Session 4-3

AI 빅데이터 수집

송민규 상무 (미디어젠)

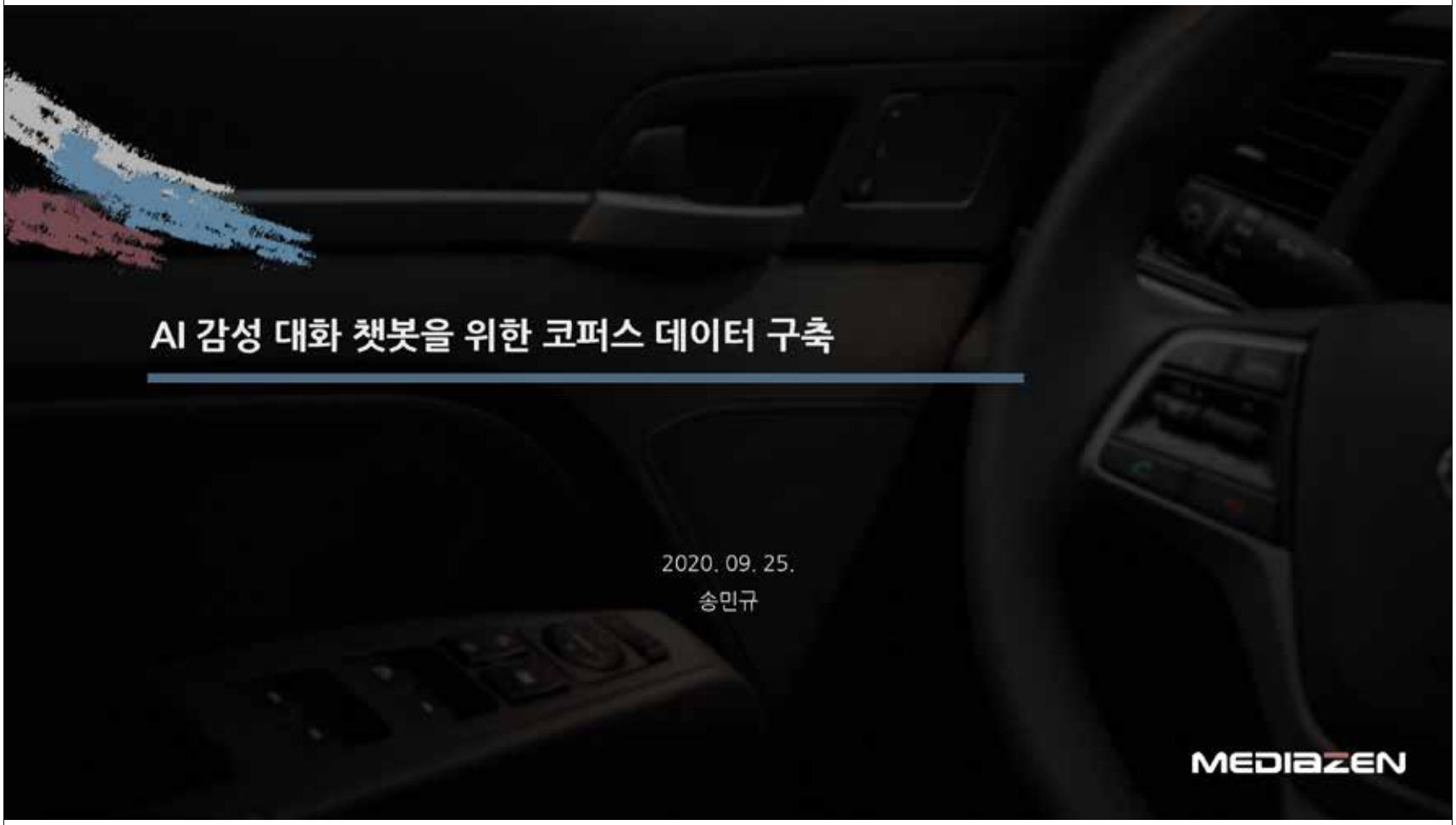
[요약문]

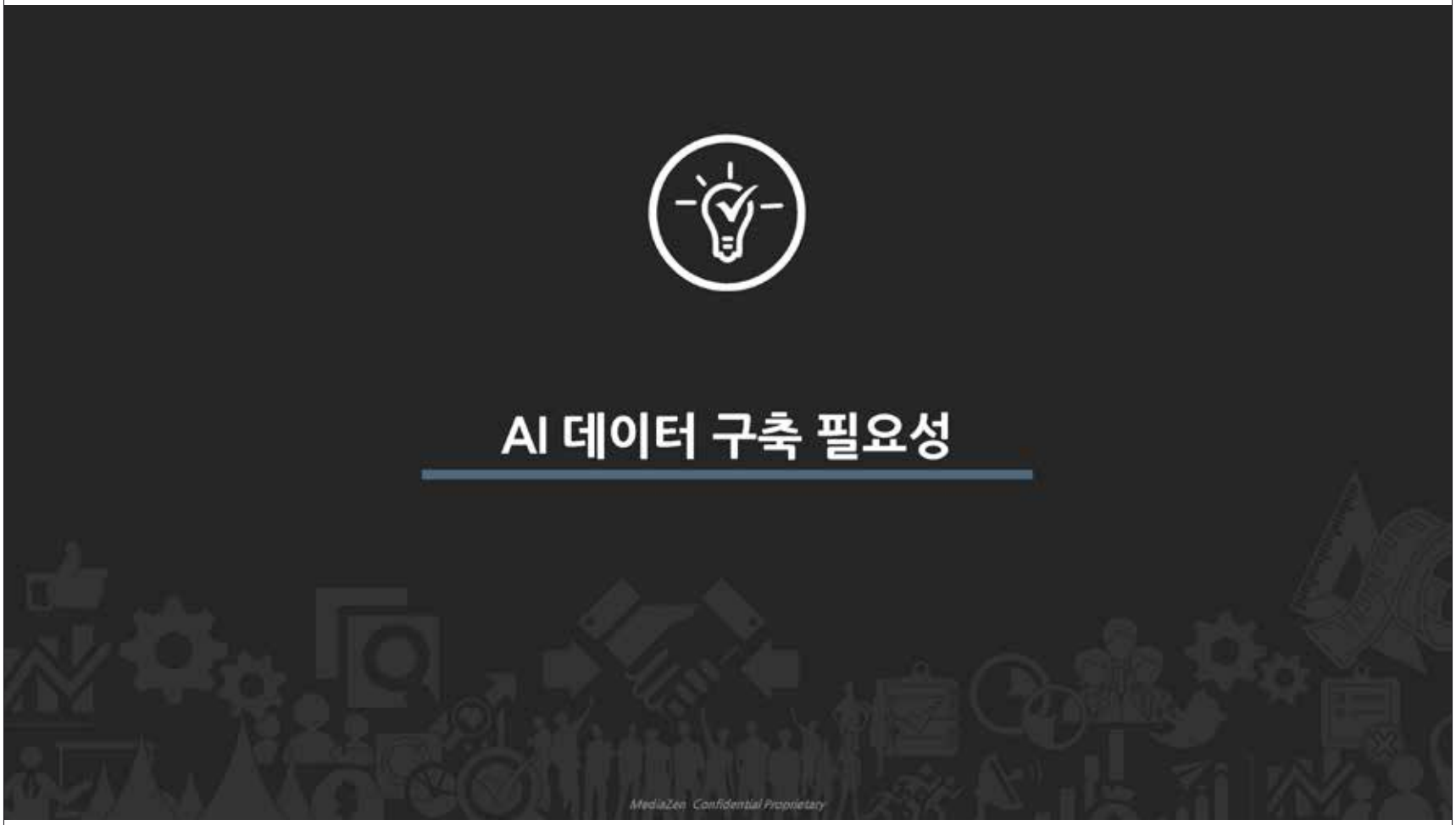
AI 성능 향상을 위해 빅데이터는 필수적으로 확보되어야 하는 중요한 자산이다.
그러나 데이터의 품질과 저작권의 문제 등 데이터 확보의 과정에서 넘어야 할 장애물들이 많이 산재해 있는 것도 주지의 사실이다.
특히 구어 및 대화 텍스트의 희소성은 자연어 처리를 위한 텍스트 기반 AI 데이터 수집에서 극복하기 어려운 문제 중의 하나이다.
본 강연에서는 인공지능 챗봇 개발을 위한 구어 및 대화 데이터 수집에서, 감성 AI 데이터 수집의 문제들을 극복하기 위한 데이터 설계 및 수집 방법, 데이터 모델링 등에 대한 그간의 연구 진행 내용들을 소개한다.

[발표자 약력]

1998년 고려대학교 국어국문학 학사
2001년 고려대학교 응용어문정보학 석사
2007년 고려대학교 국어국문학 박사
2001년~2003년 고려대학교 민족문화연구원 음성언어정보연구실 선임연구원
2003년~현재 미디어젠 상무 (ICT 및 AI 사업 담당)

관심분야 : 음성인식, 음성합성, 자연어 처리, 대화 분석, AI 데이터 구축

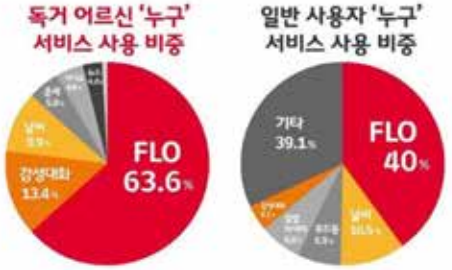

AI 감성 대화 챗봇을 위한 코퍼스 데이터 구축
2020. 09. 25.
송민규
MEDIAZEN


AI 데이터 구축 필요성
MediaZen Confidential Proprietary

AI 데이터 구축 필요성

AI 기반 감성 대화 사용 비중의 증대

"노인들은 잘 못 쓸 거다, 발음이 불확실해서 인공지능이 못 알아들을 거다. 그런 걱정이 많았습니다. 최고령자가 99세입니다. 잘 쓰고 계십니다. 나훈아 노래, 찬송가 주로 들으시고요. 성동구 사시는 97세 어르신도 있습니다. 이분도 인공지능 스피커를 아주 다양하게 쓰고 계십니다."



- 독거노인의 '감성대화' 사용 비중(13.5%)이 일반인 사용 패턴(4.1%)에 비해 월등히 높은 것이 확인됨

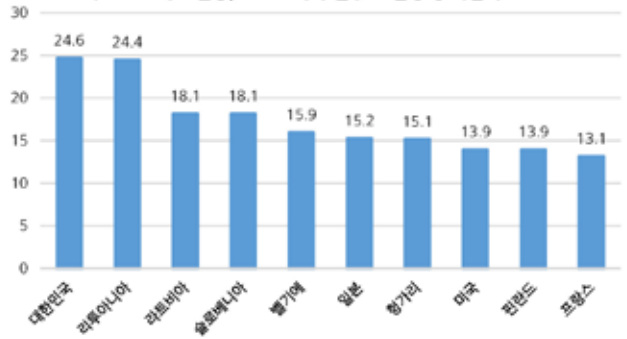
〈 출처 : 블로터 뉴스 (https://www.bloter.net/archives/345897) 〉

AI 데이터 구축 필요성

OECD 최고 자살률, 나이들수록 높아져:

- "24일 통계청이 발표한 사망원인통계 자료를 보면 2018년 자살로 숨진 사람은 13,670명으로, **하루 평균 37.5명이 스스로 목숨을 끊고 있다**는 얘기다. 우리나라의 연령표준화 자살률은 OECD 평균 11.5명의 두배가 넘고 **OECD 국가 가운데 가장 높은 자살률**을 기록했다."
- "특히, 노인일수록 자살률이 높았다. 인구 10만명 당 자살률을 비교해봤더니 10대는 5.8명, 20대 17.6명, 30대 27.5명, 40대 31.5명, 50대 33.4명, 60대 32.9명, 70대 48.9명, 80대 이상 69.8명으로 **50대 이후 늘다가 70대 이후 자살률이 눈에 띄게 높아진 것**을 알 수 있다."

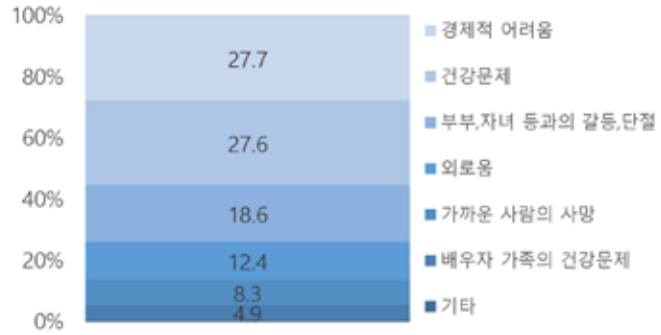
자료: UN 자료 활용, OECD 국가 인구 10만명 당 자살 수



노인이 자살을 생각한 이유:

- "한국보건사회연구원이 2018년 발표한 '노인실태조사'를 보면 **65세 노인의 21.1%가 우울증상을 지니고 있는 것**으로 나타났고, 6.7%는 자살을 생각해본 적이 있다고 답했다. 그리고 자살을 생각해본 적 있는 노인들 가운데 13.2%는 자살을 시도한 경험도 갖고 있었다."
- "조사에 따르면 심리적 **우울감을 유발하는 원인**이 될 수 있는 요소인 가족과의 갈등, 외로움, 가까운 사람의 사망이 **40%가량을 차지**했다."
- "이유의 절반정도를 차지한 경제적 어려움과 건강문제 두 가지도 우울증상으로 이어지는 것을 막는다면 자살을 방지하는데 도움이 될 것이다."

자료: 한국 보건사회 연구원, 2018년 노인실태조사



KBS News, 한국 노인 OECD 최고 자살률…노인들이 가장 힘들어 하는 것은? (2019)

AI 데이터 구축 필요성

우울증 및 돌봄 AI 서비스 동향

항 목	내 용
	▪ 부모님 생활 관리 및 우울증/치매 예방 AI 로봇 부모사랑 효돌 ▪ 여러 매체에서 어르신 돌봄 용 로봇 '효돌'의 독거노인 돌봄 효과 조명
	▪ 시니어 세대의 편의를 위한 원더풀플랫폼 AI 로봇 다솜
	▪ 우울증 치료 AI 챗봇 '워봇'
	▪ 공황장애 셀프 관리 챗봇 서비스 '토닥이'
	▪ 뇌 건강을 위한 인지훈련을 돕는 기능을 하는 카카오톡 챗봇 '새미'
	▪ 스티비 더 로봇: 양로원과 같은 시설에 있는 노인들을 위한 소셜 로봇. ▪ 초기 치매 예방등과 같이 시설에 거주하는 노인들과 다양한 상호작용을 할 수 있음.

AI 데이터 구축 필요성

AI 기반 우울증 및 노인 돌봄 서비스 관련 동향

날짜	매체	내용	노인	우울증	AI
2020. 04. 27	중앙일보	• 8년째 청소년 사망원인 1위 자살...27%는 '우울감' 경험		O	
2020. 04. 22	BLOTER	• SKT, AI 스피커 활용 노인 복지 서비스 시작	O		O
2020. 04. 3	EBN	• 코로나 우울증' AI가 상담해준다		O	O
2020. 03. 19	스포츠경향	• 코로나블루 우울함 달래주는 AI 대화친구 '심심이', 우울증 극복 여름로 인기		O	O
2020. 03. 12	로봇신문	• 원더플플랫폼 '아바딘', 2020 서울어워드 우수상품 선정	O		O
2019. 12. 31	EMD 메디컬뉴스	• 노인 우울증, 객관적 분석 알고리즘 개발	O	O	O
2019. 12. 25	로봇신문	• 가천대,우울증 예방 로봇 · 챗봇 개발 추진		O	O
2019. 12. 16	여웃소싱타임스	• 사물인터넷 활용 독거노인 돌보기 등 디지털 사회혁신 사례 6건 선정	O		O
2019. 11. 22	사이다경제	• 조커가 반한 '인공지능'...실제로 출시됐다		O	O
2019. 11. 20	time.com	• Artificial Intelligence Could Help Solve America's Impending Mental Health Crisis	O	O	O
2019. 10. 27	오다인 기자 블로그	• SK텔레콤, 독거노인 AI 돌봄서비스 성과는?	O		O
2019. 10. 7	ZDNet Korea	• 치매·ADHD·우울증 문제...약물 대신 'AI로 푼다'		O	O
2019. 10. 7	메디컬포스트	• 사망 원인 2위 '우울증' 막아줄 인공지능(AI) 서비스		O	O
2019. 09. 29	KBS 뉴스	• 한국 노인 OECD 최고 자살률…노인들이 가장 힘들어 하는 것은?	O	O	
2019. 09. 25	코메디닷컴	• 이제 정신 건강 상담도 로봇이 하는 시대		O	O
2019. 09. 9	AI Dev_인공지능 개발자 모임	• 노인 돌봄 인형 - 효돌	O		O
2019. 08. 31	한국일보	• "사랑해" "할머니 최고"… 독거노인 마음 돌보는 AI로봇	O	O	O
2019. 07. 9	시사저널	• 우울증 환자는 '자살 암시 흔적'을 남긴다		O	
2019. 07. 9	BLOTER	• AI스피커로 독거노인 돌볼 수 있음까	O		O
2019. 05. 15	IT Chosun	• "흩어진 가족 연결하는 따뜻한 플랫폼 만들겠다"…구승엽 원더플플랫폼 대표	O		O
2019. 04. 19	서울경제	• 김지희 크로스핸제 대표 '도공 동반자형 로봇, 독거노인 돌봄 효율 높일 것'	O		O
2019. 04. 16	KFM 경기방송	• [기획1] '묵숨 맺아가는 노인 우울증' - 극단적 선택 유발하는 다음의 병	O	O	
2019. 04. 3	랩스토리	• [스타트업] 치매 예방을 위한 카카오톡 챗봇 '세리'	O		O
2019. 01. 21	마인드플니스	• 우울증 : 문제의 영역		O	
2018. 09. 27	Cnet Korea	• 퓨처로봇, 수원과학대와 치매 케어 로봇 개발	O		O
2018. 08. 22	DPL(디지털 정신의학 연구소)	• 챗봇과 친구가 될 수 있을까? - 정신건강 챗봇을 알아보자(Woebot, Wysa, Tess)		O	O
2018. 03. 21	더웨어컨설팅	• 우울증 치료 AI 챗봇 '우봇', 8백만 달러 투자 유치		O	O
2018. 01. 11	brunch by 트로스트	• 인공지능이 저를 상담해 준다고요?		O	O
2017. 11. 2	DPL(디지털 정신의학 연구소)	• 우울증 치료 챗봇 Woebot. 정신건강 서비스로서 누구의, 어떤 문제를 해결할 것인가		O	O
2017. 01. 4	삼성서울병원 건강이야기	• 노인 자살, 외로운 선택을 되돌리려면…?	O	O	
2016. 06. 23	중앙일보	• 우울증 환자, 자살 생각 왜 하게 되는 걸까		O	
2016. 06. 22	테크M	• 인공지능(AI) 스피커 독거노인 치매 예방 돕는다	O	O	O

우울증과 노인 돌봄 분야에서 AI 기반 기술에 대한 기대감이 크게 나타나고 있음

AI 데이터 구축 필요성

데이터 직접 취득 및 고품질 태깅 데이터 확보 필요성

무분별한 데이터 크롤링의 저작권 문제 해결 필요

"크롤링 남의 자산 훔치는 범죄행위, 인식 변화 갖자"

대법원 "웹사이트 무단 크롤링은 불법"

잡코리아, 사람인 상대 9년 소송전서 승소
크롤링(데이터 수집) 법적 기준 정립

일반 데이터로는 고품질의 AI 시스템 개발이 어려움

조동희 外(2016), "MUSE 감성주석코퍼스 구축을 위한 분류 체계 및 태그셋 연구"

국외에서는 영어를 대상으로 한 감성주석코퍼스의 구축 연구(Wiebe 2002, Wiebe et al. 2005, Wilson 2008, Saif et al. 2013)가 꾸준히 발표되고는 있지만, 대규모 코퍼스의 경우는 이모티콘 등을 이용해 자동 분류하여 획득된 유형이며, 수작업으로 구축된 코퍼스는 그 규모에 있어 현저하게 작다.

국내의 경우는 감성주석코퍼스 구축 문제에 관심을 가진 공개적인 연구로 영어 MPQA에 기반한 한국어 감성코퍼스(Shin et al. 2012, 2013, Shin 2013, 김은형 외 2013)를 들 수 있는데, 이 경우 구축된 코퍼스 규모가 크지 않아 실질적인 오피니언 마이닝 연구에 적용되기 어려우며, 특히 감성 데이터를 대상으로 하지 않고 범용의 코퍼스인 세종 코퍼스를 대상으로 하여 수행된 것이기 때문에 현실적인 사용자 생성문(User-Generated Texts)의 속성을 반영하지 못하고 있다.

AI 데이터 구축 필요성

AI 기반 감성 대화 기술의 기대 효과

기대 효과

AI 기술의 글로벌 경쟁력 강화

- 비정형 데이터를 활용한 엔진 개발을 통해 선진국에 뒤쳐져 있던 인공지능 기술의 글로벌 경쟁력 확보

감성대화 기술 개발 활성화

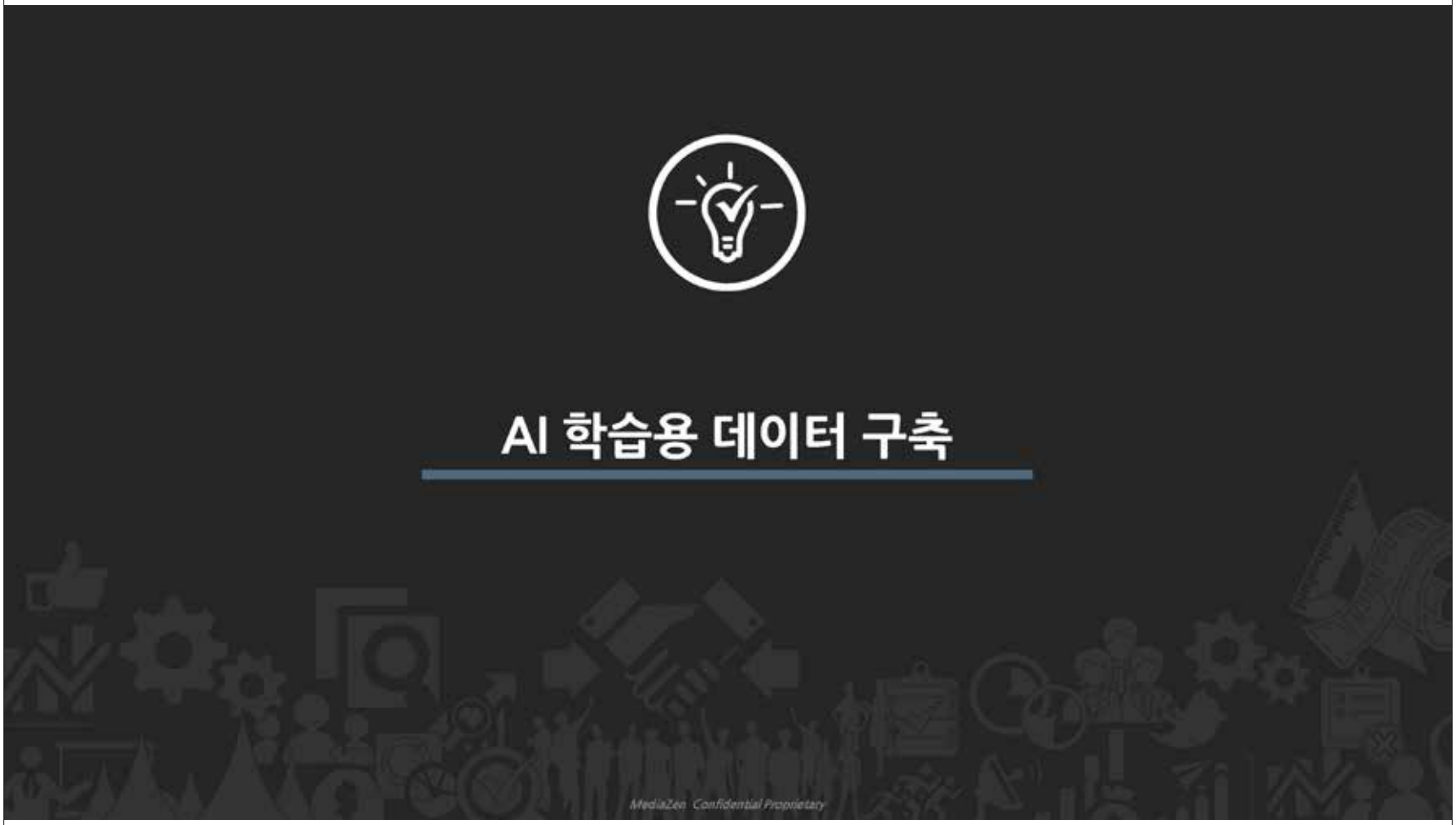
- 고품질의 감성대화 텍스트 데이터 구축을 통해 음성인식 기술의 미래 유망 분야인 감성대화 기술 개발을 활성화

음성솔루션 분야 신시장 확대

- 해외에 뒤쳐진 음성솔루션 분야에서 유망기술인 감성대화 엔진 개발 및 사업화를 통해
  다양한 영역에서의 온라인 서비스 구축과 신시장 확대가 가능

정신건강 악화 사회문제 완화

- 청소년, 청장년층, 노년층이 지닌 스트레스, 우울증 등 개별적인 문제에 대응 가능한 감성대화 서비스를 통해 정신건강 악화라는 심각한 사회문제를 완화


AI 학습용 데이터 구축
MediaZen Confidential Proprietary

AI 학습용 데이터 구축

데이터 구축 목표 및 개요

데이터 구축 목표

감정 인식을 통한
AI 기반 감성 챗봇용
코퍼스 데이터 구축
ALBERT 모델링을 위한 데이터

01

데이터 구축 범위

감성 대화 언어 의미를 구조화
언어 이해 모델링 데이터 구축
대화 응답 시나리오 동반 수집
감성 텍스트 언어 수집

02

데이터 구축 규모

일반인 1,500명을 대상으로 한
인터뷰 및 크라우드 소싱 수행
60개 감정 상태별 분류
코퍼스 약 27만 문장 수집/태깅

03

데이터 구축 일정

데이터 설계 및 인프라 구축 1개월
데이터 수집/태깅 3개월
데이터 품질 검수 1개월
총 5개월 소요 예정

04

AI 학습용 데이터 구축

감정 인식 코퍼스 구축 필요성

인간의 감성은 감정과는 구분되는 심리적 현상으로, 다음과 같은 차이가 있음.

- 감정: 생리적/신체적 반응을 동반, 두뇌의 단계적 정보처리의 결과, 하나의 대상에 대해 여러 사람이 유사한 반응을 보임.
- 감성: 강도가 낮고 생리적 변화 없음, 직관적이고 반사적, 동일한 대상이라도 개인에 따라 다르게 나타날 수 있음.

감정과 감성에 대한 용어는 다음과 같은 차이를 가진다:

용어	의미
감정 (感情)	어떤 현상이나 일에 대하여 일어나는 마음이나 느끼는 기분. feeling(s); [정서] emotion; sentiment; [격정] passion; [충동] impulse
감성 (感性)	1. 자극이나 자극의 변화를 느끼는 성질 2. 〈철학〉 이성(理性)에 대응되는 개념으로 외계의 대상을 오관(五官)으로 감각하고 지각하여 표상을 형성하는 인간의 인식 능력 [감각력] sensitivity; sensibility; sensitive faculty; the sense [감수성] susceptibility

감성과 감성 용어정리 (정현원, 2008)

개인 감성의 발생과 요인은 생활 경험이 큰 영향을 미친다:

개인의 감성발생과 요인

감성 AI 챗봇 개발 목적:

개인의 정서(Emotion), 기분(Mood), 대상에 대한 감상(Sentiment), 소감 및 의견(Opinion) 등이 감성 단어를 통해 텍스트에 나타나게 되고, 문서 상에 표현된 감성 표현을 추출하는 것

AI 학습용 데이터 구축

감정 인식을 통한 감성 대화 수행

기본 6가지 감정을 총 60가지의 세부 기준으로 분류하고, 그 중에서 부정적인 감정에 대해 인지하여 긍정적 감정으로 바뀔 수 있도록 감성 대화를 제공

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

"3 Ways to Better Understand Your Emotions" by Susan David (November 10, 2016), Harvard/McLean Institute of Coaching.

AI 학습용 데이터 구축

감정 상태 표현을 위한 페르소나

다양한 감정이 발현되는 조건을 설계하기 위한 페르소나 구성

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)/배우자 있음(기혼)/기타(별거, 이혼, 사별, 재혼)
6	결혼여부2 (자녀)	자녀 없음/자녀 있음
7	교육수준	중졸 이하/고졸 이하/대졸이상
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

AI 학습용 데이터 구축

코퍼스 데이터 수집 방법

■ 데이터 수집 방법

항 목		내 용
데이터 수집 방법	Quality methods	▪ 질적 방법은 관찰, 인터뷰, 사례 연구, 서면 문서 분석 등의 절차가 포함되고 일반적으로 이벤트와 프로세스에 대한 흐름도 및 서술 설명을 생성
	Quantity methods	▪ 양적 방법은 검사와 평가 척도 및 생리학적 측정에 의존하고 수치 결과를 산출
질문법		▪ 표본의 모든 개인에 대해 동일한 질문을 하는 방법
인터뷰		▪ 면접원의 구두 질문과 연구 참가자의 구두 응답으로 구성
설문 조사		▪ 표본에서 의도한 모집단에게 결과를 일반화하기 위해 질문, 인터뷰를 사용하여 샘플의 참가자 특성, 경험 및 의견에 대한 데이터를 수집

■ Qualitative (데이터 수집 품질 구분)

방법	설명
Observation	▪ 연구원은 주제를 연구하여 사람들이 자신이 하는 말을 하는지 여부를 이해하고 주제에 대한 암묵적 지식에 접근 할 수 있도록 하기 위해 충분히 가까이 접근
Interview	▪ 질문을 하거나 듣고 답변을 개인이나 그룹으로 구조화, 반 구조화 또는 비정형 형식으로 심층적 인 방식으로 듣고 녹음하는 작업이 포함
Focus 그룹 토론	▪ 모두가 대화의 기회를 가지고 의견의 다양성을 제공 할 수 있을 정도로 충분히 작은 그룹과 집중적이고 상호 작용하는 세션
다른 방법들	▪ 신속한 평가 절차, 무료 목록, 말뚝 정렬, 순위, 생활사 등

AI 학습용 데이터 구축

코퍼스 구축 프로세스

AI 학습용 데이터 구축

코퍼스 구축을 위한 데이터 수집 시나리오 제작 방법

코퍼스 수집 시나리오 구성 시 다양한 방법론을 적용

인터뷰

각 대상이 의사소통을 할 때 사용하는 언어를 관찰하고 기록함.
비대면 수집도 가능하므로 크라우드 소싱을 통한 대규모 수집 장점.

과제 중심 의사소통

참가자에게 일정할 역할을 부여.
Role Play 과정에서 언어 수집.
자연스러운 대화 수집이 장점.

스토리 재구성

주제에 따라 주어진 그림이나 예시 지문을 보고 개인의 감성을 표현하는 방식. 다양한 감성 대화 패턴 수집이 장점.

자기 성찰

피험자 스스로가 본 프로젝트에 참여하면서 겪은 점을 감성적으로 표현. 본인이 느낀 점을 그대로 표현한다는 점에서 개인의 주관과 인상을 자연스럽게 수집.

AI 학습용 데이터 구축

코퍼스 데이터 정제 방안

항 목	내 용
띄어쓰기	▪ 신문이나 언론사의 자료를 통해 지면 편집과 일반적인 관례를 따라서 띄어쓰기 오류 발생 - 관례('에베레스트산' (x)-> '에베레스트 산'(o)) // 지면편집('김씨'(x) -> '김 씨'(o))
복합어 처리	▪ 원시 데이터에 띄어 쓰지 않은 형태로 나타난 복합어 중 띄어 쓸 수 있는 곳에 "-"를 삽입 - 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	▪ 수를 적을 때 '만' 단위로 띄어쓰기 적용 ('만, 억, 조' 및 '경, 해, 자')
	- 한글로 표시된 단위 등은 변환 범위에 포함 하지 않음 : 5분 -> {{오}}//{{5}}분
	- 한자어, 영문, 기호로 표시된 단위 등은 다로 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 번호독식과 봉독식으로 표현 : 20개-> {{스무, 이십}}//{{20}}개

외국어 처리	▪ 한글로 변화하되 기준은 국립국어 연구원의 '외래어 한글 표기법'으로 표기 - '한글 음사' 다음 괄호 안에 적은 영문자 :빅맥(big-mac), 팅(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

특수 기호 처리 방안

일반 문자로 변환

"España"라는 국호의 이름과 영어의 동의어 "Spain"이나 "Spanish"에 대해서는 논란의 여지 가 있다.
→ Espana라는 국호의 이름과 영어의 동의어 Spain이나 Spanish에 대해서는 논란의 여지가 있다.

1946년에 제5 구 구조가 성립되었고 그 후 쑨원과 관계된 중산 구라고 개칭했다.
→ 1946년에 제5구 구조가 성립되었고 그 후 쑨원과 관계된 중산 구라고 개칭했다.

특수기호 그대로 남김

- 마침표, 느낌표, 물음표
- 기호를 빼면 문장이 어색해지는 경우 (한자, 외국어, 한글 자모, 단위, 퍼센트, "&", 통화류

闇の中の魑魅魍魎를 오랜시간 함께해왔던 신도 가네토의 각본으로 찍었고 이것이 칸 영화제 경쟁부문에 진출했다.
나 ಂಶಮ ಗೂ&ಂಶಲ 도 ZWNJ를 필요로 하는 고유명사의 예이다.

일괄 처리가 힘든 경우

- 슬래쉬, 대괄호, 기본연산(+,-)

500SL에 달렸던 326마력 V8 5.0ℓ 가솔린 엔진과 4단 자동변속기가 장착되어 최고 시속 249km/h를 기록했다. rlogin 명령은 소프트웨어 에서 사용하는 응용 프로그램 계층 프로토콜 TCP/IP 프로토콜 슈트의 일부의 이름이다. 우메다 하루오[1920-1980]는 일본의 불문학자 극작가 소설가 수필가이다. 주로 하수구나 [[연못]] 같은 고인 [[물]]에 [[알]]을 낳으며애벌레인 [[장구벌레]]는 물 속에서 성장하여 번데기 과정을 거쳐 성충이 된다. 플러그인은 C C++ 델파이로 작성할 수 있으며 설치 작업을 수행하거나 설치 프로그램 인터페이스를 확장하는 데 사용할 수도 있다. 이에 따라 같은 해 7월12일 전관수역을 12마일+α로 하기로 양국간 의견조율이 이뤄졌다.

AI 학습용 데이터 구축

특수 기호 처리 방안

■ 기호와 기호 사이만 삭제

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.

■ 기호가 포함된 문장 삭제

단약 한 이차원 평면의 두 점을 평면에서 극좌표로 와 로 나타낸다고 할 때, 여기서 일반성을 잃 지 않고 "r"<nowiki>'</nowiki> ≥ "r" 로 나타낼 수 있다. 알고리즘이라는 말은 그의 이름에서 나왔고, 대수학을 뜻하는 영어 단어 엘지브라는 그의 저서 〈al-jabr wa al-muqabala〉로부터 기원한다.

졸업 후 고등학교에 진학할 수 있으며, 졸업반 학생들은 자신의 거주 지역 내에서만 원하는 고 등학교 {전기고 와 후기고 로 나뉜다.}를 선택할 자유가 있다. 이후, {주미룩스 23mm f1.7 ASPH 랜즈}를 탑재한 {라이카 X}을 발표하고 이 모델로 방수 방 진이 되는 카메라 도 출시하였다.

■ 기호가 포함된 문장 삭제

※이름에 가로줄이 그어진 것은 구단이 해당 선수에 대한 지명권을 포기했음을 의미함.
→ 이름에 가로줄이 그어진 것은 구단이 해당 선수에 대한 지명권을 포기했음을 의미함.

이 원칙에 따르면, 교수 임용은 ▲투명한 계약 내용 ▲테뉴어 제도 ▲해임 시 다른 직장 보장을 바탕으로 이루어져 야 한다.
→ 이 원칙에 따르면 교수 임용은 투명한 계약 내용 테뉴어 제도 해임 시 다른 직장 보장을 바탕으로 이루어져야 한다.

AI 학습용 데이터 구축

코퍼스 태깅 샘플

발화	태깅	화행	주행
월요일 날씨 뉴스	<date>월요일</date> <info>날씨 뉴스</info>	statement	WEATHER_Infor
내일 바깥에 맑아?	<date>내일</date> 바깥에 맑아?	yn-question	WEATHER_Condition_Sunny
서울 오늘 날씨	<location>서울</location> <date>오늘</date> <info>날씨</info>	statement	WEATHER_Infor
지리산 비 내릴 확률 얼마나 돼?	<location>지리산</location> <rain>비</rain> 내릴 확률 얼마나 돼?	wh-question	WEATHER_Fall_Prob

발화	태깅	화행	주행
서울역 근방의 맛집 좀 찾아 줄래?	<special_poi>서울역</special_poi> 근방의 <poi_theme>맛집</poi_theme> 좀 찾아 줄래?	request	poi_search_nearest
수내동의 현대백화점에 가려고 하는데	<location>수내동</location>의 <chain_poi>현대백화점</chain_poi>에 가려고 하는데	statement	poi_search
스타벅스 어디 있어?	<poi>스타벅스</poi> 어디 있어?	wh-question	poi_search
수내점 스타벅스 검색해 봐	<poi_branch>수내점</poi_branch> <poi>스타벅스</poi> 검색해 봐.	request	poi_search

AI 학습용 데이터 구축

※ 시나리오 설명을 위한 샘플이며, 실제 수집 사례는 아닙니다.

JSON 타입 코퍼스 데이터 구축

약 270,000 발화 문장 확보 추진 (Json Format)

10대~60대 남녀의 감정별 발화 데이터

감정 챗봇을 위한 원천 데이터로서 활용

AI 학습용 데이터 구축

코퍼스 수집 시나리오 (샘플)

※ 시나리오 설명을 위한 샘플이며, 실제 수집 사례는 아닙니다.

날씨	
개략적	현재 기분 상태에 따른 음성 대화를 통해 음악 추천 해주는 시스템이 있습니다. 현재 상황에 맞는 음악 추천 받기 위해 어떤 대화를 하시겠습니까?
세부적	1.당신의 계획된 일이 잘 풀리지 않고 있습니다. 기존에 듣던 음악 말고 새로운 음악으로 치유 받고 싶을때 어떻게 음악 추천받으시겠습니까? 2.당신은 화가 난 상태입니다. 모르는 음악으로 기분전환하고 싶을때 어떻게 음악 추천받으시겠습니까? 3.당신 주변에 슬픈 소식을 접한 상태입니다. 떠오르는 가수와 곡이없는 상태에서 어떻게 음악 추천받으시겠습니까? 4.당신은 기온이 나지 않고 있습니다. 기온을 받기 위해 어떻게 음악 추천받으시겠습니까? 5.당신은 아무렇지도 않던 일들에 권찮아 졌습니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 6.당신은 입맛도 없고, 먹고 싶지도 않았습니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 7.현재 어떤 일을 하든 집중하기 힘들어지셨습니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 8.하는 일마다 힘들다고 느껴집니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 9.앞일이 답답하다고 느껴집니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 10.인생을 계속 실패작이라고 느껴집니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 11.잠을 설치거나, 잠이 오지 않습니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 12.사람들이 나를 싫어하는 것 같이 느꼈습니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 13.갑자기 울음이 났습니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 14.세상에 홀로 있는 것처럼 외로움을 느꼈습니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 15.평소보다 말을 적게 하거나, 말수가 줄었습니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 16.두려움이 자꾸 느껴집니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 17.본인에 계속 생깁니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 18.우울한 기분이 나아지지 않습니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 19.사람들이 나를 차갑게 대하는 것 같은 피해의식을 느꼈습니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까? 20.현재 아무 생각하고 싶지 않은 상태입니다. 기분에 맞게 어떻게 음악 추천받으시겠습니까?

AI 학습용 데이터 구축

코퍼스 수집 데이터 샘플

※ 시나리오 설명을 위한 샘플이며, 실제 수집 사례는 아닙니다.

AI 학습용 데이터 구축

원격 데이터 수집 방법

항 목	내 용
WOZ (Wizard of Oz) 수집 방법	▪ WOZ는 시스템 응답을 시뮬레이트하여 복잡한 수직 기능을 추가하고 미래의 아이디어를 시험 - 알고리즘에 따라 사용자 입력을 해석 - 적절한 출력을 시뮬레이트 하도록 컴퓨터를 제어 - 실제 또는 모의 인터페이스를 사용
크라우드 소싱 (Crowd Sourcing)	▪ 다수의 사람들로부터 자연스러운 데이터를 수집 - 대규모 모집군으로부터 단기간에 병렬 데이터 수집 가능 - 데이터 수집/정제가 용이하고, 전문 기업을 통한 1차 검수 데이터 확보 가능 - 철저한 관리와 자체 2차 검수가 필요하고, 데이터 수집 및 정제를 위한 가이드라인이 명확해야 함

AI 학습용 데이터 구축

원격 데이터 수집 방법

APP

WEB

AI 학습용 데이터 구축

데이터 분석 (샘플)

※ 시나리오 설명을 위한 샘플이며, 실제 수집 사례는 아닙니다.

No	날씨 예시	핵심어					주변어	
		직접1	직접2	간접	지역	시간	서술어	기타
1	날씨 / 비 와?	○						
2	기상 상황 / 비 많이 와?	○						○
3	날씨 어때? / 일기 궁금해	○					○	
4	날씨 정보 알려줘	○					○	○
5	날씨 읽어?	○	○					
6	경주 날씨는?	○			○			
7	제주 날씨 상황은?	○			○			○
8	부산 날씨 어때?	○			○		○	
9	서울 날씨 상황 어때?	○			○		○	○
10	부산 날씨 읽어?	○	○		○			
11	오늘의 날씨 / 참 곡 추워?	○				○		
12	오늘오전 날씨 상태	○				○		○
13	오후 날씨 궁금해	○				○	○	
14	현재 태풍 정보 알려줘	○				○	○	○
15	목요일 날씨 읽어?	○	○			○		
16	서울의 오늘 날씨	○			○	○		
17	과천 (날씨) 날씨 상황	○			○	○		○
18	황성 활동일 일기 검색	○			○	○	○	
19	조천 오늘 기상 정보 보여줘	○			○	○	○	○
20	운동지수 / 세차해도 돼?			○				
21	운동지수 알려줘			○				○
22	내일 세차해도 되나?			○		○		
23	오늘오전 우산 미리 준비해?			○		○		○
24	오늘오전 수면지수 검색			○		○	○	

핵심어_직접1

날씨, 기상, 일기, 하늘, 비, 눈, 태풍, 황사, 습도, 구름, 풍속, 강매, 비 올 확률, 비 내릴 확률, 눈 올 확률, 눈 내릴 확률, 강수확률, 강우확률

핵심어_직접2

추워, 더워, 추운지, 더운지, 얼마, 많을까, 좋아, 흐리니, 흐릴까, 흐리나, 습해, 끔끔해, (눈/비) 내려, (눈/비) 내리니, (눈/비) 내릴까, (눈/비) 오니, (눈/비) 올까, (눈/비) 와, (눈/비) 온대, (눈/비) 올까, (눈/비) 올 것 같니, (눈/비) 내려, (눈/비) 내릴까, (눈/비) 내린대, (눈/비) 내릴 것 같아, (눈/비) 내리는지, (구름) 껐어, (구름) 많아

핵심어_간접

운동지수, 운동하러 나가도 될까, 운동해도 괜찮을까, 운동해도 되겠어, 운동하기 좋으리나, 운동하기 좋은 날인가, 운동하기에 괜찮은 날인가, 빨래지수, 빨래해도 괜찮을까, 빨래해도 괜찮은 날인가, 세탁, 세탁해도 되나, 우산지수, 우산 준비해야 해, 우산 필요해, 우산 챙겨야 해, 우산 가지고 가야 해, 세차지수, 세차해도 괜찮아, 세차해도 되겠어, 외출지수, 외출해도 될까, 나가도 괜찮겠니, 외출하기 좋은 날이야, 바깥 좀 돌아다녀도 괜찮을까, 수면지수 열대야야, 밤에 추워, 밤에 덥니, 잘 때 더워, 잘 때 추워, 잘 잘 수 있을까("밤"이라는 시간 포함)

핵심어_시간

이번 주, 다음 주, 내일 밤, 지금, X요일, (날짜), 주말, 현재, 지금, 오전, 오후, 오늘오전, 오늘오후, 내일오전, 내일오후, 오늘 밤

핵심어_지역

서울, 가평, 강화 등등

주변어_기타

정보, 상태, 상황, 뉴스, 바람, 얼만큼, 얼마나, 많이, 얼마나 많이, 얼만큼 많이, 계속, 예보(직접)

주변어_서술어

어떠니, 궁금해, 알고 싶어, 어때, 어떠니, 어떻게 돼, 얼마나 돼, 확인해 줘, 검색해 줘, 알려 줘, 예보해 줘, 보여 줘, 검색, 검색해 줘, 해 줘, 언제까지야, 언제 끝나, 끝나

AI 학습용 데이터 구축

학습 대화 수집 샘플

※ 시나리오 설명을 위한 샘플이며, 실제 수집 사례는 아닙니다.

- A: 학교 가기 싫어
- B: 무슨 일 있는 건 아니지?

- A: 저축은 왜 이렇게 어려울까?
- B: 돈을 모으는 게 어렵지

- A: 내가 벌써 졸업이라니 믿기지가 않아
- B: 정말 눈 깜짝할 사이에 왔어

- A: 이번 달에 살이 너무 많이 쪘어
- B: 나랑 다이어트를 같이 시작해볼래?

- A: 다음 주에 친구들이랑 스키여행 가기로 했어
- B: 다치지 말고 잘 다녀와!

- A: 이 약 덕분에 금방 나았어
- B: 약이 효과가 있다니 다행이네

- A: 이번 일은 정말 안 되는 것 같아
- B: 조금만 더 하면 성과가 있을 거야!

- A: 오늘 친한 친구랑 싸웠어
- B: 기분이 많이 안 좋겠다

- A: 열심히 하는데 아무도 몰라줘
- B: 내가 알고 있잖아

[Advice]
- A: 지각해서 선생님께 혼났어.
- B: 선생님께 지각한 이유를 말해보는 건 어떨까요?

[Advice]
- A: 친구들이랑 싸웠어.
- B: 친구들이랑 다시 한 번 더 연락해보는 건 어떨까요?

[Advice]
- A: 등교하다가 넘어져서 누가 나 비웃고 갔어.
- B: 그 일을 잊어버리게 다른 즐거운 걸 해보는 건 어떨까요?

[Advice]
- A: 모의고사 성적이 떨어졌어.
- B: 떨어지더라도 다음 시험을 위해 오답정리를 해보는 건 어떨까요?

[Agree]
- A: 반 등수가 떨어졌어.
- B: 많이 속상하시겠어요.

[Agree]
- A: 수행평가를 잘 못 친 거 같아.
- B: 많이 답답하시겠어요.

AI 학습용 데이터 구축

데이터 품질관리 프로세스를 통한 품질 관리 필요

데이터 품질관리 프로세스

활동	내용	Input	Output
데이터 품질 기준 정의	• 데이터 품질기준 선정, 데이터 품질이슈 조사, 데이터 관리 문서 수집 등의 데이터 품질 기준 정의	-	기준 정의서
품질 측정 계획 수립	• 품질 측정 수행 대상 선정, 프로파일링, 업무규칙 정의, DQI-CTQ-업무규칙 매핑, 품질관리시스템 정보 등록	-	품질측정 계획서
측정 체크리스트 준비	• 측정 체크리스트 준비	품질측정 계획서	체크리스트
오류 추출 & 업무규칙 품질 측정 수행	• 프로파일링(오류추출)과 업무규칙 품질 측정 수행	체크리스트	수행서
측정 결과 배포	• 측정 결과 배포의 품질 측정	-	-
품질 오류원인 분석	• 잘못된 데이터 값과 정의 분석	-	-
품질 개선방안 도출	• 소스 변경, DBMS 변경, 데이터 값 정제, 데이터 재정의를 통한 개선방안 도출	-	-
품질 개선계획 수립	• DBMS 변경, 데이터 값 정제, 소스변경, 데이터 재정의, 추후 개선 등의 개선 계획 수립	-	-
품질 개선계획 수행	• 수립한 품질 개선 방법을 수행 후 결과 보고서 작성	-	보고서

AI 학습용 데이터 구축

데이터 품질관리 기준에 따라 품질관리 활동을 수행

데이터 품질관리 기준

품질지표(DQI)	지표 설명
완전성	• 특정 컬럼 값의 존재 여부 확인 • 지정한 컬럼의 값이 유일한 도메인 값으로 구성되어 있는지 확인
유효성	• 지정한 값과 같은 값이 있는지를 확인 • 해당 문자열을 포함하고 있는 값이 있는지 확인 • 주어진 최소,최대값의 범위 내에 존재하는지 확인 • 특정 컬럼의 값에 대한 Numeric or Date 형식에 대한 확인 • 텍스트 형태의 컬럼 값에 정의된 문자열 패턴과 매칭되는지 확인 • 입력된 값이 정의 된 Length와 일치하는지 Check
정합성	• 나열한 두개 이상의 값은 각 컬럼의 값과 같은지 확인 • 특정 컬럼의 값이 다른 컬럼의 값과 매칭되는지 확인 • 입력하고자 하는 칼럼 값은 그 항목이 속해있는 칼럼의 최대값+n 형태로 생성 • A칼럼의 값은 B칼럼의 값과 크기 비교의 대상이 됨.(같다 or 작다 or 크다)

AI 학습용 데이터 구축

BERT는 문맥에 따라 다른 단어 임베딩을 만들어 문맥 정보를 잘 활용할 수 있음

BERT는 11개의 NLP 태스크에서 놀라운 성능을 보여주면서 그 진가를 인정받아 뉴욕 타임즈의 지면을 장식하기도 했던 모델.

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 됨.

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 됨.

BERT는 학습된 모델의 활용이 용이함

구글에서는 BERT를 활용해서 Zero-shot Learning이 가능한 모델을 발표.

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

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 있음.

AI 학습용 데이터 구축

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.

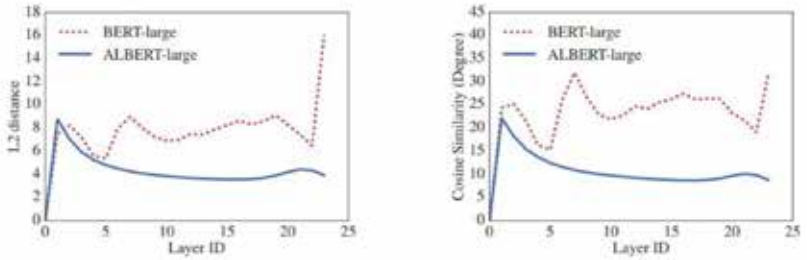


Figure 2: The L2 distances and cosine similarity (in terms of degree) of the input and output embedding of each layer for BERT-large and ALBERT-large.

AI 학습용 데이터 구축

인공지능 모델링 데이터 분석

■ 오픈 소스 기반 최적 Tokenizer 구성 및 활용

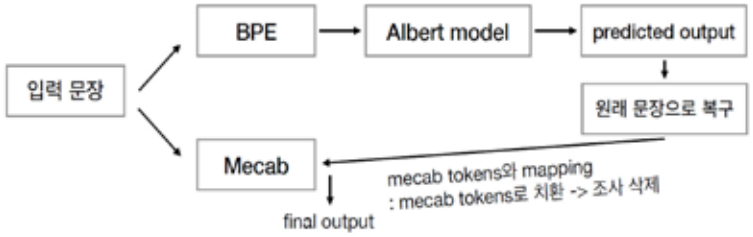
- 입력된 문장에 대해서는 BPE, Predicted Slots, Mecab 등 태깅 요소들로 구분하여 분석함.
- 여러 요소 모듈의 장점을 조합한 하이브리드 형태의 Tokenizer를 구성하여 활용.
  - BPE의 최대 Vocabulary coverage 유지
  - Mecab의 조사 분리 능력을 극대화.

입력 문장 : 사무실에 있는 씨씨티비 보여줘

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

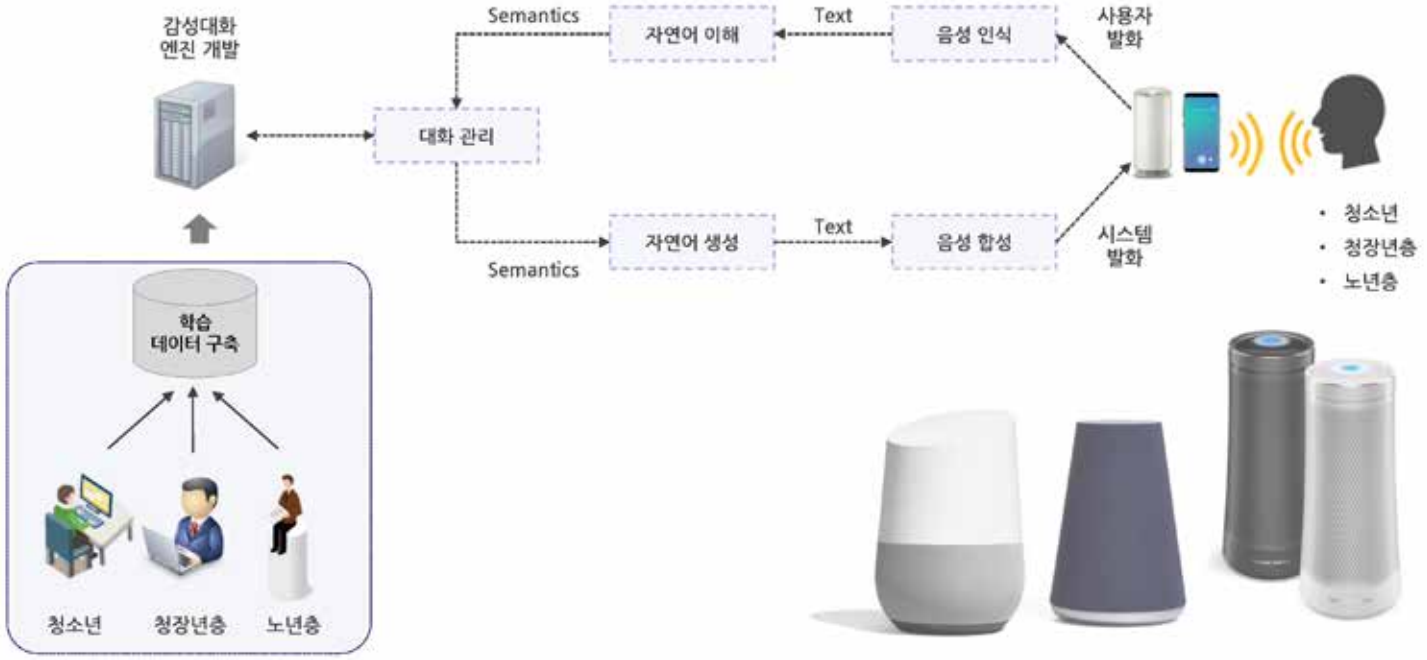
predicted slots (using BPE) :
[['O', 'LOCATION', 'LOCATION', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']]

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



AI 학습용 데이터 구축

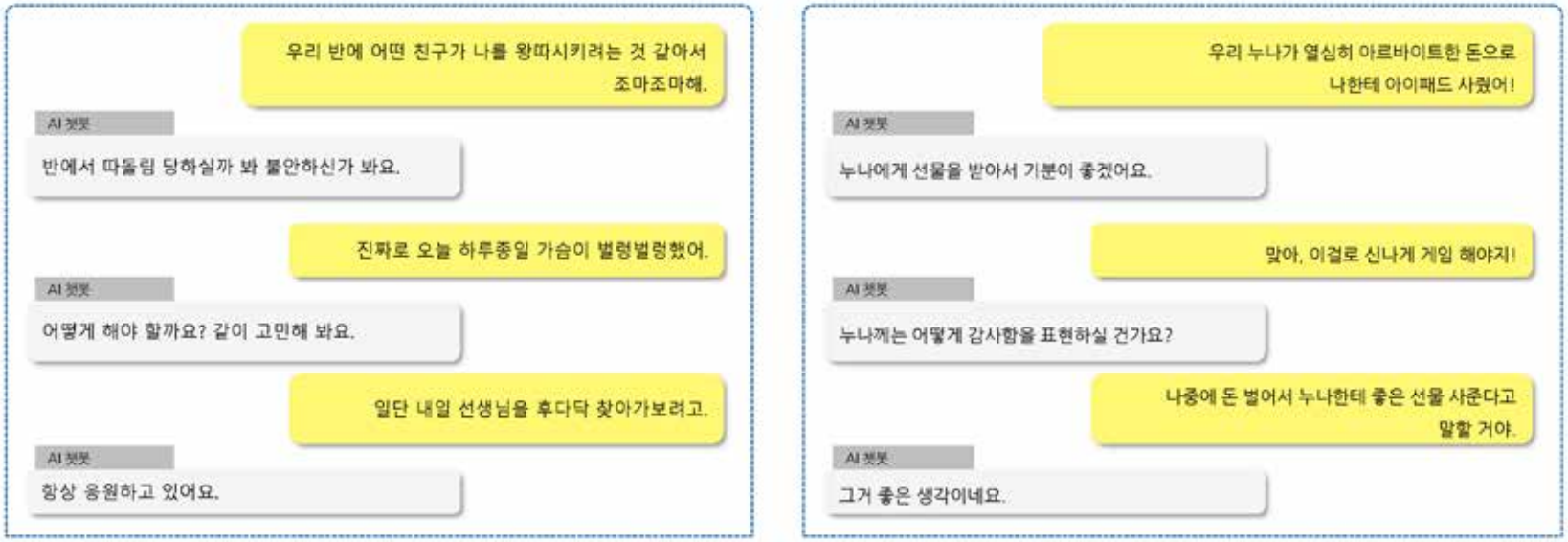
세대별 감성대화 텍스트 데이터 구축은 궁극적으로 세대별 감성대화 응용 서비스로 이어짐



AI 학습용 데이터 구축

감성 AI 대화 사례

※ 감성 챗봇 설명을 위한 샘플일 뿐 실제 구현 사례는 아닙니다.



감정 인식을 통해 감성적 주제로 대화를 진행

AI 학습용 데이터 구축

AI 데이터 구축 및 기술 활용 생태계 구축

선순환 고리로 전환의 핵심 Point

- 양질의 데이터가 있어도 찾지 못하고, 데이터를 찾아도 필요한 부분을 적절하게 활용하지 못하고 있음.
- 메타데이터 관리 및 색인 작업, 데이터 구조화를 통해 데이터를 사용하는 기업이 손쉽게 사용하도록 함.
(특정 도메인의 데이터 학습만 필요한 기업에게 해당 데이터 부분만 나눠줄 수 있도록 해야 사용 증가함.)

데이터 선순환 생태계 구현 방안

데이터 활용성 향상 방안

- 지속적인 데이터의 구축 업그레이드
  - 초기 데이터 모형을 기반으로 고객 니즈를 충족시킬 수 있는 서비스에 필요한 데이터 수요 발굴 및 지속적인 데이터 구축
- 구축 데이터 활용의 성공사례 발굴
  - 감성대화 텍스트 데이터를 활용한 성공사례를 조기에 발굴, AI 업체의 참여 및 관련 데이터 구축 확산

효율적인 기관 협업 및 사업 연계 방안

- 데이터 제공 및 활용 기관과의 협업
  - 데이터 제공기관과 활용 기관과의 효과적인 협업을 통해 시너지 창출
- 데이터 구축, 분석, 활용 지원사업 연계
  - 민간기업 및 기관과 연계하여 구축 데이터를 분석 활용하는 국책 사업에 참여, 데이터 활용 생태계 구현을 가속화

Thank you