BERTScore指標を用いた類似特許検索

Similar Patent Retrieval Using BERTScore

静岡大学　情報学部　准教授
綱川　隆司

2008 年東京大学大学院情報理工学系研究科コンピュータ科学専攻博士課程単位取得退学。博士（情報理工学）。静岡大学情報学部学術研究員、同助教、同講師を経て、2024 年より静岡大学情報学部准教授。自然言語処理の研究に従事。

合同会社 MODE・CREATE
加藤　康聡

2019 年株式会社デンソーを退社。同年、合同会社 MODE･CREATE を設立し代表に就任。“特許情報分析･共有プラットフォーム”の企画・開発に従事。https://mode-create.com/

静岡大学　大学院総合科学技術研究科情報学専攻
山本　隼輔

2024 年から静岡大学大学院総合科学技術研究科情報学専攻に在籍中。

tuna@inf.shizuoka.ac.jp（綱川）　y_katoh@mode-create.com（加藤）　TEL 053-478-1487（綱川）

1 はじめに

日本における年間特許出願件数は近年 30 万件前後で推移しており、企業等において知的財産を管理する上では常に最新の情報を参照する必要が生じる。そのため、出願人が特許を出願する際は先行技術調査を行う。一方、特許庁において出願された特許の新規性や進歩性を特許審査官が判断する際には、先行技術を示す文献を厳密に調査する必要があり、類似特許検索が重要なタスクとなっている。本研究では、予め特許審査官による審査の際に類似文書と認定された先行技術文献（引用文献）に対する被引用文献を、類似文書検索技術によって検索対象集合の中から上位に検出することを目標とし、類似文書検索に BERTScore[1] を用いる方法を提案する。従来の類似特許検索においては国際特許分類（IPC）やキーワードを手掛かりに類似例、非類似例を人手でラベル付けした特許公報群を、それぞれバランスよく同数程度を正例と負例の訓練データとして十分な件数を準備した上で、教師あり機械学習により関連性を判定する判別器を構築する方法がとられているものがある。この手法では訓練データの収集・ラベル付けコストが高い。本研究では検索対象とする特許明細書（引用文献）に対する類似文書として被引用文献（3 ～ 5 件）と、同一の技術カテゴリに属する特許公報 100 件 からなるテストサンプルを用いて、検索対象の特許明細書（引用文献）の類似文書検索において被引用文献を正解としてテストサンプルの中から上位に検出する ことで現実的な時間内で高い精度の検索結果が得られるモデルの作成を目標とする。

2 関連研究

文書間の類似度を測る手法としては、語の出現頻度に基づく TF-IDF 法による手法 [2] や LSA を用いた手法 [3] が提案されている。近年では大規模言語モデ

ル BERT に基づく手法として Sentence-BERT[4,5] や BERTScore[1] による方法が提案されている。Sentence-BERT による手法では類似文献中に含まれる類似文対を抽出し、それらを Sentence-BERT の学習に用いる必要がある。本研究では IPC 分類のセクション・クラスごとにそれぞれ 5 万件の特許公報で事前学習した BERT モデルを用いて教師なしの類似特許検索を行うため BERTScore による手法を採用している。

3 提案手法

2 つの発明を比較する上で、特許明細書の「特許請求の範囲」に記載された各請求項のテキスト間の類似性を判定する。発明間の類似箇所を特定するため、特許請求の範囲を一定の範囲の長さからなる“セグメント”に分割し、セグメント同士の BERTScore を求めることで、BERTScore の大きい箇所の組み合わせを類似箇所として抽出する。

3.1 比較対象テキストの抽出

本研究では特許明細書のうち、比較対象とするテキストを最初の独立項（第 1 請求項）とその従属項に限ることとした。ここで、独立項とは先行する請求項を引用しない請求項、従属項は引用する請求項を指す。すなわち、通常、発明の最も基本的かつ重要な特徴を記述している最初の請求項とそれを引用している請求項のみに限ることで発明の主要な概念に基づいた一貫性のある比較が可能となる。また、第 1 請求項とその関連従属項のみを対象とすることで、処理するデータ量を適切に制限し、計算効率を向上させることができる。

3.2 ストップワードの除去

BERTScore は単語単位の類似度を求めるが、特許明細書に頻出する“請求項”や“前項”などの語は発明間の類似性とは無関係であるため、これらの影響を除くため予めストップワードリストを作成し除去した。

3.3 長い請求項の分割

各請求項は 発明の特徴を網羅的に列挙し、一つの文で表現されることが多く、発明の複雑さや詳細さに応じて請求項の長さに大きなばらつきが生じる。また、慣用的に文末は「～することを特徴とする○○」などの体言止めの形式が用いられる。BERTScore は単語単位で類似度を求める手法であるため比較対象となるテキストの単語数は大きく異ならないことが望ましい。このため、請求項が長い場合は一定の範囲の“セグメント”に分割する。本研究では各セグメントが概ね 50 ～ 150 文字の範囲で 100 文字に近くなるように、かつ文の意味や構造を考慮して、各請求項で区切るとともに分割箇所がなるべく読点になるように分割した。詳細な分割アルゴリズムを手順 1 に示した。

3.4 BERTScore による類似度算出

2 つの特許明細書の特許請求の範囲の比較対象テキストの全てのセグメントの組み合わせに対して、事前学習された BERT の派生モデルを用いて BERTScore（F 値）を求める。同一セグメントの重複を除き、類似度の高い上位 5 組を類似セグメントとして抽出し、発明間の類似度はそれら上位 5 組の類似度の平均として算出する。

3.5 言語モデルのファインチューニング

事前学習された BERT モデルは一般に様々な分野の大量のテキストによる学習がなされたものであり、特許明細書独特のスタイル（例えば、専門的な技術用語の多用、複雑な文構造、特殊な記述形式による表現など）を扱う上で最適とは限らない。本研究では、BERT の派生モデルに対して特許公報を追加学習データとしてファインチューニングを行うことで特許明細書への適合性を高

手順 1 請求項分割アルゴリズム

入力：N 個の請求項テキスト T = (t1, t2, ..., tN), セグメント文字数上限 max, セグメント文字数下限 min
出力：セグメントのリスト L = ()

1. 処理テキスト t の初期値を t1 とし、T から先頭の要素を削除する。
2. t が max 文字以内のテキストであれば、t を L の末尾に追加して 5 へ。そうでなければ 3 へ。
3. t の min～max 文字目に句読点があれば、((min+max)/2) 文字目に最も近い句読点を探し、その句読点の位置を k 文字目とする。句読点がなければ k = max とする。
4. t の k 文字目までのテキストを L の末尾に追加、かつ (k+1) 文字目以降のテキストを新たに t とし、2 へ。
5. T の先頭の要素を削除する。T にまだ要素が残っていれば先頭の要素を t として 2 へ。残っていなければ終了する。

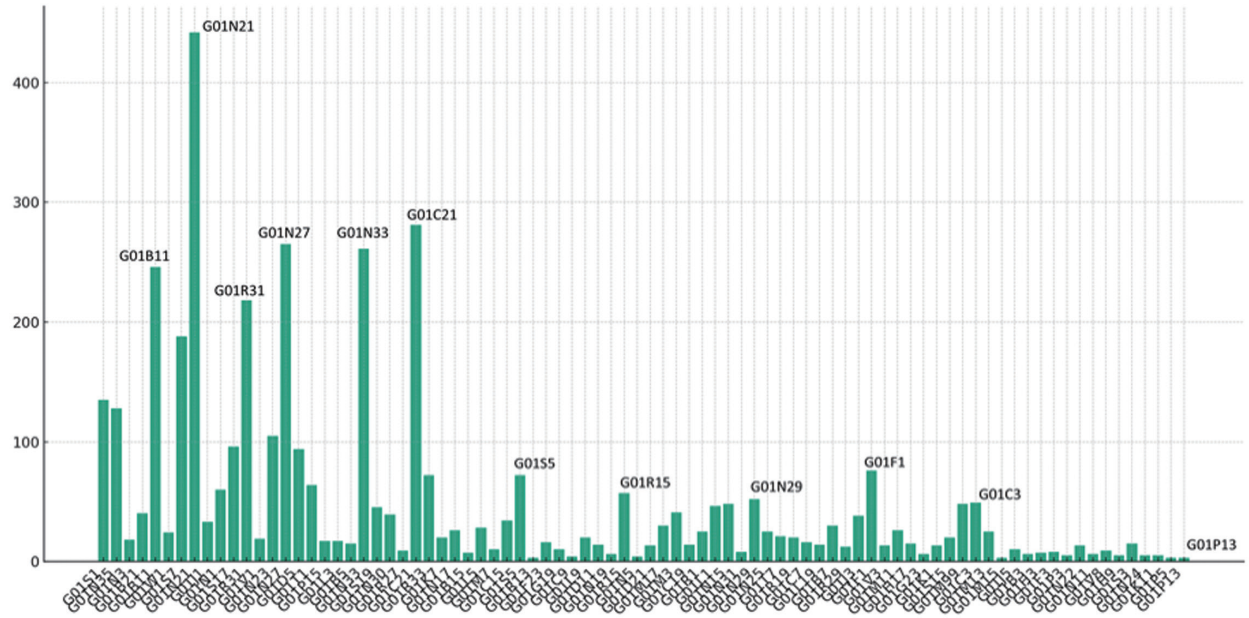

図 1　筆頭 IPC のメイングループで 100 件以上の特許公報をもつクエリの件数

めることを試みた。具体的には、特許明細書に付与された筆頭 IPC のセクション・クラスに該当するそれぞれ 5 万件の特許公報をメインクラスごとに用意したモデル(例えば、G01、G02、G03、G06 など）にファインチューニングを行うデータセットとして使用し、特定の技術分野ごとに特有の専門用語や表現を各 BERT モデルに学習させることで、類似度計算の精度向上を目指した。

4　評価実験

提案手法の評価を行うため、特許庁の審査官によって新規性・進歩性の否定に使われた先行技術文献（引用文献）[1] 群の中の 1 件を検索対象とする特許明細書（以下、クエリと呼ぶ）に対して、クエリに対する被引用文献を正解とする特許公報と、クエリの筆頭 IPC のセクション・クラス・メイングループまでが同じ 100 件程度の特許公報からなるテストサンプルを用いて、正解の特許公報を検索する実験を実施することで提案手法の性能を評価した。

4.1　テストセットの構成

評価用のテストセットは、上述のようにクエリと、クエリの被引用文献 3 ～ 5 件と、クエリの筆頭 IPC のセクション・クラス・メイングループまでが同一（例えば G01「測定；試験」の中の G01N21 や G01R31）の 100 件の特許公報から構成される集合であり、テストサンプル 1000 組を一つのテストセットとして用意した。被引用文献とは、特許審査官による新規性・進歩性の審査対象としての特許明細書であって、その審査結果である拒絶理由通知書または検索報告書には先願としての当該クエリが引用されているものを指す。評価実験においては被引用文献である特許明細書を上位に検索されるべき正解文書として扱う。このような構成により、クエリ（引用文献）と審査官によって選ばれた高い類似性を持つ特許明細書（被引用文献）と、クエリと同じ技術分野の筆頭 IPC が付与されてはいるが人為による類似性の判断を与えられていない特許明細書からなる検索対象母集団を模擬している。

実験した各メインクラスとしては表 2 に示す 7 つを用いた。

4.2　評価指標

クエリに対して類似検索結果の上位 n 位までを検索結果とした場合の適合率 (Precision)、再現率 (Recall)、およびそれらの調和平均である F 値を評価指標とする。

1　引用文献と被引用文献のペアの取得には PatentSQUARE と CyberPatentDesK を用いたが、引用文献情報は、拒絶理由通知書に限定されず検索報告書にも記載されている先行文献が含まれる。これは商用の特許検索システムでは現状、拒絶理由通知書に明記された引例の文献だけを取得できない事情による。

表 1 PR-AUC・適合率・再現率

評価指標	LINE DistilBERT	DeBERTa V2 large
PR-AUC	**0.2362**	0.2050
Top-1 適合率	**0.4712**	0.4195
Top-5 適合率	**0.2585**	0.2433
Top-20 適合率	**0.1047**	0.0991
Top-1 再現率	**0.1622**	0.1444
Top-5 再現率	**0.4451**	0.4189
Top-20 再現率	**0.7212**	0.6826

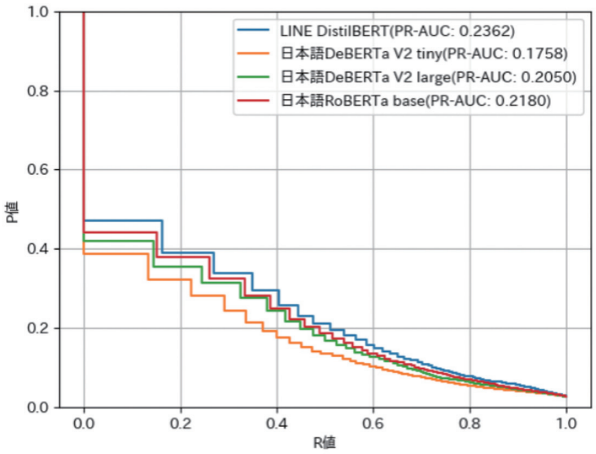

図 2 モデルごとの PR 曲線

適合率＝（検索結果のうち正解文書の数）／ n
再現率＝（検索結果のうち正解文書の数）／（正解文書の総数）

さらに、n を 1,2,3,…と変化させたときに適合率と再現率をプロットして描いた PR 曲線において、曲線の下の部分の面積割合 (PR-AUC) を k に依存しない評価指標として用いた。

4.3　モデルごとの性能比較

メインクラス G01 のテストサンプル 1000 組を用意するにあたり、クエリと同じ筆頭 IPC のセクション・クラス・メイングループが付与された特許公報を取得する際に、2017 ～ 2021 年の 5 年間で 100 件以上が取得できるメイングループを特定した。図 1 はメインクラス G01 の 84 種のメイングループにおけるこの条件を満たすクエリの件数のヒストグラム分布を示している。同様に他のメインクラスでは、G02 では 19 種、G03 では 14 種、G06 では 41 種、H01 では 81 種、H02 では 49 種、H04 では 67 種のメイングループで 1 つ以上のクエリが取得できている。特定のメイングループへの偏りを避けるため、各メインクラスの中から最低 1 つ以上取得できる全てのメイングループを漏れなく抽出してそれぞれ 1000 組のテストサンプルを取得した。これらのテストセットを用いて、BERTScore を求める上で用いる下記に挙げた言語モデルの比較実験を行った。

- LINE DistilBERT[2]
- 日本語 DeBERTa V2 tiny[3]
- 日本語 DeBERTa V2 large[4]
- 日本語 RoBERTa base[5]

類似特許検索はセグメント間類似度の計算を大量に行う必要があるため、計算効率と性能のバランスを考慮し、異なるサイズと特性を持つ複数のモデルを比較対象とした。

図 2 は各モデルによる PR 曲線を示しており、全体的に LINE DistilBERT を用いた場合に最もよい性能が得られている。また表 1 は上位 n 件 (n=1,5,20) を検索結果とした場合の適合率、再現率を示しており、例えば正解文書のうち約 45%は上位 5 位に含まれていることを示している。以降の実験では最もよい性能が得られた LINE DistilBERT を用いる。検索時においては、GPU のない汎用的な PC 環境でも 1 クエリあたり最長で数分以内に検索結果が得られることを確認した。

筆頭 IPC にメインクラス G01 が付与された特許公報 5 万件を用いて LINE DistilBERT の継続学習（特許文書へのファインチューニング）を 2 エポック実施した場合の結果、およびベースライン手法としてメインクラス G01 の特許公報 5 万件のデータを用いた TF-IDF 法による文書類似度を BERTScore の代わりに用いた場合の性能を図 3 に示した。1000 組からなるテストセットを用いて得られたこの結果は、特許明細書による DistilBERT モデルのファインチューニング、および従来のチューニング手法である TF-IDF による手法と比べて提案手法が優れていることを示唆している。

2　huggingface.co/line-corporation/line-distilbert-base-japanese
3　huggingface.co/ku-nlp/deberta-v2-tiny-japanese
4　huggingface.co/ku-nlp/deberta-v2-large-japanese
5　huggingface.co/ku-nlp/roberta-base-japanese-char-wwm

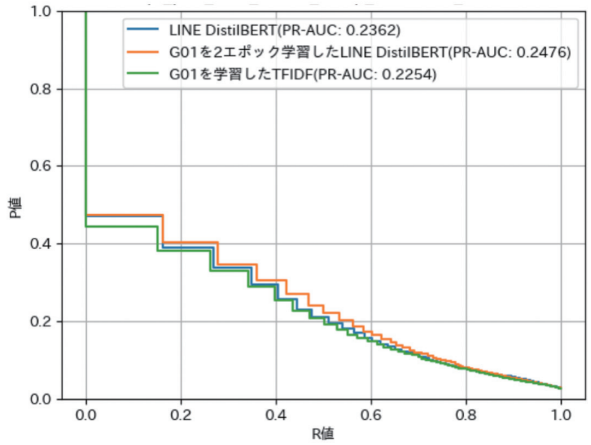

図３　ファインチューニング後のモデル、TF-IDF 手法の PR 曲線

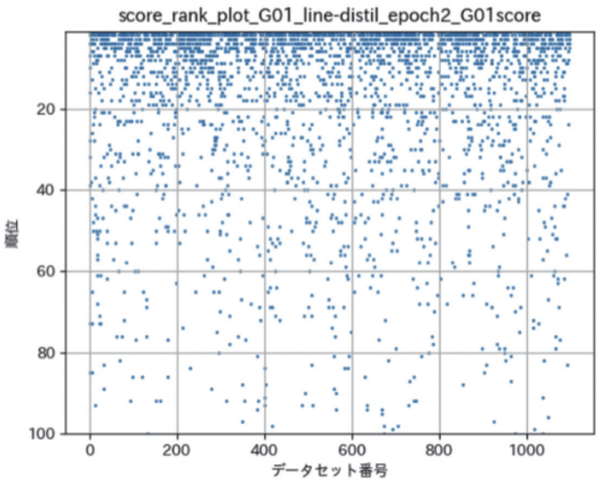

図４　正解文書の順位分布

図４は、G01 メインクラスの各テストサンプルを用いて実験したときの正解文書の順位分布を示している。横軸はテストセット中のあるテストサンプルを示し、あるテストサンプルの検索結果における正解文書の順位を縦にプロットしている。この図からも、正解文書が上位に集中して含まれていることが分かる。この結果は、本手法が高い精度でクエリに対する被引用文献を類似特許として上位にランク付けできていて、一方で、順位が下がるにつれて正解文書の割合は徐々に減少していることから、本手法が効果的に類似特許を検索できることが示されている。

さらに、評価対象の７つの各メインクラスにおいて、当該メインクラスで学習を行ったモデル、および対応するセクション（G または H）単位の特許明細書による継続学習を行ったモデルと元のモデルを比較した場合の PR-AUC を表２に示す。 結果として、どのメインクラスにおいても、元のモデルに比べて各メインクラスで学習したモデルの性能が最も高くなった。一方、セクション単位でファインチューニング（例えば、G セクションならば、用意した G01、G02、G03、G06 を順番に２エポック学習）したモデルは、元のモデルよりは性能が向上したものの、各メインクラスで個別に学習したモデルほどの性能向上は見られなかった。この結果は、より狭い範囲でのファインチューニングが、特定のメインクラスにおいてより効果的であることを示唆している。

表２ メインクラスごとの PR-AUC

メインクラス	学習前のモデル	各セクションで学習	各メインクラスで学習
G01	0. 2362	0. 2412	**0. 2476**
G02	0. 1833	0. 1935	**0. 2002**
G03	0. 2537	0. 2603	**0. 2674**
G06	0. 1809	0. 1917	**0. 1965**
H01	0. 1880	0. 1927	**0. 1983**
H02	0. 1691	0. 1697	**0. 1839**
H04	0. 1715	0. 1894	**0. 1950**

４. ４　人手評価

メインクラス G01 のテストセットのうち、無作為に 58 件を抽出して特許明細書間で求められた類似セグメ

表３ セグメント対人手評価基準

評価値	評価基準
5	非常に高い類似性がある。2 つの文同士で使用されている単語、類語が多数あり、全体のコンセプトも類似しているように感じられる。
4	かなりの共通点が認識できる。2 つの文同士で使用されている単語、類語が多数あるが、全体のコンセプトにはいくつかの違いがある。
3	いくつかの共通点が認識できる。2 つの文同士で使用されている単語、類語が少数あるが、全体のコンセプトはやや異なるアイデアや概念を扱っているように感じられる。
2	ごくわずかな共通点が認識できる。2 つの文同士で使用されている単語、類語が少数あるものの、文のコンセプトは異なっている。
1	共通する要素（単語、類語）がほとんどまたは全く認識できない。文の構造に大きな違いがある。

表 4 セグメント対人手評価結果

評価対象	評価平均
正解文書（上位 2 つと最下位 1 つ）	2.9209
50 位以下の正解文書	2.5722
正解以外の文書	1.7022

ントに対する人手評価を行った。

あるクエリに対して、正解文書[6]、正解以外の文書、および提案手法による類似度が約 100 件中 50 位以下になった正解文書のそれぞれの特許明細書とクエリの間で求められた上位 5 組までの類似セグメント対に対し、表 3 の基準で評価値を付与した。表 4 はそれぞれの条件で得られた評価値の平均を示す。

正解以外の文書の評価平均が 2 を下回り、最も類似したセグメント対であっても共通点に乏しく、一方で正解文書は評価平均が 3 に近く、内容に共通点があるセグメント対が抽出できていることが確認された。正解文書を 50 位以下に限ると評価平均が低下することから、類似するセグメントが少ない正解文書は提案手法においてもクエリとの類似度が低くなる傾向があることが示唆された。

検索結果が下位の正解文書とのセグメント対や、検索結果が上位である正解以外の文書とのセグメント対を評価者が確認した。実際の文書間の類似性とセグメント対の類似度に乖離があるケースを観察したところ、原因の一つとして、10 ～ 30 文字程度の短い請求項を一つのセグメントとして扱っている場合に類似度を求めた場合、用いている語彙が偶然似ているなど、不適切に高い類似度が得られてしまうことが挙げられた。また、比較対象となる請求項の数が少ないためにセグメント数が 1 つや 2 つになってしまう場合も、類似度を求めるセグメント数の組合せが少なくなり、類似度が低下する傾向がみられた。また、一部のケースでは正解以外の文書が検索結果のかなり上位になった場合、出願人がクエリと同じであることが少なからずみられた。これらのことから、提案手法においてはセグメント長の制約を調整する等の改善の余地があるといえる。

6　1 つのクエリに対し正解文書は最大で 5 つまであるため、評価平均を求めるための正解文書の数は 3 つを標準とし、正解文書が 3 つ以上ある場合はそのうち上位 2 つおよび最下位 1 つを評価対象とした。正解文書が 2 つ以下の場合はすべての正解文書を評価対象とした。

5 おわりに

本稿では与えられた特許明細書（クエリ）に対して、クエリに対する被引用文献を正解とする特許公報と、セクション・クラス・メイングループまでがクエリと同じ筆頭 IPC をもつ 100 件程度の特許公報からなるテストサンプルから、BERTScore を用いて類似度の高いものを正解の特許公報として現実的な時間内で検索する手法を紹介した。被引用文献を正解文書とした評価実験においては、クエリとなる特許公報 1 件に対して検索対象約 100 件中、Top-5 の適合率 26%、再現率 45% の性能が得られ、また GPU のない汎用的な PC 環境でも 1 クエリあたり最長で数分以内に検索結果が得られることを確認した。人手による主観評価ではセグメント長の制約や検索対象の除外等、手法に改善の余地があることの示唆を得た。

謝辞

本研究での BERT モデルの評価には特許公報の引用文献と被引用文献の関係に着目したが、これは本研究のテーマを綱川に提案した加藤が特許情報活用研究会で得た "被引用文献数の多い特許公報ほど価値が高い" という知見から得た着想による。この、引用 - 被引用の関係が「特許公報の学習済み言語モデルの性能評価指標」としても有用であることを示すことができ、この研究会にてご教示いただいた Japio IP リサーチフェロー桐山勉氏、ブラザー工業㈱佐久間幹雄氏、京セラ㈱富山明俊氏にこの場をお借りして謝意を表します。

本研究は公益財団法人浜松地域イノベーション推進機構より A-SAP 産学官金連携イノベーション推進事業による助成を受けています。

参考文献

[1] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi (2020). “BERTScore: Evaluating text generation with BERT,” ICLR 2020.

[2] K. Spärck Jones (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval,” Journal of Documentation, 28 (1): 11-21.

[3] S. Deerwester, Susan Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). “Indexing by Latent Semantic Analysis,” Journal of the American Society for Information Science, 41(6):391-407.

[4] N. Reimers and I. Gurevych (2019). “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” EMNLP 2019, pp.3982-3992.

[5] 加納渉，竹内孔一 (2022). “Sentence-BERT を利用したFAQ 検索におけるデータ拡張手法，” 言語処理学会第 28 回年次大会発表論文集，pp.1830-1834.