대화의 재구성 : 음성 인식 기반 텍스트 요약 및 키워드 추출 모델

팀원의 재구성
최용우 , 최희범

contents

연구 배경

- 정보량이 폭발적으로 증가하는 현대사회에서 **음성 데이터**에서는 중요한 정보를 쉽게 파악하고 처리하기 어려움
- **음성인식 텍스트 요약 모델** 구현을 목표로 함
    - Speech-to-Text 모델을 활용하여 음성을 텍스트로 변환
    - Text-Summarization 모델을 활용하여 원문요약 추출
    - Keyword-Extraction 모델 활용하여 텍스트 키워드 추출

- 본 연구에서는 한국어 대화 요약 데이터를 사용

팀 구성 및 역할

	역할	담당 업무
최희범	팀장	음성 인식 모델 연구, 문서 요약 모델 Fine-tuning
최용우	팀원	문서 요약 모델 Fine-tuning, Keyword 추출 모델 연구

프로젝트 수행 과정

구분	기간	활동	비고
사전 기획	4/15 ~ 4/22	프로젝트 주제 선정	
데이터 수집 및 전처리	4/23 ~ 4/25	필요 데이터 수집 및 모델링에 필요한 형태로 변환	
모델링	4/26 ~ 5/6	모델 구현 및 fine tuning	
총 개발기간	4/15 ~ 5/6(총 3주)		

Methods

연구 방법론

- 학습 데이터 전처리
  - **정규 표현식** 및 **불용어 처리**를 통해 학습 데이터 변환

- KoBART 모델 Fine-tuning
  - 패키지의 버전이 달라져 오류가 생기는 부분을 번안 및 KoBART 모델 **fine-tuning**

- 성능 지표 확인
  - 요약된 데이터의 **Rouge 성능 지표** 비교

Methods

Deep Learning Model

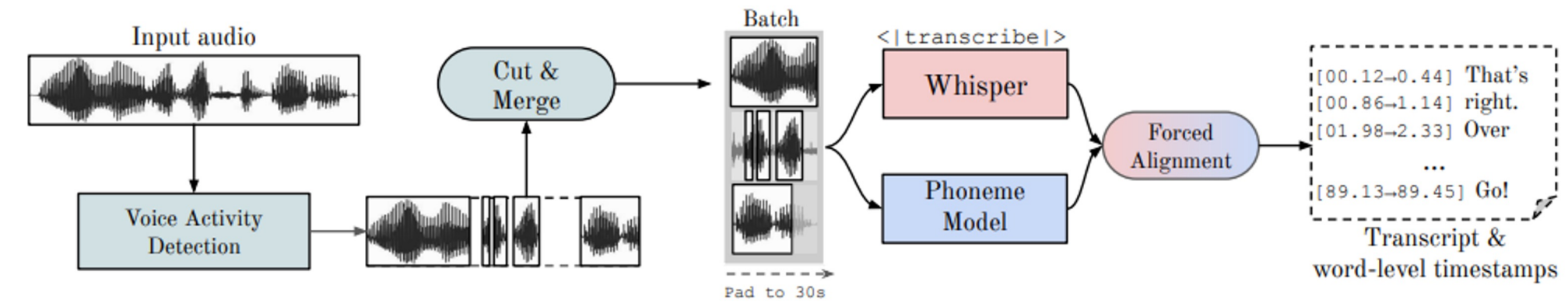
음성 데이터 텍스트 변환	→	WhisperX
텍스트 요약	→	KoBART
키워드 추출	→	KeyBERT

Methods

Whisper X

- 음성 데이터를 텍스트로 변환하기 위해 Speech-to-Text 모델인 **WhisperX** 모델 사용
  - WhisperX를 사용한 이유
    - Whisper, 화자분리 불가 / 효과적인 대화 요약을 위해서는 해당 기능 필요한 것으로 판단

      => 각 화자에 라벨링 기능을 부여한 WhisperX 모델 채택

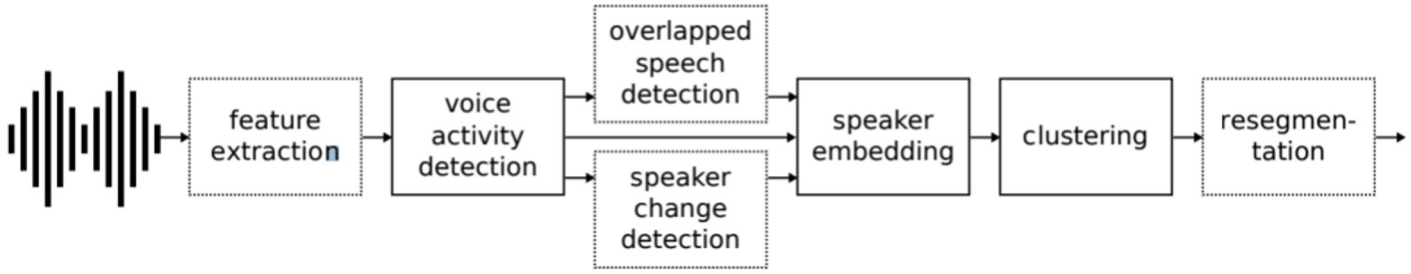


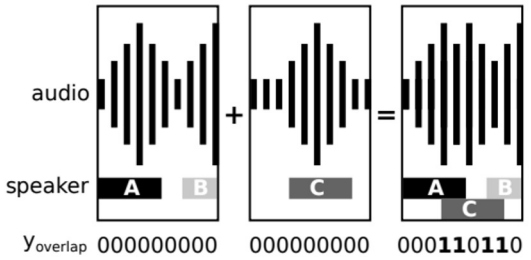
* 출처: WhisperX: Time-Accurate Speech Transcription of Long-Form Audio

Methods

Whisper X

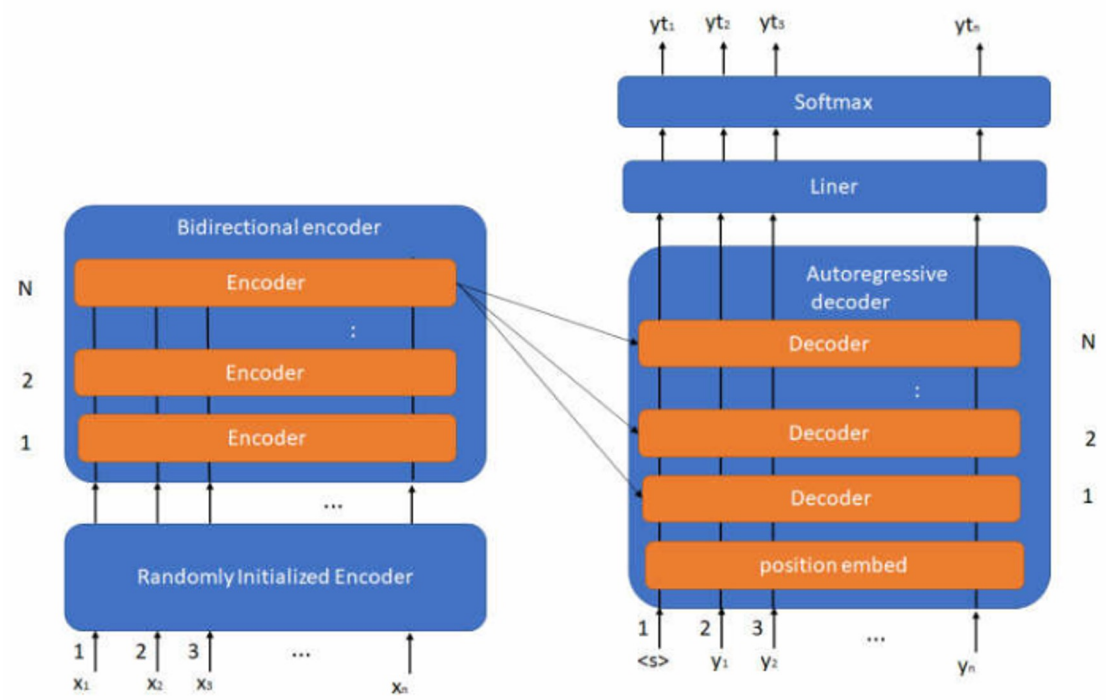
- Whisper의 경우, 자동음성인식(ASR)을 통해 최대 30초 배치로 나누어진 음성을 텍스트 문단으로 변환
- WhisperX는 pyannote.audio의 음성활동감지(VSD) 개념을 도입하여, 사전에 오디오 음성 분석 후 잡음 제거하고 각 화자의 음소를 클러스터화하여 분류된 음성을 합침. 이를 통해 계산량을 줄이면서도 음성인식 정확도를 향상시킴.





Methods

KoBART

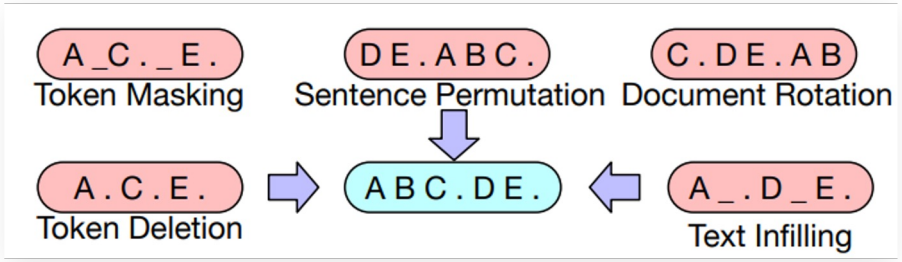


- Text-Summarization 모델로, 기존의 **Huggingface의 Bart** 모델을 한국어로 pre-training
- KoBART가 대화 데이터에 맞게 한국어 대화 요약 데이터로 Fine-tuning

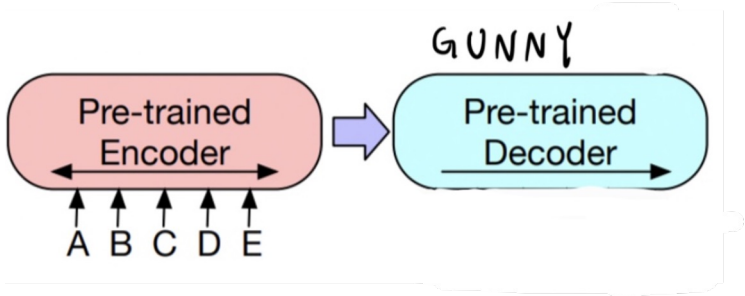
* 출처: Pseudocode Generation from Source Code Using the BART Model

Methods

KoBART



Pretraining

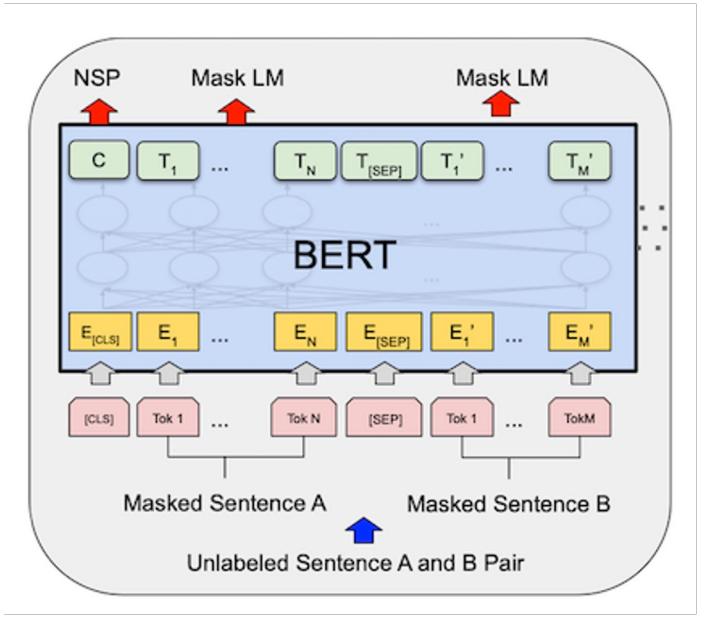


Fine-tuning

* 출처: Pseudocode Generation from Source Code Using the BART Model, velog.io/@tobigs-nlp

Methods

KeyBERT



Pretraining

- MLM(Masked Language Model)
- NSP(Next Sentence Prediction)

- 원문과 키워드의 유사도를 추출하고 핵심 키워드를 추출하기 위해 KeyBERT 활용
- KoBERT의 pre-trained 된 Embedding 사용

* 출처: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

KeyBERT

- Keyword 추출 방법

1. 사전 훈련된 BERT 모델을 사용해 벡터 형태로 문서 임베딩
2. 단어 또는 n-gram을 추출하기 위해 Bag of words 등의 기법을 활용해 텍스트 데이터로부터 단어와 n-gram 추출
3. 문서를 임베딩하는데 사용된 **동일한** 모델을 이용해 단어 또는 n-gram 임베딩
4. 문서 임베딩과 각 키워드의 임베딩 사이의 코사인 유사도 계산
5. 가장 유사도 점수가 높은 키워드가 핵심 키워드로 선택

Results

Speech-to-Text

```
1 df = pd.DataFrame(data=result['segments'])
2 df[:60]
```

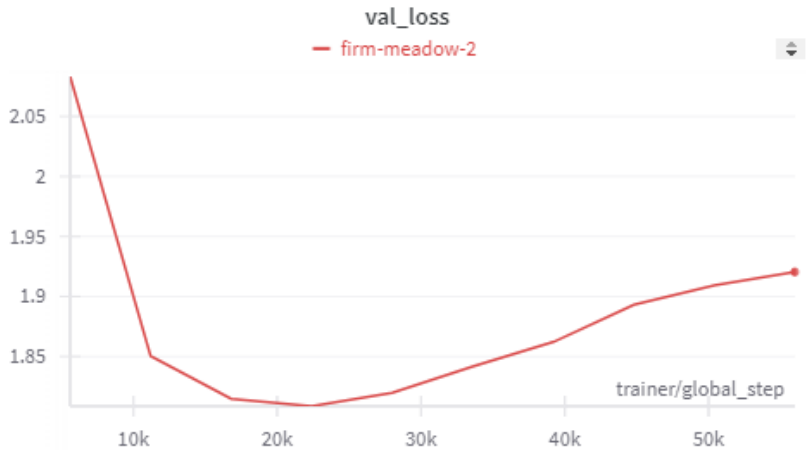
	start	end	text	words	speaker
0	0.811	1.831	여보세요.	[{'word': '여보세요.', 'start': 0.811, 'end': 1.83...	SPEAKER_01
1	2.372	2.992	네, 안녕하세요.	[{'word': '네,', 'start': 2.372, 'end': 2.532, ...	SPEAKER_00
2	3.092	4.492	최용우님 맞으시죠?	[{'word': '최용우님', 'start': 3.092, 'end': 3.632...	SPEAKER_00
3	4.512	4.553	네.	[{'word': '네.', 'start': 4.512, 'end': 4.553, ...	SPEAKER_00
4	4.633	8.274	네, 여기는 딥러닝 부트캠프 신청해주신 알파코인데요.	[{'word': '네,', 'start': 4.633, 'end': 4.733, ...	SPEAKER_00
5	8.354	8.414	네.	[{'word': '네.', 'start': 8.354, 'end': 8.414, ...	SPEAKER_00
6	8.534	11.415	네, 잠시 통화 가능하세요?	[{'word': '네,', 'start': 8.534, 'end': 8.814, ...	SPEAKER_00
7	11.515	11.875	네.	[{'word': '네.', 'start': 11.515, 'end': 11.875...	SPEAKER_01
8	12.436	15.637	네, 저희 과정 지원해주셔서 감사드리고요.	[{'word': '네,', 'start': 12.436, 'end': 13.296...	SPEAKER_00
9	15.937	16.017	네.	[{'word': '네.', 'start': 15.937, 'end': 16.017...	SPEAKER_00
10	16.277	20.599	저희 지원서 바탕으로 몇 가지 여쭤보고 면접 예약 도와드릴게요.	[{'word': '저희', 'start': 16.277, 'end': 16.497...	SPEAKER_00
11	20.679	20.739	네.	[{'word': '네.', 'start': 20.679, 'end': 20.739...	SPEAKER_01
12	21.559	22.280	네.	[{'word': '네.', 'start': 21.559, 'end': 22.28,...	SPEAKER_00
13	22.940	26.721	내일 배움 카드는 있다고 체크해주셨는데 네.	[{'word': '내일', 'start': 22.94, 'end': 23.12, ...	SPEAKER_00
14	26.902	29.182	네, 실물 카드로 갖고 계신 거죠?	[{'word': '네,', 'start': 26.902, 'end': 26.962...	SPEAKER_00
15	29.283	30.343	네네.	[{'word': '네네.', 'start': 29.283, 'end': 30.34...	SPEAKER_00
16	30.383	30.403	네.	[{'word': '네.', 'start': 30.383, 'end': 30.403...	SPEAKER_00
17	30.776	35.217	혹시 저희 과정 같은 국비과정 수강하신 이력 있으실까요?	[{'word': '혹시', 'start': 30.776, 'end': 31.056...	SPEAKER_00
18	35.237	35.997	아니요.	[{'word': '아니요.', 'start': 35.237, 'end': 35.9...	SPEAKER_01
19	36.017	36.377	없어요.	[{'word': '없어요.', 'start': 36.017, 'end': 36.3...	SPEAKER_01
20	36.397	38.218	처음이세요?	[{'word': '처음이세요?', 'start': 36.397, 'end': 38...	SPEAKER_00

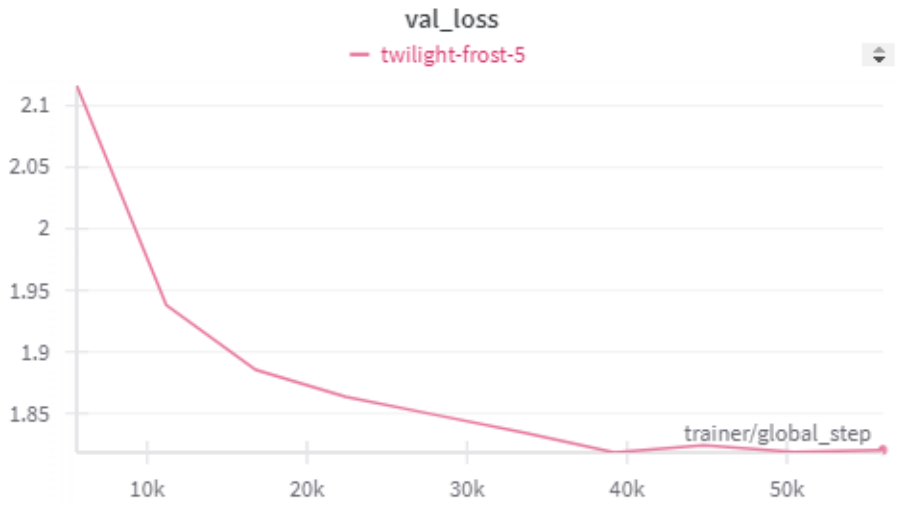
- 발화 시간, 텍스트, 어절 단위 분류, 발화자

Results

Text-Summarization

- KoBART 모델을 Fine-tuning
- 약 22만개의 원문과 요약문 쌍을 Train Data로 6만개의 원문과 요약문 쌍을 Validation Data로 사용
- Sequence의 최대 길이를 변경해보며 학습하였고, Overfitting이 되지 않도록 Dropout을 조정함





Results

원문 텍스트

여보세요. 네 안녕하세요 최용우님 맞으시죠 네 여기는 딥러닝 부트 캠프 신청해주신 알파코인데요 네 잠시 통화 가능하세요 네 저희 과정 지원해주셔서 감사드리고요. 저희 지원서 바탕으로 몇 가지 여쭤보고 면접 예약 도와드릴게요 네 어 내일 배움 카드는 있다고 체크해 주셨는데 네 네 실물 카드로 갖고 계신 거죠? 네 혹시 저희 과정 같은 국비 과정 혹시 수강하신 이력 있으실까요? 아니요 없어요. 아 처음이세요? 네 그러면 혹시 지금 거주하시는 지역은 어느 쪽이세요? 어 그 일산 쪽에 거주하고 있어요. 일산 쪽이세요 혹시 뭐 구나 동까지 말씀 주실 수 있으실까요? 저희가 통화 때문에 확인하는 부분이라서요. 아 행신력 바로 근처에 거주하고 있음. 인식력 근처로 잘 남겨드리고요. 네 어 저희가 그 지원하셨을 때 확인은 하셨겠지만 그 개강하고 2주 정도는 온라인 과정이 있고요 이후에 5개월 정도는 성수에 있는 교육장에서 지금 저 좀 교육이 진행이 되거든요. 통학은 문제 없으실까요? 네 문제 없어요. 네 그러면 저희가 마지막으로 그 면접 안내드릴 건데요 면접은 저희가 줌을 통해서 비대면으로 진행하고 있어요. 네 주문 설치되어 있으실까요? 네 그러면 혹시 월요일 오후에 면접 가능한 시간 있으실까요 어. 어 다음 주 월요일. 네네. 네 아무 때나 괜찮아요. 괜찮으세요? 그러면은 면접 시간 2시로 잡아드려도 괜찮으실까요? 네. 그 비대면 면접이라는 게. 네 응? 그냥 줌을 통해서 면접한다는 거죠? 네네 맞습니다. 아 네 알겠습니다.한 10분에서 15분 정도 진행될 거고요. 저희 과정 신청하신 동기라던가 뭐 아니면은 뭐 커리큘럼에 대한 이해도 그다음에 저희 과정 수료하시고 계획 같은 거 위주로 여쭤볼 거예요. 그래서 네 편하게 참석해 주시면 되고요 면접 참여하시는 URL은 월요일날 오전 중으로 발송될 거예요. 2시에 네 해당 URL 접속해서 면접 진행해 주시면 됩니다. 아 알겠습니다. 네 감사합니다. 네.

요약 텍스트

요약 텍스트: 딥러닝 부트 캠프를 신청해 주셔서 감사하다고 하고 면접 시간은 2시로 잡아드리기로 하고 면접에 대한 대화를 나눈다

Results

Text-Summarization

- 2000개의 Test Data 기준 모델 성능

	Max_len=180		Max_len=200		Max_len=200, Dropout	
	Rouge-1	Rouge-l	Rouge-1	Rouge-l	Rouge-1	Rouge-l
Precision	0.278	0.233	0.277	0.233	0.278	0.234
Recall	0.589	0.500	0.587	0.497	0.597	0.506
F1	0.368	0.310	0.367	0.309	0.369	0.311

Results

Keyword Extraction

- Mecab 형태소 분석기 사용
- MSS(Maximum sum Similarity), MMR(Maximum Marginal Relevance)를 사용해 다양한 키워드 추출

```
[('참석', 0.5838),
 ('커리큘럼', 0.5615),
 ('지원서', 0.56),
 ('수강', 0.557),
 ('교육장', 0.5391)]
```

활용분야

- 비즈니스 회의 및 워크숍
  - 회의 내용 실시간 텍스트 변환 및 회의 주요 내용 요약

- 의료 상담
  - 의사와 환자 간 대화를 텍스트로 변환하여 요약하여 진단, 치료계획 등 키워드를 얻고 의료 기록으로 활용

- 고객 서비스 및 지원
  - 전화나 온라인 대화를 텍스트로 변환, 요약하여 문제나 요청사항을 신속하게 해결 가능

Discussion&Conclusion

자체 평가

- 요약 모델 정확도 부족
  - 컴퓨팅 자원 부족으로 인해 충분히 많은 Epoch를 학습하지 못했고, 하이퍼 파라미터 튜닝을 충분히 하지 못함

- 실시간 키워드 추출의 어려움
  - 문맥을 파악하여 키워드를 추출하지만 속도가 느려서 실시간으로 적용하기에 어려움이 있음

➢ 튜닝을 충분히 한다면 좋은 성능을 낼 수 있을 것이고, 모델의 구조 자체를 파악하는 것도 값진 경험이 되었음

Thank You

st

Appendix

부록1. BART의 Pre-Training 방식

- Token Masking

BERT의 MLM과 동일한 방법으로 random tokens들이 추출되어 [MASK] token으로 대체

- Token Deletion

Input으로 들어가는 text에서 random tokens들이 삭제가 되는 방법, [MASK] token을 맞추는 Token Masking과 다르게
이 방법은 어느 위치에서 token이 삭제가 되었는지 맞추는 것이 목적

Appendix

부록1. BART의 Pre-Training 방식

- Sentence Permutation

단순히 문장 간의 순서를 바꾸는 식으로 original data에 noise를 주는 방법. 예시를 들면, 'ABC.DE'라는 original text가 'DE.ABC.'로 바뀐 모습을 볼 수 있음 이러한 문장 간의 구분을 단순히 ' full stop(마침표) ' 를 기준으로 하여 구분

- Document Rotation

token들의 나열로 이루어진 text에서 동일한 확률로 랜덤하게 하나의 token을 골라서, 이를 시작점으로 두고 배열하는 방법 예시를 들자면, 'ABC.DE'라는 원본 문장에서, 랜덤하게 'C'라는 token을 임의로 뽑아 이를 시작 점에 배치해두고 'C'앞에 있던 token들은 자연스레 뒤로 가서 바뀐 문장은 'C.DE.AB', 이 방법으로 하여금 모델이 document의 시작점을 인지하는 능력을 학습

Appendix

부록2. BART의 Hyperparameter

- D_model : 모델의 임베딩 차원 크기
- Encoder, Decoder_attention_heads : 인코더, 디코더의 멀티 헤드 어텐션의 헤드 수
- Activation_Function : 사용되는 활성화 함수, Default는 Gelu 함수
- Encoder, Decoder_ffn_dim : Encoder, Decoder feedforward 네트워크 차원
- Encoder, Decoder_layers : Encoder, Decoder layer 수
- Activation, attention, classification Dropout : Activation function, Attention, Classification 후 Dropout 비율을 전부 다르게 사용할 수 있음

Appendix

부록3. KeyBERT의 MSS, MMR 계산

- Maximum Sum Similarity

데이터 쌍 사이의 최대 합 거리는 데이터 쌍 간의 거리가 최대화되는 데이터 쌍
후보 간의 유사성을 최소화 하면서 문서와 후보 유사성을 최대화 하려고 함

1) 문서와 각 키워드 간 유사도 계산, 각 키워드 간의 유사도 계산
2) 문서와 키워드 간 코사인 유사도 기반하여 상위 n개 단어를 pick
3) 각 키워드 간 가장 덜 유사한 키워드 간의 조합을 계산(ex. 5개 조합의 유사도 합 최소)
4) 상위 nr_candidates 개수를 뽑아서 top_n개의 조합의 유사도 합 최소인 조합을 가져옴
5) ex)**def max_sum_sim**(doc_embedding, candidate_embeddings, words, top_n, nr_candidates)
6) nr_candidates가 클수록 다양한 키워드를 가져옴

Appendix

부록3. KeyBERT의 MSS, MMR 계산

- Maximum Marginal Relevance

키워드 추출 알고리즘은 키워드/키프레이즈를 다양화하는 데 사용할 수 있는 MMR을 구현.
먼저 문서와 가장 유사한 키워드/키프레이즈를 선택하고 그런 다음 문서와 유사하고 이미 선택된 키워드/키프레이즈와 유사하지 않은 새로운 후보를 반복적으로 선택

1) 문서와 각 키워드 간의 유사도 리스트 생성
2) 각 키워드 간의 유사도 계산
3) 문서와 가장 높은 유사도를 가진 키워드 인덱스 추출
4) top(n)-1만큼 추가적인 키워드를 선택함
    a. mmr= (1-diversity) * *candidate_similarity(각 후보 키워드와 문서 사이의 유사도)-diversity* *target_similarities(각 후보 키워드와 이미 선택된 키워드 사이의 최대 유사도), diversity가 높을 수록 다양성을 중요시함, 계산된 mmr값이 가장 높은 후보를 선택