한국어 튜터링 챗봇을 위한 말뭉치 구축

김한샘○, 최경호, 한지윤, 정해영, 곽용진
연세대학교, ㈜이르테크, 연세대학교, ㈜이르테크, ㈜이르테크
khss@yonsei.ac.kr, khchoi@iirtech.co.kr, hanjiyoon01@gmail.com, hyjung@iirtech.co.kr, silhuett@iirtech.co.kr

Building a Corpus for Korean Tutoring Chatbot

Hansaem Kim○, Kyung-Ho Choi, Ji-Yoon Han, Hae-Young Jung, Yong-Jin Kwak
Yonsei University, IIR Tech Inc., Yonsei University, IIR Tech Inc., IIR Tech Inc.

요 약

교수-학습 발화는 발화 턴 간에 규칙화된 인과관계가 강하고 자연 발화에서의 출현율이 낮다. 일반적으로 어휘부, 표현 제시부, 대화부로 구성되며 커리큘럼과 화제에 따라 구축된 언어자원이 필요하다. 기존의 말뭉치는 이러한 교수-학습 발화의 특징을 반영하지 않았기 때문에 한국어 교육용 튜터링 챗봇을 개발하는 데에 활용도가 떨어진다. 이에 따라 이 논문에서는 자연스러운 언어 사용 수집, 도구 기반의 수집, 주제별 수집 및 분류, 점진적 구축 절차의 원칙에 따라 교수-학습의 실제 상황을 반영하는 준구어 말뭉치를 구축한다. 교실에서 발생하는 언어학습 상황을 시나리오로 구성하여 대화 흐름을 제어하고 채팅용 메신저와 유사한 형태의 도구를 통해 말뭉치를 구축한다. 이 연구는 한국어 튜터링 챗봇을 개발하기 위해 말뭉치 구축용 챗봇과 한국어 학습자, 한국어 교수자가 시나리오를 기반으로 발화문을 생성한 준구어 말뭉치를 최초로 구축한다는 데에 의의가 있다.

주제어: 한국어교육(Korean Language Education), 튜터링(tutoring), 챗봇(chatbot), 말뭉치(corpus)

1. 서론

이 연구의 목적은 한국어 튜터링 챗봇 개발에 필요한 말뭉치의 구축이다. '한국어 튜터링 챗봇'은 물리적 거리나 비용의 한계를 극복하고 챗봇과의 대화를 통해서 자연스럽게 한국어 구사 능력과 언어 지식을 습득할 수 있도록 하는 대화 시스템을 의미한다. 이러한 시스템을 개발하기 위해서는 한국어 교수-학습이라는 특수한 영역에 최적화된 말뭉치가 필요하고 이를 구축할 방법론과 도구에 대해 논의해야 한다. 자연언어 기반 대화형 인터페이스는 한국어를 모국어로 하는 화자의 일반 대화 자료를 주로 사용해 왔다. 그러나 자연스러운 대화 전개를 위해서는 해당 도메인과 채팅 환경 등 실제 시스템이 작동하는 환경에서 생성된 언어자원이 필요하다. 이 연구에서 구축하여 활용하고자 하는 말뭉치는 실제 한국어 교수-학습 상황을 반영하기 때문에 챗봇이 튜터링 프로세스를 진행하는 데에 직접적인 도움을 줄 수 있다.

1960년대에 세계 최초로 Brown 말뭉치가 구축된 이후 지금까지 수많은 다양한 말뭉치가 구축되어 왔는데, 이들은 자연언어처리 분야에서의 활용이라는 관점에서 크게 세 가지 유형으로 나눌 수 있다. 첫 번째는 Brown 말뭉치로부터 영국 국가 말뭉치(BNC: British National Corpus)로 이어진 전통적인 말뭉치로 인간의 언어를 있는 그대로의 모습으로 관찰하고자 하는 데 그 특징이 있다. 전통적인 말뭉치는 형태소 분석, 구문 분석, 의미 분석과 같은 기초적인 언어 처리에 주로 사용된다. 두 번째로 게임, 영화, 소설과 같이 내러티브가 있는 콘텐츠 제작, 텍스트의 요약 및 생성 등에 사용되는 말뭉치가 있다. 이들 말뭉치는 이야기 모티브와 전개 정보, 주제에 대한 정보를 중점적으로 담고 있어서 컴퓨터가 화제의 전개와 커뮤니케이션, 창의성 등을 모방하는 데 사용된다. 마지막으로 말뭉치에 포함된 광범위한 정보를 제한하여 목표 시스템 또는 서비스 환경에서의 언어 사용을 충실하게 담아내는 말뭉치들이다. Standford의 QnA를 위한 SQUAD나 대화형 에이전트 개발에 자주 사용되는 Ubuntu 채팅 말뭉치, 트위터나 페이스북 데이터가 대표적이다. 이들 말뭉치는 목표 시스템이나 서비스의 환경 하에서 생성된 인간의 언어 사용을 담고 있어서 보편적인 언어 사용 정보로 인한 기계학습 효과의 발산을 막는다. 이러한 말뭉치들은 구축 단계부터 특정 시스템이 결합됨으로써 전산적 처리가 용이하고 일관성이 높아 챗봇, 자동 QA, 감성 분석 등과 같은 목표 지향적 시스템 및 서비스 개발에 효과적이다.

이 연구에서 구축하는 말뭉치는 마지막 유형에 해당하는 말뭉치로 한국어 튜터링용 챗봇이라는 구체적인 시스템의 개발을 위해 기계에 학습시킬 데이터로서, 말뭉치 구축용 챗봇과 한국어 학습자, 한국어 교수자 등의 화자가 시나리오를 기반으로 발화문을 생성하여 구축하는 최초의 말뭉치이다.

2. 관련 연구

Facebook 등에서는 다양한 챗봇용 언어자원을 수집하기 위한 플랫폼을 개발하여 공개하고 있다[1]. 챗봇을 이용한 언어자원 수집은 챗봇과 인간 사용자간의 대화뿐만 아니라, 챗봇과 챗봇, 챗봇들과 인간이 복합된 환경 등 다양한 상황, 테마, 발화 주체를 구성하여 진행되고 있다. 챗봇을 위한 트레이닝 데이터를 확보하기 위해 실제 대화를 전사하여 구축한 스크립트를 가공하거나, 특정 서비스를 이용하기 위한 대화를 전사하여 텍스트화하는 방식, 웹에서 수집한 대화를 가공하는 방식으로 데

이터를 구축한다[10]. 교육, 학습 분야에서도 컴퓨터를 활용한 튜터링 기술을 주목하면서 CALL이나 튜터링용 챗봇 개발에 대한 연구가 수행되어 왔다[2]. 언어 튜터링 챗봇이라는 언어 교육용 시스템 개발을 위해 특화된 말뭉치를 따로 구축한 경우는 흔하지 않다. 교육용 대화 시스템의 논리 전개와 추론을 위해 오류를 포함하는 영어 학습자 말뭉치를 통사의미적으로 분석하여 활용한 사례가 있다[11]. 국내에서도 학습자의 중간 언어 사용 양상을 구축한 한국어 학습자 말뭉치가 국가 차원과 대학 연구소 차원에서 구축된 바 있다[12], [13]. 그러나 발화자와 주제가 통제되지 않은 학습자 말뭉치는 바로 언어 처리에 활용하기 힘들고 특히 한국어 학습자 말뭉치는 문어 중심으로 구축되어 있어 대화를 전제로 하는 챗봇 시스템에는 활용도가 떨어진다. 학습자 구어 말뭉치를 대신해 활용할 만한 말뭉치로 한국어 교재 말뭉치가 있다. 연세대 언어정보연구원에서 구축하여 서비스하고 있는 한국어 교재 말뭉치는 교육 항목 전달을 위해 인위적으로 만든 정제된 대화로 구성되어 있어 오류를 포함하는 학습자 말뭉치보다 활용도가 높다[4]. 다만 발화의 길이와 수준이 학습자 등급에 따라 통제되어 있어, 대화 참여자의 발화가 담당한 기능을 화제 도입, 화제 전개, 상대방 발화 내용 확인, 발화 보충, 발화 수정, 대답 발화, 대화 지속 반응, 발화 지연, 의식적 표현 등으로 분류할 때[14], 기본적인 기능을 수행하는 발화에 치우쳐 있어 자연스러운 의사소통을 산출하기 위한 자료로는 부족하다. 이 연구와 같이 교육용 챗봇 시스템에 활용하기 위해 실제 교수-학습 현장에서의 교수자와 학습자의 상호 작용을 반영한 말뭉치를 구축하기 위한 논의는 진행된 바가 없다.

3. 교수-학습 말뭉치 구축 방법론

3.1. 교수-학습 말뭉치의 필요성

한국어 교재 말뭉치 등에서 볼 수 있는 언어 학습의 대화 상황 예시는 주로 다음과 같이 이루어져 있다.

정우: 투이 씨, 로라 씨가 이번에 승진을 했어요.
이야기 들었어요? 정말 잘 됐어요.
투이: 어제 로라씨가 저에게도 전화했어요. 정말 잘됐어요.
정우: 그래서 내일 저녁에 친구들이 모여서 로라 씨 승진을 축하해 주려고요.
투이: 좋아요. 갈게요. 저도 뭘 좀 도와드릴게요.

예시1 세종 한국어3(p36.대화2) 국립국어원

로라: 저건 색깔이 너무 화려할까요?
직원: 어머니께 드릴 선물이라면 이 색깔이 더 좋을 것 같습니다.
로라: 음, 그럼 이걸로 포장해 주세요. 그런데 혹시 어머니가 보시고 색깔을 마음에 들어하지 않으시면 교환할 수 있어요?
직원: 네, 일주일 이내에 오시면 교환이 가능합니다. 교환하러 오실 때는 반드시 영수증을 가지고 오셔야 합니다.

예시2 세종 한국어5(p70.대화2) 국립국어원

위의 예는 실제 대화를 전사한 구어 말뭉치가 아니라 대화를 연습하기 위한 스크립트로서 준구어 말뭉치의 일종이다. 이러한 대화는 실제 발화 상황을 그대로 반영할 수 없기 때문에 챗봇 시스템에 활용하는 데에 있어 여러 가지 한계점을 가지고 있다. 특정 상황에 대한 암묵적 가정 하에 대화를 진행하므로 도입부, 마무리 없이 배워야하는 교육 항목에 해당하는 4~5 턴의 짧은 대화만 주어져 있다. 대화를 시작하거나 마무리할 때에 필요한 의식적 발화가 생략되어 있다는 것이다. 교육 항목에 해당하는 어휘와 표현을 포함시키기 위해 구성한 대화이므로 교실수업에서 교사-학생의 대면을 전제로 함에도 불구하고 학습자의 감성 화행이 간과되어 관계 중심적 학습 보다는 과제 중심적 학습 지문의 흐름이 연출되어 있다[9].

이러한 한계를 극복하기 위하여 챗봇 등의 컴퓨터 대화 기반 튜터링에 필요한 교수-학습 말뭉치를 구축할 때에 도입부터 마무리까지 1:1 '챗봇:학습자' 대화 형태로 구성되어야 하며, 어휘나 표현은 학습자의 수준과 상황에 따라 달라지므로, 교과서상 제시된 맞춰진 본문과는 달리 예측 어려운 학습자 발화에 대한 대응까지 고려하는 어려움이 있다. 따라서 대화쌍도 교사인 챗봇 발화를 포함시켜야 하며, 최소 7~8 턴 이상으로 구성해야 한다. 기존의 말뭉치를 활용하기 힘들고 챗봇을 위한 말뭉치를 따로 구축해야 할 필요성이 여기에 있다.[5][6]

교수-학습 말뭉치는 지식의 전달과 숙련을 위한 구조화된 흐름 속에서 제한된 언어 사용이 이루어진다. 세종 말뭉치나 BNC와 같은 전통적 말뭉치에도 이러한 교수-학습 상황의 언어사용 자료가 포함되어 있으나, 일정한 교육적 목표와 일관된 다양한 학습 주제를 획득하기에는 어려움이 있다. 교수-학습 발화, 특히 언어학습에 대한 교수-학습 발화에는 다음과 같은 특징이 있다.

1) 발화 턴 간에 규칙화된 인과관계가 강함
2) 자연 발화에서의 출현율이 낮음
3) 일반적으로 어휘부, 표현 제시부, 대화부로 구성
4) 커리큘럼과 화제에 따른 언어자원 구성 필요

교사-학생간의 발화는 교사의 질문이나 학습 지시에 대한 학생의 반응, 질문과 수행으로 이루어지는 경우가 많다. 이러한 특성은 보편적인 언어현실을 포착하기 위한 전통적 말뭉치에서는 잘 드러나지 않는다. 대부분의 발화에도 발화 턴간에 유의미한 인과관계가 발생하나, 교수-학습과 같은 특수한 환경에서는 교사의 질문, 지시가 발화의 흐름을 주도한다.

또한 교수-학습 발화는 교육하고자 하는 내용(어휘, 표현, 문법)과 이를 이해시키기 위한 설명적 재료(화제, 문화, 경험)가 일련의 연관성을 지니고 구조화되어 나타난다. 예를 들어 [날씨]를 [묻는] [표현]을 가르치기 위해 [최근 장마철 날씨]를 주제로 한 대화 상황을 조성하여 교육을 진행한다. 그러므로, 교수-학습 발화 말뭉치는 이러한 교육적 구성과 화제에 따라 언어자원을 수집하여 분류해야 한다.

3.2. 교수-학습 말뭉치 구축의 원칙

이 연구에서는 한국어 학습을 위한 말뭉치를 구축함에 있어서 아래 4가지의 원칙을 견지한다.

1) 자연스러운 언어 사용 수집
한국어 교재에 등장하는 대화문과 같이 인위적으로 교사-학습자 발화를 작성하지 않는다. 기능을 중심으로 분류할 때 발화의 규형이 골고루 적재적소에 배치될 수 있도록 실제 학습자의 발화문과 교수자의 발화를 통한 교육적 접근을 수집한다.

2) 도구 기반의 수집
교사와 학습자의 대화가 일정한 교육 흐름에 따라 유도록 되도록 대화형 인터페이스를 제공하고, 특히 교사의 발화는 교육적 흐름에 대한 발화 의도가 자연스럽게 주석될 수 있도록 교육 흐름의 정보를 제공한다. 교수-학습 말뭉치 구축용 도구는 학습자의 반응과 교육 흐름에 따른 다양한 교수 발화를 수집하는 교수 발화 수집 모드, 학습자의 다양한 반응을 수집하는 학습자 발화 수집 모드, 수집된 발화 또는 챗봇간 발화의 문제점을 교정하는 관찰 모드가 제공된다. 또한 수집된 말뭉치로부터 발화 턴간 상관 관계 분석이 용이하도록, 형태, 구문, 발화 의도를 주석할 수 있는 자동/반자동 주석 도구를 제공한다.

3) 주제별 수집 및 분류
앞서 언급 바와 같이, 동일한 교수 흐름이라도 학습자의 흥미, 수준, 반복 정도에 따라 서로 다른 주제와 교육 내용이 필요하다. 예를 들어 '시제'에 대한 문법적 교육을 위해 사용될 수 있는 주제는 '날씨'일 수도 있고, '여행' 계획에 대한 주제일 수도 있다. 또한 각각의 주제와 상황으로부터 교육 주제로 이끌어 가는 방법 또한 다양하게 나타난다. 그러므로 교사-학습자의 발화쌍은 주제별로 분류될 것을 고려해야 한다.

4) 점진적 구축 절차

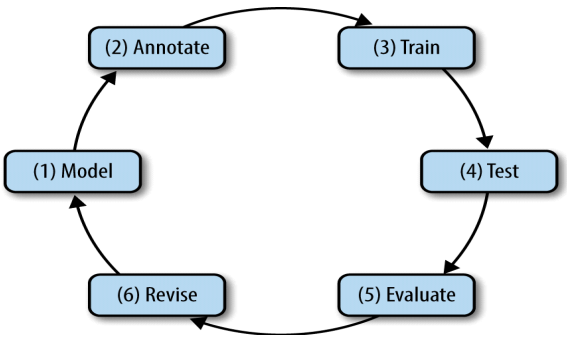

그림 1 MATTER Cycle(Pustejovsky & Stubbs 2012)

말뭉치의 구축이 점진적으로 수행되어야 함은 오래전부터 강조되어 왔다[7],[8]. 특히, 보편적인 언어현실을 반영하기 위해 자연스러움이 강조되는 전통적 말뭉치와 달리 기계학습을 전제로 하는 말뭉치는 말뭉치가 기계학습의 대상인 동시에 결과물이므로 구축 목적의 지향성과 부합하는지를 지속적으로 확인하는 과정이 필요하다. 그림 1은 Pustejovsky & Stubbs이 기계학습을 위한 말뭉치와 주석 개발을 위해 제시한 MATTER 프로세스이다. 지식 모델과 지침을 기반으로 구축한 말뭉치를 기계학습을 통한 훈련, 테스트, 평가를 거쳐 지식 모델과 지침을 개선한 뒤 데이터를 확대하는 방식으로 언어자원의 양적 희소성과 소규모 데이터를 학습 모델의 수렴에 용이하게 하는 장점이 있다. 대규모 자료 수집이 쉽지 않고 유사한 기존 말뭉치가 존재하지 않는 교수-학습 말뭉치의 구축에 적합하다.

3.3. 교수-학습 말뭉치의 설계

한국어 교육용 챗봇 개발에 필요한 준구어 말뭉치 구축은 우선 교수-학습 상황을 배경으로 한 예상 대화 시나리오를 기본으로 한다. 이 대화 시나리오는 교수해야 하는 어휘와 표현, 주제 등을 고려하여 설계된다.

대화는 교사의 발화와, 이에 대한 학습자의 예상 발화, 그에 따른 교사의 피드백이 하나의 단위로 구성되며, 이 단위들이 복수 개로 모여 일정한 흐름을 지닌 것이 대화 시나리오가 된다. 한국어 교육용 챗봇에 사용될 대화 말뭉치 구축 도구는 실제 교사와 학습자에게 각각의 발화를 수집하기 위해 이러한 학습 시나리오에 따른 흐름을 교사와 학습자에 제공하도록 설계되었다.

교사와 학습자에게 기대하는 역할이 다르기 때문에 교사와 학습자는 서로 다른 환경에서 구축에 참여하게 된다. 교사는 발화의 적절성을 판단하면서 동시에 발화를 생성한다. 발화를 생성하는 동시에 제공받은 대화 단위의 구성과 흐름이 적절한 지 판단하는 것이다. 이러한 과정을 통해 실제 교수-학습 상황과 유사한 발화를 수집하여, 실제 챗봇이 학습자에게 다양한 발화를 제공할 수 있게 된다. 또한 시나리오를 벗어나는 학습자의 발화에 대해 유연한 대응이 가능해진다. 학습자의 경우, 제공된 발화에 대한 응답을 작성한다. 이 때 수집한 학습자의 발화는 실제 대화형 인터페이스에 입력될 학습자의 발화 예측을 돕는다. 말뭉치 안에 다양한 유형의 학습자 발화가 포함될수록 대화형 인터페이스의 품질이 높아진다.

교사와 학습자의 상호작용이 반영된 말뭉치를 구축하기 위한 세부 계획은 다음과 같다. 한국어 급수 3-4급에 해당하는 중급 실력의 학습자 20명과, 경력 5년 이상의 한국어 교사 5명을 실험대상으로 한다. 총 구축 단계는 6차례로 10개 주제에 대한 서로 다른 시나리오에 대한 학습자와 교사의 발화를 수집한다. 1차와 2차는 본 수집에 앞선 파일럿 작업으로 우선 1개 주제에 대한 데이터를 수집한다. 1차에서는 예상된 시나리오에 대한 학습자의 발화를 수집한다. 2차 구축 단계에서는 1차에서 수집된 학습자의 발화를 교사에게 제공하여 기존 시나리오의 수정 발화와 학습자 발화에 대한 응답 발화를 수집한다. 1차와 2차의 발화를 토대로 보완점을 개선한 뒤, 주제를 확장하여 학습자와 교사의 발화를 교대로 수집하는 방식으로 3-6차 수집을 완료한다. 예상 수집 발화량은 10만

어절로, 각 7-8 단위로 구성된 10개 주제 시나리오에 대하여 각각 100개의 변이형, 총 1000개 내외의 변주된 시나리오를 수집하게 된다.

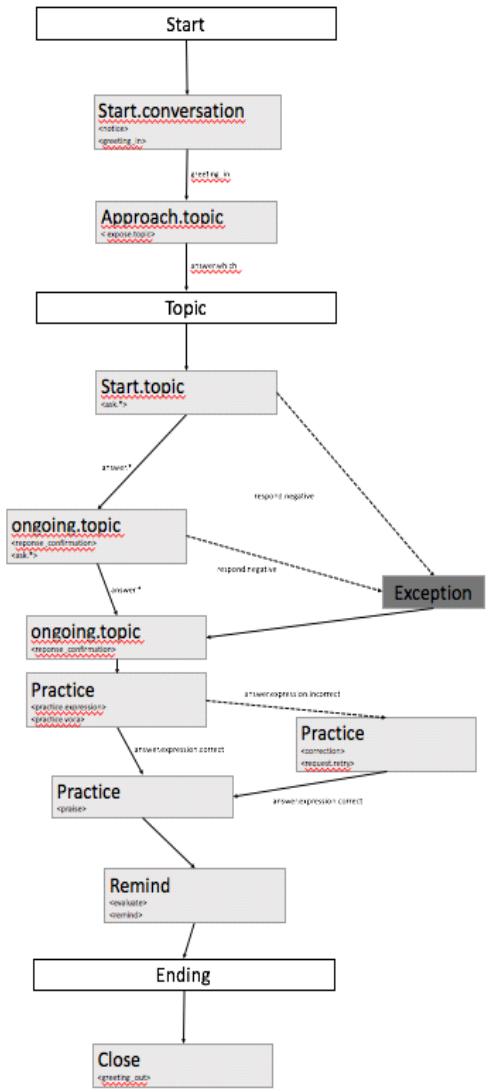

그림 2 교수-학습 말뭉치 구축도구의 교수 시나리오

4. 교수-학습 말뭉치 구축 도구

교수-학습 말뭉치는 교사-학습자간의 발화를 수집하기 위해 기본적으로 채팅용 메신저와 유사한 형태를 갖는다. 또한 언어학습 상황을 충실히 재현하도록 교실에서 발생하는 언어학습 상황을 시나리오로 구성하여 대화 흐름을 제어한다.

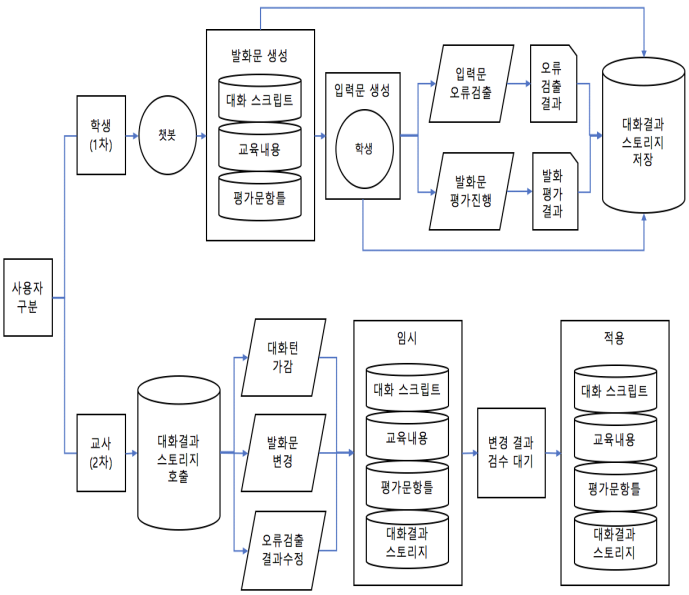

그림 3 말뭉치 구축 도구 시스템 구성도

학습자 대화 말뭉치 구축 프로세스는 크게 두 단계로 구분된다. 1차 작업의 목표는 첫째로 챗봇 발화에 대한 학습자의 응답문 데이터를 획득하기. 둘째로 챗봇 발화의 적절성에 대한 학습자의 피드백 정보를 얻기이다. 따라서 1단계에서는 기 제작된 시나리오에 의거하여 챗봇과 학습자가 대화를 진행하면서 학습자가 입력한 문장을 기록하고, 학습자가 챗봇의 발화문에 대해 추가 정보를 요청할 수 있게 하여 최종적으로 획득한 조작기록을 분석해 챗봇 발화문의 적절성에 대해 간접적으로 평가할 수 있도록 한다. 동시에 시스템에 내장된 오류검출 기능을 통해 오류라고 인식된 부분들도 함께 대화 기록 파일에 저장한다. 2차로 학습자와 챗봇의 대화를 통해 생성된 대화 기록을 기반으로 한국어 교사가 챗봇 시나리오의 적절성 및 시스템의 오류검출 결과의 타당성에 대해 검수하고 수정하는 작업을 진행한다. 검수작업 결과는 개인 작업자 별로 보관되었다가 시스템 적용 검토과정을 거쳐 최종적으로 챗봇 시스템에 적용하게 된다.

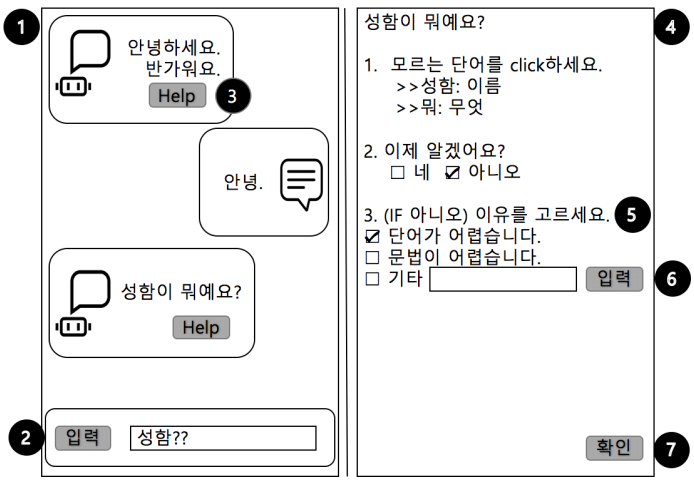

그림 4 학습자 사용화면

1차 구축작업에서 사용되는 학습자 사용화면을 살펴보면, ①영역에서 실제 챗봇과 학습자의 사이에 진행되는 대화를 볼 수 있다. 학습자는 챗봇의 발화문을 이해한 뒤 ②영역에 문장을 입력하는 방식으로 응답할 수 있다. 만약 챗봇의 발화를 이해하지 못하여 도움이 필요한 경우 ③Help 버튼을 클릭하여 ④영역에 나타난 발화문 정보요청 페이지를 통해 부족한 정보를 획득하는 작업을 진행한다. ⑤부분은 만약 학습자가 시스템에서 제공하는 정보를 획득하고도 충분하지 못하다고 생각하는 경우 발화문 적절성을 저해하는 요소가 무엇인지 파악할 수 있도록 추가된 문항이며 선택사항이 없을 경우에는 ⑥을 통해 서술형으로 응답이 가능하다. 사용자가 ⑦확인 버튼을 클릭하면 다시 ②의 입력창이 활성화 되어 대화를 진행 할 수 있다.

2차 구축작업에서 사용되는 교사 사용화면을 살펴보면 오류 검출 결과와 시나리오의 적절성을 검수하는 기능이 함께 제공된다. 다만 그림 4와 그림 5는 기능 설명을 위해 별도로 분리하였으며 실제로는 작업화면 좌측(그림 4, 5의 ①번 영역)의 대화 기록창에 제공된 기능 버튼을 클릭하는 것에 따라 우측(그림 4의 ③번, 그림 5의 ⑥번 영역)에 Toggle형태로 출력되는 작업화면이 달라지게 된다.

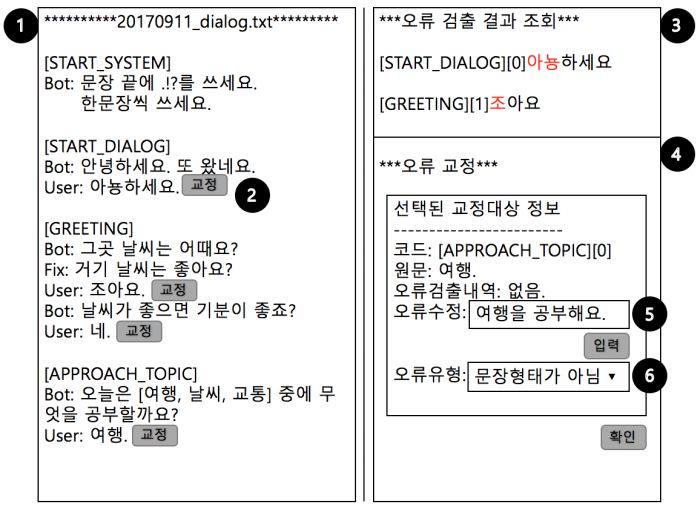

그림 5 교사 사용화면 - 오류 검출 결과 검수

그림 5는 시스템에서 자동으로 검출한 학습자 입력문의 오류내역을 한국어 교사가 검수할 수 있도록 특정 대화기록에 대한 오류검출내역을 ③번 영역에 출력해주고 교사는 검출내역에서 누락되거나 부적절한 교정을 발견하였을 경우에 ②번의 교정버튼을 클릭하여 나타난 ④번 영역의 교정작업 창에서 검수작업을 진행하게 된다. ⑤번에 적절하다고 판단되는 오류교정 내용을 서술형으로 입력하고 ⑥에서 오류의 유형을 선택한 뒤 ④번 영역 하단의 확인버튼을 클릭하여 작업결과를 저장한다.

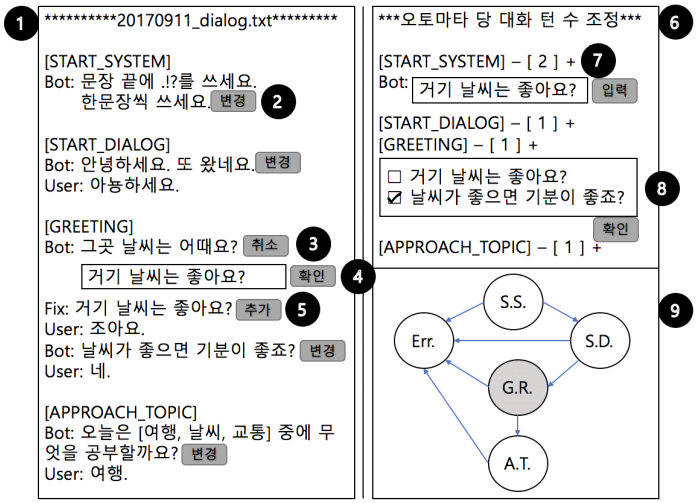

그림 6 교사 사용화면 - 시나리오 적절성 검수

그림 6은 챗봇의 발화와 학습자의 응답 기록을 한국어 교사가 살펴보고 기록된 학습자의 응답 양상과 그림 3에서 획득한 학습자의 챗봇 발화에 대한 Feedback 정보를 토대로 챗봇의 발화문을 수정하는 검수기능이 제공되는 화면이다. 교사는 챗봇 발화가 부적절하다고 판단되면 ②번의 변경버튼을 클릭한다. ④번 영역에 수정할 문정을 입력하고 확인버튼을 눌러 저장하면 ⑤번에 [Fix]라는 표지가 붙은 수정문장이 신규로 추가되어 교사가 수정한 문장을 시각적으로 확인이 가능하다. 또한 문장 끝의 추가버튼을 클릭하면 해당 단계에서 적절하다고 생각되는 수정문장을 추가로 입력이 가능하기 때문에 각 대화 시나리오 단계의 특정 대화 턴에 복수의 수정문을 입력 가능하다. ⑥번 영역에서는 각 대화 시나리오 단계의 대화 턴이 부족하거나 많다고 생각되는 경우 (+)(-)버튼을 클릭하여 턴 수를 조정할 수 있다. 대화 턴을 추가하는 경우에는 ⑦번 영역처럼 챗봇의 발화문을 교사가 직접 입력하여 추가가 가능하고, 대화 턴을 삭제하는 경우에는 ⑧번 영역처럼 기 제작된 발화문이 제공되어 그 중 한 가지를 선택해 제거가 가능하다. ⑨번은 현재 작업중인 대화 시나리오 단계를 시각적으로 강조하여 교사의 검수작업에 도움을 주고자 하였다.

5. 결론 및 향후 과제

이 연구는 한국어 튜터링 챗봇을 개발하기 위해 말뭉치 구축용 챗봇과 한국어 학습자, 한국어 교수자의 세 주체가 대화 흐름에 통제를 가한 시나리오를 기반으로 발화문을 생성한 준구어 말뭉치를 구축하는 것을 목적으로 하였다. 학습자의 감성 화행을 간과한 과제 중심적 대화문으로 구성된 짧은 대화 자료의 한계를 극복하기 위해 학습 시나리오에 따른 흐름을 교사와 학습자에 제공하도록 설계된 말뭉치 구축용 챗봇과 학습자의 대화, 이에 대한 교수자의 검증 및 응답 발화 생성 등의 단계를 거쳐 1000개 내외의 변주된 시나리오를 10만 어절 내외로 구축한다.

구축된 한국어 교수-학습 말뭉치는 교육용 목적에 최적화된 준구어 말뭉치로서 튜터링 챗봇 개발에 직접적으로 활용할 수 있는 데이터이다. 자연 발화의 출현율이 높아 말뭉치 전체가 학습 데이터로 사용될 수 있으며, 학습자의 경험을 토대로 한 다양한 발화를 포함하므로 챗봇 인터페이스의 품질을 향상시킬 수 있다. 구체적으로 제안한 말뭉치 수집 방법과 구축 도구는 챗봇용 학습 콘텐츠를 개발하고 챗봇의 대응 알고리즘을 구성하는 데에 기여할 것으로 기대한다. 말뭉치 수집이 진행되면서 데이터가 구축 목표에 부합한지 확인하는 절차를 통해 점진적으로 말뭉치의 완성도가 높아지는 과정에 대한 논의와 실제 챗봇 시스템에 활용한 결과에 대한 분석이 향후 과제이다.

감사의 글

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의지원을 받아 수행된 연구임 (No.2017-01217,말하기/쓰기 평가와 챗봇을 이용한 1:1 언어학습 튜터링 기술 개발)

참고문헌

[1] Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, Jason Weston, ParlAI: A Dialog Research Software Platform, eprint arXiv:1705.06476, 2017.

[2] Jiyou Jia, “CSIEC: A computer assisted Englishlearning chatbot based on textual knowledge and reasoning”, Knowledge-BasedSystems 22(4), p.p 249-255, Elsevier B.V. 2009.

[3] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, Bill Dolan, "A Persona-Based Neural Conversation Model", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.994-1003, Berlin, Germany, 2016.

[4] 언어정보연구원, 한국어 교재 말뭉치 https://ilis.yonsei.ac.kr/corpus/koreantext3

[5] 박범준(2010), “콘텐츠 로봇의 감성적 반응을 위한 지능형 메신저 개발”, 한국콘텐츠학회논문지 ‘10Vol.No.9.pp.13~14

[6] 조윤주 외(2009), 모바일 환경에서의 대화형 에이전트와 대화 내용에 관한 연구

[7] Biber, D. Finegan, E. “On the exploitation of computerized corpora in variation studies”, in Aijmer,K. & Altenberg, B. (ed.), 1991.

[8] James Pustejovsky, Amber Stubbs, “Natural Language Annotation for Machine Learning”, O’Reilly Media, 2012

[9] 박창균, “대화분석을 통한 말하기 교수-학습 방법 연구” 인천교대 교육대학원 초등국어 교육전공 석사학위논문,pp.23-25,1998

[10] Kadlec, R., Schmid, M., & Kleindienst, J. (2015). Im-proved deep learning baselines for ubuntu corpus dialogs. arXiv preprint arXiv:1510.03753

[11] Jia, J. (2009). CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning. Knowledge-Based Systems 22(4), 249-255.

[12] 강현화, “학습자 말뭉치의 구축과 활용연구 ‘, 소통, 2017

[13] 서상규, 유현경, 남윤진, 한국어 학습자 말뭉치와 한국어 교육, 한국어 교육, 제13권 제1호, pp. 127-156, 2002

[14] 강현주, 말하기 능력 평가에서 대화 과제 도입의 필요성, 어문논집, 71, pp.353-376, 2014