

**SK네트웍스 Family AI 과정 12기  
 데이터 전처리 인공지능 데이터 전처리 결과서**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 인공지능 데이터 전처리 결과서 |
| **제출 일자** | 2025.07.18 |
| **깃허브 경로** | <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN12-FINAL-3TEAM> |
| **작성 팀원** | 이준석 |

1. **문서 개요**

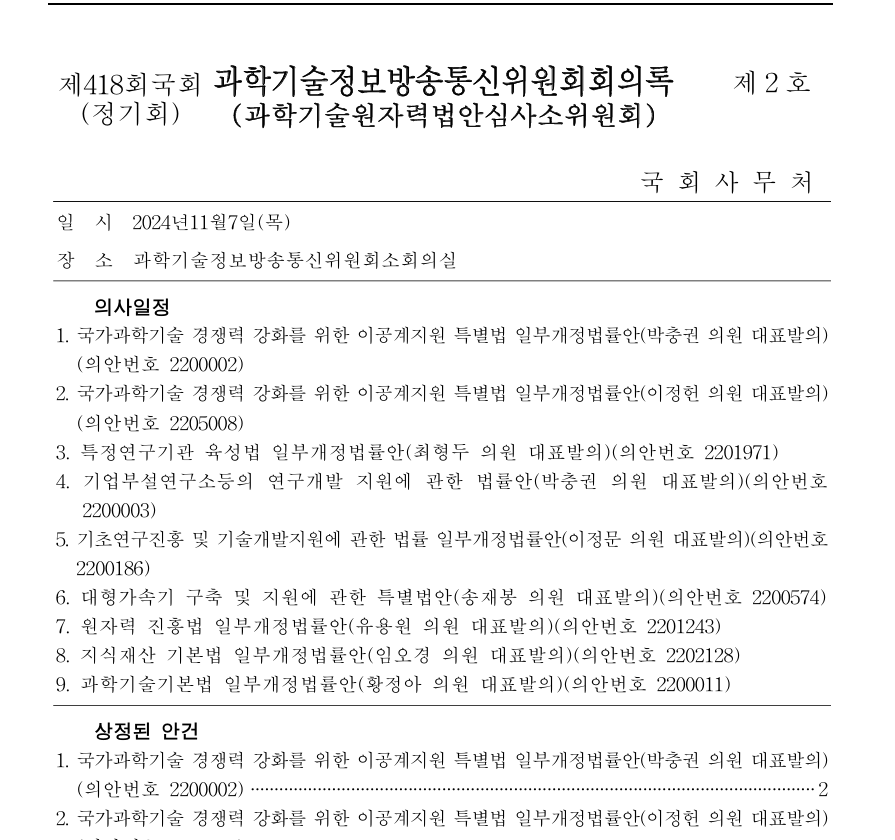
* 프로젝트명: TtalKkak – IT 기획/개발 회사를 위한 지능형 AI Agent Slack App
* 전처리 목적: 다자 회의 음성 데이터를 기반으로, 회의록을 정제하여 구조화된 기획안과 업무 분배를 가능하게 하는 인공지능 모델 학습용 고품질 데이터셋 구축
* 문제 정의:  
   회의 중 생성되는 다양한 발화 데이터는 발언자의 구분, 시간 정보, 내용 구성이 불명확한 경우가 많아 직접적인 AI 모델 학습에 적합하지 않다. 본 프로젝트에서는 국회 회의록, AMI, ICSI 등 공개된 다자 회의 데이터를 수집한 뒤, timestamp, speaker, text의 3요소로 정제하여 기획안 생성 및 업무 분배 기능을 수행할 수 있는 AI Agent 학습용 데이터셋을 구축하였다. 이로써 실제 회의 환경에서 발생하는 다양한 발화 유형을 효과적으로 반영할 수 있는 데이터 기반이 마련되었다.

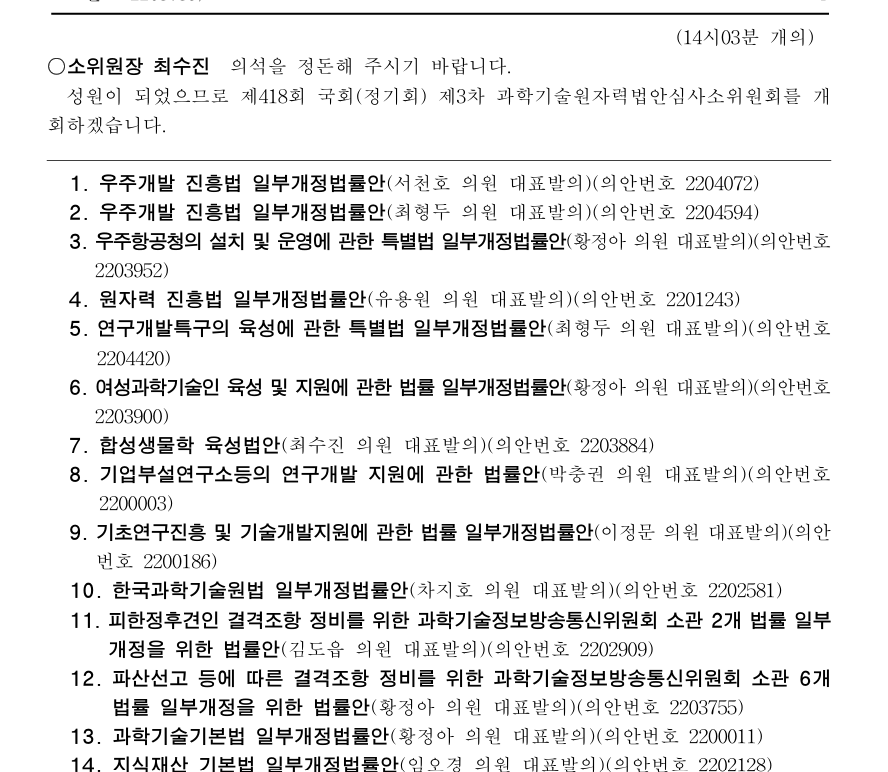
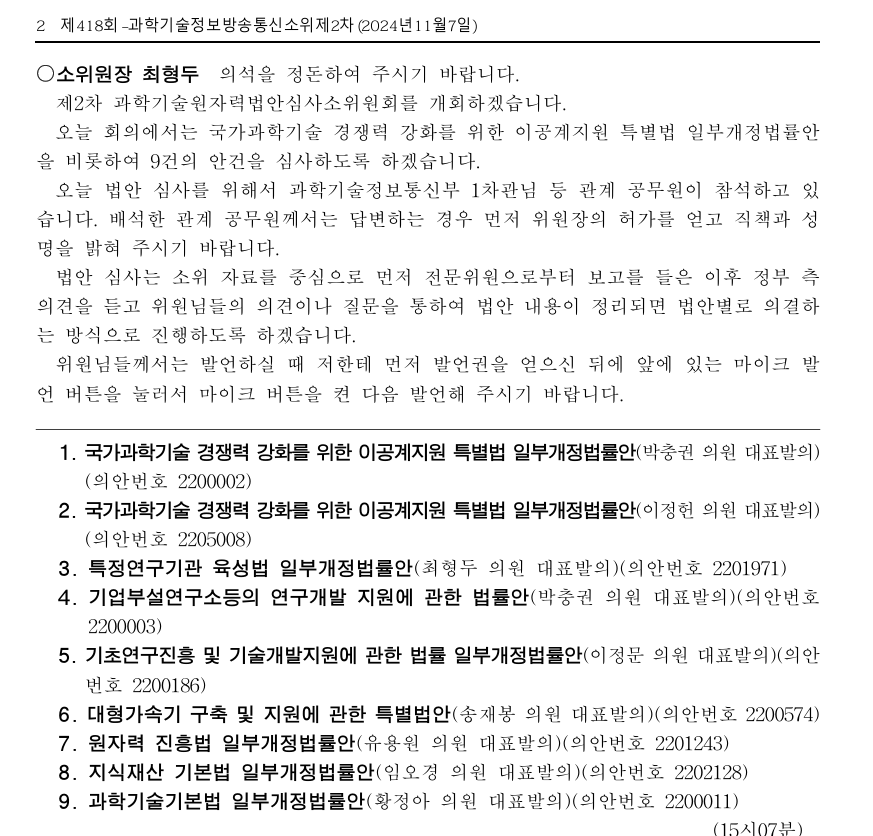
1. **데이터셋 개요**

* 데이터 출처 및 수집 방법:
* **국회 회의록**: 대한민국 국회 정보공개포털에서 수동 다운로드한 회의록 PDF 문서를 Python 스크립트로 전처리하여 구조화
* **AMI, ICSI Corpus**: The University of Edinburgh에서 공개한 영어 다자 회의 발화 데이터셋을 수동으로 수집, XML/CSV 파싱 및 정제
* **수집 방식**: 전량 수동 다운로드 후 전처리 스크립트 자동 실행
* **활용 동의**: 공개 데이터셋이므로 별도의 사용자 동의는 필요하지 않음
* 데이터 구성:

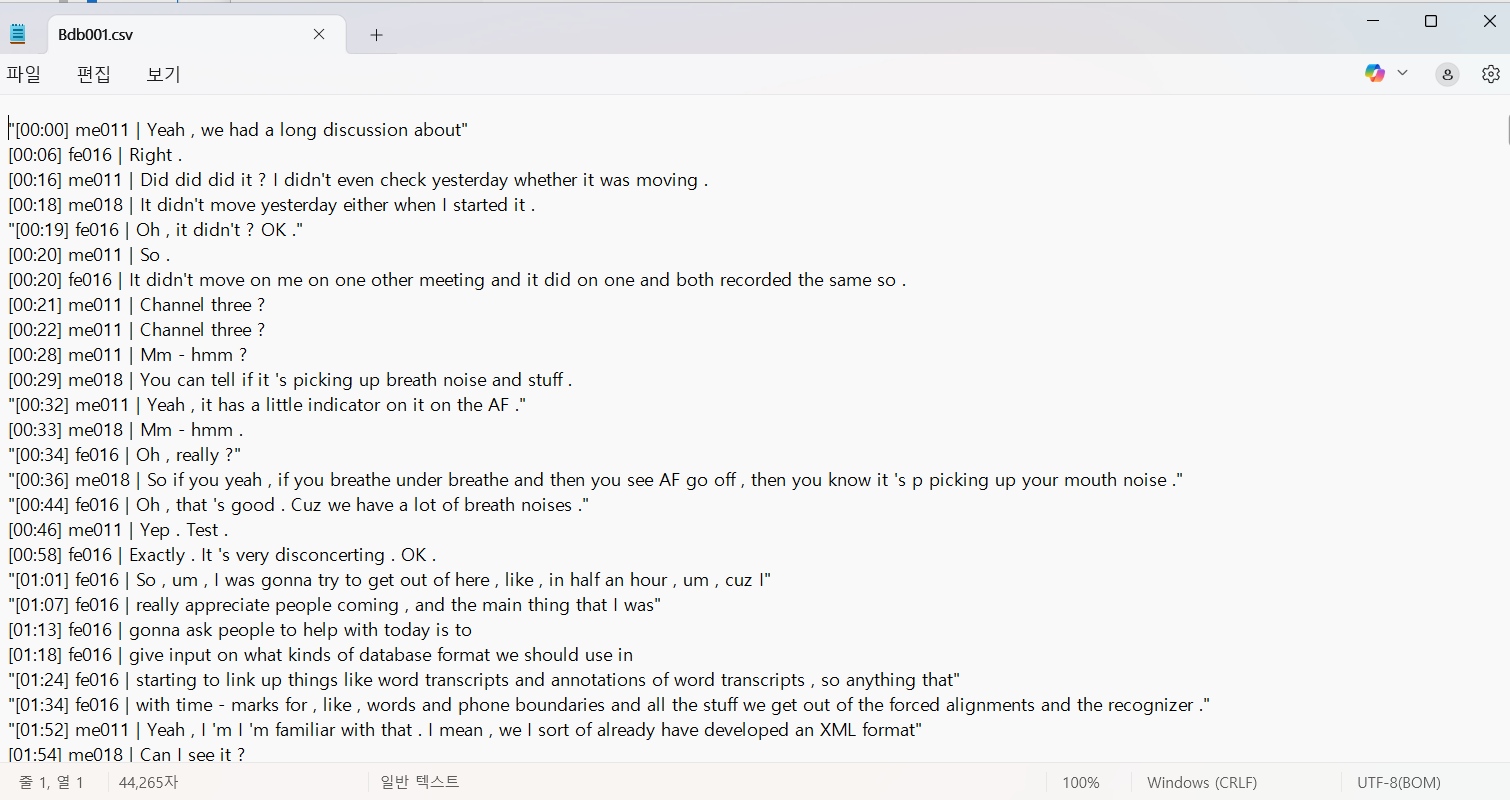
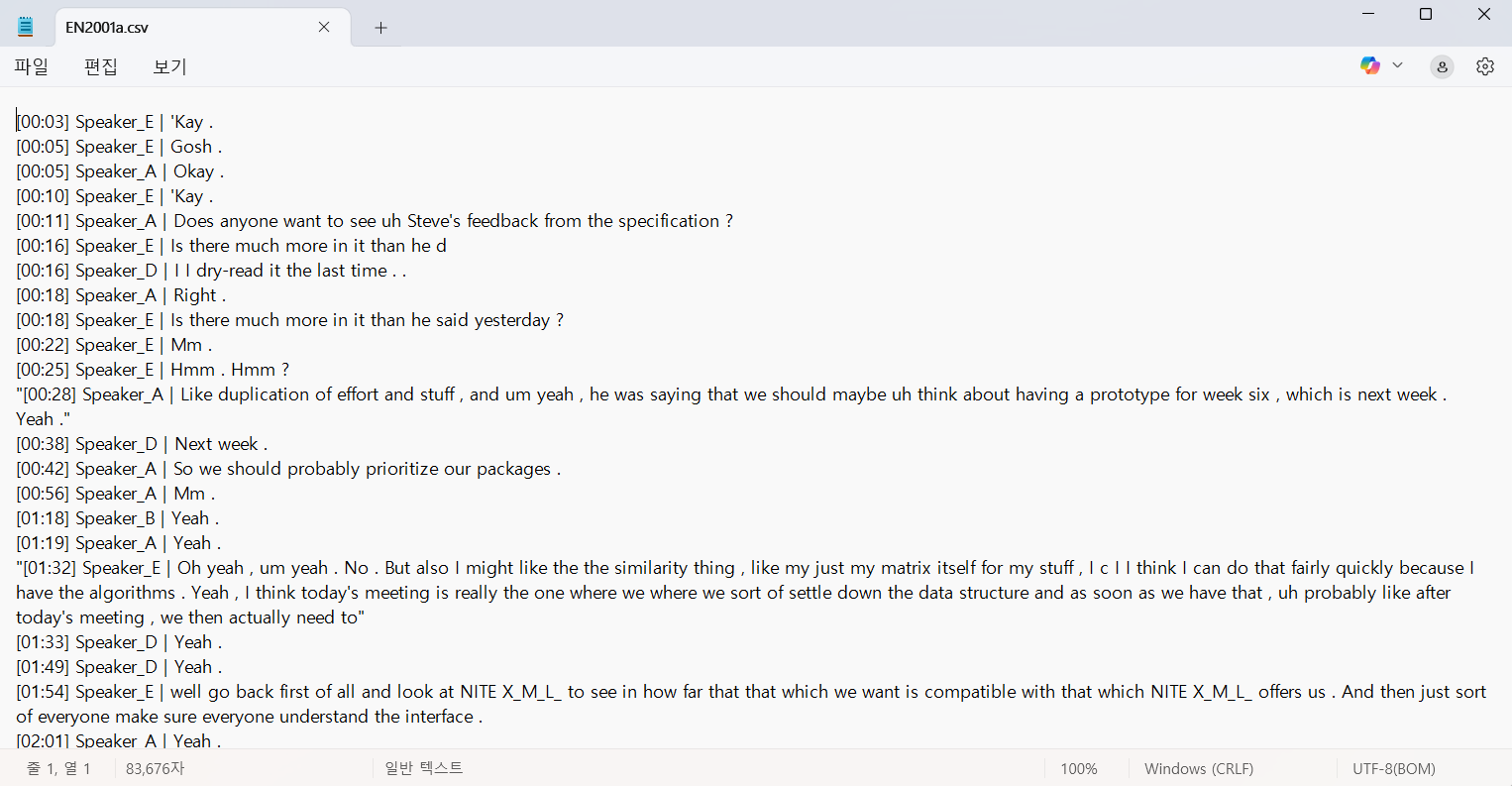
| **항목명** | **설명** | **예시** |
| --- | --- | --- |
| timestamp | 발화 시각 (분:초 또는 Null) | "00:11" / "05:28" / null   |  | | --- | |
| timestamp\_order | 발화 순서 | 29-1 / 20-2 / 39-9 |
| speaker | 발화자 식별자 | "위원장 최민희" / "Speaker\_A" / "fe016"   |  | | --- | |
| text | 발화 내용 (자연어 문장) | 2025-05-01 14:21:03 |

* 원본 데이터 샘플 (국회 회의록, AMI (CSV 형태), ICSI (CSV 형태), 5장)





<원천 데이터 (국회 회의록)>



< 원본데이터 예시 (AMI/ICSI)>

1. **전처리 프로세스 개요**

* 전체 흐름도:  
    
   ① 데이터 수집 → ② 텍스트 구조화(JSONL 변환) → ③ 결측치 및 이상치 처리 → ④ 학습용 데이터 분리
* 전처리 파이프라인 요약:

| **단계** | **목적** | **수행 작업** | **사용 도구/라이브러리** |
| --- | --- | --- | --- |
| 텍스트 구조화 | 다양한 형식의 데이터를 통일된 입력 형태로 변환 | PDF/XML/CSV 데이터를 timestamp, speaker, text 필드로 JSONL 구조화 | Python (re, xml.etree, pandas) |
| 결측치 및 이상치 처리 | 누락/무의미 발화 제거 및 데이터 품질 확보 | speaker/text 누락, 빈 문자열, ".", "@" 등 기호-only 발화, 특수기호/시스템 메시지 제거 | Python (re, pandas), 수동 검토) |
| 데이터 분리 | 모델 학습/검증/테스트 데이터셋 생성 | 전체 데이터(233,569건)를 Train/Val/Test로 무작위 분할 | sklearn.model\_selection |

1. **세부 전처리 단계**

4.1. 텍스트 구조화 (JSONL화)

* 목적: 회의 음성 데이터를 다양한 형식(PDF, XML, CSV)에서 수집한 후, 인공지능 모델 학습에 적합하도록 통일된 JSONL 구조로 변환하여 일관된 입력 포맷을 확보하기 위함.
* 수행 작업:
* 국회 회의록(PDF): PDF → TXT 변환 후, 발화 단위로 정제하여 timestamp, speaker, text 필드로 구조화
* AMI 대화록(XML): XML 파싱을 통해 발화 시점, 화자, 내용을 추출한 뒤 JSONL 포맷으로 변환
* ICSI 대화록(CSV): 정규표현식을 활용해 구조 불일치 데이터를 필터링한 뒤 JSONL로 변환
* 각 데이터셋에서 불필요한 기호(예: ◯, ※, “...”)는 제거하고 발화 단위 구문을 정제하여 구조화
* 사용 도구 및 라이브러리: Python, re (정규표현식), xml.etree.ElementTree, pandas
* 코드 예시 (Python):

# 파일 읽기 - BOM 처리

with open(input\_path, "r", encoding="utf-8-sig", newline='') as f:

file\_content = f.read()

# Windows \r\n을 \n으로 정규화

file\_content = file\_content.replace('\r\n', '\n').replace('\r', '\n')

# 혹시 남아있을 수 있는 BOM 문자 제거

if file\_content.startswith('\ufeff'):

file\_content = file\_content[1:]

# 원본 라인 수 계산 (빈 라인 제외)

original\_lines = [line.strip() for line in file\_content.split('\n') if line.strip()]

print(f" 원본 비어있지 않은 라인 수: {len(original\_lines)}"

# 발화 단위로 파싱

utterances = parse\_meeting\_transcript(file\_content)

print(f" 파싱된 발화 수: {len(utterances)}")

# 첫 번째 발화 확인

if utterances:

first = utterances[0]

print(f" 첫 번째 발화: [{first['timestamp']}] {first['speaker']}")

# JSONL 파일로 저장

with open(output\_path, "w", encoding="utf-8-sig", newline='') as f\_out:

for utterance in utterances:

json.dump(utterance, f\_out, ensure\_ascii=False)

f\_out.write("\n")

* JSONL 구조 예시:

{"timestamp": null, "timestamp\_order": "2-2", "speaker": "소위원장 김현", "text": "소위의 언론 공개는 여기까지 하겠습니다."}

{"timestamp": null, "timestamp\_order": "2-3", "speaker": "소위원장 김현", "text": "언론인들의 협조를 부탁드립니다."}

4.2 결측치 처리

* 결측치 존재 여부: 없음
* 결측 컬럼 및 비율:

| **컬럼명** | **결측률** | **처리 방법** |
| --- | --- | --- |
| speaker | 0% | 해당 행 삭제 |
| text | 0% | 해당 행 삭제 |

* 코드 예시 (Python):

if text.strip() == "":

log\_issue(log\_file\_path, session\_name, "빈 텍스트", ...)

continue

if not speaker or speaker.strip() == "":

log\_issue(log\_file\_path, session\_name, "빈 화자", ...)

continue

-> 코드 실행 시 로그 파일에 결측치 부분이 기록됨. 로그에 0건이 기록되어있으므로 결측치는 존재하지 않음.

4.3 이상치 처리

* 정의한 이상치 기준
* 발화 내용이 기호(@, ., 공백 등)만 포함된 경우
* 화자의 실제 발언이 아닌 시스템 메시지(예: "회의 개회", "산회")
* 처리 방식 및 영향:

| **항목** | **기준** | **처리 방식** | **제거 수** |
| --- | --- | --- | --- |
| Text (AMI/ICSI, 번역 후) | 기호 등 노이즈 | 해당 부분만 제거 | AMI: 6,356개  ICSI: 5,171개 |
| Text (국회데이터) | 발언 내용 외 | 해당 행 제거 | 손실률:  문자 - 3.74%  라인 - 5.7%  단어 - 2.38% |

4.4 학습용 데이터 분리

* 분리 기준 및 방법:  
  + 기준: 무작위 분할, 겹치기 없음, 전체 데이터에서 무작위로 12,400건을 추출
  + 비율: Train 64.5% / Validation 16.1% / Test 19.3%
* 분리 후 건수:

| 구분 | 데이터 수 |
| --- | --- |
| 학습 데이터 | 8,000건 |
| 검증 데이터 | 2,000건 |
| 테스트 데이터 | 2,400건 |

1. **전처리 결과 요약 및 평가**

* 전처리 후 전체 건수: 전체 233,569건 → 221,862건
* 품질 향상 지표:  
  + 이상치 제거: AMI - 6,356건, ICSI -> 5,171건 처리
  + 레이블 정리 및 불균형 개선
* 향후 활용 방안:회의 기반 LLM 모델 학습, Slack Agent 연동, 회의 흐름 분석 등 다양한 업무 자동화 모델에 활용

1. **변경 이력**

| 변경일 | 변경자 | 변경 내용 | 비고 |
| --- | --- | --- | --- |
| 2025-07-17 | 이준석 | 내용 작성 |  |
| 2025-07-18 | 이준석 | 내용 수정 |  |