮 은데 상가 주차장 에 있 거든요
그럼 언제 배달 되나요?	네 해 주 셔야 돼요
홀 장사도 해요?	김밥 메뉴 쪽 보 세요
몇 시 마감이에요?	네 맞 아요
짜장 1그릇 짬뽕 2그릇 요	네 맞 아요
정이태 잘생겼죠	그거 는 돼지고기 들어가 서 드 시 면 보통 맛 도 있 는데 가격 대로 고르 시 면 …

연구 결과

01

전처리의 중요성

동일 데이터셋을
음절 어절 형태소
분석을 하여 추출한 결과의
각각의 차이 대해 알아봄.

02

형태소분석기에 따른 성능차이

Mecab , Okt
형태소 분석기 별 성능 차이에 대해
알아보았음.

References

* https://www.youtube.com/watch?v=mxGCEWOxfe8

* https://towardsdatascience.com/day-1-2-attention-seq2seq-models-65df3f49e263

* https://omicro03.medium.com/%EC%9E%90%EC%97%B0%EC%96%B4%EC%B2%98%EB%A6%AC-nlp-30%EC%9D%BC%EC%B0%A8-gru-4fce44eb4243

* https://konlpy-ko.readthedocs.io/ko/v0.4.3/morph/#pos-tagging-with-konlpy

* 솔트룩스 한국어 자연어처리 입문 자료 _ 김성현 님

* 텐서플로 2와 머신러닝으로 시작하는 자연어 처리 서적 _ 전창욱 님 외 3인

* 텍스트마이닝 강의 _ 국민대 김남규 교수님

향후 진행계획

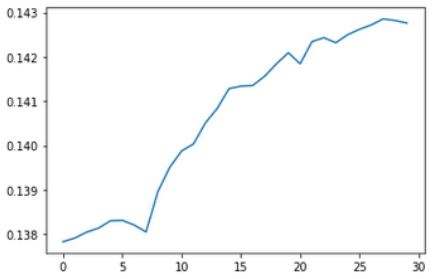
```
Epoch 7/30
3136/3136 [==============================] - ETA: 0s - loss: 0.2463 - accuracy: 0.9428
Epoch 00007: val_accuracy improved from 0.94220 to 0.94317, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3443s 1s/step - loss: 0.2463 - accuracy: 0.9428 - val_loss: 0.3773 - val_accuracy:
0.9432
Epoch 8/30
3136/3136 [==============================] - ETA: 0s - loss: 0.2274 - accuracy: 0.9437
Epoch 00008: val_accuracy improved from 0.94317 to 0.94415, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3518s 1s/step - loss: 0.2274 - accuracy: 0.9437 - val_loss: 0.3886 - val_accuracy:
0.9442
Epoch 9/30
3136/3136 [==============================] - ETA: 0s - loss: 0.2101 - accuracy: 0.9447
Epoch 00009: val_accuracy improved from 0.94415 to 0.94516, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3504s 1s/step - loss: 0.2101 - accuracy: 0.9447 - val_loss: 0.4041 - val_accuracy:
0.9452
Epoch 10/30
3136/3136 [==============================] - ETA: 0s - loss: 0.1935 - accuracy: 0.9457
Epoch 00010: val_accuracy improved from 0.94516 to 0.94619, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3520s 1s/step - loss: 0.1935 - accuracy: 0.9457 - val_loss: 0.4124 - val_accuracy:
0.9462
Epoch 11/30
3136/3136 [==============================] - ETA: 0s - loss: 0.1796 - accuracy: 0.9468
Epoch 00011: val_accuracy improved from 0.94619 to 0.94725, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3514s 1s/step - loss: 0.1796 - accuracy: 0.9468 - val_loss: 0.4193 - val_accuracy:
0.9473
Epoch 12/30
3136/3136 [==============================] - ETA: 0s - loss: 0.1693 - accuracy: 0.9478
Epoch 00012: val_accuracy improved from 0.94725 to 0.94828, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3494s 1s/step - loss: 0.1693 - accuracy: 0.9478 - val_loss: 0.4303 - val_accuracy:
0.9483
Epoch 13/30
3136/3136 [==============================] - ETA: 0s - loss: 0.1579 - accuracy: 0.9489
Epoch 00013: val_accuracy improved from 0.94828 to 0.94930, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3507s 1s/step - loss: 0.1579 - accuracy: 0.9489 - val_loss: 0.4449 - val_accuracy:
0.9493
Epoch 14/30
3136/3136 [==============================] - ETA: 0s - loss: 0.1444 - accuracy: 0.9499
Epoch 00014: val_accuracy improved from 0.94930 to 0.95035, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3514s 1s/step - loss: 0.1444 - accuracy: 0.9499 - val_loss: 0.4627 - val_accuracy:
0.9503
Epoch 15/30
3136/3136 [==============================] - ETA: 0s - loss: 0.1333 - accuracy: 0.9509
Epoch 00015: val_accuracy improved from 0.95035 to 0.95139, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3474s 1s/step - loss: 0.1333 - accuracy: 0.9509 - val_loss: 0.4696 - val_accuracy:
0.9514
Epoch 16/30
3136/3136 [==============================] - ETA: 0s - loss: 0.1209 - accuracy: 0.9520
Epoch 00016: val_accuracy improved from 0.95139 to 0.95245, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3506s 1s/step - loss: 0.1209 - accuracy: 0.9520 - val_loss: 0.4838 - val_accuracy:
0.9525
Epoch 17/30
3136/3136 [==============================] - ETA: 0s - loss: 0.1099 - accuracy: 0.9530
Epoch 00017: val_accuracy improved from 0.95245 to 0.95352, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3514s 1s/step - loss: 0.1099 - accuracy: 0.9530 - val_loss: 0.4956 - val_accuracy:
0.9535
Epoch 18/30
3136/3136 [==============================] - ETA: 0s - loss: 0.0997 - accuracy: 0.9541
Epoch 00018: val_accuracy improved from 0.95352 to 0.95456, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3481s 1s/step - loss: 0.0997 - accuracy: 0.9541 - val_loss: 0.5179 - val_accuracy:
0.9546
Epoch 19/30
3136/3136 [==============================] - ETA: 0s - loss: 0.0902 - accuracy: 0.9551
Epoch 00019: val_accuracy improved from 0.95456 to 0.95559, saving model to ./data_out/seq2seq_kkma/weights.h5
3136/3136 [==============================] - 3443s 1s/step - loss: 0.0902 - accuracy: 0.9551 - val_loss: 0.5291 - val_accuracy:
0.9556
Epoch 20/30
1407/3136 [============>.................] - ETA: 30:14 - loss: 0.0765 - accuracy: 0.9558
```

1) 꼬꼬마 tokenizer 추가하여 실험완성

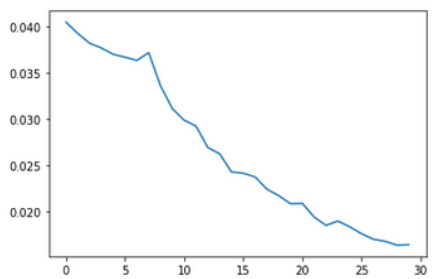
꼬꼬마 tokenizer로 한 결과는
실험 도중에 커널이 죽는 대참사가 발생하여
첨부하지 못하였습니다.

향후 진행계획

```
In [86]: output = predict("짜장면 배달 되나요?")
Input: 짜장면 배달 되나요?
Output: 네 가능합니다

In [87]: output = predict("지금 배달 되나요?")
Input: 지금 배달 되나요?
Output: 네 맞습니다

In [88]: output = predict("그럼 언제 배달 되나요?")
Input: 그럼 언제 배달 되나요?
Output: 네 가능합니다

In [89]: output = predict("홀 장사도 해요?")
Input: 홀 장사도 해요?
Output: 여기요

In [90]: output = predict("몇 시 마감이에요?")
Input: 몇 시 마감이에요?
Output: 네

In [91]: output = predict("짜장 1그릇 짬뽕 2그릇 요")
Input: 짜장 1그릇 짬뽕 2그릇 요
Output: 아니요 나가실 때 결제하시면 됩니다

In [92]: output = predict("정이태 잘생겼죠")
Input: 정이태 잘생겼죠
Output: 저희 2명이요
```



accuracy



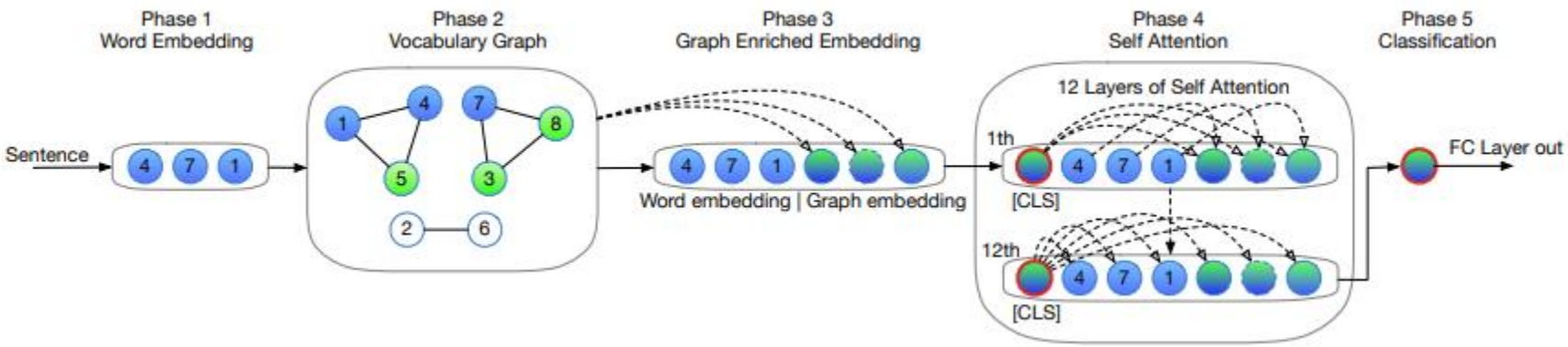
loss

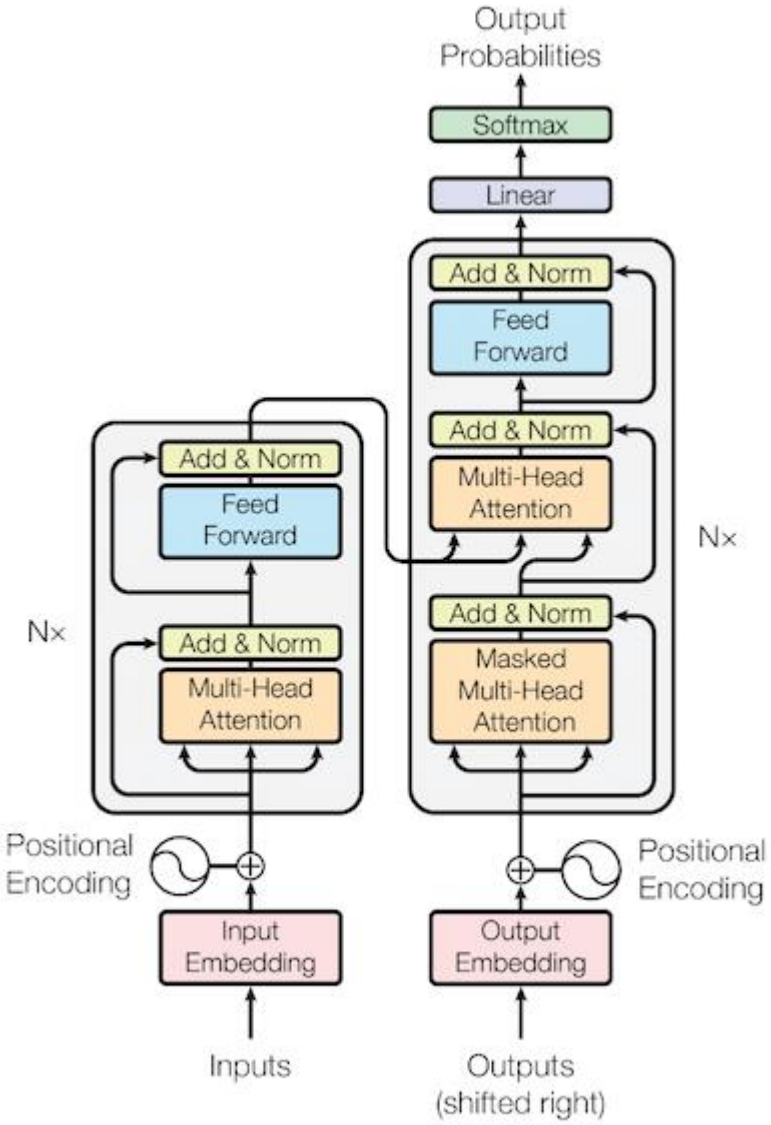
2) Transformer , GPT2 추가 실험

기존에 진행 하려했던 Transformer 과 GPT2 에서도 추가적으로 실험하여 완료해보고자 합니다. Transformer로 하고자 하였으나 다소 성능이 낮게 나온 터라 좀 더 보완이 필요하다 생각하여 추가하지 않았습니다.

3) Graph Embedding 방법 적용

기존 인코딩 방법과 다르게 Graph Embedding 방법을 적용하여 텍스트 생성 task에서 유의미한 결과를 낼 수 있을것 인가에 대해 최종적으로 연구해 보고자 합니다.





Transformer

- 멀티 헤드 어텐션

내적 어텐션 구조가 중첩된 형태.

* 스케일 내적 어텐션 , Masked 디코더 셀프 어텐션 , 인코더 디코더 어텐션

- 서브시퀀트 마스크 어텐션

RNN 과 다르게 전체 문장이 한번에 행렬 형태로 입력되는 구조

→ 자신보다 앞에 있는 단어만 참고해서 단어를 예측해야 하지만 위치에 상관없이 모든 단어를 참고해서 예측할 것이기에 자신보다 뒤에 있는 단어를 참고하지 않게 하는 기법.

- 포지션 인코딩

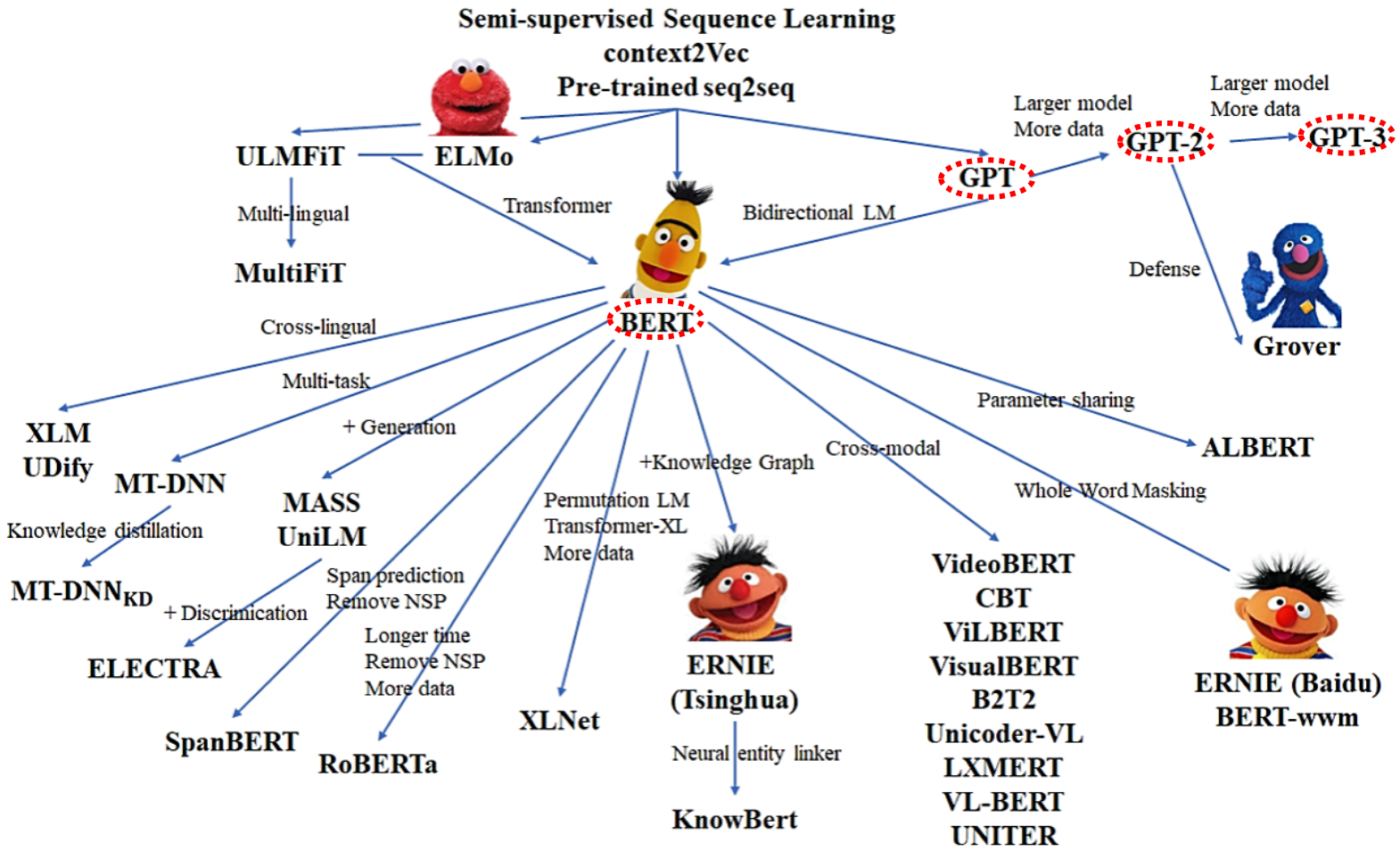
순서 정보가 반영되지 않는 문제를 해결하기 위해 사용한 기법. 피처차원에서 사인 / 코사인 함수를 활용하여 각 인덱스에 함숫값을 할당 함.

- 포지션-와이즈 피드 포워드 네트워크

한 문장에 있는 단어 토큰 벡터 각각에 대해 연산하는 네트워크

- 리지듀얼 커넥션

입력 정보 x , 네트워크 레이어를 거쳐 나온 정보 F(x) 를 더해 앞에 있던 정보 x를 보존하고자 하는 방법.


Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ULMFiT
ELMo
Multi-lingual
MultiFiT
Transformer
Bidirectional LM
GPT
Larger model
More data
GPT-2
Larger model
More data
GPT-3
Defense
Grover
BERT
Cross-lingual
Multi-task
+ Generation
+Knowledge Graph
Cross-modal
Parameter sharing
Whole Word Masking
XLM
UDify
MT-DNN
MASS
UniLM
ALBERT
Knowledge distillation
MT-DNN$_{KD}$
+ Discrimication
ELECTRA
Span prediction
Remove NSP
Longer time
Remove NSP
More data
SpanBERT
RoBERTa
Permutation LM
Transformer-XL
More data
XLNet
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
ERNIE (Baidu)
BERT-wwm