공학석사 학위논문

대화형 에이전트 개발을 위한 크라우드소싱 기반의 학습데이터 수집 방안 연구

– 수집 결과의 표현 다양성 향상을 중심으로 –

2019 년 8 월

서울대학교 융합과학기술대학원

융합과학부 디지털정보융합전공

김 병 준

대화형 에이전트 개발을 위한 크라우드소싱 기반의 학습데이터 수집 방안 연구

– 수집 결과의 표현 다양성 향상을 중심으로 –

지도 교수 이 중 식

이 논문을 공학석사 학위논문으로 제출함
2019 년 7 월

서울대학교 융합과학기술대학원
융합과학부 디지털정보융합전공
김 병 준

김병준의 공학석사 학위논문을 인준함
2019 년 7 월

위 원 장 서 봉 원 (인)

부위원장 이 교 구 (인)

위 원 이 중 식 (인)

초 록

대화형 에이전트는 사용자로부터 자연어를 입력 받아 인텐트를 파악하고 기능을 수행하는 시스템이다. 음성 인식 기술의 고도화와 거대 IT 기업들을 중심으로 개발 플랫폼을 제공함에 따라 대화형 에이전트를 이용한 서비스 개발이 보편화되고 있다.

이러한 대화형 에이전트를 개발하기 위해서는 다양하고 많은 양의 학습데이터가 필요하다. 현재 대화형 에이전트는 사용자에게 사람처럼 대화하는 상호작용 방식을 제공한다. 이에 따라 대화형 에이전트는 사용자의 대화 인텐트를 파악해야 하며, 인텐트 파악은 다양하고 많은 양의 학습데이터를 통해 학습되기 때문이다.

하지만 대화형 에이전트 개발을 위한 학습데이터를 수집하는 것은 자연어의 표현 다양성과 수집 방법의 한계로 인해 매우 어려운 작업이다. 자연어의 표현 다양성은 같은 의미를 가지면서 다른 구조를 가질 수 있음을 뜻하며, 학습데이터 수집은 이러한 특성이 고려되어야 한다. 수집할 수 있는 방법들이 일부 제안되긴 하였으나 시간, 비용, 접근성 등의 문제가 제기되고 있다.

최근 인공지능 개발이 활성화됨에 따라 크라우드소싱 분야가 발전하면서 이러한 문제를 해결할 가능성을 엿볼 수 있게 되었다. 크라우드소싱은 컴퓨터가 해결하기 어려운 문제를 사람들로부터 풀며, 적은 비용으로 다수의 사람들에게 데이터를 수집할 수 있는 장점이 있다. 실제, 학습데이터 수집과 관련하여 크라우드소싱의 활용 가능성이 제기되고 있다.

하지만 태스크 디자인 방식에 따라 크라우드소싱의 수집결과가 많은 영향을 받고, 학습데이터의 다양성이 중요함에도 불구하고 태스크 디자인 방식에 대한 이해가 부족한 상황이다. 따라서 본 연구는 표현의 다양성 향상에 초점을 맞춰 태스크 디자인 요소가 학습데이터 수집 결과에 미치는 영향을 알아보고, 효과적으로 학습데이터를 수집할 수 있는 디자인 방안을 제언하고자 한다.

이를 위해 본 연구에서는 크라우드소싱 기반의 학습데이터 수집 결과에 영향을 주는 태스크 디자인 요소들을 선정하여 이에 따른 영향을 알아보고자 일련의 3가지 실험(태스크 양, 보너스 보상 방식, Social Proof 기반 설명 방식)을 진행했다. 수집가능성이 검증된 패러프레이징 태스크를 사용하였으며, MTurk을 통해 480명의 참가자로부터 73.65달러를 사용하여 1473개의 데이터를 수집하였다. 수집한 데이터는 4가지 지표(의미적 동등성, 다양성, 에러 비율, 수행 시간)로 분석되었다.

분석 결과, 태스크 양이 늘어날수록 같은 의미를 갖는 데이터를 얻기 어려웠다. 보너스 보상 방식 측면에서는, 보너스 보상 방식을 제공할 때 수집의 효율성이 높아졌다. 마지막으로 Social Proof 기반 설명 방식 측면에서는 다양성과 효율성 사이의 트레이드 오프(Trade-off) 관계가 나타났다. 최종적으로 참가자 간 수집의 개인차, 수집 결과에 대한 압박에 대해 논의하고, 실험 결과를 종합하여 통합적인 태스크 디자인 방식을 제언하였다.

본 연구는 학습데이터의 수집 가능성을 밝히는 연구가 주를 이루는 가운데, 수집 결과를 향상시킬 수 있는 방안을 연구한다는 점에서 학술적 의의를 갖는다. 또한 대화형 에이전트의 개발이 보편화되는 시점에, 산업 분야에서 실제 겪고 있는 문제를 해결하고자 한다는 점에서 시의성과 유용성 측면의 의의를 갖는다. 마지막으로 사회심리학 이론, HCI, 공학 분야를 접목한다는 점에서 융합적 의의를 갖는다.

주요어 : 크라우드소싱, 학습데이터 수집, 대화형 에이전트, 태스크 디자인, 표현의 다양성
학 번 : 2017-24292

목 차

표 목차

그림 목차

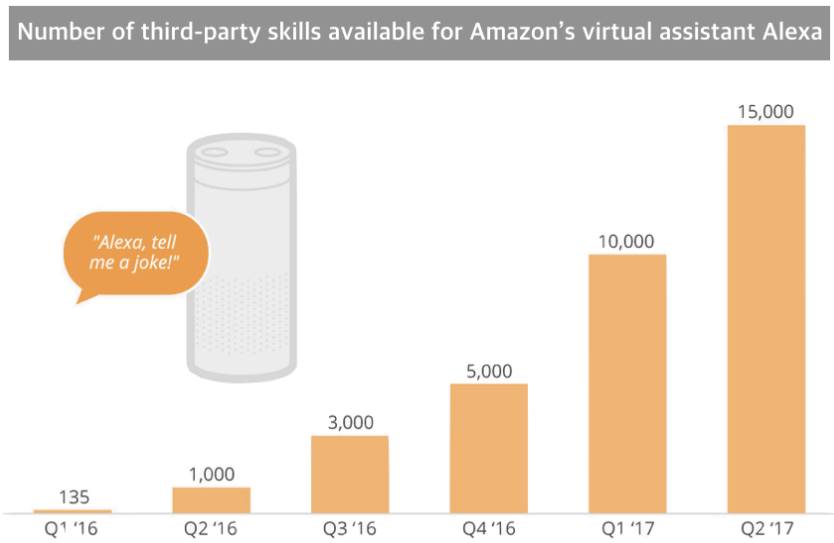
제 1 장 서 론

제 1 절 연구의 배경

대화형 에이전트(예, 아마존 에코, 구글 홈)는 사용자로부터 자연어(문자 혹은 글자)를 입력 받아 인텐트를 파악하고 기능을 수행하는 시스템이다(Jurafsky and Martin, 2014). 여기서 인텐트란 사용자가 대화한 의도 혹은 목적을 뜻한다. 예를 들어, 사용자가 대화형 에이전트에게 “음악 틀어줘” 라고 요청한다면, 대화형 에이전트는 음악 재생이라는 인텐트를 파악하고 응답에 필요한 기능을 수행하게 된다. 음성 인식 기술의 고도화와 거대 IT 기업들을 중심으로 아마존의 Alexa Skill Kit[①], 구글의 Dialogflow[②]와 같은 개발 플랫폼을 제공함에 따라 대화형 에이전트를 이용한 서비스 개발이 보편화되고 있다. 독일의 마켓 리서치 전문 기관인 Statista에 따르면 아마존의 Alexa Skill Kit을 통해 만들어진 서비스는 2017년 상반기까지 15,000개에 달하며, 분기 당 개발되는 서비스의 증가 폭이 점점 늘어나고 있다[그림 1]. 또한, 한국정보화진흥원에 따르면 서비스 개발 분야 측면에서도 전 분야에 걸쳐 대화형 에이전트의 서비스 개발이 증가하고 있으며 상담, 주문, 상품 검색 등 그 분야가 더 다양화되고 있다[그림 2]. 리서치 전문 기관인 Gartner에서 2020년까지 평균적으로 배우자보다 대화형 에이전트와 더 많은 의사소통을 나눌 것으로 전망하면서 향후에도

① https://developer.amazon.com/alexa-skills-kit
② https://dialogflow.com/

대화형 에이전트를 이용한 서비스 개발이 점점 더 증가할 것으로 추정된다.



[그림 1] 개발 플랫폼을 통한 대화형 에이전트 개발의 증가③

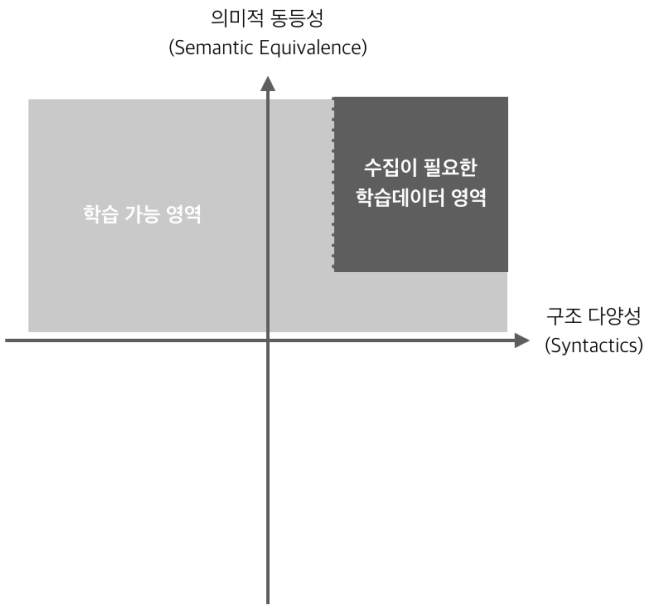
[그림 2] 다양한 서비스 분야에서 활용되고 있는 대화형 에이전트④

③ 그림 출처 : https://www.statista.com/chart/8304/alexa-skills/
④ 그림 출처 : https://www.topbots.com/100-best-bots-brands-businesses/

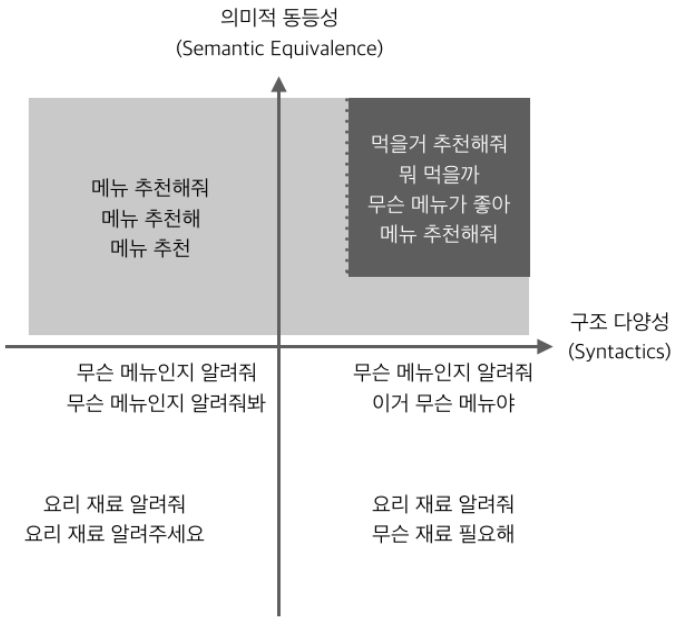
이러한 대화형 에이전트를 개발하기 위해서는 사용자 대화로부터 인텐트를 파악할 수 있도록 하기 위한 많은 양의 학습데이터가 필요하다. 대화형 에이전트는 자연어 이해(Natural Language Understanding, 이하 NLU) 모델을 기반으로 사용자 대화로부터 인텐트 파악 과정을 진행한다. 인텐트 파악은 사용자 대화(또는 쿼리)로부터 인텐트를 분류하는 작업으로, 문장을 구성하는 단어의 구조에 따라 특정 인텐트에 대한 조건부 확률로 표현된다(Kim et al, 2016). 여기서 NLU 모델이 가지고 있는 인텐트 클래스 리스트 중 가장 확률이 높은 특정 인텐트를 찾게 된다. 인텐트 클래스 리스트는 인텐트와 이에 대한 문장 셋으로 학습되며, 학습데이터의 양이 적으면 특정 인텐트 결과값 간 분산이 높아져 인텐트 파악의 정확도가 떨어지게 된다. 학습을 위한 적정 데이터 양은 쉽게 수치화할 수 없지만, 데이터 양이 풍부하고 다양할수록 인텐트 파악 성능이 향상되는 것으로 연구되었다(Banko and Brill, 2001; Tur et al, 2010).

하지만 자연어가 갖는 표현의 다양성 및 수집 방법의 한계로 인해, 대화형 에이전트를 개발하기 위한 학습데이터 수집은 매우 어려운 작업이다. 여기서 표현의 다양성이란 하나의 인텐트가 여러 형태로 표현될 수 있음을 뜻하며, 같은 의미(Semantics)를 가지면서 다른 구조나 형태(Syntactics)를 갖는 것을 의미한다(Achananuparp et al, 2008). 이로 인해 대화형 에이전트 개발에 필요한 학습데이터는 [그림 3]과 같이 일정 수준의 의미적 동등성과 구조의 다양성 조건을 만족하는 다양한 표현이 필요하다. 예를 들어 “뭐 먹을까?”, “메뉴 추천해줘”, “무슨 요리가 좋아?” 등과 같이 메뉴 추천이라는 같은

의미에 다른 형태로 표현된 데이터 수집이 요구된다[그림 4]. 이를 위해 Woz 방법론(Lathrop et al, 2004)이나 문법 기반의 문장 생성(Pieraccini and Huerta, 2005), 데이터 셋 활용(Lowe et al, 2017)등의 기법들이 제안되었으나, 시간 및 비용이 많이 소모되거나 일부 전문가에 국한된 접근성, 도메인 이슈(Labutov and Lipson, 2013)의 한계가 있는 상황이다.



[그림 3] 표현의 다양성에 따른 학습데이터 수집 영역



[그림 4] 메뉴 추천 인텐트에 대한 [그림 3]의 사분면별 예시

그러나 근래 인공지능(Intelligent Agent) 개발의 활성화와 함께 크라우드소싱(또는 Human Computation) 분야가 발전하면서 새로운 가능성이 제안되었다. 크라우드소싱은 실제 사람들로부터 데이터를 수집할 수 있어 컴퓨터가 해결하기 어려운 문제를 푸는데 활용될 수 있다. 또한 전통적으로 사람을 모집하여 리워드를 제공하는 방식보다 더 적은 비용이 소모되며, 크라우드소싱 플랫폼을 통해 손쉽게 많은 사람들을 대상으로 데이터를 수집할 수 있다는 장점이 있다(Callison-Burch and Dredze, 2010). 이러한 장점들로 인해 대화형 에이전트 개발 과정에서도 크라우드소싱 활용방안이 연구되었다. 데이터에 라벨을 매칭하는 라벨링 작업(Snow et al, 2008), 음성 파일을 전사하는 작업(Vashistha et al, 2017), 소수 언어를 번역하는 작업(Bloodgood and Callison-Burch, 2010; Zaidan and Callison-Burch, 2011) 등의

분야들이 연구되었으며, 적은 비용으로 요구하는 수준의 데이터 수집이 가능하다고 밝혀졌다.

따라서 본 연구는 대화형 에이전트 개발 초기에 부족한 학습데이터 문제를 크라우드소싱을 통해 해결하는 방안을 제안하고자 한다. 이러한 문제를 크라우드소싱으로 해결하기 위해서 선행된 연구들은 크라우드소싱을 이용하여 인텐트 파악 학습에 필요한 학습데이터 수집이 가능함을 밝히는데 초점을 맞추었다(Bapat et al, 2018). 그러나 크라우드소싱의 태스크 디자인 방식에 따라 데이터 수집 결과가 많은 영향을 받으며(Allahbakhsh et al, 2013; Daniel et al, 2018), 학습데이터의 다양성이 중요하므로 학습데이터 수집을 위해 어떻게 태스크를 디자인해야 하는지에 대한 이해가 필요하다. 그러므로 본 연구에서는 특히 표현 다양성 향상에 초점을 맞추어, 학습데이터 수집을 효과적으로 할 수 있는 태스크 디자인 방식을 연구하고자 한다. 여기서 태스크 디자인 방식이란 크라우드소싱을 통해 수집한 데이터 결과(Quality)에 영향을 미치는 태스크 디자인 요소들을 뜻하며, 선행연구들에 기반하여 요소를 선정하고 각 요소에 따라 다양한 표현 수집에 미치는 영향을 고찰하고자 하였다.

이를 위해 본 연구는 각 160명씩 총 480명을 대상으로 3가지 주요 요소(태스크 양, 추가 보너스 제공 방식, 태스크 설명 방식)에 대한 디자인 통찰을 얻고자 일련의 실험을 진행했다. 실험은 Amazon Mechanical Turk(이하 MTurk)을 사용하였으며 Bapat et al. (2018)에 의해 활용가능성이 검증된 패러프레이징 방식을 태스크로 제시하였다. 첫 번째, 태스크 양 관련 실험에서는 26.45 달러를 사용하여 529개의

학습데이터를 수집하였으며, 분석 결과 태스크 양이 증가할수록 같은 의미를 갖는 데이터를 얻기 어려웠다. 두 번째 추가 보너스 제공 방식 관련 실험에서는 22.1 달러를 사용하여 442개의 학습데이터를 수집하였다. 분석 결과, 추가 보너스 제공 방식은 수집 효율성을 높여주는 것으로 나타났다. 마지막으로 세 번째 태스크 설명 방식에서는 Social Proof 개념을 적용하여 설명 방식을 제시하였다. 25.1 달러를 사용하여 502개의 학습데이터를 수집하였으며, 분석 결과 데이터 간 다양성과 에러 사이에 트레이드 오프 관계가 나타났다. 최종적으로 실험 결과에서 해석된 개인차, 태스크 결과 충족에 대한 압박 논의를 통해 디자인 가이드 라인을 제언하고 연구의 한계 및 의의를 밝혔다.

제 2 절 논문의 구성

본 논문은 다음과 같은 논문의 구성을 따른다. 먼저, 2장에서는 대화형 에이전트의 인텐트 파악과 학습데이터, 자연어 관련 크라우드소싱 활용 연구, 수집결과에 영향을 미치는 태스크 디자인 요소 및 실험 방법과 관련하여 Social Proof 연구에 대한 이론적 배경을 살펴본다. 그 다음, 3장에서는 본 연구의 연구 문제를 구체적으로 서술하며, 4장에서 연구 문제를 풀기 위해 실험 설계한 연구 방법을 소개한다. 이후 5장에서 일련의 실험에 대한 각각의 연구 결과를 제시하였으며, 6장과 7장에서 실험으로부터 도출한 논의 및 결론이 이어 서술된다.

제 2 장 이론적 배경

제 1 절 대화형 에이전트의 인텐트 파악

본 절에서는 대화형 에이전트의 인텐트 파악에 대해 서술하고 해당 과정이 학습데이터와 어떤 관련이 있는지 알아본다. 인텐트 파악의 방식 및 학습데이터와 대화형 에이전트의 인텐트 파악에 대한 성능 관계 연구들은 풍부하고 다양한 학습데이터의 중요성을 보여준다.

대화형 에이전트가 사용자와 상호작용을 하기 위해서는 인텐트 파악 과정이 요구된다. 대화형 에이전트는 사용자에게 사람과 대화하는 것처럼 자연스러운 상호작용 방식을 제공하고 있다. 실제 대화형 에이전트 사용자들은 하나의 작동방식을 외워 사용하는 것이 아니라 사람과 대화 하듯이 자신의 대화 인텐트를 다양하게 표현하고 있다(Cooke et al, 2017; Luger and Sellen, 2016). 이로 인해 대화형 에이전트는 사용자의 특정 응답을 매칭하여 기능을 수행하는 방식이 아니라 사용자의 대화 인텐트를 파악하는 과정이 필요하다. 인텐트 파악 에러가 대화형 에이전트 사용 경험을 저하시키는 주요 요인으로 연구되기도 하였다(Myers et al, 2018).

대화형 에이전트의 인텐트 파악은 NLU 모델을 기반으로 한다. 여기서 인텐트 파악은 사용자의 대화 문장으로부터 인텐트를 분류하는 작업이다(Kim et al, 2016). 문장의 구성 요소를 토큰화하여 어휘 및 문장의 구조에 따라 특정 인텐트에 대한 조건부 확률로 표현된다[그림

5].

$$y' = \arg\max_{y} p\left(y|w_1, \ldots, w_n\right)$$

w_i 는 문장의 i번째 단어, y는 Intent

[그림 5] 인텐트 파악의 수식화⑤

특정 인텐트는 대화형 에이전트의 NLU 모델이 가지고 있는 인텐트 클래스 리스트 중 하나로써 사용자의 대화 문장으로부터 가장 확률이 높은 인텐트를 사용자 인텐트로 파악한다. 인텐트 클래스 리스트는 학습하고자 하는 인텐트 라벨과 해당 라벨과 매칭되는 문장 데이터로 학습된다. 이때, 학습데이터의 양이 적으면 특정 인텐트에 대한 조건부 확률값 사이에 분산이 높아져 인텐트 파악의 정확도가 떨어지게 된다.

학습을 위한 적정 데이터 양에 대해서는 쉽게 수치화할 수 없지만 학습데이터 양의 중요성은 연구적으로 밝혀져 왔다. Tur et al. (2010)는 NLU 학습을 위해 많은 연구들에서 사용하고 있는 항공 여행 정보 시스템 데이터(ATIS)를 기반으로 에러를 줄일 수 있는 방법들에 대해 탐구했다. 그 결과, 모델링이나 특징 디자인 과정에서 생기는 과적합 문제 및 자연어로 인한 분산 문제를 많은 양의 학습데이터를 통해 해결할 수 있다고 제안했다. 또한 Banko and Brill. (2001)은 자연어 모호성 작업을 수행하는데 있어 학습데이터 양에 따른 머신러닝 알고리즘들의 성능을 분석하였으며, 학습데이터 양이 많으면 많을수록 성능이 향상된다고 밝혔다. 최근 Kim et al. (2016)은 ATIS 데이터와 상대적으로 데이터 크기가 더 큰 마이크로소프트 코타나로부터 얻은

⑤ Kim et al. (2016) 수식 인용

데이터를 가지고 인텐트 파악 모델을 학습시켜 성능을 비교했다. 이를 통해 더 풍부하고 다양한 학습데이터로 학습시킬수록 모델의 성능이 향상됨을 밝혔다. 이와 같은 연구들은 대화형 에이전트의 인텐트 파악에 있어 다양하고 많은 양의 학습데이터가 성능에 중요하다는 사실을 보여준다. 이는 대화형 에이전트 개발 초기에 다양하고 풍부한 학습데이터를 수집할 수 있다면, 대화형 에이전트의 인텐트 파악 성능 향상에 큰 도움이 될 수 있다는 점을 시사한다.

제 2 절 자연어 학습데이터 관련 크라우드소싱 활용 연구

본 절에서는 자연어 학습데이터와 관련하여 크라우드소싱을 활용한 연구들을 살펴보고, 더 나아가 본 연구 분야인 대화형 에이전트 개발을 위한 학습데이터 수집 연구들로부터 연구의 타당성과 차별성을 설명한다.

크라우드소싱은 컴퓨터가 해결하기 어려운 문제를 사람을 통해 해결하고자 발전되었으며, Howe. (2006)에 의해 처음 명명되었다. 이후 크라우드소싱은 근 10년 동안 매우 빠르게 성장하여 다양한 영역에서 활용되고 있으나(Mao et al, 2017), 본 절에서는 본 연구와 관련하여 자연어 학습데이터를 중심으로 서술하고자 한다.

자연어 학습데이터와 관련하여 크라우드소싱은 라벨링, 전사, 번역 등의 분야에서 활용되었다. Snow et al. (2008)은 MTurk을 사용하여 크라우드소싱을 통해 비전문가로부터 신뢰할 수 있는 수준의 자연어 라벨 데이터를 얻을 수 있는지 검증하였다. 감정 인지, 단어 유사성, 맥락 파악, 사건 순서, 단어 모호성의 5가지 태스크를 실시하여 데이터

라벨을 얻고 표준 데이터(Gold standard)와의 비교를 통해 결과를 평가했다. 그 결과, 크라우드소싱을 통해 적은 비용으로 다양한 라벨링 태스크를 수행할 수 있음을 밝혔다. Vashistha et al. (2017)은 음성 전사를 위해 크라우드소싱을 활용하였는데, 인도 학생들로부터 힌두어와 인도식 영어 전사를 수행하도록 하였다. 또한 그 과정에서 태스크 디자인을 달리하여 수집 결과의 효율성을 높이는 디자인을 찾고자 하였고, 음성 전사 서비스를 지원하는 상업적 도구들보다 더 경제적으로 뛰어난 음성 전사 결과를 얻었다. 이와 비슷하게 Lee and Glass. (2011)도 크라우드소싱을 통한 음성 전사 결과를 향상시킬 수 있는 디자인 방안을 연구하여 적은 비용으로 뛰어난 수집 결과를 얻었다. 뿐만 아니라 기계 번역 분야에서도 많은 양의 학습데이터가 필요하기 때문에 크라우드소싱이 많이 활용되었다. Bloodgood and Callison-Burch. (2010)는 기계 번역 테스트를 위한 데이터를 수집하고자 크라우드소싱을 활용했다. 우르두어와 영어의 번역 문장을 수집하였고 더 적은 비용으로 기계 번역 테스트 데이터를 수집할 수 있음을 보여줬다. 이러한 연구들은 자연어 학습데이터를 수집함에 있어 크라우드소싱의 유용성을 보여준다.

자연어 학습데이터 생성과 관련해서도 크라우드소싱을 활용하는 방안이 연구되었다. Wang et al. (2012)은 크라우드소싱을 통해 학습데이터를 생성할 수 있는 유도 방법 기반의 방식을 제안했으며, 패러프레이징을 통한 학습데이터 생성 가능성을 보여줬다. 마찬가지로 Bapat et al. (2018)은 End-to-End 파이프라인의 학습데이터 생성 방식을 제안하여, 패러프레이징을 통해 학습데이터를 늘릴 수 있음을

밝혔다. 하지만 이전 연구들은 학습데이터 생성에 미치는 영향이나 수집 결과의 다양성 보다는 크라우드소싱을 이용한 자연어 학습데이터 생성 가능성을 밝히는데 초점을 맞추고 있다는 한계가 있다.

대화형 에이전트 개발을 위한 학습데이터는 다양성이 중요하고 태스크 디자인 방식에 따라 수집 결과에 큰 영향을 미침에도 불구하고 아직까지 이런 요인들이 크라우드소싱 기반의 학습데이터 생성 과정에서 어떻게 영향을 미치는지에 대한 연구는 미비한 상황이다. 일부 연구(Jiang et al, 2017)에서 태스크 디자인에 따라 수집된 학습데이터의 다양성에 영향을 끼친다는 점을 밝혀낸 바 있지만, 요소들 간의 상호작용이나 개별 디자인 요소의 통합적 제언, 학습 모델을 통한 실제 사용자 발화와의 비교를 통해 연구를 확장하고자 한다.

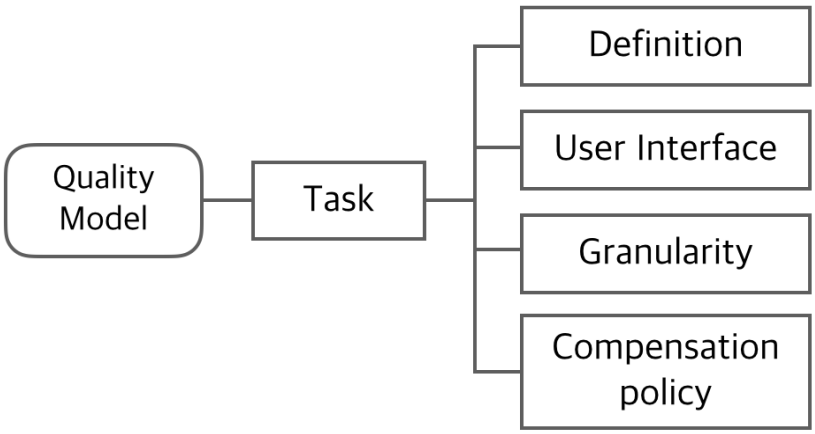
따라서 본 연구는 학습데이터 생성 가능성이 검증된 패러프레이징 방식을 기반으로, 수집 결과의 표현 다양성 향상에 초점을 맞추어 학습데이터 생성에 태스크 디자인 요소가 미치는 영향을 분석하고, 이를 바탕으로 통합적인 태스크 디자인 가이드라인을 제언하고자 한다.

제 3 절 크라우드소싱 수집 결과와 관련된 태스크 디자인 요인

본 절에서는 크라우드소싱 수집 결과에 영향을 주는 태스크 디자인 요소들에 대해 알아본다. 이를 통해 태스크 디자인 요소들을 도출하고, 도출한 요소 중 선행연구들을 통해 다양성에 영향을 미치는 요소들을 연구문제로 적용하여 연구를 설계하였다.

먼저 크라우드 소싱의 수집 결과(Quality)란 Allahbakhsh et al. (2013)에 따르면 제공된 결과가 요청자의 요구 사항을 충족시키는 범위라고 정의하고 있다. 이에 따라 본 연구는 표현의 다양성 정의에 맞춰 패러프레이징을 요청하는 인텐트와 참가자가 패러프레이징한 문장이 같은 인텐트로 판단되는 경우를 최소 충족 범위로 정하였다. 또한, 수집 결과의 향상 여부(Quality Control)는 최소 충족 범위를 만족하는 데이터 양의 증가와 충족 범위를 만족한 데이터 간의 다양성 증가로 정하였다.

Allahbakhsh et al. (2013)는 크라우드소싱 수집 결과에 [그림 6]과 같은 태스크 디자인 요인들이 영향을 준다고 밝혔다.



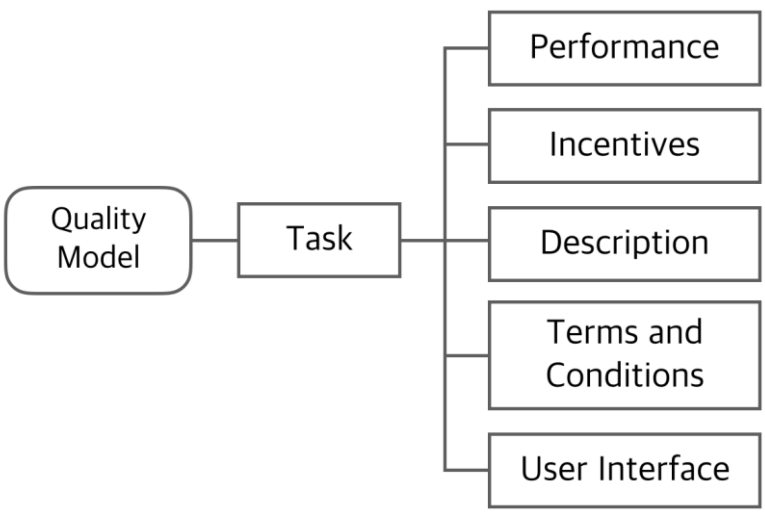
[그림 6] 수집 결과에 영향을 미치는 태스크 디자인 요소⑥

첫 번째, Definition은 참가자에게 주는 크라우드소싱 태스크에 대한 설명 정보이다. 간단하고 명확한 설명과 태스크 수행을 위한 요구

⑥ Allahbakhsh et al. (2013) 프레임워크 인용

조건들이 포함된다. 다음으로 User Interface는 참가자가 태스크에 접근할 수 있는 인터페이스를 의미한다. 세 번째, Granularity는 태스크의 복잡한 정도를 의미하며, 태스크의 양과도 연관된다. 마지막으로 Compensation policy는 태스크 수행에 대한 보상 방식을 뜻한다. 애타주의와 같은 내적 인센티브, 금전적 보상과 같은 외적 인센티브로 구성된다.

이러한 크라우드소싱 수집결과에 영향을 미치는 요인들은 최근 Daniel et al. (2018)에 의해 더 세분화되어 발전되었다. Daniel et al. (2018)은 크라우드소싱 수집결과에 영향을 미치는 요인들에 대해 연구한 결과들을 조사하여 [그림 7]과 같은 Quality Model을 제시했다.



[그림 7] 수집 결과에 영향을 미치는 태스크 디자인 요소에 대한 Quality Model[⑦]

첫 번째 요소인 Performance는 시간 및 비용 대비 수행할 수 있는

⑦ Daniel et al. (2018) Quality model 일부 인용

태스크의 양을 뜻한다. 두 번째 Incentives는 이전의 Compensation Policy와 동일하게 보상 방식을 의미한다. 세 번째, Description은 태스크에 대한 설명 방식으로 태스크 수행 방법을 포함한다. 네 번째, Terms and Conditions는 태스크 요청자와 참가자 간에 프라이버시 보호와 같은 일반적 규약을 의미한다. 마지막 User Interface는 이전과 동일하게 참가자가 접근할 수 있는 인터페이스를 뜻한다.

본 연구는 Daniel et al. (2018)에 의해 더 세분화된 Qulity Model에 기반하여 태스크 디자인 요소를 선정하였다. 먼저 Performace 측면에서는 Wang et al. (2012)이 크라우드소싱 기반의 학습데이터 생성에 있어 태스크 양과 수집 결과의 다양성 사이에 영향이 있을 수 있음을 논의했다. 또한 한 참가자에게 더 많은 패러프레이징을 얻을 수 있다면 중복된 문장을 입력하지 않아 더 많은 표현을 얻을 수 있을 것으로 기대된다. 따라서 본 연구에서는 태스크 양이 표현 다양성 수집에 미치는 영향을 보고자 하였다. 두 번째, Incentives의 경우에는 보상 방식에 따라 수집 결과에 영향을 준다는 연구(Chen et al, 2011)에 기반하여 연구 문제를 정하였다. 이에 따라 태스크 양과 관련지어 보너스 보상 방식이 표현 다양성 수집에 미치는 영향을 보고자 하였다. 보상 금액과 관련해서는 보상액이 수집 결과 향상에 영향을 미치지 않는다고 밝혀져(Mason and Watts, 2009) 모든 실험 조건에서 패러프레이징 문장 당 동일한 보상 금액을 제공했다. 세 번째 Description 측면에서는 태스크 수행에 대한 설명 방식과 충족 조건 제시가 포함되어 해당 내용을 포괄할 수 있는 Social Proof 개념을 적용하고자 하였다. 이와 관련해서는 4절에서 더 자세히 설명한다. 네

번째, Terms and Conditions 측면에서는 본 연구가 MTurk을 기반으로 진행되어 MTurk의 규약을 따르고자 하였다. 마지막으로 User Interface의 경우에는 본 연구의 태스크가 패러프레이징이므로 해당 태스크를 수행할 수 있는 인터페이스 내에 한정되었다. 참가자가 쉽게 접근할 수 있는 인터페이스로 녹음과 타이핑 방식이 제안되었으며, 녹음 방식은 음질, 외부 소음 등의 문제가 제기되어(Lane et al. 2010) 타이핑 인터페이스를 사용하였다.

제 4 절 Social Proof 효과

본 절에서는 Social Proof 효과에 대해 서술하고 본 연구의 크라우드소싱 태스크 설명방식으로써 해당 개념을 적용한 배경 및 목적에 대해 서술한다.

본 연구의 목적은 태스크 디자인 요소에 따라 크라우드소싱 기반의 학습데이터 생성 결과에 미치는 영향을 알아보고 표현의 다양성을 향상시킬 수 있는 태스크 디자인 방식을 도출하는 데 있다. Social Proof 효과는 제 3절에서 언급되었듯이 태스크 디자인 요소 중 Description 측면에서 적용하여 표현의 다양성 수집에 미치는 영향을 보고자 하였다.

Social Proof(or informational social influence)는 사람들이 주어진 상황에서 올바르게 행동하기 위하여, 다른 사람들의 행동을 참조하여 자신의 행동을 결정하는 심리적 현상을 뜻한다(Cialdini, 2009). 특히 자신의 행동 결과에 대해 불확실할 때, 다른 사람들의

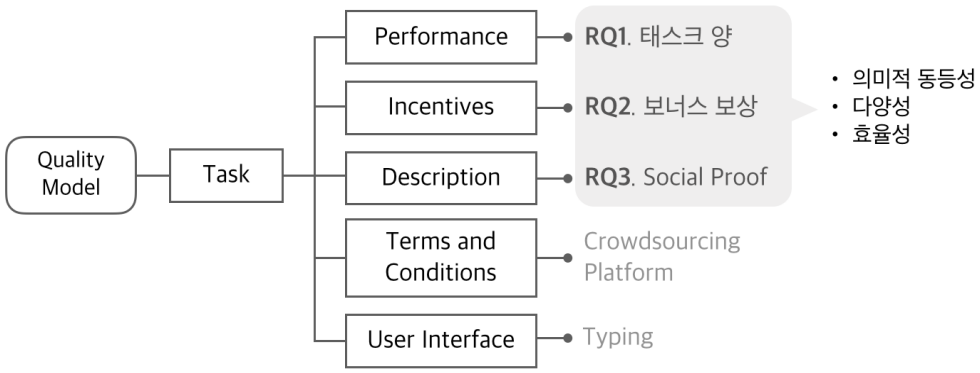
행동을 반영하려는 Social Proof 경향이 강해지는 것으로 나타났다(Wooten and Reed, 1998). 예를 들어 공연장이나 극장에서 누군가 박수를 치면 나머지 다른 청중들이 따라 박수를 치는 현상이다.

이러한 Social Proof 개념은 여러 연구들에서 사람들의 행동을 유도하는데 활용될 수 있음을 보여준다. Malu et al. (2012)는 사람들이 온라인 커뮤니티에 정보적인 내용보다 개인적인 내용을 더 올리도록 유도하기 위해 Social Proof 개념을 사용하였다. Das et al. (2014)는 사람들이 컴퓨터 보안 도구에 대해 더 인식하고 사용할 수 있도록 유도하는데 Social Proof 개념을 활용했다. 또한 Vashistha et al. (2018)은 Social Proof 개념을 사용하여 HCI 분야에서 긍정적인 피드백을 주려는 응답 편향 문제를 해결하고 비판적인 피드백을 얻고자 하였다.

일반적으로 크라우드소싱 태스크는 참가자가 처음 접하고 요청자가 요구하는 조건에 대해 불확실하기 때문에 Social Proof 효과가 반영될 것으로 가정할 수 있다. 또한 본 연구가 표현의 다양성 향상을 목적으로 하여 참가자가 다양한 표현을 입력하도록 행동을 유도하는 것이 중요하므로 Social Proof 개념을 설명 방식으로 활용하고자 한다. 이에 따라 태스크 설명 방식으로 타 참가자가 수행한 결과처럼 얻고자 하는 결과물을 제시하고, Social Proof 기반 설명 방식이 표현 다양성 수집에 미치는 영향을 고찰하고자 하였다.

제 3 장 연구 문제

본 연구의 연구 문제는 태스크 디자인 요소에 따라 [그림 8]과 같이 도식화할 수 있다. 5가지 태스크 디자인 요소(Performance, Incentives, Description, Terms and conditions, User interface)에 기반하여 제안되었다. 먼저 태스크 양이 수집된 학습데이터의 표현 다양성에 미치는 영향을 알아보고 이어 보너스 보상 방식, Social Proof 설명 방식을 적용하며 결과에 미치는 영향을 분석하는 구조를 갖는다.



[그림 8] 연구 문제 도식화[8]

먼저 연구 문제 1에서는 태스크 양을 달리 하였을 때, 수집한 데이터의 표현 다양성에 미치는 영향에 대해 알아본다. 표현 다양성 수집에 미치는 영향을 분석하고자 수집된 데이터가 수집하고자 한 데이터와 얼마나 같은 의미인지, 그리고 태스크 양에 따라 어떻게

[8] Daniel et al. (2018) Quality model 일부 인용

달라지는지 파악한다. 이후 같은 의미를 갖는 데이터 간에 서로 얼마나 다른 구조를 갖는지 분석한다. 마지막으로 데이터 수집을 얼마나 효율적으로 할 수 있는지 파악한다.

연구 문제 1. 태스크 양이 표현 다양성 수집에 어떻게 영향을 미치는가?

1.1 태스크 양에 따라 수집 데이터의 의미적 동등성이 어떻게 달라지는가?

1.2 태스크 양에 따라 수집 데이터 간 다양성이 어떻게 달라지는가?

1.3 태스크 양에 따라 데이터 수집의 효율성이 어떻게 달라지는가?

연구 문제 2에서는 태스크 양이 표현 다양성 수집에 미치는 영향을 고려하여 보너스 보상 방식을 적용하였을 때, 달라지는 변화 양상을 알아보고자 하였다. 이를 위해 연구 문제 1에서 도출한 태스크 양에 따른 표현 다양성 수집의 변화와 보너스 보상 방식을 추가 적용하였을 때의 표현 다양성 수집의 변화를 비교하고 태스크 디자인 개선 방향을 찾고자 하였다.

연구 문제 2. 추가 문장 당 보너스 보상 방식이 표현 다양성 수집에 어떻게 영향을 미치는가?

1.1 추가 문장 당 보너스 보상 방식에 따라 수집 데이터의 의미적 동등성이 어떻게 달라지는가?

1.2 추가 문장 당 보너스 보상 방식에 따라 수집 데이터 간

다양성이 어떻게 달라지는가?

1.3 추가 문장 당 보너스 보상 방식에 따라 데이터 수집의 효율성이 어떻게 달라지는가?

연구 문제 3에서는 Social Proof 개념을 이용하여 다른 참가자가 수행한 결과인 것처럼 태스크 설명을 제공하였을 때 표현 다양성 수집에 미치는 영향을 알아보고자 하였다. 연구 문제 1, 2를 통해 도출한 결과와 비교하여 해당 방식이 표현 다양성 수집에 주는 효과 및 개선 방향을 도출하고자 하였다.

연구 문제 3. Social Proof 기반 설명 방식이 표현 다양성 수집에 어떻게 영향을 미치는가?

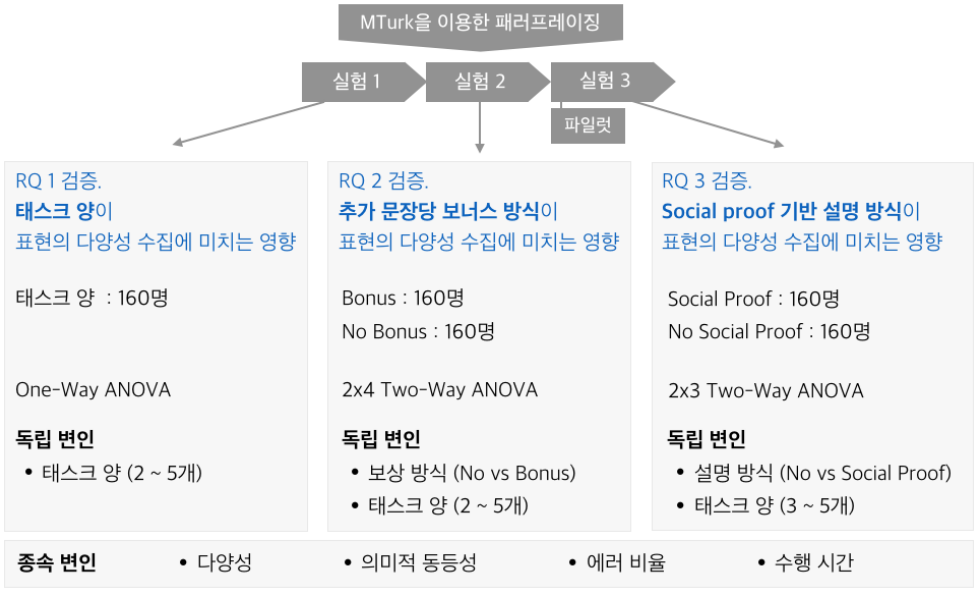
1.1 Social Proof 기반 설명 방식에 따라 수집 데이터의 의미적 동등성이 어떻게 달라지는가?

1.2 Social Proof 기반 설명 방식에 따라 수집 데이터 간 다양성이 어떻게 달라지는가?

1.3 Social Proof 기반 설명 방식에 따라 데이터 수집의 효율성이 어떻게 달라지는가?

제 4 장 연구 방법

본 연구는 크라우드 소싱 기반의 대화형 에이전트 개발을 위한 학습데이터 생성을 함에 있어, 수집 결과의 표현 다양성을 향상시킬 수 있는 태스크 디자인 제언을 목적으로 한다. 이를 위해 선행연구를 기반으로 선정한 태스크 양, 보너스 보상 방식, Social Proof 기반 설명 방식의 3가지 요인에 대한 일련의 실험을 진행했다[그림 9]. 본 장에서는 3가지 실험 설계 및 분석 방법에 대해 서술한다. 어떤 태스크를, 어떤 방식으로 요청했는지, 어떤 지표로 결과를 분석했는지에 대해 설명하고자 한다.



[그림 9] 연구 방법의 도식화

제 1 절 태스크 및 실험 절차

본 연구는 Amazon Mechanical Turk[⑨] (이하 MTurk)을 사용하여 참가자를 모집하고 영어로 태스크를 진행하였다. 본 연구의 연구문제는 태스크 디자인 방식에 따라 표현 다양성 수집에 미치는 영향을 보기 위함이므로, 크라우드소싱을 통한 학습데이터 생성 방식으로 활용가능성이 검증(Bapat et al, 2018)된 패러프레이징을 모든 실험 조건에 사용하였다. 패러프레이징 태스크란 하나의 문장을 프롬프트로 참가자에게 제공하고 해당 프롬프트에 대한 패러프레이징을 요청하는 방식이다. 본 연구에서 프롬프트로 제시된 문장은 레시피 검색 인텐트에 해당하는 문장이었다.

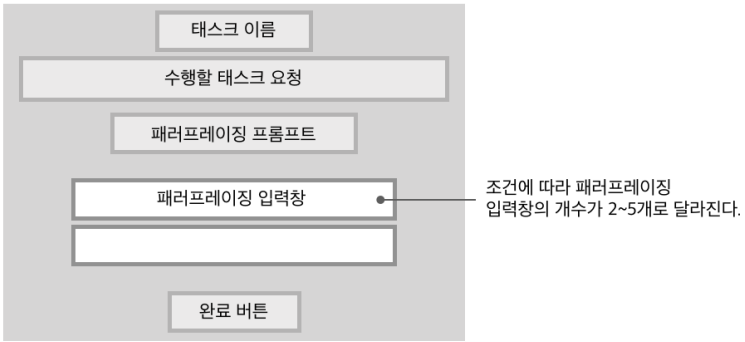
태스크 참가자들은 해당 태스크가 대화형 에이전트와 사용자 사이에 일어나는 대화를 수집하기 위한 목적임을 설명받았다. 또한 제시된 프롬프트를 패러프레이징하는 태스크 수행방식에 대해서도 설명을 받았다. 참가자들은 모든 실험 조건에 랜덤하게 할당되었으며, 학습 효과를 방지하고자 모든 실험 조건을 통틀어 한번만 참가할 수 있도록 하였다. 태스크가 영어 패러프레이징이기 때문에 다양한 대화 표현에 익숙하여야 하며, 대화형 에이전트와 사용자 사이의 대화 수집이 목적이므로 영어 의사소통이 능숙한 참가자를 모집하기 위해 참가자는 미국인으로 한정하였다. 선행 연구 및 파일럿 테스트 시 태스크 수행 시간을 종합적으로 참고하여 패러프레이징 문장 하나당 5센트를 리워드로 제공하였다.

⑨ https://www.mturk.com/

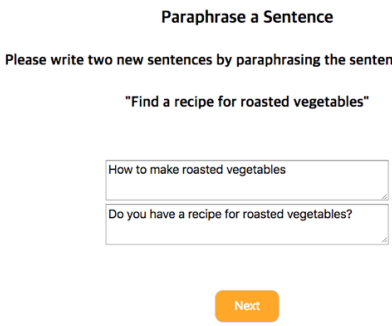
제 2 절 실험물

본 절에서는 각 실험 태스크에 사용되는 실험물에 대해 서술한다. 앞서 언급되었듯이 본 연구는 태스크 양, 보너스 보상 방식, Social Proof 기반 설명 방식에 따른 영향을 보고자 3가지 일련의 실험을 진행한다. 따라서 각 실험에 맞추어 태스크 실험물이 제작되었다. 각 실험물은 공통적으로 참가자가 패러프레이징을 입력할 수 있도록 하는 텍스트 입력창과 본 연구의 태스크 목적인 패러프레이징 요청 글이 포함되었다.

먼저 실험 1에서는 태스크 양에 따른 영향을 보기 위해 본 실험의 태스크에 해당하는 패러프레이징의 양을 달리하여 실험물을 제작했다. 참가자가 입력할 수 있는 텍스트 입력창을 2개에서 5개까지 달리하는 방식으로 패러프레이징의 양을 조절했다. 따라서 독립변인은 태스크 양(패러프레이징 양)에 해당한다. 실험물에 대한 시각화는 [그림 10]과 같다.



(a) 실험 1에 사용된 실험물 시각화

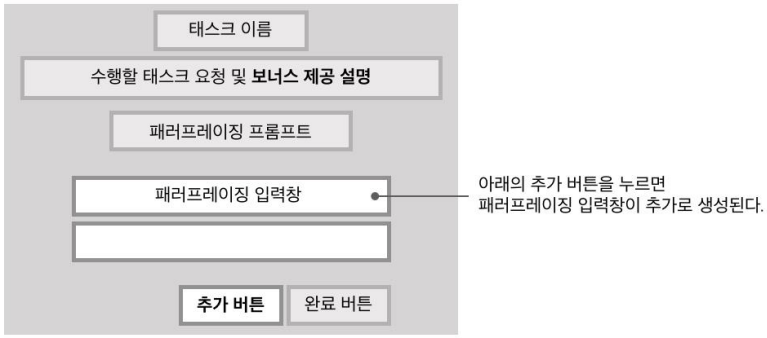


(b) 텍스트 입력창이 2개인 기본 실험물 예

(c) 조건에 따라 텍스트 입력창이 늘어난 실험물 예

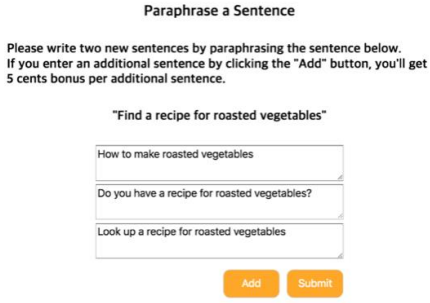
[그림 10] 실험1을 위해 제작된 실험물

다음으로 실험 2에서는 추가 문장 당 보너스 보상 방식에 따른 영향을 보기 위해 참가자가 추가로 텍스트를 입력할 수 있도록 추가 버튼을 만들었다. 참가자가 3개 이상의 패러프레이징을 입력하고자 하는 경우 추가 버튼을 누르면 텍스트 입력창이 생성된다. 실험 1과의 비교를 위하여 텍스트 입력창의 추가 생성은 최대 5개까지 지원하였다. 참가자가 추가 패러프레이징 입력 시 보너스 보상을 받는다는 사실을 알 수 있도록 보너스 제공에 대한 설명을 제공하였다. 추가 문장 당 보너스는 기본 문장 당 보상액과 동일하게 5센트를 제공하였다. 실험물에 대한 시각화는 [그림 11]과 같다.



(a) 실험 2에 사용된 실험물 시각화

(b) 추가 생성 버튼을 누르지 않은 초기 상태



(c) 추가 생성 버튼을 눌러 입력창이 늘어난 상태

[그림 11] 실험 2를 위해 제작된 실험물

마지막으로 실험 3에서는 Social Proof 기반 설명 방식에 따른 영향을 보기 위해 참가자에게 타 참가자가 수행한 결과처럼 서술한 결과물을 제공하였다. 이를 위해 참가자가 타 참가자의 수행 결과물로 인식하는지, 결과물을 얼마나 제공해야 할지에 대한 가이드를 얻고자 파일럿 테스트를 진행했다.

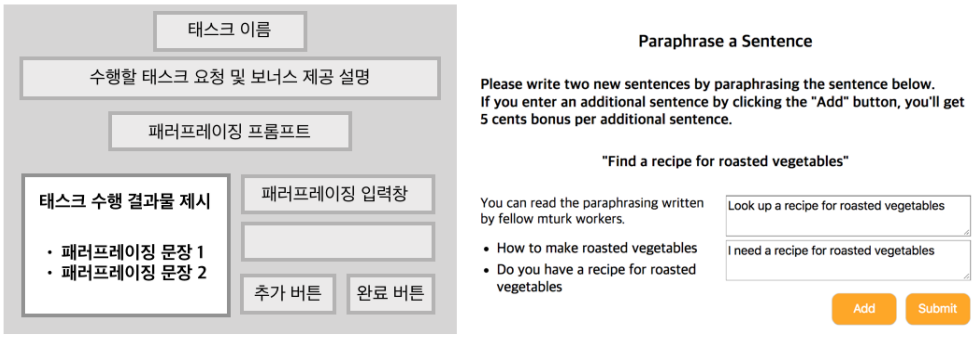
파일럿 테스트는 본 실험과 동일하게 MTurk으로 50명을 모집하여 진행되었다. 파일럿 참가자는 Social Proof 기반 설명 방식이 제공된 패러프레이징 태스크를 수행하였으며, 이후 Social Proof 기반 설명 방식에 대해 구조화 인터뷰를 진행했다. 구조화 인터뷰는 질문에 대한 답변을 텍스트 입력창에 입력하는 방식으로 진행되었다. 참가자에게 패러프레이징 태스크를 수행함에 있어 타 참가자의 결과물 제시가

도움이 되었는지, 인식을 하였는지에 대해 수집했다.

그 결과, 태스크 참가자는 크게 타 참가자가 수행한 결과처럼 제시된 결과물로부터 영감(Inspiration)과 참조(Reference)의 2가지 긍정적인 영향을 받았다. 영감을 얻은 참가자들은 제시된 결과물로부터 패러프레이징 아이디어를 얻거나 패러프레이징 유형으로부터 자신의 답변을 더 독창적으로 만드는데 도움이 되었다고 답하였다 (예 : "It sparked new ideas in my head." "(They) gave me ideas for what to write). 참조 유형의 경우에는 요청자가 요구하는 태스크 결과물에 대한 이해를 얻거나 패러프레이징 하는 방식 또는 예를 보여 준다고 답하였다 (예 : "Showed me what is expected from me" "They helped me understand what you were looking for"). 반면에 제시된 결과물이 자신의 패러프레이징 아이디어를 가져가 도움이 되지 않았다는 참가자들도 있었다. 이들은 제시된 결과물이 자신이 작성하려던 예를 포함하거나 너무 많은 아이디어를 제시하고 있다고 답하였다 (예 : "Too many, user up ideas I was having." "They just took away from ideas I might have had.").

따라서 실험 3에서는 파일럿 테스트의 결과 및 선행연구를 종합적으로 고려하여 2개의 패러프레이징 결과물 예시를 제공하였다. 또한 본 연구가 다양한 표현을 얻는데 목적이 있으므로, 제시된 결과물 예시 문장은 서로 다른 문장 구조 및 어휘를 갖도록 하였다. 실험 2와의 비교를 위하여 기본 2개 패러프레이징 문장에 추가 버튼을 눌러 문장을 입력하는 방식을 동일하게 사용하였다. 추가 문장 당 보너스도 동일하게 5센트를 제공하였다. 실험 3에서 사용된 실험물에 대한 시각화는 [그림

12]와 같다.



(a) 실험 3에 사용된 실험물 시각화

(b) 실제 실험3 참가자가 수행한 태스크 실험물

[그림 12] 실험 3을 위해 제작된 실험물 설명

제 3 절 측정 지표 및 분석 방법

본 절에서는 실험에서 측정한 데이터 지표들과 그 지표를 분석한 방법에 대해 서술한다. 본 연구에서 측정한 지표는 4가지로 구성되는데 표현의 다양성 정의에 기반한 2가지 지표(의미적 동등성, 다양성)와 효율성 분석을 위한 2가지 지표(수행 시간, 에러 비율)로 이루어진다. 각 측정 지표에 대한 자세한 설명은 이어 서술된다. 4가지 측정 지표는 각 실험의 종속 변인으로써 분석되었다.

1. 측정 지표

의미적 동등성

표현의 다양성은 의미가 같은 상태를 전제로 하므로 수집된 데이터가 프롬프트로 제시된 문장(얻고자 하는 인텐트)와 동일한 의미인지 평가되어야 한다. 이를 위해 IBM Watson Assistant API[10]를 사용하였다. 먼저 미국인 14명으로부터 실제 아마존 에코닷을 사용하면서 얻은 레시피 검색 인텐트 발화, 223개를 IBM Watson Assistant에 학습시켰다. 이후 레시피 검색 인텐트를 학습시킨 IBM Watson Assistant에 크라우드소싱으로 수집한 데이터를 입력시켜 레시피 검색 인텐트와의 신뢰도(Confidence) 점수를 비교하였다. 해당 점수는 높을수록 학습된 인텐트와 입력데이터가 같은 의미임을 뜻한다. 실제 대화형 에이전트와 사용자 사이의 발화데이터를 크라우드소싱으로 수집한 데이터와 비교함으로써 데이터 타당도를 높이고자 하였다. 또한 기존에 사람이 평가하던 방식에서 머신러닝 모델을 활용하여 평가함으로써 분석의 객관성을 확보하고자 하였다.

다양성

다양성은 표현의 다양성 정의에 기반하여 의미적으로 동등하다고 평가된 데이터(Confidence > 0.5)에 한해서 문장 구조 및 어휘의 유사성을 평가하였다. 이를 위해 가장 뛰어난 문장 구조(Syntactic) 유사성 분석 성능을 가진 것으로 연구된 spaCy API[11]의 유사도 점수를 사용하였다(Choi et al, 2015). 다양성이란 단어와의 혼동을 막고자 다양성 지표는 (1 − 유사도) 점수로 변환하였다. 다양성 점수가

[10] https://www.ibm.com/watson/kr-ko/developercloud/conversation.html
[11] https://spacy.io/

높을수록 서로 다른 표현임을 의미한다.

에러 비율

에러는 수집한 데이터 중 의미적 동등성을 만족하지 못 하거나 입력 개수를 모두 채우지 못한 경우로 정의하였다. 실제 데이터 상에서 에러로 분류된 참가자의 데이터 예시는 [표 1]과 같다. 에러 비율은 각 실험 조건별 총 데이터에서 에러가 발생한 비율을 의미한다. 실험 조건별 효율성을 분석하고자 측정하였다.

유형	예시 문장
제시한 프롬프트	Find a recipe for roasted vegetables
에러로 분류된 데이터	Roasted vegetables Dish with roasted vegetables What vegetables can be roasted together Vegetables roasted Roasted vegetable ideas

[표 1] 에러로 분류된 데이터 예시

수행 시간

수행 시간은 한 문장을 수집하는데 걸리는 시간으로, 총 태스크 수행시간을 입력한 문장 개수로 나누어 측정하였다. 데이터 편차가 커 중간값을 사용하였으며, 에러 비율과 마찬가지로 실험조건 별 효율성을

분석하고자 측정하였다.

수집한 데이터의 표현 다양성 향상을 의미하는 측정 지표별 점수를 정리하면 해당 점수 유형은 [표 2]와 같다. 의미적 동등성 지표가 높을수록 패러프레이징을 통해 얻은 데이터들이 프롬프트로 제시한(얻고자 하는) 인텐트와 같은 의미임을 뜻한다. 이와 비슷하게 다양성 지표에서도 다양성이 높을수록 같은 의미를 가진 데이터들이라도 그 데이터 사이에 더 다양한 표현(문장 구조 혹은 어휘의 다양함)을 가짐을 의미한다. 에러 비율은 높을수록 에러가 많아 학습할 수 있는 데이터가 적어지고 수행 시간도 마찬가지로 높을수록 데이터를 모으는데 오랜 시간이 소요되므로 낮을수록 태스크 효율성이 높아지는 것으로 해석한다.

측정 지표	의미적 동등성	다양성	에러 비율	수행시간
점수 유형	↑	↑	↓	↓

[표 2] 표현 다양성 향상을 의미하는 측정 지표별 점수 유형

2. 분석 방법

먼저 실험 1에서는 태스크 양에 따른 표현 다양성 수집에 미치는 영향을 보고자 하였다. 분석은 앞서 언급한 4가지 측정 지표를 종속변인으로 하여 태스크 양에 따른 변화를 보았다. 의미적 동등성과

다양성 지표는 ANOVA 테스트를 통해 태스크 양에 따라 유의미한 차이가 있는지 통계적 검정을 하였다. 에러 비율은 에러 발생 여부에 대한 카테고리 데이터이기 때문에 Chi-Square 테스트를 통해 태스크 양에 따라 에러 발생 여부의 차이가 있는지 보았다. 마지막으로 수행 시간 지표는 데이터의 왜도가 높아 비모수적 통계 방법인 Kruskal-Wallis rank sum 테스트를 통해 태스크 양에 따른 유의미한 차이를 통계적 검정하였다.

다음으로 실험 2에서는 보너스 보상 방식에 따른 표현 다양성 수집에 미치는 영향을 보고자 하였다. 보너스 보상 방식을 통해 실험 1과 비교하여 수집 결과 향상 여부와 요소 간 상호작용을 보기 위하여 태스크 양, 보너스 보상 방식을 독립변인하고, 이에 따른 4가지 지표를 종속 변인으로 분석하였다. 이에 따라 실험 1 결과를 보너스 보상 방식을 제공하지 않은 그룹으로 두고, 실험 2를 통해 수집한 결과를 보너스 보상 방식을 제공한 그룹으로 하여 분석하였다. 의미적 동등성 및 다양성, 수행시간 분석은 2가지 독립변인에 따른 그룹 간 차이를 보고자 하므로 2x4 Two-Way ANOVA 테스트를 통해 각 독립 변인에 따른 차이 및 상호 작용을 보았다. 에러 비율은 탐색적 데이터 분석을 실시하고 보너스 보상 방식에 의해 태스크 양에 따른 에러 비율이 어떻게 달라지는지 보기 위해 Chi-Square 테스트로 통계적 검정을 하였다.

마지막으로 실험 3에서는 Social Proof 기반 설명 방식이 표현 다양성 수집에 미치는 영향을 보고자 하였다. 실험 2에서의 설계와 동일하게 실험 2 결과와 비교하여 수집 결과의 향상과 요소 간

상호작용을 보고자 Social Proof 기반 설명 방식, 태스크 양을 독립변인으로 하여 4가지 측정 지표를 종속변인으로 분석하였다. 이를 위하여 실험 3의 분석에서도, 실험 2의 결과를 Social Proof 기반 설명 방식을 적용하지 않은 그룹으로 두고 실험 3을 통해 수집한 결과를 Social Proof 기반 설명 방식을 적용한 그룹으로 분석하였다. 이 때, 보너스 보상 여부 조건을 일치시키고자 보너스 보상을 받게 되는 3~5개의 수집 결과를 분석하였다. 의미적 동등성, 다양성 및 수행 시간은 2x3 Two-Way ANOVA 테스트를 통해 각 독립변인에 따른 차이 및 상호작용을 보았다. 에러 비율은 실험 2와 동일하게 분석되었다.

제 5 장 연구 결과

본 장에서는 제 4장의 연구 방법을 바탕으로 실험한 3가지 일련의 연구 결과에 대해 설명한다. 이를 통해 본 연구의 연구문제에 해당하는 1) 태스크 양, 2) 보너스 보상 방식, 3) Social Proof 기반 설명 방식이 표현의 다양성 수집에 미치는 영향을 밝히고자 하였다. MTurk을 통해 총 480명을 모집하였으며, 73.65 달러를 사용하여 1473개의 패러프레이징 문장을 수집하였다. 4가지 측정 지표(의미적 동등성, 다양성, 에러 비율, 수행 시간)을 중심으로 분석하였으며, 통계적으로 유의미한 차이를 검증하기 위하여 통계 분석 방법이 사용되었다.

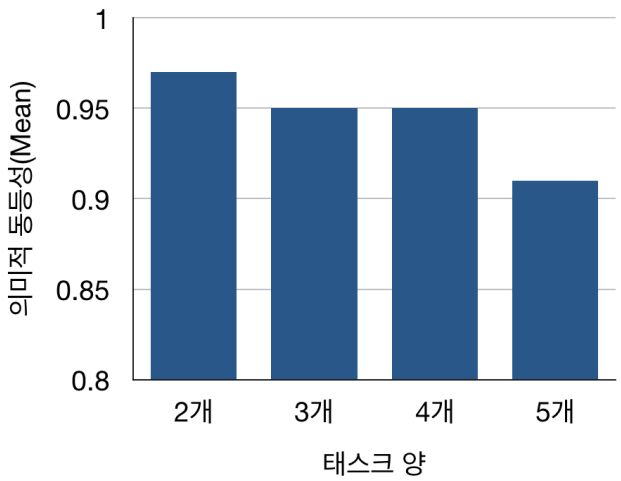
제 1 절 태스크 양에 따른 수집 결과

본 절에서는 연구 문제 1. 태스크 양이 표현의 다양성 수집에 미치는 영향에 대해 서술한다. 연구방법에서 밝혔듯 태스크는 패러프레이징이며, 따라서 태스크 양은 참가자에게 요청한 패러프레이징 양을 뜻한다.

MTurk을 통해 실험 조건당 40명을 모집하여 총 160명으로부터 560개의 패러프레이징 문장을 수집하였다. 이 중 태스크와 무관한 데이터를 입력한 8명의 데이터를 제외하고 529개의 패러프레이징 문장을 분석하였다. 529개의 학습데이터 수집을 위해 26.45달러를 사용하였다.

RQ 1.1 태스크 양이 의미적 동등성에 미치는 영향

연구 문제 1.1에 해당하는 태스크 양이 의미적 동등성에 미치는 영향을 알아보고자 태스크 양에 따른 참가자별 의미적 동등성을 평균내어 분석하였다. 이에 대한 결과를 살펴보면 태스크의 양이 '2개'인 경우 0.97(SD = 0.06), '3개'인 경우 0.95(SD = 0.07), '4개'인 경우 0.95(SD = 0.06), '5개'인 경우 0.91(SD = 0.15)로 나타났다. 이에 대한 결과는 [그림 13]을 통해 보여지는데, 태스크 양이 늘어남에 따라 의미적 동등성이 떨어지는 경향을 확인할 수 있다.



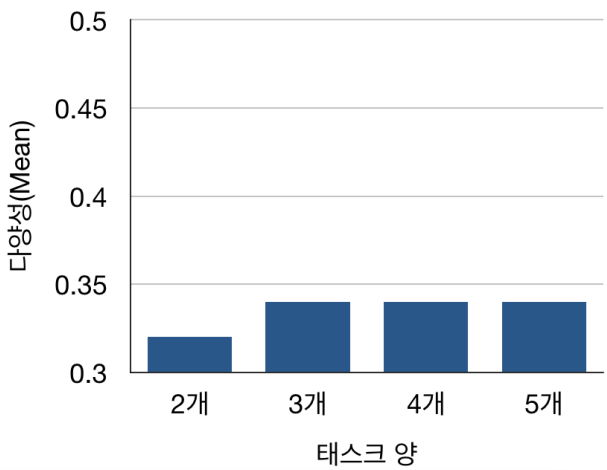
[그림 13] 태스크 양에 따른 의미적 동등성

통계적으로도 태스크 양에 따라 의미적 동등성 값 간에 유의미한 차이가 있는지 알아보고자 ANOVA 테스트를 실시하였다. ANOVA 테스트 결과, $F(3, 148) = 3.06$, $p < 0.05$로 태스크 양이 의미적 동등성에 미치는 영향이 유의미함을 보여주었다. 이후 어떤 그룹 간에

유의미한 차이가 있는지 알아보고자 사후검증 방식으로 TukeyHSD 테스트를 실시하였다. 그 결과 태스크 양이 2개인 그룹과 5개 그룹 사이에 유의미한 차이(p = 0.019)가 있다고 나타났다. 이러한 결과는 앞서 태스크 양이 늘어날수록 의미적 동등성이 떨어지는 경향과 일치한다.

RQ 1.2 태스크 양이 다양성에 미치는 영향

연구 문제 1.2에 해당하는 태스크 양이 다양성에 미치는 영향을 알아보고자 태스크 양에 따라 참가자별 다양성 점수를 평균 내어 분석한 결과 [그림 14]와 같이 나타났다. 태스크 양이 '2개'인 경우 0.32(SD = 0.21), '3개'인 경우 0.34(SD = 0.16), '4개'인 경우 0.34(SD = 0.09), '5개'인 경우 0.34(SD = 0.12)로 나타났다. 2개인 경우 다른 경우보다 값이 작지만 그 차이가 0.02로 거의 차이가 나타나지 않았다.

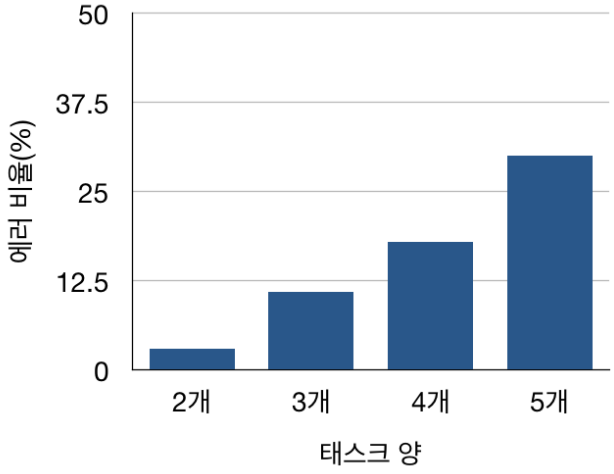


[그림 14] 태스크 양에 따른 다양성

ANOVA 테스트를 통한 통계적 유의미성 검증에서도 $F(3, 126) = 0.118$, $p > 0.05$로 유의미한 차이가 나타나지 않았다. 따라서 태스크 양은 수집 결과의 다양성에 영향을 미치지 못하는 것으로 파악된다.

RQ 1.3 태스크 양이 수집 효율성에 미치는 영향

연구 문제 1.3에 해당하는 태스크 양이 수집 효율성에 미치는 영향을 알아보기 위해 에러 비율과 수행 시간을 분석하였다. 먼저 에러 비율은 타 지표들과 달리 태스크 양에 따른 차이가 크게 나타났다. 태스크 양에 따라 에러 비율을 측정한 결과는 [그림 15]와 같다. 태스크 양이 '2개'인 경우 3%, '3개'인 경우 11%, '4개'인 경우 18%, '5개'인 경우 30%로 분석되었다. 태스크 양이 늘어남에 따라 에러 비율이 크게 증가하는 것을 확인할 수 있다.

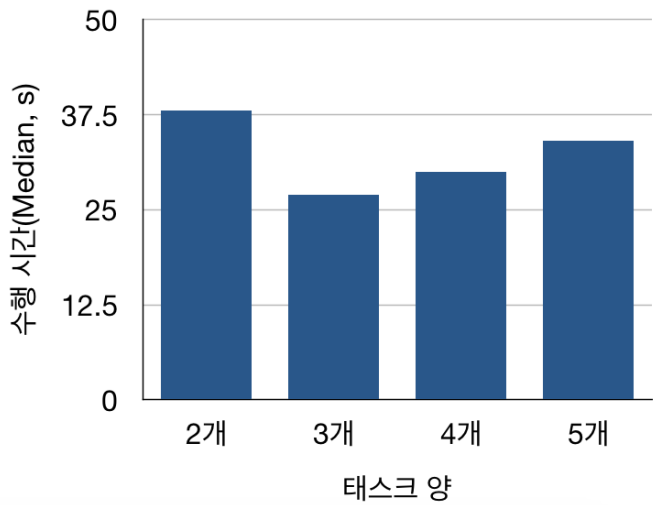


[그림 15] 태스크 양에 따른 에러 비율

통계적으로도 Chi-Square 테스트를 실시한 결과(에러 비율은 에러 여부를 판단하는 카테고리 데이터이기 때문에 Chi-Square 테스트를 사용하였다.), $\chi^2 = 8.8971$, df =3, $p < 0.05$로 태스크의 양에 따라 에러 발생 여부에 유의미한 차이가 있다고 나타났다.

태스크 양에 따른 평균 수행 시간은 [그림 16]과 같다. 태스크 양이 '2개'인 경우 38초, '3개'인 경우 27초, '4개'인 경우 30초, '5개'인 경우 34초로 분석되었다. 처음 '2개'인 경우보다 태스크 양이 많을 때 수행 시간이 더 적게 걸리는 것으로 보여진다.

하지만 통계적으로 유의미성을 검증해본 결과 그룹 간에 유의미한 차이가 나타날 정도로 차이가 있지는 않았다(수행시간은 데이터의 왜도가 높아 Kruskal-Wallis rank sum 테스트를 실시하였으며 $\chi^2 = 4.1275$, df =3, $p > 0.05$).



[그림 16] 태스크 양에 따른 수행 시간

에러 비율과 수행 시간을 종합해보면, 태스크 양이 늘어날수록 수행

시간이 감소하기는 하지만 그 차이가 유의미할 정도로 크지 않고 오히려 에러 비율이 높아지는 모습이 나타났다. 따라서 태스크 양을 늘려 한 사람에게 데이터를 더 얻을 수는 있지만 그만큼 에러 비율이 높아져 수집 효율성이 떨어지는 것으로 파악된다.

소결론

실험조건	의미적 동등성 (Mean, SD)	다양성 (Mean, SD)	에러비율 (%)	수행시간 (s)
2개	0.97 0.06	0.32 0.21	3	38
3개	0.95 0.07	0.34 0.16	11	27
4개	0.95 0.06	0.34 0.09	16	30
5개	0.91 0.15	0.34 0.12	30	34

[표 3] 태스크 양에 따른 측정 지표 요약

실험 1에서는 연구문제 1. 태스크 양에 따른 표현 다양성 수집에 미치는 영향을 알아보고자 태스크 양에 따른 4가지 지표의 변화를 분석하였다[표 3]. 분석 결과, 태스크 양에 따라 다양성이나 수행 시간 측면에서는 유의미한 차이가 나타나지 않았다. 즉, 한 참가자에게 패러프레이징을 통해 더 많은 학습데이터를 수집하여도 다양성이나 수행

시간 측면에서 문제가 되지 않는다. 반면, 태스크 양이 늘어날수록 의미적 동등성이 떨어지고 에러 비율이 높아지는 결과가 나타났다. 따라서 표현의 다양성은 같은 의미를 갖는 상태에서 문장의 구조나 어휘의 다양성을 의미하므로, 단순히 태스크 양을 늘리는 것은 한 사람에게 더 많은 데이터를 얻을 수는 있지만 표현의 다양성 조건에 따라 수집되는 학습데이터 측면에서는 효율적이지 못한 방식으로 밝혀졌다.

이러한 결과는 태스크 양이 늘어날수록 에러 비율은 증가하지만 다양성에는 차이가 없는 것으로 보아 태스크 완료 조건이 원인으로 보인다. 태스크 양이 늘어나면 그만큼 같은 의미를 갖는 다른 표현을 생성해야 하는데, 같은 의미를 갖게 만드는 문장의 양이 많아짐에 따라 구조나 어휘를 변환하는 과정에서 의미가 달라지는 것이다. 이러한 차이는 참가자마다 달리 나타나고 있어 패러프레이징을 통해 학습데이터를 수집할 수 있는 양이 참가자마다 다를 수 있음을 시사한다.

제 2 절 보너스 보상 방식에 따른 수집 결과

본 절에서는 연구 문제 2. 보너스 보상 방식이 표현의 다양성 수집에 미치는 영향에 대해 서술한다. 보너스 보상 방식과 태스크 양에 따른 변인을 둘 다 고려하여 각 변인에 따른 영향과 상호작용을 보고자 하였다.

보너스 보상을 제공한 방식은 MTurk을 통해 총 160명을 모집하여,

태스크와 무관한 데이터를 입력한 17명의 데이터를 제외하고 442개의 패러프레이징 문장을 수집하였다. 이를 위해 22.1달러를 사용하였다. 추가 버튼을 눌러 수행한 태스크의 양별 참가자 수는 [표 4]와 같다.

태스크 양	참가자 수 (명)
2개	67
3개	27
4개	18
5개	31

[표 4] 추가 버튼을 눌러 수행한 태스크 양별 참가자 수

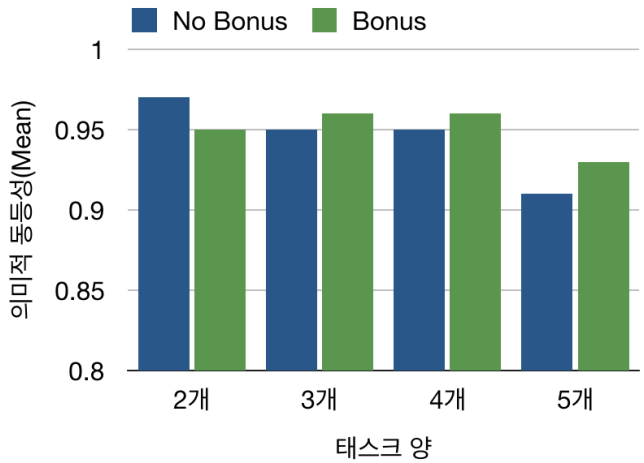
RQ 2.1 보너스 보상 방식이 의미적 동등성에 미치는 영향

보너스 보상 방식과 태스크 양, 그리고 두 요소 간에 상호작용이 의미적 동등성에 미치는 영향을 보기 위해 2x4 Two-Way ANOVA 테스트(2 : 보너스 보상 방식 여부, 4 : 태스크 양)을 실시 했다.

그 결과, 보너스 보상 방식 여부에 따라서는 의미적 동등성 지표에 유의미한 차이가 나타나지 않았다($F(1, 287) = 0.422$, $p = 0.52$). 그러나 태스크 양에 따른 의미적 동등성 차이는 유의미하게 나타났다($F(3, 287) = 2.981$, $p < 0.05$). TukeyHSD 테스트를 통해 사후검증 해본 결과, 태스크 양이 2개인 그룹과 5개 그룹 사이에서

유의미한 차이가 나타났다(p.adj = 0.03). 이는 실험 1에서 태스크 양에 따라 의미적 동등성 차이가 난 결과와 동일하며, 태스크 양이 많을 때 태스크 양이 적을 때보다 유의미하게 의미적 동등성이 낮아진다고 볼 수 있다. 두 변인 간 상호작용에 따른 의미적 동등성의 유의미한 차이는 나타나지 않았다(F(3, 287) = 1.12, p = 0.34).

이러한 결과는 보너스 보상 방식이 의미적 동등성에 유의미한 영향을 주지 못함을 보여준다. 그러나 통계적으로는 유의미한 차이가 아니지만 보너스 보상 여부에 따른 의미적 동등성 지표를 나타낸 [그림 17]을 보면, 보너스 보상 방식을 제공할 때 태스크 양이 5개에서 감소하는 의미적 동등성 폭이 줄어드는 것을 볼 수 있다. 실제 보너스 보상 방식을 제공 할 때, 의미적 동등성의 값에서도 태스크의 양이 '2개'인 경우 0.95(SD = 0.09), '3개'인 경우 0.96(SD = 0.06), '4개'인 경우 0.96(SD = 0.04), '5개'인 경우 0.93(SD= 0.08)으로 태스크 양에 따라 의미적 동등성 지표의 차이가 크게 나타나지 않았다.



[그림 17] 보너스 보상 여부에 따른 의미적 동등성 비교

RQ 2.2 보너스 보상 방식이 다양성에 미치는 영향

보너스 보상 방식과 태스크 양, 그리고 두 요소 간에 상호작용이 다양성에 미치는 영향을 보기 위해 2x4 Two-Way ANOVA 테스트(2 : 보너스 보상 방식 여부, 4 : 태스크 양)을 실시 했다.

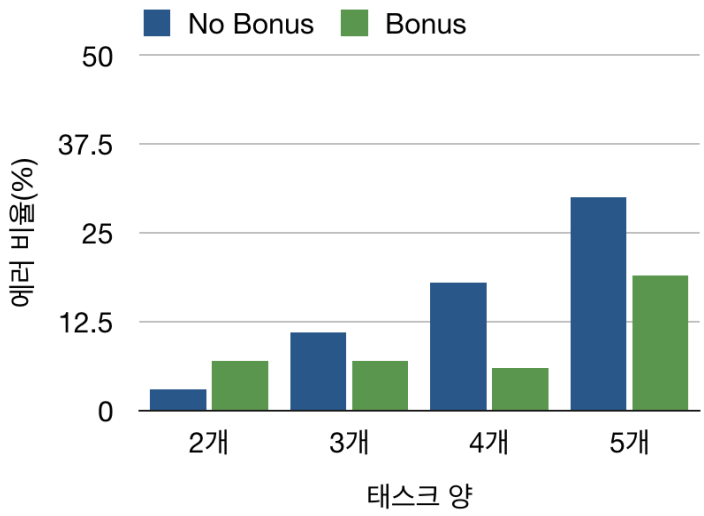
그 결과, 보너스 보상 방식 여부에 따라서는 다양성 지표의 유의미한 차이가 나타나지 않았다($F(1, 251) = 0.033$, $p = 0.86$). 태스크 양에 따라서도 다양성 지표가 유의미하게 차이가 나지 않았다($F(3, 251) = 0.524$, $p = 0.67$). 상호작용 면에서도 $F(3, 251) = 1.303$, $p = 0.27$로 유의미한 차이가 나타나지 않았다.

이러한 결과는 실험 1에서 태스크 양에 따라 나타난 다양성 영향과 동일하게 보너스 보상 방식이 다양성에 유의미한 영향을 주지 못함을 보여준다.

RQ 2.3 보너스 보상 방식이 수집 효율성에 미치는 영향

연구 문제 2.3에 해당하는 보너스 보상 방식이 수집 효율성에 미치는 영향을 알아보기 위해 에러 비율과 수행 시간을 분석하였다. 먼저 에러 비율을 보면 보너스 보상 방식은 태스크 양에 따른 에러 비율 차이를 감소시키는 것으로 나타났다. 이에 대한 분석 결과는 [그림 18]과 같다. 보너스를 제공하지 않은 그룹의 에러 비율이 태스크 양이 늘어날수록 증가하는 반면 보너스를 제공한 그룹에서는 태스크 양이 4개까지는 비슷하다가 5개부터 증가한 것을 볼 수 있다. 실제 에러

비율의 값에서도 태스크 양이 '2개'인 경우 7%, '3개'인 경우 7%, '4개'인 경우 6%, '5개'인 경우 19%로 차이가 줄어든 것을 볼 수 있다.

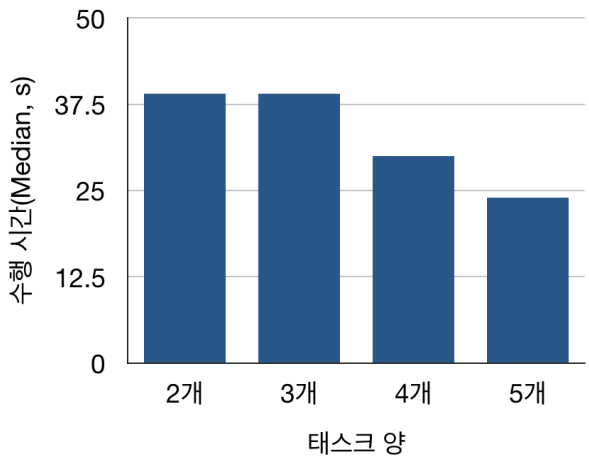


[그림 18] 보너스 보상 여부에 따른 에러 비율 비교

통계적으로도 보너스 보상 방식을 제공한 경우에는 태스크 양에 따라 에러 비율의 차이가 유의미하게 나타나지 않았다(Fisher's Exact 테스트 결과 $p > 0.05$, 기대도수가 5 미만인 값이 전체의 20% 이상 있어 Fisher's Exact 테스트를 실시 했다).

다음으로 보너스 보상 방식, 태스크 양, 두 변인의 상호작용이 수행 시간에 미치는 영향을 알아보고자 2x4 Two-Way ANOVA 테스트를 실시했다. 보너스 보상 방식, 태스크 양, 상호작용 모두 수행시간에 유의미한 차이가 나타나지 않았다(보너스 보상 방식 : $F(1,287) = 0.712$, $p = 0.4$ / 태스크 양 : $F(3, 287) = 1.046$, $p = 0.37$ / 상호작용 : $F(3, 287) = 0.615$, $p = 0.61$). 이러한 결과는 보너스 보상 방식이 수행 시간에 통계적으로 유의미한 영향을 주지 않음을 보여준다.

하지만 보너스 보상 방식을 제공했을 때 태스크 양에 따른 수행 시간을 보면 태스크 양이 '2개'인 경우 39초, '3개'인 경우 39초, '4개'인 경우 30초, '5개'인 경우 24초로 나타난다[그림 19]. 이는 태스크 양이 늘어날수록 패러프레이징 문장 하나를 수집하는데 걸리는 평균 시간이 감소함을 보여준다[그림 22].



[그림 19] 태스크 양에 따른 수행 시간(보너스 보상 방식)

통계적으로도 태스크 양이 늘어날수록 수행시간 감소에 유의미한 차이를 보였다(Kruskal-Wallis rank sum 테스트 결과, $\chi^2 = 25.871$, df = 3, $p < 0.001$).

종합해보면 보너스 보상 방식은 태스크 양이 늘어날수록 증가하던 에러 비율을 감소시키고, 수행 시간은 오히려 양이 늘어날수록 감소시키는 경향을 보인다. 이는 한 사람에게 얻을 수 있는 데이터를 늘리면서 문장 당 수집 시간도 단축시키므로 보너스 보상 방식이 수집 효율성을 높인다고 파악된다.

소결론

실험 조건	의미적 동등성 (Mean, SD)		다양성 (Mean, SD)		에러비율 (%)		수행시간 (s)	
	No Bonus	Bonus	No Bonus	Bonus	No Bonus	Bonus	No Bonus	Bonus
2개	0.97 0.06	0.95 0.09	0.32 0.21	0.36 0.18	3	7	38	39
3개	0.95 0.07	0.96 0.06	0.34 0.16	0.31 0.11	11	7	27	39
4개	0.95 0.06	0.96 0.04	0.34 0.09	0.31 0.11	16	7	30	30
5개	0.91 0.15	0.93 0.08	0.34 0.12	0.32 0.11	30	19	34	24

[표 5] 보너스 보상 방식 여부에 따른 측정 지표 요약

실험 2에서는 보너스 보상 방식이 표현 다양성 수집에 미치는 영향을 알아보고자 4가지 지표를 기반으로 분석하였다[표 5]. 그 결과, 보너스 보상 방식은 태스크 양이 늘어날수록 의미적 동등성이 감소하는 폭을 줄여주는 것으로 나타났다. 다양성 측면에서는 차이가 없었지만 태스크 양이 늘어날수록 증가하던 에러 비율을 감소시키는 것으로 밝혀졌다. 수행 시간 지표에서도 태스크 양이 늘어날수록 문장 당 수집시간이 감소하는 것으로 나타났다. 따라서 보너스 보상 방식은 수집 데이터의 다양성에 영향을 주지는 않지만 수집되어지는 양을 늘려주고 효율성을 높여준다는 점에서 효과적인 디자인 방식으로 파악된다.

실험 1에서 참가자마다 해당 태스크의 수행 능력에 차이가 있을 것으로 언급되었는데, 실험 2에서는 이러한 현상이 더 뚜렷하게 발견되었다. 태스크 양을 참가자가 자율적으로 선택하였을 때, 태스크

양이 많더라도 1) 의미적 동등성이 떨어지지 않고 2) 에러 비율이 덜 높아진 점 3) 문장 당 수집시간이 오히려 감소한 점을 통해 참가자의 패러프레이징 수행 능력 개인차 고려가 표현 다양성 수집을 위한 태스크 디자인에 중요 사항임을 알 수 있다.

제 3 절 Social Proof 기반 설명 방식에 따른 수집 결과

본 절에서는 본 연구의 마지막 연구 문제인 Social Proof 기반 설명 방식이 표현의 다양성 수집에 미치는 영향에 대해 서술한다. Social Proof 기반 설명 방식과 태스크 양에 따른 변인을 둘 다 고려하여 각 변인에 따른 영향과 상호작용을 보고자 하였다.

Social Proof 기반 설명 방식에 대한 데이터 수집은 MTurk을 통해 총 160명을 모집하여, 태스크와 무관한 데이터를 입력한 3명의 데이터를 제외하고 502개의 패러프레이징 문장을 수집하였다. 이를 위해 25.1달러를 사용하였다. 수집한 데이터 중 실험 2와의 비교와 Social Proof 기반 설명 방식과 태스크 양의 두 가지 변인에 대한 상호작용을 보기 위해 보너스 보상 여부 조건을 일치시키고자 보너스 보상을 받게 되는 3~5개의 수집 결과를 분석하였다. 따라서 Social Proof 기반 설명 방식을 제공하지 않고 76명으로부터 수집한 308개의 패러프레이징 문장, Social Proof 기반 설명 방식을 제공하고 수집한 91명으로부터의 370개의 패러프레이징이 분석되었다.

참가자별 추가 버튼을 눌러 수행한 태스크의 양은 [표 6]과 같다. Social Proof 기반 설명 방식을 적용하기 위해 2개의 문장을 예시로

제시한 방식이 참가자가 추가 버튼을 눌러 수행하는 태스크의 양에 영향을 미치는지 보았다. 실험 2, 3에서 무관한 태스크를 입력한 참가자 차이가 15명이고 각 참여 비율 경향에 차이가 나지 않아, 예시 제공 개수가 참가자가 수행하는 태스크 양에 영향을 미치지 않는 것으로 분석된다.

태스크 양	실험 2. 참가자 수 (명, %)	실험 3. 참가자 수 (명, %)
3개	27, 19	33, 21
4개	18, 13	19, 12
5개	31, 22	39, 25

[표 6] 참가자가 참여한 태스크 양(실험 2와 실험 3 비교)

RQ 3.1 Social Proof 기반 설명 방식이 의미적 동등성에 미치는 영향

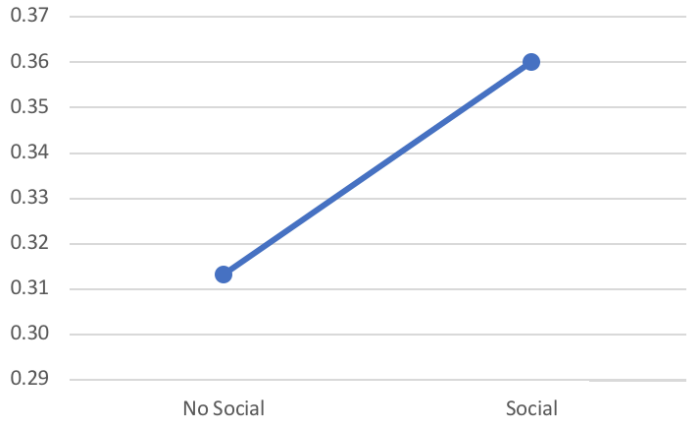
Social Proof 기반 설명 방식과 태스크 양, 그리고 두 요소 간에 상호작용이 의미적 동등성에 미치는 영향을 보기 위해 2x3 Two-Way ANOVA 테스트(2 : Social Proof 기반 설명 제공 여부, 3 : 태스크 양)을 실시했다. 그 결과, Social Proof 기반 설명 방식, 태스크 양, 상호작용 모두 의미적 동등성에 유의미한 차이를 만들지 못 하는 것으로 나타났다. Social Proof 기반 설명 방식에 따른 의미적 동등성은 F(1,

161) = 1.492, p = 0.22로 유의미한 차이를 보여주지 못 했다. 태스크 양에 따른 의미적 동등성에서도 F(2, 161) = 2.043, p = 0.13로 유의미한 차이가 없었다. 마지막으로 상호작용 측면에서도 F(2, 161) = 0.015, p = 0.99로 유의미한 차이가 나타나지 않았다.

이러한 결과는 Social Proof 기반 설명 방식이 수집 결과의 의미적 동등성에 유의미한 영향을 미치지 않음을 보여 준다.

RQ 3.2 Social Proof 기반 설명 방식이 다양성에 미치는 영향

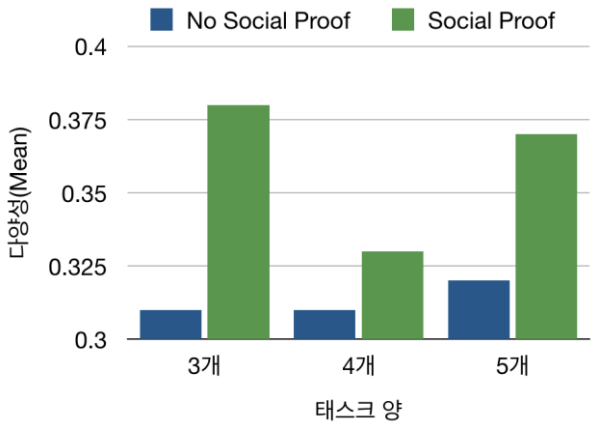
Social Proof 기반 설명 방식과 태스크 양, 그리고 두 요소 간에 상호작용이 다양성에 미치는 영향을 보기 위해 2x3 Two-Way ANOVA 테스트(2 : Social Proof 기반 설명 제공 여부, 3 : 태스크 양)을 실시 했다. 그 결과, Social Proof 기반 설명 방식을 제공하는 여부에 따라 다양성 지표에 유의미한 차이가 나타났다(F(1, 136) = 8.432, p < 0.01). 이 경우 Social Proof 기반 설명 방식을 제공할 때, 제공하지 않은 경우보다 유의미하게 더 다양한 표현을 얻을 수 있다고 볼 수 있다[그림 20].



[그림 20] Social Proof 기반 설명 방식 제공 여부에 따른 다양성 비교

이는 Social Proof 효과가 더 다양한 표현을 수집하는 데 활용될 수 있음을 보여준다. 태스크 양에 따라서는 다양성에 유의미한 차이가 나타나지 않았다($F(2, 136) = 0.474$, $p = 0.62$). 상호작용 면에서도 유의미한 차이가 나타나지 않았다($F(2, 136) = 0.48$, $p = 0.62$).

이러한 결과는 Social Proof 기반 설명 방식이 다양성에 유의미한 영향을 미치며, Social Proof 기반 설명 방식을 제공할 때 더 다양한 표현을 수집할 수 있음을 보여준다. 이는 Social Proof 기반 설명을 제공하는 여부에 따라 각 태스크 양에서의 다양성 지표를 비교하면 더 뚜렷한 차이를 볼 수 있다[그림 21].

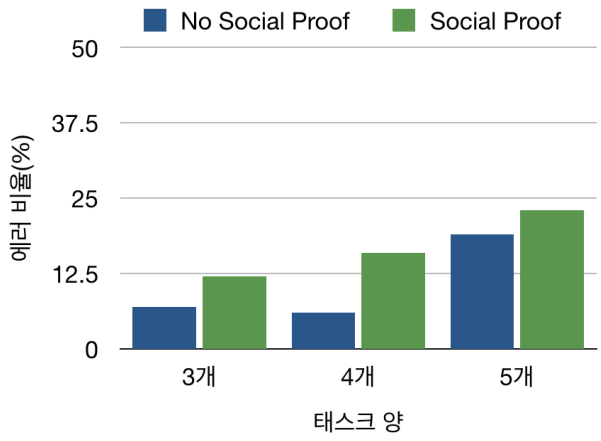


[그림 21] Social Proof 기반 설명 제공 여부에 따른 태스크 양별 다양성 비교

RQ 3.3 Social Proof 기반 설명 방식이 수집 효율성에 미치는 영향

연구 문제 3.3에 해당하는 Social Proof 기반 설명 방식이 수집 효율성에 미치는 영향을 알아보기 위해 에러 비율과 수행 시간을 분석하였다.

먼저 Social Proof 기반 설명 방식이 에러 비율에 미치는 영향을 보면 Social Proof 기반 설명 방식을 제공할 때, 제공하지 않을 때보다 에러 비율이 높아지는 것으로 나타났다[그림 22].



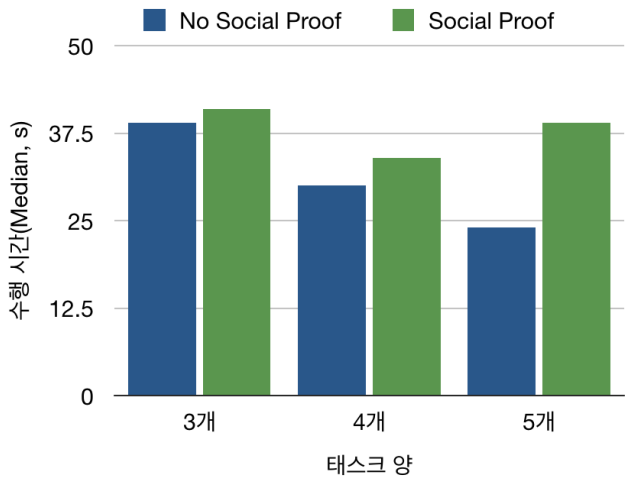
[그림 22] Social Proof 기반 설명 제공 여부에 따른 태스크 양별 에러 비율 비교

각 태스크 양별 Social Proof 기반 설명 방식에 따른 에러 비율 차이를 보면 '3개'인 경우 5%, '4개'인 경우 10%, '5개'인 경우 4%로 Social Proof 기반 설명 방식을 제공할 때 에러가 높아지는 것을 볼 수 있다.

수행 시간에 Social Proof 기반 설명 방식이 미치는 영향은 2x3 Two-way ANOVA 테스트를 통해 분석했다. 그 결과, Social Proof 기반 설명 방식, 태스크 양, 상호작용 모두 수행 시간에 유의미한 차이를 주지 못 했다. Social Proof 기반 설명 방식에 따라서는 $F(1, 161) = 0$, $p = 1.0$으로 유의미한 차이가 나지 않았다. 태스크 양에 따라서는 $F(2, 161) = 2.514$, $p = 0.08$로 유의미한 차이가 나지 않았다. 상호작용에 의한 결과에서도 $F(2, 161) = 1.159$, $p = 0.32$로 유의미한 차이가 나타나지 않았다. 이러한 결과를 통해 Social Proof 기반 설명 방식이 수행 시간에 유의미한 영향을 주지 않음을 알 수 있다.

다만 통계적으로 유의미한 차이가 나타나지 않았으나 Social Proof

기반 설명 방식 여부에 따른 태스크 양별 수행 시간을 비교해보면 Social Proof 설명 방식을 제공할 때 수행시간이 증가하는 경향을 볼 수 있다[그림 23]. Social Proof 기반 설명 방식을 제공할 때 태스크 양이 '3개'인 경우 2초, '4개'인 경우 4초, '5개'인 경우 15초의 수행 시간이 증가하였다.



[그림 23] Social Proof 기반 설명 제공 여부에 따른 태스크 양별 수행 시간 비교

에러 비율과 수행 시간을 종합하여 Social Proof 기반 설명 방식에 따른 수집 효율성을 살펴보면 에러 비율 측면에서는 Social Proof 기반 설명 방식을 제공할 때 태스크 양에 따른 에러 비율이 증가하였다. 수행 시간 측면에서도 통계적으로 유의미한 차이가 나지는 않았으나 Social Proof 기반 설명 방식을 제공할 때 수행 시간이 소폭 증가하였다. 따라서 Social Proof 기반 설명 방식은 수집 효율성을 감소시키는 영향을 보이는 것으로 파악된다.

소결론

실험 조건	의미적 동등성 (Mean, SD)		다양성 (Mean, SD)		에러비율 (%)		수행시간 (s)	
	No Social	Social	No Social	Social	No Social	Social	No Social	Social
3개	0.96 0.06	0.94 0.08	0.31 0.11	0.38 0.16	7	12	27	41
4개	0.96 0.04	0.95 0.08	0.31 0.11	0.33 0.10	7	16	30	34
5개	0.93 0.08	0.92 0.10	0.32 0.11	0.37 0.18	19	23	34	39

[표 7] Social Proof 기반 설명 방식 여부에 따른 측정 지표 요약

실험 3에서는 Social Proof 기반 설명 방식이 표현 다양성 수집에 미치는 영향을 알아보고자 4가지 지표를 기반으로 분석하였다[표 7]. 그 결과, Social Proof 기반 설명 방식은 본래 의도한 목적인 다양성을 향상시킬 수 있는 방안으로 나타났다. 하지만 다양성이 늘어나는 반면 태스크 양이 늘어날수록 의미적 동등성, 에러 비율. 수행시간 지표가 떨어지는 결과가 나타났다. 이러한 결과는 Social Proof 기반 설명 방식이 다양성과 효율성 사이에 트레이드 오프(Trade-off) 관계임을 나타낸다.

지표의 결과를 보면 다양성은 증가하면서 타 지표는 부정적인 경향을 보인다. 이는 서로 다양성이 높은 결과물을 제시한 설명 방식에 따라 Social Proof 개념이 크라우드 소싱 기반의 학습데이터 생성에도 적용됨을 알 수 있다. 다만 시간이 더 오래 걸리고, 에러 비율이

증가하는 점을 보면 이러한 다양성을 충족하고자 압박을 받는 것으로 보인다. 이러한 모습은 실험 3의 파일럿 결과를 통해서도 유추해볼 수 있다. 파일럿 결과에 따르면 참가자들은 Social Proof 기반 설명 방식으로 제시한 결과물을 태스크 수행 조건으로 참조하고 있다. 반면, 그 과정에서 제시한 결과물로 인해 자신의 아이디어를 뺏기는 경향도 있었다. 이러한 점들을 미루어볼 때, Social Proof 기반 설명 방식은 태스크의 수행 충족 조건에 대한 가이드를 제시할 수 있지만 그에 따른 참가자의 수행 부담을 고려해야 한다.

제 6 장 디자인 제언

본 연구는 크라우드 소싱 기반의 대화형 에이전트 개발을 위한 학습데이터 수집에 있어 태스크 디자인 방식이 수집 결과의 다양성에 미치는 영향에 대해 알아보았다. 이를 위해 일련의 3가지 실험을 진행하였다. 본 장에서는 각 실험에서 논의된 점과 결과를 종합하여 크라우드 소싱 기반의 대화형 에이전트 개발을 위한 학습데이터 수집에 있어 적용할 수 있는 디자인 제언을 하고자 한다.

개인차를 고려한 자율적 입력 방식

본 연구는 실험 1을 통해 태스크 양이 표현의 다양성 수집에 미치는 영향을 알아보았다. 그 결과 태스크 양이 증가할수록 수집 결과에 부정적인 영향을 주었지만 다양성에는 차이가 없는 것으로 나타났다. 이러한 결과는 참가자마다 달리 나타나 패러프레이징을 통해 학습데이터를 수집할 수 있는 양이 참가자마다 다를 수 있음을 시사했다. 뿐만 아니라 실험 2를 통해 참가자가 자율적으로 태스크 양을 추가하고 그에 따른 보상을 지급하는 방식에서 참가자의 개인차는 더 뚜렷이 나타났다. 실험 결과 면에서도 참가자가 자율적으로 태스크 양을 추가하고 그에 따른 보상을 받는 방식에서 학습데이터 수집의 효율성이 높아지는 것으로 나타났다.

따라서 크라우드 소싱을 이용하여 대화형 에이전트 개발을 위한 학습데이터를 수집하고자 하는 경우에는 1) 참가자의 태스크 수행에

대한 개인차를 고려하고 2) 이를 반영하여 참가자가 자율적으로 선택할 수 있는 입력 방식을 제공해야 한다. 이러한 방식은 간단히 추가 버튼을 만들고 수행한 태스크 양에 맞추어 보상을 주는 것으로도 효과를 얻을 수 있다.

반응형 결과물 제시 방식

본 연구의 실험 3을 통해 Social Proof 개념을 적용하여 얻고자 하는 결과물 예시를 제공함으로써 수집 결과의 다양성을 향상시킬 수 있음을 알 수 있었다. 하지만 이러한 다양성 향상과 더불어 수집의 효율성이 저하되는 트레이드-오프 관계가 나타나 해당 방식을 그대로 태스크 디자인에 적용하는 데에 한계가 있다.

그러나 일련의 실험을 통한 지표 분석과 파일럿 과정에서 얻은 참가자의 구조화 인터뷰 정보를 통해 이에 대한 해법의 실마리를 찾을 수 있다. 먼저 지표 분석을 통해 앞서 언급했듯 참가자마다 태스크 수행(여기서는 패러프레이징 수행)에 차이를 갖고 있음을 논의했다. 이러한 차이는 인터뷰 내에서도 나타나 결과물 제시가 영감이나 참조 역할을 하는 긍정적인 영향을 주기도 하면서 자신의 아이디어를 가져가는 부정적인 영향을 보이기도 했다. 이와 같은 내용들을 바탕으로 볼 때, 모두에게 처음부터 Social Proof 기반의 결과물 예시를 제공하는 것은 참가자의 개인차를 고려하지 못하게 됨을 뜻한다. 따라서 참가자의 개인차를 고려하여 결과물 예시 제공이 요구되는데, 참가자가 자율적으로 추가 태스크를 수행한다면 이는 태스크 수행을 함에 있어 더

나은 능력을 가진다고 볼 수 있다. 이로부터 모든 참가자에게 처음부터 결과물 예시를 제공하는 것이 아니라 기본 태스크 이외에 추가적인 상호작용을 할 때 그에 반응하여 결과물 예시를 제공하는 방식을 제언한다.

제 7 장 결론

제 1 절 연구 결과의 요약

본 연구는 크라우드 소싱 기반의 대화형 에이전트 개발을 위한 학습데이터 수집에 있어 표현의 다양성을 향상시킬 수 있는 효과적인 태스크 디자인 방식을 찾고자 진행되었다. 이를 위하여 수집 결과에 영향을 미치는 태스크 디자인 요소들을 살펴보고 본 연구에 활용할 수 있는 3가지 요소(태스크 양, 보너스 보상 방식, Social Proof 기반 설명 방식)을 선정하여 일련의 실험을 진행했다. 그 결과 MTurk을 통해 총 480명을 모집하였으며, 73.65 달러를 사용하여 1473개의 패러프레이징 문장을 수집할 수 있었다. 4가지 지표를 중심으로 각 실험을 분석한 연구 결과는 다음과 같다.

먼저 태스크 양에 따른 표현 다양성 수집에 미치는 영향을 보고자 하였다. 태스크 양에 따라 수집되는 데이터의 다양성 차이는 없었지만 같은 의미를 갖는 정도가 감소하거나 에러 비율이 높아지는 결과가 나타났다. 이는 태스크 양이 늘어날수록 한 사람에게 수집할 수 있는 데이터 양은 많아지지만 그에 따라 에러 비율도 높아져 효율적이지 못한 방식으로 밝혀졌다. 다음으로 보너스 보상 방식이 표현 다양성 수집에 미치는 영향을 본 결과, 다양성 면에서는 차이가 나타나지 않았지만 태스크 양이 늘어나면서 나타나는 의미적 동등성 감소를 줄이는 것으로 나타났다. 또한 태스크 양 증가에 따라 에러 비율의 증가 폭도

감소시켰으며, 문장 당 수집 시간이 줄어드는 경향을 보였다. 이는 고정적인 태스크 양에 맞춰 입력하는 방식보다 참가자가 자율적으로 태스크 양을 선택하여 입력하고 그에 따른 보상을 받는 방식이 학습데이터 수집에 효과적임을 보여준다. 마지막으로 Social Proof 기반 설명 방식이 표현 다양성 수집에 미치는 영향을 보고자 하였다. 수집된 데이터의 다양성이 증가하여 Social Proof 효과가 적용되는 것으로 보이나, 태스크 양이 늘어남에 따라 에러 비율 및 수행 시간도 증가하여 다양성과 효율성 간의 트레이드-오프 관계가 나타났다.

최종적으로 실험 과정에서 논의된 참가자의 개인차, 결과물 제시에 따른 압박과 더불어 실험 결과를 종합하여 개인차를 고려한 자율적인 입력 방식, 반응형 결과물 제시 방식에 대한 태스크 디자인 제언을 하였다.

제 2 절 연구의 한계

본 연구는 해석 가능한 분석 지표들로 크라우드소싱 기반의 대화형 에이전트 개발을 위한 학습데이터 수집에 있어 태스크 디자인 요소들이 미치는 영향을 분석하였지만, 태스크 참가자 측면에 대한 이해가 부족하다는 한계가 있다. 일부 참가자의 수행 결과에 대한 인터뷰 정보를 얻긴 하였으나 파일럿 조사에 기반하여 결과를 해석하는데 한계가 있다. 또한 실제 대화형 에이전트 사용자의 발화와 수집 데이터를 비교하여 분석의 객관성을 확보하고자 하였으나 이에 따라 여러 분야의 인텐트를 보지 못한 한계가 있다. 따라서 태스크 참가자

혹은 실험 참가자 측면의 정보를 수집할 수 있는 방식을 마련하고, 여러 분야의 인텐트로 확장하여 후속 연구가 추가적으로 진행될 수 있을 것이라 생각한다.

제 3 절 연구의 의의

본 연구는 크게 학술적 의의, 시의성과 유용성, 융합적 의의를 갖는다고 볼 수 있다. 먼저 학술적 의의 측면에서는 크라우드소싱 기반의 대화형 에이전트 개발을 위한 학습데이터 수집 분야에 있어, 기존 연구들이 학습데이터의 수집 가능성을 밝히는 데 초점을 맞춘 반면, 수집 결과를 향상시킬 수 있는 방안으로 연구 범위를 확장한다는 점에서 학술적 의의를 둘 수 있다.

또한 시의성과 유용성 측면에서는 대화형 에이전트의 개발이 보편화되고 있는 시점에 개발 과정에서 겪고 있는 학습데이터 수집 문제를 해결하고자 한다는 점에서 의의를 갖는다고 볼 수 있다. 본 연구는 크라우드소싱 기반의 대화형 에이전트 개발을 위한 학습데이터 수집에 있어 수집 결과의 표현 다양성을 향상시킬 수 있는 태스크 디자인 방식을 연구한다. 이는 실질적으로 대화형 에이전트 개발 단계에서 겪는 학습데이터 수집 문제를 도울 수 있을 것으로 기대한다.

마지막으로 본 연구는 Social Proof라는 사회심리학 이론을 적용하고자 했다는 점, 태스크 디자인이라는 HCI적 접근과 대화형 에이전트 개발이라는 공학 분야를 접목한다는 점에서 융합적 의의를 갖는다고 볼 수 있다.

참고 문헌

Achananuparp, P., Hu, X., & Shen, X. (2008, September). The evaluation of sentence similarity measures. In *International Conference on data warehousing and knowledge discovery* (pp. 305-316). Springer, Berlin, Heidelberg.

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., & Dustdar, S. (2013). Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, *17*(2), 76-81.

Banko, M., & Brill, E. (2001, July). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting on association for computational linguistics*(pp. 26-33). Association for Computational Linguistics.

Bapat, R., Kucherbaev, P., & Bozzon, A. (2018, June). Effective Crowdsourced Generation of Training Data for Chatbots Natural Language Understanding. In *International Conference on Web Engineering* (pp. 114-128). Springer, Cham.

Bloodgood, M., & Callison-Burch, C. (2010, June). Using Mechanical Turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk* (pp. 208-211). Association for Computational Linguistics.

Callison-Burch, C., & Dredze, M. (2010, June). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 1-12). Association for Computational Linguistics.

Chen, J. J., Menezes, N. J., Bradley, A. D., & North, T. (2011). Opportunities for crowdsourcing research on amazon mechanical turk. *Interfaces*, *5*(3).

Choi, J. D., Tetreault, J., & Stent, A. (2015). It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Vol. 1, pp. 387-396).

Cialdini, R. B. (2009). *Influence: Science and practice*(Vol. 4). Boston, MA: Pearson education.

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, *51*(1), 7.

Das, S., Kramer, A. D., Dabbish, L. A., & Hong, J. I. (2014, November). Increasing security sensitivity with social proof: A large-scale experimental confirmation. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security* (pp. 739-749). ACM.

Jiang, Y., Kummerfeld, J. K., & Lasecki, W. S. (2017). Understanding task design trade-offs in crowdsourced paraphrase collection. *arXiv preprint arXiv:1704.05753*.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). London: Pearson

Kim, J. K., Tur, G., Celikyilmaz, A., Cao, B., & Wang, Y. Y. (2016, December). Intent detection using semantically enriched word

embeddings. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*(pp. 414-419). IEEE.

Labutov, I., & Lipson, H. (2013). Re-embedding words. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 489-493).

Lane, I., Waibel, A., Eck, M., & Rottmann, K. (2010, June). Tools for collecting speech corpora via Mechanical-Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 184-187). Association for Computational Linguistics.

Lathrop, B., Cheng, H., Weng, F., Mishra, R., Chen, J., Bratt, H., ... & Bei, B. (2004). A Wizard of Oz framework for collecting spoken human-computer dialogs: An experiment procedure for the design and testing of natural language in-vehicle technology systems. In *Proc. ITS*.

Lee, C. Y., & Glass, J. (2011). A transcription task for crowdsourcing with automatic quality control. In *Twelfth Annual Conference of the International Speech Communication*

Association.

Lowe, R. T., Pow, N., Serban, I. V., Charlin, L., Liu, C. W., & Pineau, J. (2017). Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, *8*(1), 31-65.

Luger, E., & Sellen, A. (2016, May). Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5286-5297). ACM.

Malu, M., Jethi, N., & Cosley, D. (2012, February). Encouraging personal storytelling by example. In *Proceedings of the 2012 iConference* (pp. 611-612). ACM.

Mao, K., Capra, L., Harman, M., & Jia, Y. (2017). A survey of the use of crowdsourcing in software engineering. *Journal of Systems and Software*, *126*, 57-84.

Mason, W., & Watts, D. J. (2009, June). Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 77-85). ACM.

Myers, C., Furqan, A., Nebolsky, J., Caro, K., & Zhu, J. (2018, April). Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 6). ACM.

Pieraccini, R., & Huerta, J. (2005). Where do we go from here? Research and commercial spoken dialog systems. In *6th SIGdial Workshop on Discourse and Dialogue*.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008, October). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*(pp. 254-263). Association for Computational Linguistics.

Tur, G., Hakkani-Tür, D., & Heck, L. (2010, December). What is left to be understood in ATIS?. In *Spoken Language Technology Workshop (SLT), 2010 IEEE* (pp. 19-24). IEEE.

Vashistha, A., Okeke, F., Anderson, R., & Dell, N. (2018, April). 'You Can Always Do Better!: The Impact of Social Proof on

Participant Response Bias. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 552). ACM.

Vashistha, A., Sethi, P., & Anderson, R. (2017, May). Respeak: A Voice-based, Crowd-powered Speech Transcription System. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1855-1866). ACM.

Wang, W. Y., Bohus, D., Kamar, E., & Horvitz, E. (2012, December). Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *Spoken Language Technology Workshop (SLT), 2012 IEEE* (pp. 73-78). IEEE.

Wooten, D. B., & Reed, A. (1998). Informational influence and the ambiguity of product experience: Order effects on the weighting of evidence. *Journal of consumer psychology*, *7*(1), 79-99.

Zaidan, O. F., & Callison-Burch, C. (2011, June). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 1220-1229). Association for Computational

Linguistics.

<기사 및 보고서>

한국정보화진흥원. 2018. 인공지능 기반 챗봇 서비스의 국내외 동향분석 및 발전 전망.

BBC R&D. 2017. TALIKING WITH MACHINES : PROTOTYPING VOICE INTERFACES FOR MEDIA

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, *14*(6), 1-4.

Abstract

Study on collecting training data for development of conversational agent based on crowdsourcing - Focusing on the improvement of diversity of expression -

Kim Byungjoon

Department of Digital Contents Convergence

The Graduate School

Seoul National University

The conversational agent is a system that receives the natural language from the user and understand the intent for performing the function. With the advancement of speech recognition technology and the development platform of IT companies, service development using a conversational agent is

becoming popular.

To develop such a conversational agent, a large amount of training data is required. Currently, conversational agents provide a way for users to interact as if they were human beings. Accordingly, the conversational agent needs to understand the user's intent and Understanding intent is learned through various and large amount of training data.

However, collecting training data for the development of conversational agents is a very difficult task because of the diversity of expressions and the limitations of collections methods in natural language. Diversity of expressions means having different structures with the same meaning, collecting training data should take characteristic into consideration. Although some methods of collecting are proposed, problems such as time, cost, and accessibility are raised.

With the recent development of artificial intelligence, crowdsourcing has developed and the possibility of solving these problems can be seen. Crowdsourcing has the advantage of solving problems that are difficult for a computer to solve from people and collecting data to a large number of people at low cost. In practice,

the possibility of using crowdsourcing in relation to the training data acquisition is raised.

However, although quality of crowdsourcing is influenced greatly by the task design method and diversity of training data is important, understanding of task design method is insufficient. Therefore, this paper focuses on improving the diversity of expression, examines the effect of task design elements on training data collection, and then suggests a design method that can collect training data effectively.

For this purpose, this paper selects three design elements(task amount, bonus compensation method, social proof based explanation method) to explore the effect of task design elements and conducts 3 experiments of three design elements. The paraphrasing task that possibility of training data acquisition is proven was used, 1473 data were collected from MTurk using $73.65. The collected data were analyzed with four indicators(semantic equivalence, diversity, error rate, and execution time).

As a result of analysis, it was difficult to get data with the same meaning as the amount of task increased. In terms of bonus compensation method, the efficiency of collection increased when

offering bonus compensation. Finally, in terms of the social proof-based explanations, there is a trade-off relationship between diversity and efficiency. Individual differences in collecting among participants and pressure on collecting results were discussed, and an integrated task design method was suggested.

This paper has academic significance in that it studies the possibility of improving the quality of collecting, mainly focusing on the study of the possibility of collecting training data. In addition, it has significance in terms of timeliness and usefulness in trying to solve the problem that is actually experienced in the industrial field. Finally, there is significance in terms of convergence in that it combines social psychology theory, HCI and engineering.

Keywords : Crowdsourcing, Training data Collection, Conversational Agent, Task Design, Diversity of Expression

Student Number : 2017-24292