Text Summarization With T5

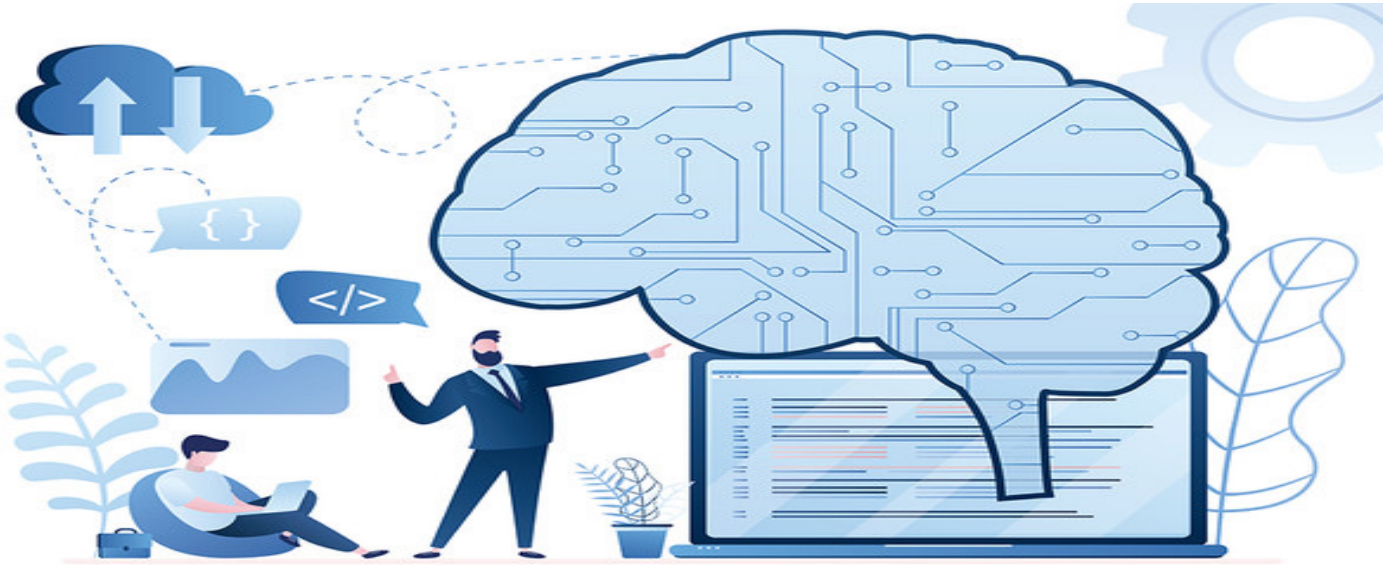
Team : 옥계동

CONTENTS

1
Introduction

Introduction

문서요약이란?

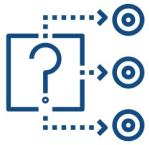


문서 요약이란, 주어진 문서로 부터 특정 사용자나 작업에 적합한 축약된 형태의 문서로 재생성하는 작업을 말한다.
이를 통해서 복잡도를 줄이면서 필요한 정보를 유지하는 것이 문서 요약의 주목적이다.

Introduction

Goal of analysis

간결하고 논리 정연하다.



중요한 Topic을 잡아낸다.

유창하고 일관성 있다.

동일한 정보에 대해서 반복하지 않아야 한다.

주어진 문서에 대해서 단순하게 키워드를 뽑아내는 Topic Modeling과는 다르게 Text Summarization은 Text Generation의 하위 테스크입니다. Summary를 사람이 한 것처럼 만들기 위해서는 문서에 대한 요약이 필요하며 Summary 자체의 특성을 알아야 합니다.

Introduction

Text Summarization의 세분화

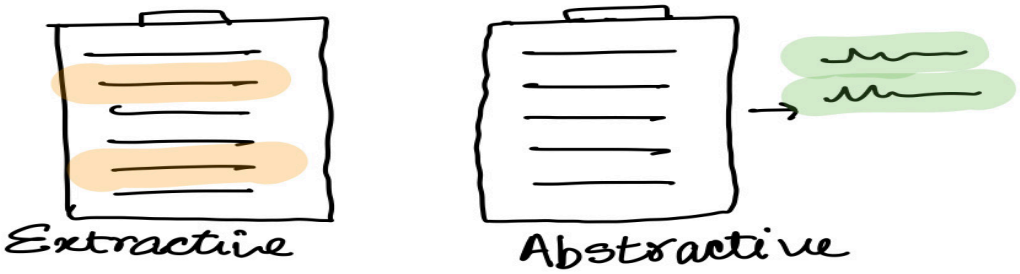
Extractive vs Abstractive

▶ Extractive는 문서의 내용을 변경하지 않고 주요 문장을 추출하는 것 입니다.

▶ Abstractive Summarization은 문서에 있던 문장들을 그대로 사용하지 않고 Paraphrase해서 요약하는 방법입니다.

▶ Extractive Summarization의 문제점은 문장을 선택함으로써 전체 Summary가 부자연스러울 수 있고, 내용이 의미하는 바를 모델이 알지 못할 가능성이 높다는 것 입니다. 이는 문장을 paraphrase하지 못하기 때문입니다.

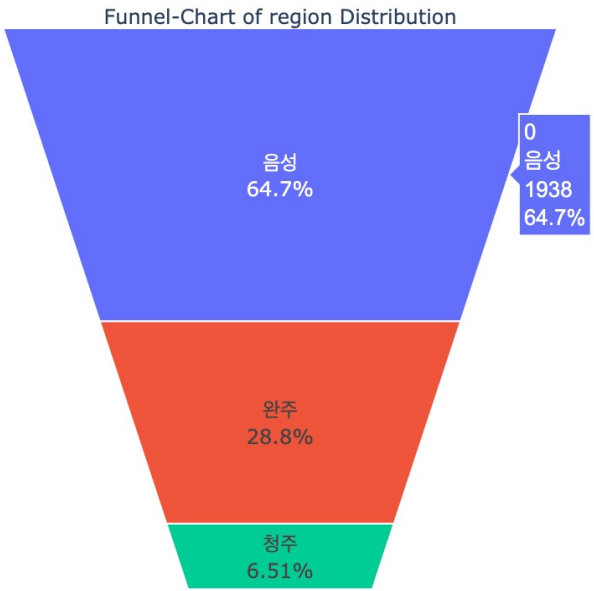
▶ 최근에는 Extractive에서 Abstractive로 Text Summarization의 흐름이 바뀌었습니다

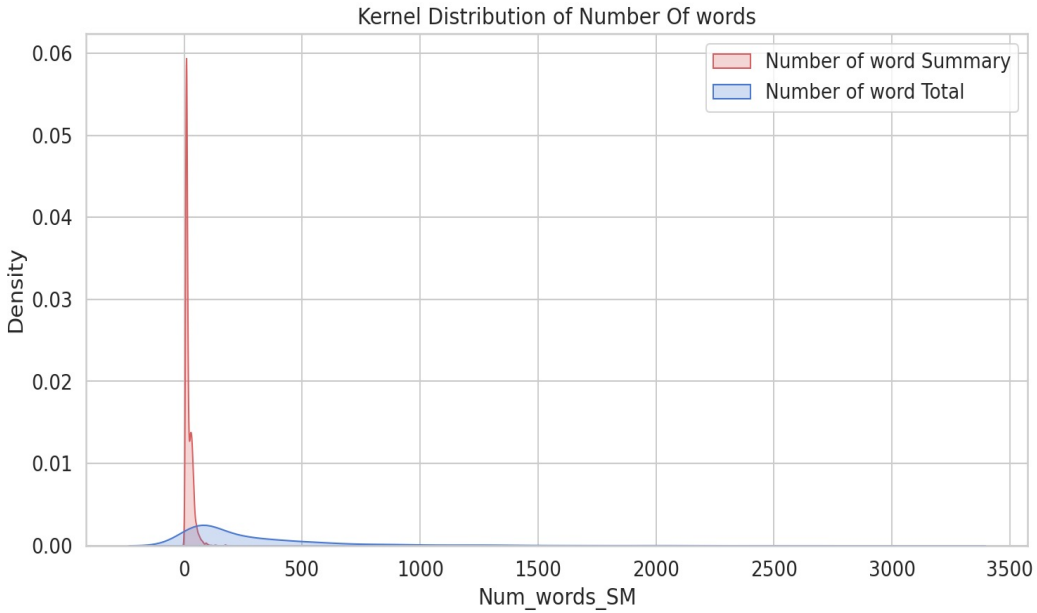


2
EDA

EDA

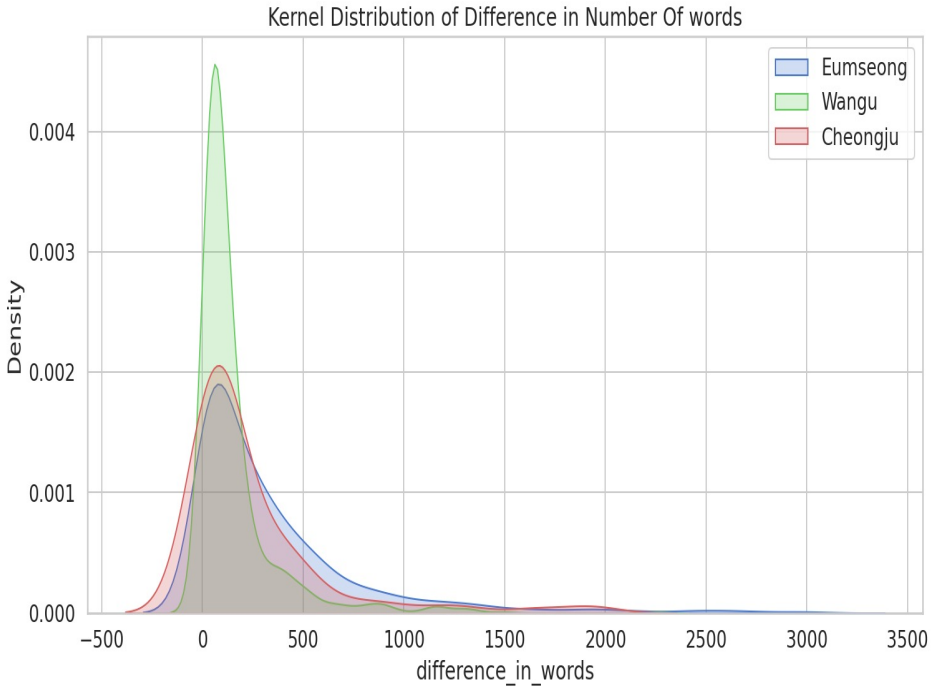
Data distribution

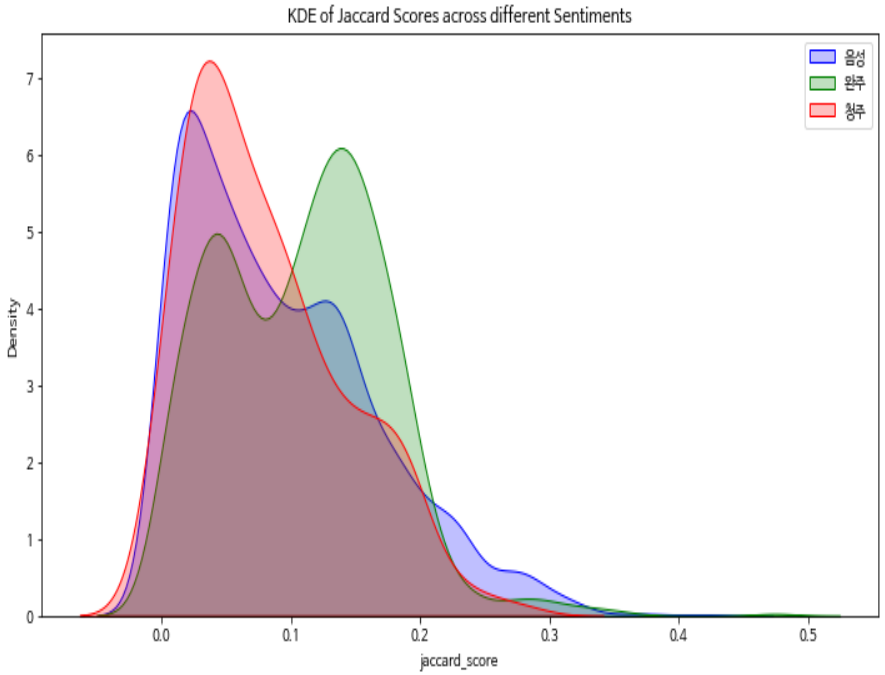




EDA

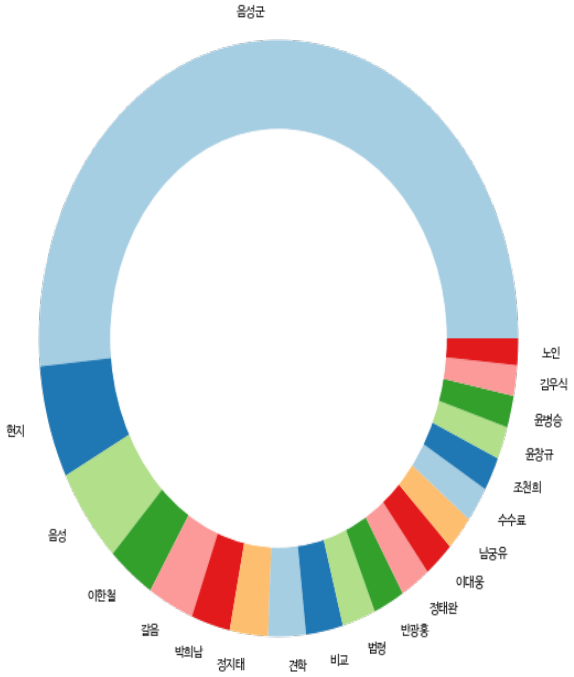
Data distribution

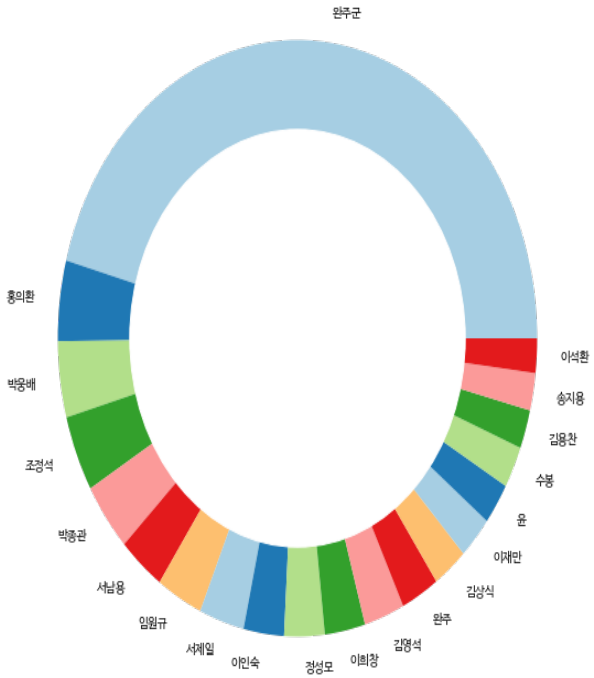


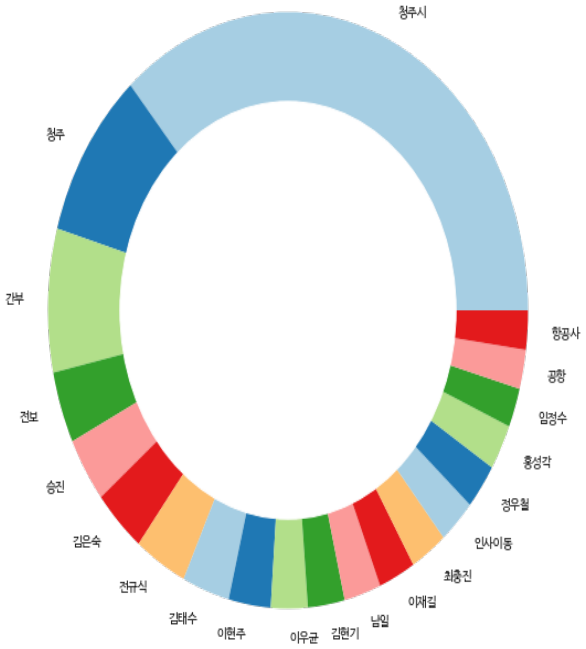


EDA

The only word by region

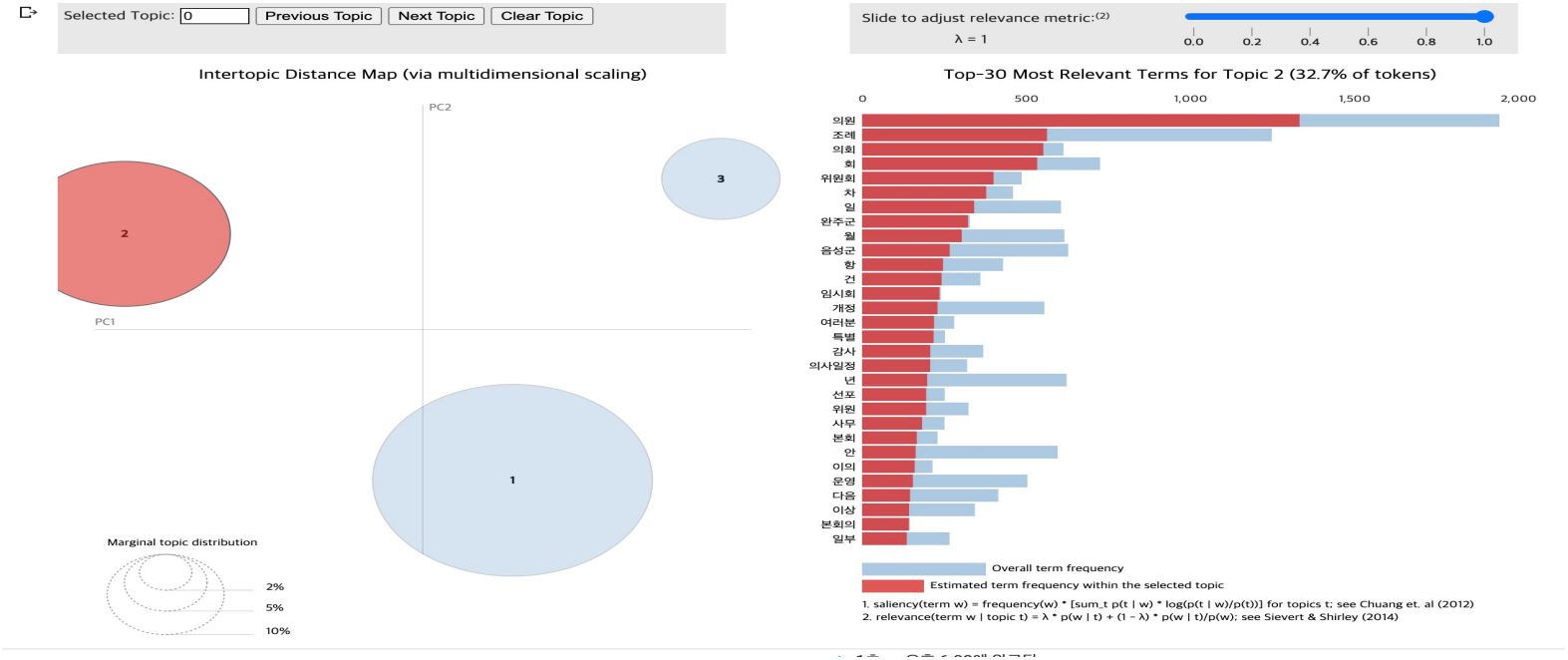






EDA(LDA)

Topic Modeling



Topic 1 : 사업 추진에 관한 녹취록

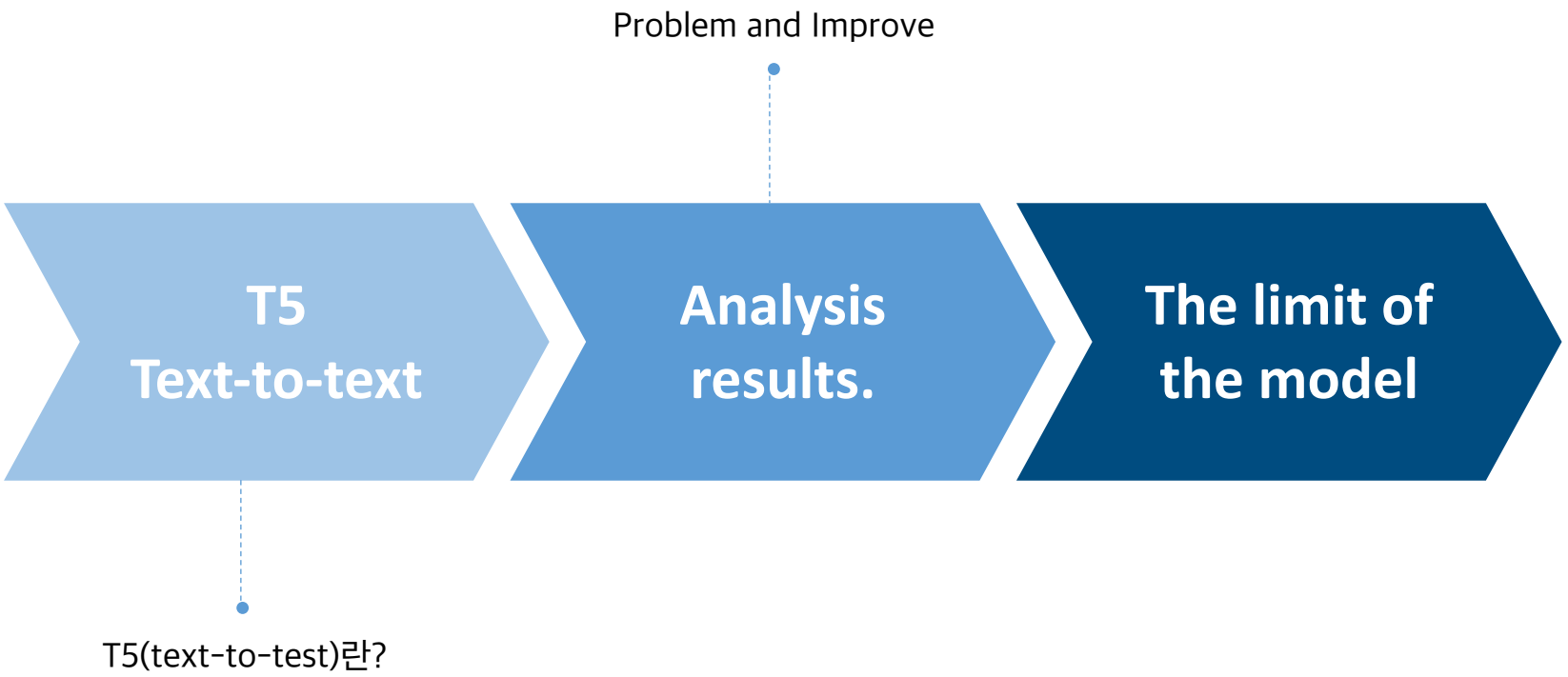
Topic 2 : 임시회 및 감사에 관한 녹취록

Topic 3 : 예산 운용에 관한 녹취록

3 Model

Model

Flow chart



Model

T5(text-to-text)

구글에서 제안한 T5 구조는 상당히 특이한 형태로 문제들을 기술합니다.
기존의 MT-DNN 등의 모델들이 배치 단위에서 실 훈련 데이터를 변경하며 다양한 문제를 모델에 투입하였다면, T5는 모든 문제를 문장 형태로 추상화 한 다음에, 그 추상화된 문장을 푸는 것을 훈련시킨 구조입니다.

Model

T5(text-to-text)

Model config

vocab_size : 35100

dropout_rate: 0.1

batch_size : 5

feed_forward_proj : relu

max_length : 512

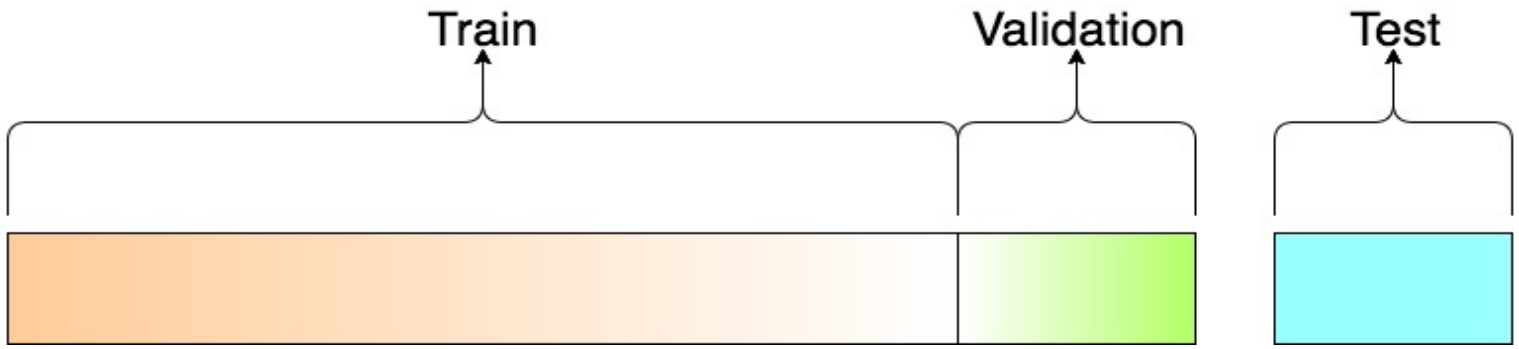
num_layers : 12

Summarization : { "early_stopping": true, "length_penalty": 1.5, "max_length": 150, "no_repeat_ngram_size": 3,"num_beams": 2, "prefix": "summarize: " },

항목	BERT	XLNet	RoBERT	MT-DNN	T5
Parameter Size	Base:110M Large:340M	Base:110M Large:340M	Base:110M Large:340M	Base:110M Large:340M	220M
Train Data Size	16GB	Base: 16GB(BERT) Large: 113GB	Base: 16GB(BERT) Large: 160GB	16GB(BERT) + GLUE	750GB (C4)
GLUE Score	80.5	88.4	88.5	85.1	89.7

Model

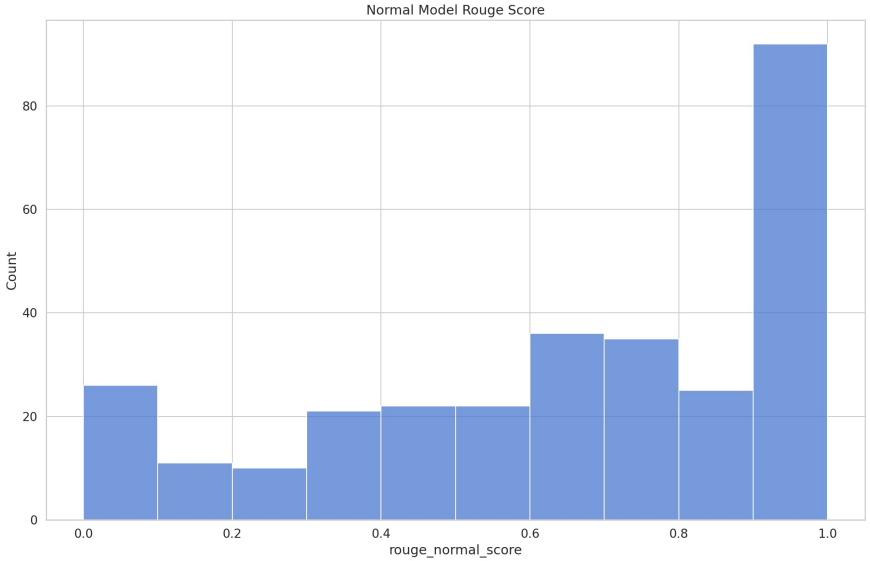
Analysis results

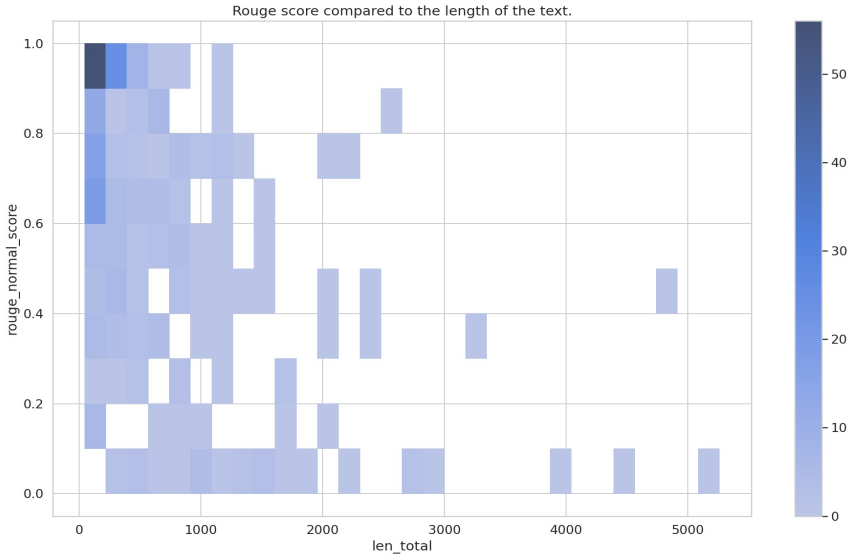


- Train data(total, summary) : 전체 train 데이터의 90%
- Validation data(total, summary) : 전체 train 데이터의 10%
- Test data(total) : 전체 test 데이터

Model

Analysis results





- 문서의 길이가 긴 data의 경우 Rouge score가 낮은 걸 확인할 수 있다.
- 문서의 길이가 짧은 data의 경우 전반적으로 rouge score가 높지만 여전히 요약하지 못하는 문서가 다수 존재한다.
- Rouge score : 0.657

Model

Analysis results and problem

region	Rouge score	Text length
완주	0.741399	367
음성	0.646854	672
청주	0.404990	369

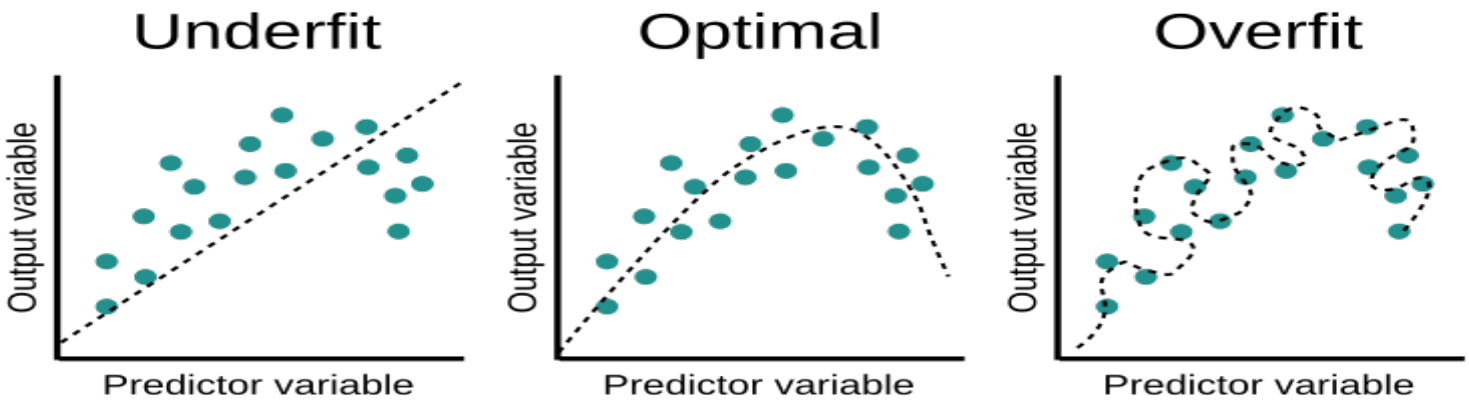
topic	Rouge score	Text length
Topic 1 (사업 추진)	0.620722	255
Topic 2 (임시회 및 감사)	0.480144	458
Topic 3 (예산 운용)	0.686118	626

청주 지역에서 나온 요약 문서에 대한 Rouge score값이 매우 떨어진다.
음성 지역의 경우 Rouge score의 값이 낮지만 문서 길이가 평균적으로 길기 때문이라고 생각된다.

Topic을 기준으로 살펴본 결과 topic2에 해당되는 문서의 rouge score값이 낮은 것을 확인할 수 있다.

Model

Analysis results and problem



Underfitting

Summary : 의회사무국 간부공무원 소개.

T5_Summary : 제49회 청주시의회 임시회는 이재길 의원 외 열두 분의 의원으로부터 조례안 처리를 위한 집회요구가 있어 집회함.

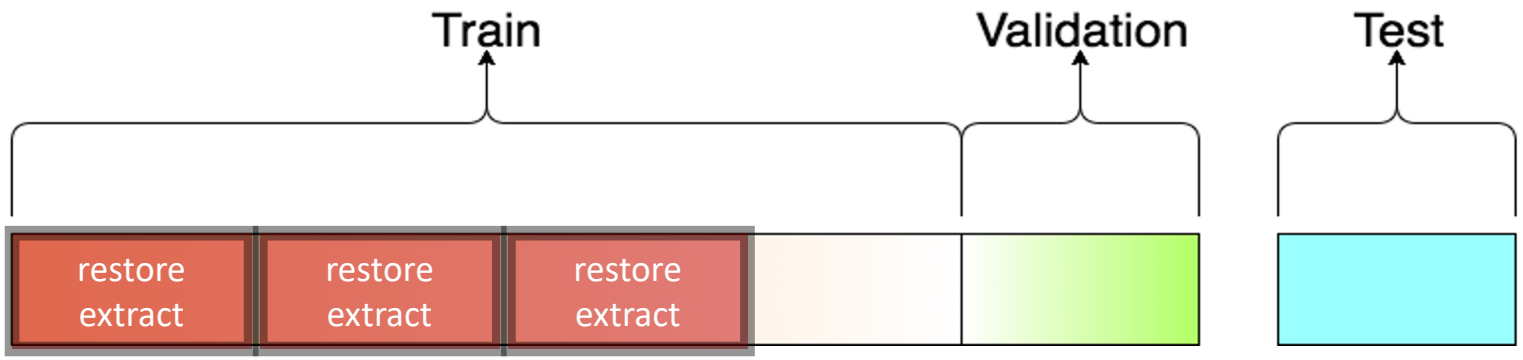
Overfitting

Summary : 과장님과 현장에 나가서 창고를 늘려 퇴비생산량을 늘리는 게 아니라 냄새를 없애는 방법을 찾고 증축 을 하게끔 유도할 것

T5_Summary : 제321회 음성군 행정사무감사 시정 및 건의사항 조치결과 보고.

Model

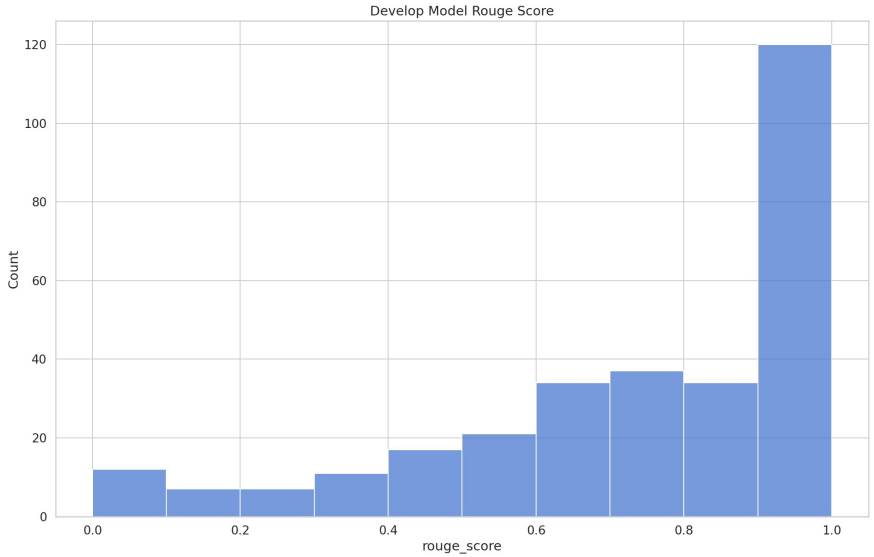
Analysis results and improvement.

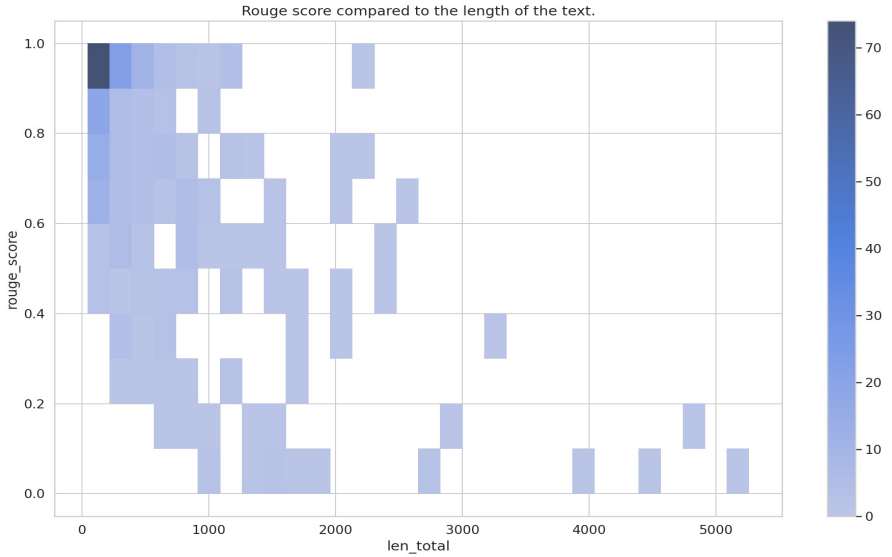


- Train data(total, summary) : random 하게 데이터의 20% 를 제거(복원 추출)하고 학습한다.
  학습 epoch을 기존 학습보다 더 많이 한다.
- Validation data(total, summary) : 전체 train 데이터의 10%
- Test data(total) : 전체 test 데이터

Model

Analysis results and improvement.





- 문서의 길이가 긴 data의 경우 Rouge score가 개선이 있다.
- 문서의 길이가 짧은 data의 경우 문서의 길이 1000이하인 문서에 대해서 Rouge score가 많이 개선되었다.
- Rouge score : 0.7496

Model

Analysis results and improvement.

region	Rouge score	Text length
완주	0.845376	367
음성	0.712283	672
청주	0.734819	369

topic	Rouge score	Text length
Topic 1 (사업 추진)	0.771890	255
Topic 2 (임시회 및 감사)	0.739215	458
Topic 3 (예산 운용)	0.765075	626

전체적인 Rouge score의 값이 개선되었다

문제가 되었던 청주, 음성 지역의 rouge score이 개선되었다.

topic 2 의 rouge score의 값이 크게 개선되었다.

Model

Analysis results and improvement.

Underfitting -> fitting

Summary : 의회사무국 간부공무원 소개.

T5_Summary : 의회사무국 간부공무원 소개

Overfitting -> fitting

Summary : 완주군 도시계획 조례안은 도시계획법과 같은 법 시행령이 개정됨에 따라, 종전 건축법의 적용을 받던 도시계획 구역내의 건축물이 도시계획법조례의 적용을 받게 되면서 해당 조례에서 정하도록 위임된 사항을 규정하고자 제정되었음. 해당 안건은 가결됨

T5_Summary : 완주군도시계획조례안은 도시계획법과 같은 법 시행령의 개정으로 종전 건축법의 적용을 받던 도시게획 구역내의 건축물이 도시계획법의 규정을 적용토록 변경됨에 따라 동 법률에서 조례에 정하도록 위임된 사항을 정하고자 제정됨. 해당 안건은 가결됨.

Model

The limit of the model

Gpu의 메모리 용량 제한

Gpu의 메모리 용량의 한계로 input data의 max_length의 길이를 늘리는데 한계가 있고, Batch_size의 크기에도 제한 사항이 있다.

문서의 길이가 긴 경우

모델을 개선하여 짧은 문서에 대한 정확도는 올라갔지만 긴 문서의 길이가 2000이 넘어가는 경우 정확도가 크게 낮다.

데이터의 불균형

지역간의 데이터 수의 차이가 많이 나며, 데이터가 적은 지역의 문서를 요약하는 경우 성능이 좋지 않다.

HyperParameter Tuning

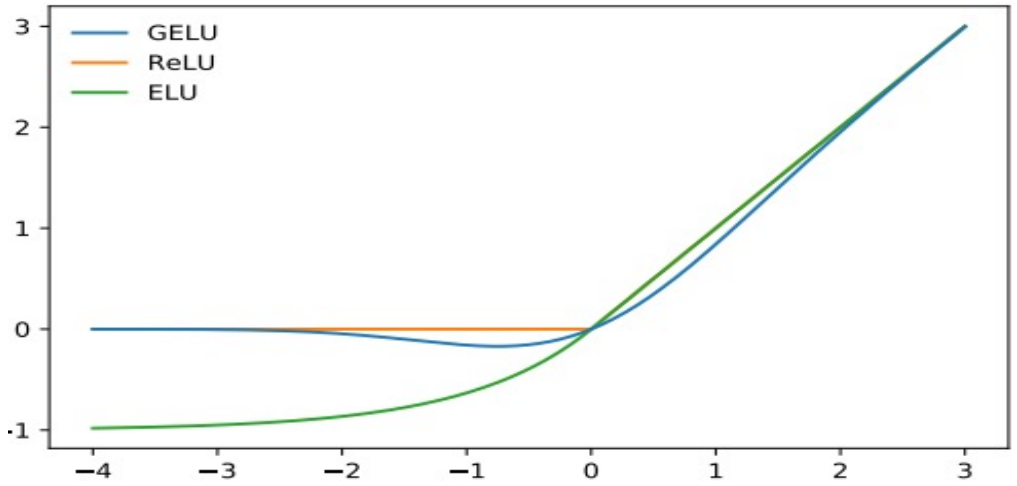
Learning_rate, epoch, Regularizaztion 등의 파라미터 튜닝을 다양하게 진행하지 못했다.

4

Development potential

Development potential

Activation function



활성화 함수(GELU)

GELU 활성화 기능을 사용하는 모델의 성능은 ReLU 또는 고급 버전 ELU(Exponential Linear Unit) 활성화 기능을 사용하는 모델의 성능과 비슷하거나 그 이상이다. GELU는 BERT, ROBERTa, ALBERT 및 기타 상위 NLP 모델과 호환된다.

T5 모델에 사용한 활성화 함수는 ReLU이다.
이것을 GELU로 바꿔 모델링을 진행할 경우 성능 향상을 기대할 수 있다.

Development potential

EDA: Easy Data Augmentation

EDA: Easy Data Augmentation

- SR: Synonym Replacement, 특정 단어를 유의어로 교체
- RI: Random Insertion, 임의의 단어를 삽입
- RS: Random Swap, 문장 내 임의의 두 단어의 위치를 바꿈
- RD: Random Deletion: 임의의 단어를 삭제

Data Augmentation을 통해서 training data의 수를 늘려준다.

THANK YOU!