대한민국 대전환 한국판뉴딜

2020년 인공지능 학습용 데이터 교육영상 “감성 대화 말뭉치” 소개 및 활용

2021.03

미디어젠㈜
상무 송민규

과학기술정보통신부 NIA 한국지능정보사회진흥원

2021.03

INDEX

01. 인공지능 학습용 데이터 소개

02. 저작도구 활용법

03. 데이터 학습 방법 및 시연

04. 학습된 모델 결과 확인 방법 및 시연

05. 서비스 개발 시 학습 된 모델 활용 방법

01 인공지능 학습용 데이터 소개

데이터 구축 목적

- 대화 속에서 드러나는 개인의 감정을 인식하여 감성적인 대화 표현을 출력함으로써, 정서적 커뮤니케이션이 가능한 인공지능 대화형 챗봇 구현이 가능한 텍스트 데이터를 구축
- 우울증 등 부정적인 감정 상태의 대화가 감지되면, 긍정적 방향으로 대화를 이끌어갈 수 있는 대화체 구어 코퍼스 데이터 구축

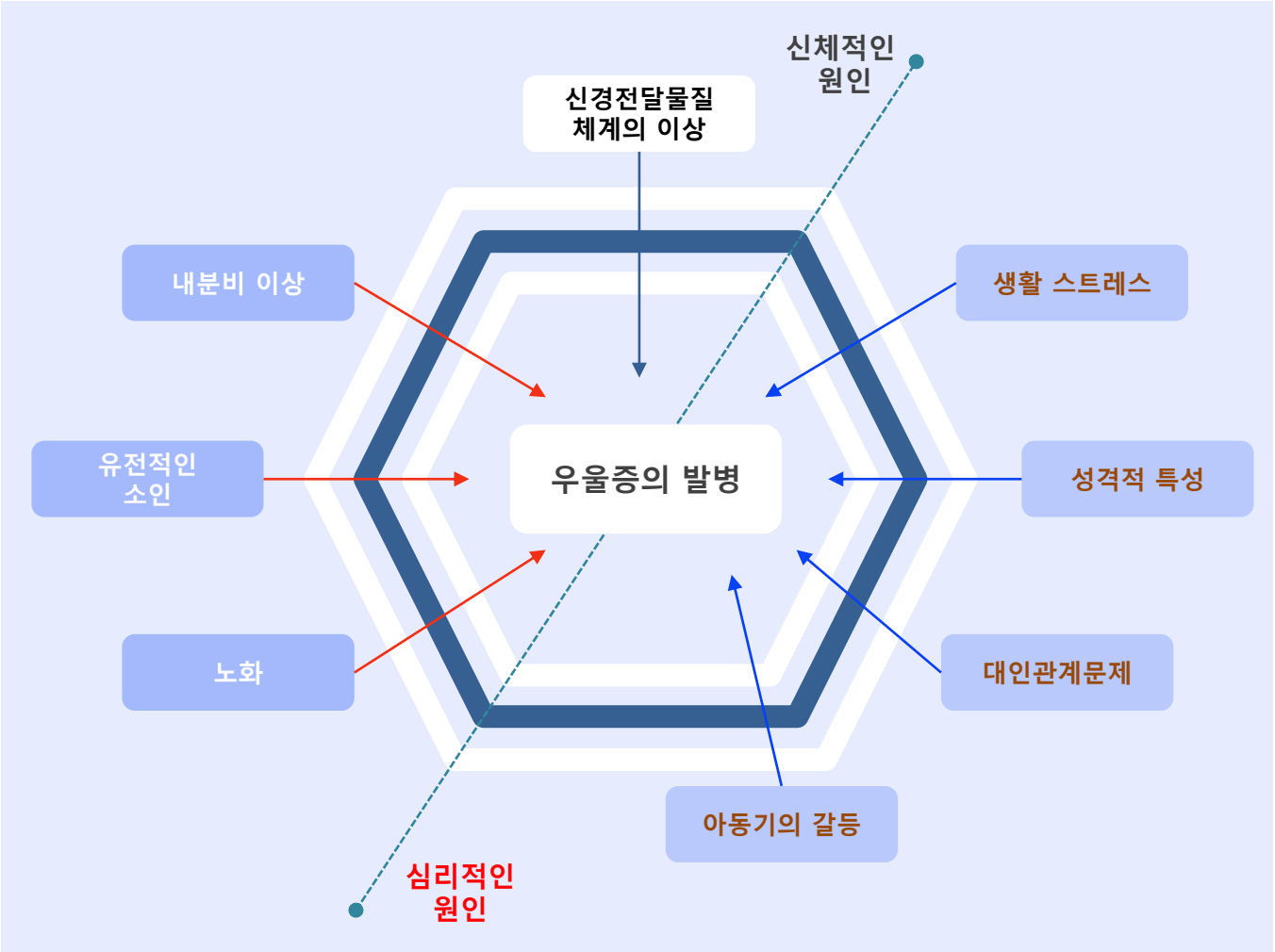
데이터 구성

- 6가지 감정 대 분류가 포함된 코퍼스 수집 (분노, 슬픔, 불안, 상처, 당황, 기쁨)
- 각 감정 대 분류를 총 60가지의 세부 감정으로 구분하여 데이터 수집 프레임을 생성
- 일반인 화자 약 2,700여 명으로부터 상세 시나리오에 따른 발화 패턴을 수집
- 감정 상태 정확성 및 대화 내용의 일관성 측면에서 데이터 품질 검수 수행

데이터 규모

- 총 27만 문장의 텍스트 코퍼스 구축 완료
- 총 1만 문장 규모의 테스트용 음성 DB를 구축하여, 감정 인식 평가 수행 지원
- 데이터에 대한 공인 검증 시험 수행

01 인공지능 학습용 데이터 소개



01 인공지능 학습용 데이터 소개

감정 인식을 통한 감성 대화 수행

60가지 감정 분류						
기본	분노	슬픔	불안	상처	당황	기쁨
1	툴툴대는	실망한	두려운	질투하는	고립된	감사하는
2	좌절한	비통한	스트레스 받는	배신당한	남의 시선을 의식하는	사랑하는
3	짜증내는	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

01 인공지능 학습용 데이터 소개

감정 상태 표현을 위한 페르소나 고려 대상

페르소나 분류		
No.	**대분류**	**상세 분류 (택1)**
1	**나이**	청소년 / 청년 / 중장년 / 노년
2	**성별**	남 / 여 / 기타
3	**원가족관계1 (보호자)**	양부모가정 / 한부모가정 / 조부모가정 / 기타(보육원)
4	**원가족관계2 (형제자매)**	형제자매없음 / 있음(n남n녀, 장남, 차녀, 등 기재)
5	**결혼여부1 (배우자)**	배우자 없음(미혼),
6	**결혼여부2 (자녀)**	자녀 없음 / 자녀 있음
7	**교육수준**	중졸 이하 / 고졸 이하 / 대졸 이상
8	**직업군**	전문직 / 생산직 / 사무직 / 서비스직 / 자영업 / 학생 / 기타
9	**월평균 가구소득 (만원)**	100만원 미만 / 100~200만원 미만 / 200~300만원 미만 / 300 ~ 400만원 미만 / 400만원 이상
10	**건강상태 – 신체**	양호 / 암질환 / 심장질환 / 뇌혈관질환 / 치매 / 당뇨 / 수면장애 / 선천적 장애 / 후천적 장애
11	**건강상태 - 정신**	양호 / 우울 / 불안 / 중독 (알코올, 마약) / 자살시도 또는 자해 / 섭식장애

01 인공지능 학습용 데이터 소개

심리 상담 전문가 자문을 통한 우울증 원인 항목 도출 (130개에서 60개로 단순화)

연령기		대분류	1	2	3	4	5	6	7	가지 수
청소년기	1	학업문제	주관적 성적 평가	진로/취업문제	미디어 중독(게임, 인터넷, SNS)					15개
	2	친구관계	학교폭력	왕따	이성친구과의 갈등	동성친구와의 갈등	또래 압력			
	3	가족관계	양육자의 지나친 간섭,	양육자의 무관심,	양육자의 과잉보호,	학대적, 방임적 태도,	가정불화	재혼가정	한부모가정	
청년기	1	정체성혼미	인생 목표 부재,	현실과 이상의 괴리,	낮은 자존감, 낮은 자기효능감(자신감),	부모로부터의 독립과 의존(양가감정),	부정적 자아상			15개
	2	스트레스와 불안	과도한 경쟁,	암울한 취업시장,	전망 없는 미래,	상대적 열등감,	학업스트레스,	진로고민,		
	3	관계적 문제	친구와의 갈등,	부모와의 갈등,	가정불화,	낮은 사회적/정서적지지				
중년기	1	경제적 문제	낮은 소득수준 및 빈곤,	소득의 감소,	생계부양에 대한 부담,	직무 스트레스				15개
	2	건강상태	중증질환, 만성질환,	장애유발 질환, 예후가 좋지 않은 질환,	지각된 건강상태,	갱년기 (비생산성에 따른 우울),	노화에 의한 위기감, 무가치감			
	3	가족 관련	배우자 사별,	이혼,	가정생활 스트레스,	가족에 대한 책임,	가족갈등, 부정적 가족갈등 대처방식,	다중역할(부모봉양, 자녀 독립과 결혼)		
노년기	1	건강약화	만성질환, 신체적 및 인지적 기능 약화, 건강쇠약,	일상생활 활동 제한,	배우자 부양 및 돌봄					15개
	2	경제적 어려움	퇴직,	경제력 상실, 소득 상실,	노동력 상실,	노후준비 미비,	사회적 역할(신분과 지위)상실			
	3	고독과 무위	노년기 이혼,	대인관계 축소,	유대감 상실,	의사소통 단절, 소외, 외로움,	학대받는 느낌,	배우자 및 친족의 죽음,	사회적 지지 감소	

01 인공지능 학습용 데이터 소개

상세 검수 절차 및 내용

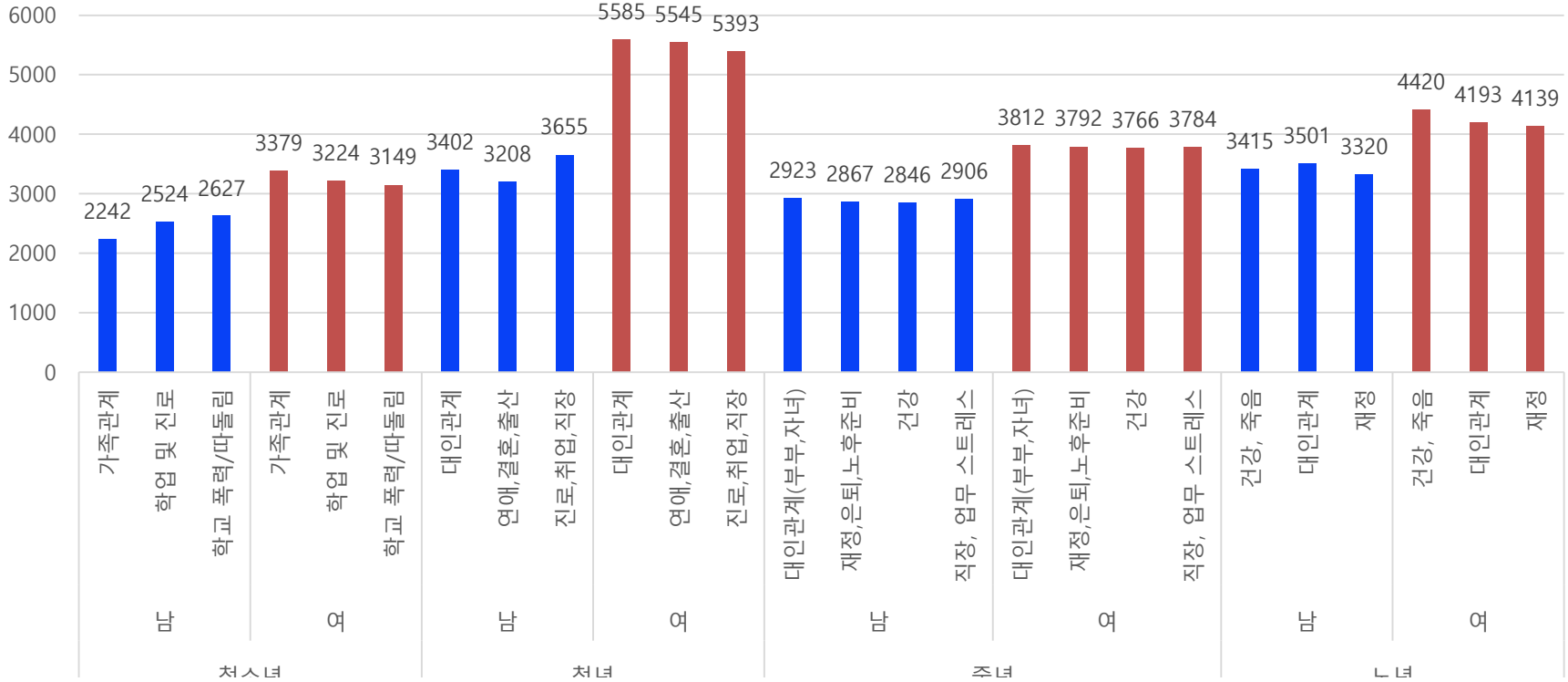
크라우드 소싱을 통한 원시데이터 수집

- **주어진 상황에 맞지 않는 부분은 별도 검수를 통해 삭제 또는 수정**
- **수집된 데이터는 JSON 타입으로 저**
- **장d Depth를 기준으로 수집**
  정제 담당자, 상용 근로자, 전문가 등의 검수를 받으며, 페르소나와 감정 상태에 알맞는 대화를 수집

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 개연성 있게 진행되는지 확인
불특정 다수를 위한 가이드 제공	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

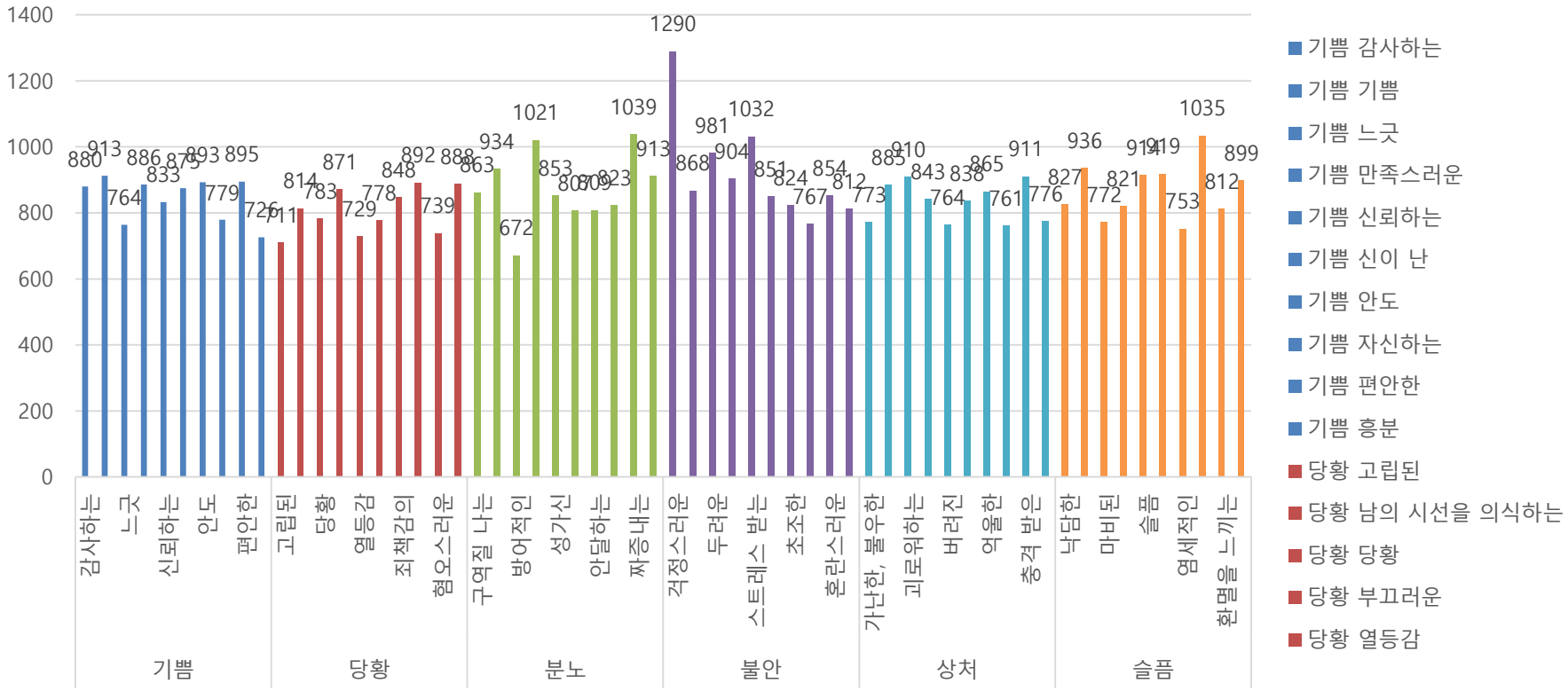
01 인공지능 학습용 데이터 소개

페르소나별 데이터 분포



01 인공지능 학습용 데이터 소개

감정 별 데이터 분포



01 인공지능 학습용 데이터 소개

어노테이션/라벨링 Persona_Emotion ID 규칙

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

* 연령(A), 성별(G), 시스템응답(C)의 기본 ID가 배정

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

* 감정 상태(Emotion)의 경우 상황(Situation : S), 질병(Disease : D), 감정(Emotion : E)의 ID가 배정

01 인공지능 학습용 데이터 소개

어노테이션/라벨링 Persona_Emotion ID 규칙

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

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

구분	항목	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

* 프로파일 ID 할당 샘플

01 인공지능 학습용 데이터 소개

Json 포맷 샘플


```
{
"profile":
    "profile-id": "Pro_13409",
    "persona": {
        "persona-id": "A04_G02_C01",
        "human": ["A04", "G02"],
        "computer": ["C01"]},
    "emotion": {
        "emotion-id": "S04_D02_E38",
        "type": "E38",
        "situation": ["S04", "D02"]
        }
    },
"talk": {
    "id": {
        "profile-id": "Pro_13409",
        "talk-id": "Pro_13409_00009"
        },
    "content": {
        "HS01": "이번에 새로 옮긴 요양원에서 생활하고 있는데 새로운 요양원인지라 다소 긴장되고 조심스러워.",
        "SS01": "새로운 요양원에서 생활하시는 데 긴장감을 느끼고 계시군요.",
        "HS02": "예전 요양원과는 생활방식도 다르고 사람들도 달라서 혹시라도 실수하게 될까 봐 조심스러워.",
        "SS02": "어떻게 하면 지금 상황에서 조금이라도 덜 스트레스 받을 수 있을까요?",
        "HS03": "얼른 요양원안에서 친구를 만들어야겠어. 친구를 만들면 새로운 요양원에 빨리 적응할 수 있을 것
        같아.",
        "SS03": "마음이 맞는 친구 분을 만드셔서 빠르게 새로운 요양원에 적응하시길 바라요."
        }
    }
},
```

01 인공지능 학습용 데이터 소개

인공지능 모델 소개 (ALBERT)

ALBERT는 BERT를 개선한 효율적인 모델

BERT와 같은 Pre-trained language representation 모델은 일반적으로 모델의 크기가 커지면 성능이 향상되지만, 모델이 커짐에 따라 다음의 문제가 발생.

- **Memory Limitation** - 모델의 크기가 메모리량에 비해 큰 경우 학습시 OOM(Out-Of Memory) 발생
- **Training Time** - 학습하는데 오랜 시간이 소요됨
- **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 모델임.

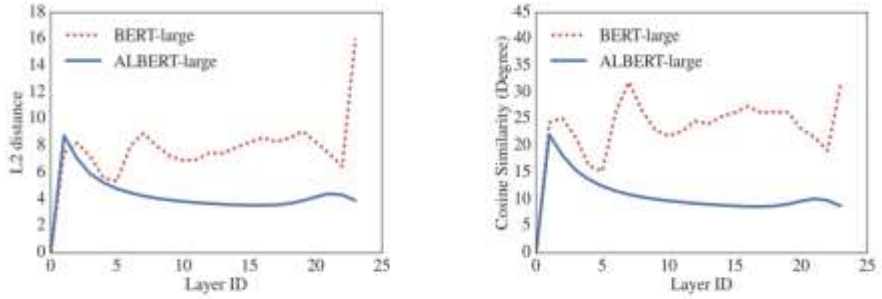


Figure 2: The L2 distances and cosine similarity (in terms of degree) of the input and output embedding of each layer for BERT-large and ALBERT-large.

01 인공지능 학습용 데이터 소개

인공지능 모델 소개 (ALBERT)

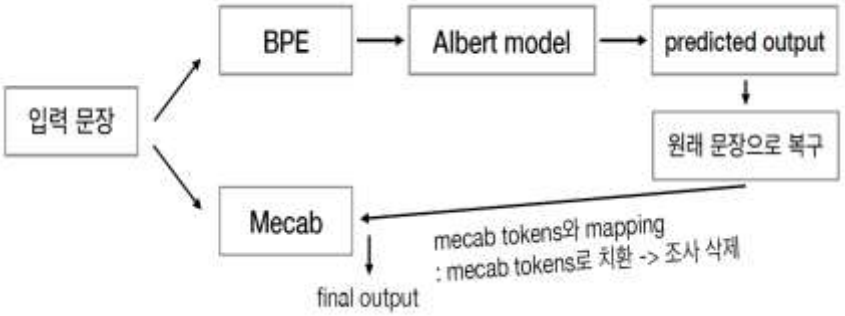
■ 오픈 소스 기반 최적 Tokenizer 구성 및 활용

- 입력된 문장에 대해서는 BPE, Predicted Slots, Mecab 등 태깅 요소들로 구분하여 분석함.
- 여러 요소 모듈의 장점을 조합한 하이브리드 형태의 Tokenizer를 구성하여 활용.
  - BPE의 최대 Vocabulary coverage 유지
  - Mecab의 조사 분리 능력을 극대화.

입력 문장 : 사무실에 있는 씨씨티비 보여줘

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

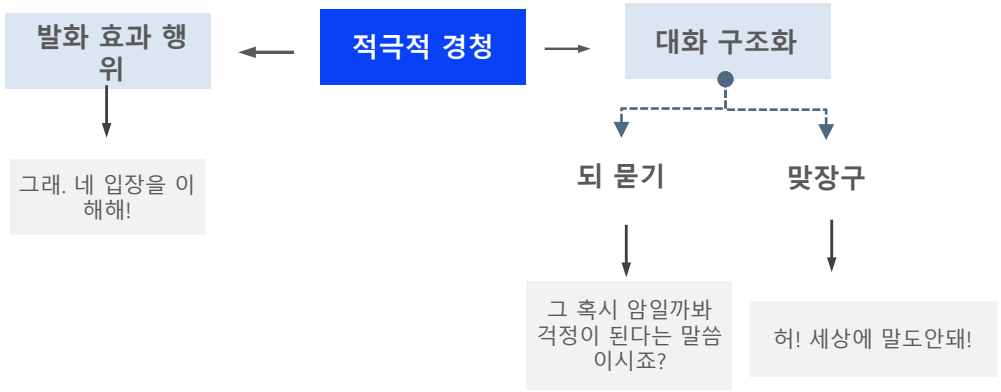
predicted slots (using BPE) :
[['O', 'LOCATION', 'LOCATION', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']]
Mecab :
{'사무실에': '**사무실_NNG^에_JKB**', '있는': '있_VV^는_ETM', '씨씨티비': '씨씨_IC^티비_NNG', '보여줘': '보여줘_VV+EC+VX+EC'}



01 인공지능 학습용 데이터 소개

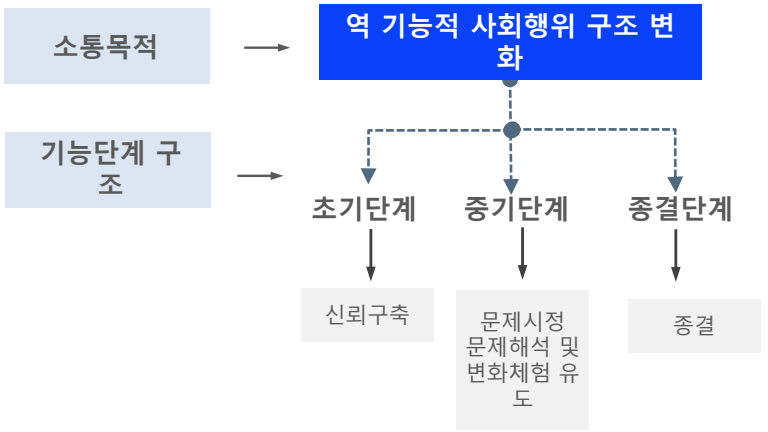
대화 전략 : 적극적 경청과 소통 단계

발화효과 행위와 대화구조화 행위의 전제조건



< 적극적 경청의 행위 유형>

적극적 경청은 특히 신뢰 구축과 문제를 시정하는 초기단계에서 중요한 역할을 함:



< 가족치료 대화의 소통목적 및 기능단계 구조>

가족치료 대화의 기능단계:

초기: 관계형성을 통한 신뢰구축과 이에 기초하여 내담자 및 가족구성원들의 문제점을 탐색하는 단계

중기: 변화를 위한 전략으로 가족이 변화되고 내담자 문제가 치유되도록 돕는 단계

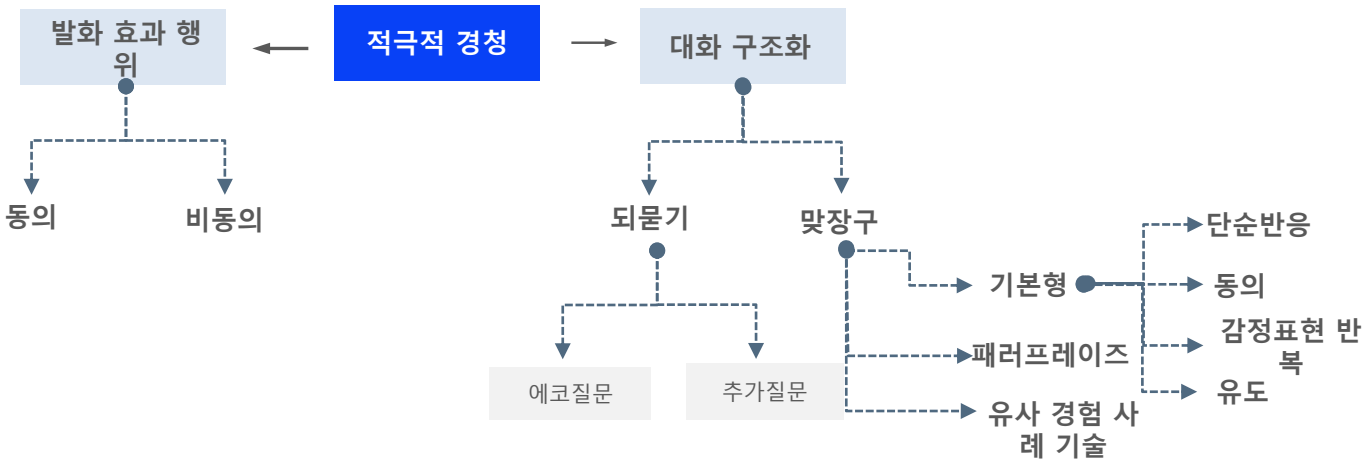
종결: 종결가능성에 대해 탐색하여 내담자에게 계약내용을 상기시켜 종결통보를 위한 계획을 수립.

조용길(2012), "건설적 상담대화와 적극적 경청", 독어학, pp267-287.

01 인공지능 학습용 데이터 소개

대화 전략 : 적극적 경청 (공감, 되묻기, 맞장구)

적극적 경청의 행위 유형과 상응하는 표현 유형:



< 적극적 경청 행위 및 표현유형 >

적극적 경청의 세부유형:

적극적 경청의 전제요건들이 갖추어졌을 때 긍정적인 의미를 가짐. 상대방의 입장 그의 생각과 감정을 적절히 반영해야 하고,

상대방이 대화를 주도하도록 해야하며 이해하는 자세가 필요함.

조용길(2012), "건설적 상담대화와 적극적 경청", 독어학, pp267-287.

01 인공지능 학습용 데이터 소개

감성 AI 대화 사례

감정 인식을 통해 감성적 주제로 대화를 진행

02 저작도구 활용법

획득·정제 도구

- **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 작성.**

- **어노테이션/라벨링 절차 활용**
  텍스트를 수집하면서 곧바로 어노테이션이 진행되는 형식이므로, 아래의 그림과 같이 상세 상황에 대해 순서대로 대화의 질의와 응답을 작성하는 절차로 데이터가 수집.
  구어 데이터를 매우 효율적으로 수집할 수 있는 구조로 데이터 구축이 진행.

<어노테이션/라벨링 절차>

수집된 대화 데이터는 각각의 페르소나에 해당되는 ID와 수집 텍스트가 쌍으로 저장

페르소나와 감정 태그는 자동으로 부여되게 되므로 별도의 작업이
없이도 자연스러운 태깅 수행

데이터 수집 툴을 통해 자동으로 수행

<어노테이션/라벨링 기준>

03 데이터 학습 방법 및 시연

데이터 학습 방법 시연

영상 별도 제출 예정

04 학습된 모델 결과 확인 방법 및 시연

학습 모델 결과 확인 시연

영상 별도 제출 예정

05 서비스 개발 시 학습 된 모델 활용 방법

서비스 개발 모델 활용 시연

영상 별도 제출 예정

대한민국 대전환 한국판뉴딜

감사합니다

미디어젠㈜
상무 송민규

E-MAIL :
minks@mediazen.co.kr
PHONE : 02-6429-7104

과학기술정보통신부 NIA 한국지능정보사회진흥원

2021.03