

RNN 문장 임베딩과 ELM 알고리즘을 이용한 금융 도메인

고객상담 대화 도메인 및 화행분류 방법

오교중^o, 박찬용, 이동건, 임채균, 최호진

한국과학기술원, 전산학부

{aomaru, pparty, hagg30, rayote, hojinc}@kaist.ac.kr

RNN Sentence Embedding and ELM Algorithm Based Domain and Dialogue

Acts Classification for Customer Counseling in Finance Domain

Kyo-Joong Oh^o, Chanyong Park, DongKun Lee, Chae-Gyun Lim, Ho-Jin Choi

KAIST, School of Computing

요약

최근 은행, 보험회사 등 핀테크 관련 업체에서는 챗봇과 같은 인공지능 대화 시스템을 고객상담 업무에 도입하고 있다. 본 논문에서는 금융 도메인을 위한 고객상담 챗봇을 구현하기 위하여, 자연어 이해 기술 중 하나인 고객상담 대화의 도메인 및 화행분류 방법을 제시한다. 이 기술을 통해 자연어로 이루어지는 상담내용을 이해하고 적합한 응답을 해줄 수 있는 기술을 개발할 수 있다. TF-IDF, LDA, 문장 임베딩 등 대화 문장에 대한 자질을 추출하고, 추출된 자질을 Extreme learning machine(ELM)을 통해 도메인 및 화행 분류 모델을 학습한다.

주제어: 대화 시스템, 자연어 이해, 문장 임베딩, 화행 분류

1. 서론

올해 들어 국내 은행, 증권, 신용평가 등 금융 관련 도메인에서는 비대면 상담 서비스를 위한 챗봇 시스템의 도입이 활발히 일어나고 있다. 챗봇은 시간과 장소에 구애 받지 않고 고객상담 서비스를 제공할 수 있다. 이를 통해 콜센터 내 야간 및 주말 고객상담 인력을 줄일 수 있으며, 단순 반복 질문에는 자동으로 대처하고, 상담사는 보다 복잡한 상담 사례에 집중할 수 있다. 현재 하나은행의 핀크, 우리은행의 위비톡 등 주요 은행을 시작으로, 카카오뱅크의 챗봇, K뱅크의 핑거뱅크 등 인터넷 전문은행에서도 챗봇의 도입이 가속화 되고 있다.

현재까지 개발된 고객상담 챗봇 시스템은, 메뉴 기반으로 답변 가능한 질문을 객관식으로 제공하거나, 사전에 정의된 시나리오에 기반하여 준비된 질문을 하고 사용자의 답변과 유사한 시나리오의 응답을 제공하고 있다. 이 같은 서비스 형태는 모바일뱅킹의 새로운 인터페이스 방식일 뿐이며, 상품 소개와 정보 조회 등의 단순한 서비스만 제공 가능하다. 지속적으로 고도의 자연어처리 기술과 기계학습 기반의 인공지능 기술을 통하여 사람처럼 대화를 하는 대화 시스템의 연구와 개발이 필요하다.

이 같은 인공지능 기술에 기반한 대화 시스템을 개발하기 위해서는, 전화 등 음성 대화를 문자형 데이터로 바꿔주는 STT 기능, 기존의 자연어처리(NLP), 기계학습 기술에 기반한 상담 대화의 자연어 이해(NLU) 기술, 응답 내용을 결정하고, 자연어 대화를 생성(NLG)하는 응답 기술 등 높은 수준의 대화 기술이 필요하다.

본 논문에서는 대화형 시스템을 위한 자연어 이해 기술에 초점을 맞춘다. 그 중에서 비정형 정보에 해당하는 고객상담 대화의 도메인 및 화행 분류 기술에 대해 다룬다. 도메인 및 화행 분류는 질의에 정확하게 응답하기 위해서 수행되어야 하는 기술이다.

본 논문에서 제안하는 도메인 및 화행 분류 정보는 금융 도메인 고객상담 서비스에 특화된 자연어 대화 문장에 대한 비정형 정보이며 기계학습 기반의 접근 방법을 적용하여 분석한다. 전처리 과정으로써 고객상담 대화 말뭉치를 워드임베딩(Word2Vec) 모델을 통해 형태소 수준의 언어 모델을 학습하고, TF-IDF로 추출된 주요 키워드, LDA모델을 통해 얻은 분류 카테고리 자질을 입력 정보로 ELM 알고리즘을 사용한 분류 모델을 학습하여 입력된 상담 대화의 도메인과 화행을 분류한다.

2. 관련 연구

챗봇이란, 텍스트나 음성 등을 이용하여 인간과 대화를 수행하는 컴퓨터 프로그램을 말하며, 기존의 토크봇(Talbots), 채터봇(chatterbot), IM봇, 대화형 에이전트(interactive agent), 인공 대화 개체(artificial conversational entity) 등을 지칭한다. 챗봇 기술은 기존의 대화 시스템(dialog system) 응용에서 고객 상담, 정보 획득과 같은 실증적인 문제를 풀기 위해서 시작되었으며, 의미, 화행, 화용 분석 등 복잡한 자연어 처리와 인공지능 기술을 필요로 한다.

최근 온라인 쇼핑몰과 유통 업계에서 대화형 시스템을 도입하여 사용자에게 맞춤형 상품을 추천하거나[1], 상담원을 대신하여 빈번한 상담에 대한 답변을 해주는 대화형 시스템[2]을 개발한 사례가 있다. 그 외에 법률 상담[3]과 심리치료와 같은 헬스케어 분야[4] 등 다양한 산업 분야에서도 챗봇 기술을 도입한 응용 시스템의 연구 개발 사례가 시도되고 있다.

금융 분야에서도 챗봇과 관련된 연구 및 개발이 이루어지고 있다. [5]에서는 बैं킹 시스템을 위한 인공지능 챗봇을 최초로 제시하였으며, [6]에서는 बैं킹서비스를 돕기 위하여, 자연어 이해 및 생성 등 챗봇 시스템의 여러 접근 방법을 정리하였다. [7]에서는 고도의 자연어 이해 기술을 적용하여 담화 구조를 가진 대화를 관리하고, 이를 챗봇 서비스에 적용한 연구 사례가 발표되었다. 이처럼 자연어 이해 기술은 챗봇을 구현하는데 있어 반드시 필요한 기술이다.

본 논문은 챗봇의 응답 모델을 구현하기 위한 초기 단계로, 자연어 이해 기술 중 하나인 대화의 도메인과 화행을 분석하는 방법에 관해 기술한다. 본 연구팀은 기존 연구 [8]에서 tripadvisor.com의 여행지에 관한 사용자 리뷰를 바탕으로 여행 의도를 분석하는 방법을 연구하였으며, 이를 기반으로 여행지를 추천하는 기술을 개발하였다. 본 연구에서는 기존에 연구한 의도 분류 방법을 금융 도메인의 자연어 대화 데이터에 적용하여, 대화 도메인 뿐만 아니라 화행까지 분류함으로써 챗봇의 응답 성능을 향상시키는데 기여한다.

본 연구에서는 전처리 기술로 Word2Vec[9] 알고리즘을 적용한 워드임베딩 기술을 사용한다. 은행 고객상담 대화 말뭉치로부터 신경망을 통해 언어 모델을 학습하며, 단어나 구를 원하는 크기의 벡터로 표현할 수 있다. 고객상담 말뭉치에 사용된 어휘들의 언어 모델을 도메인에 특화하여 학습하고, TF-IDF로부터 추출된 중요 키워드 자질과 입력 문장을 벡터 형태의 학습 자질로 바꾸는데 사용되었다.

이렇게 얻어진 벡터 형태의 학습 자질을 이용하여 지도 학습 방법의 분류 모델을 구현한다. 본 논문에서는 학습기반의 분류 알고리즘으로 ELM 알고리즘 [10]을 사용한다. 얇은 신경망 혹은 적은 수의 은닉 층을 가지는 신경망을 이용하여 출력 층과 은닉 층 사이의 가중치는 한번만 연산함으로써, 신경망 모델의 학습 속도가 느린 단점을 보완한 분류 모델로, 최근 여러 연구에서 SVM과 비슷하거나 더 좋은 분류 결과를 보이는 것으로 보고되고 있다.

3. 고객상담 대화 화행분류

본 논문에서는 금융 도메인 고객상담 대화의 도메인과 화행 분류 방법을 제안한다. 크게 전처리, 자질 추출, 도메인 및 화행분류 3개의 단계로 이루어진다. 그림 1은 고객상담 대화에 대한 도메인 및 화행 분류 방법에 대한 개념도이다. 도메인 분류는 대분류와 세부 분류 2단계로 나뉘지며, 두 분류 결과에 따라 화행이 분류된다.

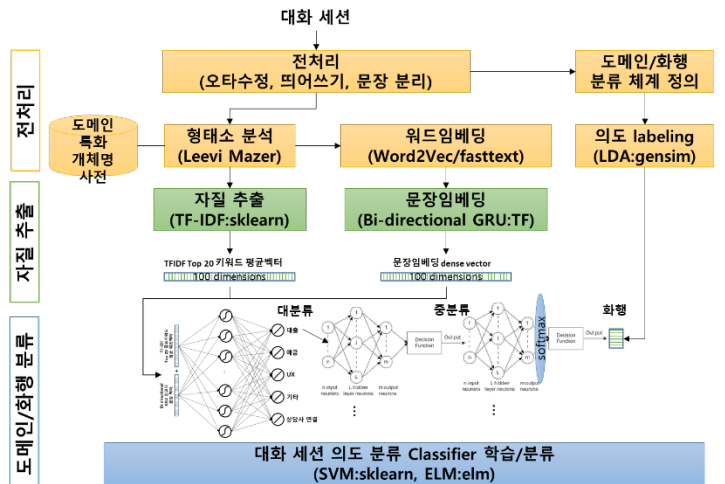


그림 1 고객상담 대화 화행 분류 방법

3.1 전처리 과정

이 단계에서는 고객상담 대화 데이터의 내용을 비식별화 및 정제하고, 언어모델 학습을 위한 말뭉치 구축과 형태소 분석 및 개체명 사전 구축 등, 자질 추출을 위한 전처리 과정이다. 고객상담 대화 말뭉치는 메신저나 대화 플랫폼에서 이루어지는 자연어 문장으로 이루어져 있으며 오타, 띄어쓰기 오류, 개인 정보가 포함된 대화가 많이 포함된다. 또한 문장을 문장 부호 기준으로 간단하게 분리하기 어려우며, 한 문장이 여러 대화에 걸쳐 쪼개지기도 한다. 따라서 연속된 대화를 처리하고 문맥의 흐름에 따라 대화를 분리해야 한다.

그림 2는 고객상담 대화 데이터의 예시이다. 그림2와 같이 연속된 대화에 대해서는 문장을 결합하고 문맥 상 질문이 변경되었을 때는 대화 세션을 분리하여 질의-답변 말뭉치를 구축하여야 한다.

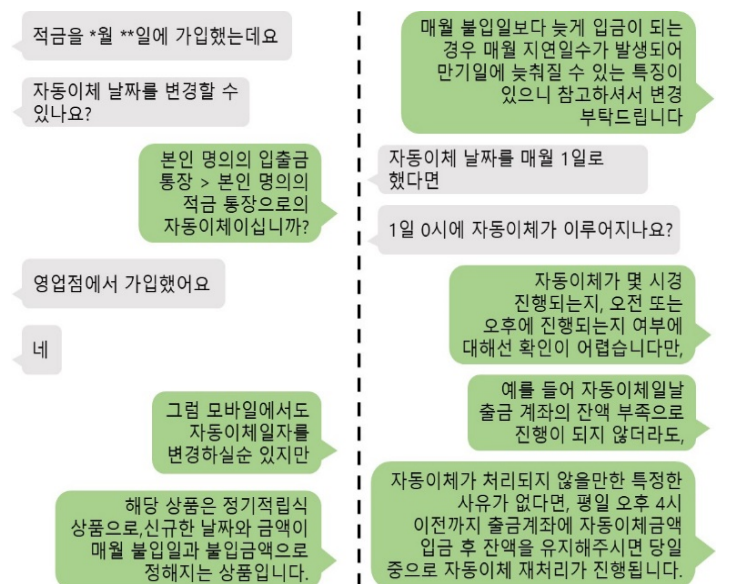


그림 2 금융 도메인 고객상담 대화 예시

다음 단계에서 대화 문장의 벡터 화와 도메인 및 화행 분류를 위해서는 학습 자질로 정해진 크기의 실수 벡터를 사용해야 한다. 이를 위해 Word2Vec 모델을 통한 언어모델 학습이 전처리 과정으로 필요하며, 본 연구에서는 형태소 분석을 통한 형태소 수준의 워드임베딩 모델 [11]을 생성하였다. 워드임베딩 학습에 이용된 말뭉치는 그림2와 같은 고객상담 대화와 한글 위키피디아의 분류: 금융 관련 문서에 등장한 한글 30만 문장이 사용되었다.

3.2 도메인 및 화행 분류 체계 정의

이 단계는 지도 학습을 위한 학습 데이터를 생성하는 단계이다. 도메인 및 화행 분류 카테고리를 정의하고 학습 기반의 분류 모델을 위한 도메인 및 화행 레이블링 데이터를 구축한다. 이를 위해 우선 도메인 및 화행의 분류 체계를 정의 해야한다.

본 연구팀은 Latent Dirichlet allocation(LDA) 모델을 사용하여 데이터 기반으로 클러스터를 정의하였다. 기존의 고객센터에서 사용하던 상담 분류 체계에 기초하여, 4가지 대 분류(대출, 예금, UX, 기타)에 대하여 각 대 분류 별 7~15개의 중분류로 이루어진 초기 분류체계를 가지고, 대 분류로 나누어진 대화 데이터를 LDA 클러스터링 (클래스 개수 40개)을 통해 초기 분류 체계와 매핑 또는 세분화하는 방법을 거쳐 최종적으로 도메인 세부 분류와 화행 수준의 구분까지 수행하였다. 이 방법은 도메인 전문가의 개입없이 고객센터 서비스 제공자의 입장에서 분류 체계를 구축하는 방법으로, 데이터만 있으면 비전문가도 도메인 분류와 화행 분류 체계를 정의할 수 있다.

그림 3은 실제로 수행한 대출에 해당하는 대분류의 세부 도메인과 화행 분류 체계를 보여준다. 붉은색으로 표시한 화행은 특정 세부 분류에 특화된 화행을 나타낸다.

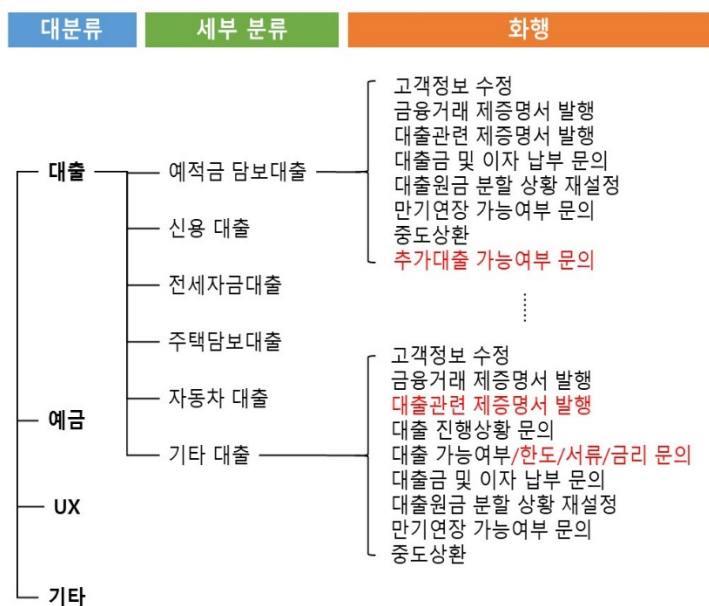


그림 3 도메인 및 화행 분류 체계 및 대출 분야 예시

3.3 학습 자질 추출 및 문장 임베딩

본 논문에서는 화행분류를 위한 입력 자질로 TF-IDF, GRU 문장인코딩 결과를 사용한다.

TF-IDF(Term frequency and Inverse Document Frequency)는 분석하려는 대화 문장이나 문서에서 다른 대화 세션에 비해 중요도가 큰 키워드를 추출한다. 이를 통해 중요 키워드 자질을 화행분류 학습에 반영할 수 있다. 추출된 Top-20 중요 키워드의 워드임베딩 평균 벡터를 입력 벡터의 일부로 사용한다. 추가로 문장에 사용된 주요 개체명의 워드 벡터도 TF-IDF 자질과 함께 반영한다. 주요 개체명은 응용 도메인과 서비스에 맞추어 별도로 제작된 बैं킹 서비스를 위한 개체명 사전으로, 예금 또는 대출 상품명 등 직접적으로 대화의 도메인을 반영하는 정보이다.

마지막 RNN 문장 인코딩 자질은 문장의 의미적 자질을 화행 분석에 사용하기 위해 사용된다. 그림 4는 문장 임베딩에 사용한 RNN 인코더 모델을 나타낸다. Bidirectional GRU dynamic RNN 모델을 사용하였으며, 입력층에는 각 단어의 워드임베딩 결과가 순차적으로 입력된다. 최종적으로 Forward layer와 backward layer의 마지막 GRU 셀의 RNN state을 이어 붙인 dense 벡터를(200차원) 문장 임베딩 결과로 사용한다.

RNN 문장 인코더의 학습 방법은 그림 5와 같다. Bidirectional GRU로 이루어진 sequence to sequence 모델을 그림 5와 같이 같은 문장으로 학습한다. 고객 상담 대화 말뭉치의 모든 문장에 대하여 loss가 수렴할 때까지 반복 학습한 후, 인코더 부분만 분리하여 새로운 입력 문장에 대해 문장 벡터를 추출한다.

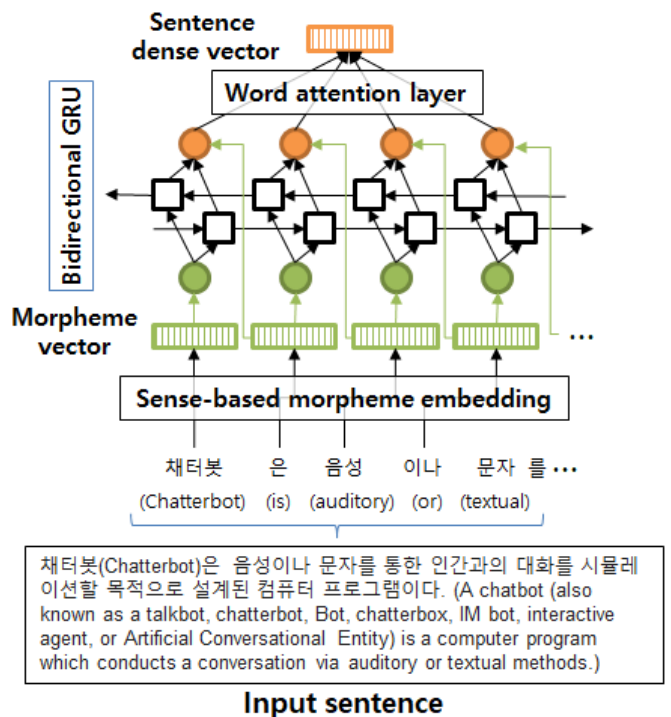


그림 4 문장 임베딩을 위한 RNN 인코더 개념도

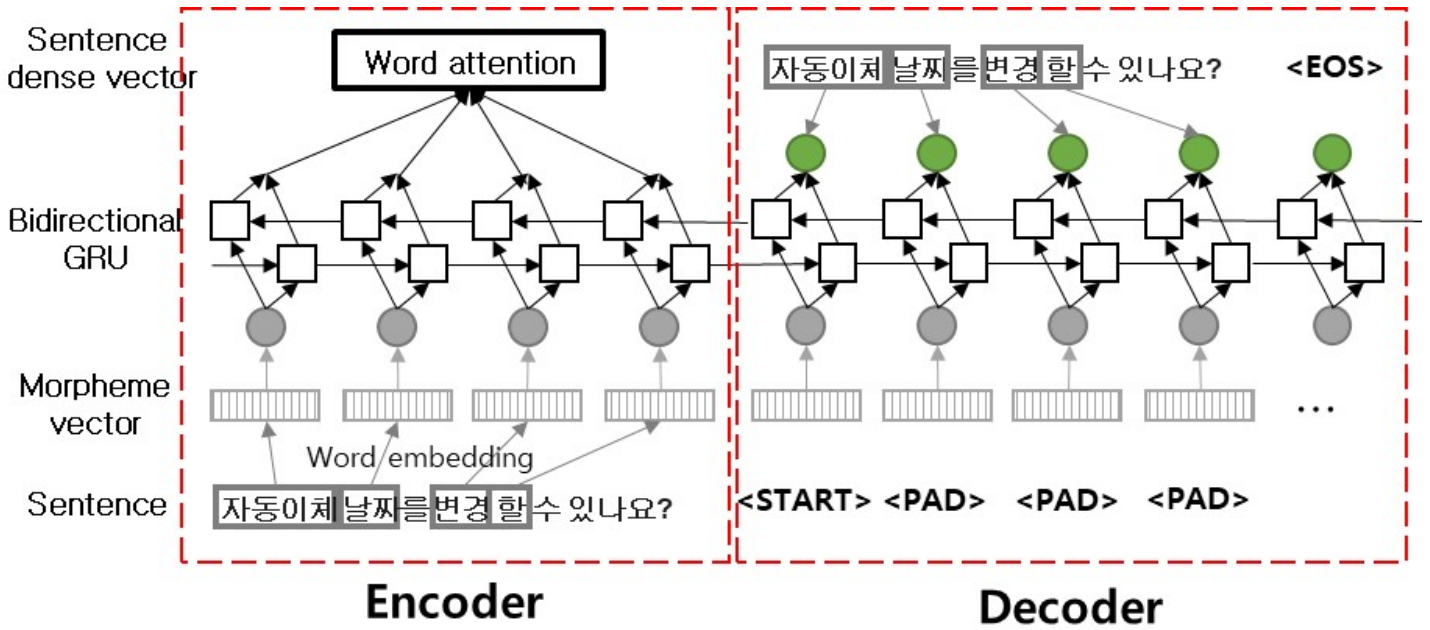


그림 5 Bidirectional GRU RNN 인코더 학습 과정

3.4 지도 학습 기반 대화 문장 도메인 및 화행 분류

Extreme Learning Machine(ELM)은 단일 또는 소수의 은닉 층으로 구성된 전방 전달 신경망(feed forward neural network)의 한 종류로 주로 분류, 회귀 분석, 클러스터링 등에 많이 이용된다. 임의의 가중치를 갖는 입력 층과 은닉 층과, 한 차례의 역전사 과정을 통해 가중치를 갖는 은닉 층과 출력 층 구조의 신경망으로 구성된다. 낮은 오류로 선형 분류 모델을 학습하는 SVM의 장점과, 다차원의 공간을 접거나 구부리는 신경망 분류 모델의 장점을 차용하여 multi-class classification에서 속도와 정확도 면에서 좋은 성능을 보이는 분류 모델이다.

내 세부 화행을 분류하기 위한 총 3개의 ELM classifier가 사용되었으며 대분류, 세부 분류, 화행 각각의 레이블링 된 데이터를 통해 학습되었다.

첫번째 ELM classifier는 그림 6와 같이 5개의 대분류(대출, 예금, UX, 기타, 상담원 연결)에 대한 분류를 수행한다. 초반 분류 과정에서 상담사 연결 확률을 계산함으로써, 챗봇이 응답할 수 없는 대화에 대해서 빠르게 상담사에게 넘긴다.

두번째 ELM classifier는 입력 문장의 벡터와 대분류 자질을 함께 이용하여 세부 분류를 분류한다. 대출에는 6개, 예금에는 15개의, UX은 5개, 기타에는 4개의 세부 분류 체계가 있다. 출력 층에서 softmax 활성화 함수를 사용하여 [0, 1] 사이의 정규화된 수치를 얻음으로써 multi-label 분류를 할 수 있도록 모델을 구성하였다.

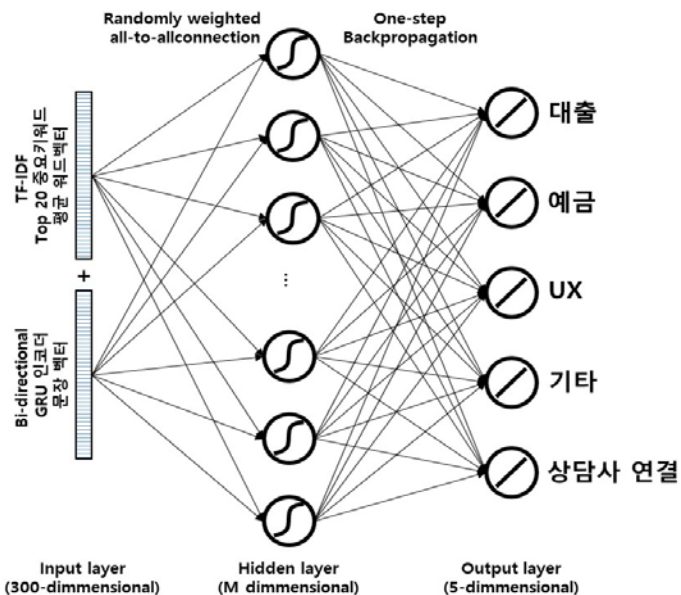
세번째 ELM classifier는 첫번째 대분류와 세부 분류까지 입력 자질로 활용하여 화행 분류를 수행하는 모델이다. 화행의 경우 대분류, 세부 분류와 독립적으로 대화의 행위적 측면에 대한 정보이므로 세부 분류의 아래 단계에서 수행된다.

4. 실험 설계

4.1 실험 데이터 설명

기존의 은행 및 금융권에서는 고객과의 상담 대화를 공개하지 않기 때문에 기계 학습과 관련된 연구 사례가 많지 않았다. 본 연구는 모바일 뱅킹 앱에서 발생한 비식별화 처리된 고객상담 대화 데이터의 일부 샘플을 이용하여 수행하였으며, 추후 더 많은 양의 고객상담 데이터를 정제하여 화행분류 모델의 성능 평가 및 고도화 연구를 추가적으로 수행될 것이다.

그림 5에서도 알 수 있듯이, 상담 고객이 원하는 은행



본 논문에서는 대화 세션의 도메인 대분류와 대분류

그림 6 ELM 알고리즘 기반 도메인 분류 모델

업무로는 기능을 찾기 위한 상담, 특정 상품에 대한 구체적인 조건이나 설명 제공, 고객의 금융 정보(자동이체 금액, 납일 일, 이체 한도 상품 등) 조회/변경/해지 등, 상품 추천, 오류 신고 및 확인 등이 있다.

본 논문에서 문장 임베딩 모델에 사용된 한글 말뭉치는 일부 대화 데이터와 금융 도메인 관련 한글 위키피디아 문서, 뉴스 기사에 등장한 문장 30만 문장으로 학습하였다. 도메인 및 화행 분류 모델의 성능 평가를 위한 평가셋으로는 대분류 별로 50개씩 총 200개의 고객상담질의 문장이 사용되었다.

4.2 부분적 실험 결과

본 논문에서 제시한 고객상담 대화 화행분류 연구는 현재 수행 중인 연구로, 이 논문에서는 부분적 실험 결과인 대분류 분류 결과만을 공개한다.

표 1은 ELM 알고리즘을 사용한 분류 모델의 대분류 정확도이다. 추후 연구에서 보다 많은 평가 셋 대화 세션 데이터를 정제하여 화행분류를 수행한 결과를 공개하고자 한다.

표 1 은행업무 대분류 분류 결과

평가항목	대출	예금	UX	기타	평균
평가 문장 수 (개)	50	50	50	50	50
분류 일치 문장 수 (개)	44	46	34	41	41.3
정확도(%)	88	92	68	82	82.5

5. 결론

본 논문은 은행 업무과 관련된 고객 상담을 위한 챗봇을 개발하기 위한 자연어 이해 과정의 일부로, 응답 결정하기 위하여 대화의 도메인과 화행을 분류하는 방법에 대해 기술하였다.

LDA 클러스터링을 통해 대화 데이터의 도메인과 화행 분류 체계를 정의하고, 전처리 과정으로 금융 도메인에 특화된 언어 모델을 워드임베딩 모델을 통해 학습하고, 대화 세션 내 TF-IDF로 추출된 주요 키워드와 개체명을 통해 얻은 입력 자질과, RNN 인코더를 통해 문장 벡터를 사용하여 ELM 알고리즘을 사용하여 도메인과 화행을 분류한다.

이 논문은 현재 진행 중인 연구 개발로 학습데이터를 일부 샘플링 한 데이터를 기반으로 평가셋을 구축하여 작성되었다. 추후 연구를 통해 고객상담 대화 말뭉치 길의 별로 정제하고, 대화 화행 분류 결과에 따른 응답 방법을 결정하는 대화 모형에 대한 실험과, 고객상담 말뭉치 전수 데이터를 사용하여 응답 성능의 평가를 진행할 예정이다.

감사의 글

본 연구는 미래창조과학부 산업융합원천기술개발사업의 “휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발” (과제번호 2013-0-00131)과 ICT유망기술개발지원사업 “지능형 대화 서비스를 위한 화용 및 문맥 분석 기반 대화솔루션 개발” (과제번호 2017-0-00868) 과제의 지원으로 수행되었음.

참고문헌

- [1] A. Iftene and J. Vanderdonckt, MOOCBuddy: a chatbot for personalized learning with MOOCs, In proc. of RoCHI, vol. 91, 2016.
- [2] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, R, A New Chatbot for Customer Service on Social Media, In proc. of ACM CHI 2017, pp. 3506-3510, May 2017.
- [3] J. Aguilar, K. Berbos, K. Cayube, R. Sagum, and B. Comendador, PHILEX: Philippine Land Law Expert Chatbot, In proc. of ICCRD 2013, ASME Press, 2013.
- [4] K. Oh, D. Lee, B. Ko, and H. Choi, A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation, In proc. of IEEE MDM 2017 pp. 371-375, May, 2017.
- [5] A. Dole, H. Sansare, R. Harekar, and S. Athalye, Intelligent Chat Bot for Banking System, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol. 4 no. 5, 2015.
- [6] K. B. Shah, M. S. Shetty, D. P. Shah, and Pamnani, R, Approaches towards Building a Banking Assistant. International Journal of Computer Applications, vol.166, no.11, 2017.
- [7] B. Galitsky and D. Ilvovsky, Chatbot with a Discourse Structure-Driven Dialogue Management, In proc. of EACL 2017, 2017.
- [8] K. Oh, Z. Kim, H. Oh, C. Lim, and G. Gweon, Travel intention-based attraction network for recommending travel destinations, In proc. of IEEE BigComp 2016, pp. 277-280, , Jan. 2016.
- [9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, In Advances in neural information processing systems, pp. 3111-3119, 2013.
- [10] G. B. Huang, Q. Y. Zhu, and C. K. Siew, Extreme learning machine: theory and applications, Neurocomputing, vo.70, no.1, pp.489-501, 2006.
- [11] 이동건, 오교중, 최호진, 허정, 질의응답 시스템에서 형태소임베딩 모델과 GRU 인코더를 이용한 문장유사도 측정, 제28회 한글 및 한국어 정보처리 학술대회 논문집, 2016.