합성 텍스트 생성을 위한 ChatGPT 기반 의료 텍스트 증강 도구 개발

공진우[O], 김기연[*], 김유섭[**], 오병두[***]
[O]한림대학교 융합소프트웨어학과,
[*]한림대학교 융합소프트웨어학과,
[**]한림대학교 정보과학대학 소프트웨어학부,
[***]한림대학교 뇌혈관질환 선도연구센터
e-mail: {kongjw0110[O], dekmj001[*]}@gmail.com, {yskim01[**], bdoh[***]}@hallym.ac.kr

Development of ChatGPT-based Medical Text Augmentation Tool for Synthetic Text Generation

Jin-Woo Kong[O], Gi-Youn Kim[*], Yu-Seop Kim[**], Byoung-Doo Oh[***]
[O]Dept. of Convergence Software, Hallym University,
[*]Dept. of Convergence Software, Hallym University,
[**]Division of Software, College of Information Science, Hallym University,
[***]Cerebrovascular Disease Research Center, Hallym University

● 요 약 ●

자연어처리는 수많은 정보가 수집된 전자의무기록의 비정형 데이터에서 유의미한 정보나 패턴 등을 추출해 의료진의 의사결정을 지원하고, 환자에게 더 나은 진단이나 치료 등을 지원할 수 있어 큰 잠재력을 가지고 있다. 그러나 전자의무기록은 개인정보와 같은 민감한 정보가 다수 포함되어 있어 접근하기 어렵고, 이로 인해 충분한 양의 데이터를 확보하기 어렵다. 따라서 본 논문에서는 신뢰할 수 있는 의료 합성 텍스트를 생성하기 위해 ChatGPT 기반 의료 텍스트 증강 도구를 개발하였다. 이는 사용자가 입력한 실제 의료 텍스트로 의료 합성 데이터를 생성한다. 이를 위해, 적합한 프롬프트와 의료 텍스트에 대한 전처리 방법을 탐색하였다. ChatGPT 기반 의료 텍스트 증강 도구는 입력 텍스트의 핵심 키워드를 잘 유지하였고, 사실에 기반한 의료 합성 텍스트를 생성할 수 있다는 것을 확인할 수 있었다.

키워드: 합성 의료 텍스트(Synthetic Medical Text), 텍스트 증강(Text Augmentation), 챗지피티(ChatGPT)

I. Introduction

자연어처리 (Natural Language Processing, NLP)는 방대하게 수집된 병원의 전자의무기록 (Electronic Medical Record, EMR)에 포함된 비정형 데이터로부터 유의미한 정보의 추출이나 분석 등이 가능해 의료 분야에서 큰 잠재력을 가지고 있다 [1]. 예를 들어, 의료진의 의사결정을 지원하거나 환자에게 더 나은 진단 또는 치료 등을 지원할 수 있다. 하지만 전자의무기록은 개인정보와 같은 민감한 정보가 포함되어 있어 충분한 양의 데이터를 확보하기 어렵다.

데이터 증강 (Data Augmentation) 기술은 적은 양의 데이터를 가진 일부 도메인에서 합성 데이터를 생성하여 학습 데이터를 생성한다 [2]. 또한, 이는 데이터의 접근성과 익명성 등을 보장할 수 있는 장점이 있다. 그러나 합성 데이터는 실제로 수집된 데이터가 아니기때문에 오류가 포함될 수 있어 신뢰성이 매우 중요하다. 즉, 합성 데이터는 실제 데이터와 유사해야 학습 데이터로써의 의미가 있다. 따라서 데이터 증강 기술은 의료 분야의 인공지능 구현에 필요하다.

본 논문에서는 최근 OpenAI에서 개발한 대화형 인공지능인 ChatGPT [3]에 기반한 의료 텍스트 증강 도구를 개발하였다. 이를 위해, 우리는 ChatGPT의 적합한 프롬프트를 탐색하였다. 그리고 전자의무기록 중 CT (Computerized Tomography) 영상의 판독문[1]

1) This study was performed in accordance with the Declaration of Helsinki, and it was approved by the Institutional Review Board at Chuncheon Sacred Heart Hospital (IRB No. 2021-10-012)

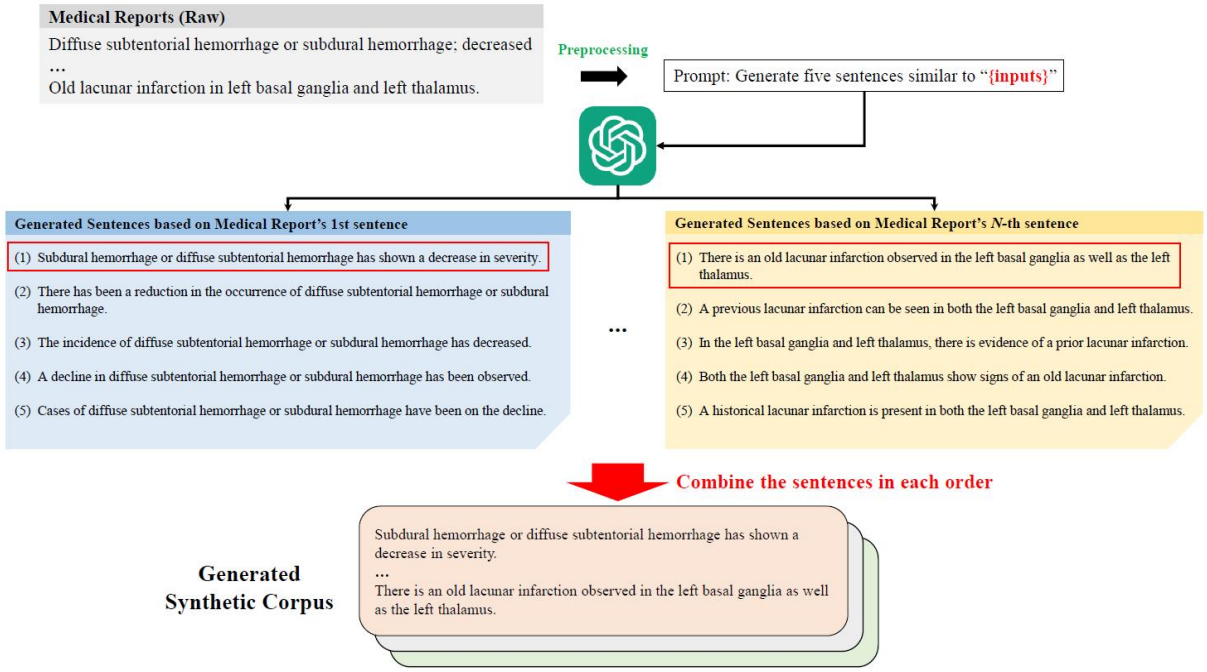

Fig. 1. 의료 합성 텍스트 생성을 위한 ChatGPT 기반 의료 텍스트 증강 도구의 전체 흐름도

(복문 형식)을 입력하고, 이 문장 내 핵심 키워드를 잘 유지하면서 생성할 수 있는지 확인하였다.

II. Proposed Methods

먼저, 의료 텍스트를 문장 단위로 분할한 후 전처리를 수행한다. 이후 각 문장은 우리가 설정한 프롬프트에 추가한 후 ChatGPT의 입력으로 사용된다. 우리가 설정한 프롬프트는 다음과 같다: "Generate five sentences similar to "{input}"". 여기서, "{input}"에 의료 텍스트의 각 문장이 입력된다. 이때 우리는 한 문장에 대해 최대 5가지의 유사한 문장을 생성하도록 설정하였다. 왜냐하면 생성할 경우의 수가 많으면 형태가 유사한 문장이 다수 등장했기 때문이다. 이렇게 생성된 합성 데이터의 예시는 다음의 표 1과 같다.

Table 1. CT 영상 판독문의 한 문장과 합성 텍스트 비교

Original	Diffuse subtentorial hemorrhage or subdural hemorrhage; decreased
Syntethic	Subdural hemorrhage or diffuse subtentorial hemorrhage has shown a decrease in severity.
	A decline in diffuse subtentorial hemorrhage or subdural hemorrhage has been observed.

마지막으로, 각 문장에 대해 생성된 합성 텍스트들은 동일한 순서(또는 차례)의 문장들을 결합하여 하나의 합성 판독문으로 생성된다. 이는 사용자의 선택에 따라 텍스트 파일과 엑셀, 그리고 csv 파일로 저장된다.

III. Conclusion

본 논문에서는 신뢰할 수 있는 의료 합성 텍스트를 생성하기 위해 ChatGPT의 응용 가능성을 확인하였다. ChatGPT는 의료 분야의 입력 문장과 유사한 합성 텍스트를 생성하는 것을 확인할 수 있었다. 향후 우리는 ChatGPT 기반 의료 텍스트 증강 도구를 API로 개발하고, 뇌질환 특화 언어모델을 개발할 것이다.

ACKNOWLEDGEMENT

이 성과는 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1A5A8019303).

REFERENCES

[1] Amin-Nejad, A., Ive, J., and Velupillai, S. (2020, May). Exploring transformer text generation for medical dataset augmentation. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 4699-4708).

[2] Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. Journal of big Data, 8, 1-34.

[3] OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat