TUSNLP at the NTCIR-18 MedNLP-CHAT Task:
Utilization of External Medical Knowledge and Hybrid Approach of BERT and ChatGPT

Aoi Ohara[1*], Nanami Murata[1*], Ami Yuge[1*], and Rei Noguchi[1* †]

[1]Department of Business Economics, School of Management, Tokyo University of Science

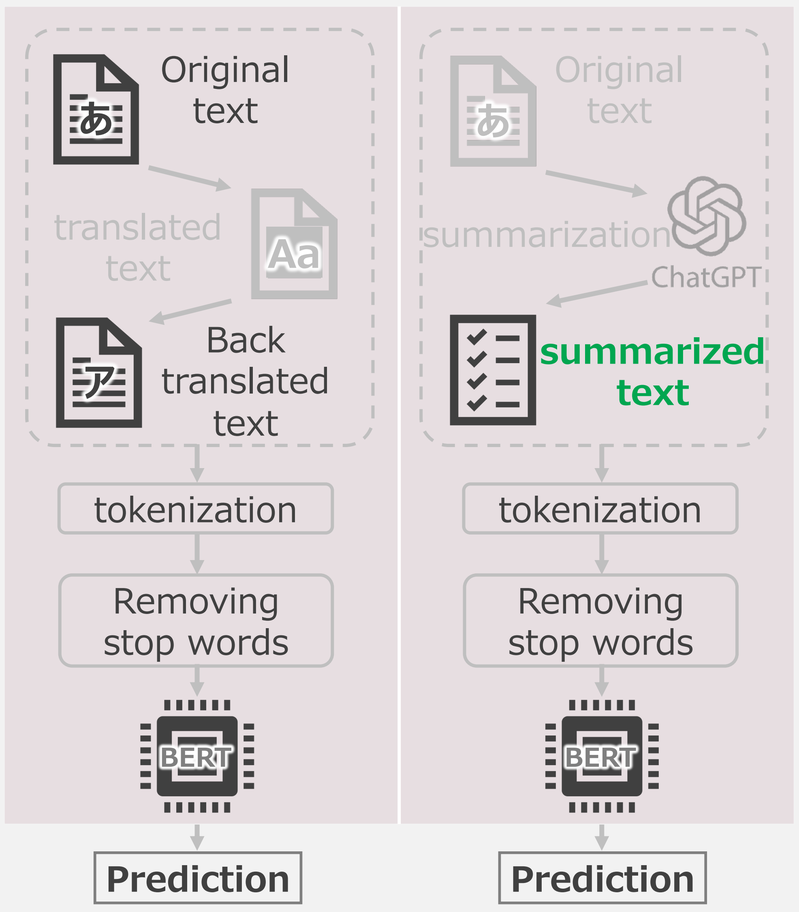
*These authors have contributed equally to this work and share first authorship. / †Corresponding author.

Introduction

Medical chatbots are highly expected to reduce the burden on physicians and improve patient accessibility. However, generative AI-based chatbots are known to output plausible misinformation, known as hallucination, which can cause serious issues of medical safety and ethics. In this study, we propose three model systems for detecting medical, legal, and ethical risks in chatbot responses.

Methods

BERT model

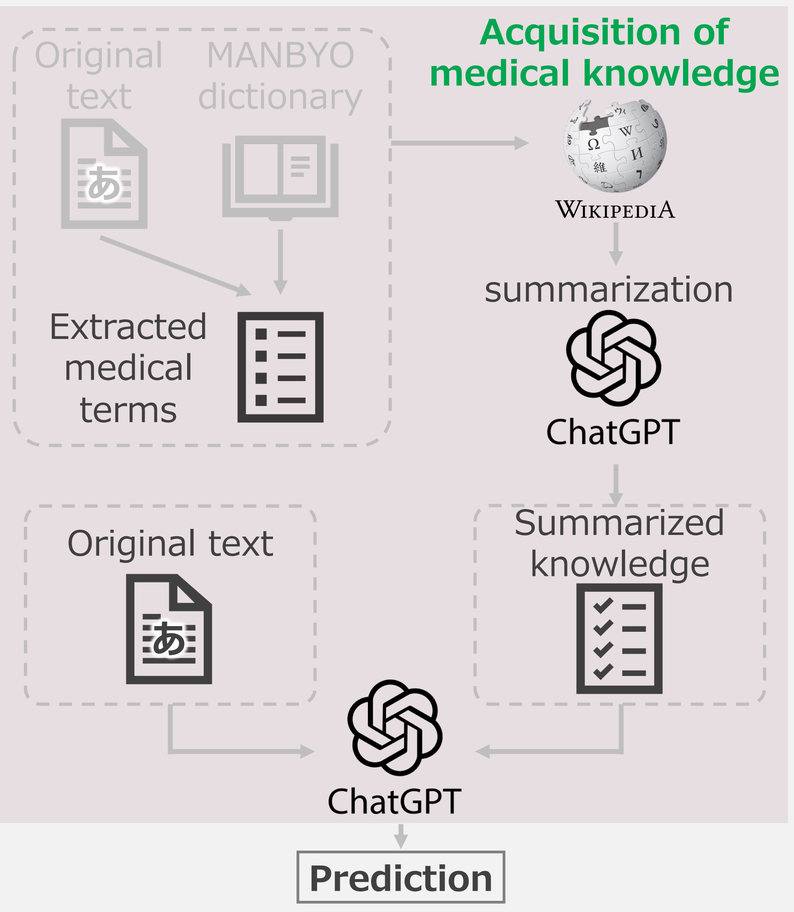


- Developed BERT models that directly predict the risks.
- Experimented with some models:
  - **Data augmentation** by back-translation
  - **Summarization** by ChatGPT (Extractive / Abstractive / Medical-specific)
  - **Two types of tokenizers** (JMedRoBERTa / MeCab)
  - **Removal of stopwords**

Tabel 1 Overview of the BERT Models

Model Name	Back-Translation	Summarization	Tokenizer	Stopwords Removal
Model_A	No	No	JMed	No
Model_B	Yes	No	JMed	No
Model_C	No	No	MeCab	No
Model_D	No	Extractive	MeCab	No
Model_E	No	No	MeCab	Yes
Model_F	No	Extractive	MeCab	Yes
Model_G	No	Abstractive	MeCab	Yes
Model_H	No	Medical-Specific	MeCab	Yes

ChatGPT model

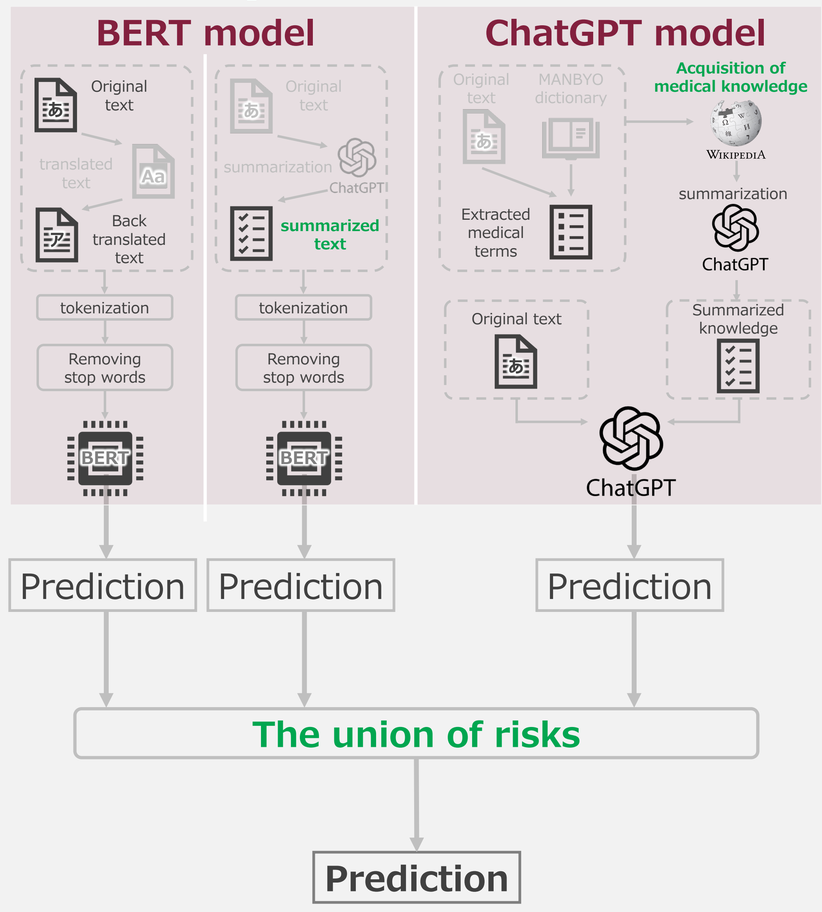


- Developed ChatGPT models that predict the risks using **medical knowledge** extracted from Wikipedia like "**RAG**" system.
- Experimented with some models:
  - **GPT-3.5** / **GPT-4o-mini** model
  - **Three types of prompts**
    - Detecting incorrect information (_harmful)
    - Detecting overstatement (_toomuch)
    - Referring "Risknote" (_Note)

Tabel 2 Results of the ChatGPT Models

Model Name	accuracy	precision	recall	F-measure
3_toomuch	0.6900	0.6000	0.0938	0.1622
3_harmful	0.6238	0.3929	0.3438	0.3667
4mini_toomuch	0.5400	0.3600	0.5625	0.4390
4mini_harmful	0.6809	0.5278	0.5938	0.5588
4mini_Note	0.6200	0.4348	0.6250	0.5128

Hybrid model



- Integrated the risk prediction results of BERT and ChatGPT models by taking **the union set** of the True/False outputs from each model.
- Selected **Model_G** for the BERT model as part of the hybrid model, because of the best results.
- Selected **4mini_harmful** for the ChatGPT model as part of the hybrid model.
- In addition, three types of "**system role**" were added to the prompts for the ChatGPT model, depending on the kind of risk to be predicted:
  - Physician
  - Expert in medical ethics
  - Legal expert in the medical field

Results

Tabel 3 Evaluation results in formal run

System	Risk	Accuracy	Recall	Precision	F1-measure
BERT	Medical	0.524	0.456	0.417	0.407
	Legal	0.817	0.500	0.500	0.489
	Ethical	0.913	0.604	0.618	0.610
ChatGPT	Medical	0.595	0.509	0.550	0.422
	Legal	0.659	0.569	0.539	0.524
	Ethical	0.889	0.533	0.533	0.533
Hybrid	Medical	0.524	0.465	0.447	0.435
	Legal	0.635	0.579	0.542	0.518
	Ethical	0.865	0.637	0.577	0.593

- The BERT model performed well for both legal and ethical risks. In particular, F1 in ethical risk was ranked in the top three.
- The ChatGPT model produced the best results for medical risk, suggesting that external medical knowledge contributes to accuracy improvement.
- The hybrid model yielded an overall improvement in recall value compared to the other systems, indicating that it reduces false negatives by combining the results from the BERT and ChatGPT models.

Conclusion

- Our three systems demonstrated a certain level of accuracy and usefulness.
- Since the hybrid model improved the recall value, the BERT and the ChatGPT models may have different points of focus for prediction, and a detailed analysis of this difference may provide hints for further accuracy improvement.