汎用言語モデルを用いた効率的な類似特許検索

山本隼輔 [1]　加藤康聡 [2]　綱川隆司 [1]
[1] 静岡大学情報学部情報科学科
[2] 合同会社 MODE・CREATE

概要

本研究では汎用言語モデルを使用した教師なしの類似特許検索手法を提案する。予め絞り込んだ類似特許候補の中から、特許明細書中の「請求項」同士のテキスト間の BERTScore による類似度を用いて検索を行う。また、検索性能向上のため汎用言語モデルを分野ごとにファインチューニングする。評価実験において TFIDF による類似文書検索手法と比較し、提案手法の効果を確認した。

1　はじめに

知的財産の管理は社会で重要な役割を担っているが、毎年 28 万を超える出願がある特許を管理する上では多くの課題が存在している。特許の出願時には、先行技術調査や審査において類似する特許があるか探す必要があり、類似特許検索が重要なタスクとなっている。そこで、本研究では汎用言語モデルを用いて求めた特許文書の類似度に基づく類似特許検索手法を提案する。従来の類似特許検索手法は、検索対象特許に対して人手で種となる正解データをある程度用意しなければならない教師あり学習が用いられてきた。それに対し、提案手法では国際特許分類のセクションごとのコーパスを用意して言語モデルを再学習することによって、正解データを用意する必要がなく教師なしで類似特許検索を効率的に高い精度で行うことを目的とする。提案手法の有効性を示すため、特許公報とその被引用文献から構築したテストセットを用いて評価実験を行う。

2　国際特許分類とは

特許は国際特許分類 (以下 IPC) によって分類されている [1]。IPC は階層構造を持ち、上位から「セクション」、「クラス」、「サブクラス」、「メイングループ」、「サブグループ」の 5 つの階層がある。また 1 つの特許公報につき、通常複数の IPC が付与される。その中で、最初に記載されている筆頭 IPC がその発明を代表する分類になっており、本研究では筆頭 IPC のみを用いる。

3　関連研究

従来から文章間の類似度を求める研究は行われており、TFIDF のコサイン類似度を用いた手法 [3] や、LSA を用いた手法 [4] が利用されてきた。近年では、大規模言語モデルを用いたものが利用されており、SentenceBERT を用いて文章を分類する手法 [5] や、BERTScore を用いて文章の類似度を計算する手法 [6] が存在している。また、大規模言語モデルを用いた類似特許検索も行われており、[7] では、教師あり学習で類似特許検索を行っている。

4　BERTScore

BERTScore は、BERT モデルを用いて 2 文間の類似度を計算する一手法である [8]。入力する 2 文はトークン単位に分割され、それぞれ対応する埋め込みベクトルを抽出し、それらのペアの中からコサイン類似度の高い組み合わせを求めて文間の類似度を計算する。BERTScore には Recall, Precision および F 値のスコアがあるが、本研究では F 値を用いる。

5　提案手法

本研究での類似特許検索を行う上では、特許公報の「特許請求の範囲」に記載される各請求項の内容が似ているものを検索するという考えの基で行う。その上で、2 つの特許の請求項からそれぞれ得られた文章を、BERTScore を用いて類似度を計算していく。ここで、一般的にその特許を構成する最も重要な要素を最初の請求項とすることが多いため、本研究では最初の独立項とその従属項[1)]のみを、類似度を求めるための文章として使用する。また、特許の類似箇所を判定するため、請求項で文を分割する。

1)　請求項がいくつ書かれているかは特許公報ごとに異なるが、請求項 2 以降は、その前に記述した請求項を引用することが多々ある。前の請求項を引用しているものを従属項と呼び、引用していないものを独立項と呼ぶ。

表 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. L 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。

さらに、類似特許検索において「請求項」、「前記」などの、文章の内容を表す上で重要ではないと判断したものはストップワードとして予め削除する。ここで、一つの請求項に非常に長い文が現れることがあるが、一度に長い文章を比較すると内容は関係なくても類似度の高い単語が見つかる可能性が高くなり、検索の精度が落ちてしまう。そのため、概ね 150 文字を超える請求項については、各セグメントの長さが 100 文字前後になるようにさらに分割する。区切り方のアルゴリズムは表 1 に示す。max は 150 文字、min は 50 文字とした。以下、分割したテキストを「セグメント」と呼ぶ。

2 つの特許から抽出した各セグメントのすべての組み合わせについて BERTScore による類似度を求める。それらの組み合わせのうち、同一セグメントの重複を除き、スコアの高い上位 5 つの組を類似セグメントとし、それらの類似度の平均を特許間の類似度として求める。

さらに、類似特許検索の精度を上げるために汎用言語モデルのファインチューニングを行う。BERTScore は単語単位で類似度を計算するため、元のモデルの単語の理解に加え、特許に出てくる単語に関する知識を獲得することで特許の分野に適合した検索が期待できる。

5.1　モデルの学習方法

メインクラスごとに特許公報 5 万件をコーパスとして用意し、汎用言語モデルのファインチューニングを行う。特許は請求項だけではなく、「詳細な説明」から拒絶理由を探すこともあるため、「詳細な説明」も学習する対象として用意する。なお、学習の際のエポック数は 2 である。他の設定は初期値のままとした。

5.2　被引用文献

図 1 のように、ある特許公報 A の出願の審査において拒絶理由通知書等で特許 B が引用されている場合、A を B の「被引用文献」という。拒絶理由通知書で引用されることは、審査において A と B の類似性が指摘されていることを意味している。そこで、本研究ではこの関係を用いて B の類似特許検索において、B の全ての被引用文献を正解集合として評価実験を行う。

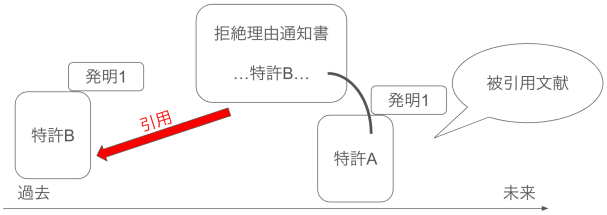

図 1　被引用文献の説明

6　特許公報収集方法

特許公報収集には特許検索システムを使用する。データを取得する際に「請求項」と「詳細な説明」を含める。特許同士の類似度を比較する際には、特許の「請求項」のみを用い、モデル学習の際には「請求項」と「詳細な説明」を用いる。

7　実験

7.1　実験概要

BERTScore と汎用言語モデルによる類似特許検索の性能評価方法について述べる。まず被引用文献が 1 件から 5 件存在する特許公報を検索対象のデータ（以下、「教師データ」と呼ぶ）として用意する。そして教師データの被引用文献を、教師データに対して検索されるべき正解データとする。さらに、教師データの筆頭 IPC とメイングループまでが同じ特許公報を約 100 件ほど用意し、それらを「検査データ」として扱う。これらの教師データと正解データ、検査データを合わせて、1 つのテストセットと呼ぶ。1 つの教師データに対し、各正解データおよび各検査データとの類似度を求め、大きい順にソートした検索結果を得る。検索結果の中で正解データがより上位にあれば高い性能を持っているといえる。本研究では各検索順位における再現率、適合率を求めて得られる PR 曲線の AUC (Area Under Curve) を検索性能評価指標として用いる。

また本研究では、G01、G02、G03、G06、H01、H02、H04、以上 7 つの IPC（セクション、メインクラス）の特許公報を用いた。これらを選んだ理由は、他の IPC（セクション、メインクラス）と比べて特許公報の件数が多く、学習に必要なデータが十分確保できるからである。

7.2 モデルごとの性能比較

まずどのモデルが一番類似特許検索に適しているかを確かめるために、予備実験として G01 の 1000 件のテストセットを用いてモデルごとの性能比較を行った。性能比較を行ったモデルは、「LINE DistilBERT」[9]、「日本語 DeBERTa V2 tiny」[10]、「日本語 DeBERTa V2 large」[11] の 3 つである。結果を図 2 に示す。

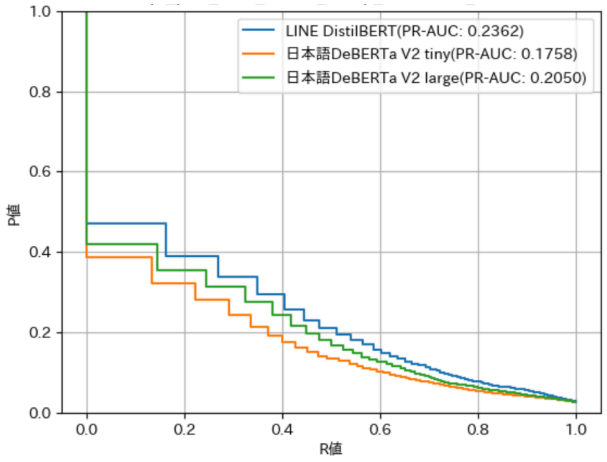

図 2 モデルごとの性能比較

一番性能が高かったものは「LINE DistilBERT」だったため、以降本モデルを性能比較に用いる。

AUC の値がどの程度の性能を示しているのかを確かめるために、AUC の値と検索結果の正解データがどの程度高い順位にあるのかの関係を表 2 に示す。代表として、「LINE DistilBERT」と「日本語 DeBERTa V2 large」の結果を示した。例えば「LINE DistilBERT」の「Recall at 5」の結果は 0.4451 だが、これは 5 位まで確認すると、全体の正解データのうち約 44.5%が存在することを示している。

7.3 学習効果とセクションごとに学習したモデルとの性能比較

学習による効果と、モデルの学習を行う際、G01 などのメインクラスのカテゴリーで学習させる方が良いのか、G や H などのセクションのカテゴリーで学習させる方が良いのかを検証する。

G01、G02、G03、G06、H01、H02、H04 の同じ IPC（セクション、メインクラス）ごとに特許公報 5 万件をコーパスとして用意し、それぞれファインチューニングを行った汎用言語モデルを用意する。また、G セクションの 4 つのメインクラスおよび、H セクションの 3 つのメインクラスのそれぞれについてコーパスを結合してファインチューニングを行った汎用言語モデルも用意する。そして、それぞれのモデルで対応するメインクラスの 1000 件のテストセットを用いて性能比較を行った結果を表 3 に示す。また、比較対象として、ファインチューニング前の元の LINE DistilBERT モデルの性能も示す。G または H のセクション全体の単位で学習したモデルは、元のモデルよりも性能が高い結果が得られたため、学習による効果があると言える。しかしメインクラスごとにばらつきはあるが、各メインクラスで学習をしたモデルが一番高い性能を示した。

表 2 AUC の値と検索結果の正解データの順位の関係

	LINE DistilBERT	日本語 DeBERTa V2 large
AUC	0.2362	0.2050
Precision at 1	0.4712	0.4195
Precision at 5	0.2585	0.2433
Precision at 20	0.1047	0.0991
Recall at 1	0.1622	0.1444
Recall at 5	0.4451	0.4189
Recall at 20	0.7212	0.6826

表 3 IPC ごとの AUC 比較

IPC	学習結果	G/H セクションモデル	元のモデル
G01（測定）	**0.2476**	0.2412	0.2362
G02（光学）	**0.2002**	0.1935	0.1833
G03（写真）	**0.2674**	0.2603	0.2537
G06（計算）	**0.1965**	0.1917	0.1809
H01（電気素子）	**0.1983**	0.1927	0.1880
H02（電力の発電）	**0.1839**	0.1697	0.1691
H04（電気通信）	**0.1950**	0.1894	0.1715

7.4 TFIDF との性能比較

提案内容が、既存手法である TFIDF よりも優れているかを確かめるために、TFIDF と LINE DistilBERT の性能を比較する。それぞれ G01 の特許 5 万件を学習したのち、G01 の 1000 件のテストセットを用いて性能比較を行った。LINE DistilBERT を学習する際のエポック数は 2 である。また比較対象として元の LINE DistilBERT の性能も示す。結果を図 3 に示す。G01 を学習した LINE DistilBERT は、TFIDF よりも高い性能を示していた。また元の LINE DistilBERT も TFIDF より高い性能を示しており、本提案内容が TFIDF よりも優れていると言える。

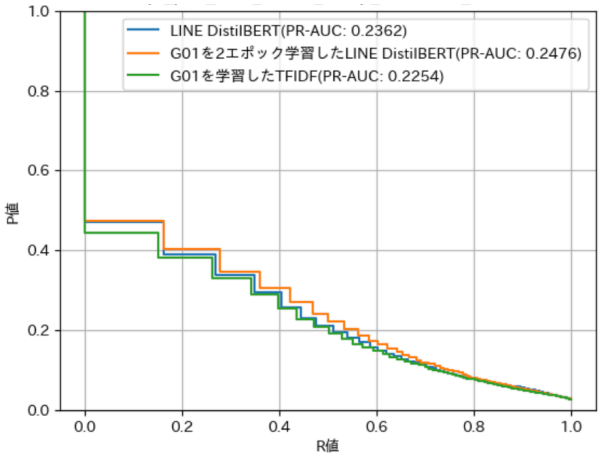

図 3　TFIDF と LINE DistilBERT の性能比較

8　検索性能の人手評価

G01 のテストセットのうち、58 件を用いて検索性能の人手評価を行った。検査データであっても、教師データの被引用文献にはならなかっただけで、実は教師データと酷似している内容であり、それゆえに順位が高くなっている可能性などを確かめることができる。

人手評価の方法について説明する。教師データと正解、検査データとの類似度を測る際、似た文章の組み合わせが最大 5 つ存在する。正解データのうち上位 2 つおよび最下位 1 つ、検査データのうち上位 3 つのデータにおいて、教師データと似た文章として挙げられた部分が、本当に似ているのかを人手で確認する。正解データが 1〜2 個の場合は、全ての正解データを対象とする。文章の類似度を評価する際は、五段階評価を用いる。五段階評価の内容を、表 4 に示す。得られた評価を、正解データ、検査データごとに平均を取る。また、検索対象の約 100 件のうち順位が 50 位以下となった正解データの平均も取る。評価の結果を表 5 に示す。結果は、正解データ、50 位以下の正解データ、検査データの順で評価平均が高いことがわかった。これは正解データにおいて、他のデータと比べて、より似たセグメントを検索できていることを示している。また、この結果だけ見ると、順位が高い検査データの評価は比較的低いため、検査データの順位が高くても、必ずしも似たセグメントを検索できているとは限らないことがわかる。

検索された特許公報のセグメントを実際に確認し、正解データなのに低い順位にあるものや、検査データなのに高い順位にある原因を確かめた。その

表 4　五段階評価の内容

	評価内容
評価 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。
評価 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。
評価 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。
評価 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。
評価 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 5　人手評価の結果

	評価平均
正解データ	**2.9209**
50 位以下の正解データ	2.5722
検査データ	1.7022

原因の一つとして、特許公報のセグメントが 10 文字から 30 文字と非常に短いものが存在するということが挙げられる。これらは請求項番号が 90 などと大きい場合に、短いセグメントが存在していると思われる。セグメントが短い場合、正しく類似度を計算できないため、極端に高い、もしくは低い類似度が算出されていたと思われる。また、セグメント数が 1 つや 2 つなど少ない場合、類似度を求めるセグメントの組み合わせが少なくなってしまうため、正しい類似度を算出できない可能性がある。正解データが 2 位や 3 位など高い順位にある時、それよりも高い位置に検査データがある場合は、内容が他の検査データと比べて似ていることが多く、さらにはそれは教師データの特許公報と出願人が同じ場合が多かったように思えた。

9　おわりに

本研究では、汎用言語モデルを使用し、教師なしで類似特許検索を行う方法を提案した。結果ファインチューニングによって検索性能は上がり、従来手法である TFIDF よりも高い性能を示した。

今後の課題としては、「発明の概要」、「発明を実施するための形態」の文章も類似度比較の対象とすることにより、性能が上がるかを検証したい。また類似度比較の際、セグメントが短い場合は比較の対象外にしたり、周辺の請求項と繋げて一つのセグメントにしたりすることで性能が上がるかを検証したい。

謝辞

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。

人手評価においては、静岡大学情報学部綱川研究室のメンバーの協力を得ました。ここに感謝の意を表します。

参考文献

[1] 特許庁.”特許分類の知識”. 発明推進協会アジア太平洋工業所有権センター.2013

[2] 松田聡.”特許申請・特許出願件数の推移”. 松田国際特許事務所.2023-08-29.https://www.matsuda-pat.com/tokkyo-nagare/kensuu.html

[3] 渡邊博之.TF・IDF 法を用いた類似レポート判定に関する一検討.2006

[4] 沈暁鵲, 森元史朗, 三原徹治. 潜在的意味検索 LSA における縮約効果. 電気関係学会九州支部連合大会.2009

[5] 加納渉, 竹内孔一.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.2022

[6] 吉田基信, 松本和幸, 吉田稔, 北研二.BERT を用いた SNS 上における攻撃的文章訂正システム. 情報処理学会第 84 回全国大会.2022

[7] 星野雄毅, 内海祥雅, 中田和秀.Contrastive Learning を利用した類似特許検索. 言語処理学会 第 29 回年次大会 発表論文集.2023

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

[9] LINE. https://huggingface.co/line-corporation/line-distilbert-base-japanese

[10] 京都大学. https://huggingface.co/ku-nlp/deberta-v2-tiny-japanese

[11] 京都大学. https://huggingface.co/ku-nlp/deberta-v2-large-japanese