

**SK네트웍스 Family AI과정 12기  
 데이터 수집 및 저장 수집 데이터 보고서**



| **산출물 단계** | 데이터 수집 및 저장 |
| --- | --- |
| **평가 산출물** | 수집 데이터 보고서 |
| **제출 일자** | 2025.07.11 |
| **깃허브 경로** | <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN12-FINAL-3TEAM> |
| **작성 팀원** | 이준석 |

**수집 데이터 개요**

**A. 국회 상임·특별위원회 회의록**

| **데이터명** | **수집 대상** | **수집 목적** | **사용 예정 기능** | **세션 수 /**  **발화 수** | **형식** | **출처/저작권** |
| --- | --- | --- | --- | --- | --- | --- |
| 국회 상임·특별위원회 회의록 | 국회 위원회 회의 발언 내용 | 회의 발언 기반 업무 분배 자동화 / 다자 대화 구조 파악 | 발언자 분리, 회의 흐름 추출, 작업(Task) 추출, 역할 분류 모델 테스트 | 52 / 49,003 | JSONL (정제) | 대한민국 국회 정보공개 자료 |

### B. AMI 다자 대화록 (한글 번역)

| **데이터명** | **수집 대상** | **수집 목적** | **사용 예정 기능** | **세션 수 /**  **발화 수** | **형식** | **출처/저작권** |
| --- | --- | --- | --- | --- | --- | --- |
| AMI 다자 대화록 | 사전에 구성된 소규모 회의 참가자 간 실시간 다자 대화 | 정형화된 회의 구조 기반으로 역할 분류, 요약, 발화 흐름 분석 등 LLM 실험용 레퍼런스 구축 | 발화자 역할 분류, 회의 요약, 다자 대화 이해/분석 모델 검증 | 154 /  89,773 | JSONL (정제) | The University of Edinburgh (공개 데이터셋) |

### C. ICSI 다자 대화록 (한글 번역)

| **데이터명** | **수집 대상** | **수집 목적** | **사용 예정 기능** | **세션 수 /**  **발화 수** | **형식** | **출처/저작권** |
| --- | --- | --- | --- | --- | --- | --- |
| ICSI 다자 대화록 | 연구실 기반 자연스러운 대화 (20개 팀, 실제 연구자 간 회의) | 자연어 회의에서의 잡음·중복·누락 대응 실험 및 발화 필터링, 회의 이해도 향상을 위한 전처리 알고리즘 실험 | 발화자 역할 분류, 회의 요약, 다자 대화 이해/분석 모델 검증 | 75 / 94,793 | JSONL (정제) | The University of Edinburgh (공개 데이터셋) |

**수집 방법 및 자동화 절차**

### A. 국회 상임·특별위원회 회의록

* **수집 방식**  
  + 국회 정보공개포털 수동 다운로드 후 문서 파일 업로드
* **수집 도구 또는 스크립트 설명**
  + **사용한 언어/라이브러리**: Python 기반 스크립트로 PDF -> TXT 변환 후 OPENAI API로 TXT 정리 및 TXT -> JSONL 변환
  + **자동화 여부 및 주기**:  
    - 수집: 수동 (정보공개자료 다운로드)
    - 전처리: 반자동 (스크립트로 OPENAI API 사용 후 결과물 출력)
  + **오류 발생 시 예외 처리 전략**:  
    - 발언 외의 내용 삭제
    - 정형 구조 위반(발언자 누락, 줄바꿈 오류 등)이 있을 경우 수동 검토

### B. AMI 다자 대화록

* **수집 방식**  
  + AMI Corpus 전체 수동 다운로드 후 원본 XML 기반 구조 복원
* **수집 도구 또는 스크립트 설명**
  + **사용한 언어/라이브러리**: Python 스크립트로 XML → CSV → JSONL 변환
  + **자동화 여부 및 주기**:  
    - 수집: 수동
    - 정제: 자동화 (모든 발화를 timestamp, timestamp\_order, speaker, text 4개 필드로 구조화)
  + **오류 발생 시 예외 처리 전략**  
    - UTF-8 BOM 인코딩 오류로 인한 첫 필드 누락이 있을 경우 코드 내 보정 처리

### C. ICSI 다자 대화록

* **수집 방식**  
  + ICSI Corpus 전체 수동 다운로드 후 CSV 기반 회의록 정제
* **수집 도구 또는 스크립트 설명**
  + **사용한 언어/라이브러리**: Python 스크립트로 XML → CSV → JSONL 변환
  + **자동화 여부 및 주기**:  
    - 수집: 수동
    - 정제: 자동화 (모든 발화를 timestamp, timestamp\_order, speaker, text 4개 필드로 구조화)
  + **오류 발생 시 예외 처리 전략**:  
    - UTF-8 BOM 인코딩 오류로 인한 첫 필드 누락이 있을 경우 코드 내 보정 처리

**데이터 설명 및 구성**

### A. 국회 상임·특별위원회 회의록

#### 데이터 정의 및 원본 문제점

#### 국회 상임·특별위원회 회의록은 대한민국 국회 정보공개포털에서 제공되는 공식 회의 자료로, 고위 공직자와 위원, 증인 등의 발언 내용을 포함한 공개 문서임.

#### 원본은 PDF 형식의 회의록 문서로 제공되며, 발언자의 구분이 명확하지 않거나, 발언 내용과 회의 절차 문구가 혼재되어 있어 다음과 같은 문제가 있음.

#### 발언자와 발언 내용의 명확한 경계가 없음: 줄바꿈과 띄어쓰기 없이 연속적으로 기록되어 있어 정규표현식 기반의 복원이 필수적임

* 줄바꿈 기준 미정형: 발언이 단락 없이 이어지거나 누락되는 경우가 있어 전처리 없이 모델 입력에 사용 불가능
* 화자 구분 없이 통합된 텍스트 구조: 대화형 구조로 학습하기 부적합

→ 이러한 구조적 문제를 해결하기 위해, 본 프로젝트에서는 모든 발화를 timestamp, speaker, text의 3개 필드로 구조화하여 JSONL 포맷으로 정제함.

**데이터를 일부 열어봤을 때 판단한 사용 근거**

예시)위원장 최민희 | "이상으로 산회를 선포하겠습니다."

→ 이처럼 각 발화가 시간, 화자, 텍스트로 명확히 분리되어 있어, 화자 구분, 발화 시점 인식, 내용 처리 등 모델 입력에 필요한 구조가 완비되어 있음.

또한 발언들이 공식 회의체의 구어체에 가까운 정제된 문장으로 구성되어 있어, 다음과 같은 실험에 적합:

* LLM 기반 요약
* 역할 분류 및 발화자 성격 파악
* 다자 회의 흐름 추적 및 의제 전이 분석

특히, timestamp가 포함되어 있어 시퀀스 기반 모델 학습, 시간 축 기준 발화 분석, 정책 논의 흐름 추적 등 다양한 실험에 활용 가능함.

**파일 및 필드 설명**

| **파일명** | **필드명** | **데이터 타입** | **설명** | **예시** |
| --- | --- | --- | --- | --- |
| 제22대국회 제426회(임시회) 제1차 과학기술정보방송통신위원회(정보통신방송법안심사소위원회) (2025.06.27.)\_split.jsonl | timestamp | string | 시간 | None |
| 제22대국회 제426회(임시회) 제1차 과학기술정보방송통신위원회(정보통신방송법안심사소위원회) (2025.06.27.)\_split.jsonl | timestamp\_order | string | 다문장의 순서 | 265-6 |
| 제22대국회 제426회(임시회) 제1차 과학기술정보방송통신위원회(정보통신방송법안심사소위원회) (2025.06.27.)\_split.jsonl | speaker | string | 화자 | 소위원장 김현 |
| 제22대국회 제426회(임시회) 제1차 과학기술정보방송통신위원회(정보통신방송법안심사소위원회) (2025.06.27.)\_split.jsonl | text | string | 회의 내용 | “이상으로 산회를 선포하겠습니다” |

**데이터 양**

* 전체 수집 데이터 건수: 52건
* 추출된 고품질 데이터 건수 : 52건

**저장 위치 및 포맷**

* 저장 경로 : [SKN12\_Ttalkkak/데이터/국회 회의록 (전처리 후) at main · Keindus8404/SKN12\_Ttalkkak](https://github.com/Keindus8404/SKN12_Ttalkkak/tree/main/%EB%8D%B0%EC%9D%B4%ED%84%B0/%EA%B5%AD%ED%9A%8C%20%ED%9A%8C%EC%9D%98%EB%A1%9D%20(%EC%A0%84%EC%B2%98%EB%A6%AC%20%ED%9B%84))
* 저장 포맷: JSONL (정제 전: PDF, TXT , 정제 후: JSONL)
* 개인정보 포함 여부: 고위 공직자 및 증인 등의 공식 발언만 포함. 주민번호·연락처 등 민감 개인정보는 포함되지 않음 (공개 회의록 기반)
* 출처 및 사용권: 대한민국 국회 정보공개포털
* 검토자 및 검토 일자: 2025-07-11

**데이터 품질 및 정합성 관리 방안**

* 발언 내용 외의 기타 내용 (표, 법령 내용, 화면 내용 기록 등)은 삭제
* 개회와 산회 기준으로 잡고 그 안의 발언 내용에 집중

### B. AMI 다자 대화록

**데이터 정의 및 원본 문제점**

* AMI 다자 대화록은 The University of Edinburgh에서 공개한 다자 회의 발화 데이터로, 총 154개 세션으로 구성되어 있음.
* 원본은 XML 기반으로 제공되며, 발화 내용이 시간·화자·텍스트가 혼합된 형태로 되어 있어 기계 판독이 불가능하며, 구분 기호가 일관되지 않아 정규표현식을 활용한 전처리가 필수적임.
* 또한 일부 발화는 시간 정보가 누락되거나 줄바꿈 위치가 일관되지 않아, 직접 학습 데이터로 사용하기엔 부적합한 상태였음.

→ 이러한 구조적 한계를 극복하기 위해, 본 프로젝트에서는 모든 발화를 timestamp, speaker, text의 세 필드로 완전하게 분리하여 JSONL 포맷으로 정제.

**데이터를 일부 열어봤을 때 판단한 사용 근거**

예시) [00:03] Speaker\_E | 'Kay .

* 이처럼 각 발화가 명확하게 분리되어 있어, 화자 구분, 발화 시점, 텍스트 처리 등 모든 요소가 모델 입력에 적합한 상태로 정리되어 있음
* 또한 발화가 비교적 짧고 구어체로 구성되어 있어, LLM 기반 요약, 역할 분류, 대화 흐름 분석 등 실제 회의 시나리오를 반영한 실험에 매우 유용함.
* 시간 정보가 포함되어 있어 시퀀스 모델 학습이나 시간 기반 주제 전이 분석에도 활용 가능함.

**파일 및 필드 설명**

| **파일명 또는 테이블명** | **필드명** | **데이터 타입** | **설명** | **예시** |
| --- | --- | --- | --- | --- |
| EN2001a.jsonl | timestamp | string | 시간 | 00:11 |
| EN2001a.jsonl | timestamp\_order | string | 다문장의 순서 | 5-1 |
| EN2001a.jsonl | speaker | string | 화자 | Speaker\_A |
| EN2001a.jsonl | text | string | 대화 내용 | "스티브의 사양에 대한 피드백을 보고 싶은 분 계신가요?" |

**데이터 양**

* 전체 수집 데이터 건수: 154건
* 추출된 고품질 데이터 건수 :154건

**저장 위치 및 포맷**

* 저장 경로: [**SKN12\_Ttalkkak/데이터/영어 회의록 (AMI Corpus, 전처리 후) at main · Keindus8404/SKN12\_Ttalkkak**](https://github.com/Keindus8404/SKN12_Ttalkkak/tree/main/%EB%8D%B0%EC%9D%B4%ED%84%B0/%EC%98%81%EC%96%B4%20%ED%9A%8C%EC%9D%98%EB%A1%9D%20(AMI%20Corpus%2C%20%EC%A0%84%EC%B2%98%EB%A6%AC%20%ED%9B%84))
* 저장 포맷: JSONL (정제 전: CSV, 정제 후: JSONL)
* 인코딩: UTF-8 with BOM
* 개인정보 포함 여부: 미포함
* 출처 및 사용권: The University of Edinburgh (공개 사용 허가 라이선스 포함)
* 검토자 및 검토 일자: 2025-07-10

**데이터 품질 및 정합성 관리 방안**

* 중복 제거 기준: 동일 발화의 반복적 저장/수집을 제거하여 데이터 품질 유지
* 정합성 검증 방법: 시간/화자/내용 필드 누락 여부 자동 점검
* 결측치나 Null 처리: 삭제 처리

### C. ICSI 다자 대화록

**데이터 정의 및 원본 문제점**

* ICSI 다자 대화록 역시 에든버러 대학에서 공개한 영어 회의 대화 데이터로, 총 75개 세션으로 구성.
* 원본은 CSV 형태이나, 발화 내용이 timestamp, speaker, text가 아닌 하나의 문자열로 혼합되어 있고, 줄마다 구조가 달라 자동 파싱이 불가능.
* 특히 UTF-8-BOM 인코딩 문제로 인해 첫 번째 필드가 누락되거나 잘못 해석되는 경우가 있었고, "text": ""와 같은 빈 필드가 산발적으로 존재.

→ 이러한 구조적 한계를 극복하기 위해, 본 프로젝트에서는 모든 발화를 timestamp, speaker, text의 세 필드로 완전하게 분리하여 JSONL 포맷으로 정제.

**데이터를 일부 열어봤을 때 판단한 사용 근거**

예시) [05:28] fe016 | waveforms

* 이처럼 각 발화가 명확하게 분리되어 있어, 화자 구분, 발화 시점, 텍스트 처리 등 모든 요소가 모델 입력에 적합한 상태로 정리되어 있음.
* 또한 발화가 비교적 짧고 구어체로 구성되어 있어, LLM 기반 요약, 역할 분류, 대화 흐름 분석 등 실제 회의 시나리오를 반영한 실험에 매우 유용함.
* 시간 정보가 포함되어 있어 시퀀스 모델 학습이나 시간 기반 주제 전이 분석에도 활용 가능함.

→ 다만, 일부 발화에서 발생하는 빈 필드에 대해서는 제거 필요

**파일 및 필드 설명**

| **파일명 또는 테이블명** | **필드명** | **데이터 타입** | **설명** | **예시** |
| --- | --- | --- | --- | --- |
| Bdb001\_processed.jsonl | timestamp | string | 시간 | 05:28 |
| Bdb001\_processed.jsonl | timestamp\_order | string | 시간 | 29-1 |
| Bdb001\_processed.jsonl | speaker | string | 화자 | Speaker\_B |
| Bdb001\_processed.jsonl | text | string | 대화 내용 | "동일하게 유지되는 파형, 변경될 수도 있고 변경되지 않을 수도 있는 전사본, 그리고" |

**데이터 양**

* 전체 수집 데이터 건수: 75건
* 추출된 고품질 데이터 건수 :75건

**저장 위치 및 포맷**

* 저장 경로: [SKN12\_Ttalkkak/데이터/영어 회의록 (ICSI Corpus, 전처리 후) at main · Keindus8404/SKN12\_Ttalkkak](https://github.com/Keindus8404/SKN12_Ttalkkak/tree/main/%EB%8D%B0%EC%9D%B4%ED%84%B0/%EC%98%81%EC%96%B4%20%ED%9A%8C%EC%9D%98%EB%A1%9D%20(ICSI%20Corpus%2C%20%EC%A0%84%EC%B2%98%EB%A6%AC%20%ED%9B%84))
* 저장 포맷: JSONL (정제 전: CSV, 정제 후: JSONL)
* 인코딩: UTF-8 with BOM
* 개인정보 포함 여부: 미포함
* 출처 및 사용권: The University of Edinburgh (공개 사용 허가 라이선스 포함)
* 검토자 및 검토 일자: 2025-07-10

**데이터 품질 및 정합성 관리 방안**

* 중복 제거 기준: 동일 발화의 반복적 저장/수집을 제거하여 데이터 품질 유지
* 정합성 검증 방법: 시간/화자/내용 필드 누락 여부 자동 점검
* 결측치나 Null 처리: 삭제 처리

**분석결과**

### A. 국회 상임·특별위원회 회의록

* **총 발화 수:** 약 49,000건
* **평균 발화 길이:** 약 20~40단어 (공식 회의 발언 특성상 비교적 장문)
* **발화 유형:**
* 정책 질의 및 문제 제기
* 기관 또는 부처 측의 응답/설명
* 제안/요구 및 발언 정리
* **주요 분석 결과:**
* 발화 흐름이 비교적 논리적으로 구성
* 화자 정보가 정확하게 표기
* 공식 문서체 문장 구조
* 회의 기록 특성상 명료한 목적성과 논리 흐름이 있어 LLM fine-tuning 데이터로 적합

### B. AMI 다자 대화록

* **총 발화 수:** 약 89,000건
* **평균 발화 길이:** 약 7~10단어
* **발화 유형:**
* 단순 응답
* 정보 전달 및 설명 발화
* 질문/반론
* **주요 분석 결과:**
* 시간 순서 정렬이 완벽하게 보존
* 화자별 발화 패턴 분석 시, 회의 내 역할 차이가 명확히 드러남
* 구어체 특성이 자연스러워 LLM fine-tuning 데이터로 적합

### C. ICSI 다자 대화록

* 총 발화 수: 약 94,000건
* **평균 발화 길이:** 약 1~2단어
* **발화 유형:**
* 짧은 추임새 및 동의
* 정보 전달 및 설명 발화
* 질문/반론
* **주요 분석 결과:**
* 일부 화자/내용 누락 발화 존재 → 전체 구조에는 큰 영향 없음
* 전처리 후에도 구조상 누락 필드가 산발적으로 분포

**결측치 및 보간 처리 방안**

### A. 국회 상임·특별위원회 회의록

* **결측치 발생 여부:** 드물게 존재 (speaker 또는 발화 내용 일부 누락)
* **평균 발화 길이:** 약 20~40단어 (공식 회의체 특성상 비교적 장문)
* **보간 처리:**

**-** Speaker나 text 누락 시 제거

### B. AMI 다자 대화록

* **결측치 발생 여부:** 없음 (speaker, timestamp, text 모두 정제 완료)
* **평균 발화 길이:** 약 7~10단어
* **보간 처리: 필요 없음**
* 원본의 비정형 포맷은 정규표현식 기반 스크립트로 완전 구조화
* 의미 단절 없이 모든 발화가 정렬

### C. ICSI 다자 대화록

* **결측치 유형:**
* "text": "" (비어 있는 발화)
* **평균 발화 길이:** 약 3-5단어
* **보간 전략:**

**-** "@", ".", "" 만 포함된 발화는 **삭제**

**변경 이력 및 보완 내역**

| **변경일** | **변경자** | **변경 내용** | **비고** |
| --- | --- | --- | --- |
| 2025-07-04 | 이준석 | 문서 생성 및 수정 |  |
| 2025-07-08 | 이준석 | 문서 수정 및 정리 |  |
| 2025-07-10 | 이준석 | 문서 수정 |  |
| 2025-07-11 | 이준석 | 문서 수정 |  |