DECO-LGG 반자동 증강 학습데이터 활용 멀티태스크 트랜스포머 모델 기반 핀테크 CS 챗봇 NLU 시스템

유광훈○, 황창회, 윤정우 & 남지순
한국외국어대학교, DICORA 연구센터
rhkdgns2008@naver.com, hch8357@naver.com, skyjw1211@gmail.com, jeesun.nam@gmail.com

Multitask Transformer Model-based Fintech Customer Service Chatbot NLU System with DECO-LGG SSP-based Data

Gwang-Hoon Yoo○, Chang-Hoe Hwang, Jeong-Woo Yoon & Jee-Sun Nam
DICORA, Hankuk University of Foreign Studies

요 약

본 연구에서는 DECO(Dictionnaire Electronique du COreen) 한국어 전자사전과 LGG(Local-Grammar Graph)에 기반한 반자동 언어데이터 증강(Semi-automatic Symbolic Propagation: SSP) 방식에 입각하여, 핀테크 분야의 CS(Customer Service) 챗봇 NLU(Natural Language Understanding)을 위한 주석 학습 데이터를 효과적으로 생성하고, 이를 기반으로 RASA 오픈 소스에서 제공하는 DIET(Dual Intent and Entity Transformer) 아키텍처를 활용하여 핀테크 CS 챗봇 NLU 시스템을 구현하였다. 실 데이터을 통해 확인된 핀테크 분야의 32가지의 토픽 유형 및 38가지의 핵심 이벤트와 10가지 담화소 구성에 따라, DECO-LGG 데이터 생성 모듈은 질의 및 불만 화행에 대한 양질의 주석 학습 데이터를 효과적으로 생성하며, 이를 의도 분류 및 Slot-filling을 위한 개체명 인식을 종합적으로 처리하는 End to End 방식의 멀티태스크 트랜스포머 모델 DIET로 학습함으로써 DIET-only F1-score 0.931(Intent)/0.865(Slot/Entity), DIET+KoBERT F1-score 0.951(Intent)/0.901(Slot/Entity)의 성능을 확인하였으며, DECO-LGG 기반의 SSP 생성 데이터의 학습 데이터로서의 효과성과 함께 KoBERT에 기반한 DIET 모델 성능의 우수성을 입증하였다.

주제어: 챗봇 NLU 시스템, 멀티태스크 트랜스포머 모델, 반자동증강 학습데이터, DECO 전자사전, LGG 언어자원

1. 서론

본 연구에서는 DECO(Dictionnaire Electronique du COreen) 한국어 전자사전[1]과 LGG(Local-Grammar Graph) 프레임[2]에 기반한 반자동 언어데이터 증강(SSP: Semi-automatic Symbolic Propagation)[3] 방식에 입각하여 핀테크 분야의 CS(Customer Service) 챗봇 개발을 위한 NLU(Natural Language Understanding) 주석 학습 데이터를 생성하고, 이 데이터를 기반으로 RASA 오픈 소스에서 제공하는 DIET(Dual Intent and Entity Transformer) 아키텍처와 KoBERT를 활용하여 챗봇 NLU 시스템을 구현하는 과정을 소개하는 것을 목적으로 한다. 실 데이터에 기반한 DECO-LGG 데이터 생성 모듈은 핀테크 CS NLU 처리를 위한 고객 질의 및 불만 화행(Speech act)에 대한 정교한 주석 학습 데이터를 효과적으로 생성하며, 이를 토대로 본 연구에서는 의도 분류(Intent Classification) 및 Slot-filling을 위한 개체명 인식을 종합적으로 처리하는 End to End 방식의 멀티태스크 트랜스포머 모델로 학습함으로써 그 성능을 입증하였다.

챗봇이란 사람과 컴퓨터가 상호작용하는 대화형 시스템을 일컫는 것으로, 전달매개, 지식제공범위, 목적 등에 따라 여러 방식으로 분류가 가능하다[4]. 전달매개의 경우 텍스트와 음성으로 분류할 수 있고, 지식제공범위의 제약에 따라 개방형 도메인(Open Domain)과 폐쇄형 도메인(Closed Domain)으로 나눌 수 있으며, 서비스의 목적 유무에 따라, 목적지향형 챗봇과 비목적형 챗봇으로 분류할 수 있다. 챗봇 개발을 위해서는 NLU(Natural Language Understanding)와 NLG(Natural language generation)를 모두 포괄한 복잡한 자연어 처리가 요구된다[5].

챗봇 서비스는 자연어 처리 산업에서 가파르게 성장하고 있는 유망 분야 중 하나로, 전 세계 챗봇 시장은 2019년도부터 연평균 약 30%의 성장률을 기록함에 따라 2024년에는 94억 달러 규모에 이를 것으로 전망되는, 연구 부가가치가 매우 높은 산업이다[6]. 챗봇은 음식 배달부터 비행기/호텔 예약까지 다양한 사람들의 니즈에 맞는 서비스를 제공하고 있으며, 특히 불필요한 인건비를 줄이면서 24시간 CS 대응을 지원할 수 있게 되었다[7]. 산업에서 효과적으로 활용되고 있는 CS 챗봇의 경우, 특정 목적을 설정한 폐쇄형 도메인 챗봇인 경우가 대부분이며, 이때 빠른 응답을 바라는 고객 요구에 맞춰 질의 및 불만 의도를 파악하는 효과적인 NLU 처리가 매우 중요하다.

이러한 NLU 학습을 위해 대용량의 학습 데이터가 필요한데, 여기서 의도 분류 및 슬롯에 대한 주석 처리가 필수적이다. 그런데 무분별한 데이터 수집으로 인해 생길 수 있는 개인정보 이슈 문제와 맞물려서, 이러한 연구를 위한 학습데이터 구축에 요구되는 원시데이터의 수집이 쉽지 않아, 한국어 NLU 데이터 생성 모듈 및 효과적인 학습 모델에 대한 중요성이 높아지고 있다.

현재 이와 같은 CS 챗봇 NLU 연구의 필요성에도 불구하고, 대부분의 연구는 영어를 기반으로 진행되고 있으며, 더욱이 한국어의 경우, 개인정보 수집 이슈로부터 자유로운 다양한 데이터 생성 모듈에 대한 연구는 부족한 상황이다. 이와 같은 문제를 해결하기 위해, 본 연구는 토스(TOSS) 핀테크앱 등 6개의 핀테크 애플리케이션의 사용자 불만 리뷰와 네이버 지식인의 금융 관련 질의문 데이터를 크롤링하여, 핀테크 CS 분야에서 핵심적으로 나타날 수 있는 32가지의 토픽 표현 및 38가지의 이벤트를 분류하고, 이와 공기할 수 있는 10가지 담화소를 분석 및 검토하여 각각의 카테고리를 구상하였다. 이를 통해 이벤트와 담화소 조합의 언어 패턴 문법을 면밀히 분석하였고, 핀테크 CS 챗봇 NLU를 위한 DECO-LGG SSP기반 데이터 생성 모듈을 구축하여, 질의 및 불만 화행에 대한 방대한 패턴의 주석 학습 데이터를 효과적으로 생성하였다. 생성된 데이터를 기반으로 학습된 NLU 모듈에 대해 RASA 오픈 소스에서 제공하는 멀티태스크 트랜스포머 모델 DIET을 활용하여 핀테크 분야에서 반복적으로 나타나는 다양한 이벤트에 대한 채팅 테스트셋으로 성능을 확인한 결과, DIET-only F1-score 0.931(Intent)/0.865(Slot/Entity), DIET+KoBERT F1-score 0.951(Intent)/0.901(Slot/Entity)의 결과를 획득하였다. 즉 DIET+ KoBERT의 의도 분류 및 슬롯/개체명 인식 부분 모두 더 높게 나타남을 확인하였으며, 이를 통해 DECO-LGG 기반 SSP 데이터와 DIET+KoBERT 모델의 효율성을 입증하였다.

2장에서는 딥러닝 기반 챗봇 NLU 연구 및 대화처리 데이터 구축 선행 연구에 대해 살펴보고, 3장에서 언어자원 구축 방법론과 실 데이터를 통해 나타난 질의 및 불만 화행의 패턴 문법의 양상 및 DECO-LGG 데이터 생성 모듈 구성에 대해 논의한다. 4장에서는 RASA 오픈 소스에서 제공하는 멀티태스크 트랜스포머 모델인 DIET와 그 활용 관련 사항을 서술한 후, 5장에서 본 연구 방법론의 성능을 평가하기 위해 한국어 화자가 핀테크 CS 챗봇과 대화하는 상황을 설정하여 구성한 평가용 정답지를 사용하였다. 6장에서 평가 결과에 대한 정리를 통해, 본 연구의 의의와 향후 연구에 대한 방향을 논의하였다.

2. 관련 연구

딥러닝 기술의 발전으로 CS를 포함한 다양한 목적의 챗봇 NLU를 위한 연구가 진행되었다. 초기에는 이전의 hidden state를 계속 반영하여 새로운 hidden state를 생성함으로써 문자열 시퀀스 처리에 적합한 RNN (Recurrent Neural Network) 모델과, 시퀀스가 길어지는 경우 나타나는 원거리 의존성 문제를 처리하기 위해 LSTM (Long Short-Term Memory) 모델을 같이 NLU에 활용한 연구 등이 소개된 바 있다[8]. 근래의 챗봇 NLU 연구 동향으로는, 비지도 방식으로 언어모델을 학습한 결과를 실제 응용 태스크에 fine-tuning하는 방식이 다양한 언어 처리 태스크에 좋은 성능을 보여주고 있으며, 특히 트랜스포머 모델 기반으로 pre-trained 언어모델인 BERT를 기반으로 NLU를 수행한 전이 학습(transfer learning) 모델이 활발히 연구되고 있다.

전이 학습에 기반한 End to End 방식의 연구로, 하나의 모델을 통해 Slot-filling과 의도 분류를 동시에 처리하는 Joint BERT 모델이 소개되었다[9]. Joint BERT의 경우, 의도 분류를 위해 hidden state의 첫 토큰으로 [CLS]을 사용하고 타 토큰의 hidden state를 활용하여 Slot-filling을 위한 개체명을 인식한다. 이렇게 하나의 모델로 멀티태스크를 수행하는 프레임워크는 의도 분류 정보와 슬롯에 해당하는 개체명과의 높은 연관성을 활용함으로써 베이스라인을 능가하는 괄목할 만한 성능을 보이는데, 예를 들어 “LoL 플레이 할래”라는 사용자 의도에 대해서 Play_game이라는 분류 라벨을 지정할 때, “LoL(League of Legends)”의 슬롯 지정이 Game이 될 상관관계가 언어구성상 높다는 특징을 활용한 방식이다.

이러한 End to End 방식의 멀티태스크 트랜스포머 모델의 장점에 따라, 챗봇 개발에 필요한 다양한 파이프라인을 제공하는 RASA 오픈 소스에서 DIET 모델이 소개되었다. 여기서는 의도 분류와 Slot-filling을 하나의 모델에서 처리하여 효율을 높이는 멀티태스크 트랜스포머 방식을 활용한 것과 더불어, 작업 도메인에 해당하는 양질의 학습 데이터를 사용한다면 전이학습의 BERT 모델과 달리 기학습된 모델을 필수적으로 요구하지 않고도 학습 시간 대비 뛰어난 효율성을 보일 수 있음을 입증하였다[10]. 또한 딥러닝 기반 NLU 서비스의 성능을 평가한 [11]에 따르면, DIET 소개 이전에도 RASA의 NLU 모델은 영어 질의문 처리에 있어 LUIS와 더불어 F-measure가 0.94로, 각각 0.74, 0.68의 F-measure를 보인 Watson과 API.ai의 결과보다 높은 성능을 보임으로써 그 성능이 입증된 바 있다.

RASA NLU 파이프라인을 활용한 한국어 챗봇 개발 연구로 [12]에서는 환자와 의사간의 의사소통을 증진시키는 챗봇을 개발하여 의학 상식 사전, 알람 예약, 병원 오는 길 보여주기 등의 다양한 기능을 구현하였지만, 한국어를 이해하는 챗봇 NLU 외의 기능성에 초점이 맞춰 있어, 제한된 데이터와 의도 분류에 기반한 한계점을 보인다. 이 외에 RASA의 한국어 NLU 파이프라인 최적화를 위해서 [13]에서는 Mecab 토크나이저를 사용하고, RASA에서 제공하는 다양한 NLU 모듈을 실험하여 한국어 처리에 최적화된 NLU 파이프라인 및 하이퍼 파라미터를 제시하였다.

챗봇의 대화 처리 데이터셋은 주로 목적 지향형 챗봇의 slot-filling task를 수행하기 위한 DSTC(Dialog State Tracking Challenge)와 같은 구조화된(structured) 대화 코퍼스 데이터 연구가 진행되었으며[14], Twitter로부터 추출한 비구조화된(unstructured) 대화 코퍼스 데이터 연구[15] 등이 있다. 대부분의 연구는 영어를 중심으로, 대표적인 영어의 대화 처리 데이터로는 MultiWOZ 데이터셋이 소개되었다[16]. 이는 크라우드소싱(crowdsourcing)을 이용한 Wizard-of-Oz (WOZ) 방식으로, 불특정 다수로부터 주석 처리를 수행하는 Amazon Mechanical Turk 플랫폼을 활용하여 구축되었다. MultiWOZ 데이터셋의 경우, 단일 도메인(예: bus timetable, restaurant booking 등)의 대화 코퍼스로부터 여러 개의 도메인(예: restaurant information, tourist information 등)을 연계하는 멀티도메인으로 구축되었다. 대화시스템 개발에서 대화 데이터셋의 구축량과 구축 품질은 대화시스템의 성능을 결정하는 상당히 중요한 요소로서 특히 신경망 기반 대화시스템에서는 학습용 데이터셋의 양과 질에 따라 대화의 품질이 좌우되기 때문에, 실제 사용자의 언어 사용의 전형을 담아낸 대화 데이터셋의 구축은

매우 중요한 이슈가 된다[17]. 이와 같이 챗봇 성능에 결정적 영향을 미치는 대화처리 데이터셋의 경우, 대부분 영어를 중심으로 구축되고 있으며, 개인정보 이슈에서 자유로운 NLU 주석 데이터 생성 모듈 연구나 한국어 데이터 반자동 구축 방법론에 대한 연구는 찾아보기 어려운 실정이다.

본 연구에서는 금융 분야의 한국어 질의 및 불만, 요청 화행 데이터를 분석하여, 토픽 명사구를 구성하는 32가지의 자질 유형 및 38가지의 이벤트, 그리고 10가지 담화소 패턴 유형에 대한 기술을 토대로, 전형적이며 다양한 문장 패턴을 생성하였다. 이를 통해 한국어 CS 챗봇 NLU을 위한 대용량의 주석 언어자원을 구축하였다. 이를 기반으로 최적화된 파이프라인을 사용한 End to End 방식의 멀티태스크 트랜스포머 모델을 활용하여, 한국어 NLU 처리를 위한 전반적인 프로세스를 실험하고 그 효율성을 입증하였다.

3. 한국어 질의 및 불만 화행 처리를 위한 DECO-LGG SSP 데이터 생성 모듈

3.1. 분석 데이터와 연구 방법론

핀테크 도메인에서 나타나는 질의 및 불만 화행의 유형을 분석하기 위해서는 실제 한국어 사용자가 금융 서비스에 대해 문의한 내용과 관련 서비스에 대한 평가 데이터가 필수적이다. 실제 서비스되고 있는 핀테크 CS 챗봇의 대화처리 데이터를 활용하는 것이 가장 이상적이나, 이와 같은 데이터는 기업의 중요한 자산일 뿐 아니라 개인정보 보안과 같은 문제로 공개되지 않는 것이 일반적이다. 따라서 본 연구에서는 구글 플레이스토어(Google Playstore)에서 제공하는 어플리케이션 중 한국에서 사용되는 핀테크 관련 어플리케이션 6개를 임의로 선정하였으며, 2018~2021년 사이에 작성된 사용자 리뷰를 수집하고, 별점 3점 미만의 데이터를 분석함에 따라 불만/요청 화행문의 양상을 검토하였다. 또한 질의문이 직접적으로 드러난 문형의 검토를 위해 금융 관련 키워드를 기반으로 '네이버 지식IN'에서 관련 텍스트를 수집하였다. 분석을 위해 수집된 데이터의 규모는 약 5만여 문장(약 127만 토큰)으로 구성되었다.

어플리케이션	Playstore	지식IN	합계	총계	토큰 수
토스	8,756	8,084	16,840	48,457	1,272,230
KB스타뱅킹	6,377	835	7,212		
카카오뱅크/ 카카오페이	8,246	6,957	15,203		
우리원뱅크/ 우리은행	1,409	6,401	7,809		
뱅크샐러드	1,393	-	1,393		

표 1. 핀테크 도메인의 화행 분석을 위한 데이터의 규모

챗봇 학습을 위한 언어 자원의 구축을 위해, 부분 문법 그래프(Local Grammar Graph: LGG) 프레임을 사용하였다. 텍스트에서 실현되는 언어 국지적인 현상을 그래프 형식으로 표상하는 LGG SSP 스키마는 유한상태 오토마타(Finite-State Automata: FSA) 및 트랜스듀서(Finite-State Transducer: FST)로 컴파일되어 코퍼스에서 나타나는 핵심 패턴을 포착하고 주석할 수 있으며, 동시에 이러한 패턴에 기반한 새로운 텍스트 생성에도 활용될 수 있는 반자동 언어데이터 증강 모듈이다. LGG의 실질적인 구현과 적용을 위해 본 연구에서는 Unitex[18] 플랫폼을 사용하였다. LGG SSP 모듈은 한국어의 어휘, 통사, 의미 정보 및 극성 정보를 처리하는 DECO 기계가독형 한국어 전자사전을 기반으로 작동한다. FSA/FST 문법으로 컴파일된 LGG는 순환 전이망(recursive transition network) 문법 형태로 구성된 DECO 기계가독형 전자사전이 제공하는 언어 정보를 기반으로 적형에 해당되는 패턴을 코퍼스에서 인식하거나, LGG 문법을 기반으로 연구자가 형식화한 텍스트를 생성할 수 있다.

3.2. 질의 및 불만 화행 데이터 생성 DECO-LGG SSP 모듈

3.2.1. 토픽(TOPIC)을 구성하는 대상/자질 의미클라스

실제 질의 및 불만 화행 데이터의 언어 구조를 언어학적 기법으로 분석해 보면, 표 2와 같이 "토픽(Topic) - 이벤트(Event) - 담화소(Discourse)" 유형으로 구조화할 수 있다.

유형	Topic	Event	Discourse
질의화행	적금 통장	개설하려는데	어떻게 해야 하나요?
불만화행	어플	설치	계속 오류 나서 짜증남

표 2. 질의 및 불만 화행문의 의미적 모듈 구성

토픽(topic)을 나타내는 대상/자질 어휘와 사용자의 의도 및 문장 내 실질적인 서술성을 표현하는 이벤트(event) 어휘는 도메인의 특징에 따라 상호 제약관계가 나타난다. 토픽을 구성하는 대상어(entity)는 '우리WON뱅킹/토스/카카오뱅크'와 같은 기관명과 '햇살론/저금통' 등과 같은 상품 및 서비스명으로 구성된다면, 토픽의 자질어(feature)는 일련의 일반명사구 및 전문용어 유형으로 구성된다. 이들의 일부 예를 보이면 다음과 같다.

번호	정규화	레이블	예시
1	예금	DEPOSIT	예금/신규적금
2	대출	LOAN	대출/담보대출
3	청약	SUBSCRIPTION	청약/주택청약
4	외환	FOREX	외환
5	펀드	FUND	펀드/신탁투자
6	증권	STOCK	증권/주식
7	보험	INSURANCE	보험/보험상품
8	세금	TAX	세금/세액
9	연금	PENSION	연금/연금상품
10	계좌	BANK_ACCOUNT	계좌/통장/결제통장
11	이자	INTEREST	이자/이자율/이율

표 3. 토픽의 자질어(FEATURE) 분류의 일부 예시

토픽을 구성하는 어휘 표현은 사용자 '의도'의 대상(target) 및 그 자질(feature) 부류를 의미하는데, 이는 이러한 의도 이벤트 술어의 명사구 논항적 성질을 띠는 어휘 유형이다. 표 3의 자질분류 카테고리는 모두 32가지로 분류되며, 이때 이들은 다시 상·하위어 분류 체계(taxonomy)로 구조화될 수 있다. 위의 용어들은 각 카테고리별 레이블과 함께 모두 LGG 패턴 문법 형식으로 자원화된다.

3.2.2. 의도(INTENT)를 구성하는 이벤트술어 의미클라스

이벤트는 사용자가 특정 대상에 대한 실질적인 의도를 표

현하고자 하는 핵심적인 의미를 담고 있는 서술성 성분으로, “개설”과 같이 “하다/되다” 등의 접사와 결합하는 서술성 명사와 “만들다” 등의 단일 용언으로 구성된다. 본 연구에서는 수집된 데이터에서 추출된 고빈도 서술어 성분을 바탕으로 다음과 같이 하위분류가 수행되었다.

번호	상위 범주		핀테크 이벤트 유형
1	ON(생성/시작/접근)		신청, 가입, 설정, 연동, 설치, 다운로드, 발생, 업데이트, 업로드, 발급, 로그인, 개설, 승인
2	PROCESS (변경/과정/처리)	UP(상승)	상향, 증가
3		DOWN(하강)	감축, 감면
4		IN(수신)	입금, 매수
5		OUT(발신)	송금, 매도, [선/제]결제, 상환, 인출
6		UTILIZE(활용)	이용, 투자운용
7		REPLACE(대체)	변경, 환전, 환불
8		CHECK(확인)	조회, 표시, 인증
9	OFF(소멸/끝/분리)		해지, 수신거부, 탈퇴, 삭제, 차단

표 4. 이벤트(Event) 카테고리

상위 범주는 생성/시작/접근(ON) (예: ‘만들다’ 류), 변경/과정/처리(PROCESS) (예: ‘바꾸다’ 류), 소멸/끝/분리(OFF) (예: ‘없애다’ 류)와 같이 시간적 프로세스 양상을 기반으로 분류되었다. 이와 같은 범주 유형화는 연관된 유사 어휘를 확장하고 형식화하는데 유용하게 사용되며, 이외 도메인에도 확장 적용될 수 있다. 예를 들어 미디어(Media)와 같은 콘텐츠 관련 도메인에서는 생성/시작(ON)의 범주로 노래/영상 등을 시작하는 의미의 “재생” 및 “틀어” 등과 같은 어휘가 포함되는 형식으로 구조화된다.

핀테크 이벤트 카테고리는 총 9개의 중간 범주와 38개의 하위 범주로 구성되어 그림 1과 같은 패턴 문법으로 기술된다.

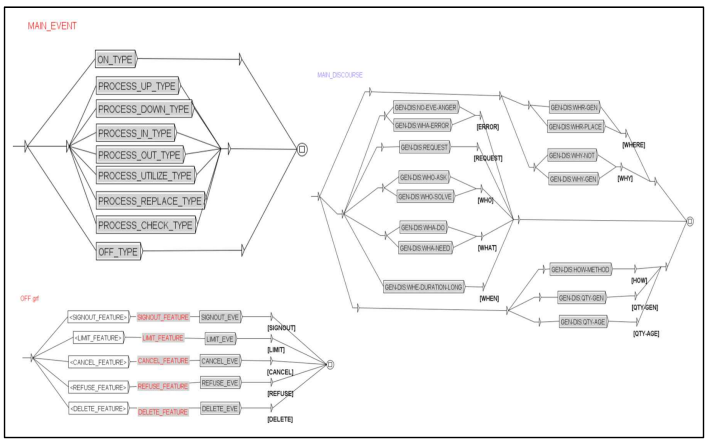

그림 1. 이벤트 LGG 모듈의 일부 예

위의 패턴 문법에서는 서술성 명사 결합형 또는 단일 용언 구성형에 대해서, 결합 가능한 토픽 어휘 표현과의 연동 관계에 따라 각 이벤트의 서술절이 인식 및 생성될 수 있도록, 총 961개의 표현 경로가 모듈화되었다.

3.2.3. 담화소(DISCOURSE)를 구성하는 의미클라스

담화소의 경우, 이벤트 어휘로 전달되는 문장의 핵심적 의미에서 화행 및 담화적 역할을 구체화하는 언어적 구성으로, 특정 도메인에 제한되는 형태가 아닌 범도메인적 특징을 지닌다. 담화소는 표 5와 같이 명령문(Imperative), 평서문(Declarative), 의문문(Interrogative)의 문형 구조 내에 실현될 수 있으며, 요청 담화적 관점에서는, ‘처리요청’, ‘정보요청’, 그리고 ‘인식요청’으로 분류될 수 있다.

문형/요청	처리 요청	정보 요청	인식 요청
명령문 (Imperative)	A계좌로 현금 이체해	현금 이체하는 방법 알려 줘	오늘 이체한 거 기억해 둬
평서문 (Declarative)	A계좌로 현금 이체하고 싶어	현금 이체하는 방법이 궁금해	오늘 분명히 이체했다.
의문문 (Interrogative)	A계좌로 현금을 이체해 주겠어?	현금 이체 어떻게 하지?	오늘 이체한 거 기억할 수 있지?

표 5. 문형과 요청화행에 따른 분류

표 5의 예는 화행의 구조가 특정 문형에 한정되지 않고 다양하게 실현될 수 있다는 사실을 보여주는데, 일례로 ‘정보요청’ 화행의 경우, 의문문의 문형 구조뿐만 아니라 평서문과 명령문 형태로도 실현될 수 있음을 확인할 수 있다. 본 연구에서는 3가지 문형으로 실현되는 질문 및 불만 요청 화행의 입력문 NLU 처리를 위해, 표 6과 같이 의문사를 포함한 전형적인 질의문(6WH, QTY) 담화소 구성과, 언표내적으로 직접 요청 의도를 표현하거나 불만 감정을 드러내는 것을 통해 요청 의도를 간접 화행으로 표현하는 불만 화행 구성(ERROR/REQUEST) 등 모두 10가지 유형을 LGG 패턴 문법으로 기술하여 모듈화하였다.

번호	상위유형	하위유형	명칭	예시
1	6WH	WHO	누구	누구?/이름이 뭐?
2		WHE	언제	언제?/어느 시간?
3		WHR	어디서	어디?/어떤 주소?
4		WHA	무엇	무엇?/어떤거?
5		WHY	왜	왜?/어떤 이유로?
6		HOW	어떻게	어떻게?/어떤 방법?
7	QTY	QTY-AGE	나이	몇-살(세)?
8		QTY-GEN	수수료/금리	얼마?/몇-N?
9	ERR	ERR/ANG	오류/불만	오류 나/짜증나
10	REQ	REQUEST	요청/희망	-하고 싶어

표 6. 담화소(Discourse) 카테고리

LGG 패턴 문법으로 형식화된 담화소 표현 데이터는, 단순히 의문사 키워드 등을 기반으로 질의문을 처리하는 기제와는 다르다. 가령 “(이벤트 당첨금 수령을 직접 하고 싶은데) *어느 주소로 가야 하는지 알려 주세요*”와 같은 경우는, 표면적 의문사의 활용과 다른 기능을 하는 다단어 의문 표현이 실현되었으나, [WHERE]의 의미 기능을 수행하고 있다. 반면 “(공인인증서 오류라니) *정신머리를 어디다 두고 앱을 만든거냐?!*”와 같은 경우는, 화자가 관용어 표현을 통해 불만 화행 [ANGER]을 표현하는 형태의 예를 보인다.

본 연구에서 구현한 LGG 패턴 문법은 이상과 같은 담화소 유형들을 포함하여 전체 40,384개 유형의 담화소 시퀀스를 인식/생성 가능하게 한다. 이와 같은 방식으로 ‘토픽-이벤트-담화소’ 구성 시퀀스를 주석 처리하면, 생성 가능한 데이터 규모는 총 60,927,364,097개에 이르게 된다. 이때 모델을 학습하는 시간 및 하드웨어의 성능을 고려하여 챗봇 상황에서 나타나는 경제성에 입각한 언어 활용에 따라 비격식체, 지시문, 직접 화행 표현 등에 가중치를 적용하여, 학습에 최적화된 서브 데이터셋을 생성할 수 있도록 하였다.

4. RASA DIET 시스템

DIET(Dual Intent and Entity Transformer)는 복합 챗봇 프레임워크 API인 RASA에서 제공하는 NLU 아키텍처로, 의도 분류와 Slot-filling을 위한 개체명 인식을 종합적으로 수행할 수 있는 End to End 방식의 멀티태스크 트랜스포머 모델이다. 기학습된 언어 모델 없이도 성능과 학습 속도 면에서 효율성을 보이며[10], 개발 목적에 따라 GloVe 및 BERT와 같은 기학습된 단어 임베딩을 사용하여 다양한 요소들을 활용할 수 있도록 유연성 높은 파이프라인이 제공된다.

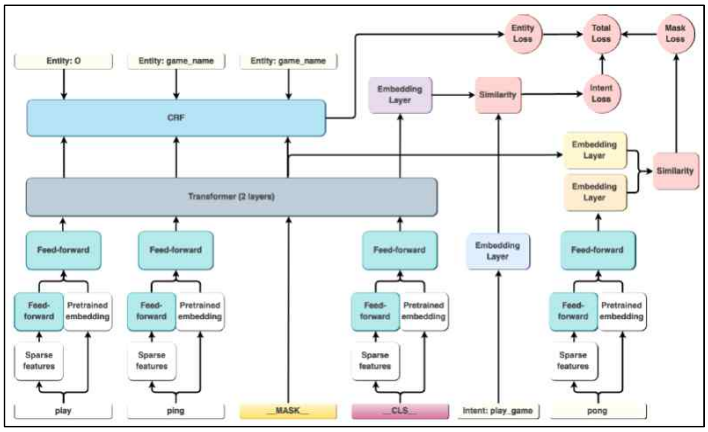

그림 2. DIET 모델 구조

DIET 모델의 경우 그림 2를 통해 설명된 구조와 같이 기학습된 임베딩 또는 토큰의 희소 피처를 피드포워드 신경망 구조(Feed-Forward Neural Network: FFNN)를 거친 n개의 레이어 트랜스포머 구조를 통해 개별 토큰의 임베딩을 계산하고, 이를 Conditional Random Field(CRF)를 통해 개체명 인식 및 손실 함수를 계산하여 학습한다. 또한 동시에 의도 분류를 위한 특수 클래스 토큰(__CLS__)을 개별 토큰과 같은 경로를 통해 전체 입력 문장을 요약하는 임베딩을 도출하여, 이를 정답과 비교하는 유사성 함수 처리를 통한 의도 분류 손실을 계산한다. 특히 DIET 모델의 FFNN은 중퇴율(Dropout rate)이 설정되어 있는 가벼운 학습 모델 구조로 학습 효율성을 증진시킨다.

또한 "__MASK__" 토큰을 사용한 마스킹을 통해 트랜스포머 구조 언어 모델을 학습하는데, 최종적으로 마스크/개체명/의도 손실값을 통합한 Total Loss를 산출하여 모델을 학습하는 형식으로써 의도 분류 및 개체명 인식을 종합적으로 처리한다. 이와 같은 멀티태스크 모델의 경우, 의도와 개체명 정보에 큰 상관관계가 있기 때문에 개체명 인식에 있어 개별적 개체명 인식 모델보다 더 높은 성능을 보인다.

DECO-LGG를 통해 생성된 마크다운(Markdown) 주석 처리된 학습 데이터를 저장하면 RASA API에서 그림 3과 같은 JSON 형식으로 변환되어 처리된다.

이때 Slot-filling을 정확하게 수행하기 위한 한국어 토크나이징이 선행되어야 하는데, 본 연구에서는 한국어 처리를 위해 Python 모듈 KoNLPy에서 제공하는 오픈소스 한국어 처리기의 토크나이저 Otk를 사용하였다. 또한 GeForce RTX 3090 (24GB), RAM 32GB, Intel Core i7-8700K CPU 환경에서 0.001 학습률의 Adam optimizer[10], 500 Epoch, 4-레이어 트랜스포머 구성 등 [13]에서 입증한 표 7과 같은 최적화 파이프라인을 참고하여 학습을 진행하였다.

```
{
  "text": "주식 매수하려는데 오류 뜨네요",
  "intent": "ERROR_BUY",
  "entities": [
    {
      "start": 0,
      "end": 2,
      "value": "주식",
      "entity": "STOCK"
    }
  ]
},
```

그림 3. JSON 형식의 데이터 처리

Parameter	DIET-Base	DIET-Opt
Epoch	300	500
Transformer layers	2	4
Transformer size	256	256
Weight sparsity	0.8	0.7
Embedding dimension	20	30
Hidden layer size	[256, 128]	[512, 128]

표 7. DIET-Opt 일부

5. 실험

5.1. 실험 자원

본 연구에서는 DECO-LGG 기반 SSP 생성 데이터를 활용하여 구현된 멀티태스크 트랜스포머 모델의 성능 평가를 위하여, 현재 구축되어 있는 대규모 주석데이터에서 일부 서브셋을 추출하여 실험용 학습데이터로 구성하였다.

이를 위해 이벤트-담화 구성의 의도 유형 중 실 데이터에서 가장 사용 빈도가 높은 107개 의도에 해당하는 쿼리 총 32,100개를 추출하였다. 이를 1)최적화된 파이프라인 DIET 모델과 2)기학습된 KoBERT[1] 임베딩을 활용하는 DIET+KoBERT로 각각 학습하고, 해당 의도의 챗봇 상황을 기반으로 한국어 화자들이 작성한 슬롯 유형 38개(총 1,047개)를 포함한 챗봇 입력문 800개(총 토큰 4,024개) 규모의 챗봇 NLU 테스트 데이터로 성능을 평가하였다. 관련 데이터 및 코드는 [2]DICORA 홈페이지를 통해 공개하였다.

5.2. 성능 결과

본 연구에서 사용한 실 채팅 테스트 데이터의 의도/슬롯별 데이터 수는 일정하지 않기 때문에, 데이터 비율을 고려한 Weighted avg를 활용하여 성능을 측정하였다. 실험 결과 표 8과 같이 DIET-only 모델보다 DIET+KoBERT 모델이 의도 분류 및 슬롯/개체명 인식에서 F1-score에서 모두 높은 성능을 보여주는 것을 확인할 수 있다.

1) https://huggingface.co/monologg/KoBERT
2) http://dicora.kr/?page_id=3918

Type	Metric	Intent	Slot(Entity/Feature)
DIET-only	F1	0.931	0.865
	Recall	0.924	0.789
	Precision	0.939	0.957
DIET+KoBERT	F1	**0.951**	**0.901**
	Recall	**0.944**	**0.843**
	Precision	**0.958**	**0.968**

표 8. 성능 평가표

이를 통해 기학습된 KoBERT 임베딩을 기반으로 DIET 모델을 활용하면, 특히 슬롯/개체명 인식 태스크에 더욱 효과적이라는 사실과 함께, DECO-LGG 기반 SSP 생성 데이터셋이 모델 학습에 효율적인 양질의 데이터셋으로 기능한다는 것을 확인할 수 있다.

6. 결론

본 연구는 DICORA-LINITO 프로젝트의 핀테크 CS 챗봇 LIFI[19] 구현을 위한 대규모 학습데이터셋을 구축하는 연구의 일환으로 진행되었다.

본 연구에서는 핀테크 CS 챗봇의 NLU 모듈 학습을 위한 주석데이터 생성을 위해, [토픽(Topic) - 이벤트(Event) - 담화소(Discourse)]의 모듈별 구성의 DECO-LGG 언어자원에 기반하여 SSP 방법으로 학습 데이터를 증강하는 방법론을 제시하였다. 이렇게 생성된 학습데이터를 통해, 챗봇 사용자의 질의 및 불만, 요청 화행의 의도를 이해하는 NLU 모델의 학습이 가능하게 된다. 본 연구에서 생성한 학습데이터는 다양하면서도 전형적인 챗봇 NLU 데이터의 특성을 가지기 때문에, End to End 방식의 효과적인 멀티태스크 트랜스포머 모델 DIET 아키텍처를 학습시키는 데에 최적화되어 있음을 확인할 수 있다.

본 연구의 결과를 통해, 숙련된 소수 전문 인력으로, 보다 효율적으로 양질의 데이터를 생성함으로써, 개인정보 보호와 같은 이슈로부터 자유로우며, 동시에 크라우드소싱 방식의 불특정 다수에 의존하는 수동식 챗봇 데이터 구축 방식의 고비용·저효율의 대안으로 큰 효율을 보일 수 있다는 점을 확인할 수 있다. 이에 따라 본 연구는 향후 더 다양한 도메인으로 확장되어, NLG를 포함한 통합적 대화처리 데이터셋 구축 및 복합적 대화처리 모델 연구를 위한 중요한 토대가 될 수 있을 것으로 기대된다.

참고문헌

[1] 남지순, 코퍼스 분석을 위한 한국어 전자사전 구축방법론. 도서출판 역락. 2018.

[2] Gross, M.. The Construction of Local Grammars. Finite-State Language Processing, The MIT Press, 1997.

[3] 남지순, 금융앱 챗봇의 사용자 의도분석 모델 개발을 위한 학습데이터 이중 증강 방법. DICORA-TR-2021-12. 디코라연구센터. 한국외국어대학교. 2021.

[4] 정천수, 정지환, 포스트 코로나19 언택트 시대 대응을 위한 AI 챗봇 구축방법에 관한 연구, 한국IT서비스학회지, 19(4), 31-47, Aug. 2020.

[5] Shawar, B. A., Atwell E., Using corpora in machine-learning chatbot systems, International journal of corpus linguistics, 10(3), pp. 489-516, 2014.

[6] 이한얼, 쓸수록 진화하는 '챗봇', 3년 후 10兆 시장 간다, 지디넷코리아, 2021.

[7] 최수정, 성목경, 최진주, 박준성, 룰 기반 챗봇을 활용한 교통정보 서비스 개발, 한국정보과학회 학술발표논문집, pp. 1868-1869, 2017.

[8] Ravuri S., Stolcke A., Recurrent Neural Network and LSTM Models for Lexical Utterance Classification, in Proc. Interspeech , Dresden, Germana, Sept., pp. 135-139, 2015.

[9] Chen Q., Zhuo Z., Wang W., BERT for joint intent classification and slot filling. CoRR, abs/1902.10909, 2019.

[10] Bunk T., Varshneya D., Vlasov V., Nichol A., Diet: Lightweight language understanding for dialogue systems, arXiv preprint arXiv:2004.09936, 2020.

[11] Braun D., Mendez A. M., Matthes F., Langen M., Evaluating Natural Language Understanding Services for Conversational Question Answering Systems, Proceedings of the 18th Annual SIGdial Meeting on Discourse and DialogueAt: Saarbrücken, Germany, 2017.

[12] 류연준, 박세리, 성현규, 이진수, 김웅섭, 인공지능 챗봇을 기반으로 한 환자-의사 소통 증진 소프트웨어, 한국정보처리학회 학술대회논문집, 27(2), pp. 501-504, 2020.

[13] Hwang M., Shin J., Seo H., Im J., Cho H., KoRASA: Pipeline Optimization for Open-Source Korean Natural Language Understanding Framework Based on Deep Learning, Mobile Information Systems, vol. 2021, 9 pages, 2021.

[14] Williams J., Raux A., Ramachandran D., Black A., The dialog state tracking challenge, in Proceedings of the SIGDIAL 2013 Conference, pp. 404-413, 2013.

[15] Lowe R., Pow N., Serban I., Pineau J., The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems, in SIGDIAL, Prague, Czech Republic, Spet. pp. 285-294, 2015.

[16] Budzianowski P., Wen T., Tseng B., Casanueva I., Ultes S., Ramadan O., Gašić M., Multiwoz-a Largescale Multi-domain Wizard-of-Oz Dataset for Taskoriented Dialogue Modelling, in Proc. Conf. Empirical Methods Natural Language Process, Belgium, pp. 5016-5026, 2018.

[17] 권오욱, 홍택규, 황금하, 노윤형, 최승권, 김화연, 김영길, 이윤근, 심층 신경망 기반 대화처리 기술 동향, Electronics and Telecommunications Trends, 2019.

[18] Paumier, S.. Unitex Users' Manual. UPEM, 2003.

[19] http://linito.kr.