AI에 적합한 일반상식 문장의 자동 생성을 위한 정량적, 정성적 연구

신현규, 송영숙[0]
서울대학교, 경희대학교,
realo0404@gmail.com, klanguage1004@gmail.com

CommonAI: Quantitative and qualitative analysis for automatic-generation of Commonsense Reasoning sentence suitable for AI

Hyeon Gyu Shin, YoungSook-Song[0]
Seoul national university, KyungHee university

요 약

본 논문에서는 인공지능이 생성하는 일상 대화의 품질 향상을 위해 상식 추론을 정의하고 설문을 통해 정량적, 정성적 분석을 진행하였다. 정량적 평가에서는 '주어진 문장이 AI에게 학습시키기에 적합한가'라는 수용성 판단을 요청한 질문에서 40대 이상의 연령이 20, 30대와 유의미한 차이를 보였다. 정성적 평가에서는 '보편적 사실 여부'를 AI 발화 기준의 주요한 지표로 보았다. 이어서 '챗봇' 대화의 품질에 대한 설문을 실시했다. 이를 통해 일상 대화를 사용한 AI 챗봇의 대화 품질을 높이기 위해서는 먼저, 질문의 요구에 적절한 정보와 공감을 제공해야 하고 두 번째로 공감의 정도가 챗봇의 특성에 맞는 응답이어야 하며 세 번째로 대화의 차례에 따라 담화의 규칙을 지키면서 대화가 진행되어야 한다는 결론을 얻을 수 있었다. 이 세 가지 요건이 통합적으로 적용된 담화 설계를 통해 완전히 인공지능스러운 대화가 가능할 것으로 여겨진다.

주제어: AI, 챗봇, 상식 추론, 일상대화, 대화 분석, 오픈 도메인 대화 시스템

1. 서론

상식적인 내용으로 사고를 하고 상식적으로 조합 가능한 문장을 산출하는 것은 인공지능 생성 모델의 주요한 기능 중에 하나다. 그런데 기능에 맞게 설계되었는지와 설계된 대로 작동하고 있는지 확인하기 위해서는 무엇이 인공지능에게 요구되는 상식인가를 다루는 상식의 개념과 범위가 명확해야 할 것이다. 또한 이를 통해 한국어에서 상식적인 내용을 담은 문장이 챗봇과 같은 대화 시스템에 적용되기 위해서는 무엇을 고려해야 하는지에 대한 순차적 담화 설계가 필요하다. 따라서 본 논문에서는 일반상식 대화 분석을 통해 구축한 데이터가 챗봇 등의 인공지능 모델에 대화 품질에 기여하는지를 정량적, 정성적으로 분석하여 오픈 도메인 대화 시스템(open domain dialogue system) 또는 인공지능 챗봇의 대화 품질 향상에 기여코자 한다. 먼저, 문제 정의를 위해 그간 일반 상식 추론에서 다룬 문제를 종합해 한국어에 적합한 상식 추론이란 무엇인지 다룬다. 이를 위하여 상식 추론의 영역을 다룬 국내외 선행 연구를 종합해 정리하고 일반적으로 상식 추론을 정의한 후 상식 추론에 적합하다고 생각하는 문장을 설문을 통해 정량적, 정성적으로 분석한다.

이제까지 일반상식은 모든 사람이 공유하는 가장 일반적이고 널리 적용 가능한 일상생활에 대한 지식으로 이해되었다[1]. 또한 보통 사람들에게는 '상식'이라는 용어가 '잘된 판단(good judgement)'과 동의어로 여겨지나, 인공지능 커뮤니티에서는 수백만 가지 기본적 사실과 이해를 가리키는 기술적 의미로 사용되고 있고 그 대상에는 일상의 공간적(spatial), 물리적(physical), 사회적(social), 시간적(temporal), 심리적(psychological) 측면이 포함된다[1].

인공지능에서 다루는 일반상식의 분류에 대한 좀 더 분화된 논의도 있다[2][3]. 특히 [2]에서는 그간의 이미지 캡셔닝에서 다루어 왔던 문장들은 "완전한 인공지능(AI-complete)스럽지 않다"고 비판했다. 따라서 "완전한 인공지능"에 가깝기 위해서는 다음과 같은 열린 질문에 자연스럽게 답할 수 있어야 한다고 주장한다.

표 1. 질문 유형 분류

세밀한 인식	이 피자엔 어떤 종류의 치즈가 있나요?
물체 감지	얼마나 많은 수의 자전거가 있나요?
행동 인식	남자가 울고 있나요?
지식기반 추론	이것은 채식주의자용 피자인가요?
상식 추론	**이 사람은 시력이 20/20인가요?", "이 사람은 일행을 기다리고 있나요?**

위 분류에서 "세밀한 인식, 물체 감지, 행동 인식"은 사물을 보고 인식하는 영역이라면 "지식 기반 추론과 상식 추론"은 인식한 것을 기반으로 보편타당한 추론을

[0] 교신저자(Corresponding author)

이끌어내는 능력을 의미한다고 할 수 있다. 즉 질문의 유형에서 일상의 객체를 감지하고 인식하는 것과 가장 그럴듯한 판단을 이끌어내는 영역을 구분하여야 하고 이로부터 납득 가능한 문장을 생성할 수 있어야 한다. 위의 연구 [2]는 양질의 데이터를 선별하기 위한 구분 책략도 밝히고 있다.

표 2. 요구 답변의 특성에 따른 질문 분류

지나치게 쉬운 질문	저 고양이는 무슨 색이야? 지금 이 장면에는 몇 개의 의자가 있어요?
상식적 정보만을 이용해서 대답할 수 있는 질문	수염은 어떤 구성 성분으로 이루어져 있나요?
상식적 정보뿐 아니라 상식적 지식이 필요한 질문	사진의 동물이 어떤 소리를 낼 것 같아요?

상식적 지식이 필요한 질문을 이끌어내기 위해서는 지나치게 쉬운 질문이나 상식적 정보만으로 대답하는 것이 아닌, 일상생활에서 두루 쓰이는 용어 및 문장을 이해하고 이로부터 기대되는 감정과 상황 변화에 대한 보편적 지식을 산출해 냈는가를 중요한 요인이 될 것이다. 그런데 한국어에서는 아직 일상생활 중요도(life-essentiality)를 반영한 어휘 특성이나 도메인 등에 대한 정의가 구체적으로 논의되지 않았다. 따라서 2장에서는 일반상식의 도메인으로 사용할 수 있는 일상생활의 세부 분야를 분류하고자 한다.

2. 한국어 일반상식 데이터의 특성

2.1. 도메인 특성

일반 상식 추론 데이터로는 그림 추론, 일상생활 데이터의 논리적, 인과적 추론, 질의응답 등이 구축되었으나 그 범주가 구체화되어 있지는 않다. 가령, MS-COCO[12] 데이터 세트의 경우 일상생활에서 흔히 볼 수 있는 대상(object)에 대한 기본적인 캡셔닝 정보를 제공하고 있을 뿐이다. 따라서 따라서 본 논문에서는 인공지능 분야는 아니지만 외국어 교육에서의 범주와 일상 대화 말뭉치의 도메인 세분화를 위해 구분한 범주를 비교하여 살펴보고자 한다. 아래 표3은 AI HUB에 공개되어 있는 한국어 대화 요약[4]과 모두의 말뭉치에 공개되어 있는 일상 대화 말뭉치2020[14] 그리고 2017년 국제 통용 한국어 표준 교육과정 적용 연구에서 제시된 유럽 공통 참조와 국제 통용 말뭉치의 목록을 정리한 것이다.

표 3. 일상대화 대분류와 세분류

AI HUB 한국어 대화 요약	국립국어원 일상 대화 말뭉치 2020	유럽 공통 참조 (14범주)	국제 통용 2단계 (17범주)	2017년 국제 통용 3단계 (17범주)	국제 통용 3단계 항목(85개)
미용과 건강	건강/다이어트	건강	건강	건강	신체, 위생, 질병, 치료, 보험
교육	회사/학교/아르바이트	교육	교육	교육	학교 교육, 교과목, 진로
일과 직업			일과 직업	일과 직업	취업, 직장 생활, 업무
주거와 생활	반려동물	일상생활	생활	일상생활	가정생활, 학교생활
상거래 (쇼핑)	선물	쇼핑	물건 사기	쇼핑	쇼핑 시설, 식품, 의복, 가정용품, 가격
	가족/반려동물	가족, 집, 환경	가족	주거와 환경	장소, 숙소, 방, 가구·침구, 주거비, 생활 편의 시설, 지역, 지리, 동식물
개인 및 관계	성격/연애/결혼	개인 신상	취미와 여가	개인 신상	이름, 전화번호, 가족, 국적, 고향, 성격, 외모, 연애, 결혼, 직업, 종교
여가와 오락	방송/연예/영화/스포츠/레저	여가, 오락	휴일 / 책 과 문학	여가와 오락	휴일, 취미 · 관심, 라디오·텔레비전, 영화· 공연, 전시회 · 박물관, 독서, 스포츠
여행	여행지(국내/해외)	여행		여행	관광지, 일정, 집, 숙소
행사 및 모임		대인관계		대인관계	친구 · 동료 · 선후배 관계, 초대, 방문, 편지, 모임
식음료	먹거리	식음료		식음료	음식, 음료, 배달, 외식
		(공공)서비스		공공 서비스	우편, 전화, 은행, 병원, 약국, 경찰서
전공전문 지식			전문 분야	전문 분야	언어학, 과학, 심리학, 철학
기후	계절/날씨	날씨		기후	날씨, 계절
예술 문화생활			예술	예술	문학, 음악, 미술
시사, 사회			뉴스, 시사 문제/미디어	사회	정치, 경제, 범죄, 제도, 여론, 국제 관계
교통				교통	위치, 거리, 길, 교통수단, 운송, 택배
공공 서비스	꿈(목표)	언어/장소	한국과 한국어 /인사/ 수과학		

표 3을 통해서 범주화 시기와 쓰임이 다름에도 불구하고 각 데이터의 분류가 상당 부분 일치함을 확인할 수 있다. 특히 건강, 교육, 직업, 주거, 가족, 식음료, 여가, 행사, 기후 등은 일상생활을 토대로 구축된 범주의 핵심 구성 항목이었다. 본 논문에서는 위의 주요 도메인별 상식 문장의 항목을 중심으로 일상생활 중요도(life-essentiality)를 파악했다.

2.2. 어휘 빈도

빈도를 중심으로 일상생활 중요도(life-essentiality)를 파악하기 위해서는 먼저, 한국어 대화 요약 데이터의 어휘를 품사별[1]로 추출하였다. 다음으로는 특정 도메인에는 고빈도로 등장하지만 다른 도메인에서는 상대적으로 저빈도인 명사의 빈도를 추출하고, 이를 다시 의식주와 연결되는 필수적인 키워드인 밥, 집, 옷과 상대적으로 덜 필수적으로 여겨지는 게임, 다이어트, 날씨 등으로 분류하였다. 추출된 도메인별 키워드의 빈도는 다음과 같다.

표 4. 도메인별 키워드 분포표

	개인 및 관계	미용과건강	상거래	시사교육	식음료	여가생활	일과직업	합
밥	1439	107	56	52	1664	195	246	3759
집	3274	212	473	294	2206	745	509	7713
날씨	80	18	13	4	42	55	14	226
옷	582	77	839	11	13	283	59	1864
게임	199	2	27	16	8	1518	28	1798
다이어트	146	369	29	5	259	37	4	849

필수 영역: 밥, 집, 옷 / 주변 영역: 날씨, 게임, 다이어트

위의 표4에서 추출된 어휘가 포함된 문장의 일상생활 중요도를 살펴보고 이러한 차이가 문장 내적인 것인지 아니면 사회적인 차이도 포함하고 있는지를 확인하기 위

1) 키워드 추출을 위한 형태소 분석기는 baikalnlpy를 사용했다.(URL: https://pypi.org/project/baikalnlpy/)

해 한국어 대화 요약 데이터[4]를 사용해 성별, 연령별 '일반상식' 수용성을 묻는 설문을 수행했다.

3. 상식적인 문장 테스트

한국어 대화 요약 데이터는 한국어 대화 데이터에서 대화 주제 분류와 생성 요약문 어노테이션을 한 후에 AI 모델링을 통해 데이터의 유효성을 검증한 데이터이다.

수용성 설문에 참여한 사람들은 모두 국어학 또는 한국어 교육학 등 관련 분야 전공자들로 한국지능정보사회진흥원의 인공지능 학습용 데이터 구축사업 '15-1 일반상식 문장 교정 데이터'에서 상식 문장을 정제 및 검수한 크라우드 워커들이다. 20대, 30대, 40대 이상의 남녀 각 6명, 총 36명에게 6개 영역에서 10문장씩, 총 60문장에 대한 판단을 물었다.

조사는 구글 설문지를 이용했고, 문항은 '주어진 문장이 AI에게 학습시키기에 적합한가'의 수용성을 T/F의 이진 척도로 묻는 객관식 필수 문항과 판단 근거를 묻는 주관식 선택 문항으로 구성했다. 문장은 필수 영역과 주변 영역에 속하는 문장들을 무작위로 섞어 제시했다.

3.1. 정량적 분석 결과

정량적 분석에 사용하는 객관 문항에서는 60문항에 대한 36명의 답변 결과(2,160개)를 분석한다.

3.1.1. 연령별 차이

일상생활 중요도에 따라 도메인을 필수/주변 영역으로 양분할 때, 20대는 필수 영역에서의 긍정 답변이 269개, 부정 답변이 91개, 주변 영역에 대한 답변은 긍정 258개, 부정 102개였다. 30대 역시 필수 영역에서는 긍정 255개, 부정 75개, 주변 영역은 각각 230개, 100개였다. 반면에 40대 이상의 답변에서는 필수 영역에 대한 긍정 답변 197개, 부정 답변 163개, 주변 영역에 대한 긍정 답변이 208개, 부정 답변이 152개로 주변 영역에서의 긍정 답변이 더 많았고, 전체적으로 부정 응답의 비율이 높았다.

필수 및 주변 영역에서의 응답과 연령의 상관관계를 유의수준 0.05로 검증했을 때, 유의 확률이 0.00에 수렴하여, 세대 변인과 응답 사이에 상관관계를 확인했다.

표 5. 연령별 긍정 응답 수와 비율

	필수 영역				주변 영역				총계
	밥	집	옷	계	날씨	게임	다이어트	계	
20	94	89	86	**269**	106	69	83	**258**	**527 (73%)**
30	93	94	88	**275**	95	72	88	**255**	**530 (74%)**
40 이상	67	67	63	**197**	82	50	76	**208**	**405 (56%)**

다음 (1)은 응답자 총 36명 중 34명이 긍정적으로 답한 문장들이다. 부정 답변은 30대, 40대 이상에서 각 한 명이었고, 20대 응답자들은 모두 상식적이라고 답변했다.

(1) ㄱ. "작년이나 재작년과 달리 올해는 날씨가 덥다."
ㄴ. "다이어트를 하는데 오늘 유독 배가 고파서 힘들다."
ㄷ. "집에서 오랜만에 창문을 다 열어놓고 환기를 하고 있다."

위의 예를 통해 일상적으로 일어날 수 있는 문장이면서 문법적으로 판단했을 때도 정문인 문장을 상식적인 문장으로 판단했음을 알 수 있다. 긍정적으로 평가한 사람들이 가장 적은 문장은 "오늘 것은 내 취향이 아닌 것 같은데 기본 카드게임에 스토리가 있는 것 같다."로 20대에서 2명, 30대에서 2명, 40대 이상에서 1명으로 총 5명만이 상식적인 문장이라고 대답했다.

(2) ㄱ. "방 탈출 게임인 크라임 신을 하고 싶어 하는 이야기를 하고 있다."
ㄴ. "오늘도 H&M(호앤므)에서 예쁘고 귀여운 옷을 샀다고 하니 그만 사라고 한다."

위의 (2)는 연령별 차이를 가장 크게 보인 예다. (2ㄱ)에 대해 40대 이상에선 1명만이 상식적인 문장으로 평가한 반면 20대에서 5명, 30대에서 8명이 긍정적으로 평가했다. 부정적으로 본 견해는 "크라임 신"과 "방탈출 게임"은 비교적 최근에 보급되었기 때문에 일상대화에 자주 사용하는 용어가 아니라는 의견이었다. 본 설문과는 별개로 실시한 상식 문장 사전 조사에서 '초등학교, 수우미양가' 처럼 최근에는 사용하지 않는 어휘나 초등학교에서는 사용하지 않는 채점 방식에 대해 초등학생 자녀를 둔 학부모는 상식으로 판단하기 어렵다는 견해가 있었다. 이로 미루어 보아 연령별로 어휘의 수용성 판단이 다를 수 있을 뿐만 아니라 개인적인 문화 차이 역시 상식 문장 판단에 영향을 줄 수 있을 것으로 판단된다.

(2ㄴ)에 대해서도 40대 이상에서는 1명만이 상식적인 문장으로, 20대와 30대에서는 각각 8명, 5명이 이 문장을 상식적인 것으로 꼽았다. 부정적 평가의 근거로 40대 이상 응답자는 문장이 사실 전달이 아닌 개인의 의견 표시라는 점을 지적했다. 전체 부정 평가 근거 중 40대 이상에서는 '개인적인 판단에 의해서 결정하면 되는 문제니까', '매우 개인적인 취향에 해당함' 등 문장이 개인의 영역에 속하기 때문에 상식적인 문장에 해당하지 않는다는 의견이 많았다. '개인'(11회), '사실'(5회), '생각'(3회), '취향'(3회)과 같은 어휘가 평가 근거로 자주 사용되는바, 이 연령대에서는 보편적 사실 여부가 상식 판단에서 주요한 근거로 뽑혔다. 30대 응답자에서 문장의 성격에 대한 판단에서 '개인'의 생각이나 경험이라는 점이 부정 판단의 근거로 뽑힌 예는 2회에 그쳤다. 20대 응답자는 이러한 점을 지적하지 않았다.

3.1.2. 성별 차이

여기에서는 성별에 따른 중요 도메인과 덜 중요한 도메인의 차이가 있는가를 확인했다. 영역에 따른 차이가 보이지 않았고, 영역별로 성별과 응답은 유의 확률이 각기 0.64, 0.47을 초과하여, 필수 영역과 주변 영역 모두에서 성별은 응답형 사이의 상관관계도 볼 수 없었다.

3.2. 정성적 분석 결과

여기서는 세부적인 사항을 분석하기 위하여 주관식 판단을 대답 유형별로 분류하여 분석하였다. 형식에 대한 것과 내용에 대한 것으로 대별해 살폈다. 분류 외적인 요소로는 어휘 사용 빈도를 고려해야 한다는 의견이 있었다. 먼저, 어휘 사용의 적절성과 관련하여 3.1.에서 제시한 예문을 중심으로 논의를 구체화하고자 한다.

표 6. 문장 판단 요인

문장 내적 요인		문장 외적 요인	
어휘적 적절성	보편적 내용	문법적 적절성	담화 맥락

(3) ㄱ. 방 탈출 게임인 크라임 신을 하고 싶어 하는 이야기를 하고 있다.
ㄴ. 요즘 날씨가 엄청 춥고 눈도 오는 한파이다.

위의 예를 상식적 문장이라고 판단한 근거로 “크라임 신이라는 특정 게임명이 등장하지만 문장 내에서 게임명이라는 것을 알 수 있음.”과 같이 밝혔다. 반면에 비상식적 문장으로 판단한 근거로는 “게임은 프로그램 개발에 따라 신조어가 바로 사어가 될 수 있기 때문에 AI에게 학습시키기에는 시간적으로 문제가 있을 수 있습니다.”와 같은 견해가 있었다. 여기서 일반상식에 적합한 문장인지에 대한 판단에 특정 어휘가 영향을 준다는 점을 알 수 있다. 그렇다고 해서 단순히 쉬운 어휘를 사용하는 문장이 곧 일반상식 문장이라고 단정할 수는 없다. (3ㄴ)의 ‘한파’는 기초 어휘는 아니나 응답자는 “한파가 무엇인지에 대한 정보는 충분히 상식적인 정보”라고 답했다. 즉, 고유명사 내지 특정성이 강조된 명사가 일반상식에 적합한 문장으로 용인되기 위해서 전체 문장이 모두 저연령층도 알 만큼 쉬운 어휘로 구성되어야 한다는 경직된 사고보다는 두루 받아들여지는 타당한 내용을 담고 있는가를 우선 판단할 필요가 있겠다.

그러면 내용의 보편성이 문장의 상식성을 담보할까.

(4) ㄱ. 김밥에 든 단무지랑 햄버거에 든 피클은 안 먹는다며 취향을 존중해달라고 한다.
ㄴ. “언니 옷을 사러 갔다가 가디건이랑 얇은 엘레쎄 니트를 구매했다.”

(4ㄱ)의 예에서 “김밥에는 단무지가, 햄버거에는 피클이 들어 있으므로 상식적이라고 답했다”는 답변은 있었지만 ‘김밥, 햄버거’가 상식에 해당하지 않는 어휘라는 응답은 없었다. 반면 (4ㄴ)에 대해서는 “엘레쎄는 한국인에게도 상식 단어가 아닌 것 같습니다.”와 같은 의견이 있었다. 문장이 의식주 등 일반적인 정보를 나타낸다고 해서 모두 일반상식에 적합하다고 볼 수는 없으며, 상표명 등 상식 수준의 어휘를 사용했더라도 그 어휘가 한국인 일반에게 익숙한 것인가는 별도로 논의해야 한다는 것을 알 수 있다. 이를 통해 일반상식의 보편성과 개별성을 구분하여 한국인이 보편적으로 많이 사용하는 어휘들의 결합 양상이 먼저 파악해야 할 것이다.

다음으로, 문법적 적절성과 같은 형식적 요인 역시 영향을 미친다고 할 수 있다. 다음 문장을 보자.

(5) ㄱ. 3시에 비 오기 전에 일찍 가자고 하였다.
ㄴ. 옷 사기 전에 보여주겠다는데 다음에는 여름 티가 시급해서 여름 티를 사려 한다.

위의 예는 문장 자체가 인공지능 학습에 적절한가를 분석하기 전에 문법적 적절성과 관련하여 상식적이지 않다고 판단한 문장들이다. (5)에 대해 “일상 생활에서 자주 사용하는 복문은 아니라고 생각합니다. 상황은 충분히 나타날 수 있는 상황이지만 AI가 복문을 만드는 데 어려움이 있을 것 같습니다”, “문장 성분 생략으로 의미가 불분명함. 복문을 만들면서 의미가 애매해져서 학습시킬 만한 문장은 아니라고 생각합니다.” 등의 답변이 있었다. 단문과 복문에 따른 문장의 길이 및 용인 가능성 등도 학습에 적절한 문장인가를 판단하는 주요 요인이 될 수 있다고 결론 내릴 수 있다.

마지막으로, 연결어미를 중심으로 한 앞뒤 문장이 담화 맥락을 반영하여 유기적으로 연결되어 있는가는 다음 예시를 통해 살펴볼 수 있다.

(6) ㄱ. 점심에 초밥을 먹으러 추리하게 나갈 순 없는데 귀찮다고 하였다.
ㄴ. 여름에는 집이 너무 시끄러워서 살면 안 되겠다고 하자 호텔도 잠자리가 불편해서 잘 못 잤다는 이야기를 한다.

위 예에 대하여 ‘점심에 초밥을 먹으러 가자는데 추리하게 나갈 순 없어서 귀찮다고 대답했다’. 혹은, ‘추리하게~’ 부분을 빼고 ‘점심에 초밥을 먹으러 가자는데 귀찮다고 하였다.’로 만들어야 한다는 견해가 있었다. 또한 “현재 문장은 인과관계가 맞지 않는 비문이다, 시끄러워서 잘 수 없다 정도면 모를까”, “집 이야기하다가 갑자기 호텔 이야기?” 등 문장 내용이 상식적인가보다 ‘상황 맥락과’ 앞뒤 문장의 연결이 이해 가능한 문장인가에 대한 판단 여부가 상식적인 문장 판단에서 우선하였다. 여기서 어떤 내용을 담은 문장을 학습할 것인가에 대한 데이터 중심적 고려와 함께 정문을 생성하는 능력이 있는가라는 학습 능력에 대한 판단 역시 일반상식에 매우 중요한 요소임을 알 수 있다. 이에 4장에서는 상식적 대화가 친밀한 담화로 이어지기 위한 정량적 분석 외에도 정성적 방법론을 모색한다.

4. 상식적 대화에서 친밀한 담화로

3장에서 응답자들이 상식적 문장이라고 판단한 조건은 주위에서 친근하게 많이 들을 수 있는 어휘를 사용한 것, 한국인에게 보편적인 내용을 담을 것, 문법적으로 틀림이 없는 정문일 것, 담화 맥락을 이해할 수 있는 유기적인 문장 구조 등의 판단이 우선 작용했다. 가령 일상생활 데이터에 있는 다음 문장의 경우는 상호작용이 가능한 상식(interactive common sense)을 적용하면 적절한 상호 작용이 가능한 문장으로 변경할 수 있다.

(7) “배추에 맞았지만 게임상의 일이라서 화를 안 냈다.”
적절한 상호 작용
⇒ ① 일반적으로 배추로 맞는다면 화가 난다.
⇒ ② 게임상에서 일어난 일이라면 화를 낼 필요가 없다
⇒ ③ 어린이가 맞았다면 우선 다친 곳이 없는지 파악한다.
⇒ ④ 상처가 있다면 긴급 연락을 취한다.
부적절한 상호 작용
⇒ ① 누군가 맞았다면 인공지능 서비스는 상황에 관계없이 경찰에 신고한다.
⇒ ② 윤리적 판단을 할 수 없는 인공지능이 대답할 수 있는 문제가 아니다.

부적절한 상호 작용으로 판단한 문장 중 ①은 상식 추론이 되지 않아 하나의 상황에서는 하나의 판단밖에 할 수 없는 경직된 상호 작용이라 하겠다. ②는 챗봇 대화 등에서 자주 볼 수 있는 대답 회피형 상호작용이다.

상식적 문장이 적절한 상호작용이 가능한 데이터로 기능하는지를 확인하기 위해선 공개된 일반상식 데이터·

대화 데이터를 사용해야 한다. 하지만 아직 공개된 데이터가 없어 본 연구는 3장의 일상생활 요약 데이터를 오픈 도메인 대화 시스템(Open-domain Dialog Systems)의 질의응답형 대화로 변형해 2차 설문을 실시했다.

일상대화는 담화 내에서 개인적 정보 교환이나 자신의 마음을 솔직히 전달하는 과정이므로 친밀감 형성이라는 순기능이 있다[5]. 따라서 대화 품질 측정을 위해서는 챗봇 대화 비교를 통해 사람에게 기대하는 내용과 정도를 인공지능 챗봇 모델에게 기대하는지 등에 대한 기초적인 조사부터 이루어져야 하겠다. 사람들이 기대하는 챗봇 대화 분석을 위해 2018년 8월에 공개된 한국어 챗봇 대화 데이터(Chatbot_data_for_Korean v1.0)를 사용했다.[2] Chatbot_data_for_Korean은 총 11,876개의 대화로 구성되어 있는데 그중 주요 도메인과 보조 도메인에 나오는 핵심어와 연계되어 있는 대화만을 추출해, 3장에서 사용한 문장들과 비교가 가능하도록 대응쌍을 이루었다.

(8) Q : 옷장이 점점 줄어들어
Aㄱ. 이전 챗봇 응답: 지난 계절 옷을 잘 정리해 보세요.
Aㄴ. 상식 어휘 적용 챗봇 응답: 그래도 가디건이랑 얇은 니트류는 또 사고 싶죠?!

한국어 대화 요약 데이터에서 (8-Aㄴ)은 “언니 옷을 사러 갔다가 가디건이랑 얇은 엘레쎄 니트를 구매했다” 하는 진술문이었으나 지나치게 구체적이어서 상식적이지 않다고 평가된 상표명은 제외하였고 인용 표현 역시 챗봇과 대화하는 형식으로 수정하였다. 3장에서 기술한 것과 같이 복문이나 비문도 상식 판단에 영향을 주는 것으로 판단되어 핵심 개념어를 중심으로 문장을 재구성하여 비교 문항과의 문장 길이 차이를 최소화하고자 했다.

조사 참여자는 3장의 문장 테스트 참여자들로, 담화 맥락이 AI 챗봇과의 대화 상황에서 더욱 친숙한 답변을 고르도록 했다. 여기에 제3의 선택지로 ‘기타’ 항목을 두어 폐쇄형 질문 구성의 단점을 보완했다. 또한 적절한 문장쌍에 대한 선택형 질문 60개와 별도로 어떠한 내용을 담은 문장이 더 친밀했는지 등을 묻는 주관식 답변을 추가해 인식을 조사했다.

4.1. 정량적 분석 결과

총 36명이 각각 60문항에 대하여 대답한 결과인 2,160개의 응답을 대상으로 한다. 단, 제3의 응답 선택은 정량적 분석에서 제외하고 이전 대화 챗봇의 문장과 가공된 상식 문장 사이의 선택에 초점을 맞춘다.

4.1.1. 연령별 차이

3장에서의 분석과 마찬가지로 응답과 세대 변인은 유의미한 상관관계를 가지고 있음을 확인하였다. 필수 및 주변 영역 모두에서 세대와 응답 사이의 유의 확률은 0.00에 수렴했다. 특히 20대 및 30대의 평가와 40대 이상의 평가 사이에서 큰 차이가 보였다.

표 7. 연령별 각 문장 선택 집계

		필수 영역				주변 영역				총계
		밥	집	옷	계	날씨	게임	다이어트	계	
20	이전	43	62	52	157	50	75	70	195	352
	상식	72	50	62	184	70	39	47	156	341
30	이전	34	62	60	156	57	67	70	194	350
	상식	80	50	56	186	58	50	40	148	334
40 이상	이전	56	65	78	199	66	88	82	236	435
	상식	51	44	34	129	49	27	30	106	235

먼저 세대와 무관하게 다수의 답변이 쏠린 예를 확인해 보자.

(9) Q : 옷은 많은데 입을 게 없어.
Aㄱ. 옷은 사도 사도 모자라요.
Aㄴ. 주문한 옷이 택배로 와 있는데 걸 데가 없을 것 같군요.
(10) Q : 꽃게탕 진짜 밥도둑
Aㄱ. 기분 좋아 보이세요.
Aㄴ. 게딱지 양념에 흰쌀밥을 비벼 먹어도 맛있어요.

위 (9)의 경우 36명 중 35명이 (9-Aㄱ)을, 단 한 사람만이 (9-Aㄴ)을 선택했다. 반대로 (10)은 35명 중 32명이 (10-Aㄴ)을, 3명이 (10-Aㄱ)을 택했다. (9-Aㄱ)과 (10-Aㄴ)은 정보 제공적 성격에서는 차이를 보이지만 상대의 발화에 공감을 표시하는 문장이라는 점에서 공통점을 가진다. 반면에 (9-Aㄴ)은 택배 도착 등의 정보를 전달하지만 “옷은 많은데 입을 게 없어.” 라는 푸념에 대해 공감을 나타내지 못하여 선택되지 않은 것으로 보인다. (10-Aㄱ) 역시도 상대의 발화에 공감 또는 동조하지 않는다는 점이 유효하게 작용하여 배제되었다. 이에 반하여 (10-Aㄴ)은 상대가 긍정적으로 평가한 ‘꽃게탕’과 관련된 부가적인 정보를 제공함과 동시에 ‘맛있다’라는 술어를 통해 상대 의견에 동조함을 표현한다. 이런 측면에서 전 세대에서 ‘공감’을 친숙한 문장의 요건으로 삼았음을 알 수 있다.

(11) Q : 신혼집 인테리어 어떻게 하지?
Aㄱ. 하나씩 꾸며보세요.
Aㄴ. 저는 가짜 나무로 만든 트리를 만들어 놓고 싶어요.

연령 차이는 (11)에서 가장 두드러졌다. 대화에 대해 20대 참여자 11명 중 9명은 (11-Aㄴ)을 택해 명확한 선호를 보였다. 30대 참여자는 11명 중 5명이 이를 선택했는데, 40대 이상은 12명 중 단 한 사람만이 (11-Aㄴ)을 선택했다. 이러한 경향은 3.1.1.에서 본 40대 이상 응답자의 부정 평가 근거와 맥이 닿는다. 앞선 조사에서 높은 연령대일수록 개별적인 의견이나 취향에 바탕을 둔 문장은 AI가 학습할 상식으로는 부적절하다고 지적했다. 따라서 이를 학습한 일상대화에서도 행위 능력이 없는 AI가 신혼집 인테리어와 같은 정보성 질문에 자신의 취향을 밝히는 응답을 하는 것에 대해서 부정적이었다고 할 수 있다.

4.1.2. 성별 차이

3장에서와 마찬가지로 도메인과 성별 차이 사이에는 상관관계를 확인할 수 없었다. 필수 영역의 응답에 대한 성별 사이의 유의 확률은 0.65, 주변 영역에서는 0.95를

2) 한국어 챗봇 데이터Chatbot_data_for_Korean v1.0)데이터는 다음 링크에서 내려 받을 수 있다. URL:https://github.com/songys/Chatbot_data. 이 데이터로 만든 챗봇 모델은 [6]의 317-318쪽에서 시퀀스 투 시퀀스 기본 모델에 어텐션(attention)기법을 추가한 모델로 학습했을 때 데이터 크기에 비해 스코어가 상당히 높은 데이터인 것으로 기술되어 있다.

초과하였다. 이로써 성별 변인이 문장 평가에 유의미한 변인으로 작용하지 않음을 확인하였다.

4.2. 정성적 분석 결과

“결정을 망설이는 질문에서는 더 나은 선택을 하도록 도와주는 답변이, 단순히 걱정만 하는 질문에서는 긍정적인 말(위로, 격려)을 해주는 답변이, 정보가 필요해 보이는 질문에는 구체적인 정보를 제시해주는 답변이 친밀하게 느껴졌습니다.”의 의견이 보여주듯이 응답자들은 선행 문장의 성격에 따라 (1) 공감, (2) 정보 제공 및 조언 등을 담은 답변을 더욱 친숙하게 여겼다. 앞에서 본 (10)과 더불어 아래 (12)-(13)에서는 모두 일상 대화 데이터에서 추출한 어휘가 포함된 문장인 ‘Aㄴ’이 다수의 선택을 받았다.

(12) Q : 개강 옷 예쁘게 입어 볼까?
Aㄱ. 개시해보세요.(5명)
Aㄴ. 어떤 옷이 있는지 말해 주면 어떻게 매치하는 게 좋을지 알려 줄게요.(30명)
(13) Q : 날씨 좋은데.
Aㄱ, 하늘 보고 한 번 웃어봐요. 기분이 바뀔 거예요.(4명)
Aㄴ. 날씨도 좋고 구름도 예쁘죠?(31명)

(12-Aㄴ)은 Aㄱ이 포함하고 있는 ‘허락’의 기능을 포함하면서 상대에게 새로운 정보를 추가적으로 제공한다. 이러한 점은 챗봇 이용의 여러 목적 중 하나인 정보 제공 기능과 연결된다. ‘챗봇이라고 생각하니 사람이라고 생각할 때에 비해 좀더 유용한 정보를 포함해서 답변해 주는 것이 친밀하다고 생각된다.’ 등의 의견에서도 대화 상대로서의 챗봇에 정보 제공을 기대한다는 점을 밝힌다. 응답자들은 날씨가 따뜻해졌다는 문장이 주어졌을 때 ‘나들이 갈 곳을 찾아드릴까요?’, (9)의 주어진 문장에 대해 ‘신혼집에 어울리는 인테리어를 추천해 드릴게요.’ 등 ‘추천’(6회), ‘찾다’(2회) 등의 어휘가 포함된 의 제3 답변을 제시했다, 위의 (12-Aㄴ) 역시 정보 제공과 선택에 대한 조언 등의 기능을 포함한다.

그러나 문장이 더 많은 정보를 담고 있다고 해서 친밀감을 느끼지는 않는다. 이는 아래 예를 통해 알 수 있다,

(14) Q : 비 맞아서 옷 젖었어.
Aㄱ. 감기 조심 하세요.(31명)
Aㄴ. 옷이 급하게 필요하네요.(4명)

앞서 본 (13-Aㄴ)은 일종의 공감 표현으로, ‘날씨가 좋다’라고 전달받은 정보를 되풀이하면서 공동의 지식 체계를 공유한다. 여기에 더해 추가적인 정보(‘구름이 예쁘다’)를 제공함으로써 친밀감을 형성한다. 31명이 선택한 (14-Aㄱ)의 답변은 상대의 건강 상태에 대해 걱정한다.

생물의 공감(empathy)의 기저에는 상대의 주관적 상태에 대한 접근 메커니즘이 있어, 공감 능력(empathic capacity)의 핵심으로 꼽힌다[7]. 두 주체가 비슷하고 사회적으로 가까울수록 둘 사이의 동일시가 쉬워지고, 결과적으로 공감과 도움을 촉진시킨다[8]. 설문 응답자들도 공감 발화, 상대 발화에 심정적으로 동조하는 발화에서 친밀함을 느낀다고 응답했다. 응답자들은 더위에 대한 발화에 대해 ‘저도 녹아내릴 것 같은 날씨예요’, 다이어트 실패에 대한 발화에 ‘저도 해야 되는데 하다가 늘 실패하곤 해요’ 등의 제3 답변을 제안했다. 이처럼 ‘저도’(4회), ‘나도’(2회) 등 서로 다른 개체의 상황을 아우르는 조사 ‘도’를 사용하여 공감을 표시하여 친밀함을 표현하는 방향을 제안했다. 또는 ‘맞아요’, ‘그러게(요)’ 등 의견에 대한 동의 표현도 제시되었다.

한 가지 흥미로운 점은 상대의 말에 대해 공감과 조언을 표시하는 개체로서의 AI의 발화와 달리 대화 상대자로서 상대에게 동일시되지 않고 개인적 경험을 발화하는 데에는 거부감을 보였다는 점이다.

(15) Q : 게임하다 시간 다 갔어.
Aㄱ. 게임할 때는 시간이 더 빨리 가요.(33명)
Aㄴ. 저는 하다말다 하니까 실력이 안 느는 것 같아요.(3명)
(16) Q : 고구마만 먹고 다이어트 해야지
Aㄱ. 너무 무리하지는 마세요.(32명)
Aㄴ. 이제 만나면 고구마 같은 다이어트 식단으로 먹어야겠네요.(4명)

사람의 경우 친밀감을 형성하기 위하여 자신이 개방하는 것이 중요하지만 챗봇과의 대화라는 것을 인지하는 경우에는 (15-Aㄴ), (16-Aㄴ)과 같이 챗봇이 개별 인격체로서 개인적 경험을 이야기하는 대화는 친밀감을 이끌어내지 못했다. “암기된 답변을 읊는 것 자체가 친밀감을 떨어뜨린”다는 응답자 평가는 이 점과 관련된다. 이보다는 ‘다른 게임해보세요’(23명)와 같은 수준에서의 제안이나 개인의 경험이 아닌 희망을 이야기하는 ‘저도 하고 싶네요’(31명) 등이 긍정적인 답변으로 인정되었다. 이는 인공지능 챗봇을 사람과 동일시하여 대화 모델을 설계하는 경우 사용자의 경험에는 부정적 영향을 미칠 수 있음을 뜻한다. 가벼운 수다에 해당하는 질문에 정보를 제공하는 경우 역시 마찬가지다.

요약하면 일상대화를 사용한 대화 품질을 높이기 위해서는 세 가지 선결 조건이 지켜져야 하겠다. 먼저 질문의 요구에 적절한 정보와 공감을 제공해야 한다. 두 번째로 공감의 정도가 챗봇의 특성에 맞는 응답이어야 한다. 세 번째로 대화의 차례에 따라 담화의 규칙을 지키면서 대화가 진행되어야 한다. 인공지능 챗봇임을 분명히 하고 행위 능력이 없음을 대화하는 사람이 알고 있는 경우에 “이제 만나면 고구마 같은 다이어트 식단으로 먹어야겠네요”와 같이 대답하는 경우는 위의 세 가지 이유에서 모두 긍정적으로 평가되지 못한다. 반면에 발화 차례에서 사용자가 챗봇을 개별 인격체로 허가를 한 질문을 하는 경우에는 긍정적 피드백을 받을 수 있다. 예컨대 ‘밥 먹었니?’와 같은 경험을 묻는 질문에 대해서는 ‘배고프지 않아요’(5명)보다 ‘오늘은 순두부찌개 먹고 싶어서 시켜 먹었어요.’(31명)와 같이 일상생활 용어를 반영한 구체적 문장을 선택한 사람이 월등히 많았다. 따라서 위의 세 가지 조건들이 통합적으로 적용된 담화 설계가 된 이후에야 “완전히 인공지능(“AI-complete”)스러운 대화[2]가 가능할 것으로 판단된다.

5. 결론

최근 NYU 및 Johns Hopkins 대학의 자연어 처리 연구자 480명을 대상으로 한 조사의 기술 문서가 공개되었다[9]. 그중 향후 10년간 가장 영향력 있는 연구 분야는 무엇인지에 대해 응답자들은 하드웨어(hardware)나 데이터 크기(data scaling)보다 "문제 정의와 과제 설계"가 더 중요하다고 보았다. 과제 설계에는 다양한 방법과 응용이 있을 수 있지만 데이터 설계에서의 문제 정의는 데이터의 특성을 파악하는 것에서부터 비롯되어야 할 것이다. 일반상식에 대한 연구도 마찬가지다. 현실 데이터에 바탕을 둔 문제 정의와 효율적인 과제 설계가 이루어진 후에 데이터 크기를 키우는 단계적 설계가 필요하겠다.

이렇게 구축된 데이터와 모델은 챗봇이나 한국어 일반상식 대화 등 인공지능 모델을 산업화하는 데에도 중요하게 사용할 수 있을 것이다. 가령, 현재 인공지능 분야의 연구 방향 중 하나인 비윤리적인 내용이 있는 데이터를 탐지하는 연구 역시 지금과 같이 문제가 있을 만한 데이터를 삭제하는 방식보다는 사회 구성원의 대다수가 옳다고 생각하는 상식적인 내용을 학습하도록 한 후 인간과 상호작용하도록 해야 할 것이다.

그동안 상식 문장에 대한 정의는 단편적인 경우가 많았고 한국어에 적합한 상식 문장에 대한 고민 역시 저조한 상황이었다. 그런 가운데 일반상식 문장의 자동 생성 및 평가가 가능한 정도의 기술력이 요구되고 있다. 그러나 이렇게 구축된 데이터나 모델이 상식 문장의 자동화를 이루었다고 판단하기는 어렵다. 따라서 본 논문은 일반상식이 인공지능이 일상생활에서 두루 쓰이는 구체적 용어에 대해 이해하고 이로부터 기대되는 감정과 상황 변화에 대한 보편적 지식을 산출할 수 있는가를 다루는 영역이며, 이러한 측면에서 일반상식은 지식의 증강이나 추론과는 차이가 있다고 정의하고 이를 바탕으로 선별된 문장을 중심으로 설문을 실시했다. 정량적 평가에서 일반상식에 대한 연령별 차이는 뚜렷했으나 성별 차이는 그렇지 않았다. 정성적 평가에서는 어휘 하나하나를 두고 상식 정도를 판단하는 것보다는 문장 층위에서 다수의 한국인에게 받아들여지는 보편적 내용을 담고 있는가를 판단하는 것이 의미있다고 보았다. 또한 문장의 내용 못지않게 유기적인 내용으로 구성된 정문이 일반상식에 적합한 데이터로 여겨진다는 것을 알 수 있었다.

담화 설계에서는 질문에 요구에 따라 적절히 정보와 공감을 제공할 때 긍정적 피드백을 받을 수 있었다. 공감을 제공할 때도 그 응답이 챗봇의 특성에 맞는 공감적 언어를 사용하고 대화의 차례에 따라 담화의 규칙을 지키면서 대화가 진행될 때 긍정적인 것으로 평가되었다.

참고문헌

[1] Liu, H. & Singh, P., ConceptNet: A Practical Commonsense Reasoning Toolkit, 211-226, BT Technology Journal 22(4), 2004.

[2] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L. & Parikh, D., VQA: Visual Question Answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433), 2015.

[3] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S.R., GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461. 2018.

[4] AIHUB, 한국어 대화 요약, URL: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=117, 2020.

[5] Jiaxin Pei & David Jurgens, Quantifying Intimacy, in Language, https://aclanthology.org/2020.emnlp-main.428, 2020.

[6] 전창욱 · 최태균 · 조중현 · 신성진, 텐서플로 2와 머신러닝으로 시작하는 자연어 처리, 위키북스, 2020.

[7] Preston, S.D, & de Waal, F.B.M., Empathy: its ultimate and proximate bases. Behav. Brain Sci. 25, 1-72, 2002.

[8] de Waal, F.B.M., Putting the Altruism Back into Altruism: The Evolution of Empathy, The Annual Review of Psychology 2008;59, 279-300, 2008.

[9] Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R.Y., Phang, J. & Bowman, S. R., What Do NLP Researchers Believe? Results of the NLP Community Metasurvey, arXiv:2208.12852, 2022.

[10] Hwang, J.D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A., & Choi, Y., COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. AAAI, 2021.

[11] Jaehyung Seo, Seounghoon Lee, Chanjun Park, Yoonna Jang, Hyeonseok Moon, Sugyeong Eo,Seonmin Koo & Heuiseok Lim, A Dog Is Passing Over The Jet? A Text-Generation Dataset for Korean Commonsense Reasoning and Evaluation, Findings of the Association for Computational Linguistics: NAACL 2022, pages 2233 - 2249 July 10-15, Association for Computational Linguistics, 2022.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," European conference on computer vision, pp. 740-755, 2014.

[13] 김중섭 · 김정숙 · 이정희 · 김지혜 · 박나리 · 박진욱 · 이수미 · 강현자 · 장미정 · 홍혜란, 〈국제 통용 한국어 표준 교육과정 적용 연구〉. 국립국어원, 2017.

[14] 국립국어원, 국립국어원 일상 대화 음성 말뭉치 2020(버전 1.2).URL:https://corpus.korean.go.kr, 2021.

[15] 송영숙, 한국어 챗봇 데이터 Chatbot_data_for_Korean v1.0). URL: https://github.com/songys/Chatbot_data, 2018