# 규칙기반 데이터 증강기법을 활용한 한국어 증상발화 데이터 구축

전성원[1], 이동준[2], 이동호[3]
[1]한양대학교 인공지능융합학과 바이오인공지능융합전공 석사과정
[2]한양대학교 국방정보공학과 학부생
[3]한양대학교 인공지능융합학과 바이인공지능융합전공 교수
seacom0601@hanyang.ac.kr, dong270@hanyang.ac.kr, dhlee72@hanyang.ac.kr

# Construction of Korean symptom articulation data using rule-based data augmentation technique

Seong-Won Jeon[1], Dong-Jun Lee[2], Dong-Ho Lee[3]
[13]Dept. of Applied Artificial Intelligence, Hanyang University,
Major in Bio Artificial Intelligence
[2]Dept. of Millitary Information Engineering, Hanyang University

요 약

건강정보 검색 요구가 증가하면서 다양한 건강정보 검색 서비스가 제공되고 있다. 하지만 최근의 건강정보 검색 서비스는 정형화 된 전문적인 의료정보와 그 해석을 제공하기 때문에 사용자는 이러한 정보를 스스로 이해하여 원하는 건강정보를 검색해야 한다. 사용자의 검색 피로를 줄이고 원하는 정보를 정확하게 얻을 수 있는 건강정보 검색 시스템 개발을 위하여 사용자의 비의료적 표현인 한국어 증상발화 데이터 구축이 선행되어야 한다. 이러한 데이터 구축은 많은 시간과 비용이 필요하기 때문에 이를 줄이기 위한 규칙기반 데이터 증강기법을 제시하고, 이를 활용하여 한국어 증상발화 데이터를 증강하였다. 증강된 데이터의 유효성을 보이기 위하여 KoBERT 기반의 증상분류 실험을 진행하였으며, 증강된 데이터가 그 전의 데이터보다 F1 스코어가 더 높음을 확인할 수 있었다.

## 1. 서론

최근 건강정보 검색 수요가 증가하면서 다양한 건강정보 검색 시스템이 개발되고 되고 있다. 현재 제공되는 대부분의 건강정보 검색 시스템은 정형화 된 전문적인 의료정보 대한 해석을 사용자가 직접 이해하여 원하는 건강정보를 찾는 형태의 서비스이다. 이러한 경우, 사용자는 어려운 전문용어 때문에 원하는 정보를 얻기 어렵다. 이를 개선하기 위해 사용자의 비의료적 표현에서 사용자가 원하는 건강정보 검색결과를 제공할 수 있어야 한다.

본 논문에서는 사용자의 비의료적 표현으로 자신의 증상 및 관련 질병을 검색할 수 있는 시스템을 가정하여, 이에 필요한 한국어 증상발화 데이터를 최초로 정의하였다. 또한 실제 환자-의사의 대화 데이터를 기반으로 직접 증상발화 데이터를 구축하였다. 이를 기반으로 다양한 데이터 증강 규칙을 찾아 규칙기반 데이터 증강기법을 제시하고, 증강된 한국어 증상발화 데이터를 구축하였다. 이렇게 구축한 데이터의 유효성 검증을 위해 KoBERT 모델을 활용하여 증상분류 실험을 진행하여 정확도와 F1 스코어를 기준으로 성능을 분석하였다.

## 2. 관련연구

해외에서는 다양한 원격진료 시스템에서 개인정보를 제외한 진료 데이터를 공개하고 있다. 이 데이터는 환자-의사 대화, 최종 진단결과 등을 포함하고 있다. MedDialog[1]는 이 중 환자-의사 대화를 영어와 중국어에 대해 구축한 의료 대화 데이터이다.

국내에는 네이버 지식인의 의료전문가 답변 중 질문을 데이터로 구축한 데이터[2]와, 의료 키워드를 포함한 데이터를 환자의 주관적 증상텍스트[3]로 구축한 사례가 있다. 하지만 MedDialog와 같은 환자의 증상발화 데이터는 존재하지 않는다.

딥러닝은 충분한 학습을 위해 많은 데이터가 필요하여 많은 시간과 비용을 야기한다. 이를 해결하기 위해 데이터의 양을 효율적으로 증가시키는 다양한 데이터 증강기법들이 연구되고 있다.

자연어 데이터 증강기법은 생성모델 혹은 규칙기반의 두 가지 방식으로 구분할 수 있다. 생성모델은 많은 데이터를 짧은 시간에 생성할 수 있다. 하지만 검수 작업이 반드시 필요하고 많은 자원을 소모해야 한다. 반면 규칙기반은 규칙에 따라 데이터를 생성하기 때문에 검수 작업을 줄일 수 있다.

AEDA[4]는 문장에 다양한 문장부호를 삽입하는 기법을 제시하여 자연어처리에서 활용되는 대부분의 모델에 유의미한 성능향상을 보였다. [5]는 자연어 특징에서 도출한 기법들을 제시하였다. 하지만 데이터 특성에서 도출된 규칙이 아니기 때문에 특정 분야의 데이터를 증강하기에는 적합하지 않다. 한편 [6]은 정도부사를 삽입하는 방식으로 한국어 데이터 증강기법을 제안하고 있다. 하지만 정도부사의 수가 적어 1% 미만의 정확도 향상을 보였다. 본 연구에서는 일반적인 자연어 특징에서 도출되는 단순한 규칙 이외에도 증상발화의 특징에서 추가적인 규칙을 도출하여 한국어 증상발화 데이터에 적합한 증강기법을 제안한다.

## 3. 한국어 증상발화 데이터

### 3.1 증상 선택 및 수집

아산병원에서 제공하고 있는 증상 1,005개 중 희귀질환, 정신질환, 소아질환 증상을 제거하여 총 305개의 증상을 데이터 구축에 활용하였다.

### 3.2 학습용 증상발화 데이터 수집

환자의 증상발화의 정의 및 작성방법에 대한 교육을 이수한 연구원 5명이 작성하였다. 증상의 상세설명 및 MedDialog를 참고하여 21,860개 학습용 증상발화 데이터를 작성하였다.

### 3.3 테스트 증상발화 데이터 수집

데이터의 유효성 검증을 위해 테스트용 증상발화 데이터를 작성하였다. 학습용 데이터를 작성하지 않은 연구원 3명이 환자의 증상발화의 정의 및 작성방법에 대한 교육을 이수한 뒤 4,183개 테스트용 증상발화 데이터를 작성하였다.

<표 1> 증상 305개에 대한 한국어 증상발화 데이터

| 수집 용도 | 총 데이터 개수 |
|---|---|
| 학습용 | 21,860 |
| 테스트용 | 4,183 |

## 4. 규칙기반 데이터 증강기법

### 4.1 데이터 특성을 고려한 증강기법 선택

일반적인 자연어 데이터와 달리 증상발화는 단어 하나가 달라지더라도 다른 증상이 될 수 있다. 따라서 검수 작업이 반드시 필요하다. 또한 생성모델로 데이터를 증강하는 경우 많은 자원이 필요하다. 검수 작업과 자원에 대한 부담을 줄이기 위해 단어변형이 없는 규칙을 도출하여 3절에서 구축한 한국어 증상발화 데이터를 증강하였다.

### 4.2 규칙기반 데이터 증강기법

1) 문장 변형률(α) 정의

원래의 데이터에서 변형할 비율(α)을 정의하였다. 변형률은 0에서 1사이의 값을 가질 수 있다. 본 논문에서는 원래 데이터의 길이가 짧음을 고려하여 데이터 증강 시에 α 값을 α=0.05로 사용하였다.

2) 문장부호/신체부위 교체

{변형률 x 교체 가능 단어 수} 값을 교체 단어 수로 정의하였다. 문장부호 형태소, 신체부위 단어를 무작위로 교체하였다. 이를 통해 1,565개 데이터를 생성하였다.

문장부호 교체의 경우, 한국어에서 사용하는 문장부호 코퍼스를 정의하여 AEDA보다 다양한 데이터를 생성하였다. 신체부위 단어 교체의 경우, 신체부위 코퍼스를 정의하여 하위 신체부위 단어로 교체하였다. 이를 통해 한 증상이 다양한 부위에 발생하는 표현을 생성하였다.

3) 정도부사/문장부호/자모음 삽입

{변형률 x 어절 수} 값을 삽입 단어 수로 정의하였다. 정도부사, 문장부호 혹은 자모음을 어절 사이에 무작위로 삽입하였다. 이를 통해 2,871,793개 데이터를 생성하였다.

정도부사 삽입은 [6]을 참고하여 증상발화에 알맞은 14개 정도부사를 코퍼스로 정의하였다. 이를 통해 다양한 증상 정도 표현을 생성하였다. 문장부호 및 자모음 삽입의 경우, 구어체 표현을 나타낼 수 있는 문장부호 및 자모음 코퍼스를 정의하였다. 이를 통해 구어체 표현을 생성하였다.

4) 조사/문장부호 삭제

{변형률 x 형식 형태소 수} 값을 삭제 단어 수로 정의하였다. 조사, 문장부호 형태소를 무작위로 삭제하였다. 이를 통해 34,710개 데이터를 생성하였다.

기존 단어 삭제 기법[5]은 핵심어를 삭제하여 내용이 없는 문장이 생성되기도 하였다. 본 논문에서는 형식 형태소만을 삭제하여, 한국어에서 발생할 수 있는 생략된 문장을 생성하였다.

5) 단어 위치 변경

{변형률 x 어절 수} 값을 위치 변경 단어 수로 정의하였다. 무작위로 어절을 선택하여 그 어절의 위치를 변경하여 74,259개 데이터를 생성하였다.

기존 위치변경 기법[5]은 조사가 잘못 배치되어 다른 의미의 문장이 생성되기도 했다. 따라서 본 논문에서는 어절 단위로 재배치하여 의미의 변형과 문법적 오류를 최소화 하였다.

<표 2> 규칙 별 한국어 증상발화 데이터 생성 예시

| 수집 유형 | 학습용 데이터 | 증강된 데이터 |
|---|---|---|
| 문장부호 교체 | 관절이 아파요. | 관절이 아파요! |
| 신체부위 교체 | 관절이 아파요. | 손목이 아파요. |
| 정도부사 삽입 | 관절이 아파요. | 관절이 아주 아파요. |
| 문장부호 삽입 | 관절이 아파요. | 관절이, 아파요. |
| 자모음 삽입 | 관절이 아파요. | 관절이 아파요.ㅠ |
| 조사 삭제 | 관절이 아파요. | 관절 아파요. |
| 문장부호 삭제 | 관절이 아파요. | 관절이 아파요 |
| 단어 위치변경 | 관절이 아파요. | 아파요. 관절이 |

<표 3> 규칙 별 증강한 한국어 증상발화 데이터 수

| 수집 유형 | 총 데이터 개수 |
|---|---|
| 문장부호/신체부위 교체 | 1,565 |
| 정도부사/문장부호/자모음 삽입 | 2,871,793 |
| 조사/문장부호 삭제 | 34,710 |
| 단어 위치변경 | 74,259 |
| 총 합계 | 2,982,327 |

## 5. 실험 및 분석

학습용 데이터와 증강된 데이터의 유효성을 확인하기 위해 KoBERT[7]를 사용하여 데이터 수집 유형 별 정확도와 F1 스코어를 비교하였다.

<표 4> 데이터 별 KoBERT 증상분류 실험 성능

| 데이터 | 정확도 | F1 스코어 |
|---|---|---|
| 학습용 데이터 | 0.75 | 0.76 |
| 증강된 데이터 | 0.78 | 0.8 |

학습용 데이터로 학습한 경우는 0.75의 정확도, 증강된 데이터로 학습한 경우는 0.8의 정확도를 보였다. 하지만 한국어 증상발화 데이터는 불균형한 데이터이기 때문에 추가적으로 F1 스코어를 도출하여 증상을 옳게 분류한 성능을 비교하였다.

학습용 데이터로 학습한 경우의 F1 스코어는 0.76로, 유의미한 분류 성능을 보였다. 하지만 증상 수에 비해 데이터 수가 적기 때문에 실제 증상검색에 사용될 수 있을 정도의 분류 성능 향상이 필요하다. 증강한 데이터로 학습한 경우의 F1 스코어는 0.8로, 학습용 데이터의 F1 스코어에 비해 0.4 향상된 성능을 보였다.

## 6. 결론 및 향후 연구

본 논문에서는 국내 최초로 한국어 증상발화 데이터를 정의하고 구축하였다. 학습용 및 증강된 데이터 모두 유의미한 분류 성능을 보였으며 F1 스코어를 기준으로 증강된 데이터는 학습용 데이터에 비해 0.4 향상되었다. 이를 통해 한국어 증상발화 데이터가 비의료적 표현으로부터 증상을 분류할 수 있음을 나타내고 규칙기반 데이터 증강기법을 통해 유의미한 데이터 증강이 가능함을 보였다.

다만 데이터 수에 비해 증상이 현저히 많기 때문에 성능향상을 위해서는 추가적인 데이터 증강이 필요하다. 향후 연구에서는 유의어 사전을 구축하여 다양한 표현을 증강하고, 증강된 데이터의 라벨을 검수하는 기법으로 대량 증강이 가능한 규칙기반 증상발화 데이터 기법을 연구하고자 한다.

## 사사문구

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. RS-2022-00155885, 인공지능융합혁신인재양성(한양대학교 ERICA))을 받아 수행된 연구임.
이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 202300000000924).
본 연구는 2023년 과학기술정통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음(2018-0-00192)

## 참고문헌

[1] Zeng, Guangtao, et al. MedDialog: Large-scale Medical Dialogue Datasets. Pro. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. p.924-9250.
[2] 이서희, 강주영. 환자의 주관적 증상 텍스트에 대한 진료과목 분류 모델 구축. 한국빅데이터학회 학회지. 6. 1. pp.51-62. 2021.
[3] 이태훈, 김영민, 정은지, 나선옥. 의료 조언을 위한 질문 의도 인식: 학습 데이터 구축 및 의도 분류. 정보과학회논문지. 48. 8. pp.878-884. 2021.
[4] Karimi, et al. AEDA: An Easier Data Augmentation Technique for Text Classification. Findings of the Association for Computational Linguistics: EMNLP 2021. 2021. p.2748 - 2754.
[5] 조진욱, 정민수, 이정훈, 정윤경. 한국어 텍스트 데이터를 위한 변형적 데이터 증강 방법론. 한국정보과학회 학술발표논문집. 47. 2. pp.592-594. 2020.
[6] 김지환, 조원익, 이동준, 김남수. 정도 부사를 이용한 텍스트 분류 작업에서의 데이터 증강에 관한 연구. 한국통신학회 학술대회논문집. 2022. 2. pp.1569-1570. 2022.
[7] SKTBrain. KoBERT:Korean BERT pre-trained cased. https://github.com/SKTBrain/KoBERT. 2019