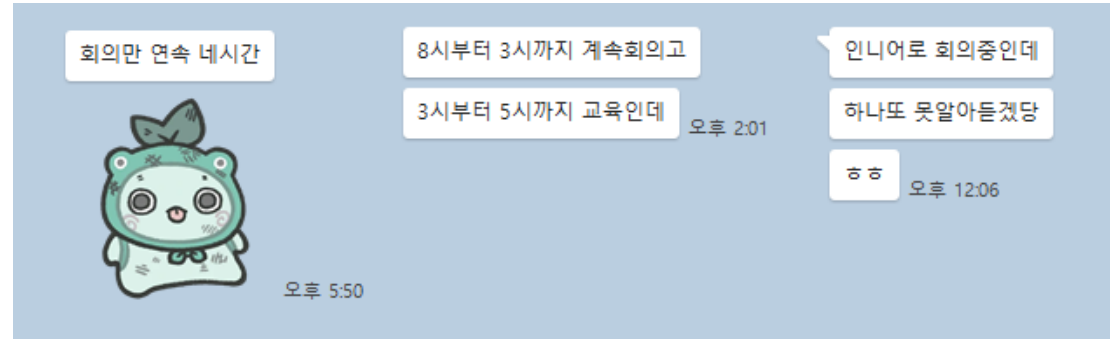
A photograph of a person's hands writing on a white document with a black pen. The person is wearing a dark long-sleeved shirt and a thin black wristband. The image is semi-transparently darkened to serve as a background for the text.

회의록 요약문 생성 모델 Text Summarization Model

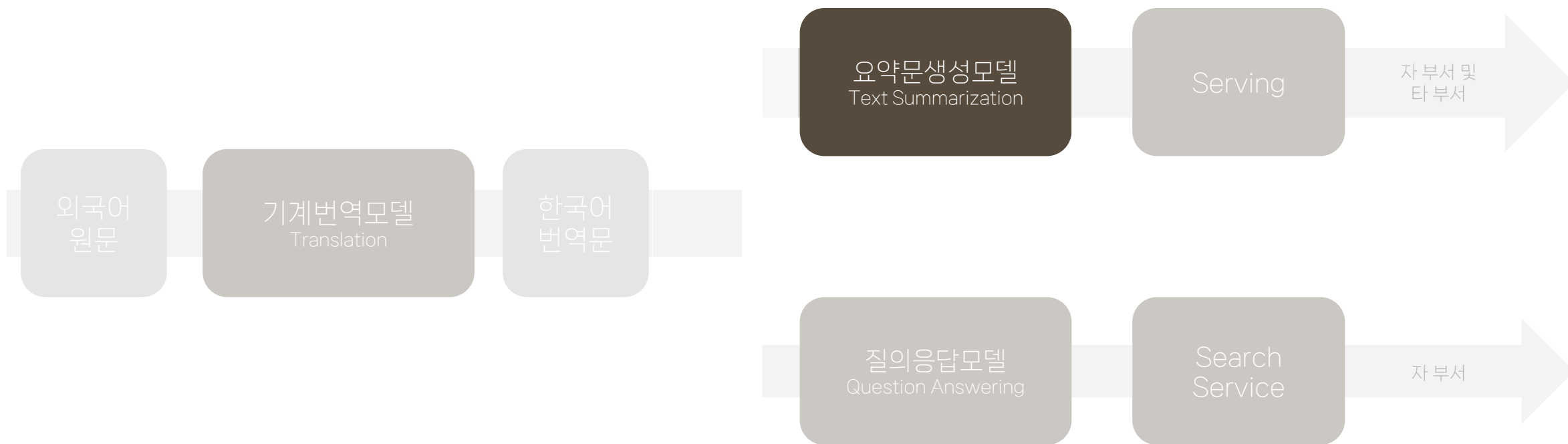
Deployment Background



과중한 반복업무, 언어장벽, 수기 방식 회의록의 무한 굴레

Service 구상안





Model

Text Summarization Task

추출적 요약
Extractive Summarization

TextRank

문장 단위로 중요도를 scoring한 후,
이를 기반으로 선택하고 조합하여 summary하는 비지도학습

문장을 핵심적으로 정리하는
뉴스기사, 소논문 등의 요약 task에 적합

추상적 요약
Abstractive Summarization

KoBART

원문을 기반으로 하되, 새로운 텍스트를 생성해내는 NLG natural language generation 방식(지도학습)으로, Transformer/BERT 등이 해당

BART Bidirectional and Auto-Regressive Transformers
- 현 Text Summarization 분야의 SOTA 모델
- 입력 텍스트 일부에 노이즈를 추가 text Infilling 하여
이를 다시 원문으로 복구하는 autoencoder 형태로 학습하는 구조

KoBART

- Text Infilling 노이즈 함수를 사용하는 BART 모델에 한국어 텍스트를 학습한 한국어 encoder-decoder 언어 모델

목표 및 가설 설정

실험

1

추출적 요약인 TextRank 알고리즘과
추상적 요약인 KoBART 모델의 성능 비교



가설

높은 추상화수준을 가진 KoBART 모델의 성능
이 TextRank보다 높을 것이다

2

직접 학습시킨 KoBART 모델과
전이학습TL한 KoBART 모델의 성능 비교



데이터의 크기 상, 전이학습 모델의 성능이 직접
학습 모델보다 높을 것이다

3

Tokenizer에 따른 성능 비교



Tokenizer에 따라 결과가 달라질 수 있기에 유
의미한 성능 차이가 있을 것이다

평가

정량적 평가 Rouge-1, 2, L score
정성적 평가 Human Evaluation

DataSet & Model

Data 데이콘 AI 기반 회의 녹취록 요약 경진대회

- ✓ **Source** 실제 국회에서 녹취된 내용을 문서화
- ✓ **Structure** 회의 별, 안건 별 분할

Data Preprocessing

- ✓ 각 모델 및 토크나이저의 요구 format에 맞춰 가공
- ✓ Textrank 특수기호 삭제
- ✓ Tokenizer 요구조건에 따라 1024글자 이상의 데이터 제거

최종 DataSet

- ✓ **Train** 1674
- ✓ **Val** 559
- ✓ **Test** 458 (rows)

Library

- ✓ HuggingFace Library에서 제공하는 모델 및 tokenizer 중
- ✓ **Model** digit82/koBART-summarization
- ✓ **Tokenizer** gogamza/KoBART-summarization, digit/KoBART-summarization

Id:1000,#회의번호

context:{AGENDA_1: #안건번호 및 회의내용
['의사일정 제3항 의사일정변경의 건을 상정합니다. 당초 7월 3
일 제2차 본회의시 상임위원회 위원 선임의 건과 상임위원장 선
거를 4항과 5항으로 상정하여 처리하고자 하는데 의원여러분 이
의 없으십니까?'],
'(『없습니다』 하는 의원 있음.)',
'의의가 없으므로 가결되었음을 선포합니다. 다음은 상임위원회
위원 선임과 상임위원장 선거입니다만 위원 선임을 하여야 하기
때문에 잠시 정회를 선포합니다.'],

label:{'AGENDA_1': #요약문
'제2차 본회의에 상임위원회 위원 선임의 건과 상임위원장 선거
를 상정하기로 함. 해당 안건은 가결됨.'},

...

Model Result – 1) TextRank vs KoBART

TextRank

Metric	sub	Score
Rouge-1	f1-score	0.1503
	precision	0.1949
	recall	0.1462
Rouge-2	f1-score	0.1011
	precision	0.1350
	recall	0.0993
Rouge-L	f1-score	0.1490
	precision	0.1932
	recall	0.1449

Summary TextRank	
2420	당초 7월 3일 제2차 본회의시 상임위원회 위원 선임의 건과 상임위원장 선거를 4항과 5항으로 상정하여 처리하고자 하는데 의원 여러분 이의 없으십니까
1724	다음은 지방자치법 제64조제2항과 회의규칙 제46조 제1항의 규정에 의하여 제102회 완주군의회 임시회 회의록에 서명하실 의원님을 선임하고자 합니다.
1124	존경하는 의장님, 그리고 동료 의원 여러분!완주군의회 예산결산 특별위원회 구성 결의안에 대하여 제안설명 드리도록 하겠습니다.

Label
제2차 본회의에 상임위원회 위원 선임의 건과 상임위원장 선거를 상정하기로 함. 해당 안건은 가결됨.
제102회 완주군의회 임시회 회의록 서명의원으로 박웅배 의원과 김순길 의원을 선임함.
예산결산특별위원회위원은 의장을 제외한 9인으로 구성함. 특별위원회는 2012년 7월 10일부터 2013년 7월 9일까지 1년간 2011년도 완주군 세입, 세출결산 및 예비비지출 승인안과 2012년도부터 2013년도 예산안 등 관련 안건을 의사일정에 따라 심의하고자 구성함. 해당 안건은 가결됨.

KoBART - 직접학습

Metric	sub	Score
Rouge-1	f1-score	0.8063
	precision	0.8189
	recall	0.8193
Rouge-2	f1-score	0.7279
	precision	0.7404
	recall	0.7365
Rouge-L	f1-score	0.8008
	precision	0.8137
	recall	0.8132

Summary KoBART(MyDL)	
7월 3일 제2차 본회의시 상임위원회 위원 선임의 건과 상임위원장 선거를 4항과 5항으로 상정하여 가결됨.	
제102회 완주군의회 임시회 회의록 서명의원으로 박웅배 의원과 김순길 의원이 선임됨.	
완주군의회 예산결산 특별위원회 구성 결의안은 2011년도 완주군 세입, 세출결산 및 예비비지출 승인안과 2012년도부터 2013년도 예산안 등 관련 안건을 종합적이고 객관적으로 심사하기 위해 <지방자치법>제56조와 「완주군의회 위원회 조례」 제7조 제2항에 의거 구성하려는 것으로 ... 해당 안건은 가결됨.	

Model Result – 2) MyDL vs PreDL

KoBART - 직접학습

Metric	sub	Score
Rouge-1	f1-score	0.8063
	precision	0.8189
	recall	0.8193
Rouge-2	f1-score	0.7279
	precision	0.7404
	recall	0.7365
Rouge-L	f1-score	0.8008
	precision	0.8137
	recall	0.8132

2420

Summary MyDL	
7월 3일 제2차 본회의시 상임위원회 위원 선임의 건과 상임위원장 선거를 4항과 5항으로 상정하여 가결됨.	
제102회 완주군의회 임시회 회의록 서명의원으로 박웅배 의원과 김순길 의원이 선임됨.	

1724

KoBART - 전이학습

Metric	sub	Score
Rouge-1	f1-score	0.3105
	precision	0.2439
	recall	0.5013
Rouge-2	f1-score	0.2135
	precision	0.1638
	recall	0.3638
Rouge-L	f1-score	0.3021
	precision	0.2372
	recall	0.4878

Summary PreDL	
7월 3일 본회의시 상임위원회 위원 선임의 건과 상임위원장 선거를 4항과 5항으로 상정하여 처리하고자 하는 의사일정변경의 건이 가결되었음을 선포하며 잠시 정회를 선포한다.	
지방자치법 제64조제2항과 회의규칙 제46조제1항의 규정에 따라 제102회 완주군의회 임시회 회의록에 서명하실 의원님은 박웅배 의원님과 김순길 의원님을 선임하고자 하는데 의원 여러분 이의 없으십니까?(『없습니다』 하는 의원 있음)그러면 두 분 의원님께서 본 회기동안 수고하여 주시기 바랍니다.	

Model Result – 2) MyDL vs PreDL

KoBART - 직접학습

Metric	sub	Score
Rouge-1	f1-score	0.8063
	precision	0.8189
	recall	0.8193
Rouge-2	f1-score	0.7279
	precision	0.7404
	recall	0.7365
Rouge-L	f1-score	0.8008
	precision	0.8137
	recall	0.8132

Summary MyDL	
1124	완주군의회 예산결산 특별위원회 구성 결의안은 2011년도 완주군 세입, 세출결산 및 예비비지출 승인안과 2012년도부터 2013년도 예산안 등 관련 안건을 종합적이고 객관적으로 심사하기 위해 <지방자치법>제56조와 「완주군의회 위원회 조례」 제7조 제2항에 의거 구성하려는 것으로 특별위원회 명칭은 '완주군의회 예산결산 특별위원회'로 하고, 구성인원수는 의장님을 제외한 9인으로 하며, 존속기간은 2012년 7월 10일부터 2013년 7월 9일까지로 하고자 제안됨. 해당 안건은 가결됨.

KoBART - 전이학습

Metric	sub	Score
Rouge-1	f1-score	0.3105
	precision	0.2439
	recall	0.5013
Rouge-2	f1-score	0.2135
	precision	0.1638
	recall	0.3638
Rouge-L	f1-score	0.3021
	precision	0.2372
	recall	0.4878

Summary PreDL	
	완주군의회 예산결산 특별위원회는 2011년도 완주군 세입·세출결산 및 예비비지출 승인안과 2012년도부터 2013년도 예산안 등 관련 안건을 종합적이고 객관적으로 심사하기 위해 '지방자치법' 제56조와 「완주군의회 위원회 조례」 제7조 제2항에 의거 구성하려는 것으로 특별위원회 명칭은 '완주군의회 예산결산 특별위원회'로 하고, 구성인원수는 의장님을 제외한 9인으로 하며, 존속기간은 2012년 7월 10일부터 2013년 7월 9일까지로 하고자 한다.

Label
예산결산특별위원회위원은 의장을 제외한 9인으로 구성함. 특별위원회는 2012년 7월 10일부터 2013년 7월 9일까지 1년간 2011년도 완주군 세입, 세출결산 및 예비비지출 승인안과 2012년도부터 2013년도 예산안 등 관련 안건을 의사일정에 따라 심의하고자 구성함. 해당 안건은 가결됨.

Model Result – 3) Tokenizer 비교

Tokenizer digit82/kobart-summarization

Metric	sub	Score
Rouge-1	f1-score	0.8063
	precision	0.8189
	recall	0.8193
Rouge-2	f1-score	0.7279
	precision	0.7404
	recall	0.7365
Rouge-L	f1-score	0.8008
	precision	0.8137
	recall	0.8132

Summary Tokenizer1	
2420	7월 3일 제2차 본회의시 상임위원회 위원 선임의 건과 상임위원장 선거를 4항과 5항으로 상정하여 가결됨.
1724	제102회 완주군의회 임시회 회의록 서명의원으로 박웅배 의원과 김순길 의원이 선임됨.

Label
제2차 본회의에 상임위원회 위원 선임의 건과 상임위원장 선거를 상정하기로 함. 해당 안건은 가결됨.
제102회 완주군의회 임시회 회의록 서명의원으로 박웅배 의원과 김순길 의원을 선임함.

Tokenizer gogamza/kobart-summarization

Metric	sub	Score
Rouge-1	f1-score	0.8063
	precision	0.8189
	recall	0.8193
Rouge-2	f1-score	0.7279
	precision	0.7404
	recall	0.7365
Rouge-L	f1-score	0.8008
	precision	0.8137
	recall	0.8132

Summary Tokenizer2	
2420	7월 3일 제2차 본회의시 상임위원회 위원 선임의 건과 상임위원장 선거를 4항과 5항으로 상정하여 가결됨.
1724	제102회 완주군의회 임시회 회의록 서명의원으로 박웅배 의원과 김순길 의원이 선임됨.

c.f. KoBART 직접학습모델에 적용함

Conclusion

- 추출적 요약인 TextRank 알고리즘과 추상적 요약인 KoBART 모델의 성능 비교
 - 높은 추상화수준을 가진 KoBART 모델이 TextRank 에 비해 유의미하게 성능이 높음을 알 수 있음
- 직접 학습시킨 KoBART 모델과 전이학습TL한 KoBART모델의 성능 비교
 - 가설과는 달리, 직접학습 모델이 전이학습 모델보다 유의미하게 성능이 높았음
 - 그러나 데이터셋 특성 상 과적합 우려가 있어, 다양한 데이터셋을 확보하여 학습 및 성능평가를 통해 일반화 성능을 높여야 할 필요성이 있음
- Tokenizer에 따른 성능 비교
 - KoBART-summarization 전용의 토큰라이저는 제작자가 다르더라도 요약문 생성에 큰 영향을 미치지 않았음
 - KoBART용이 아닌 일반적인 토큰라이저를 적용했을 때의 성능 차이를 확인할 필요성이 있음

TextRank

Metric	sub	Score
Rouge-1	f1-score	0.1503
Rouge-2	f1-score	0.1011
Rouge-L	f1-score	0.1490

Summary TextRank

2420

당초 7월 3일 제2차 본회의시 상임위원회 위원 선임의 건과 상임위원장 선거를 4항과 5항으로 상정하여 처리하고자 하는데 의원 여러분
이의 없으십니까

1724

다음은 지방자치법 제64조제2항과 회의규칙 제46조제1항의 규정에
의하여 제102회 완주군의회 임시회 회의록에 서명하실 의원님을 선
임하고자 합니다.

KoBART - 직접학습

Metric	sub	Score
Rouge-1	f1-score	0.8063
Rouge-2	f1-score	0.7279
Rouge-L	f1-score	0.8008

Summary KoBART(MyDL)

7월 3일 제2차 본회의시 상임위원회 위원 선임의 건과 상임위원장
선거를 4항과 5항으로 상정하여 가결됨.

제102회 완주군의회 임시회 회의록 서명의원으로 박용배 의원과 김
순길 의원이 선임됨.

KoBART - 전이학습

Metric	sub	Score
Rouge-1	f1-score	0.3105
Rouge-2	f1-score	0.2135
Rouge-L	f1-score	0.3021

Summary KoBART(PreDL)

7월 3일 본회의시 상임위원회 위원 선임의 건과 상임위원장 선거를
4항과 5항으로 상정하여 처리하고자 하는 의사일정변경의 건이 가
결되었음을 선포하며 잠시 정회를 선포한다.

지방자치법 제64조제2항과 회의규칙 제46조제1항의 규정에 따라
제102회 완주군의회 임시회 회의록에 서명하실 의원님은 박용배 의
원님과 김순길 의원님을 선임하고자 하는데 의원 여러분 이의 없으
십니까?(「없습니다」 하는 의원 있음)그러면 두 분 의원님께서 본
회기동안 수고하여 주시기 바랍니다.

한계점

- 질의응답 모델을 목표로 했으나, 구현하지 못 한 점
- 데이터셋 자체의 한계) 데이터셋 특성 상 특정 형태의 요약에 과적합 우려가 있어 더욱 다양한 데이터셋을 추가로 학습 및 성능평가에 반영
- 추후 Metric으로 한국어 Text Generation 평가에 적절한 RDASS, ROUGE-SU 등을 구현하여 평가
 - 한국어는 문장구조, 조사유무 등이 자유로운 교착어이기에, 형태학적 유사도보다 다른 지표가 고려될 필요성이 있음 ("나? 먹었지, 밥" vs "나? 밥 먹었지")
 - RDASS 문서(d), 정답 요약 문장(r), 모델이 생성한 요약 문장(p) 각각의 임베딩 벡터를 통해 코사인 유사도를 계산하여 유사도의 평균값을 구하는 지표
 - ROUGE-SU skip-gram과 unigram을 동시에 고려하는 지표
- KoBERT 와 KoGPT 등의 다른 딥러닝 모델과의 성능 비교

A background image showing a person's hands writing on a notepad with a pen. A coffee cup is visible on the left side of the frame. The image is dimmed to serve as a background for the text.

감사합니다