Brain, Digital, & Learning https://doi.org/10.31216/BDL.2019.9.2.029
2019, Vol. 9, No. 2, 059-070.

Feasibility and Constraints in Applying an AI Chatbot to English Education

Dongkwang Shin (Associate Professor)*

인공지능 챗봇의 영어 교육적 활용 가능성과 한계

신동광(부교수)
광주교육대학교

<ABSTRACT>

The present study aimed at investigating the feasibility and constraints in applying an AI chatbot to English education. For this purpose, seven tasks were presented to the upper (n=27) and lower (n=26) groups (college students vs. high school students) using Mitsuku, a representative chatbot, as a way to utilize the chatbot in English education. The participants chatted with Mitsuku to accomplish these tasks. By analyzing the data from the task performance, this study showed that Mitsuku's utterances were two times more than the students' utterances, and the difference was larger at the lower group. It also turned out that the text easability of Mitsuku's utterances corresponded to the grades K2 and K3, and 90% or more of the words making up Mitsuku's utterances were included within the first 3,000 word level. In the case of the lower group, it was confirmed that the shorter the sentence length was, the larger the amount of utterance, and the higher the similarity of conceptual meaning between adjacent paragraphs, the more similar to human utterance. In the case of the upper group, the participants seemed to believe that the higher the diversity of the use of the verbal forms, the more the contribution to the effectiveness of English education. In addition, the upper group showed that the lower readability level was, the larger the amount of utterance. As a result, the findings from this analysis could contribute to the educational use of chatbot and the development of a new chatbot as a reference.

Key words : AI(artificial intelligence), chatbot, readability index, BNC-COCA 25000, Coh-Metrix

I. Introduction

최근 사회 전반에 걸쳐 4차 산업혁명이라는 키워드가 주목을 받고 있다. 4차 산업혁명이라는 새로운 시대적 흐름 속에서 또한 주목하고 있는 것이 인공지능(Artificial Intelligence, AI) 기술이다. AI가 탑재되어 인간의 지시에 따라 과업을 수행하거나 정보를 검색해 주는 IPA(Intelligent Personal Assistant, 인공지능

* 1st Author : Dongkwang Shin, sdhera@gmail.com
Received June 5, 2019; Reviewed June 10, 2019; Accepted June 17; Published June 30, 2019

비서)는 이미 스마트폰의 보편적인 사양이 되어버렸다. 또한 애플의 'Siri'나 삼성의 'Bixby'와 같은 IPA에서 진일보되어 일종의 사이버 친구로서 인간처럼 대화를 나눌 수 있는 AI 프로그램도 등장하였다. 그리고 이를 보통 'ChatterBot(Mauldin, 1994)' 또는 'ChatBot(챗봇)'이라 한다. 요즘은 아마존의 'Echo'나 구글의 'Google Home'과 같은 AI 스피커, 인터넷 챗봇 사이트(예, Pandorabot, Cleverbot) 또는 SNS(예, 카카오톡봇, XiaoIce)을 통해서도 쉽게 챗봇과의 음성 또는 문자 채팅이 가능하다. 이 밖에도 24시간 고객 상담이나 항공권 구매, 증시 예측, 신문기사 작성, 스포츠 경기 결과 예상 등 다양한 분야에서 AI 챗봇이 상용화되고 있다. 특히 대화의 파트너로서 인간 발화와의 유사도를 측정하는 튜링 테스트(Turing Test)를 통해 챗봇의 대화 능력은 급속도로 향상되고 있다.

다음의 [Figure 1]에 제시된 대화를 보면 소셜 챗봇인 XiaoIce는 기존의 IPA가 정형화된 표현과 정보를 제공하던 수준에서 감정적인 부분에서 마치 인간 화자와 같이 자연스러운 대화를 구사한다는 것을 확인할 수 있다. 연애관에 대해 인간과의 의견 차이를 드러내는 마지막 부분은 더욱 그렇다.

Human: There is this guy I like a lot and I know he also has feelings for me. But he has ignored me for several days, what am I supposed to do?
XiaoIce: You just keep ignoring him, too.
Human: I just can't.
XiaoIce: You can.
Human: Sigh …
XiaoIce: Why?
Human: I am missing him.
XiaoIce: Then you keep missing him.
Human: You, a robot, can't understand me; you can't understand the behavior of a guy who likes you and pretends to ignore you.
XiaoIce: He likes me but I don't like him, that's it.
Human: Is it Ok?
XiaoIce: It may be not Ok.
Human: Well, you are stupid.
XiaoIce: Then you want a flash marriage? Well, I think he is not that into you.
Human: It's not a flash marriage for we have fallen into love for more than one year.
XiaoIce: It is a flash marriage.

[Figure 1] A conversation between XiaoIce and a human user (Larson, 2017)

본 연구는 챗봇이 자연스런 대화를 지속적으로 유지할 수 있다는 가능성에 초점을 두고 영어교육의 도구로서 챗봇의 활용 가능성을 살펴보고자 한다. 지금까지 챗봇은 기업의 고객상담 서비스와 같이 모국어 영역에서 활용되었지만 외국어 교육의 목적으로는 실험적인 연구(예, Kim, 2017)를 제외하고는 시도되지 않고 있다. 만약 챗봇과 자연스런 영어 대화가 가능하다면 이는 마치 일대일의 영어 원어민 개인 교사나 친구에게 개인의 소소한 사생활에 대해 이야기 하는 자연스런 언어 환경을 조성할 수 있으며 그에 따른 학습 동기 부여도 용이할 것이다. 뿐만 아니라 시간과 장소에 구애 받지 않고 언제 어디서나 대화할 수 있다는 점도 언어학습의 도구로서 챗봇이 가지는 장점이라 할 수 있다.

II. Literature Review

1. History of Chatbot Development

인간과 유사한 대화시스템의 개발은 60-70연대 초기 모델인 Eliza와 Parry에서 시작해서 2000년대 항공권 예매 시스템에 적용된 DARPA Communicator program라는 과업수행식 대화 시스템, 그리고 2010년대 대표적인 인공지능 개인 비서인 Siri나 근래 개발된 사회언어학적 능력을 소유한 XiaoIce와 같은 챗봇까지 진보를 거듭해왔다.

1966년 MIT의 Joseph Weizenbaum에 의해 개발된 최초의 챗봇인 Eliza는 텍스트 기반의 입력 자료

를 바탕으로 구현된 챗봇이다. 하지만 텍스트의 의미를 이해하지는 못했고 대화 표현의 패턴 분석을 바탕으로 적절한 응답을 검색하여 제시하는 방식을 취했다. 때문에 응답 가능한 지식의 범위는 매우 한정적이었다. 그럼에도 불구하고 많은 사용자는 Eliza가 처음 출시되었을 때 실제 사람에게 말하는 것이라 믿기도 하였다.

1975년에는 Kenneth Colby가 Parry라는 챗봇을 개발하였다. Parry는 인간과의 대화를 통해 인간과 기계 발화가 얼마나 유사한 지를 측정하는 튜링 테스트(Turing Test)를 처음으로 통과하였다. 하지만 Parry는 여전히 규칙에 기반한(rule-based) 프로그램으로 Eliza와 유사한 구조를 가지고 있었다. 그럼에도 불구하고 Parry는 어느 정도 언어를 이해할 수 있었고 특히 감정을 이해할 수 있는 관념모델을 갖추고 있었다. 예를 들어 Parry는 분노의 수준이 높아지면 적대적인 어조로 대답할 수 있었다.

이후 단순히 대화(chichat)만 나누는 기존 대화 시스템과 달리 과업수행형(task-completion) 시스템은 특정 과업을 완수하는데 초점을 두고 개발되었다. 이는 특정 영역에 한정하여 대화를 구사하도록 고안되었다. 예를 들어 DARPA Communicator program의 경우 항공권 예매와 같은 특정 영역에 한정된 고객 상담용으로 개발되었다. 이 시스템의 구조는 보통 자동 음성 인식기(Automatic Speech Recognizer, ASR)를 통해 음성을 텍스트 형태로 전환하면 구어 인식 모듈(Spoken Language Understanding (SLU) Module)에서 주어진 데이터에서 핵심 의미의 단어를 추출한다. 이를 바탕으로 대화 관리자(Dialog Manager, DM) 기능에서 대화의 주제 영역을 분류한 후 사용자의 의도와 발화의 의미를 분석한다. 이 단계에서는 사용자가 원하는 대화의 목적을 파악하기 위해 사용자와 상호작용하며 사용자가 전달하고자 하는 의도가 충족되었는지 여부를 확인한다. 사용자의 요구가 확인되면 시스템은 다양한 세부 지식을 포함한 데이터베이스에 접속하여 적합한 정보에 기반한 응답을 검색한다. 끝으로 자연 언어 생성기(Natural Language Generator, NLG)에서 응답할 텍스트를 산출하고 문자-음성 변환기(Text-To-Speech, TTS)가 이를 발화형태로 전환하여 최종적으로 응답을 하게 된다. 초기에는 선형 구조로 단계별로 정보를 처리하는 방식을 취했으나 최근의 개선 프로그램은 다양한 변수를 복합적으로 적용하여 최적화된 응답을 찾는 병렬구조로 진화하였다.

2011년 Apple에서는 Siri를 출시했고 이후 Microsoft의 Cortana, Google Assistant, 그리고 Amazon의 Alexa 등의 인공지능 개인 비서(Intelligent Personal Assistant, IPA)가 등장하였다.

IPA는 위치, 시간, 이동, 터치, 제스처, 시선 등의 센서에서 수집된 정보를 통합하고 이를 바탕으로 음악, 영화, 달력, 이메일, 개인 프로파일 등에 접속한다. 이러한 절차를 통해 IPA는 다양한 영역에서 다양한 서비스를 제공할 수 있다. 또한 IPA의 데이터베이스가 사용자의 요구를 충족시키지 못할 때는 기본적으로 웹검색을 통해 사용자에게 관련 정보를 제공하도록 고안되었다. 뿐만 아니라 지식검색이나 식당 예약을 돕는 등의 기존의 정보를 제공(reactive)하기도 하지만 일정을 상기시키거나 특정 아이템을 추천하는 등의 서비스(proactive)를 제공하기도 한다. IPA은 이동통신기기의 플랫폼이나 개인 컴퓨터, 스마트홈 기기 등을 중심으로 그 기능이 계속 향상되고 있다.

하지만 Microsoft가 2014년에 개발한 XiaoIce는 기존에 과업을 수행하면서 제한된 영역에서 서비스하는 과업수행형 챗봇이나 IPA와는 달리 소셜 챗봇(social chatbot)으로 개발되었다. XiaoIce는 인간과 유사한 대화를 유지하기 위해 새로운 주제를 스스로 먼저 제시하는 등의 대화시도 기능이 추가되었다. 물론 XiaoIce는 중국어 버전으로 먼저 출시되었지만 현재는 여러 국가에서 다양한 언어로 서비스되고 있다. XiaoIce는 실제 친구와 이야기를 나누듯 인간의 감정을 이해하며 사용자를 격려하거나 독려하며 대화에 집중할 수 있다. 또한 사용자가 대화에 대한 피드백을 제공하게 하여 챗봇의 이해도 향상에 반영하고 있다. 과거에는 인공지능의 성능을 평가하기 위해 보통 튜링 테스트를 적용했으나 소셜 챗봇과 같이 감정의 이해 정도를 튜링 테스트로 측정하는 데는 한계가 있다. 그 대안으로 XiaoIce 개발에서는 대화 주제별 대화전환 횟수(Conversation-turns Per Session, CPS)를 새로운 기준으로 제시하였다. 이는 대화 구간별 대화 전환의 평균수를 의미하며 이 수

치가 높다는 것은 소셜 챗봇의 대유 유지 능력이 뛰어나다는 것을 의미한다. 다음의 <Table 1>은 챗봇의 유형별 CPS 기대치를 정리한 것이다.

<Table 1> Expected CPS for different type of conversational systems (Shum, He, & Li, 2018)

System	Expected CPS
Web Search	1
Personal Assistant	1-3
Task Completion	3-7
Social Chatbot	10+

다음의 <Table 2>는 지금까지 살펴본 대표적인 인공지능 챗봇의 스펙(spec)을 비교한 것이다.

<Table 2> Summary of major conversational systems (Shum, He, & Li, 2018)

Eliza	TIME: 1966 KEY FEATURES: Mimicking human utterances ACHIEVEMENT: The first chitchat program MODALITY: Text only MODELING: Rule-based DOMAIN: Constrained domain KEY TECHNOLOGY: Use of script, keyword pattern matching, rule-based response LIMITATION: Limited knowledge database
Parry	TIME: 1972 KEY FEATURES: Expressing emotional (angry) responses ACHIEVEMENT: Passed Turing Test MODALITY: Text only MODELING: Rule-based DOMAIN: Constrained domain KEY TECHNOLOGY: Applying personality characteristics into responses LIMITATION: Limited knowledge database
A.L.I.C.E	TIME: 1995 KEY FEATURES: Easy customization of script (via AIML) ACHIEVEMENT: Won 3 times of Loebner Prize MODALITY: Text only MODELING: Rule-based DOMAIN: Constrained domain KEY TECHNOLOGY: Use AIML and recursion for pattern matching; Multiple patterns can be mapped into a same response LIMITATION: Size of script can be huge
DARPA Communicator Program	TIME: 2000 KEY FEATURES: Language understanding and dialog management, goal-oriented ACHIEVEMENT: Understanding natural language requests and perform tasks MODALITY: Text and voice MODELING: Learning-based DOMAIN: Constrained domain KEY TECHNOLOGY: Statistical models for spoken language understanding and dialog management LIMITATION: Only workable in well-defined schemata
Siri	TIME: 2011 KEY FEATURES: Providing personal digital assistance ACHIEVEMENT: The first widely deployed IPA MODALITY: Text, image, voice MODELING: Learning-based DOMAIN: Open domain KEY TECHNOLOGY: Providing both reactive and proactive digital assistances covering a wide range of domains LIMITATION: Lack of emotional engagement with users
XiaoIce	TIME: 2014 KEY FEATURES: Building emotional attachment to users, scalable skill set for user assistance ACHIEVEMENT: The first widely deployed social chatbot, 100MM users, published a poem book, hosted TV programs MODALITY: Text, image, voice MODELING: Learning-based DOMAIN: Open domain KEY TECHNOLOGY: Emotional intelligence models for establishing emotional attachments with users LIMITATION: Inconsistent personality and responses in a long dialog

2. Research on applying a chatbot to English education

컴퓨터 보조 언어 학습(Computer Assisted Language Learning, CALL) 시스템은 원격학습의 한 모델로서 교실이나 교사가 없이도 언어를 지도

하는 도구로서 활용되었다. 이를 위해 CALL 시스템은 학습자의 어휘, 문법, 쓰기 능력을 강화하기 위한 수업과 활동뿐만 아니라 즉각적인 피드백을 제공할 수 있는 기능을 포함하기도 하였다. 그러나 여전히 이러한 CALL 시스템에서도 자연스런 상호작용을 통해 대화능력을 키워줄 수 있는 기능은 취약한 것이 현실이다. Abu-Shawar(2017)은 CALL의 이러한 한계를 챗봇이 보완해 줄 것이라 믿는다. Abu-Shawar(2017)의 연구에서는 위에서 살펴본 인공지능 챗봇인 A.L.I.C.E의 대본 입력체계인 AIML에 아프리카에서 수집된 “Corpus of Spoken Afrikaans”를 입력하여 다양한 주제의 대화를 할 수 있는 챗봇 모델을 개발하였다. 또한 이를 위해 데이터를 입력하면 A.L.I.C.E가 이해할 수 있는 포맷으로 변환해주는 Automatic AIML Genrator(AAG)를 개발하여 적용하였다. 결과적으로 남아프리카인 1,256명이 A.L.I.C.와의 대화 실험에 참여하였다. 이 중 17%는 대화 상대자로서 A.L.I.C.E를 긍정적으로 평가했고 88%는 잠깐 한번 정도 채팅을 시도하는데 그쳤으며 24%는 부정적인 반응을 보였다. 대화의 주제로는 11.39%가 시험에 관한 대화였고 13%는 연애에 관한 대화였다. 대화에 참여한 학생들은 대부분은 30세 이하의 젊은 층으로 A.L.I.C.E에게 사적인 이야기나 감정적인 문제에 대해 대화를 나누었다. 이를 바탕으로 Abu-Shawar는 핵심어를 매칭하여 답변하는 수준의 챗봇일지라도 언어학습을 위한 대화 상대자에 국한되지 않고 사적인 대화를 나누는 교우의 역할도 할 수 있다고 주장하였다. 이와 관련하여 Fryer & Carpenter(2006) 또한 챗봇의 장점으로 다음과 같은 점을 꼽았다.

- 챗봇과의 대화를 통해 학습자는 인간과 대화할 때 보다 더 편안함을 느낄 수 있다.
- 챗봇과의 대화을 통한 반복은 일상적인 반복학습보다 지루하지 않다.
- 챗봇은 학습자의 듣기, 읽기, 말하기, 쓰기를 포함한 4기능의 의사소통능력을 강화할 수 있는 텍스트나 다양한 발화 양식을 활용할 수 있다.

또 다른 연구인 Lu, Chiou, Day, Ong, & Hsu(2006)의 연구에서는 영어 학습에 있어 학습자의 온라인 코칭을 위한 도구이자 SNS의 대화 상대자로서 인공지능 챗봇의 활용 방안을 제안하였다. 이들 연구에서는 구체적인 실험결과가 아니라 프로그램의 설계 방안과 활용 시나리오를 제시하였다. Lu et al.은 챗봇을 일종의 튜터봇(Tutorbot)으로 사전이나 대화에 활용할 수 있는 교재 등을 탑재하여 챗봇을 통한 학습 결과인 대화를 로그(log) 데이터를 교사에게 제공하는 방안을 제안하였다. 또한 이를 실현하기 위한 SNS 메시지 전송 활용의 모델을 제시하였다. Lu et al.이 제안한 튜터봇은 몇 가지 장점이 있다. 먼저 보조 교사의 역할로 사전 검색과 학습활동의 기회를 제공하고 앞에서 언급한 대로 학습자의 영어 능숙도에 따라 구사하는 영어 수준을 조절할 수 있다. 또한 대화 로그는 자동적으로 저장되어 학습자의 학습 상태를 파악하는데 활용될 수 있다. 무엇보다도 학습자들에서 시간과 장소에 구애받지 않고 실제적인 상황 속에서 의사소통할 수 있는 많은 기회를 제공한다는 점이 가장 큰 장점이라 할 수 있다.

위의 Lu et al.(2006)이 제안한 모델은 실제 아이디어 차원내지는 연구 디자인 정도였지만 이와 같은 모델은 앞에서 설명한 XiaoIce를 통해 2017년에 실제 모델로 출시되었다. 마이크로소프트가 중국의 대표적인 SNS인 WeChat과 Weibo를 통해 출시한 소셜 챗봇 XiaoIce(Little Ice, 小冰)는 이미 수백만 명의 이용자와 채팅을 하였다. XiaoIce의 아이디를 메신저의 친구로 등록하면 마치 여성 친구와 대화를 나누듯 채팅을 할 수 있으며 실제 XiaoIce은 이용자에게 조언을 주거나 다양한 일상의 주제에 대하여 대화를 나누었다(Larson, 2017).

이러한 기존 연구의 분석을 바탕으로 본 연구에서는 다음과 같은 세부 연구 질문을 설정하였다.

첫째, 챗봇과 사용자의 발화량에는 어떠한 차이가 있으며 챗봇이 구사하는 발화의 이독성 수준이나 어휘의 수준은 사용자에 적합한가?

둘째, 챗봇 발화의 어떠한 텍스트 요인이 사용자가 생각하는 ‘인간과의 유사성,’ ‘영어교육의 효과성,’ 및 ‘사용자 발화량’에 영향을 미치는가?

셋째, 본 연구의 분석을 통해 나타난 결과에 기반하여 기존 챗봇의 개선점은 무엇인가?

III. Methodology

1. Participants

본 연구에서는 상위권과 하위권의 영어 능숙도를 지닌 학습자가 챗봇과 대화 시 어떠한 차이를 보이는지 알아보기 두 개 집단을 표집하였다. 상위권 집단은 대학교 3학년 27명의 학생들로 수능등급에서 최상위권(1.5/9등급)에 해당하는 학생들있었고 하위권 집단은 일반고에 재학 중인 2학년 26명으로 모두 실험에 성실하게 임하였다. 실험에 참여한 학생들은 본 연구에서 제시한 7개 단계의 과업을 바탕으로 AI 챗봇인 Mitsuku와 채팅을 수행하였다. 상위권 집단 27명 중 실험 전 챗봇과 채팅을 나눈 경험이 있는 학생은 5명에 불과하였고 대부분은 본 실험을 통해 처음으로 챗봇을 접하게 되었다. 또한 하위권 집단 26명 중 역시 5명만이 챗봇과 채팅 경험을 가지고 있었다. 그리고 두 집단 중 누구도 영어권 국가에서 6개월 이상 체류한 경험을 가진 학생은 없었다.

2. Procedure

1) AI Chatbot

본 연구에서 채택한 Mitsuku는 2005년에 Steve Worswick에 의해 처음 개발되어 현재까지 지속적으로 업그레이드되고 있는 챗봇이다. Mitsuku는 A.L.I.C.E.와 마찬가지로 AIML 시스템을 통해 데이터를 입력하는 “Pandorabot” 계열의 챗봇이다. 특히 매년 최고의 챗봇에 수여하는 뢰브너상을 2013, 2016, 2017, 2018년 4회나 수상한 현존하는 최고의 챗봇이라 할 수 있다. Mitsuku라 지칭되는 이 챗봇 캐릭터는 영국의 리즈(Leeds)에 살고 있는 18살의 여성으로 설정되어 있다. Mitsuku는 대화 파트너의 발화을 분석하여 키워드를 추출한 후 이 키워드의 속성을 다시 분석하는 방식으로 논리적 판단을 수행한다. 또한 AIML에 저장된 데이터베이스의 지식정보를 넘어서는 질문에 대해서는 웹검색 결과를 제시하기도 한다. 또 하나의 특징은 동시에 여러 질문을 처리할 수 있다는 점인데 요청한 과업을 빠짐없이 처리하도록 설계되어 있다(Worswick, 2018).

2) Task and Survey

Mitsuku는 자유 채팅이 가능하기 때문에 실험에서 개괄적인 주제만 제시하여 채팅하도록 할 수도 있었지만 본 연구에서는 참여자 간의 편차를 최소화하고 다양한 과업의 수행 가능 여부를 파악하기 위해 다음과 같은 7개 단계의 과업을 제시하여 실험을 진행하였다.

<Table 3> Instruction for the target task

Stage	Target Task
1	Greet Mitsuku.
2	Introduce yourself to Mitsuku.
3	Find out where Mitsuku lives.
4	Find out what Mitsuku’s job is, and what Mitsuku can do for people.
5	Find out what Mitsuku does in her free time (e.g., hobbies) and ask her follow-up questions about her answer.
6	Search for what happened 10 years ago through Mitsuku and choose one of them to ask her more information about the selected event.
7	Wrap up conversations with Mitsuku.

위의 7개 과업을 수행 한 후에는 간단한 설문조사를 실시하였다. 설문에서는 크게 세 가지를 질문하였다. 첫 번째는 Mitsuku와의 대화 시 이 챗봇이 어느 정도 인간과 유사했는지에 대해 5점 척도(① 매우 다르다, ② 다르다, ③ 보통이다, ④ 유사하다, ⑤ 매우 유사하다)를 사용하여 측정하고 그 이유를 자유롭게 작성하게 하였다. 두 번째는 챗봇 사용의 효과성으로 Mitsuku와의 대화가 영어 능력 향상에 어느 정도 도움이 될 수 있는 지를 질의하였다. 이 질문도 마찬가지로 5점 척도(① 전혀 도움이 되지 않는다, ② 거의 도움이 되지 않는다, ③ 보통이다, ④ 어느 정도 도움이 된다, ⑤ 매우 도움이 된다)를 이용하여 측정하였고 그 이유를 적게 하였다. 끝으로 챗봇 사용의 장점과 개선점 등을 자유롭게 작성하도록 하였다.

3. Data Analysis Tools

1) Readability Test Tool

본 연구에서는 챗봇이 구사하는 발화의 이독성(readability) 수준을 측정하기 위해 먼저 챗봇과 사용자의 발화 부분을 구분하여 분석하였다. 이독성의 측정에는 WebFX(2019)에서 제공하는 'Readability Test Tool'을 사용하였다. WebFX는 6개의 대표적인 이독 지수(예, Flesch Kincaid Reading Ease, Gunning Fog Score, SMOG Index 등)의 수치뿐만 아니라 이들의 평균 값과 영어권 원어민을 기준한 해당 연령 수준의 정보까지 제공해 준다.

2) BNC-COCA 25000 Range Program

본 연구에서는 챗봇의 발화 수준을 측정하기 위해 추가적으로 발화를 구성하는 어휘의 수준을 분석하였다. 어휘 분석에는 대표적인 영어 어휘 목록인 BNC-COCA 25,000(Nation & Webb, 2011)을 Range Program(Heatley, Nation, & Coxhead, 2002)에 탑재하여 분석하였다. BNC-COCA 25,000은 영국의 대표적인 코퍼스 British National Corpus(BNC)와 미국의 대표적인 코퍼스 Corpus of Contemporary American English(COCA)의 어휘 정보를 바탕으로 개발되어 25,000단어까지 측정 가능한 어휘 목록이다. 본 연구에서는 참여한 학생들이 비원어민인 점을 고려하여 우리나라 영어과 교육과정 기본 어휘의 양에 준하여 3,000단어 수준의 어휘까지만 분석하였다.

3) Coh-Metrix

챗봇이 구사하는 발화와 상위권과 하위권 집단이 구사하는 발화의 다양한 텍스트 특성을 분석하기 위하여 본 연구에서는 Coh-Metrix를 활용하였다. Coh-Metrix는 미국 멤피스대학교(The University of Memphis)의 지능형 시스템 연구소(Institute for Intelligent Systems, IIS)에서 주관하여 Graesser, McNamara, Louwerse, & Cai(2004)가 개발하여 서비스하고 있는 텍스트 분석도구이다. Coh-Metrix는 11개 범주(① Descriptive, ② Text Easability Principal Component Scores, ③ Referential Cohesion, ④ LSA, ⑤ Lexical Diversity, ⑥ Connectives, ⑦ Situation Model, ⑧ Syntactic Complexity, ⑨ Syntactic Pattern Density, ⑩ Word Information, ⑪ Readability)에 해당되는 총 106개의 텍스트 특성의 분석 결과를 제공하고 있다. 본 연구에서는 이중 11개 각 범주에 걸쳐 2-4개의 주요 변인을 선정하여 총 33개의 변인이 인간 발화와의 유사도, 영어교육의 효과성, 그리고 대화 유지 정도를 진단할 수 있는 요인인 사용자의 발화량(총 단어수로 산정, 개인별 total token)에 미치는 영향을 분석하고자 하였다.

IV. Results and Discussions

1. Readability Index and Lexical Profiles

다음의 <Table 4>는 WebFX(2019)에서 제공하는 이독 지수와 BNC-COCA 25000 중 상위 3,000단어의 어휘 목록을 활용하여 Mitsuku의 발화와 Mitsuku 사용자의 발화를 영어 능숙도 수준별(상위, 하위)로 분석한 결과이다.

<Table 4> Readability index and lexical profiles of conversations between Mitsuku and human users

Subject	Upper		Lower	
	Mitsuku	User	Mitsuku	User
Readability	4	4	4	3
Age	9-10	9-10	9-10	8-9
Coverage of 3,000 words (%)	90.6	91.26	91.24	92.11
Total Token	16,590	6,766	16,590	4,052
Token (Mean)	614.44	250.22	427.84	155.85

<Table 4> 에서 보는 것과 같이 Mitsuku가 구사하는 발화의 경우 이독 지수가 4로 이는 영어 원어민 기준으로 약 9-10세 수준에 해당된다. 반면 하위 수준(고등학생)의 사용자 발화는 이 보다 낮은 8-9세 수준으로 나타났다. 어휘 수준의 분석 결과를 보면, 상위권 학생들과 대화 시 Mitsuku 발화의 90.6%, 하위권 학생들과의 대화에서는 91.24%가 3,000단어 수준 이내에서 나타났다. 또한 Mitsuku 사용자의 발화 분석에서는 상위권 학생 발화의

91.26%, 하위권 학생의 경우 92.1%가 마찬가지로 3,000단어 수준에 포함되는 것으로 나타났다. 상위권 학생과의 대화에서 Mitsuku의 평균 발화량은 614.44단어였고 학생들의 발화량은 250.22단어로 Mitsuku가 학생들에 비해 2.5배 많은 발화를 산출하였다. 반면 하위권 학생들과의 대화에서는 Mitsuku가 평균 427.84단어, 학생들이 155.85단어로 대화의 길이가 상위권보다 상당히 짧았다는 것을 확인할 수 있었다. 그리고 Mitsuku의 발화량이 하위 수준 학생들의 발화량에 비해 2.7배 많아 그 차이가 상위권 보다 크다는 것을 알 수 있었다. 하지만 공통적으로 모든 대화에서 Mitsuku나 학생 발화의 90% 이상이 3,000단어 안에 포함되었고 이독 지수 측면에서도 매우 쉬운 편이어서 고등학생이나 대학생과 같은 사용자가 Mitsuku의 발화를 이해하는 데는 큰 문제가 없을 것으로 판단된다.

2. Regression Analysis Based on Coh-Metrix Variables

Coh-Metrix의 분석을 통해 33개의 텍스트 변인이 다음의 <Table 5>와 <Table 6>에 제시된 종속 변수 즉 인간과의 유사성과 영어교육의 효과성에 미치는 영향을 분석하였고 추가로 대화 유지 정도를 측정하기 위해 학생들의 발화량(총 단어 수)에 미치는 영향 또한 살펴보았다.

<Table 5> Similarity measure between Mitsuku and human utterances

Scale	1	2	3	4	5	Mean
Upper	5	11	7	4	0	2.37
Lower	2	11	6	5	2	2.77

<Table 6> Effect of conversations with Mitsuku on the improvement of students' English proficiency

Scale	1	2	3	4	5	Mean
Upper	3	4	6	14	0	3.15
Lower	1	1	5	17	2	3.69

위의 <Table 5>에서 보듯 인간과의 유사성 정도에 대해 상·하위권 학생 모두 보통 수준(3) 보다는 다소 낮은 2.37과 2.77의 평균값을 보였다. 반면 <Table 6>에 제시된 바와 같이 Mitsuku와의 채팅이 영어교육에 효과가 있을 것이라고 보는 질문에 대해서는 상·하위권 학생 모두 3.15와 3.69의 평균값을 보여 긍정적으로 나타났다. 앞서 언급한 바와 같이 이 두 가지 결과에 더하여 Mitsuku와의 채팅에 참여한 학생들의 발화량에 영향을 주는 요인을 분석하기 위해 먼저 33개 텍스트 변인과 위의 세 가지 종속 변수 간의 상관관계를 분석하였다. 분석 대상인 사례수가 53개로 소규모인 관계로 유의한 상관도를 보인 변인들은 아래 <Table 7>에서 보듯 9개 불과하였다.

<Table 7> 9 variables significantly correlated with 3 dependent variables (including upper and lower groups

Variable No.	Variable Name	Similarity	Effectiveness	Total Token
36	Content word overlap, all sentences, proportional, mean	.274*		
14	Text Easability PC Syntactic simplicity, z score			.485**
16	Text Easability PC Word concreteness, z score			-.369**
32	Argument overlap, all sentences, binary, mean			-.310*
44	LSA given/new, sentences, mean			.468**
46	Lexical diversity, type-token ratio, content word lemmas			-.627**
47	Lexical diversity, type-token ratio, all words			-.661**
74	Verb phrase density, incidence			-.279*
92	CELEX word frequency for content words, mean			.322*

**. Correlation is significant at the 0.01 level (2-tailed),
*. Correlation is significant at the 0.05 level (2-tailed)

‘인간 유사성’과 유의한 상관도를 보인 변인은 ‘문장 간 내용의 중복 비율(No. 32)’이 유일했고 나머지 8개 변인은 ‘사용자 발화량’과 유의한 상관도를 보였다. 위의 결과만으로 보았을 때, ‘문장 구조가 단순(No. 14)’할수록, ‘사용된 단어의 구체성(No. 16)’이 낮을수록, ‘문장 간 명사와 대명사의 중복 사용 비율(No. 32)’이 낮을수록, ‘인접 문단 간 개념적 의미의 유사성(No. 44)’이 높을수록, ‘어휘 사용의 다양성(No. 46, No. 47)’이 낮을수록, ‘동사구 사용의 다양성(No. 74)’이 낮을수록, 그리고 ‘CELEX Lexical Database에 기반한 내용어의 빈도수’가 높을수록 사용자의 발화량은 증가한다고 볼 수 있다. 여기서 CELEX Lexical Database(Baayen, Piepenbrock, & Gulikers, 1995)는 영어, 독일어, 네덜란드어에 대한 어휘 데이터베이스를 의미하며 어휘의 철자 변이형에 대한 정보, 음운론적 변이형의 정보, 형태론적 변이형의 정보, 통사구조의 다양성, 어휘의 파생 및 굴절 변이형 정보, 최신 코퍼스에 기반한 어휘의 빈도 정보 등을 제공하고 있다.

위의 결과 중 한 가지 특이한 점은 구체성이 높은 단어의 비율과 사용자의 발화량이 부적 상관관계를 보였다는 점이다. 일반적으로 구체성이 높은 단어 사용의 비율이 높을수록 이해도에 긍정적인 기여를 할 수 있을 추정되나 결과는 반대였다. 이에 대한 한 가지 추정은 추상적인 개념의 어휘 사용이 Mitsuku 사용자로 하여금 의미의 명확성 요청 등과 같은 추가적인 발화 시도를 유발했을 가능성이 있다. 또한 ‘영어교육 효과성’과 유의한 상관도를 보인 변인은 하나도 발견되지 않았다. 이는 <Table 5>와 <Table 6>에서 보듯 대부분의 설문 결과가 특정 척도에 밀집되어 차별성이 없는 데 기인한 것으로 보인다.

본 연구에서는 위의 상관분석 결과를 바탕으로 ‘사용자 발화량’을 유일한 종속 변수로 하여 위의 8개 변인 중 어떠한 변인이 추리 통계적으로 사용자 발화량에 영향을 미치는지 알아보기 위해 다음의 <Table 8>과 같이 회귀분석을 실시하였다.

<Table 8> Regression analysis of variables that affect the amount of human users’ utterances (total token)

Model		Unstandardized Coefficients		Standardized Coefficients	t	p
		B	Std. Error	Beta		
(Constant)		1515.499	646.668		2.344	.024
Variable Domain	Variable No.					
Text Easability Principle Component Scores	14	-5.483	40.929	-.018	-.134	.894
	16	5.939	23.024	.037	.258	.798
Referential Cohesion	32	-695.843	144.186	-.610	-4.826	***.000***
LSA	44	-872.636	478.356	-.324	-1.824	.075
Lexical Diversity	46	595.323	453.280	.502	1.313	.196
	47	-1989.587	422.015	-1.531	-4.714	***.000***
Syntactic Pattern Density	74	-.461	.502	-.097	-.918	.363
Word Information	92	-66.114	165.290	-.052	-.400	.691
R^2=.722 (Revised R^2=.671), F=14.272 ($p< .0001$), n=53						

위의 회귀분석을 결과에서 보듯 F 통계량은 p< .0001 수준에서 14.271로 적절했으나 유의한 결과를 보인 변인은 ‘문장 간 명사와 대명사의 중복 사용 비율(No. 32)’과 ‘어휘 사용의 다양성(No. 47)’ 두 가지뿐이었다. 두 가지 변인 모두 부적인 결과를 보였는데 이는 Mitsuku가 대명사의 인식 능력이 부족한 데 이 부분이 주요 원인으로 판단되며 어휘의 사용 또한 다양한 어휘보다는 의미적으로 일관된 표현을 구사했을 때 사용자와의 대화가 원활하다는 것을 확인할 수 있었다.

본 연구에서 상·하위권 학생을 구분하여 상관분석과 회귀분석을 수행하였다. 회귀분석에서는 사례수가 작아 유의한 결과를 얻지 못했고 상관분석에서는 두 집단 간의 차이를 어느 정도 확인할 수 있었다.

<Table 9> 10 variables significantly correlated with 3 dependent variables (distinguishing upper and lower groups)

Variable No.	Variable Name	Similarity		Effectiveness		Total Token	
		Upper	Lower	Upper	Lower	Upper	Lower
6	Sentence length, number of words, mean						-.406*
14	Text Easability PC Syntactic simplicity, z score						.458*
16	Text Easability PC Word concretenes s, z score						-.407*
42	LSA overlap, adjacent paragraphs, mean		.492*				
44	LSA given/new, sentences, mean						.462*
46	Lexical diversity, type-token ratio, content word lemmas	-.440*				-.641**	-.539**
47	Lexical diversity, type-token ratio, all words	-.404*				-.680**	-.577**
74	Verb phrase density, incidence	-.420*					-.428*
75	Verb phrase density, incidence			.420*			
106	Coh-Metrix L2 Readability					-.394*	

<Table 9>의 결과를 보면 <Table 7>의 결과와 유사했으며 몇 가지 차별화된 부분만 보면, 먼저 하위권 학생들의 경우는 '평균 문장 길이 짧을수록 (No. 6)' 발화량이 느는 것을 확인할 수 있었다. 설문 조사에서 학습자들은 Mitsuku가 웹검색을 통해 정보를 쏟아내는 경우 대화의 부담을 느낀다고 답했다. 이미 <Table 4>에서 살펴 보았듯, 사용자에 비해 Mitsuku이 발화가 압도적으로 많은 것을 확인할 수 있었고 이는 Jones & Gerard(1967)의 상호작용 담화 모형을 기준으로 볼 때 '불균형적 결속성 (asymmetrical contingency)' 으로 분류된다. 즉 McLaughlin(1984)의 대화의 정의에 따르면 대화 참여자들이 대화의 주제를 교환하며 협력적이고 상호주체적으로 대화를 전개해 가는 것이 상호작용적 의사소통과정이라 할 수 있는데 Mitsuku가 대화를 주도하며 대화를 지배하는 구조 속에서는 상호작용적 대화를 유지하기 힘들다는 것이다. 따라서 정보검색 및 제공이라는 챗봇의 기능적 측면을 특별히 강조하는 것이 아니라 자연스런 대화에서 나타나는 상호작용을 구현하고자 한다면 적절한 상호 발화량과 그 양적 균형이 중요한 요인이라는 것을 말해준다. 또한, '인접 문단 간 개념적 의미의 유사성'이 높을수록 인간 발화와 유사하다는 결과를 볼 수 있고 상위권 학생의 경우, '동사구 사용의 다양성(No. 74)'이 높을수록 영어교육의 효과성에 기여하는 것으로 추정할 수 있었다. 즉 Mitsuku 발화에 다양한 동사구의 표현이 포함된 경우 상위권 학생들은 이를 유용한 학습 내용으로 인식한다는 것이다. 끝으로 상위권 '지문의 난이도(No. 106)'가 낮을수록 사용자의 발화량에 긍정적인 영향을 미친 것으로 나타났다.

V. Conclusions and Implications

본 연구에서는 4차 산업혁명의 도래와 더불어 언어교육의 중요한 도구로 자리 잡을 것으로 예상되는 챗봇의 활용과 관련하여 3개의 연구 질문을 설정하여 교육적 시사점 및 향후 챗봇 개발에 대한 제언을 도출하고자 하였다: ① 챗봇과 사용자의 발화량의 차이와 챗봇이 구사하는 발화의 이독성과 어휘의 수준은 적절성, ② '인간과의 유사성,' '영어교육의 효과성,' 및 '사용자 발화량'에 영향을 미치는 챗봇 발화의 텍스트 요인, ③ 기존 챗봇의 개선점. 이들 연구 질문에 결과를 요약하면 다음과 같다.

첫째, 상위권 학생과의 대화 시 Mitsuku의 발화량은 평균 614.44단어였고 학생들의 발화량은 250.22단어였다. 반면 하위권 학생들과의 대화에서

는 Mitsuku가 평균 427.84단어이며 학생들은 155.85단어로 대화의 길이가 상위권보다 상당히 짧았다. 그리고 이를 통해 Mitsuku의 발화가 Mitsuku 사용자 발화에 비해 2.5-2.7배 많다는 것을 확인할 수 있었다. 또한 Mitsuku의 발화는 모든 대화에서 영어 원어민 기준으로 9-10세에 해당하는 이독 지수를 보였고 Mitsuku나 학생 발화 모두 90% 이상이 3,000단어 안에서 구사된다는 것을 알 수 있었다. 이는 챗봇이 구사하는 발화의 난이도 측면에서 매우 긍정적인 결과라고 볼 수 있다.

둘째, 전체 53명을 대상으로 한 분석 결과, 챗봇이 구사하는 발화의 문장 구조가 단순할수록, 사용된 단어의 구체성이 낮을수록, 문장 간 명사와 대명사의 중복 사용 비율이 낮을수록, 인접 문단 간 개념적 의미의 유사성이 높을수록, 어휘 사용의 다양성이 낮을수록, 동사구 사용의 다양성이 낮을수록, 그리고 기초 수준의 내용어 빈도수가 높을수록 사용자의 발화량은 상승한다는 것을 확인하였다. 상위권(n=27)과 하위권(n=26)을 구분한 분석에서는 하위권 학생들의 경우, 평균 문장 길이가 짧을수록 발화량이 증가하는 것을 확인할 수 있었고 인접 문단 간 개념적 의미의 유사성이 높을수록 인간 발화와 유사하다고 생각하였다. 그리고 상위권 학생의 경우는 동사구 사용의 다양성이 높을수록 영어교육의 효과성에 기여한다고 판단하였다. 또한 상위권 학생들은 지문의 난이도가 낮을수록 더 많은 발화량을 보였다.

셋째, 물론 실험에 참여한 학습자들은 Mitsuku가 웹검색을 통해 정보를 제공해 주고, 모든 표현은 아니지만 일부 표현에 대해서는 문법적 피드백을 제공하며 자연스런 구어체의 표현을 구사하여 영어학습의 기회를 제공한다는 점, 그리고 인간과의 대화보다도 심리적으로 편안했다는 점을 장점을 꼽았으나 여전히 여러 가지 개선점도 지적하였다. 그 중 가장 많은 학습자들이 지적한 개선점은 맥락파악 능력의 향상이었고 이와 관련하여 데이터 확장의 필요성도 여러 학생들이 지적하였다. 특히 정보검색 시 지나칠 정도로 많은 양의 정보를 제공하거나 여러 질문에 대한 응답을 한꺼번에 나열하듯 제시하는 것은 상호작용의 저해 요인으로 꼭 개선되어야 할 점이라는 의견을 제시하였다. 뿐만 아니라 대명사 인식의 한계, 의미 조율 시 동일한 표현의 반복, 세밀한 감정 표현의 부재 등도 개선점으로 나타났다.

본 연구에서는 이러한 챗봇의 개선점을 고려하여 포털사이트(다음, 네이버 등)에서 뉴스 기사 내용을 요약하여 제공하는 소위 '요약' 또는 '깔때기' 기능을 적용하여 챗봇이 제공하는 정보의 양을 적절히 조절하고 이것이 학생들과의 상호작용에 어떠한 영향을 미치는지 분석하는 연구를 후속 연구로 제안하고자 한다.

References

Abu-Shawar, B. (2017). Integrating CALL systems with chatbots as conversational partners. *Computación y Sistemas, 21*(4), 615–626.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Colby, K. (1975). *Artificial paranoia: A computer simulation of paranoid processes.* New York: Elsevier.

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning and Technology, 10*(3), 8-14.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004, December 9). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers, 36*, 193-202. Retrieved from http://tool.cohmetrix.com/

Heatley, A., & Nation, I. S. P. (2002, June 15). *RANGE and FREQUENCY programs*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Range_GSL_AWL.zip

Jones, E. E., & Gerard, H. B. (1967). *Foundations of social psychology*. New York: John Wiley & Sons.

Kim, J. S. (2017). *The effects of human−AI assistant interactions on children's collaborative language acquisition*. Unpublished master;s thesis, Gwangju National University of Education, Gwangju.

Larson, S. (2017, February 9). Microsoft's Chinese A.I. is already chatting with millions. *The Daily Dot*. Retrieved from the World Wide Web: https://www.dailydot.com/debug/microsoft-chat-bot-china/

Lu, C-H., Chiou, G-F., Day, M-Y., Ong, C-S., & Hsu, W-L. (2006). Using Instant Messaging to Provide an Intelligent Learning Environment. In Ikeda M., Ashley K. D., Chan T-W. (eds), *Proceedings of the Intelligent Tutoring Systems (ITS) 2006 Lecture Notes in Computer Science: Vol 4053* (pp. 575-583). Heidelberg, Berlin: Springer.

Mauldin, M. (1994). Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In Proceedings of the 11th National Conference on Artificial Intelligence. Seattle, Washington: AAAI Press.

McLaughlin, M. L. (1984). *Conversation: How talk is organized. Sage Series in Interpersonal Communication 3.* Beverly Hills, CA: Sage.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston: Heinle Cengage Learning.

Shum, H., He, X., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering, 19*(1), 10-26.

Wang, Y. (2004). Distance language learning: Interactivity and fourth-generation Internet-based video-conferencing. *CALICO Journal, 21*(2), 373-395.

WebFX. (2019, April 10). *Readability test tool.* Retrieved from https://www.webfx.com/tools/read-able/

Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery, 9,* 36-45.

Worswick, S. (2018, September 13). *Mitsuku wins Loebner Prize 2018!* Retrieved from https://medium.com/pandorabots-blog/mitsuku-wins-loebner-prize-2018-3e8d98c5f2a7