의료 조언을 위한 질문 의도 인식: 학습 데이터 구축 및 의도 분류*

이태훈 [O1] 김영민 [1,2†] 정은지 [1] 나선옥 [1]

[1] 한양대학교 산업융합학부, [2] 한양대학교 기술경영전문대학원

{leeth5225, yngmnkim, eunji0729, nso94}@hanyang.ac.kr

Question Intent Detection for Medical Advice: Training Data Construction and Intent Classification

Tae-Hoon Lee [O1] Young-Min Kim[1,2] Eunji Jeong[1] Seon-Ok Na[1]

[1]Division of Interdisciplinary Industrial Studies, Hanyang University

[2]Graduate School of Technology & Innovation Management, Hanyang University

요 약

대부분의 과업 지향 대화 시스템에서는 의도 인식과 개체 인식이 선행되어야 한다. 본 연구에서는 의료 조언이라는 신규 분야에 대한 대화 시스템 구축을 위해 사용자 질문의 의도를 인식하는 분류 문제를 다룬다. 최종 목적에 해당하는 의료 조언을 위해 필요한 의도 카테고리를 정의하는 것에서부터 학습데이터 수집 및 구축, 레이블링을 위한 가이드라인을 상술한다. 질문 의도 인식을 위해 BERT 기반의 분류 모델을 사용했으며 한국어 처리를 위해 변형된 KorBERT도 적용한다. 딥러닝 기반의 모델이 본 연구에서 구축한 중규모의 학습 데이터에서도 좋은 성능을 보이는 것을 검증하기 위해 일반적으로 많이 쓰이는 SVM도 비교 모델로 활용하였다. 실험 결과 8개의 의도 카테고리에 대한 f1 점수가 SVM, BERT, KorBERT에서 각기 69%, 78%, 84% 였으며 향후 데이터 보강을 통해 최종 성능을 높일 예정이다.

1. 서 론

딥러닝 기술의 비약적인 발전에 힘입어 다양한 분야에서 기계학습 기술이 상용화되고 있다. 대화 시스템(dialogue system)도 이러한 기술의 수혜를 받고 있는 주요 분야 중 하나이다[1]. 또한 음성 비서 역할을 수행하는 인공지능 스피커의 성장으로 보다 정교한 대화 시스템에 대한 시장의 요구도 증가하고 있다. 그러나 높아진 수요와 기대에 비해 그에 걸맞는 성능이나 서비스는 아직 제한적인 상황이다.

과업(목적) 지향 대화 시스템은 소통을 위한 챗봇과는 달리 대상 분야에 대한 철저한 분석을 통해 제공할 상세 서비스를 결정하고 대화 프로세스를 구축하는 과정이 필요하다. 더불어 학습을 위한 데이터 수집이 선행되어야 하는데 원하는 목적에 맞는 신규 대화 데이터를 충분히 수집하는 것은 어려운 일이다. 심층 신경망 기반의 end-to-end 시스템이 활발히 연구되고 있지만 아직은 상용화되기 어려운 것도 방대한 학습 데이터가 필요하기 때문이기도 하다[2]. 따라서 대부분의 과업 지향 대화 시스템에서는 발화를 분류하는 '의도 인식'과 주요 개념을 추출해내는 '개체 인식' 작업이 필수적이며 이를 기반으로 대화 프로세스가 이뤄진다[3].

본 연구는 의료 조언이라는 신규 분야에 대한 대화 시스템 구축의 첫 단계인 의도 인식을 다룬다. 최종 목적인 의료 조언을 위해 필요한 의도 카테고리의 정의에서부터 학습 데이터 수집 및 구축을 수행하고, 레이블링을 위한 가이드라인 등을 상술한다. 또한 최근 언어 처리에서 뛰어난 성능을 보이고 있는 BERT [4] 기반의 모델을 분류에 활용하여 의도 인식을 수행한다. BERT는 일반적으로 대규모의 데이터에 적합한데, 본 연구에서 활용하는 중규모의 데이터에서도 좋은 성과를 내는 것을 실험을 통해 살펴본다. 검증을 위해 중소규모 데이터 분류에 가장 좋은 성과를 보이는 기법 중 하나인 SVM도 비교 모델로 활용한다.

BERT는 Google에서 2018년 하반기에 개발한 언어 표현 모델로, 양방향의 트랜스포머 인코더(Transformer encoder)를 쌓아 대규모 코퍼스에 대한 언어 표현을 학습하는 방법이다. 레이블링 없이 학습을 수행한 후, 원하는 세부 학습 목적에 맞춰 심층 신경망의 가중치를 재학습하여 모델을 구축한다. 대부분의 언어 처리 과업에서 최고의 성능을 보였으며 현재 XLNET[5], ALBERT[6] 등 다양한 종류의 최신 기술이 BERT의 원리를 기반으로 한 모델이라고 볼 수 있다.

신규 분야에서 대화 시스템을 구축하기 위해서는 적절한 데이터 수집이 필수적이지만 의료 조언을 위한 대화 데이터는 구축이 매우 어렵다. 따라서 본 연구에서는 최대한 이와 비슷한 질의응답 데이터를 웹에서 수집하여 이를 대신한다.

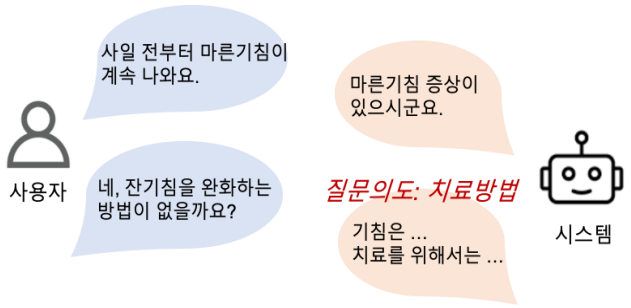
* 본 논문은 산업통상자원부 국제공동기술개발사업 사업으로 지원된 연구결과 입니다(P0000536, 인공지능 디지털 헬스케어 및 챗봇 로봇서비스).

† 교신 저자 : yngmnkim@hanyang.ac.kr

사용자 질문에 대한 의도 인식을 수행하며 문장을 분리하여 각 문장에 대한 의도를 파악하는 것이 목적이 된다.

2. 의료 조언을 위한 질문 의도 인식

본 연구에서 지향하는 의료 조언 시스템의 간단한 예를 그림 1에 소개한다. 사용자의 발화로부터 '증상' 등 주요 개체를 인식하고 질문의 의도를 파악하여 적절한 답변을 제공하는 것이다. 이 논문에서는 전체 시스템 중 **사용자 질문에 대해 의도를 인식**하는 부분만을 다룬다.



[그림 1] 의료 조언 대화 시스템 예제

학습 데이터 구축을 위해 실제 대화 데이터 대신 네이버 지식인 질의응답 중 의료 전문가가 답변한 내용을 수집했다. 네 개의 진료과(신경과, 신경외과, 내과, 이비인후과)에서 각기 답변이 가장 많은 2인의 의사를 선택, 그들이 답변한 QA 셋을 대상으로 했다. 2009년 1월에서 2018년 8월 사이에 작성된 데이터를 수집했으며 이 중 50여 가지 질병에 대해 진단한 경우만을 모아 총 4562개의 QA 쌍을 구축했다. 이 중에서 사용자 질문들만을 모아 문장 단위로 분리하고 각 문장에 대해 실제 질의문인 경우 그 문장을 열 개의 카테고리 중 하나로 레이블링 하였다.

사용자의 질문에는 구체적이지 않은 질문도 포함되어 있어 챗봇이 질문의 의도를 파악하기 어렵게 만든다. 따라서 육하원칙의 요소가 포함된 구체적인 질문 유형을 우선적으로 범주화했다. 예를 들어 '원인 질문'은 '무엇을'이나 '왜'에, '치료 방법'은 '어떻게'에, '진료과 질문'은 '어디서'에 해당한다. 이에 해당하지 않는 질문에 대해서는 '전반적 조언'로 레이블링 했다. 전반적 조언으로 분류한 문장들 중 뚜렷한 차이를 보이는 문장 유형을 그룹으로 묶어 범주화했다. 예를 들어 '병원 방문 필요성'이나 '의심 질병 질문'의 경우에는 각 유형에 해당하는 질문 빈도수가 잦았고 사용자의 질문 의도가 명확하다고 판단하여 각각을 새로운 클래스로 정의했다. 한편 질병 자체에 대한 세부적인 질문 같은 경우에는 이미 존재하는 클래스들과는 확연한 차이가 있어 '의료 정보' 클래스를 생성했다. 마찬가지 이유로 '심각 여부' 클래스도 생성했다. 표 1은 위 과정을 통해 얻어진 열 가지 카테고리의 명칭과 그 정의 및 레이블링 가이드 라인에 해당한다.

[표 1] 10가지 질문 의도

의도 카테고리	정의 및 레이블링 가이드 라인
원인 질문	증상 원인 및 인과관계 문의 (현 상태에 대한 원인, 진단명 질문의 경우)
치료 방법	치료 및 증상 호전 방법 문의 (약 복용, 식단, 응급처치 등 묻는 경우)
병원 방문 필요성	내원 필요 여부 문의 (동일 의도의 질문도 포함 ex. 검사받아야 할까요)
진료과 질문	내원할 진료과 문의 (해당 진료과가 맞는지 확인 및 추천을 묻는 경우)
의심 질병 질문	의심 질병의 일치 여부 문의 (현 증상이 의심 중인 질병과 동일한지 묻는 경우)
의료 정보	질병에 대한 세부 질문 (검사 내용, 비용 등 상세 정보를 묻는 경우)
심각 여부	질병의 심각성 문의 (건강 상태 염려에 대한 질문의 경우 모두 포함)
전반적 조언	증상에 대한 조언 (위 분류들에 해당하지 않고, 현재 상태에 대한 조언, 의견을 구하는 모든 경우)
기타	의료 조언 이외 문의 (사진 첨부 질문, 일반인 대상 질문 등 텍스트 기반 의료 조언의 범위를 벗어나는 경우)
제외	제외된 질문 (문장에 질문이 있으나 문장이 너무 길거나 질문 의도가 명확하지 않은 경우)

의도 인식을 위한 분류 모델로는 BERT 기반의 모델을 사용한다. WordPiece를 입력으로 사용하는 기본 BERT와 한국어를 위해 변형된 KorBERT[7]를 사용하여 각기 분류 모델을 학습 후 성능을 평가한다. BERT의 경우 미리 학습을 완료한 멀티 언어 모델을 초기값으로 사용하고, KorBERT도 제공되는 언어 모델을 초기값으로 설정한다. BERT는 특별한 전처리 없이 WordPiece 단위로 입력 토큰을 분리하는 반면, KorBERT는 형태소 분석을 통해 한국어에 적합한 형태소 단위로 입력 토큰을 구분하게 되므로 성능은 더 좋아질 가능성이 높다. 한편, 비교 모델로 SVM도 테스트 한다.

3. 실 험

3절에서는 2절에서 제시한 데이터 구축 가이드 라인이 유효한지 확인하기 위해 앞서 언급한 세 모델을 사용하여 의도 인식 수행 및 결과에 대해 평가한다. 10개의 카테고리 중 질문 의도에 해당하는 8개 카테고리만 실험 데이터로 사용했다. 표 2는 각 카테고리에 해당하는 데이터의 개수와 비율을 보여준다. 실험에 사용된 총 데이터 수는 5,493개이다. 데이터는 8:2의 비율로 학습셋과 검증셋으로 나누었다.

[표 2] 클래스 종류와 데이터 개수

클래스	개수	비율	클래스	개수	비율
원인 질문	1,948	35.5%	의심 질병	613	11.2%
치료 방법	595	10.8%	의료 정보	458	8.3%
병원 방문	501	9.1%	심각 여부	358	6.5%
진료과 질문	226	4.1%	전반적 조언	794	14.5%
합계			5,493		

모델 간 비교를 위해 중규모에 데이터 분류에 좋은 성능을 보이는 SVM을 사용하여 의도 분류를 진행했다. 전처리로

형태소 분석을 수행하였으며 분리된 형태소 각각에 해당 품사를 붙여 입력토큰으로 활용한다. 형태소 분석기는 Kkma, 한나눔, Okt 등을 사용했으며 그 중에 Kkma를 사용한 SVM 모델이 micro-avg F1 0.69 정도로 가장 좋은 성능을 보였다.

[표 3] BERT-base 의도 분류 테스트 결과

클래스	precision	recall	f1-score
원인 질문	0.87	0.87	0.87
치료 방법	0.80	0.67	0.73
병원 방문 필요성	0.78	0.84	0.81
진료과 질문	0.73	0.67	0.70
의심 질병 질문	0.82	0.79	0.81
의료 정보	0.51	0.49	0.50
심각 여부	0.61	0.79	0.69
전반적 조언	0.77	0.75	0.76
macro-avg	0.74	0.74	0.73
micro-avg	0.78	0.78	0.78

표 3은 BERT에 대해 5-fold cross validation으로 성능을 평가한 결과이며 따라서 5회 실험의 평균값이다. BERT는 micro-avg F1 0.78의 성능을 보였다. 클래스가 8개인 것에 비해 데이터의 규모가 크지 않다는 것을 감안하면 의미있는 결과라 할 수 있다.

결과를 살펴보면 유독 '진료과 질문', '의료 정보', '심각 여부' 클래스의 분류 성능이 낮았다. 의료 정보는 진료과 질문의 데이터 수보다 2배 이상 많은데도 분류 정확도가 가장 낮다. 의료 정보에는 특정 검사에 대한 필요성이나 진단 비용, 치료 가능성 등 다양한 질문이 섞여있어서 분류 성능을 저하시켰을 것으로 보인다.

진료과 질문은 precision에 비해 recall이 낮은 편이다. 레이블이 '진료과 질문'인 문장이 오분류된 경우 중에 '병원 방문 필요성'으로 오분류된 경우는 대략 69%이다. 병원과 관련 단어가 두 클래스에서 자주 등장하는 것이 오분류의 원인으로 추측한다. 반면, 심각 여부는 recall에 비해 precision이 낮다. BERT 모델이 심각 여부로 분류했는데 오답인 경우 중 약 19%는 의료 정보로 분류한 경우였다. 의료 정보에 모호한 문장들이 다수 존재해서 BERT가 모호한 문장을 의료 정보 클래스로 예측하는 경향이 존재하는 것으로 보인다.

[표 4] KorBERT-morp 의도 분류 테스트 결과

클래스	precision	recall	f1-score
원인 질문	0.90	0.89	0.90
치료 방법	0.85	0.79	0.82
병원 방문 필요성	0.85	0.89	0.87
진료과 질문	0.86	0.89	0.87
의심 질병 질문	0.86	0.86	0.86
의료 정보	0.64	0.63	0.64
심각 여부	0.71	0.83	0.77
전반적 조언	0.82	0.82	0.82
macro-avg	0.81	0.82	0.82
micro-avg	0.84	0.84	0.84

추가적으로 동일한 데이터셋을 사용하여 KorBERT에 대해 5-fold cross validation을 진행했고 결과는 표 4와 같다. KorBERT는 micro-avg F1 0.84의 성능을 보이며 BERT와 비교해 0.06 가량 증가했다. 또한 모든 클래스에 대해서 분류 성능이 향상됐다. KorBERT는 형태소 분석을 통해 한국어의 특성을 반영하기 때문에 어절 단위로 토큰을 구성하는 BERT보다 우수한 성능을 기록한 것으로 보인다. 하지만 한국어에 최적화된 KorBERT를 사용했음에도 여전히 의료 정보에 대한 분류 성능이 다른 클래스에 비해 낮다. 의료 정보 클래스의 가이드라인을 재정립할 필요가 있어 보인다.

4. 결 론

본 연구는 의료 조언이라는 신규 도메인에서의 대화 시스템 구축 첫 단계로서, 사용자 질문에 대한 의도 인식을 다룬다. 의료 조언을 위한 실제 대화 데이터 수집은 사실상 불가능에 가까우므로 대신 웹에 공개된 의료 상담 QA 데이터를 활용하였으며 이를 사용한 학습 데이터 구축에 대해 상세히 소개했다. 또한 최근 언어 분석에서 큰 성과를 보이고 있는 BERT 기반의 모델들을(BERT, KorBERT) 의도 분류에 적용함으로써 본 연구에서 구축한 중규모의 데이터에도 뛰어난 성능을 보임을 실험으로 입증하였다. 향후 데이터에 대한 정비를 통해 분류 성능을 향상시키고 개체 인식 결과와 결합하여 최종 대화 로직을 개발할 예정이다.

참 고 문 헌

[1] Gao, J., Galley, M and Li, L., Neural Approaches to Conversational AI, SIGIR '18, pp.1371~1374, 2018.

[2] Li, X., Chen, Y., Li, L., Gao, J and Celikyilmaz, A., End-to-End Task-Completion Neural Dialogue Systems, The 8th International Joint Conference on Natural Language Processing, pp. 733-743, 2017.

[3] Chen, H., Liu, X., Yin, D. and Tang, J., A Survey on Dialogue Systems: Recent Advances and New Frontiers, ACM SIGKDD Explorations, pp.25~35, 2017.

[4] Devlin, J., Chang, M., Lee, K. and Toutanova, K., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL-HLT, pp.4171~4186, 2019.

[5] Yang, Z. Dai, Z. Yang, Y. Carbonell, J., Salakhutdinov, R. and Quoc V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding, NeurIPS, 2019

[6] Lan, Z., Chen, M. Goodman, S. Gimpel, K. Sharma, P. and Soricut, R., ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, arXiv, 2020

[7] 한국전자통신연구원, 한국어 언어모델(KorBERT) (No.2013-2-00131, 휴먼 지식증강 서비스를 위한 지능 진화형 WiseQA 플랫폼 기술 개발).