CUAI 하계 컨퍼런스 NLP 2팀

2023.07.25

발표자 : 이혜연

스터디원 소개 및 만남 인증

스터디원 1 : 김건호

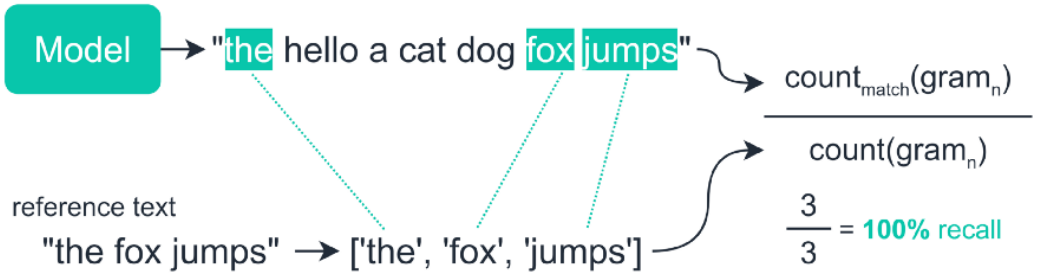
스터디원 2 : 이혜연

스터디원 3 : 최동욱

주제 확정

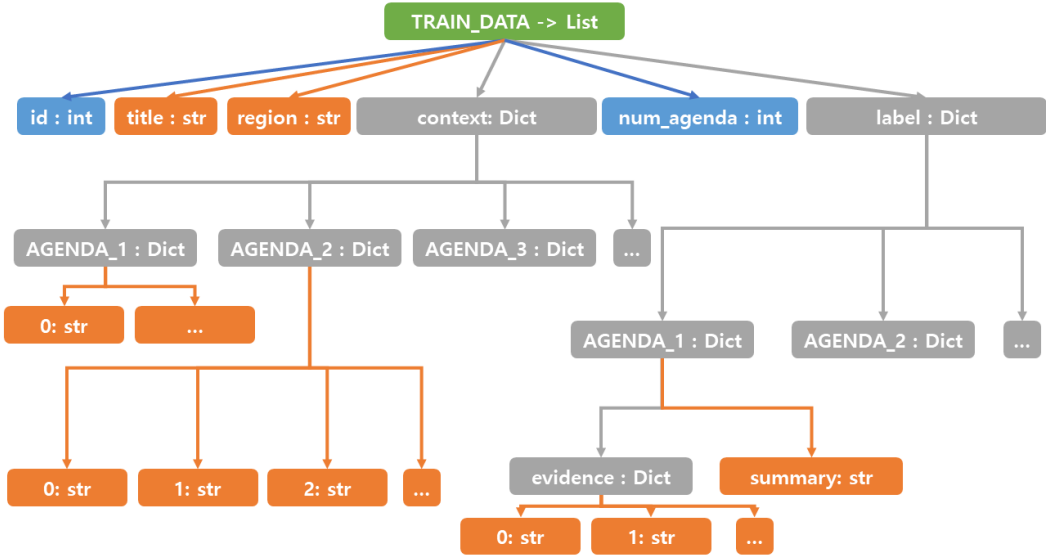
📍 **문서화된 회의 녹취록에서 핵심 내용을 요약하는 생성요약 AI 모델 개발**

ROUGE(Recall-Oriented Understudy for Gisting Evaluation) ?



데이터셋

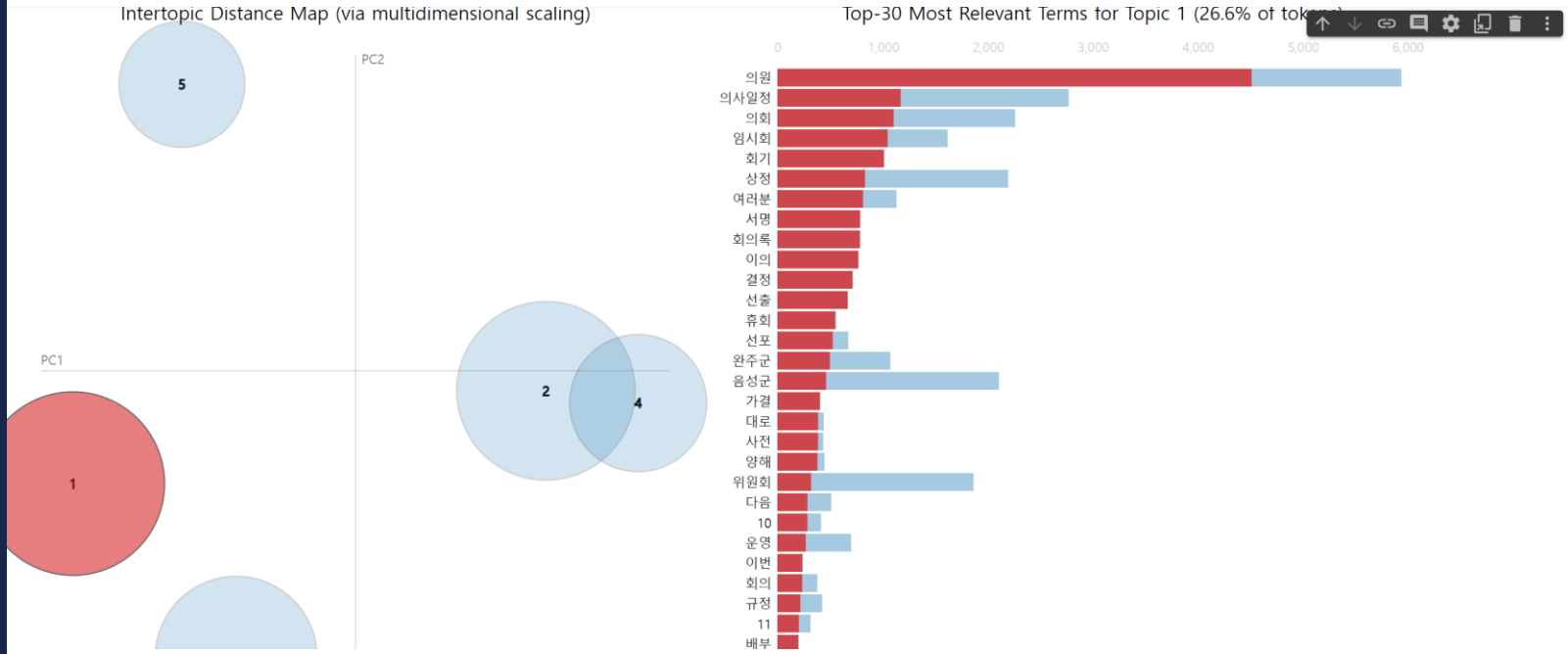
Context 내용
→ summary



train

	uid	title	region	context	summary	total
1000	1000	제207회 완주군의회(임시회) 제 1 차 본회의회의록	완주	의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제207회 완주군의회 임시회 제...	제207회 완주군의회 임시회 제1차 본회의 개의 선포.	제207회 완주군의회(임시회) 제 1 차 본회의회의록 완주 의석을 정돈하여 주시기 ...
1001	1001	제207회 완주군의회(임시회) 제 1 차 본회의회의록	완주	의사팀장 수고하셨습니다. 먼저 의사일정 제1항 제207회 완주군의회 임시회 회기 결...	제207회 완주군의회 임시회 회기는 8월 26일부터 9월 4일까지 10일간으로 가결됨.	제207회 완주군의회(임시회) 제 1 차 본회의회의록 완주 의사팀장 수고하셨습니다....
1002	1002	제207회 완주군의회(임시회) 제 1 차 본회의회의록	완주	다음은 의사일정 제2항 제207회 완주군의회 임시회 회의록 서명의원 선출의 건을 상...	제207회 완주군의회 임시회 회의록 서명의원으로 최등원 의원과 박웅배 의원이 선출됨.	제207회 완주군의회(임시회) 제 1 차 본회의회의록 완주 다음은 의사일정 제2항 ...
1003	1003	제207회 완주군의회(임시회) 제 1 차 본회의회의록	완주	다음은 의사일정 제3항 본회의 휴회의 건을 상정합니다. 상임의원회 의정활동을 위하여...	8월 27일부터 9월 3일까지 8일간 휴회가 가결됨. 제2차 본회의는 9월 4일 오...	제207회 완주군의회(임시회) 제 1 차 본회의회의록 완주 다음은 의사일정 제3항 ...
1004	1004	제251회 완주군의회(제1차 정례회) 제1차 본 회 의 회 의 록	완주	의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제251회 완주군의회 제1차 정...	제251회 완주군의회 제1차 정례회 제1차 본회의 개의 선포.	제251회 완주군의회(제1차 정례회) 제1차 본 회 의 회 의 록 완주 의석을 정돈...

토픽모델링을 통한 주제 분류



1. 안건 상정 / 2. 개정 / 3. 결산 / 4. 보고 / 5. 개의 선포

데이터 전처리

- 불용어 처리

```
import re
# context 전처리 최종 코드 (명사와 숫자가 나오도록 토큰화)
train['context_pre'] = train['context'].map(lambda x : re.sub(r'[^₩s₩w]', ' ', x)) # Text Cleaning
train['context_pre'] = train['context_pre'].map(lambda x : re.sub(r"[^ㄱ-ㅎㅏ-ㅣ가-힣0-9]", " ", x)) # 한글과 숫자 이외의 특수문자는 공백으로 변환
train['context_pre'] = train['context_pre'].map(lambda x: m.pos(x) if x.strip() else [])
train['context_pre'] = train['context_pre'].map(combine_tokens_with_xpn)
train['context_pre'] = train['context_pre'].map(combine_number_and_noun)
train['context_pre'] = train['context_pre'].map(lambda tokens: [word for word, pos in tokens if pos.startswith('N') or pos in tag])
```

제 207 회 → 제(XPN) + 207(SN) + 회(N)
2020 년 → 2020(XN) + 회(N)

```
train['context_pre']
```

```
1000    [의석, 정돈, 성원, 제207회, 완주군, 의회, 임시회, 제1차, 본회, 개의,...
1001    [의사, 팀장, 수고, 의사일정, 제1항, 제207회, 완주군, 의회, 임시회, 회...
1002    [다음, 의사일정, 제2항, 제207회, 완주군, 의회, 임시회, 회의록, 서명, ...
1003    [다음, 의사일정, 제3항, 본회, 휴회, 건, 상정, 상임, 원회, 활동, 8월,...
1004    [의석, 정돈, 성원, 제251회, 완주군, 의회, 제1차, 정례회, 제1차, 본회...
```

```
stopwords = ['다음', '이상', '여러분', '말씀', '수고', '내용', '우리', '저희']
```

데이터 전처리

- Mecab의 Custom dict

기존 Mecab의 형태소 분석기로 분석했을 때 오류

예산_NNG 결산_NNG 특별_NNG 위원회_NNG 최_XPN 등원_NNG 본회_NNG 의_JKG

Summary column 값 확인 → 관련 domain의 단어 및 이름 추가

Summary 추출 → SoyNLP의 noun extractor, word extractor 활용

```
제207회 완주군의회 임시회 제1차 본회의 개의 선포.
제207회 완주군의회 임시회 회기는 8월 26일부터 9월 4일까지 10일간으로 가결됨.
제207회 완주군의회 임시회 회의록 서명의원으로 최등원 의원과 박웅배 의원이 선출됨.
8월 27일부터 9월 3일까지 8일간 휴회가 가결됨. 제2차 본회의는 9월 4일 오전 10시에 개의.
제251회 완주군의회 제1차 정례회 제1차 본회의 개의 선포.
제251회 완주군의회 제1차 정례회 회기는 6월 3일부터 6월 17일까지 15일간으로 가결됨.
제251회 완주군의회 제1차 정례회 회의록 서명의원은 윤수봉 의원과 유의식 의원이 선출됨.
2020년도 제2회 추가경정예산안 제안설명.
6월 4일 1일간 본회의 휴회가 가결됨. 제2차 본회의는 6월 5일 오전 10시에 개의.
제210회 완주군의회 임시회 제1차 본회의 개의 선포.
제210회 완주군의회 임시회 회기는 2월 16일부터 2월 25일까지 10일간으로 가결됨.
제210회 완주군의회 임시회 회의록 서명의원으로 최상철 의원과 이인숙 의원이 선출됨.
```

Word Extraction

```python
from soynlp.word import WordExtractor

word_extractor = WordExtractor(min_frequency=100,
    min_cohesion_forward=0.05,
    min_right_branching_entropy=0.0
)

word_extractor.train(sents) # list of str or like
words = word_extractor.extract()
```

데이터 전처리

- Mecab의 Custom dict

Cohesion, Branching Entropy가 높은 단어 위주 선별

cohesion_forward	R_branching_entropy	word
0.9842273252251338	0.7000406470407772	본회의
0.9643577426082242	2.7433781479191106	음성군
0.958074887465177	2.0824081734192936	가결됨.
0.9579576853020536	0.8182782883093869	완주군
0.9277890783909317	1.7512921719369776	특별위원회
0.9273721882046156	1.0593227349082772	임시회
0.9224597861865156	0.9714958068290026	완주군의회
0.9042290466335725	2.6083171278519037	서명의원으로
0.8981481481481481	1.4801958489844318	변경
0.8980558274020201	3.172441960474515	일부개정조례안은
0.8910566892716278	2.399376628878858	채택됨.
0.8654957473266583	0.8380944476445864	예산결산특별위원회

count	score	noun
523	1.0	음성군의회
473	1.0	완주군의회
239	1.0	특별위원회
223	1.0	일부개정조례안
137	1.0	행정사무감사
111	1.0	예산결산특별위원회
97	1.0	추가경정예산안
80	1.0	개정조례안
56	1.0	2019년
51	1.0	2018년
48	1.0	2020년
39	1.0	2013년
36	1.0	청주시의회
36	1.0	전부개정조례안

→ 이 중 도메인과 밀접하며, 고유명사에 가까운 것들 추가 + Count와 score 순으로 정렬

데이터 전처리

- 띄어쓰기 PyKoSpacing

Example

1004	1004	제251회 완주군의회(제1차 정례회) 제1차 본 회 의 회 의 록	완주	의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제251회 완주군의회 제1차 정...	제251회 완주군의회 제1차 정례회 제1차 본회의 개의 선포.

'본 회 의 회 의 록' → '본회의 회의록'

PyKoSpacing 패키지

```
from pykospacing import Spacing
spacing = Spacing()
```

```
train['title'] = train['title'].str.replace(" ", '').apply(spacing)
```

데이터 전처리

- 띄어쓰기 PyKoSpacing

```
train['title'][:100]
```

```
1000              제207회 완주군의회(임시회) 제 1 차 본회의회의록
1001              제207회 완주군의회(임시회) 제 1 차 본회의회의록
1002              제207회 완주군의회(임시회) 제 1 차 본회의회의록
1003              제207회 완주군의회(임시회) 제 1 차 본회의회의록
1004       제251회 완주군의회(제1차 정례회) 제1차 본 회 의 회 의 록
1005       제251회 완주군의회(제1차 정례회) 제1차 본 회 의 회 의 록
1006       제251회 완주군의회(제1차 정례회) 제1차 본 회 의 회 의 록
1007       제251회 완주군의회(제1차 정례회) 제1차 본 회 의 회 의 록
1008       제251회 완주군의회(제1차 정례회) 제1차 본 회 의 회 의 록
1009              제210회 완주군의회(임시회) 제 1 차 본회의회의록
1010              제210회 완주군의회(임시회) 제 1 차 본회의회의록
1011              제210회 완주군의회(임시회) 제 1 차 본회의회의록
1012              제210회 완주군의회(임시회) 제 1 차 본회의회의록
1013          제229회 완주군의회(임시회) 제1차 본 회 의 회 의 록
1014          제229회 완주군의회(임시회) 제1차 본 회 의 회 의 록
1015          제229회 완주군의회(임시회) 제1차 본 회 의 회 의 록
1016          제229회 완주군의회(임시회) 제1차 본 회 의 회 의 록
1017       제178회          완주군의회(임시회) 제 1 차 본회의회의록
1018       제178회          완주군의회(임시회) 제 1 차 본회의회의록
1019       제178회          완주군의회(임시회) 제 1 차 본회의회의록
1020       제178회          완주군의회(임시회) 제 1 차 본회의회의록
1021              제288회     본회의 제2차(2017.04.28.)
1022              제288회     본회의 제2차(2017.04.28.)
1023      제168회 전라북도 완주군의회(임시회) 제1호 본 회 의 회 의 록
1024      제168회 전라북도 완주군의회(임시회) 제1호 본 회 의 회 의 록
1025      제168회 전라북도 완주군의회(임시회) 제1호 본 회 의 회 의 록
1026      제168회 전라북도 완주군의회(임시회) 제1호 본 회 의 회 의 록
1027              제194회 완주군의회(임시회) 제 1 차 본회의회의록
1028              제194회 완주군의회(임시회) 제 1 차 본회의회의록
1029              제194회 완주군의회(임시회) 제 1 차 본회의회의록
1030              제194회 완주군의회(임시회) 제 1 차 본회의회의록
1031          제238회 완주군의회(임시회) 제1차 본 회 의 회 의 록
1032          제238회 완주군의회(임시회) 제1차 본 회 의 회 의 록
1033          제238회 완주군의회(임시회) 제1차 본 회 의 회 의 록
1034          제238회 완주군의회(임시회) 제1차 본 회 의 회 의 록
```

```
train['title'][:100]
```

```
1000              제207회 완주군의회(임시회) 제1차 본회의 회의록
1001              제207회 완주군의회(임시회) 제1차 본회의 회의록
1002              제207회 완주군의회(임시회) 제1차 본회의 회의록
1003              제207회 완주군의회(임시회) 제1차 본회의 회의록
1004       제251회 완주군의회(제1차 정례회) 제1차 본회의 회의록
1005       제251회 완주군의회(제1차 정례회) 제1차 본회의 회의록
1006       제251회 완주군의회(제1차 정례회) 제1차 본회의 회의록
1007       제251회 완주군의회(제1차 정례회) 제1차 본회의 회의록
1008       제251회 완주군의회(제1차 정례회) 제1차 본회의 회의록
1009              제210회 완주군의회(임시회) 제1차 본회의 회의록
1010              제210회 완주군의회(임시회) 제1차 본회의 회의록
1011              제210회 완주군의회(임시회) 제1차 본회의 회의록
1012              제210회 완주군의회(임시회) 제1차 본회의 회의록
1013          제229회 완주군의회(임시회) 제1차 본회의 회의록
1014          제229회 완주군의회(임시회) 제1차 본회의 회의록
1015          제229회 완주군의회(임시회) 제1차 본회의 회의록
1016          제229회 완주군의회(임시회) 제1차 본회의 회의록
1017          제178회 완주군의회(임시회) 제1차 본회의 회의록
1018          제178회 완주군의회(임시회) 제1차 본회의 회의록
1019          제178회 완주군의회(임시회) 제1차 본회의 회의록
1020          제178회 완주군의회(임시회) 제1차 본회의 회의록
1021            제288회 본회의 제2차(2017.04.28.)
1022            제288회 본회의 제2차(2017.04.28.)
1023      제168회 전라북도 완주군의회(임시회) 제1호 본회의 회의록
1024      제168회 전라북도 완주군의회(임시회) 제1호 본회의 회의록
1025      제168회 전라북도 완주군의회(임시회) 제1호 본회의 회의록
1026      제168회 전라북도 완주군의회(임시회) 제1호 본회의 회의록
1027              제194회 완주군의회(임시회) 제1차 본회의 회의록
1028              제194회 완주군의회(임시회) 제1차 본회의 회의록
1029              제194회 완주군의회(임시회) 제1차 본회의 회의록
1030              제194회 완주군의회(임시회) 제1차 본회의 회의록
1031              제238회 완주군의회(임시회) 제1차 본회의 회의록
1032              제238회 완주군의회(임시회) 제1차 본회의 회의록
1033              제238회 완주군의회(임시회) 제1차 본회의 회의록
1034              제238회 완주군의회(임시회) 제1차 본회의 회의록
```

모델

Transformer, KoBART 등을 활용

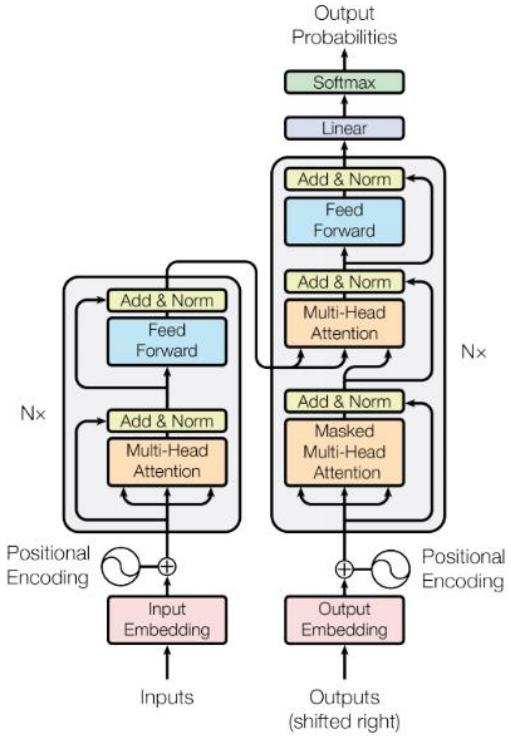
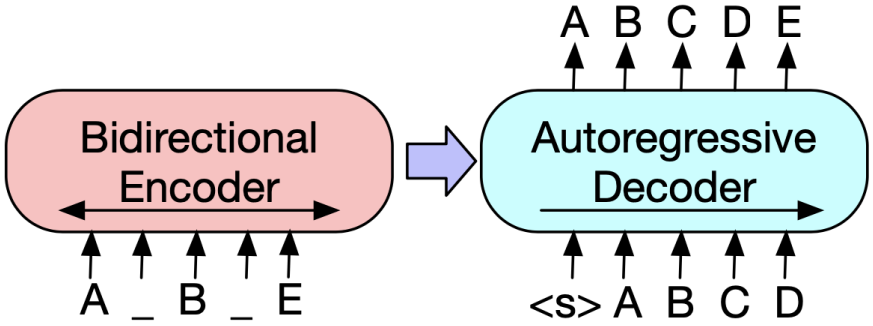


Figure 1: The Transformer - model architecture.



감사합니다