Vol.2021-NL-248 No.2
Vol.2021-CVIM-226 No.2
2021/5/20

組み込み・除外判定を機械読解により実現した系統的レビュー

佐々木 裕[1]　三輪 誠[1]　安部 賀央里[2]　頭金 正博[2]

概要：本報告では，糖尿病薬に関する医学・薬学文献を対象とした系統的レビューにおいて，組み込み・除外基準に基づく判定を機械読解により実現する方法を提案する．PubMed 検索により絞り込まれたアブストラクトに対して，フルレビューの対象にすべきかどうかの助言を，機械読解を込み込んだ能動学習により行う．実験の結果，組み込み・除外基準に基づく組み込み・除外判定に機械読解が有効であることが明らかとなった．

キーワード：系統的レビュー，機械読解，能動学習，組み込み・除外基準

Judging Inclusion/Exclusion Criteria with Machine Reading for Systematic Reviews

Yutaka Sasaki[†1]　Makoto Miwa[†1]　Kaori Ambe[†2]　Masahiro Tohkin[†2]

Abstract: This report proposes a Machine Reading approach to judging inclusion/exclusion criteria in Systematic Review of biomedical abstracts relevant to diabetes medicines. After searching PubMed, Active Learning with the Machine Reading functionality iteratively suggests abstracts that should be included in or excluded from the subsequent full paper review process. Experimental results show that the proposed Machine Reading approach is effective in the judgement of the inclusion/exclusion criteria.

Keywords: Systematic Review, Machine Reading, Active Learning, Inclusion/Exclusion Criteria

1. はじめに

系統的レビュー（Systematic Review）[1]とは，網羅的に根拠文献を収集・分析する調査手法のひとつであり，調査対象を規定した基準に合致するあらゆる実証的根拠（学術論文）を網羅的に収集・統合し，分析することで，科学的な結論を導く方法である．生命科学の分野では，臨床的疑問（Clinical Question）を解決するために，調査対象となる学術文献をすべて収集することで，バイアスを最小限とし信頼性の高い知見の提示が可能となる．

たとえば，「あるワクチン X の作用に関して人種差があるのではないか」という臨床的疑問を持った場合，ワクチン X を接種した結果に関する学術文献を過不足なく収集し，報告事項を分析し，統計的な結論を得る．「エビデンス（科学的根拠）に基づく医療」（Evidence-Based Medicine）の実践が重要視されるなか，系統的レビューはエビデンスレベルが最も高い手法とされている．

系統的レビューは次のようなプロセスで行われる．

(1) 研究対象とする臨床的疑問の決定
(2) 研究論文の網羅的な収集および選択
(3) 各研究のデータ抽出・妥当性評価
(4) 統計学的解析
(5) 結果の解釈

研究論文の網羅的な収集および選択のためには，選択基準を設定することがポイントとなる．分析の対象とするための条件である「組み込み基準」（Inclusion Criteria），および対象外とするための条件である「除外基準」（Exclusion Criteria）を設定することにより，バイアスを最小限にして，対象となる論文を系統的に選択する．

系統的レビューのプロセスにおいて，ステップ(2)の研究論文の収集と選択は，専門家にとって非常に負荷のかかる部分である．分析の対象となりそうな文献を 1 件ずつ専門家が読み，組み込み基準に合致し，除外基準に合致しないことを判定する．このような現状を改善するため，本研究では，ステップ(2)を対象に，文献が分析対象となるかの判定に深層学習がどの程度貢献できるかを検証する．

特に，本研究では，候補文献集合から分析対象を選択するステップを能動学習（Active Learning）[2]として実現する過程で，対象文書に組み込み基準が書かれているか，また除外基準に合致する記述がないかの判定を機械読解（Machine Reading）の問題として解く．分析対象とすべき文献と分析対象とならない文献を専門家に提示することで，

1 豊田工業大学
Toyota Technological Institute
2 名古屋市立大学
Nagoya City University

Vol.2021-NL-248 No.2
Vol.2021-CVIM-226 No.2
2021/5/20

表 1　組み込み基準，除外基準の定義文

	Inclusion criteria	Exclusion criteria
Study design	randomized controlled trial (RCT)	open-label study,　pooled analysis of multiple RCT, meta-analysis, pre-clinical study, rationale of study design without results, cross over study, post-hoc analysis, publications aimed to report baseline characteristics only, review/commentary
Duration	treatment 12 weeks or over	treatment less than 12 weeks
Participants	patients with T2DM, adult (over 18 years old)	healthy volunteers, pediatrics, elderly, renal impairment, hepatic impairment, impaired glucose tolerant, kidney implantation
Interventions	SGLT2/DPP-4 inhibitors (once-daily) as monotherapy or add-on to other oral agents	not SGLT2 inhibitor/not DPP-4 inhibitors, twice-daily or once-weekly, initial combination, concomitant medication not stable
Comparators	placebo	active comparator, no-comparator
Outcome	HbA1c and FPG observed at the same time-point after 12 weeks or more treatment	either HbA1c or FPG is not reported after 12 weeks or more treatment

専門家のフィードバックを得て，文献の選択を効率化する点を特徴とする．系統的レビューの文献選択に能動学習を用いる際に問題となるのは，学習データ数が本質的に少なく，また選択対象が変われば，新たにゼロから能動学習をしなければならない点である．組み込み・除外基準への合致度合を判定できれば，能動学習の効果を高めることができる．

2. 系統的レビューデータ構築

名古屋市立大学大学院薬学研究科医薬品安全性評価学分野において，「2 型糖尿病薬（DPP-4 阻害薬，SGLT2 阻害薬）における有効性の民族差」を臨床的疑問に設定して，系統的レビューを以下のように実施し，その成果を学術論文[3][4]として発表している．本研究では，その実データを対象として系統的レビューの実験を行う．実際のレビュー対象文献の絞り込みは下記のよう実施された．

＜PubMed アブストレビュー手順＞
① PubMed にアクセス
② Advanced 検索により下記を検索
DPP-4 阻害薬：(((dipeptidyl peptidase 4 inhibitor) OR gliptins) AND placebo) AND type 2 diabetes
SGLT2 阻害薬：(((((((((sodium-glucose transporter 2) OR sglt2) OR sglt2 inhibitor) OR ipragliflozin) OR dapagliflozin) OR empagliflozin) OR tofogliflozin) OR canagliflozin) OR luseogliflozin) AND placebo
③ Article types を Randomized clinical trial に限定し文献を取得
④ 取得した文献のアブストを薬学の専門家が読み，組み込み基準，除外基準で研究の対象となる文献を選別

組み込み・除外基準は表 1 のように設定した．まず，文献の対象文献のアブストラクトの中に表の除外条件のうち１つでも合致する記載があれば，アブストラクト判定表に除外した理由を記入し，その文献は除外した．不明なもの，情報がないものはすべてフルペーパーレビューを行った．その結果，DPP-4 に関しては，検索により得られたアブストラクト 406 件のレビューを 2 人で独立して行い，151 件が最終的なフルペーパーレビュー対象と判断された．SGLT2 に関しては，663 件のうち 125 件がフルペーパーレビューの対象と判断された．

この大変な作業を機械読解技術により軽減できれば将来的に系統的レビューにおける研究の大幅な効率化が期待できる．本論文では，PubMed 検索により得られたアブストラクト集合から分析対象論文を選択する問題を，文献と選択基準を対象とした機械読解問題として解き，機械読解のファインチューニングを行うことで能動学習を実現する手法を提案し，実験により対象データに対してどのような効果が得られるかを検証する．本実験の対象としたアブストラクトは Biopython[1]を用いて PubMed から XML 形式でダウンロードした．ダウンロードしたアブストラクト中に含まれる特殊記号等の UTF-8 のリテラル表現（\xe2\x80\x89 など）は UTF8HTML[2]ツールにより ASCII 記号または HTML の特殊記号表現（≤ 等）に変換した．

3. 準備

3.1 機械読解

機械読解[5]は，広義の QA 技術の一種であり，与えられたパッセージに書かれている内容を問う質問に対して答える技術である．大量文書や知識ベースから質問にマッチする答えを探し出すタイプの QA 技術とは異なり，与えられ

1 https://biopython.org/
2 https://github.com/ssk-coin/UTF8HTML

Vol.2021-NL-248 No.2
Vol.2021-CVIM-226 No.2
2021/5/20

たパッセージに対する理解を問う点が特徴である．与えられたパッセージが物語の場合もあるため，事実と異なっていても，書かれていることに従って答えなければならない．例えば，「日本の首都は時代とともに遷都してきた．… 現在は名古屋である．」とパッセージに書かれていれば，このパッセージの読解としては，「日本の首都は名古屋である」には Yes と答えるのが正解となる．そのため，FrameNet[3]のような外部事実知識を援用して質問に答えることが有効とは限らない．本報告では，選択基準の定義文がアブストラクトに対してどの程度整合するかを深層モデルにより判定することで，系統的レビューを支援することを目的とする．

3.2 文脈に基づく事前学習モデル

近年，Transformer を用いた文脈に基づく単語埋め込みモデルである BERT が様々なタスクで用いられ，高い性能を示している．本報告では，BERT の拡張版である RoBERTa[6]の Sequence Classification モデルを事前学習モデルとして用いる．機械読解のタスクに対応するため，RoBERTa モデルに対して機械読解データセット BoolQ により 2 値分類の学習を行った結果のモデルを用いる．

3.3 BoolQ

BoolQ[7]はパッセージと質問文のペアを与え，その質問をパッセージ中に書かれている内容と照らし合わしたときに正しいか間違っているかの正解ラベルを与えたデータセットである．答えが Yes/No の 2 値の場合に絞った機械読解のデータセットであることが特徴である．

例えば，下記のようなデータが含まれている[7].

質問: Have the San Jose Sharks won a Stanley Cup?
パッセージ:. . . The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016. . .
A: No

この例では，（アイスホッケーの）Stanley Cup で Sharks が優勝したかどうかをパッセージ全体の内容を読み解いて答える問題となっている．BoolQ のデータセットはテキスト含意（TE: Textual Entailment）に類似しているが，TE が 2 つの短い文間の厳密な論理的関係を判定する問題であるのに対し，BoolQ は複数の文からなるパッセージの意味内容を問う読解問題の答えを予測する問題となっている．

BoolQ データセットは，9.4k の訓練データ，3.2k の開発データ，3.2k のテストデータからなる．質問は平均 8.9 トークン，パッセージは平均 108 トークンからなる．

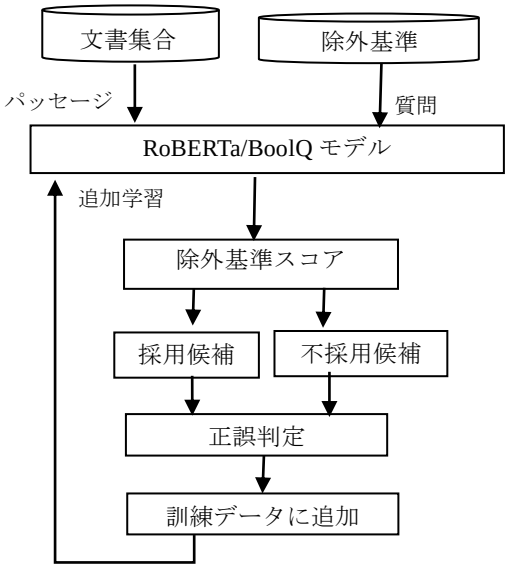

図 1　提案手法の概観図

4. 提案手法

図 1 に提案手法の概要を示す．まず，Huggingface[4]の RoBERTa Sequence Classification 事前学習モデル（roberta-base）に対して，前述の BoolQ データにより 2 値の機械読解の学習を適用する．

学習されたモデルを用いて，アブストラクトの各パラグラフをパッセージとし，除外条件全体を 1 つの質問文としたときの Yes のスコアを算出する．最もスコアが大きいパラグラフを含むアブストラクトを不採用候補，最もスコアが小さいパラグラフを含むアブストラクトを採用候補として提示する．つまり，系統的レビューの実施者に採用して良さそうな候補と採用する可能性が低そうな候補を提示して，最終的なレビューの対象となるかどうかを判定してもらい，文献の採用に関する Yes/No の正解のフィードバックを受ける．実験では，人が介在するのではなく，この部分は教師ラベルを用いてシミュレートする．実作業を想定した場合，系統的レビューの実施者はあらかじめフルペーパーレビューの対象になりそうか，なりそうにないかの事前の情報があることで，判定が効率化される．

ここで組み込み基準を用いないで，除外基準だけを対象としているのは，PubMed の検索絞り込みの時点で，組み込み基準に合致しているアブストラクトを選択しており，組み込み基準に対するアブストラクトの整合性についてはあまり違いがなく，組み込み基準に関してはほぼ満たしているので，追加学習をする効果が少ないためである．

次に，フィードバックされたアブストラクト採否の教師情報に従って，提案した 2 件のアブストラクトの対象パラグラフと除外基準のペアを追加学習のための訓練データ

3 https://framenet.icsi.berkeley.edu/fndrupal/

4 https://huggingface.co/models

Vol.2021-NL-248 No.2
Vol.2021-CVIM-226 No.2
2021/5/20

セットに追加する．[5] 除外基準に対する整合性についてモデルを適応させることが目的であるので，追加学習では BoolQ のデータは用いない．追加された訓練データを用いて，BoolQ と同様の学習を行い，モデルを更新する．これにより能動学習を実現する．能動学習による系統的レビューの支援は既に提案されている[8]が，除外基準とアブストラクトの間の機械読解モデルを追加学習により更新しているところが特徴である．

機械読解モデルが更新された後は，アブストラクトと除外基準に対する機械読解のスコアを用いて，次の候補を提示する．1 度提示したアブストラクトは対象アブストラクトから外していき，アブストラクトがなくなった時点で終了する．

5. 実験

5.1 実験設定

下記の手法を比較対象とする．

- w/ AL； 提案手法
- w/o AL： BoolQ で学習したモデルに追加学習を行わずに採用・不採用候補を提示する方法
- DPP-4+SGLT2： DPP-4 のデータを用いて提案手法で学習したモデルを初期モデルとして SGLT-2 データでの能動学習を実施する手法

BoolQ の学習は，最大長 256，バッチサイズ 32，エポック数 5，学習率 1e-5，最適化 AdamW で行った．その他は Transformer のデフォルトパラメータのまま変更していない．BoolQ の訓練データにより RoBERTa 事前学習モデルを訓練した結果，BoolQ 開発データに対して，文献[6]で報告されている 80%と同等の性能が達成できていることを確認した．

追加学習は学習の影響が過剰にならないように 1 エポック，バッチサイズ 10，学習率 1e-7 とした．その他は BoolQ の学習と変更していない．評価は，毎回採用・不採用候補を提示したときの，各繰り返しまでの提案文献の正解率（Accuracy）により行う．採否の比率によらず，ランダムな場合のベースラインは 50%である．最高の正解率は，DPP-4 の場合は，87.2% (=(151+203)/406), SGLT2 の場合は，68.9% (=(125+332)/663)である．採用数が不採用数より少ないため，2 件ずつ提示すると採用（正例）の数が提示回数より少なくなるためこのような上限が生じる．理想的な場合は，採用データ数までは正解率 100%が継続し，その後．上限値に向かって線形に減少していくことになる．

5 ここで，DPP-4/SGLT2 データセットのアブストラクト採否に関する Yes/No の正解ラベルは，除外基準の機械読解についての正例・負例とは逆になっていることに注意されたい．

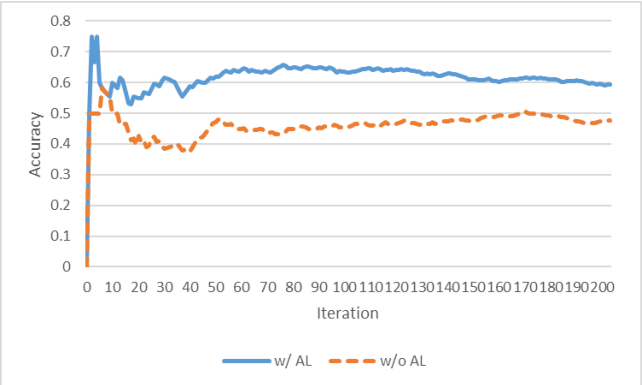

図 2 DPP-4：能動学習(AL) の有無の正解率

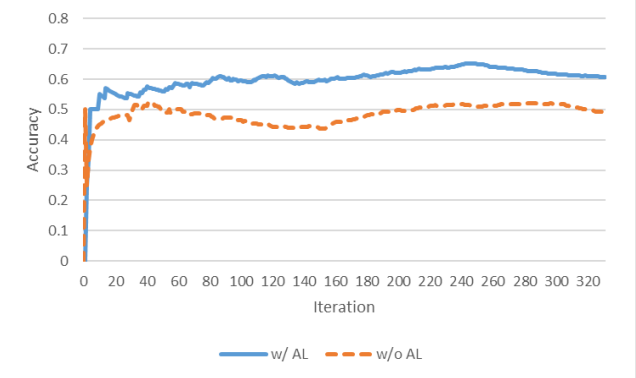

図 3 SGLT：能動学習(AL)の有無の正解率

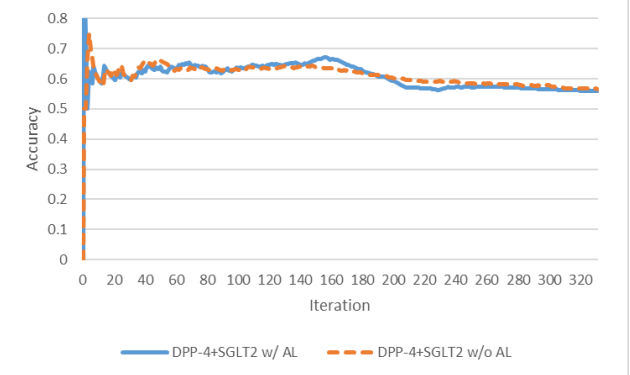

図 4 DPP-4 で学習したモデルで SGLT2 の能動学習を実行したときの正解率

5.2 実験結果

DPP-4 データに関する実験結果を図 2 に，SGLT2 データに関する実験結果を図 3 に示す．1 回の繰り返しで 2 件の候補を提示するので，全データ数の 1/2 の回数繰り返しながら累積の正解率を測定した．どちらの実験でも追加学習を行った場合（w/ AL）は，フィードバックの効果によりランダムベースラインの 50%を超えて 60%程度で能動学習が実現できている．一方，BoolQ で学習後のモデルのままで能動学習を行わずに同じ実験を行う（w/o AL）場合は，ほぼランダムベースラインと同じ正解率になっている．DPP-4 と SGLT2 は基本的に同じ 2 型糖尿病に対する阻害剤の効果に関する系統的レビューデータの収集であるが，選

Vol.2021-NL-248 No.2
Vol.2021-CVIM-226 No.2
2021/5/20

択されているアブストラクトは異なっている．2 つの異なるデータセットによる実験で同様の結果が出たことで，本手法の効果が裏付けられたと考えられる．

次に，図 4 に DPP-4 データによる追加学習を終えたモデルを初期モデルとして SGLT2 データによる能動学習を行った結果を示す．最初の立ち上がりが早くなり，また DPP-4 で学習してあるため能動学習を適用しなくて正解率が 60%程度を維持しているが，単独の場合（図 3 実線）よりも最終的な正解率が少し低下した．

6. 関連研究

分類型の系統的レビューについては，Aphinyanaphong and Aliferis [9]は SVM を用いて系統的レビューシステムを構築した．Frunza ら[10]は高齢者の健康管理サービスの普及戦略についての系統的レビューデータを対象に，医学シソーラス UMLS の概念や情報利得を利用して選択基準を適用する方法を述べた．我々も SVM を用いた選択基準に基づくレビューデータの分類の研究[11]を行っていたが，本研究のように能動学習は行っておらず，系統的レビューに関するデータセットも拡充されている．

能動学習による系統的レビューについては，Hashimoto ら[8]が，系統的レビューの能動学習にパラグラフベクトルと LDA を利用する方法を提案している．この研究では，選択基準文を推薦文献の選択のための明示的な基準として能動学習に取り込んでいない．一方，本研究は選択基準を機械読解の質問とみなし，アブストラクト中のパラグラフとの整合性を判定することで推薦文献を選択し，能動学習を行っている．

7. 考察

本研究で対象にしたデータは深層学習を実行するには比較的小規模であるが，人手で文献を精査するという系統的レビューの性質上，収集できるデータ数は限られる．今回，選択基準に対して機械読解を適用した能動学習により文献提示の正解率が向上することが分かったことで，少ない系統的レビューデータに対して深層学習を適用できることを示すことができた．ただし，BoolQ による学習だけでは，除外基準にアブストラクトが合致しているのかの判定はできていなかった．これは，BoolQ の質問文は一般の英語であり，今回対象とした学術文献や選択基準とは，文長や文体が異なることが原因と考えられる．

BoolQ データによる学習の効果を確認するため，RoBERTa の Sequence Classification モデルから直接能動学習を行った結果を図 5，図 6 に示す. DPP-4 データについては，能動学習の効果があまり得られなかった．SGLT2 データについては，正解率の上昇が BoolQ による学習を行った

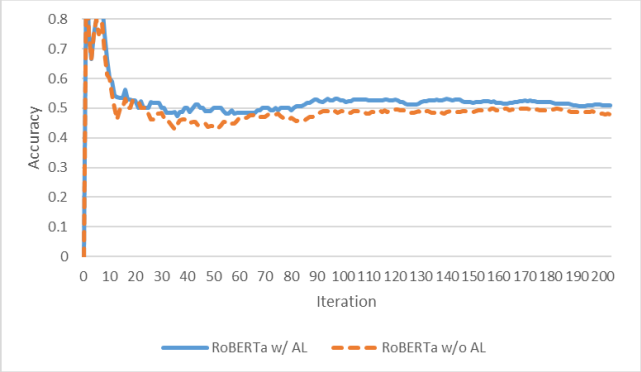

図 5 DPP-4：RoBERTa（BoolQ の学習なし）に対する能動学習(AL)の有無の正解率

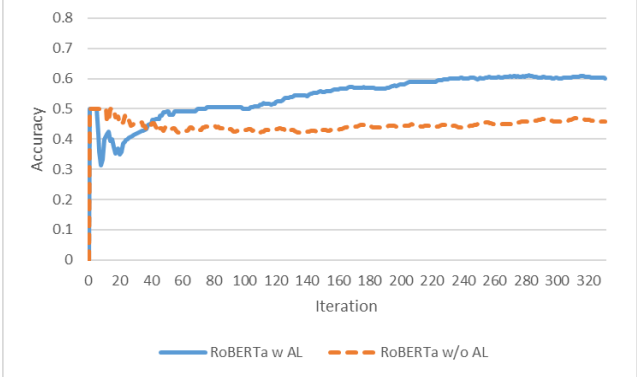

図 6 SGLT2：RoBERTa（BoolQ の学習なし）に対する能動学習(AL)の有無の正解率

場合（図 3 実線）に比べて遅れている．このことから，BoolQ により 2 値の機械読解学習を行うことで，能動学習が安定的に効果を発揮できることが確認された．DPP-4 データと SGLT2 データで能動学習の効果の傾向が異なる理由については今後の分析が必要であるが，SGLT2 データが DPP-4 データよりもデータ数が多いことが一因ではないかと推測している．

能動学習においてフィードバックを受けるアブストラクトの提示方法はいくつか考えられるが，提案手法では最も除外する可能性が高い文献と最も採用する可能性の高い文献のペアを提示して，フィードバックを得る方法を選択した．最も採用の可能性が高い 1 文献のみを提示する方法は，追加学習がアブストラクト全体を除外しない方向にのみ進行していきうまくいかない．別の方法として，除外基準のスコアがボーダーとなっている中位ランクの文献を推薦することで除外基準に対する機械読解の性能が早く向上することが期待できるが，実用場面を想定すると，ボーダーライン上の文献を提示するということは，系統的レビューの担当者は判断が難しい文献を提示されることになり，どのように作業者を支援するのかの実用場面の工夫が必要となる．系統的レビューでは，最終的にはすべてのアブストラクトを判定者が判定する必要があるため，学習の収束の速度よりも，作業者にとってのメリットを考慮して提示方

法を決める必要がある.

また，採否のフィードバックに基づく追加学習により機械読解のモデルを更新する際，本手法では学習データは除外基準全体とアブストラクトの各パラグラフのペアを単位としている．機械読解のタスクとして考えると，直感的には除外基準の各文とアブストラクトとの整合性を学習すれば良さそうだが，実際には教師ラベルは除外基準全体についているため，除外基準の個別の項目とアブストラクトの間の真の含意関係の正解ラベルは付与できない. たとえば，正解が不採用であるアブストラクトについて，除外基準のどの項目にアブストラクトが合致して不採用となっているかの正解ラベル付けがされていない．正解が採用のアブストラクトについては，除外基準を満たさないため，除外基準の各項目とアブストラクトの間の読解関係は負例とできるが，除外基準の各項目に対しての正例が存在しないため学習がうまくいかない.

8. おわりに

選択基準（除外基準）を質問とみなし，アブストラクトをパッセージとみなした機械読解を用いた能動学習により，薬物系統的レビューの性能を向上させることができることを示した. RoBERTa の Sequence Classification 事前学習モデルに BoolQ データセットの訓練を実施することで，選択基準に関する機械読解の追加学習による能動学習が安定的に実現できることを示すことができた．今後，事前学習モデルの性能向上や能動学習の効果の向上により，正解率を向上させ系統的レビューの支援に役立つシステムを構築していきたい.

謝辞

本研究の一部は, 名古屋市立大学特別研究奨励費【1922008】により支援を受けた.

参考文献

[1] D. Gough, S. Oliver, J. Thomas, An Introduction to Systematic Reviews, *Sage*, 2012.

[2] Burr Settles, Active Learning Literature Survey, *Computer Sciences Technical Report 1648*, University of Wisconsin–Madison. 2010.

[3] Y. Ito, K. Ambe, M. Kobayashi, M. Tohkin, Ethnic Difference in the Pharmacodynamics-efficacy Relationship of Dipeptidyl Peptidase-4 Inhibitors between Japanese and non-Japanese Patients: A Systematic Review, *Clinical Pharmacology and Therapeutics*, 102(4):701-708, 2017.

[4] Y. Ito, K. Ambe, T. Hayase, M, Kobayashi, M. Tohkin. Comparison of efficacy of dipeptidyl peptidase-4 inhibitors and sodium-glucose co-transporter 2 inhibitors between Japanese and non-Japanese patients: a meta-analysis, *Clinical and Translational Science*, 13(3), pp.498-508, 2020.

[5] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, Ming Zhou, Gated Self-Matching Networks for Reading Comprehension and Question Answering, In Proc. of *55th Annual Meeting of Association for Computational Linguistics (ACL)*, pp. 189–198, 2017.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv:1907.11692v1, 2019.

[7] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, Kristina Toutanova, BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions, In Proc. of *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 2924–2936, 2019.

[8] Kazuma Hashimoto, Georgios Kontonatsios, Makoto Miwa, Sophia Ananiadou, Topic detection using paragraph vectors to support active learning in systematic reviews, *Journal of Biomedical Informatics*, 62:59‐65, 2016.

[9] Y. Aphinyanaphongs and C.F. Aliferis, Text Categorization Models for Retrieval of High Quality Articles, *JAMIA* 12:207‐216, 2005.

[10] Oana Frunza, Diana Inkpen, and Stan Matwin, Building Systematic Reviews Using Automatic Text Classification Techniques, In Prof. of *23rd International Conference on Computational Linguistics*, pp.303–311, Beijing, 2010.

[11] 佐々木裕，三輪誠，安部賀央里，頭金正博，薬物の系統的レビューにおける選択基準ベクトルの利用，第 24 回言語処理学会年次大会, B6-4, 2018.