시니어 사용자를 위한 언어 모델 기반 질환 증상 인식 방법

박민경*, 최진우*, 황보택근**†
*가천대학교 인공지능 헬스케어 연구센터
**가천대학교 컴퓨터공학과
parkminkyung32@gmail.com, {jwchoi, tkwhagbo}@gachon.ac.kr

A Symptom Recognition Method of Diseases for Senior User Based on Language Model

Min-Kyung Park*, Jin-Woo Choi*, Taeg-Keun Whangbo**†
*A.I. Healthcare Research Center, Gachon University
**Dept. of Computer Engineering, Gachon University

요 약

2025년 초고령 사회로 진입할 것으로 예상됨에 따라 고령화 시대에 발생하는 문제점들을 IT기술을 응용하여 지능적으로 해결할 수 있는 인공지능 헬스케어 솔루션이 주목받고 있다. BIS리서치의 보고서에 따르면 헬스케어 산업의 챗봇 시장 규모가 2029년 약 4억 9,800만 달러로 성장할 것으로 예상된다. 따라서 시니어 사용자를 위한 기술 연구가 적극적으로 필요한 시점이다. 본 논문에서는 사전학습한 언어모델과 BiLSTM기반 신경망 모델을 이용하여 시니어 사용자에게 특화된 질환 증상 인식 모델 구현에 관한 범위 및 방법에 관해 기술한다. 이는 시니어 대상 건강관리 챗봇 솔루션에 도입하여 시니어 사용자에게 자주 발생하는 질환들을 조기에 발견할 수 있도록 지원하여 위험의 발생 예방에 도움을 주는 서비스가 될 것으로 전망한다.

1. 서론

통계청의 2019 고령자 통계에 따르면, 우리나라 65세 이상 고령 인구는 급속히 증가하여, 2019년에는 고령사회가 되었으며, 2025년은 초고령 사회로 진입할 것으로 전망되고 있다. 이에 따라 인공지능 헬스케어 솔루션이 인구 고령화와 만성질환 환자 급증에 따른 삶의 질 저하에 대한 선제적, 예방적 대응의 핵심기술로 주목받고 있다. 특히 인공지능 기술과 자연어처리 기술을 접목한 '챗봇' 기술은 사용자와 대화를 나누며 다양한 의학적 요구를 해결할 수 있는 창구의 기능을 할 것으로 기대되고 있다. 하지만 아직 챗봇은 사용자의 요청과 의도의 뉘앙스를 완벽히 이해하지는 못하며 응답에 관한 자유도가 낮다는 한계가 있다. 일반 사용자가 사용하기에도 한계가 있는 챗봇을 시니어 사용자에게 적용하려고 한다면 다른 접근 방식을 이용해야 한다고 판단하였다. 따라서 본 연구에서는 기존의 대화식 명령을 이용하는 챗봇이 아닌 시니어 사용자의 발화를 통해 증상을 예측하고 나아가 관련된 질병에 대한 정보를 받을 수 있는 '시니어 건강 관리' 분야에 집중한 딥러닝과 자연어처리 기반의 '시니어 대상 건강관리 챗봇 솔루션' 개발 내용 중 '시니어 사용자 질환 증상 인식' 신경망 모델을 개발하였고 이를 기술한다.

2. 모델 구현 범위 및 방법

2.1 언어 모델 학습 데이터

대용량의 코퍼스를 사전 학습시켜 단어 임베딩을 생성하는 언어모델을 구축하기 위해, '네이버 지식인'의 건강 카테고리 질문 글을 덤프 데이터로 사용하였다. 헬스케어 분야는 챗봇이 응답을 자연스럽게 하려면 방대한 지식과 학습이 요구되는데, 실제 의료데이터는 접근에 한계가 있다. 개방형 데이터나 서비스 또는 크롤링 기술을 통해 활용 가능한 콘텐츠 확보를 위해 '네이버 지식인'의 건강 카테고리 질문 글 중 '의사 답변'이 등록된 게시글을

크롤링하여 챗봇에 활용할 수 있는 형태로 구축하였다. 보통 자연어처리를 위한 코퍼스 구축에는 Wikipedia같은 문어체로 이루어진 덤프 데이터를 많이 사용하지만, 대화체로 작성된 '네이버 지식인' 데이터가 챗봇 발화 및 응답 시스템을 구축하기에 적절할 것으로 판단하였다. 한 문장씩 데이터 셋을 구성하면 문장당 토큰의 수가 작으므로 모델이 문맥을 파악하기 힘들어 학습이 잘 진행되지 않는다. 따라서 한 게시 글 당 하나의 데이터로 구성하였으며 Word2Vec 모델 훈련 데이터로 사용한 게시글은 총 175,884개이다.

2.2 질환 증상 학습 데이터

가천대 길병원 의료진들을 대상으로 설문조사 및 진료협력센터 자료를 통해 확보한 '질환DB 데이터'는 시니어 다빈도 질환에 관련된 증상명 36가지의 '증상명', '정의', '동의어'가 수록되어 있다. 이 중 '증상명'과 '동의어'를 키워드로 사용해 지식인 데이터(2009-2020)를 추출하였다. Supervised Learning을 진행하기 위해 추출한 게시 글 중 Question 컬럼은 입력데이터로 Keyword 컬럼은 Label로 사용하였다.

<표1> 훈련 데이터 예시

구분	내용
Index	4
Question	며칠 전부터 심장이 마치 긴장되거나 떨리듯 한 느낌과 가슴이 터질 것 같은 통증과 함께 식은땀도 나고 10 분 가까이 숨이 조여 오는 듯 했습니다.
Keyword	심계항진

KoNLPy의 'Kkma' 형태소 분석기를 사용해 데이터를 정제하였고, 한 문장 내 토큰 수는 대부분 500안팎이며 주로 0-200 사이인 것을 고려해 길이를 512로 제한하여 Outlier를 무시하였다. 이렇게 총 115,711개의 훈련데이터를 생성하였다.

2.3 언어 모델

언어모델 생성을 위해 텍스트 분류 문제에 많이 사용하는 인공 신경망 기반의 텍스트 임베딩 방법론 중 하나인 Word2vec을 이용하였다. Word2vec은 유사한 의미가 있는 어휘는 유사한 문맥에서 등장한다는 Distributional Hypothesis에 기반하여 인공 신경망을 이용해 단어(토큰)를 연속적인 벡터 공간으로 임베딩하는 방법이다.

$$p(o|c) = \frac{\exp(u_o{}^T v_c)}{\sum_{w=1}^{W} \exp(u_w{}^T v_c)}$$

(식1) Distributional Hypothesis

Word2vec은 중심단어(c)가 주어졌을 때, 주변단어(o)가 등장할 조건부 확률인 (식1)을 최대화하는 쪽으로 학습이 진행된다. 총 175,884개의 데이터를 KoNLPy의 'Kkma' 형태소 분석기를 사용해 추출한 토큰의 수는 총 30,830개 였으며 Word2vec 모델의 Hyper parameters는 다음 <표2>와 같다.

<표2> Word2Vec Model Hyper Parameters

Iter	Size	Window	Workers	Min_count	Sg
300	256	10	4	10	1

Size는 임베딩 벡터의 차원을 뜻하는데, 차원이 높을수록 Embedding word의 품질이 향상되고 주로 128, 256차원으로 지정한다.[1] 실험에서는 200차원에서 300차원 사이로 지정할 때 좋은 성능을 보였다. min_count는 단어 등장 최소 빈도수를 의미하며 3에서 5, 10으로 늘려가며 실험을 진행하였으며 min_count의 수가 높아질수록 Task Model의 val_loss 값이 개선되었다. Sg는 0이면 CBOW, 1이면 Skip-grams 방식을 사용한다. 여러 논문에서 성능 비교를 진행하였을 때, 전반적으로 Skip-grams 성능이 좋다고 알려져 있다[1]. 또, Window는 10으로 설정하였고 이는 Skip-grams를 사용할 경우 권장 값이다[2]. 다른 파라미터는 Task Model의 성능에 눈에 띄는 영향이 없었다.

2.4 신경망 학습 모델

RNN의 Gradient vanishing/exploding 문제에서의 취약점을 개선한 'Bidirectional LSTM'모델을 분류 모델로 사용하였다. BiLSTM(양방향 장단기 기억)계층은 시계열, 시퀀스 데이터의 스텝 간의 양방향 장기 종속성을 학습하고 모델이 전체 시퀀스로부터 학습하도록 할 때에 유용하므로 모델로 선정하게 되었다[3]. 긴 문장에서 포함한 단어의 주변 정보를 균형 있게 담기 위한 BiLSTM 레이어와 각 Feature Map의 상의 노드의 평균값을 뽑아 차원을 줄이는 GlobalMaxPool1D레이어로 모델을 구성하였다. 모델 출력층의 활성화(activation), 손실(loss) 함수는 Multi-class Classification 문제에 사용하는 'Softmax', 'Categorical Cross-entropy'로 Keras API를 사용하였고 batch크기는 64로 학습하였다. 과적합을 방지하기 위해 Keras API EarlyStopping을 loss기준으로 적용하였고 epoch 40에서 훈련을 멈추어 복잡도를 낮게 파라미터를 조정하였고 처음 지정한 epoch 300 모두 수행한 후 훈련을 마쳤다.

3. 결과 및 분석

3.1 학습, 검증 및 시험 데이터 결과

<표3> Model Performance Evaluation

Accuracy	Loss	Val_acc	Val_loss	Test_acc
0.9442	0.1914	0.9381	0.8704	0.8349

<표4> Classification Report

	Precision	Recall	F1-score
Macro avg	0.92	0.90	0.91
Weighted avg	0.93	0.93	0.93

모든 클래스의 평균 Precision, Recall, F1-score는 <표4>와 같으며 각 클래스의 수를 고려하지 않는 Macro 평균 보다, 클래스 수별 가중치를 둔 Weighted 평균의 값이 더 컸다. 각 클래스의 f1-score, recall, precision을 고려하여 최종 모델을 선정하였다. 본 문제는 다중 클래스 문제를 다루고 있어 Accuracy 성능보다는 Precision, Recall의 가중조화평균(Weight harmonic Average)인 F-score를 지표로 활용해야 한다. Scikit-learn의 metrics 패키지를 통해 각 클래스별 precision, recall, f1-score를 포괄적으로 살펴보았고 훈련데이터는 각 class의 개수가 다른 Imbalanced 데이터므로 f1-score에 중점을 두어 최종 모델을 선정하였다.

3.2 결과 분석

실험 결과는 토큰의 수가 많은 문장으로 학습하여도 적은 토큰의 수의 발화 데이터도 잘 인식한다는 것을 확인하였다는 것에 의의가 있다. 훈련데이터는 평균 70개의 토큰으로 우리가 챗봇에 사용하는 데이터와는 다소 차이가 있다. 챗봇 발화에 사용되는 문장은 3~20개 정도의 토큰을 가진 길이의 문장일 것이다. 선정한 모델로 예측 모듈을 만들고, '요즘에 잦은 기침이 있어.', '어제는 잇몸에서 피가 났어.' 등 실제 챗봇에 사용하듯 문장을 입력으로 넣고 테스트를 진행해 보았을 때, 긴 문장으로 테스트했을 때와 큰 차이 없이 잘 진행되었다.

또한 36개의 클래스 중 24개의 클래스의 f1-score가 0.9이상으로 측정되었다. 하지만, 데이터를 수집하는 과정에서 클래스 불균형이 생기고 실제로 테스트를 진행해 보았을 때도 수가 적은 클래스의 예측은 성능이 다른 클래스에 비해 떨어진다는 연구[4]처럼 실제 테스트에서도 이를 확인하였다. 따라서 향후 연구로 재샘플링 기법을 이용하여 이를 보완하거나 데이터 구축이 가능하다면, 문제 자체를 Multi-Label Classification 문제로 바꾸어 연구를 진행하려고 한다.

4. 결론

본 연구에서는 '시니어 대상 건강관리 챗봇 솔루션'을 위해 시니어 사용자의 발화 데이터를 얻고, 발화 속의 증상을 분류하는 신경망 모델 연구에 관해 기술하였다. 연구를 통해 최종적으로 선정된 Word2vec 언어 모델과 BiLSTM 신경망 모델의 조합으로 구성된 증상 분류 모델의 활용방안은 다음과 같다.

건강보험공단 자료를 바탕으로 시니어 다빈도 질환 상위 100개 중 질환명이 특정되는 질환들 41가지를 선정하고 질환 별 '정의', '원인', '진료과', '진단', '증상', '증상 설명', '치료', '동의어', '관련 질환' 컬럼으로 구성된 활용데이터를 이용하여, 시니어 사용자의 발화에서 분류된 증상들의 조합으로 시니어 사용자들에게 의심되는 질환을 알려주고 질환의 정보에 관해 알려주어 건강관리를 도와주는 헬스케어 챗봇을 구현할 예정이다. 이는 AI기술과 사회적 이슈가 융합된 신성장 동력 서비스로써 좋은 예가 될 것이며 시니어 사용자들의 다양한 의학적 요구를 해소하며 나아가서는 의료비 감소, 모니터링 시스템 대체로 인건비 등에 대해 비용 절감이 가능한 기술이 될 것으로 기대된다.

Acknowledgement

본 연구는 경기도의 경기도 지역협력연구센터 사업의 일환으로 수행하였음.

[GRRC-가천2017(B04), 인공지능기반 의료상담 챗봇 최적화 솔루션 개발]

참고문헌

[1] Mikolov, Tomas; et al, ICLR 2013 conference submission, "Efficient Estimation of Word Representations in Vector Space". arxiv.org/abs/1301.3781, 2013

[2] "Google Code Archive - Long-term storage for Google Code Project Hosting", https://code.google.com/archive/p/word2vec/

[3] Tao Chen; et al. "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN", Elsevier, 2017

[4] 서민지 외, 클래스 불균형 문제가 있는 다중클래스 텍스트 분류에서의 특징 선택 방법. 대한산업공학회지, 45(2), 93-100, 2019