生物医学言語モデルを用いた大規模テキストマイニングによる疾患予防薬探索

Search for preventive medicines by large-scale text mining using biomedical language models

研究課題代表者: 澤田　隆介（岡山大学学術研究院医歯薬学域）
副代表者: 山西　芳裕（名古屋大学大学院情報学研究科）
共同研究者: 星野　哲也（名古屋大学情報基盤センター）

１．研究目的

現在日本では超高齢社会によって、平均寿命と健康寿命のギャップや医療費の高騰、介護の問題が生じている。これらの課題を改善するために疾患を治療する医療だけでなく、疾患を未然に防ぐ疾患の予防が重要である。しかし、一般的に医薬品の研究開発は疾患の診断や治療が目的であり、予防効果に関する医薬品の研究開発は発展途上である。一方で近年の科学術の発展から、様々な疾患に対する予防効果の知見が、研究論文などで発表されてきている。予防効果に関して、公開されているデータを大規模に収集し有効利用することは、今後の予防薬開発の進展に必要不可欠である。

そこで本研究では、疾患予防医薬品開発につながる疾患と予防薬の関係を、大量の文献アブストラクトテキストデータ中から取得することを目的とした。近年発展が著しい AI 言語モデルベースのテキストマイニング手法を用いて、公開されているアブストラクトデータに対して大規模テキストマイニングを行い、予防薬と疾患の関係の抽出を試みる。

２．AI テキストマイニングモデルの構築

AI 言語モデルに対して追加でデータを学習させるファインチューニング学習を行い、予防効果の関係を抽出するモデルの構築を行う。本研究では、生物医学言語モデルの BioBERT(*1*)を AI テキストマイニングモデルとして利用した。抽出したい疾患-予防薬の関係の文章データセットを、PubMed(*2*)から手動で検索して、100 セット用意した。このデータセットを BioBERT に学習させ、PubMed の一部のデータに適用して、予防効果の候補を抽出した。次に、得られた予防効果の候補を目視で確認して、予防効果を示す新しい疾患-予防薬データセットを作成した。このデータセットを再び BioBERT に学習させて、候補を抽出するという構築プロセスを繰り返し、予防効果関係抽出モデルを構築した。最終的に学習データ 4,000 文章セットほどで十分な精度が得られるモデルが構築できた。

３．大規模テキストマイニングの実施

次に、構築した AI モデルを PubMed のアブストラクテキストデータと全体と PubMed Central の論文本文テキストデータに適用した。大規模に AI モデルを適用する必要があるため、計算資源としては不老 Type II GPU ノード群を利用した。

PubMed の FTP サイトよりアブストラクト全データを入手し、AI モデルが適用可能なフォー

マットに整備した（3,800 万件, 2.7 億文章）。これらの文章データを AI モデルに適用した結果、予防効果を示す候補文章として 170 万文章得ることができた。さらに、統合医療言語システム UMLS(*3*)を用いて、候補文章に含まれる化合物と健康状態を検索した結果、健康状態と化合物の予防効果の関係 112 万ペア(2 万種類の化合物、2 万種類の健康状態)の関係を抽出することができた。抽出した結果に関して、KEGG DRUG と KEGG DISEASE(*4*)を利用し、薬物と疾患の関係に絞ると、予防効果関係 3 万ペア(3,000 種類の薬物、1,000 種類の疾患)となった(図 1)。さらに、アルツハイマー病に注目し、抽出結果がどの程度あっているかを、実際に文献の内容を見て確認した(図 2)。予防効果があると抽出された 81 種類の薬物中に関して、文献の内容を確認したところ、44 種類の薬物は予防効果を示すことが文献で示されていることが確認できた。おおよそ半分程度は正しい関係が抽出できていた。

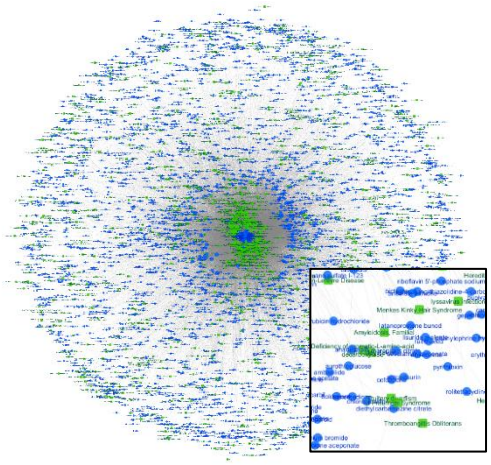

図 1．AI テキストマイニングモデルにより、文献から抽出した疾患と予防薬の関係の図。青い丸が薬物、緑の四角が疾患を示している。予防効果の関係は線で示している。

薬物ID(KEGG)	薬物名(KEGG)	薬効(KEGG)	薬物クラス(KEGG)	リファレンス(PubMed)	最大予測スコア	文献中で確認できた
D03822	リバスチグミン (JAN);	認知症治療薬, アセチ	精神神経系用薬;	15336465	1.94	TRUE
D00070	葉酸 (JP17); Folic acid	貧血治療薬, 造血薬,		16774490	1.93	TRUE
D03365	ニコチン; Nicotine (USP)	禁煙補助薬	その他; DG01718	19184661	1.93	TRUE
D01813	イノシトールヘキサニコチン	末梢血管拡張薬; ビタ		16767098	1.93	TRUE
D09568	ナプロキシノド; Naproxcinod	抗炎症薬	抗炎症薬; DG01908	15202730	1.92	FALSE
D00330	フルルビプロフェン (JP17);	鎮痛薬, 抗炎症薬, シ	抗炎症薬; DG01908	15202730	1.92	FALSE
D00109	アスピリン (JP17); アスパロ	鎮痛薬, 抗炎症薬, 解	循環器系用薬;	10581086	1.92	TRUE
D00141	インドメタシン (JP17);	鎮痛薬, 抗炎症薬, 解	抗炎症薬; DG01906	10581086	1.92	TRUE
D00126	イブプロフェン (JP17);	鎮痛薬, 抗炎症薬, 解	抗炎症薬; DG01908	10581086	1.92	TRUE
D08409	プラステロン; Prasterone	子宮頸管熟化促進薬,	その他; DG01590	2955123	1.92	TRUE
D01530	トコフェロールニコチン酸エ	抹小循環系賦活薬, ビ		10681271	1.92	TRUE
D02331	ビタミン E; Vitamin E (USP)	ビタミン E補充薬		10681271	1.92	TRUE
D00567	セレコキシブ (JAN);	鎮痛薬, 抗炎症薬, シ	抗炎症薬; DG01504	15951464	1.91	FALSE
D00040	コレステロール (JP17);	乳化剤		18819689	1.91	TRUE
D00754	サリドマイド (JAN);	抗生物質, 抗悪性腫瘍		18632006	1.91	FALSE
D09875	プロピオン酸カルシウム;	保存剤		15876979	1.91	TRUE
D00629	ニトレンジピン (JP17);	血圧降下薬, 血管拡張	循環器系用薬;	15876979	1.91	TRUE
D00473	プレドニゾン水和物;	抗炎症薬, グルココル	抗炎症薬; DG01955	19183775	1.90	TRUE
D00003	酸素 (JP17); Oxygen	医療用ガス		11811904	1.90	FALSE
D00191	ビタミン B12; Vitamin B12	ビタミン B12補充薬		17052662	1.90	TRUE
D00166	シアノコバラミン (JP17);	貧血治療薬, 造血薬,		17052662	1.90	TRUE
D02068	塩酸タクリン; Tacrine	認知症治療薬, アセチ	精神神経系用薬;	7875258	1.90	TRUE
D03753	ペリンドプリル; Perindopril	血圧降下薬, アンジオ	循環器系用薬;	15128472	1.90	TRUE
D00118	ナプロキセン (JP17);	鎮痛薬, 抗炎症薬, 解	抗炎症薬; DG01908	17111043	1.90	FALSE

図 2．テキストマイニングの結果の例。アルツハイマーに予防効果がある薬剤を文献より抽出した例。

また、構築した AI モデルを PubMed Central 文献データ 90 万件(1.6 億文章)にも適用した。結果として健康状態と化合物の予防効果の関係 48 万ペア(1 万種類の化合物、1 万種類の健康状態を含む)を抽出した。さらに、この中から薬物と疾患の関係に絞ると、予防効果関係 1 万ペア(1,500 種類の化合物、700 種類の疾患を含む)となった

文献から抽出した予防効果の関係は最終的に Web サイトで公開することを考えているが、現段階ではモデルの抽出精度が十分とは言えない。今後として、学習データの増強や手法に改善などを行い、抽出精度を向上させる予定である。

参考文献

1. J. Lee *et al.*, BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234-1240 (2020).
2. J. White, PubMed 2.0. *Med Ref Serv Q* **39**, 382-387 (2020).
3. O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* **32**, D267-270 (2004).
4. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-D361 (2017).