機械学習と日米の添付文書を用いた 日米医薬品コード間のマッピングの試み

川上 幸伸[*1, *2], 松田 卓也[*1], 高田 春樹[*1], 飛鷹 範明[*2], 田中 守[*2], 木村 映善[*1]
[*1] 愛媛大学 大学院医学系研究科 医療情報学講座,
[*2] 愛媛大学 医学部附属病院 薬剤部

A machine learning approach to mapping drug entries on Japanese and U.S. Label documents.

Yukinobu Kawakami[*1, *2], Takuya Matsuda[*1], Haruki Takata[*1], Noriaki Hidaka[*2], Mamoru Tanaka[*2], Eizen Kimura[*1]
[*1] Ehime University Graduate School of Medicine Department of Medical Informatics,
[*2] Ehime University Hospital Department of Pharmacy

現在、日本と海外の医薬品のコードをマッピングしたオープンデータは殆ど存在しない。日本の医薬品を海外で作成されているナレッジグラフに接続できれば、研究上大きな恩恵を受けることが予想される。一方、マッピング対象の医薬品は多数存在するため、人手によるマッピングでは開発と維持が困難である。そこで、文字列による単純な比較でのマッピング効率を上回るため、機械学習と日米の添付文書を用いた医薬品コード間のマッピングを試みた。DeepL で英語に翻訳した日本の添付文書と、米国の添付文書に対して PubmedBERT と、それをさらにファインチューニングした Sentence-BERT で解析し、コサイン類似度における文章ベクトル間の類似度を評価した。結果、PubmedBERT と比較して Sentence-BERT の方が高い精度を示し、成分レベルでのマッチング率は 63%であった。今後、更なる精度の向上が課題である。

Key words：Drug Label, Drug Codes , RxNorm, BERT, Machine Learning,

1. はじめに

海外では様々な医薬品に関するデータベースにおいて、医薬品が RxNorm[1]にマッピングされている。RxNorm は、NLM(National Library of Medicine)が開発している医薬品名称の正規化を提供する Terminology であり、国際的な研究にも使用されている。海外の医薬品データベースに日本の医薬品を連携できれば、研究上大きな恩恵を受けることが予想される。しかし、わが国の医薬品情報と RxNorm の概念間のマッピング成果は殆ど公開されていない。我々の先行研究[2]にて、日本の医薬品名と RxNorm の概念間でのマッピングを部分的に実施したが、人的資源の制限によりマッピングを終えていない。本研究は先行研究における文字列による単純な比較でのマッピング効率を上回る手法として、日米の添付文書間の文章ベクトルの類似度が最も高いものをマッピング候補とすることを着想した。

2. 方法

2.1 添付文書からのデータ抽出・前処理

日本の添付文書は PMDA（独立行政法人医薬品医療機器総合機構）から、米国の添付文書は NLM が管理している DailyMed から取得する。マッチング率を高めるために、添付文書の中から医薬品の特定に寄与する文章が多いと思われ、かつ日米の添付文書で共通している“DESCRIPTION”（日本の“有効成分に関する理化学的知見”に相当）、“CLINICAL PHARMACOROGY”（日本の“薬効薬理”、“薬物動態”に相当）の項目を抽出する。日本の添付文書から改訂日・YJ（個別医薬品）コード・項目データを、米国の添付文書から改訂日・setId（医薬品固有番号）・項目データを抽出する。また DailyMed で setId と RxNorm の概念コードをマッピングしたデータが公開されているため、日本の添付文書と DailyMed の添付文書のマッチ

ングができれば YJ コードから RxNorm の概念コードへマッピングすることが可能である。まずは成分レベルでのマッチング性能を検証する。評価対象とする日本の医薬品として YJ コード先頭 4 桁の薬効分類番号を用いて循環器官用薬の添付文書に絞込み、さらに無作為に 100 件抽出する。英語で文章ベクトルを算出・比較するために、抽出した 100 件を機械翻訳ツールである DeepL を用いて英語に翻訳する。

2.2 機械学習

機械学習のモデルとして BERT (Bidirectional Encoder Representations from Transformers)を選択した。BERT は Google により提案されたニューラル言語モデルで、医療分野に特化して事前学習されたモデルも公開されている。今回は Pubmed の抄録と PubmedCentral の全文を用いて事前学習されている PubmedBERT[3]と、それをさらに AllNLI データセットでファインチューニングした Sentence-BERT[4]を使用する。

2.3 マッチングの評価

PubmedBERT と Sentence-BERT で日米の添付文書の項目データの文章ベクトルを算出する。各日本の添付文書に対し、文章ベクトルのコサイン類似度が最も高かった米国の添付文書と成分レベルで一致しているか目視で検証する。

3. 結果

DailyMed から米国の添付文書 44,763 件を取得し、解析対象は setId が重複する改定日が古いデータと対象項目がないデータを除外した 41,565 件となった。日本の循環器官用薬のサンプル 100 件中 28 件は DailyMed に該当する成分の医薬品が存在しないため、判定より除外し、マッチング評価対象は 72 件となった。米国の添付文書で成分が一致したのは、PubmedBERT で 72 件中 24 件（33%）、Sentence-BERT で 72 件中 45 件（63%）であった。

4. 考察

本研究では、PubmedBERT と比較して Sentence-BERT の方が高い精度を示した。この理由として、Sentence-BERT は良質な文章ベクトルの生成を目的としたファインチューニングが行われており、より類似度評価に適した文章ベクトルが得られた可能性がある。精度の向上のため、日米の医薬品を成分レベルで連結できた医薬品データベースの情報から Sentence-BERT 用の正解データ作成を検討している。一般的な文章である AllNLI データセットよりも医薬品に特化した文章データセットを用いることで、より適切な文章ベクトルを生成するファインチューニングが期待できると考えている。安全性の観点からも高い精度が必要であり、マッチング候補に対する目視確認も重要と考える。今後、成分レベルのマッチング性能を向上させることで、目視確認の負荷低減を目指している。

DeepL の翻訳の精度については Takakusagi らが報告しており、医学論文においても高い精度を示している[5]。しかし、検証しているサンプル数が少なく、主観的な評価方法であることを踏まえ、今後の結果によっては多言語モデルである Multilingual-BERT を医学用語でファインチューニングして検証することも検討している。

参 考 文 献

[1] Nelson SJ, Zeng K, Kilbourne J, et al. Normalized names for clinical drugs: RxNorm at 6 years. Journal of the American Medical Informatics Association. 2011;18(4):441-8.

[2] 木村 映善, 川上 幸伸, 松田 卓也. 日本の医薬品の RxNorm マッピングの試み.医療情報学 41(Suppl). 2021:605-608.

[3] microsoft. BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext; [https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext(cited 2022-Feb-15)]

[4] Nils R, Iryna G. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 2019; arXiv：1908.10084.

[5] Takakusagi Y, Oike T, Shirai K, et al. Validation of the Reliability of Machine Translation for a Medical Article From Japanese to English Using DeepL Translator. Cureus. 2021;13(9):e17778.