●○ 요약 데이터 과제

한국어 대화 요약 데이터

●○ 개요: 한국어 대화 요약 데이터셋이란?

한국어 대화 요약 데이터셋은 자연언어처리(NLP, Natural Language Processing) 분야에서 아직 걸음마 단계에 있는 대화 요약(dialogue summarization) 기술의 연구 개발에 활용할 수 있는 학습 데이터셋으로 ㈜바이브컴퍼니에서 구축하였으며, 총 35만건의 SNS 대화 원문과 요약문으로 구성되어 있다.

대화 텍스트는 다수의 대화 참여자들 간의 상호작용을 통해 생성되며, 비대면 상황에서 컴퓨터를 매개로 하여 이루어지는 SNS 대화의 경우에도 구어적 속성을 지녔으므로 문장 성분의 생략, 축약적 표현의 사용 등이 빈번히 이루진다. 따라서 대화 텍스트는 문어 중심으로 개발된 기존의 자연어처리 기술의 적용에 한계가 있어, 특화된 데이터셋의 구축이 필수적으로 요청된다.

코로나바이러스감염증-19 사태가 장기화되면서 비대면 의사소통의 중요성은 더욱 부각되고 있으며, 대규모의 대화 텍스트가 데이터화되고 있다. 대화 텍스트의 가치를 증대하기 위해서는 텍스트 검색에 기반한 전통적인 정보 탐색 기술의 한계를 넘어선 대화 요약 기술의 도입이 요구되고 있으며, 본 데이터셋의 구축과 공개로 관련 기술의 활발한 연구와 개발이 이루어질 수 있을 것으로 기대된다.

●○ 데이터셋의 구성

본 데이터셋은 아래에 설명하는 기준을 충족하는 대화 원문과 이에 대응하는 요약문 35만건으로 구성되어 있다. 본 데이터셋의 규모는 최근 공개된 영어 대화 요약 데이터셋인 SAMSum Corpus이 16,000개 대화로 이루어진 것에 비해 훨씬 더 큰 규모를 지니고 있다.

데이터 종류	포함 내용	제공 방식
대화 요약 데이터셋	대화 원문 및 요약문 35만건	JSON 파일(UTF-8 인코딩)

●○ 데이터셋의 설계 기준과 분포

본 데이터셋은 아래와 같은 기준에 따라 구축되었다.

- 대화의 양적 속성
  - 말차례 수: 말차례(turn-taking)는 대화에 참여하는 화자(speaker)가 전환되는 것을 의미하며 본 데이터셋에서는 대화 한 건당 말차례가 최소 4회 포함되도록 하였다.
  - 발화 수: 발화(utterance)는 구어체 대화에서 문장(sentence)에 해당하는 단위로 본 데이터셋에서는 대화 한 건당 발화가 최소 8회가 포함되도록 하였다.
- 대화의 유형
  - 대화에는 일상 대화, 토론, 회의록, 상담 내역 등 다양한 유형이 있는데, 본 데이터셋에서는 SNS로 이루어지는 가장 일반적인 대화 형태인 일상 대화와 토론 대화, 그리고 질의응답 대화를 구분하였다.
- 대화의 내용
  - 일상 속에서 이루어지는 대화는 다양한 주제를 다루는데, 본 데이터셋은 아래에 보인 주제 분류를 적용하였다.

주제	키워드 예시
개인 및 관계	이름, 전화번호, 가족, 국적, 고향, 성격, 외모, 개인의 기호(선호), 직업, 종교, 반려동물, 연애(관), 결혼(관), 이상형, 인간 관계, SNS
주거와 생활	숙소, 방, 가구, 침구, 주거비, 생활 편의 시설, 지역, 지리, 가전 제품, 자취, 잡안 일, 육아, 부동산, 주거시설, 이사, 생활비, 자동차
상거래(쇼핑)	쇼핑 시설 및 장소, 식품, 의복, 가정용품, 물건 및 가격, 택배, 중고거래, 서비스, 교환 및 환불, 구매 후기
식음료	식사, 음식, 음료, 배달, 외식, 맛집, 식사 메뉴, 야식, 디저트, 요리
공공 서비스	우편, 전화, 통신, 휴대전화, 인터넷 서비스, 은행, 관공서
여가와 오락	휴일, 취미, 동아리 및 동호회 활동, 관심사, 방학, 휴가, 행사, 술, 웹서핑
일과 직업	취업, 스펙, 직장 생활, 업무, 회식, 급여, 계약, 협상, 회의
행사 및 모임	초대, 방문, 소개팅, 약속, 가족 및 친척 행사, 공적 행사, 사적 모임(친목 모임)
미용과 건강	신체, 위생, 부상 및 질병, 치료 및 수술, 보험, 병원, 운동, 미용, 다이어트, 건강 검진, 약품 및 건강 보조 식품(용품)
기후	날씨, 계절
여행	여행 장소 및 경로, 여행 계획(일정, 숙소, 교통편, 여행 경비), 여행팁, 기념품, 여행사 및 여행 상품
교통	위치, 거리, 길, 이동 수단, 이동 경로, 대중교통(지하철, 버스, 택시)
교육	학교 교육, 교과목, 진로, 학원, 진학, 입시, 시험, 자격증, 성적, 자기 계발, 외국어 학습, 스터디
시사, 사회	정체, 경제, 사회, 사건 및 사고, 법과 제도, 여론, 국제 관계, 재해 및 재난
예술, 문화 생활	문학, 음악, 미술, 공연, 전시, 스포츠 관람, 엔터테인먼트
전공/전문 지식	학문 및 학술 분야, 학회 및 세미나

- 개인정보 비식별화
  - 일상적인 대화에 흔히 포함될 수 있는 개인정보에 대하여 아래와 같은 기준을 적용하여 비식별화를 수행하였다.

범주	항목	레이블링	예시
이름	실명	#@이름#	(원문) 소연이 넘 고마오♥ (정제) #@이름# 넘 고마오♥
	실명(변형)		
	특수 애칭, 별명, 대화명, 필명		
	일반 애칭, 별명		자기야, 여보 등 김연아, 빌 게이츠 등
	공인 실명		
온라인	아이디	#@계정#	(원문) sample@sample.com으로 보내 (정제) #@계정#으로 보내
	이메일 주소		
	URL		
각종 번호 및 비밀번호	고유 식별 번호 (주민번호, 학번, 사번 등)	#@신원#	(원문) 응 학번은 200101-1234567 (정제) 응 학번은 #@신원#
	전화번호	#@전번#	(원문) 언니 번호 010-1234-5678이야 (정제) 언니 번호 #@전번#이야
	금융 번호 (계좌, 카드번호 등)	#@금융#	(원문) 신한 110-234-456-789 김연아 (정제) #@금융#
	일련번호	#@번호#	(원문) 사업자등록번호 123-45-67890 (정제) 사업자등록번호 #@번호#
	(구매자) 식별 번호		
	사업자 등록 번호		
	비밀번호		
장소	상세 주소(동 이하)	#@주소#	(원문) 배송지는 서구 연희동 123로요 (정제) 배송지는 서구 #@주소#로요
	거주 아파트 및 건물명		
	거주지 역명 (지하철역, 기차역 등)		도곡역 3번출구로 오세요 연세대 앞 정류장이야 롯데리아에서 만날래?
	방문 장소(비정기적)		
	상호명		
출신 및 소속	출신 및 소속 학교	#@소속#	(원문) 한국대학교에 재학중입니다 (정제) #@소속#에 재학중입니다
	출신 및 소속 직장		
	출신 및 소속 부대		
기타	위에서 언급하지 않은 항목이지만 비식별화 필요	#@기타#	

본 데이터셋의 핵심 구성 요소인 대화 요약문의 작성 기준은 아래와 같다.

- 대화문 요약문 작성 원칙
  아래의 두 가지 조건 중 하나를 충족 하도록 요약한다.
  - 육하원칙 요소 중 2가지 요소 이상을 포함하며 “누가”에 해당하는 내용을 알 수 있으면 한 문장으로 요약한다.
  - 대화문 내용에 등장하는 키워드 2개 이상을 포함하여 한 문장으로 요약한다.
- 대화 유형별 요약 형식
  - 일상 대화: 육하원칙을 바탕으로 요약하며, “ ~ 하다/한다” 형식을 따른다.
  - 토론 대화: 토론하는 주체와 토론 주제를 포함하여 요약하며, “(누가) ~에 대해서 토론한다”의 형식을 따른다.
  - 질응 응답 대화: “무엇”에 대한 질의응답인지 명시적으로 드러나도록 요약하며, “(A가) ~을 질문하고 (B가) 답변한다”의 형식을 따른다.

본 데이터셋의 구성 분포는 다음과 같다.

- 대화 참여자 성별 분포
  - 추후 공개
- 대화 참여자 연령 분포
  - 추후 공개
- 대화 참여자 거주 지역 분포
  - 추후 공개
- 대화 주제 분포
  - 추후 공개

●○ 데이터 구조

본 데이터셋에 포함된 데이터 항목은 메타 정보를 포함하는 헤더와 대화 원문과 요약문으로 구성된 본문으로 구성되었다.

- 헤더(메타 정보)
  - 대화참여자(화자) 수
  - 화자 정보(성별, 연령대, 거주지)
  - 대화 유형(일상 대화, 토론 대화, 질의응답 대화)
  - 대화 주제(16개 주제, 기타 및 다중 분류 가능)
- 본문
  - 대화 원문
    - 발화자ID
    - 발화 발생 및 날짜 및 시간
    - 발화문
  - 대화 요약문

●○ 데이터 예시

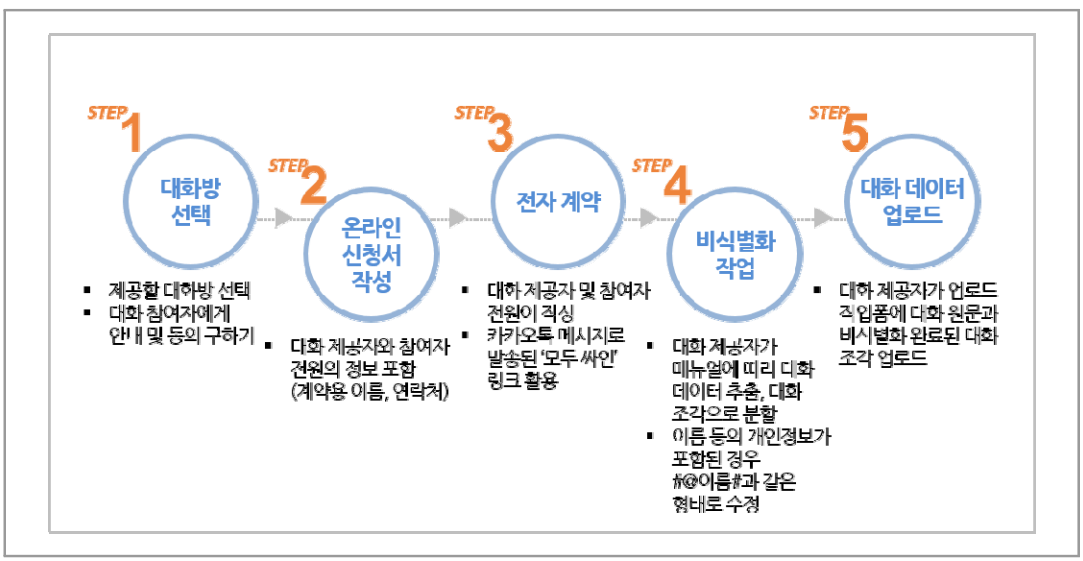
JSON 형식으로 인코딩한 본 데이터셋의 예시를 보이면 아래와 같다.


```
{
    "header": {
        "dialogueID": "D000001",
        "dialogueType": "일상대화",
        "topic": "개인 및 관계",
        "numberOfParticipants": 3,
        "numberOfTurns": 2,
        "numberOfUtterances": 5,
    },
    "body": [
        "utterances": [
            {
                "utteranceID": "U1",
                "participantID": "P1",
                "date": "20200723"
```

```
                    "time": "10:52",
                    "text": "내일이 #@MASK# 엄마 생일임.",
                {
                    "utteranceID": "U2",
                    "participantID": "P1",
                    "date": "20200723"
                    "time": "10:52",
                    "text": "저녁에 외식할까 함."
                },
                {
                    "utteranceID": "U3",
                    "participantID": "P1",
                    "date": "20200723"
                    "time": "10:52",
                    "text": "시간 ㅇㅋ?"
                },
                {
                    "utteranceID": "U4",
                    "participantID": "P2",
                    "date": "20200723"
                    "time": "10:53",
                    "text": "강의만 제때 끝나면 저녁에 시간돼요",
                },
                {
                    "utteranceID": "U5",
                    "participantID": "P3",
                    "date": "20200723"
                    "time": "11:15",
                    "text": "저도 가능해요",
                }
            ]
        },
        "summary": "가족 구성원들이 내일 저녁 엄마 생일 외식 시간에 대해 대화한다"
    }
```

●○ 데이터 구축 과정

본 데이터셋은 2020년 10월부터 크라우드 소싱에 의해 카카오톡 대화를 수집하고, 정제, 가공 및 검수의 과정을 거쳐 구축되었다. 구축 초기에는 온라인 설문 도구와 스프레드시트를 이용하였으며, 후반기에는 한국어 대화 요약 데이터셋에 특화된 데이터 저작 도구를 이용하였고, 개인정보처리 동의와 저작권활용계약 체결은 온라인 서명 플랫폼을 활용하였다.



수집 · 정제된 데이터는 자동화된 형식 요건 및 중복 검사를 거친 후 교차 검수에 의한 내용 검수, 요약, 샘플링에 의한 종합 평가 검수를 거쳐 최종적으로 데이터셋에 포함되었다.

●○ 데이터 구축 담당자

수행기관(주관): ㈜바이브컴퍼니
(전화: 02-565-0531, 이메일: biz@vaiv.kr)