


인공지능 데이터 구축·활용 가이드라인

- 감성 대화 맞춤형 AI 데이터 구축 분야 -

인공지능 데이터 구축	AI 데이터 구축	
가이드라인 작성	미디어젠	송민규
가이드라인 버전	Ver 1.4 (2020-12-23)	

목 차

1. 데이터 구축 개요	1
2. 임무 정의	2
2.1 임무 정의	2
2.2 데이터 구축 유의사항	2
3. 획득·정제	3
3.1 원시데이터 선정	3
3.2 획득·정제 절차	5
3.3 획득·정제 기준	8
3.4 획득·정제 조직	9
3.5 획득·정제 도구	10
4. 어노테이션/라벨링	11
4.1 어노테이션/라벨링 절차	11
4.2 어노테이션/라벨링 기준	11
4.3 어노테이션/라벨링 조직	13
4.4 어노테이션/라벨링 도구	14
5. 검수	14
5.1 검수 절차	14
5.2 검수 기준	15
5.3 검수 조직	17
5.4 검수 도구	17
5.5 기타 품질관리 활동	18
6. 활용	18
6.1 활용 모델	18
6.1.1 모델 학습	18
6.1.2 서비스 활용 시나리오	18
6.2 접근	18
6.3 유지보수	18
붙임1 인공지능 데이터 명세서 양식	27
붙임2 인공지능 데이터 명세서 작성 예시	27

1. 데이터 구축 개요

1.1 데이터 구축 개요

본 프로젝트는 AI 기반 감성 챗봇용 세대별 감성대화 텍스트 데이터의 구축이 주목적이며, 궁극적으로는 딥러닝 기반 언어 이해(NLU) 엔진인 ALBERT 모델링을 위한 데이터를 수집하는 것이다.

본 구축 데이터를 통해 한국인의 감성대화 엔진 또는 챗봇을 개발하려는 기업, 연구기관, 연구자 등에게 기술 개발의 리소스를 제공하며, 나아가 독거노인 돌봄 서비스 등을 제공하고 있는 기관 역시 궁극적인 수혜자가 될 수 있다.

본 데이터 구축은 시나리오 설계, 데이터 수집, 데이터 태깅, 데이터 검수의 과정으로 진행되며, 중간 결과물을 통한 NLU 엔진 모델링을 수행하여 모델 성능을 확인한다. 이어 대규모 데이터 구축을 수행한 뒤 공인 인증 기관의 품질 검증 과정을 거쳐 구축 데이터 및 NLU 모델을 공개한다.

단계		수행 주체	내용
구축	어노테이션	클라우드 용역 3사	어노테이션과 라벨링은 동일한 개념이며, 페르소나 및 감정 상황의 조합이 설계된 후 코퍼스 수집에 착수
	라벨링		
	코퍼스 수집	클라우드 용역 3사	미리 설계되고 어노테이션된 케이스에 대한 데이터 수집
	1차 검수	클라우드 용역 3사	기초적인 문법 검사
	2차 검수	주관기관	설계 상황에 대한 내용 검수
	태깅	클라우드 용역 3사	별도의 도구를 사용하여 검수
	3차 검수	주관기관	모델링을 위한 적정성 검토
	모델링	주관기관	인공지능 학습 과정
활용		해커톤 대회	불특정 다수를 위한 가이드 제공

2. 임무 정의

2.1 임무 정의

감정인식을 통한 감성 AI 대화 기술은 데이터의 확보가 매우 어려운 분야에 속한다. 기존의 빅데이터가 크롤링 등 대규모 데이터의 자동 수집 및 분류에 의존하고 있다면, 감정인식을 위한 데이터는 구어 중심의 직접 제작 데이터가 필요하기 때문이다. 인터넷이나 기타 출판물에서 얻을 수 있는 언어 데이터는 대부분 문어체 텍스트이며, 대화체는 극히 일부만 수집이 가능하다. 게다가 감정의 상태를 포함하고 있는 텍스트는 더욱 희소하기 때문에 이에 대한 딥러닝 모델링이 진행되기가 어려웠다.

본 프로젝트에서 수집하는 데이터는 감정의 상태에 따른 구어 텍스트를 직접 수집하여 저작권 문제가 없는 다량의 코퍼스이며, 정제와 검수를 거쳐 고품질의 텍스트 수집을 수행한다.

딥러닝 기반의 ALBERT 알고리즘을 이용한 모델링이 가능한 수준으로 대략 27만 문장의 데이터를 취득하는 것이 목적이다. ALBERT 알고리즘을 활용하면, 형태소 단위의 세부 태깅을 하지 않고 문장 단위의 정합성만 검수하면 데이터 모델링에 큰 문제가 없다는 점이 장점이며, 문장에서 의미와 의도를 추출하는 확률이 기존 통계 모델링 기법(CRF+ 등)에 비해 월등히 높은 성능을 보인다는 점이 특징이다.

그러나 이러한 알고리즘의 발전에도 불구하고, 데이터의 부족으로 인해 감정인식 및 감성 대화 분야는 다른 일반 대화에 비해 인공지능으로 해결하기 어려운 분야였기 때문에, 본 프로젝트를 통해 이에 대한 데이터 부족 문제를 일부나마 해결할 수 있을 것으로 기대된다.



2.2 데이터 구축 유의사항

본 프로젝트에서는 데이터를 직접 제작하여 취득하므로 법적인 문제 발생 소지가 매우 적다. 그렇다 하더라도 대본 작성에 참여하는 크라우드 워커의 이름 등 신상 노출 가능성이 있으므로, 개인 정보가 포함되지 않도록 주의한다.

크라우드 워커의 이름을 이니셜 처리 하는 등 비식별화를 통해 개인 정보 유출을 방지한다.

클라우드 워커 모집 시 정보 공개 동의서 작성 및 AI Hub를 통한 데이터 공개 여부를 고지한다.

3. 획득·정제

3.1 원시데이터 선정

원시데이터는 텍스트 형식으로 피험자에게서 바로 수집을 하기 때문에, 별도의 변환 과정은 필요하지 않다. 원시데이터는 피험자로부터 직접 수집을 원칙으로 하며, 해당 데이터는 클라우드 소싱 플랫폼을 활용하여 동시에 수집을 수행한다.

감정의 표현은 연령, 성별, 사회적 지위, 질병 이력, 학교 등 다양한 원인에 의해 달라질 수 있으므로, 다양한 상황을 제시하여 데이터 수집이 가능하도록 설계한다.

원시데이터 규모는 총 27만 문장의 텍스트 코퍼스이며, 상황 및 감정 상태를 태깅하여 의미 분류를 할 수 있도록 가공한다.

60가지 감정 분류						
기본	분노	슬픔	불안	상처	당황	기쁨
1	툼툼대는	실망한	두려운	질투하는	고립된	감사하는
2	좌절한	비통한	스트레스 받는	배신당한	남의 시선을 의식하는	사랑하는
3	짜증나는	후회되는	취약한	고립된	외로운	편안한
4	방어적인	우울한	혼란스러운	충격 받은	열등감	만족스러운
5	악의적인	마비된	당혹스러운	불우한	죄책감	흥분되는
6	안달하는	염세적인	회의적인	희생된	부끄러운	느긋한
7	구역질 나는	눈물이 나는	걱정스러운	억울한	혐오스러운	안도하는
8	노여워하는	낙담한	조심스러운	괴로워하는	한심한	신이 난
9	성가신	환멸을 느끼는	초조한	버려진	혼란스러운	자신하는

감정의 상태는 위의 그림과 같이 총 60가지로 구분하며, 해당 감정 상태에 대한 개개인의 페르소나를 생성하여 대화 코퍼스를 수집한다.

페르소나는 연령과 성별로 크게 구분하며, 사람의 대화에 대한 챗봇 시스템의 응답으로 구성된다.

구분	항목	상세
페르소나	연령 (A)	청소년
		청년
		중년
		노년
	성별 (G)	남성
		여성

상황을 만드는 요인은 상황에 대한 세부 항목, 질병, 감정 상태 등에 따라 분류된다.

구분	항목	상세
상황	상황 (S)	상황 세부 항목
	질병 (D)	질병 세부 항목
	감정 (E)	감정 세부 항목

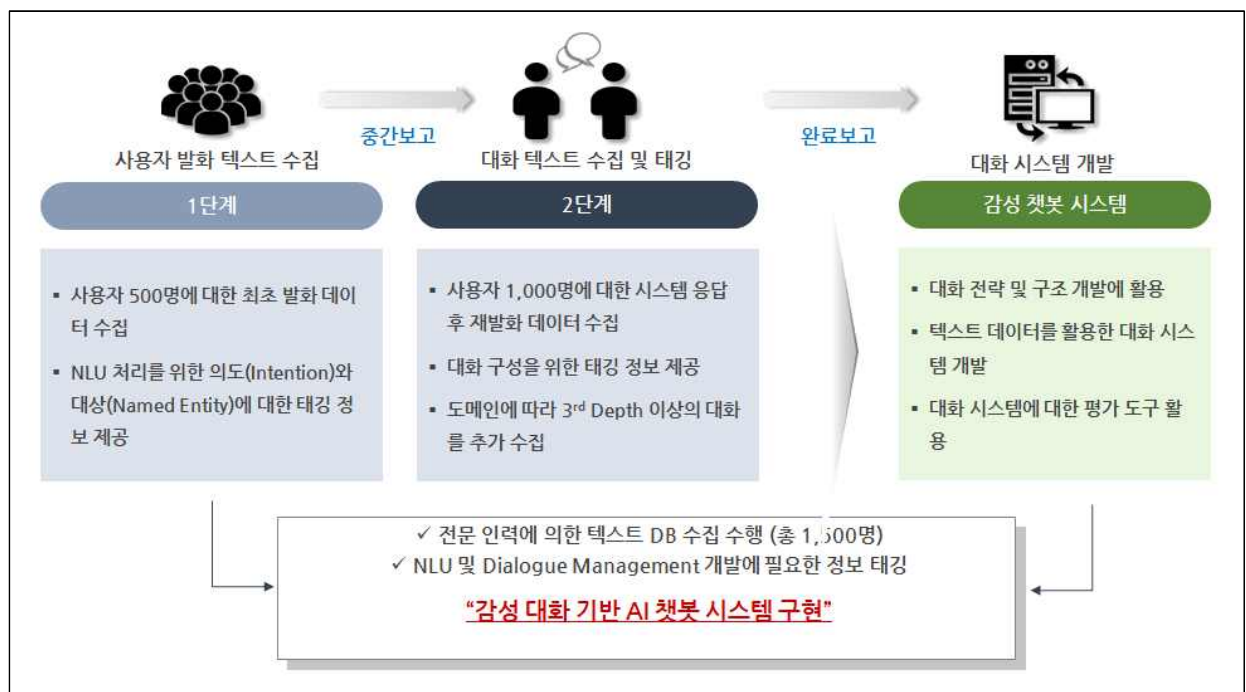
대화의 구조는 위와 같은 3턴의 대화(사람과 시스템의 총 6 발화)로 이루어진다. 필요에 따라서는 4턴의 대화(총 8발화)까지 확장하여 수집한다.

구분	항목
대화 턴	1 : 사람 대화 (감정 상태)
	2 : 시스템 응답 (응답 호응)
	3. 사람 대화 2 (진전된 대화)
	4. 시스템 응답 2 (응답 호응)
	5. 사람 대화 3 (진전된 대화)
	6. 시스템 응답 3 (마무리 대화)

3.2 획득·정제 절차

원시데이터는 클라우드 소싱 방식을 통해 수집하며, 피험자 개개인이 주어진 상황에 맞는 데이터를 직접 입력한다. 개인별 주어진 상황에 대한 감정 상태의 표현을 기재하므로, 수집 단계에서 특별한 공통 관리가 필요하지는 않다.

다만, 주어진 상황에 맞지 않는 부분은 별도 검수를 통해 걸러내거나 수정한다.



대화 코퍼스는 3rd Depth를 기준으로 수집하며, 사람의 질문과 챗봇 응답의 패턴을 유지한다. 수집된 데이터는 정제 담당자, 상용 근로자, 전문가 등의 검수를 받으며, 페르소나와 감정 상태에 알맞는 대화를 수집한다.

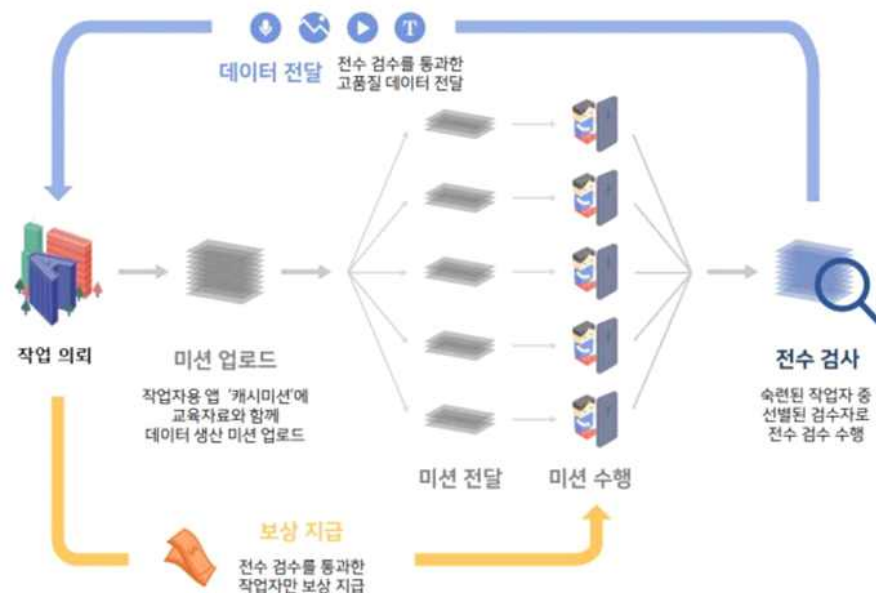
수집된 데이터는 JSON 타입으로 저장되며, 이후 다양한 서비스에 활용 가능하도록 한다.

```

1 {
2   "profile":
3     "profile-id": "Pro_13409",
4     "persona": {
5       "persona-id": "A04_G02_C01",
6       "human": ["A04", "G02"],
7       "computer": ["C01"]},
8     "emotion": {
9       "emotion-id": "S04_D02_E38",
10      "type": "E38",
11      "situation": ["S04", "D02"]
12    }
13  },
14  "talk": {
15    "id": {
16      "profile-id": "Pro_13409",
17      "talk-id": "Pro_13409_00009"
18    },
19    "content": {
20      "HS01": "이번에 새로 옮긴 요양원에서 생활하고 있는데 새로운 요양원인지라 다소 긴장되고 조심스러워.",
21      "SS01": "새로운 요양원에서 생활하시는 데 긴장감을 느끼고 계시군요.",
22      "HS02": "예전 요양원과 는 생활방식도 다르고 사람들도 달라서 혹시라도 실수하게 될까 봐 조심스러워.",
23      "SS02": "어떻게 하면 지금 상황에서 조금이라도 덜 스트레스 받을 수 있을까요?",
24      "HS03": "얼른 요양원 안에서 친구를 만들어야겠어. 친구를 만들면 새로운 요양원에 빨리 적응할 수 있을 것 같아.",
25      "SS03": "마음이 맞는 친구 분을 만드셔서 빠르게 새로운 요양원에 적응하시길 바라요."
26    }
27  }
28 },

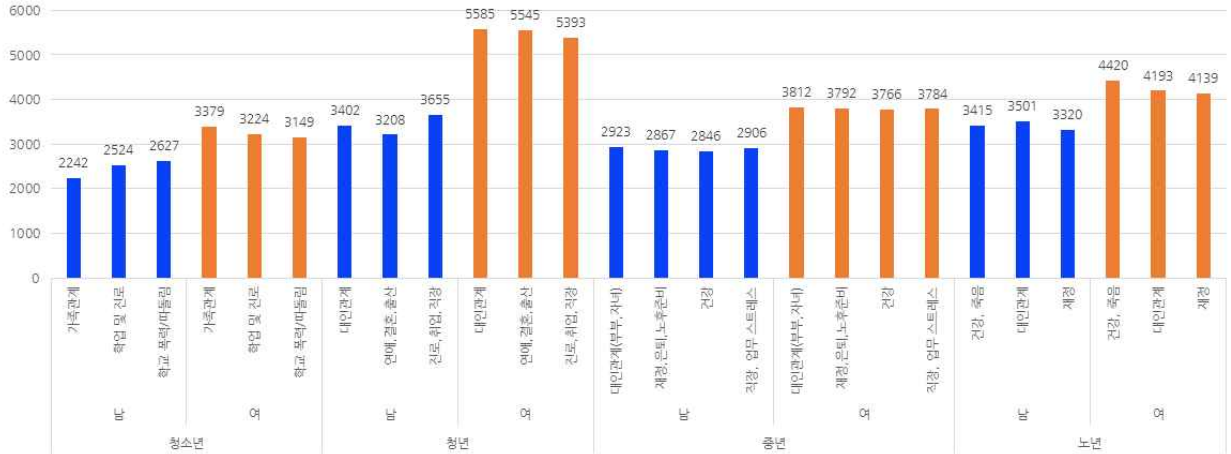
```

코퍼스 데이터 수집은 클라우드 소셜 플랫폼을 통해 전문 업체를 통해 수집되며, 각각의 개인으로부터 수집된 데이터를 취합하고 검사하여 원시데이터를 축적한다.



연령, 성별, 상황 별 데이터 수집량과 분포는 다음과 같다.

<데이터 분포>



감정 상태별 데이터 수집량과 분포는 다음과 같다.

<감정별 데이터 분포>



3.3 획득·정제 기준

수집된 데이터는 다양한 후처리를 수행하여 정제한다. 문자의 변환, 특수기호 처리, 오타자 수정 등 다양한 정제 과정을 거치는데, 본 프로젝트의 목적이 텍스트 코퍼스를 수집하는 것이기 때문에 주로 일반적인 텍스트 정제 방법론이 활용된다.

<수집 데이터 정제 기준>

(1) 일반적인 특수기호(※, ▲ 등)는 일반 문자로 변환하거나 삭제한다.

※ "España"라는 국호의 이름과 영어의 동의어 "Spain"이나 "Spanish"에 대해서는 논란의 여지가 있다.

→ Espana라는 국호의 이름과 영어의 동의어 Spain이나 Spanish에 대해서는 논란의 여지가 있다.

(2) 마침표, 느낌표, 물음표는 그대로 남겨둔다. 기호를 빼면 문장이 어색해 지는 경우는 특수기호(한자, 외국어, 단위, 퍼센트, 통화류 등)를 그대로 남긴다.

→ 闇の中の魑魅魍魎를 오랜시간 함께해왔던 신도 가네토의 각본으로 찍었고 이것이 칸 영화제 경쟁부문에 진출했다.

(3) 기호와 기호 사이는 삭제한다.

또한 그는 “정 씨(정성일)는 적절치 못한 처신으로 외가는 물론 정 전 총리 측과도 완전 결별 하다시피 했다.

→ 또한 그는 정 씨는 적절치 못한 처신으로 외가는 물론 정 전 총리 측과도 완전 결별하다시피 했다.

(4) 맞춤법에 맞지 않는 문장은 수정한다.

어제 내가 산 옷이 마음에 안 들었어.

→ 어제 내가 산 옷이 마음에 안 들었어.

(5) 문장 길이는 최소 2어절에서 최대 17어절까지 허용한다.

(6) 한 발화에서 두 문장까지 허용한다.

(7) 문장 부호는 한 발화에 3개 이상 사용하지 않는다.

(8) 기계가 사용자를 지칭할 때는 “OO님”으로 지칭한다. 다만 맥락에 따라 필요할 경우 “사용자님”이라는 표현도 가능하다.

(9) 사용자는 반말로 작성하고, 시스템 응답은 존댓말로 작성한다.

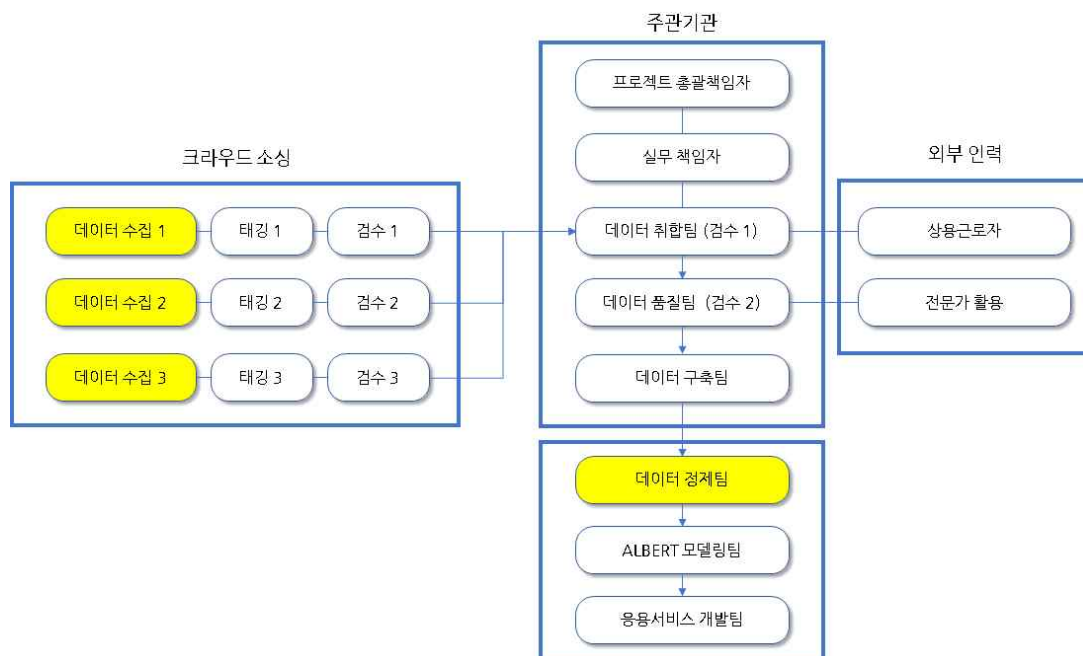
(10) 숫자 및 영어는 한글로 표기한다.

(11) 문장은 자연스러운 구어체로 작성한다.

본 프로젝트의 코퍼스 데이터는 ALBERT 알고리즘의 모델링을 목표로 하고 있기 때문에, 문장 단위의 정제가 진행되며, 의미의 분별이 분명하도록 띄어쓰기, 복합어 처리, 숫자 처리, 외국어 처리 등의 작업을 수행한다. 일괄처리가 필요한 경우 아래의 조건에 따라 처리한다.

항 목	내 용
띄어쓰기	<ul style="list-style-type: none"> 신문이나 언론사의 자료를 통해 지면 편집과 일반적인 관례를 따라서 띄어쓰기 오류 발생 <ul style="list-style-type: none"> 관례('에베레스트산' (x)-> '에베레스트 산'(o)) // 지면편집('김씨'(x) -> '김 씨'(o))
복합어 처리	<ul style="list-style-type: none"> 원시 데이터에 띄어 쓰지 않은 형태로 나타난 복합어 중 띄어 쓸 수 있는 곳에 "-"를 삽입 <ul style="list-style-type: none"> 복합어가 사전에 등재된 경우는 분리하지 않음 접두어 또는 접미어가 붙어 형성 사이 시옷이 들어 있음 '1 음절짜리' 단어로 구성 및 띄어쓰기를 하는 경우 중의성이 증가. 한 단어로 처리 (타수, 결승골, 골세례 등) 수 부류사 '개' + "1 음절짜리 명사" 회사명, 브랜드명과 같은 고유명사는 원문 유지
숫자 처리	<ul style="list-style-type: none"> 수를 적을 때 '만' 단위로 띄어쓰기 적용 ('만, 억, 조' 및 '경, 해, 자') <ul style="list-style-type: none"> 한글로 표시된 단위 등은 변환 범위에 포함 하지 않음 : 5분 -> {{오}}/{{{5}}}분 한자어, 영문, 기호로 표시된 단위 등은 따로 변환 : 3m -> {{삼}}/{{{3}}}{{미터}}/{{{m}}} 고유명사나 연어로 분류 될 수 있는 것은 동일 변환 범위에 포함 : 샤넬 No.5 -> 샤넬 {{넘버 파이브}}/{{{No.5}}} 숫자의 읽는 방법은 번독식과 봉독식으로 표현 : 20개-> {{스무, 이십}}/{{{20}}}개
외국어 처리	<ul style="list-style-type: none"> 한글로 변화하되 기준은 국립국어 연구원의 '외래어 한글 표기법'으로 표기 <ul style="list-style-type: none"> '한글 음사' 다음 괄호 안에 적은 영문자 :빅맥(big-mac), 텡(ting) 한국어 다음 의미의 명확성을 위해 적은 영문자 : 생명 윤리(bioethics) 2 단어 이상의 영문 고유 명사 영어 구문 이상의 단위 이메일 주소, ID, 웹 URL, 컴퓨터 경로명, 파일명 등 결합된 고유명사는 전체를 한 단위로 처리

3.4 획득·정제 조직



3.5 획득·정제 도구

원시데이터를 획득하는데 사용하는 도구는 직접 제작하여 사용한다. 기본 상황에서 연령, 성별, 상황키워드, 신체 질환 등의 페르소나 정보가 주어지고, 이에 매핑되는 감정 상태와 상세 감정 상태를 활용하여 상세 상황에 대한 질문 및 응답을 작성한다.

이때 기본 상황이 수집자의 페르소나이며, 이 페르소나가 그대로 데이터 어노테이션의 정보가 된다. 또한 기본 감정 및 상세 감정 상태 정보 또한 어노테이션 정보로 곧바로 활용된다.

The screenshot shows a web browser window with the URL `corpus.adsound.com`. The page title is "(주)애드사운드 감정 대화 텍스트 수집". The main content area has a header instruction: "제시된 기본 상황의 입장에서 상세 상황 조건에 맞춰서 대화를 작성해주세요." and a red button labeled "작성 가이드 보기 ##클릭##".

Below the instruction, there are two sections for inputting basic and detailed situations:

- 기본 상황 (Basic Situation):** A grid of input fields for "연령대" (Age Group) with the value "청소년(10대)", "성별" (Gender) with "남성", "상황키워드" (Situation Keyword) with "가족관계", "신체질환" (Physical Condition), "감정" (Emotion) with "슬픔", and "상세감정" (Detailed Emotion) with "낙담함".
- 상세 상황 (1 / 12) (Detailed Situation):** A yellow highlighted text box containing the sentence "형이 원하는 대학에 떨어졌다."

Below these sections, there are three user input areas, each with a "대화 작성해 주세요." (Please write the dialogue) prompt:

- 사용자1 (기본/상세 상황에 근거하여 대화 시작하기):** Prompted by a yellow speech bubble icon.
- 봇 1 (공감 혹은 초침 맞추기):** Prompted by a blue speaker icon.
- 사용자2:** Prompted by a yellow speech bubble icon.
- 봇 2 (공감 혹은 행동유발 질문):** Prompted by a blue speaker icon.
- 사용자3:** Prompted by a yellow speech bubble icon.

4. 어노테이션/라벨링

4.1 어노테이션/라벨링 절차

텍스트를 수집하면서 곧바로 어노테이션이 진행되는 형식이므로, 아래의 그림과 같이 상세 상황에 대해 순서대로 대화의 질의와 응답을 작성하는 절차로 데이터가 수집된다.

구어 데이터를 매우 효율적으로 수집할 수 있는 구조로 데이터 구축이 진행된다.

The screenshot shows a web browser window with the URL `corpus.adasound.com`. The page title is "(주)에드사운드 감정 대화 텍스트 수집". Below the title, there is a red button labeled "작성 가이드 보기 ##클릭##". The main content area is divided into several sections:

- 제시된 기본 상황의 입장에서 상세 상황 조건에 맞춰서 대화를 작성해주세요.**
- 기본 상황**: A table with fields for "연령대" (Age Group), "성별" (Gender), "상황키워드" (Situation Keyword), and "신체질환" (Physical Condition). The values are: 연령대: 청소년(10대), 성별: 남성, 상황키워드: 가족관계, 신체질환: (empty).
- 감정**: A table with fields for "감정" (Emotion) and "상세감정" (Detailed Emotion). The values are: 감정: 슬픔, 상세감정: 낙담함.
- 상세 상황 (1 / 12)**: A yellow bar with the text "형이 원하는 대학에 떨어졌다."
- 사용자1 (기본/상세 상황에 근거하여 대화 시작하기)**: A text input field with the placeholder "대화를 작성해 주세요."
- 봇 1 (공감 혹은 초침 맞추기)**: A text input field with the placeholder "대화를 작성해 주세요."
- 사용자2**: A text input field with the placeholder "대화를 작성해 주세요."
- 봇 2 (공감 혹은 행동유발 질문)**: A text input field with the placeholder "대화를 작성해 주세요."
- 사용자3**: A text input field with the placeholder "대화를 작성해 주세요."

4.2 어노테이션/라벨링 기준

사용자 프로필과 감정 상태에 맞게 수집된 대화 데이터는 각각의 페르소나에 해당되는 ID와 수집 텍스트가 쌍으로 저장되어야 한다.

미리 사전에 사용자 페르소나와 감정 상태가 정의되어 있기 때문에, 이러한 페르소나와 감정 태그는 자동으로 부여되게 되므로 별도의 작업이 없이도 자연스러운 태깅이 수행된다.

이러한 태깅은 데이터 수집 툴을 통해 자동으로 수행되므로 별도의 태깅 작업이 없어도 무방하다.

Persona_Emotion ID는 아래와 같은 규칙으로 지정된다.

연령(A : Age), 성별(G : Gender), 시스템응답(C : Computer)의 기본 ID가 배정된다.

구분	항목	상세	ID
페르소나	연령 (A)	청소년	A01
		청년	A02
		중년	A03
		노년	A04
	성별 (G)	남성	G01
		여성	G02
	시스템 응답 (C)	응답	C01

대화 시나리오의 경우 상황(Situation : S), 질병(Disease : D), 감정(Emotion : E)의 ID가 배정된다.

구분	항목	상세	ID
대화 시나리오 (감정 상태)	상황 (S)	상황 세부 항목	S01~S13
	질병 (D)	질병 세부 항목	D01~D02
	감정 (E)	감정 세부 항목	E10~E69

대화 턴에 대해서는 사람 대화(Human Speech : HS), 시스템 응답(System Speech), Turn (01~06) 값으로 ID가 지정된다.

구분	항목	상세	ID
대화 턴	사람 대화	Human Speech	HS
	시스템 응답	System Speech	SS
	대화 턴	Turn	01~06

프로파일 ID 할당 샘플

페르소나	프로파일	감정	연령	성별	상황	질병	감정
A01-G01-S 01-D01-E01	Pro_01	A01-G01-S 01-D01-E01	A01	G01	S01	D01	E01

JSON 타입 변환 테이블

profile	persona-id		"Pro_03802"
	persona	persona-id	"A02_G01_C01"
		human	["A02", "G01"]
		computer	["C01"]
	emotion	emotion-id	"S06_D02_E31"
		type	"E31"
		situation	["S06", "D02"]
talk	id	profile-id	"Pro_03802"
		talk-id	"Pro_03802_00028"
	content	HS01	"이번 프로젝트에서 내가 발표 실수를 해서 우리 팀이 감점을 받아서 너무 미안해."
		SS01	"실수하시다니 정말 죄송한 마음이 크겠어요."
		HS02	"내 능력이 부족한 거 같은데 그만 다녀야 하려나 봐."
		SS02	"능력을 올리려면 어떤 방법이 있을까요?"
		HS03	"퇴근 후 여가에 회사 일을 더 열심히 해서 피해가 가지 않도록 해야겠어."
		SS03	"꼭 좋은 결과 있길 바라요."

JSON 타입 샘플

```
{
  "profile": {
    "persona-id": "Pro_03802",
    "persona": {
      "persona-id": "A02_G01_C01",
      "human": [
        "A02",
        "G01"
      ],
      "computer": [
        "C01"
      ]
    },
    "emotion": {
      "emotion-id": "S06_D02_E31",
      "type": "E31",
      "situation": [
        "S06",
        "D02"
      ]
    }
  },
  "talk": {
    "id": {
      "profile-id": "Pro_03802",
      "talk-id": "Pro_03802_00028"
    },
    "content": {
      "HS01": "이번 프로젝트에서 내가 발표 실수를 해서 우리 팀이 감점을 받아서 너무 미안해.",
      "SS01": "실수하시다니 정말 죄송한 마음이 크겠어요.",
      "HS02": "내 능력이 부족한 거 같은데 그만 다녀야 하려나 봐.",
      "SS02": "능력을 올리려면 어떤 방법이 있을까요?",
      "HS03": "퇴근 후 여가에 회사 일을 더 열심히 해서 피해가 가지 않도록 해야겠어.",
      "SS03": "꼭 좋은 결과 있길 바라요."
    }
  }
}
```


페르소나 Index ID

구분(대)	구분(중)	구분(소)	ID
Persona	연령(Age)	청소년	A01
		청년	A02
		중년	A03
		노년	A04
	성별(Gender)	남성	G01
		여성	G02
	컴퓨터	기본	C01

상황 Index ID

상황(Situation)	가족관계	S01
	학업 및 진로	S02
	학교폭력/따돌림	S03
	대인관계	S04
	연애, 결혼, 출산	S05
	진로, 취업, 직장	S06
	대인관계(부부, 자녀)	S07
	재정, 은퇴, 노후준비	S08
	건강	S09
	직장, 업무 스트레스	S10
	건강, 죽음	S11
	대인관계(노년)	S12
	재정	S13

질병 Index ID

질병(Disease)	만성질환 유	D01
	만성질환 무	D02

감정 Index ID

감정(Emotion)	분노	E10
	툼툼대는	E11
	좌절한	E12
	짜증내는	E13
	방어적인	E14
	악의적인	E15
	안달하는	E16
	구역질 나는	E17
	노여워하는	E18
	성가신	E19
	슬픔	E20
	실망한	E21
	비통한	E22
	후회되는	E23
	우울한	E24
	마비된	E25
	염세적인	E26
	눈물이 나는	E27
	낙담한	E28
	한열을 느끼는	E29
	불안	E30
	두려운	E31
	스트레스 받는	E32
	취약한	E33
	혼란스러운	E34
	당혹스러운	E35
	회의적인	E36
	걱정스러운	E37
	조심스러운	E38
	초조한	E39
	상처	E40
	질투하는	E41
	배신당한	E42
	고립된	E43
	충격 받은	E44
	가난한, 불우한	E45
	희생된	E46
	억울한	E47
	괴로워하는	E48
	버려진	E49
	당황	E50
	고립된(당황한)	E51
	남의 시선을 의식하는	E52
	외로운	E53
	열등감	E54
	죄책감의	E55
	부끄러운	E56
	혐오스러운	E57
	한심한	E58
	혼란스러운(당황한)	E59
	기쁨	E60
	감사하는	E61
	신뢰하는	E62
	편안한	E63
	만족스러운	E64
	흥분	E65
	느긋	E66
	안도	E67
	신이 난	E68
	자신하는	E69

4.3 어노테이션/라벨링 조직

텍스트 취득을 위한 수집 시나리오에는 이미 개인의 페르소나 정보가 포함되어 있고, 해당 페르소나는 각 감정과 매핑이 되어 있으므로, 자연스럽게 데이터 수집과 함께 어노테이션 작업이 병행 진행되는 구조로 진행된다.

따라서 클라우드 소싱 기업들의 데이터 수집 시 자동적으로 태깅이 이루어지며, 검수의 과정에서 해당 어노테이션 정보와 수집 문장 정보가 일치하는지를 확인한다.

데이터 수집 시, 주의 사항에 대해 철저한 교육을 진행하며 각각의 감정 상태에 대한 정합성 있는 데이터 구축을 요구한다.



4.4 어노테이션/라벨링 도구

사용자 프로파일이 곧 어노테이션 정보를 포함하고 있으므로, 별도의 어노테이션이 필요하지는 않다. 그러나 각 문장에 대한 중복 및 검수 처리를 위해 10가지의 단어 유형에 대해서는 별도의 태깅을 진행한다.

어제 강원도 속초에 당일치기 여행을 갔다 왔는데 너무 힘들었어.



단어 클릭 시, 10개 개체 class 중 선택
(사람, 국가, 지명, 브랜드명, 기타호칭, 전화번호, 연도, 시간, 수량, 수치)

태깅 틀은 데이터 최종 공개 시 소스코드까지 같이 공개 예정이다.

5. 검수

5.1 검수 절차

데이터의 검수는 클라우드 소싱 시스템을 통해 데이터를 수집한 뒤 클라우드 소싱 기업에서 1차 검수를 진행하며, 주관기관의 검수팀이 상용근로팀과 함께 2차 검수를 수행한다. 또한 최종 검수 결과는 전문가 초빙을 통해 최종 검수를 수행한다.

데이터의 규모가 방대하기 때문에, 클라우드 소싱 수집 단계의 1차 검수가 매우 중요하며, 이에 대한 절차는 다음과 같이 각 클라우드 소싱 기업의 자체 절차를 따른다.



5.2 검수 기준

검수의 과정은 1차 검수, 2차 검수, 3차 검수의 3단계로 구성되어 있다.

검수	검수 목적	검수 항목	검수 내용
1차 검수(형태)	수집한 문장의 형태가 올바른지 확인	맞춤법	국립국어원 표준국어대사전 기준 검수
		특수 기호 표기	온점, 느낌표, 물음표를 제외한 특수 기호 제거
		어투	사람 : 반말, 시스템 : 해요체
2차 검수(내용)	Persona/Emotion에 맞는 문장인지 확인	연령	작성 내용이 제시한 연령대와 맞는지 확인
		성별	작성 내용이 제시한 성별과 맞는지 확인
		상황	작성 내용이 제시한 상황과 맞는지 확인
		감정	작성 내용이 제시한 감정과 맞는지 확인
		대화의 개연성	하나의 대화가 개연성있게 진행되는지 확인
3차 검수(모델링 적격성)	모델링 가능 여부 확인	개체명	개체명 태깅의 오류/적절성 확인
		중의적인 의미	한 문장 내에 있는 중의적인 표현의 의미 확인
		모호한 표현	한 문장 내에 있는 모호한 표현의 정확한 의미확인

1차 검수는 클라우드 소싱 기업에서 각각 원시 텍스트 데이터를 수집한 후 맞춤법 및 특수기호 표기 등에 대한 처리를 수행하며, 기본적인 말투 등 구어 대화 문장으로서의 적절성을 검수한다.

<1차 검수 기준>

- 맞춤법은 국립국어원 기준에 맞추어 검수를 하며, 잘못된 표기법은 정상적으로 수정한다.
- 특수기호는 정제 기준에 맞추어 작성되었는지를 확인한다.
- 데이터 수집 시 연령, 성별 등에 맞지 않는 어투를 쓰고 있는지 확인한다.
- 문장 길이는 최소 2어절에서 최대 17어절까지 허용한다.
- 문장 부호는 온점, 느낌표, 물음표만 허용한다.
- 한 발화에서 두 문장까지 허용한다.
- 문장 부호는 한 발화에 3개 이상 사용하지 않는다.
- 기계가 사용자를 지칭할 때는 "OO님"으로 지칭한다. 다만 맥락에 따라 필요할 경우 "사용자님"이라는 표현도 가능하다.
- 사용자는 반말로 작성하고, 시스템 응답은 존댓말로 작성한다.
- 숫자 및 영어는 한글로 표기한다.
- 욕설, 비속어 등이 포함된 문장은 삭제한다.

2차 검수는 본 과제 수집 데이터의 특성 상 주어진 감정 및 상황 조건에 따른 내용의 검수를 위주로 진행된다. 연령, 성별, 상황, 감정, 대화의 개연성 등 주어진 페르소나 및 감정 상황에 맞는 내용의 대화가 이어지고 있는지를 검수한다.

<2차 검수 기준>

- 수집된 데이터가 연령에 맞는 내용인지 다시 확인한다.
- 수집된 데이터가 성별에 맞는 어투로 작성되었는지 다시 확인한다.
- 수집된 데이터가 주어진 상황에 내용적으로 부합되는지 확인한다.
- 수집된 데이터의 맞춤법이 잘 반영되어 있는지 다시 확인한다.
- 사용자의 첫 번째 발화가 주어진 감정 상태를 잘 반영하는지 확인한다.
- 사용자의 첫 발화 및 응답에서 복수의 감정 상태가 반영된 경우 이를 통과 대상에서 제외한다.
- 사용자의 발화가 전체 대화의 맥락에서 벗어나는 경우 이를 통과 대상에서 제외한다.
- 시스템의 응답이 사용자 발화와 대화 의미가 통하지 않거나 모호한 경우 이를 통과 대상에서 제외한다.
- 시스템의 응답이 심리적 대화 전략에 맞지 않는 경우 이를 통과 대상에서 제외한다. 통과 대상 전략은 반영하기, 호응하기, 맞장구치기, 공감하기, 단순 반응, 동의, 되묻기, 감정 표현 반복, 유발 질문하기 등이며, 통과 비대상 전략은 충고하기, 단정하기, 강요하기, 부정적 의견 표현하기, 이유 묻기, 해결책 제시하기, 권유하기 등이다. 단, 시스템 응답이 심리적 대화 전략에 맞다고 하더라도, 공격적이거나 부정적 뉘앙스를 내포한 경우 통과 대상에서 제외한다.
- 사용자 이외의 다른 인물에 초점을 맞춘 대화는 통과 비대상으로 처리한다.
- 신조어의 사용은 원칙적으로 금지하나, 각 연령 및 성별에서 대중적으로 사용되는 신조어의 경우에 한하여 일부 허용한다. (수포자 등)

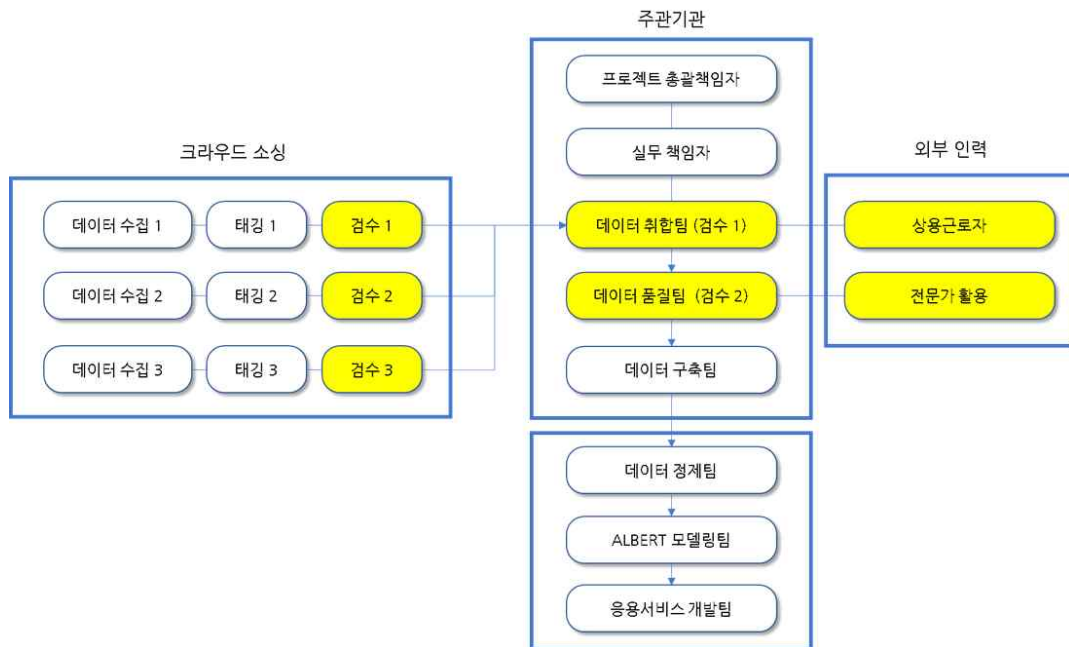
3차 검수는 ALBERT 엔진 학습을 위한 대화 모델링 관점의 검수로, 문맥적 의미는 올바르지만 인공지능 시스템의 의미값 해석에 영향을 줄 수 있는 사항들이 있는지 확인하는 과정이다. 중의적 표현이나 모호한 표현, 기타 모델링에 장애가 될 수 있는 사항들이 있는지 확인한다.

<3차 검수 기준>

- ALBERT 학습에서 모델링 오류를 발생시키는 문장은 제외 처리한다.
- 문장 의미 모호성이 확인되는 문장은 별도 선별하여 제외한다.

5.3 검수 조직

데이터 검수는 클라우드 소싱 기업의 1차 검수를 거친 결과물에 대해 주관기관이 2차, 3차 검수를 진행하는 조직 체계를 갖추고 있다.



5.4 검수 도구

1차 검수는 데이터 수집 도구를 활용하면서 데이터 수집을 진행할 때 해당 툴에서 직접 검수를 진행한다.

2차 검수는 별도의 툴 없이 엑셀 등의 문서 편집기를 활용하여 내용 검수를 진행한다. 맞춤법은 웹 기반의 맞춤법 검사 사이트를 활용하거나, 국립국어원의 맞춤법 규정을 참고한다.

3차 검수는 내용적 모호성에 대해서는 엑셀 등의 문서 편집기를 활용하여 검수하고, 모델링 적합성은 ALBERT 모델링 툴을 활용하여 인공지능 학습을 하는 과정에서 통과 비대상 목록을 선별한다.

5.5 기타 품질관리 활동

<오류 발생 예방 활동>

- 용역 수집의 경우, 데이터 품질 및 오류 유형에 대한 피드백을 정기적으로 제공하고, 주간 회의를 통해 수집량 및 품질을 확인한다.
- 상용 근로자의 경우 데이터 품질 관리팀과 실시간 질의 응답을 통해 데이터 검수 오류 발생을 미연에 방지한다.

<데이터 오류 발생 시 대처 사항>

- 오류 발생 원인 분석 수행한다.
- 분석 결과에 대해 가이드 수정 추가하여 반복 발생 방지한다.
- 자주 발생하는 오류에 대한 주의 감독 강화한다.

<데이터 망실 시 백업 계획>

- 데이터 저장용 서버 외에 백업 서버를 활용하여 데이터 보존 수행한다.

<품질 오류 발생 시 대응>

- 검수 과정에서 품질 오류 다량 발생 시 정제/검수자 재교육을 수행한다.
- 오류 케이스에 대한 전체 정제/검수자 공지 및 가이드 수정을 강화한다.

6. 활용

6.1 활용 모델

6.1.1 모델 학습

본 프로젝트 산출물의 코퍼스 데이터는 BERT의 소용량 알고리즘인 ALBERT를 활용할 계획이다.

BERT는 문맥에 따라 다른 단어 임베딩을 만들어 문맥 정보를 잘 활용할 수 있다. BERT는 11개의 NLP 태스크에서 놀라운 성능을 보여주면서 그 진가를 인정받아 뉴욕 타임즈의 지면을 장식하기도 했던 모델이다.

BERT 활용의 한 사례인 Span Prediction을 이용하면 Slot Value의 시작점에 대한 확률과 종료점에 대한 확률을 계산하고 이를 바탕으로 Slot Value를 추측하게 된다. 이렇게 되면 최소한 Slot의 Value들을 미리 알고 있지 않아도 된다.

즉, <택시호출, 목적지, 예술의 전당>에서 “예술의전당”이라는 목적지는 미리 알지 못해도 사용자의 얘기에서 추출할 수가 있으므로, <택시호출, 목적지>에 해당하는 모듈만 따로 학습해 놓으면 된다.

또한 BERT는 학습된 모델의 활용이 용이하다. 구글에서는 BERT를 활용해서 Zero-shot Learning이 가능한 모델을 발표했다.

다음의 두 비행기 예약 서비스는 Flight Service A에서 비행기 검색 의도는 SearchFlight이라고 정의되고, light Service B에서는 FindFlight라고 정의된다. 실상 이들이 하는 일은 거의 같지만, 기존의 시스템에서는 다른 서비스로 간주해 다른 데이터를 가지고 학습을 하게 된다.

만약 BERT가 이 두 의도가 비슷하다는 것을 알려준다면, Flight Service A 서비스를 위해 이미 학습이 끝난 모델을 거의 그대로 Flight Service B 서비스를 위해 사용할 수 있다.



6.1.2 ALBERT는 BERT를 개선한 효율적인 모델이다. BERT와 같은 Pre-trained language representation 모델은 일반적으로 모델의 크기가 커지면 성능이 향상되지만, 모델이 커짐에 따라 다음의 문제가 발생한다.

- (1) Memory Limitation - 모델의 크기가 메모리량에 비해 큰 경우 학습시 OOM(Out-Of-Memory) 발생
- (2) Training Time - 학습하는데 오랜 시간이 소요됨
- (3) Memory Degradation - Layer의 수 혹은 Hidden size가 너무 커지면 모델 성능 감소

ALBERT는 모델을 최적화하고 학습 방법을 개선해 성능 유지하면서 모델의 크기는 줄인 경량화된 버전의 BERT로, 현재 SQuAD2.0의 최상위권을 차지하고 있는 진보된 모델이다.

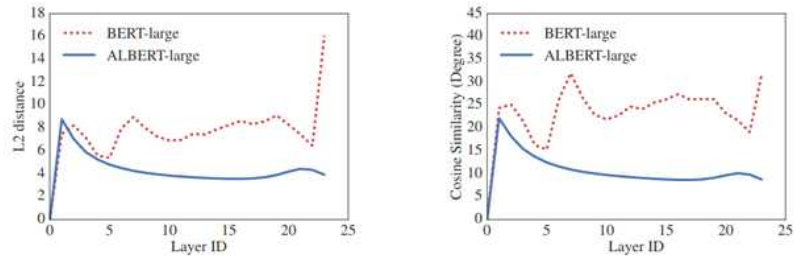


Figure 2: The L2 distances and cosine similarity (in terms of degree) of the input and output embedding of each layer for BERT-large and ALBERT-large.

ALBERT데이터 정제를 위한 Tokenizer를 활용한다. 입력된 문장에 대해서는 BPE, Predicted Slots, Mecap 등 태깅 요소들로 구분하여 분석하고, 여러 요소 모듈의 장점을 조합한 하이브리드 형태의 Tokenizer를 구성하여 활용한다. BPE의 최대 Vocabulary coverage 유지, Mecab의 조사 분리 능력을 극대화한다.

입력 문장 : 사무실에 있는 씨씨티비 보여줘

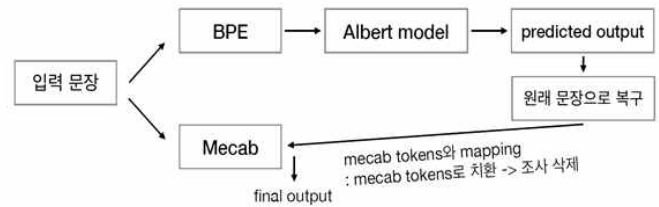
BPE : ['[CLS]', '사무', '실에', '있는', '씨', '씨', '티', '비', '보', '여', '줘', '[SEP]']

predicted slots (using BPE) :

['[O', 'LOCATION', 'LOCATION', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']]

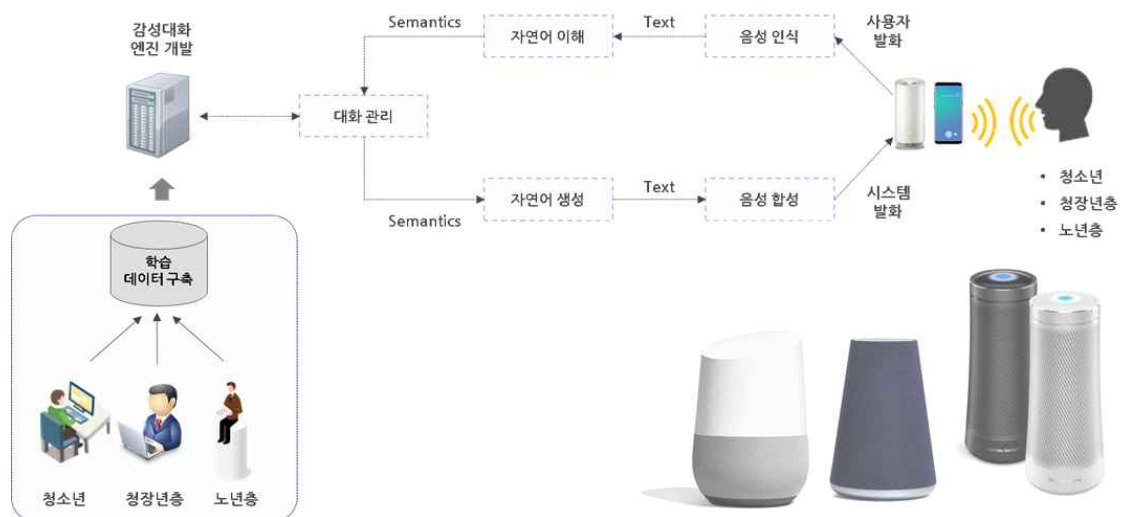
Mecab :

{ '사무실에': '사무실_NNG^에_JKB', '있는': '있_VV^는_ETM', '씨씨티비': '씨씨_IC^티비_NNG', '보여줘': '보여줘_VV+EC+VX+EC' }



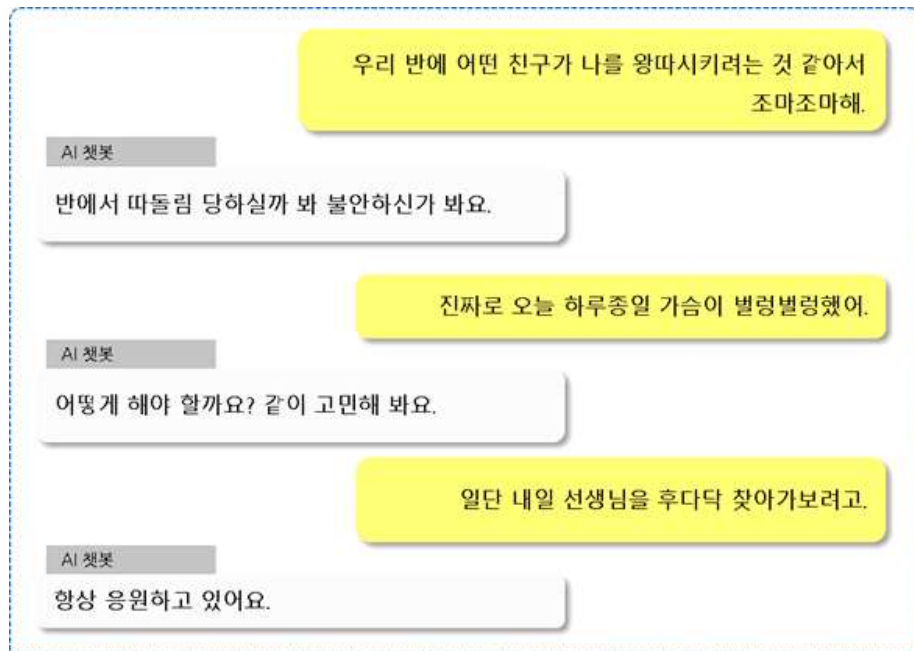
6.1.3 서비스 활용 시나리오

본 과제 산출물인 코퍼스 데이터를 통해 자연어 이해(ALBERT) 부분을 활용한 감성 대화 챗봇을 구현할 예정이다. 이와 더불어 주관기관이 보유 중인 음성인식(STT), 음성합성(TTS) 등의 엔진 기술을 활용한 감정 인식 기반 감성 대화 챗봇이 응용 서비스로 구현될 예정이다.



별도의 GUI를 통해 AI 챗봇과 사람의 대화를 아래와 같이 표현하고, 대략 3턴 내외의 감성 대

화를 이어나가는 대화 시스템이 구현될 예정이다.



6.2 접근

본 프로젝트에서 수집된 모든 데이터는 한국정보화진흥원과 AI Hub를 통해 공개될 예정이다.

6.3 유지보수

본 프로젝트 종료 후에도 지속적인 오류의 수정과 모델링의 개선을 수행할 계획이며, 추후 수정된 내용을 재배포 예정이다.

본 컨소시엄은 이번 과제 산출물을 통한 사업화를 목표로 하고 있기 때문에, 지속적인 데이터 수정 및 보완 작업을 진행 예정이다.

사용자 오류 또는 수정 데이터는 지속적으로 재배포할 예정이다.

데이터 품질 문제 발생 시 주관 기관 품질 책임자 주관으로 해당 데이터 복구 및 보완 활동을 추진한다.

붙임1

인공지능 데이터 명세서 양식

데이터 이름	감성 AI 대화 코퍼스	
데이터 포맷	JSON	
활용 분야	감성 AI 챗봇	
데이터 요약	텍스트 형식의 코퍼스 데이터로, 60가지의 감정을 개인 페르소나에 매핑한 감정 상태별 구어 텍스트 대화 데이터임	
데이터 출처	직접 제작	
데이터 이력	배포버전	V 1.4
	개정이력	신규
	작성자/ 배포자	송민규 (총괄책임자)
데이터 구성	감성 대화 말뭉치 텍스트 285,601 문장 (JSON)	
어노테이션 포맷	페르소나, 감정상태 등	
데이터 통계	데이터 구축 규모	27만 문장 구어 텍스트
	데이터 분포	남성:여성 = 4.4:5.6 청소년:청년:장년:노년 = 0.7:1.2:1:1 감정 상태 = 6개 기본 감정(총 60개 세부 감정)
	기타 활용 통계	성별, 연령, 상황별 분포 통계 감정 상태별 분포 통계
기타 정보	대표성 (Coverage)	전국 대상 클라우드 소싱 (지역, 연령 등 배분)
	독립성	직접 제작으로 법적 문제 없음 (클라우드 소싱 시 동의 절차)
	유의사항	감성 AI 관련 기술 발전에 기여 기대
	관련 연구	- 김연화, 김형주, 김봉완, 이용주, "공동 이용을 위한 음성 인식 및 합성용 음성코퍼스의 발성 목록 설계", 2002. - 조철우, 박인서, 이용주, 김봉완, "배우에 의한 한국어 정서음성 데이터베이스 수집", 2004. - 소아람, 박기남, 임희석, "식당예약 및 추천을 위한 한국어 대화 코퍼스 구축 연구", 한글 및 한국어 정보처리학술대회, 한국정보과학회 언어공학연구회, 2018