令和５年度
人工知能技術等を活用した文字商標検索に関する実証的研究事業

調査報告書

令和６年３月

株式会社 NTT データ

目次

1. 事業概要

1.1. 背景

産業財産権を取り巻く環境は、ニュー・ノーマルに向けた加速的なデジタル化、研究開発や企業活動のグローバル化、中小企業・地域における知的財産戦略の重要性に対する認識の高まりによる利用拡大等、様々な観点から多様化・複雑化している。その中で、特許庁の担う商標審査業務においても、制度の複雑化や先行する他人の登録商標及び商標登録出願に係る商標（以下、「先行商標」という。）の調査における調査対象商標の件数増加等に起因する業務処理量の増加が課題の一つとなっている。こうした環境変化とそれに伴う業務量の増加に適切に対応していくためには、最新の技術を取り入れ、更なる業務の高度化・効率化を図っていくことが求められている。

そのような環境下において、特許庁では「特許庁における人工知能(AI)技術の活用に向けたアクション・プラン」に基づき、アジャイル開発により、指定商品・役務調査や先行図形商標調査に対して、人工知能技術を適用したシステム導入や改善を継続的に進めており、特に先行図形商標の調査に対しては機械学習コンペティションを通じた精度向上等を通じて、継続的に審査業務の効率化及び審査品質の確保に取り組んでいる。

他方、先行文字商標の調査に対しては人工知能技術の適用が未実施であった中、2022年に公表された「特許庁における人工知能（AI）技術の活用に向けたアクション・プラン（令和 4～8 年度版）」では、自然言語処理技術の進展に伴い、先行商標の文字要素に着目した調査（文字商標検索）業務についても、新たな技術の活用可能性がある新規事業として報告されている。

これらの背景を踏まえ、本事業を通して文字商標検索に対して人工知能技術等の適用可能性を検証することは、更なる商標審査業務の効率化・高度化にむけて有益であると考える。

1.2. 目的

前述の背景を踏まえ、本調査では商標審査業務のうち、先行文字商標の検索業務を実施する際の業務課題を解決するために適用可能性のある人工知能等の技術的な手法を調査・選定し、実際の商標情報等を用いて構築した検証用のシステムについて精度分析を行うことにより、業務課題を解決するための最適な手法について検討することを目的とする。

1.3. 事業全体の流れ

本事業の実施スケジュールを図 1.3-1 に示す。

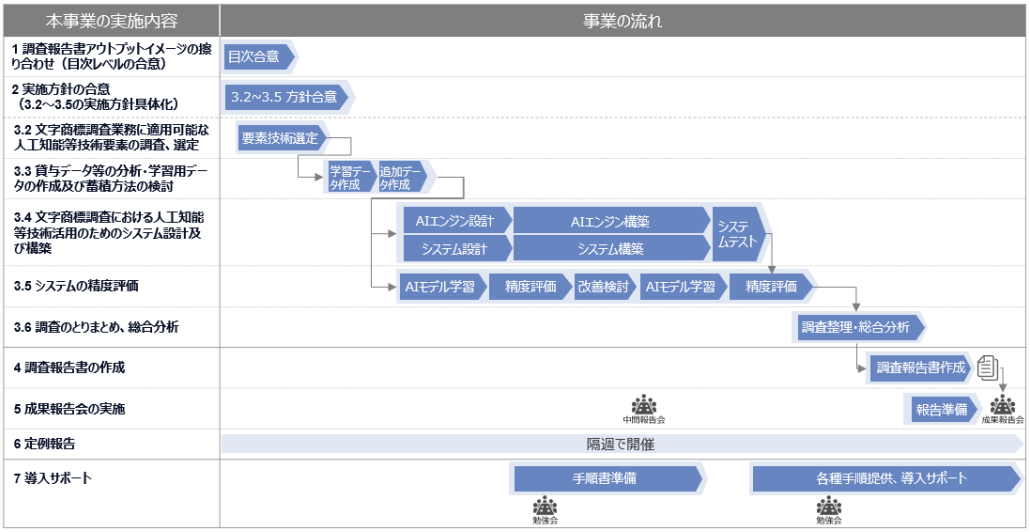


図 1.3-1 実施スケジュール

1.4. 本事業の内容

本事業で実施する調査の内容は、以下の通りである。

(1) 文字商標調査業務に適用可能な人工知能等技術要素の調査、選定

「先願に係る他人の登録商標等の調査における課題」及び「審査判断の均質性を検討するための調査における課題」における計 4 つの観点それぞれについて、適用可能な人工知能等技術を調査する。また調査後に、最も課題解決に適している技術要素を、 4 つの観点それぞれについて技術検証の対象として選定する。

(2) 貸与データ等の分析・学習用データの作成及び蓄積方法の検討

特許庁から貸与されたデータ等を分析し、学習データ、評価データとして必要なデータを抽出・作成する。さらに、人工知能等技術を用いてデータの学習を行うために学習が容易に実施できるよう、データを整備・加工する。

(3) 文字商標調査における人工知能等技術の活用のためのシステム設計及び構築

2.5 で選定した人工知能等技術要素を用いて、出願商標の情報をもとに先行文字商標を検索し、検索結果を表示するシステムの設計、構築を行う。

(4) システムの精度評価

本事業で構築するモデル及びシステムについて、評価データ等を用いた精度評

価を実施する。

(5) 調査のとりまとめ、総合分析

本事業で実施した調査や作業内容をとりまとめる。また実施結果を踏まえ、今後必要となる技術や課題を総合的に分析する。

1.5. 事業の運営体制

本事業は、先行文字商標の検索業務の業務課題解決に向けた最適な手法を検討するための実証的研究事業である。本事業の目的を達成するためには、人工知能に係る高度な技術力に加え、商標審査業務での活用を見据え、検証の結果を多角的な視点から分析する能力が求められる。

この要求に対応するために、図 1.5-1 に示す運営体制にて事業を遂行した。具体的には、モデル及びシステムを構築するクラウド環境に関する幅広い知見を持つ AI 基盤担当、人工知能技術に関する高度な専門性を持ち、かつ、特許庁業務の知識を有する調査・検証担当、特許庁システムに関する知見及び人工知能技術を用いたシステム開発に関する幅広い知識を有するシステム構築担当、それぞれの専門性を有した人員で担当チームを構成した。

また、このように多岐にわたる専門性を有したチームで構成されることから、プロジェクト運営及びリスク管理・対策を行うプロジェクト管理担当を別途設置した。

さらに、業務従事者とは別に、人工知能技術に関する学会等役員を歴任する専門家による技術監修者を設置し、事業に対する客観的かつ専門的な評価・助言を得られる体制とするとともに、必要に応じて人員補助等を担うバックアップ体制を構築した。

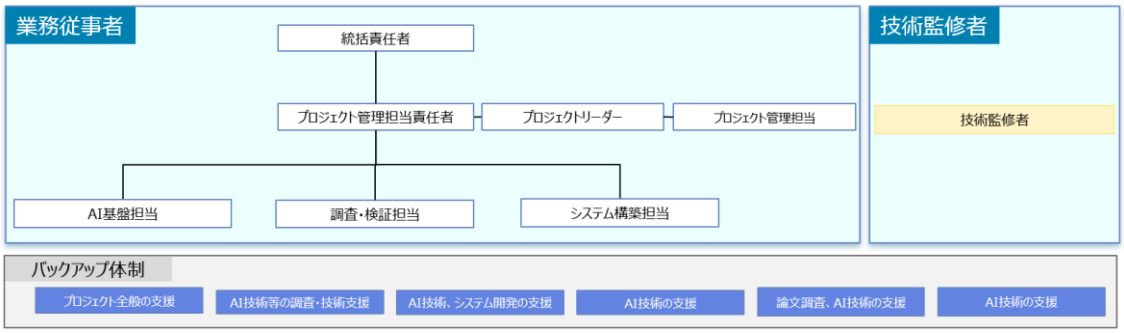


図 1.5-1 運営体制

2. 文字商標調査業務に適用可能な人工知能等技術要素の調査、選定

2.1. 調査の目的と位置づけ

文字商標調査業務における業務課題解決に向けて、検証対象となる技術要素を選定するため調査を実施する。前提となる業務課題及び検証観点を表 2.1-1 に示す。各業務課題に対して適用可能な人工知能等技術を調査し、候補となる技術要素を選定する。さらに、技術要素の候補の中から最も課題解決に適している技術要素を、各業務課題の技術検証の対象として選定する。

表 2.1-1　前提となる業務課題及び検証観点

(1) 先願に係る他人の登録商標等の調査における課題	
業務課題 1「同一又は類似する商標の検索」	過去の審査判断例を分析し、「同一又は類似」と判断される可能性の高い先行文字商標を検索結果の上位に表示すること 例１： 出願商標「Japax」（検索用商標[1]「JAPAX」、 称呼[2]「ジャパックス、ジャパエックス」） 先行商標「JapaX」（検索用商標「JAPAX」、 称呼「ジャパックス、ジャパエックス」） 例２： 出願商標「バカンテス」（検索用商標「バカンテス」、 称呼「バカンテス」） 先行商標「バカンタス」（検索用商標「バカンタス」、 称呼「バカンタス」）
業務課題 2「構成文字要素が入れ替わる商標の検索」	出願商標の検索用商標における構成文字要素と一部が一致しており、かつ、前後の構成文字要素を入れ替えたにすぎない先行文字商標を検索できるようにすること 例１： 出願商標「University of Kasumigaseki」 先行商標「Kasumigaseki University」 例２： 出願商標「ジャスミン喫茶」 先行商標「喫茶ジャスミン」

[1] 商標の構成中に含まれる文字要素を検索用に起こしたデータ

[2] 商標の構成中に含まれる平仮名、片仮名、漢字、アルファベット等の文字要素から生ずる自然な読みをカタカナで表記したデータ

業務課題3「観念が同一又は類似する商標の検索」	出願商標の構成文字から生じる観念と同一又は類似する観念が生じる先行文字商標を検索できるようにすること 例1： 出願商標「でんでんむし物語」 先行商標「かたつむり物語」 例2： 出願商標「フグの子料理」 先行商標「子フグ料理」
(2) 審査判断の均質性を検討するための調査における課題	
業務課題4「審査の均質性調査のための検索」	出願商標が、「姓氏」、「地名」又は「一般名称」の文字要素のみを組み合わせた構成からなる商標について、同一又は類似する構成からなる先行文字商標を効率的に検索できるようにすること 例1： 出願商標「佐藤商会」 先行商標「山田商会」、「TANAKA 商会」、「佐藤ストア」、「小林書店」・・・ 例2： 出願商標：「東京りんご」 先行商標：「東京メロン」、「TOKYO BEER」、「蕎麦北海道」、「コロッケ沖縄」・・・

2.2. 調査範囲と調査方法

主要な国際学会（ACL、TACL、NAACL、EMNLP 等）及び弊社のナレッジサイトを対象に、業務／技術それぞれの観点に関連するキーワードをもとに調査論文を抽出し、候補となる技術要素を選定した。（図 2.2-1）

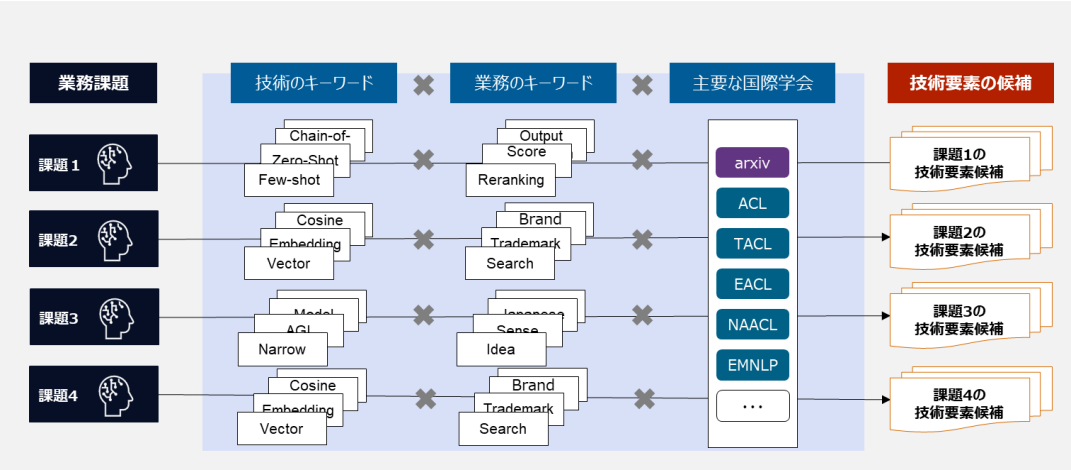


図 2.2-1 観点別キーワード及び信頼性の高い情報源を活用した技術要素の調査

2.3. 候補となる技術要素及び採用技術要素選定の観点

2.2 にて抽出した候補となる技術要素について、業務適用性及び技術的実現性の観点から検証対象とする技術要素を選定した。

2.4. 技術要素調査結果

2.4.1. 業務課題 1「同一又は類似する商標の検索」調査結果

2.2 の調査方法で抽出した候補となる技術要素を表 2.4.1-1 に示す。

表 2.4.1-1 業務課題 1 における候補となる技術要素一覧

#	技術要素	論文名	技術の概要
1	Ranking SVM	Optimizing Search Engines using Clickthrough Data	検索エンジンの精度改善を目指し、現行の検索結果とクリックログを学習データとして、分類問題で使用されるサポートベクターマシン(SVM)をランキング学習に応用した手法。
2	LSTM	Generating Sequences With Recurrent Neural Networks	テキストの埋込表現(ベクトル)を算出する際に、単語(token)の並びである時系列情報を加味し、単語間の依存関係も踏まえた特徴を学習するための技術であり、RNN の一手法。
3	Attention	Attention Is All You Need	RNN は時系列情報を内部記憶として保存するアーキテクチャであることに対し、Attention では入力情報の依存関係を外部記憶に直接参照するアーキテクチャにすることで、学習能力や処理能力を向上させた手法。
4	deep metric learning	Deep Metric Learning: A Survey	深層学習のモデルに対し、類似データは類似度が大きく、非類似データは類似度が小さくなるようなベクトル空間に埋め込むための学習を実現する手法。
5	sentence embedding	Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks	BERT に対して、Deep Metric Learning の手法を用いて追加学習することで、類似文検索等に適した BERT を構築するための手法。

2.4.2. 業務課題２「構成文字要素が入れ替わる商標の検索」調査結果

2.2 の調査方法で抽出した候補となる技術要素を表 2.4.2-1 に示す。

表 2.4.2-1 業務課題２における候補となる技術要素一覧

#	技術要素	論文名	技術の概要
1	MeCab	Applying Conditional Random Fields to Japanese Morphological Analysis	日本語の形態素解析を実現する技術、ツールであり、CRF を用いた系列ラベリングにより形態素への分割と品詞情報等の推定を実現する手法。
2	SentencePiece	SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing	言語モデルにテキストを入力する前処理において、従来は形態素解析で分割していたことに対して、SentencePiece は subword という部分文字列の単位に分割する。コーパスから統計的に計算した値を使用して、言語モデルが扱いやすい単位に自動的に分割する手法。
3	NER	NER-BERT: A Pre-trained Model for Low-Resource Entity Tagging	固有表現抽出(NER)に BERT を適用することで、固有表現抽出のタスクを学習させるための学習データが少ない場合にも、高精度を得るための手法。

2.4.3. 業務課題３「観念が同一又は類似する商標の検索」調査結果

2.2 の調査方法で抽出した候補となる技術要素を表 2.4.3-1 に示す。

表 2.4.3-1 業務課題３における候補となる技術要素一覧

#	技術要素	論文名	技術の概要
1	word2vec	Efficient Estimation of Word Representations in Vector Space	単語を、その単語の意味を含んだベクトル表現に変換するための手法。「単語の意味は周囲の単語によって形成される」という分布仮説に基づき、周辺の単語からマスクした単語を予測する教師なし学習により、ベクトル表現を学習することを実現した手法。

2	fastText	Enriching Word Vectors with Subword Information	word2vec では単語単位でテキストを処理していたが、単語の部分文字列である subword 情報も考慮したアーキテクチャとすることで、部分一致する文字の類似性や未知語の意味も考慮できるモデルを実現した手法。
3	BERT	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	Attention のみで構成されたニューラルネットワーク構造である Transformer を用いて、入力文書を双方向から処理するモデルを構築し、大量のコーパスから教師なし学習により事前学習する手法を実現することで、多くの言語処理タスクに汎用的に使用できる言語モデルの構築を実現した手法。
4	ALBERT	ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS	BERT をベースに、レイヤー間でのパラメータの共有化等を実施することで、パラメータ数を削減しつつ同等のモデルを実現した手法。
5	RoBERTa	RoBERTa: A Robustly Optimized BERT Pretraining Approach	BERT をベースに、事前学習に使用するコーパス量の増量や学習回数の増加、バッチサイズ等のハイパーパラメータの変更等を実施することで、アーキテクチャは同等のまま精度を改善した手法。
6	DeBERTa	DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION	BERT をベースに、トークン同士の位置関係を処理しやすくする等の工夫を加えることで精度向上を達成した技術。高難度のベンチマークである SuperGLUE で初めて人間を超えたスコアを記録した手法。

2.4.4. 業務課題 4「審査の均質性調査のための検索」調査結果

2.2 の調査方法で抽出した候補となる技術要素を表 2.4.4-1 に示す。業務課題 4 で用いる処理を実現する手法は業務課題 2、業務課題 3 それぞれの処理手法と同様となることが想定されており、抽出した候補となる技術要素を表 2.4.2-1 及び表

2.4.3-1 に示す。

表 2.4.4-1 業務課題 4 における候補となる技術要素一覧

#	技術要素	論文名	技術の概要
1	LLM (Large Language Model)	A Survey of Large Language Models	テキストを生成するタスクにおいては汎用的に使用可能な技術であり、LLM に入力するプロンプト（指示や質問）を工夫することによって本タスクで目的とする類似した文字要素を含んだデータについても生成できる可能性があることから採用する。

2.5. 技術検証の対象とする技術要素の選定結果

2.5.1. 業務課題 1 「同一又は類似する商標の検索」選定結果

業務課題 1 への技術適用にあたっては、「文字表記から音の類似性や関係性を考慮して類似性を判定する」必要がある。本特性と技術の有用性の観点から技術検証の対象とする技術要素を選定した結果を表 2.5.1-1 に示す。

表 2.5.1-1 業務課題 1 における技術要素の選定結果

#	技術要素	業務適用性	技術実現性	選定結果	選定理由
1	Ranking SVM	◯	◯	採用	モデルの入力とする特徴情報を設計する必要がある旧来技術であるものの、類似の観点は 15 の種別（以下、「称呼基準」という。）で定められており、その情報を明示的に拡張したモデルを構築することで、**現行業務の判断基準に沿ったモデルが構築できる可能性がある**ため採用する。
2	LSTM	◯	×	不採用	子音と母音の関係性による音の表現の特徴等を学習する上で有効性が見込める方式である。一方で、**LSTM の課題を解決した Attention の方が、多くのタスクにおいて高い精度が得られる**ことが報告されていることから、LSTM については不採用とする。
3	Attention	◯	◯	採用	子音と母音の関係性による音の表現の特徴等を学習する上で有効性が見込める方式であり、**近年の自然言語処理タスクにおいて多く使われており有効性が報告されている**ため、本検証においても採用する。

4	Deep Metric Learning	◯	◯	採用	**深層学習モデルを検索やランキングのタスクおいて、精度と性能の面で有効性が報告されており**、本タスクにおいても有効性が見込めるため採用する。
5	Sentence Embedding	×	◯	不採用	本手法が適用されるタスクは、単語や文脈の意味を捉えた類似文検索に用いられており、**音の関係性や類似性を捉えることに利用した事例は確認できておらず有効性が未知数である**ため不採用とする。

2.5.2. 業務課題２「構成文字要素が入れ替わる商標の検索」選定結果

業務課題２への技術適用にあたっては、「商標を構成文字要素に分割した上で、位置の入れ替えを考慮したマッチングを行う」必要がある。本特性と技術の有用性の観点から技術検証の対象とする技術要素を選定した結果を表 2.5.2-1 に示す。

表 2.5.2-1 業務課題２における技術要素の選定結果

#	技術要素	業務適用性	技術実現性	選定結果	選定理由
1	MeCab	×	×	不採用	**構成文字要素と形態素の単位は一致しないことが想定され、所望する出力は得られない**ことから不採用とする。
2	SentencePiece	×	×	不採用	**構成文字要素と subword の単位は一致しないことが想定され、所望する出力は得られない**ことから不採用とする。

3	NER	◯	◯	採用	形態素解析と異なりモデルを学習するコストは発生するが、**抽出する固有表現の単位は学習データの形式を調整することでカスタマイズ可能であり、構成文字要素の単位に分割できる可能性がある**ことから採用する。

2.5.3. 業務課題３「観念が同一又は類似する商標の検索」選定結果

業務課題３への技術適用にあたっては、「単語間の観念上の意味を考慮して類似性を判定する」必要がある。本特性と技術の有用性の観点から技術検証の対象とする技術要素を選定した結果を表 2.5.3-1 に示す。

表 2.5.3-1 業務課題３における技術要素の選定結果

#	技術要素	業務適用性	技術実現性	選定結果	選定理由
1	word2vec	◯	×	不採用	事前にコーパスから単語の意味を学習し、単語間の類似度を算出できる方式であり、本タスクにおいても有効性が見込める。一方で、**word2vec を改良した fastText の方が一般的に高い精度が得られることが報告されている**ため、word2vec については不採用とする。
2	fastText	◯	◯	採用	**事前にコーパスから単語の意味を学習し、単語間の類似度を算出できる方式であり、本タスクにおいても有効性が見込める**ため採用する。

3	BERT	◯	◯	採用	**事前にコーパスから単語の意味を文脈も考慮して学習し、単語間の類似度を算出できる方式であり、本タスクにおいても有効性が見込める**ため採用する。
4	ALBERT	×	◯	不採用	本検証では入力となる文字商標のテキストはテキスト長が短く、データ件数も一定量であり、BERT を使用した場合でもリソース面での問題は発生する見込みはなく**モデルサイズ削減が必要になることは想定されない**ため、不採用とする。
5	RoBERTa	◯	×	不採用	後述する DeBERTa と比較すると、**DeBERTa の方が様々なベンチマークで高精度となっている**ため、不採用とする。
6	DeBERTa	◯	◯	採用	BERT を改良した手法であり、**BERT 同様に単語間の類似度を算出可能な方式として本タスクにおいても有効性が見込める**ため採用する。ただし、処理に必要となるリソースが BERT と比較して高くなるため、精度と性能の両観点で BERT と比較しながら検証を進める。

2.5.4. 業務課題４「審査の均質性調査のための検索」選定結果

業務課題４への技術適用にあたっては、「商標を構成文字要素に分割した上で、位置の入れ替えを考慮したマッチングを行う」「単語間の観念上の意味を考慮して類似性を判定する」必要がある。本特性と技術の有用性の観点から技術検証の対象とする技術要素を選定した結果を表 2.5.4-1 に示す。2.4.4 でも述べた通り、ここ

で採用した技術を実現させるための処理手法は業務課題２、業務課題３それぞれの処理手法と同様であるため、技術要素の選定結果は表 2.5.2-1 及び表 2.5.3-1 に示す。

表 2.5.4-1 業務課題４における技術要素の選定結果

#	技術要素	業務適用性	技術実現性	選定結果	選定理由
1	LLM	○	○	採用	テキストを生成するタスクにおいては汎用的に使用可能な技術であり、**LLM に入力するプロンプトを工夫することによって本タスクで目的とする類似した文字要素を含んだデータについても生成できる可能性がある**ことから採用する。

2.6. 選定した技術要素を利用した AI モデルの各手法

選定した技術を利用した AI モデルの各手法について表 2.6.1-1 に示す。

表 2.6.1-1 選定した技術要素を利用した AI モデルの各手法

対象課題	モデルの手法	前処理・特徴量設計の方法	モデル内部で使用する技術	学習処理
業務課題1	1-(1)	人手により設計・処理構築	Ranking SVM	Ranking SVM
	1-(2)	ローマ字表記に変換する前処理のみ(深層学習手法のため、特徴量設計は不要)	Attention	deep metric learning
業務課題2	2-(1)	形態素解析	NER	NER
	2-(2)	Tokenizer(言語モデル(BERT)のツールの中で実装されている処理を使用)	NER-BERT	NER-BERT
	2-(3)	(ルール処理のため前処理なし)	共通部分文字列の検出	(ルール処理のため学習なし)

業務課題3	3-(1)	形態素解析または NER	事前知識を活用したルール処理	(ルール処理のため学習なし)
	3-(2)	NER または NER-BERT	fastText	なし
	3-(3)	Tokenizer(言語モデル(BERT)のツールの中で実装されている処理を使用)	BERT	deep metric learning
	3-(4)	Tokenizer(言語モデル(DeBERTa)のツールの中で実装されている処理を使用)	DeBERTa	deep metric learning
業務課題4	4-(1)	NER または NER-BERT	事前知識を活用したルール処理	(ルール処理のため学習なし)
	4-(2)	NER または NER-BERT	fastText	なし
	4-(3)	Tokenizer(言語モデル(BERT)のツールの中で実装されている処理を使用)	BERT	deep metric learning
	4-(4)	Tokenizer(言語モデル(DeBERTa)のツールの中で実装されている処理を使用)	DeBERTa	deep metric learning

3. 貸与データ等の分析・学習用データの作成及び蓄積方法の検討

3.1. 貸与データ等の分析

特許庁からの貸与データについて、モデル構築にあたって学習に利用可能なデータであるか分析した。

3.1.1. 分析対象

特許庁からの貸与データ、及び外部サイトから得たデータを分析対象とし、データが持つ特徴が検証予定の手法に適したデータであるか、また業務課題１～４に対して検証に必要なデータが確保できるのか確認した。詳細なデータ情報を以下の表 3.1.1-1 に示す。

表 3.1.1-1 分析対象としたデータ

対応課題	データカテゴリ	詳細
全課題共通	特許情報標準データ	特許庁が保有する特許・実用新案・意匠・商標に関する書誌・経過情報等について、情報の更新日単位でまとめられたバルクデータのうち、特許庁アジャイルシステムで利用しているデータを利用した。
課題１	先願に係る他人の登録商標等の調査における文字商標検索用データセット	過去に審査官が類似と判断した情報から、特許庁が作成した審査判断のデータで、どの称呼基準に該当するか付与されている。
課題２		審査または審判において類似、非類似と判断された事例を元に作成した検索クエリと正解のペアからなるデータで、特許庁の担当者が追加で作成したデータも含む。
課題３		審査または審判において類似、非類似と判断された事例を元に作成した検索クエリと正解のペアからなるデータで、特許庁の担当者が、観念の類似性が争点となったものと判断したデータ。
課題４	審査判断の均質性を検討するための調査における文字商標検索用データセット	・氏（漢字・読み方）リストデータ ・「姓氏」、「地名」、「商品役務の一般名称」及び各組合せや「業種名」との組み合わせの審決判断データ ・地名リストデータ ・商品役務の一般名称リストデータ

業務課題4における「地名リスト」データは政府統計の総合窓口の外部サイトより取得した。また、「商品役務の一般名称リスト」データは特許庁サイトの「類似商品・役務審査基準」より取得した。このデータは、商標登録出願の指定商品又は指定役務が他人の商標登録の指定商品又は指定役務と類似関係にあるか否かを審査するにあたり、審査官の統一的基準として用いられているものである。

3.1.2. 分析手法

特許庁からの貸与データから検証に必要であるデータ量が確保できるのか確認するためにデータ件数を分析した。その際に学習、評価の際に重複するデータは繰り返し使用ができないために重複を排除したうえでデータ件数を分析した。以下の表 3.1.2-1 にそれぞれの業務課題におけるデータの分析観点を示す。

表 3.1.2-1 それぞれの業務課題における貸与データの分析観点

業務課題	詳細
課題1	・検索クエリとなる称呼と正解となる称呼が重複しているデータを排除したデータ数 ・(上記重複を排除した上で) 称呼基準ごとのデータ数、正解データ数
課題2	・重複排除後の総データ数 ・NERで単語に分割できる可能性がある商標か
課題3	・重複排除後の総データ数
課題4	・審決判断等データ_総データ数 ・審決判断等データ_観点ごとのデータ数 ・氏リスト_総データ数 ・地名リストデータ_総データ数 ・商品役務の一般名称リストデータ_総データ数

3.1.3. 分析結果

【業務課題1】

特許庁からの貸与データにはクエリとなる商標の称呼と正解となる商標の称呼が完全に一致しているデータが含まれていることが分かった。完全一致のペアは学習データとして有効性がないことから完全一致のデータは排除した。

以上の分析の結果、モデルの学習に利用可能なデータが **18,208 件**であったため、検証に必要であるデータ量が確保できたことを確認した。

【業務課題２】

分析の結果、モデルの学習に利用可能なデータが **117 件**であったため、検証に必要であるデータ量が確保できたことを確認した。商標の表記方法には以下の表 3.1.3-1 に示すようないくつかのパターンがあることが判明した。

表 3.1.3-1 業務課題２における貸与データ分析結果

パターン
アルファベット（例：TECHTOOL）
漢字（例：開運大吉）
ひらがな（例：じゅく）
カタカナ（例：ツールテック）
複数の文字種別の組み合わせ（例：NTT ドコモ）

業務課題 2 の検索を実現するアルゴリズムを検討するうえで、構成文字要素への分割が可能なデータなのかを分析した結果、分割が可能な場合の構成要素は各構成要素が意味を持つ日本語や英語の単語等、定義に基づいた文字列の場合であることが分かった。

一方で構成要素が意味を持たない単一の文字や略語であるデータも含まれており、それらは明確な分割基準がないために分割が難しいことが想定された。そのため、構成文字要素に分割したうえで類似性を算出する、NER を使った手法の有効性は低いと考えられた。

これを踏まえて本事業においては NER を不採用とした。そのため構成要素への事前の明示的な分割は実施せず、検索機能は文字単位で入れ替えが発生しているかを検出し、スコア化するルールベース処理にて実装することとした。ルールベース処理では学習が不要で、学習データを準備する必要性がなくなったために抽出データは全て評価データとして用いることとした。

表 3.1.3-2 構成文字要素に分割したうえで類似性を算出する方法

検索クエリ	正解	NER を使った手法の対応可否
開運風水	風水開運	○
STAR JET	JETSTAR	○
雷神	神雷	×：雷神、神雷ともに 1 単語として処理され、構成文字要素の分割が難しい
ピスカル	カルピス	×：カルピスは 1 単語であり、構成文字要素の分割が難しい

HOTEL UNO	UNOHOTEL	×：UNO が普通名詞ではないため、構成文字要素の分割が難しい

【業務課題３】

分析の結果、モデルの学習に利用可能なデータが **512 件**であったため、検証に必要であるデータ量が確保できたことを確認した。一方で、一部重複するデータが含まれていたため、重複データは排除した。

【業務課題４】

分析の結果、モデルの学習に利用可能なデータが **17,838 件**であったため、検証に必要であるデータ量が確保できたことを確認した。一方で、一部重複するデータが含まれていたため、重複データは排除した。

3.2. 貸与データを用いた学習データ等の抽出・作成

3.2.1. 学習データ等の抽出対象

3.1 で分析対象としたデータ（表 3.1.1-1）を対象とし、後のモデル検証において学習データ及び評価データに用いるデータを抽出した。また、検証の際、検索対象の候補データとなる検証用 DB のデータは特許庁アジャイルシステム DB に登録されている特許情報標準データから抽出した。検証用 DB の構成は 5.1.4 の図 5.1-2 に示す。

3.2.2. 抽出手法

抽出対象データからモデル学習に有効であるデータを抽出する。抽出対象となったデータには一部重複するものが含まれているため、重複したデータは除去した。抽出ができたデータは、定量的に検証結果の評価が可能とするために評価データは 100 件を基準とし、この基準を「評価に必要なデータ数」と考えることとした。

また、業務課題１においては称呼基準が、業務課題４においては構成の分類がそれぞれ複数存在したため、その基準や観点ごとに評価や分析を実施し、データを抽出した。

ここで抽出ができたデータを抽出データと呼ぶこととする。

3.2.3. 抽出結果

貸与データから抽出することができた抽出データの件数を以下に示す。

【業務課題１】

抽出した全データ数は重複したものを除き、**18,208** 件となった。各称呼基準における抽出データを表 3.2.3-1 に示す。

その中で、**称呼基準 10、13~15** については受領したデータ件数が極端に少ないため、評価データの準備が難しく、定量的な評価が難しいために**検証の対象から外した。**

また、称呼基準ごとに分析した時、称呼基準によっては抽出データ件数が極端に少ない場合が存在した。そのため学習データのパターンの不足が考えられたため、3.3 にて学習データの拡張を実施した。

表 3.2.3-1 業務課題 1 におけるデータ抽出結果

	抽出データ件数	検証対象
全件	**18,208 件**	—
称呼基準 1	11,823 件	検証対象
称呼基準 2	520 件	検証対象
称呼基準 3	424 件	検証対象
称呼基準 4	1,052 件	検証対象
称呼基準 5	854 件	検証対象
称呼基準 6	663 件	検証対象
称呼基準 7	691 件	検証対象
称呼基準 8	1,723 件	検証対象
称呼基準 9	170 件	検証対象
称呼基準 10	5 件	検証対象外
称呼基準 11	225 件	検証対象
称呼基準 12	46 件	検証対象
称呼基準 13	0 件	検証対象外
称呼基準 14	9 件	検証対象外
称呼基準 15	3 件	検証対象外

【業務課題２】

抽出データ数は **117** 件であった。業務課題２についてはルールベース処理にて検証を実施するため、抽出したデータは全て評価データとして用いることができることから検証に必要なデータ数を確保できた。

【業務課題３】

抽出データ数は**512**件であり、検証に必要なデータ数を確保することができた。

データを一定量抽出することができたことと、利用したBERT、DeBERTaの各モデルは事前学習済みモデルを利用することから、ファインチューニング時の学習データ件数が限られていても一定の精度が期待できるために本検証においては**データ拡張を実施せず**にモデル検証を実施することとした。

【業務課題４】

抽出データ数は**17,838**件であった。表 3.2.3-2 に構成の分類ごとのデータ数を示す。

表 3.2.3-2 業務課題 4 におけるデータ抽出結果

構成の分類	抽出データ件数
全件	**17,838 件**
商品役務の一般名称	15,344 件
地名	971 件
地名+商品役務の一般名称	865 件
地名+業種名	188 件
氏	223 件
氏+商品役務の一般名称	242 件
氏+業種名	406 件

業務課題３と同様に、データを一定量抽出することができたため、**データ拡張を実施せず**にモデル検証を実施することとした。

3.3. 学習データ拡張

精度の高いモデルを構築することを目的として、3.2 にて学習データの件数が不足していると判断された業務課題について、学習データの拡張を行った。データ拡張の手法には、品質の安定した学習データを一定量以上生成するために本事業ではルールベースによるデータ拡張を採用した。また、ルールベースでのデータ拡張に加え、今後のさらなる精度向上や継続的な精度改善の可能性を検討するため、生成 AI を用いた自動データ拡張についても検証を行った。

3.3.1. 実施手順

学習データ拡張にあたっては、業務課題ごとに必要な学習データが異なることから業務課題ごとに異なる手順でデータ拡張を行う。業務課題ごとの実施手順を以下に示す。

【業務課題 1】

業務課題 1 については、貸与データに対してルールベースにて音の削除や差し替えを実施した。称呼基準ごとにルールが詳細に定められているため、各称呼基準に沿ったスクリプトを作成し、貸与データからランダムに選んだ 100 件に対してルールに合致するデータを生成した。ただし、3.2.3 に記載の通り称呼基準 10、及び 13～15 については対象外とした。

【業務課題 2】

業務課題 2 については、学習が不要なアプローチを採用しているため、学習データ拡張を実施していない。

【業務課題 3】

業務課題 3 については、3.2 にて学習データの拡張は不要と判断しているが、ルールベースでの拡張が困難である観念や意味が類似する商標の学習データについて、生成 AI による学習データ拡張の実現可能性を検証した。検証に利用した生成 AI のモデルを表 3.3.1-1 に示す。

表 3.3.1-1 学習データ拡張に採用した生成 AI

モデル名	gpt-4 (1106-preview) GPT-4 Turbo プレビュー
トークン数（最大）	入力: 128,000 出力: 4,096
学習データ	2023 年 4 月までの情報を事前学習済

生成 AI を用いて学習データを拡張するためには、プロンプトと呼ばれる指示文を入力し出力する必要がある。業務課題 3 の学習データ拡張に利用したプロンプトを表 3.3.1-2 に示す。

表 3.3.1-2 業務課題 3 における学習データ拡張検証手順

#	方針	プロンプト	備考
3-①	ある商標を入力文として、類似文（類似の商標）を生成する	以下の制約条件をもとに、入力文の各単語に対して意味が類似した文を 1 件書いてください。 #制約条件:1.クォーテーションマークを入れない;2.1 行に 1 件出力する;3.出力の長さは入力文と同程度 #入力文:{商標}	【商標】 受領した学習データを利用。

【業務課題 4 】

業務課題 4 については、検証には事前学習済みのモデルを用いることもあり、3.2 における抽出データ量は十分と判断した。追加の検証として将来的に LLM による学習データの拡充によりモデルの精度向上の実現が可能なのかを確認するため、本検証時点では商用利用向けの LLM の学習データとして利用することはできないものの、LLM(ChatGPT)により生成したデータを学習データとすることで観念・意味的な類似商標の学習データ拡張の実現可能性を検証した。検証の具体的な手順は表 3.3.1-3 に示す。

表 3.3.1-3 業務課題 4 における学習データ拡張検証手順

#	方針	プロンプト案	入力データ
4-①	ある商標を入力文として、類似文（類似の商標）を生成する	以下の制約条件をもとに、入力文を構成する各単語にカテゴリが類似した単語を 10 件書いてください。 #制約条件:1.クォーテーションマークを入れない;2.1 行に 1 件出力する;3.出力の長さは入力文と同程度 #入力文:{商標}	【商標】 受領した学習データを利用。

4-②	ある商標とその構成文字要素の分類を入力として、分類と入力文に類似する類似文（類似の商標）を生成する	以下の入力文は、以下のカテゴリの単語の組み合わせです。以下の制約条件をもとに、入力文とカテゴリに類似した単語を10件書いてください。 #制約条件:1.クォーテーションマークを入れない;2.1行に1件出力する;3.出力の長さは入力文と同程度 #入力文:{商標} #カテゴリ:{構成文字要素の分類}	【商標】 受領した学習データを利用。 【構成文字要素の分類】 以下をカテゴリとして利用。 - 商品役務の一般名称 - 地名 - 地名＋業種名 - 地名＋商品役務の一般名称 - 氏 - 氏＋業種名 - 氏＋商品役務の一般名称
4-③	構成文字要素の分類を入力に、同様のカテゴリの商標を生成する	以下の制約条件に従って、以下のカテゴリの商標を10件作成してください。 #制約条件:1.クォーテーションマークを入れない;2.1行に1件出力する #カテゴリ:{構成文字要素の分類}	【商標】 受領した学習データを利用。 【構成文字要素の分類】 以下をカテゴリとして利用。 - 商品役務の一般名称 - 地名 - 地名＋業種名 - 地名＋商品役務の一般名称 - 氏 - 氏＋業種名 - 氏＋商品役務の一般名称
4-④	ある商標を入力に、その商標を構成する単語のカテゴリを推定し、推定したカテゴリの単語を含む商標を生成する	以下の入力文を構成する各単語のカテゴリを特定してください。特定したカテゴリで、以下の制約条件に従って商標を10件作成してください。 #制約条件:1.クォーテーションマークを入れない;2.1行に1件出力する;3.出力の長さは入力文と同程度;4.生成した商標のみを出力する #入力文:{商標} #出力:{生成した商標}	【商標】 受領した学習データを利用。

3.3.2. 追加学習データ作成結果

3.3.1 の手順で追加学習データを作成した結果を以下に示す。

【業務課題 1】

業務課題 1 について、追加学習データを全 1000 件作成した。データ作成結果は表 3.3.2-1 に示す。

表 3.3.2-1 業務課題 1 における学習データ作成結果

	貸与データ件数	追加作成データ件数	データ件数
全件	**18,208 件**	**1,000 件**	**19,208 件**
称呼基準 1	11,823 件	0 件	11,823 件
称呼基準 2	520 件	100 件	620 件
称呼基準 3	424 件	100 件	524 件
称呼基準 4	1,052 件	100 件	1,152 件
称呼基準 5	854 件	100 件	954 件
称呼基準 6	663 件	100 件	763 件
称呼基準 7	691 件	100 件	791 件
称呼基準 8	1,723 件	100 件	1,823 件
称呼基準 9	170 件	100 件	270 件
称呼基準 10	検証対象外		
称呼基準 11	225 件	100 件	325 件
称呼基準 12	46 件	100 件	146 件
称呼基準 13	検証対象外		
称呼基準 14	検証対象外		
称呼基準 15	検証対象外		

具体的な作成データ例を以下表 3.3.2-2 に示す。

表 3.3.2-2 業務課題 1 における作成データ例

#	称呼基準	元データ	生成データ
例 1	2	ハナショウ	ハハナショウ
例 2	3	アイフィッティング	アヒフィッティウグ
例 3	4	レーレス	レーレ
例 4	5	キネティクス	キメティク

例 5	6	ブルーオーシャンズ	フルーオージャンス
例 6	7	ホームワークス	ホームアークス
例 7	8	コーライジンリキ	コーライキュンリキ
例 8	9	エキスパンド	エキスパンズオ
例 9	11	コミュテク	コミュミュテク
例 10	12	ドクターマン	ダクターマン

3.3.3. 追加学習データ作成検証結果

課題 3、4 における、LLM による学習データ拡張の実現可能性検証の結果を以下に示す。

【業務課題 3】

業務課題 3 について、学習データを 100 件拡張した。「意味が類似したデータを出力できているか」を基準に、入力商標と出力データの意味が類似していれば〇、一部類似していれば△、類似していなければ×と作業者の目視により判定している。データ集計結果は表 3.3.3-1 に示す。

81%の出力において意味が類似するデータであったものの、一般的に特定の意味を持たない入力商標に対しては意味を捉えることができず、単に文字列として類似しているデータが出力された。また、同一の入力商標に対する出力の中でも一部意味が類似しているとは言い難い出力も存在した。

表 3.3.3-1 業務課題 3 における学習データ作成検証結果

#	作成件数	評価結果		
		〇	△	×
3-①	100 件	81.0%	12.0%	7.0%

意味が類似するデータを出力できた例を表 3.3.3-2 に示す。

例 1 では、入力商標が「日本 IP 研究所」で、IP(Intellectual Property)に意味が類似する文字として、「特許」や「著作権」等を含むデータが出力された。

例 2 では、入力商標が「ROI DES ROIS」で ROIS に対して、王子を意味する PRINCE、公爵を意味する DUC 等の単語が出力された。入力商標がフランス語であるため、出力もフランス語になっていることが分かる。

表 3.3.3-2 業務課題 3 において意味が類似するデータを出力できた例

例 1	例 2
入力商標：日本 IP 研究所	入力商標：ROI DES ROIS

生成データ	生成データ
日本特許研究室	ROI DES PRINCES
東京知的財産センター	ROI DES DUCS
日本著作権研究会	ROI DES COMTES
大阪発明研究所	ROI DES BARONS
日本商標研究所	ROI DES MARQUIS
神戸特許研究室	ROI DES NOBLES
札幌知財研究センター	ROI DES SEIGNEURS
福岡特許研究所	ROI DES CHEVALIERS
名古屋著作権研究会	ROI DES VICOMTES
広島商標研究所	ROI DES EMPEREURS

意味が類似するデータを出力できなかった例を、表 3.3.3-3 に示す。

例 1 では、入力商標が「ARAZINE」だが、ARAZINE の意味を捉えきれず、文字列が類似したデータが出力された。

例 2 では、入力商標が「MAHO」だが、MAHO(魔法)の意味は捉えられているものの、「オモテ」「ケイ」「ミコト」等、意味が類似していないと思われるデータも出力された。

表 3.3.3-3 業務課題 3 において意味が類似するデータを出力できなかった例

例 1	例 2
入力商標：ARAZINE	入力商標：MAHO
生成データ	生成データ
ARAMIN	ジュツ
ARADIN	マナ
ARAZON	セト
ARAZIM	カンナギ
ARATIN	ケミ
ARAVIN	マボロシ
ARAZUN	オモテ
ARAXIN	ケイ
ARAKIN	ミコト
ARALIN	マヤカシ

【業務課題４】

業務課題４について、学習データを４つの手順で合計 370 件拡張した。

4-①の手順では、学習データを 100 件拡張した。「構成が類似したデータを出力できているか」を基準に、入力商標及びその構成文字要素の分類と出力データが類似していれば〇、一部類似していれば△、類似していなければ×と作業者の目視により判定している。4-①のデータ集計結果は表 3.3.3-4 に示す。

4-①の手順では、存在しない名称データを生成した事例（△列）以外は、すべて構成が類似するデータを出力した。

表 3.3.3-4 業務課題４における学習データ作成検証結果①

#	作成件数	〇	△	×
4-①	100 件	99.0%	1.0%	0.0%

構成が類似するデータを出力できた例を表 3.3.3-5 に示す。

例１では、入力商標が「軽井沢不動産協会」でカテゴリが「地名＋業種名」のところ、地名＋業種名が出力されている。また、業種名の中でも意味が近いもの（住宅連盟、開発連合、建設協議会等）が出力されている。

例２では、入力商標が「オーヨシ」でカテゴリが「氏」のところ、氏のうちカタカナ表記かつ長音記号を含むデータが出力されている。「オー」から始まるデータが多く、入力商標の表記が大きく影響する可能性が考えられる。

表 3.3.3-5 業務課題４の手順 4-①において構成が類似するデータを出力できた例

例１	例２
入力商標：軽井沢不動産協会 カテゴリ：地名＋業種名	入力商標：オーヨシ カテゴリ：氏
生成データ	生成データ
南青山住宅連盟	オーマサ
白馬土地組合	オーヤマ
箱根建設協議会	オーワダ
湘南宅地同盟	オータニ
那須開発連合	オーモリ
奥多摩不動産協会	オーカワ
浅間山地所協議	オーハマ
葉山住宅団体	オーナカ
伊豆土地協会	オーハラ
日光宅地組織	オーサト

部分的に構成が類似しないデータを出力した例を表 3.3.3-6 に示す。

入力商標が「MEPIRIZOLE」でカテゴリが「商品役務の一般名称」のところ、存在しない一般名称である「SIMETICONAZOLE」を出力した。これは存在するSIMECONAZOLE（シメコナゾール）と、入力と構成は似ているが存在しない一般名称を LLM が生成する場合があることが分かった。

表 3.3.3-6 業務課題 4 の手順 4-①において部分的に構成が類似しないデータを出力した例

例
入力商標：MEPIRIZOLE カテゴリ ：商品役務の一般名称
生成データ
CEFACLOR
FLUCONAZOLE
ITRACONAZOLE
KETOCONAZOLE
NAFTIFINE
PANTOPRAZOLE
RABEPRAZOLE
SIMETICONAZOLE
TERBINAFINE
VORICONAZOLE

4-②の手順では、学習データを 100 件拡張した。「構成が類似したデータを出力できているか」を基準に、入力商標及びその構成文字要素の分類と出力データが類似していれば〇、一部類似していれば△、類似していなければ×と作業者の目視により判定している。4-②のデータ集計結果は表 3.3.3-7 に示す。

4-②の手順において、文字構成が類似しているがカテゴリが異なる事例(×列)以外は、すべて構成が類似するデータを出力した。

表 3.3.3-7 業務課題 4 における学習データ作成検証結果②

#	作成件数	〇	△	×
4-②	100 件	99.0%	0.0%	1.0%

構成が類似するデータを出力できた例を表 3.3.3-8 に示す。

例 1 では、入力商標が「MEPIRIZOLE」でカテゴリが「商品役務の一般名称」

のところ、医薬・化学物質系が出力された。

例 2 では、入力商標が「上海小篭包」でカテゴリが「地名＋商品役務の一般名称」のところ、地名＋食品名が出力された。

例 3 では、入力商標が「オーヨシ」でカテゴリが「氏」のところ、4-①と異なり、長音を含まない氏が出力された。

表 3.3.3-8 業務課題 4 の手順 4-②において構成が類似するデータを出力できた例

例 1	例 2	例 3
入力商標：MEPIRIZOLE カテゴリ：商品役務の一般名称	入力商標：上海小篭包 カテゴリ：地名＋商品役務の一般名称	入力商標：オーヨシ カテゴリ：氏
生成データ	生成データ	生成データ
METRONIDAZOLE	北京ダック	タナカ
AMOXICILLIN	長崎カステラ	ササキ
CIPROFLOXACIN	富士宮焼きそば	コバヤシ
CLINDAMYCIN	博多ラーメン	ナカムラ
AZITHROMYCIN	仙台牛タン	ワタナベ
TETRACYCLINE	京都抹茶	イシカワ
CEPHALEXIN	大阪たこ焼き	ヨシダ
LEVOTHYROXINE	札幌スープカレー	サイトウ
PENICILLIN	金沢カレー	フジモト
IBUPROFEN	横浜シューマイ	ミヤザキ

4-③の手順では、学習データを 70 件拡張した。「構成が類似したデータを出力できているか」を基準に、入力商標及びその構成文字要素の分類と出力データが類似していれば〇、一部類似していれば△、類似していなければ×と作業者の目視により判定している。4-③のデータ集計結果は表 3.3.3-9 に示す。

4-③の手順において、入力に商標を使用しなかったものの、構成が類似するデータを 84.3%出力した。

入力に商標を使用せず、カテゴリのみ入力したことで、商標の内容に引きずられずに、指定したカテゴリに属する幅広いデータを生成できたと考えられる。

表 3.3.3-9 業務課題 4 における学習データ作成検証結果③

#	作成件数	〇	△	×
4-③	70 件	84.3%	0.0%	15.7%

構成が類似するデータを出力できた例を表 3.3.3-10 に示す。

例 1 では、カテゴリが「氏＋商品役務の一般名称」のところ、貸与データに示されていた商品役務の一般名称とはやや異なり、世間一般で使われていると思われる名称に近い出力となった。

例 2 では、カテゴリが「氏＋業種名」のところ、氏＋業種名が出力されており、多岐に渡る業種のデータを生成できることを確認した。

表 3.3.3-10 業務課題 4 の手順 4-③において構成が類似するデータを出力できた例

例 1	例 2
カテゴリ：氏＋商品役務の一般名称	カテゴリ：氏＋業種名
生成データ	生成データ
佐藤のスマートウォッチ	山田製麺所
伊藤の有機野菜宅配	佐藤電機製作
山本のハイブリッド自転車	鈴木自動車工業
鈴木のヘルシースムージー	田中建設コンサルタント
高橋のスキンケアラボ	伊藤薬品研究所
田中のエコクリーニング	渡辺アパレルデザイン
渡辺のアート教室	小林農園運営
伊東の旅行プランニング	加藤食品加工
小林のフィットネスジム	斎藤情報システムズ
加藤のペットシッターサービス	中村金属工芸

構成が類似するデータを出力できなかった例を表 3.3.3-11 に示す。

例 1 では、カテゴリが「地名＋業種名」のところ、「リオデジャネイロカーニバルズ」という地名＋業種名ではないデータが出力された。

例 2 では、カテゴリが「氏」のところ、「氏」を「名字」という意味ではなく、家系や親族集団のような意味として解釈したと思われる出力となった。

表 3.3.3-11 業務課題 4 の手順 4-③において構成が類似するデータを出力できなかった例

例 1	例 2
カテゴリ：地名＋業種名	カテゴリ：氏
生成データ	生成データ
東京エレクトロニクス	伯爵の秘密
パリファッションズ	公爵の選択
ロンドンバイオテック	王族の恩恵
ニューヨークファイナンス	貴族の紋章
ケープタウンエナジー	領主の宝
シドニーマリンズ	継承者の誉れ
サンフランシスコソフトウェア	侯爵の風格
ベルリンアートコレクティブ	騎士団の誓い
モスクワマシナリーズ	名家の系譜
リオデジャネイロカーニバルズ	族長の智慧

4-④の手順では、学習データを 100 件拡張した。「構成が類似したデータを出力できているか」を基準に、入力商標及びその構成文字要素の分類と出力データが類似していれば〇、一部類似していれば△、類似していなければ×と作業者の目視により判定している。4-④のデータ集計結果は表 3.3.3-12 に示す。

4-④の手順において、入力としてカテゴリを指定せず、入力商標からカテゴリを推定した上でデータを出力し、構成が類似するデータを 91%出力した。プロンプトの影響により、造語も出力された。また、元のカテゴリから詳細なカテゴリを推定し、出力している事例が見られ、元のカテゴリを推定した上で、元のカテゴリに関する幅広いデータを生成できたと考えられる。

表 3.3.3-12 業務課題 4 における学習データ作成検証結果④

#	作成件数	〇	△	×
4-④	100 件	91.0%	0.0%	9.0%

構成が類似するデータを出力できた例を表 3.3.3-13 に示す。

例 1 では、入力商標が「仕出しの佐久間」のところ、仕出しに対して、弁当や食事を配達するサービスといったカテゴリを推定して出力された。

例 2 では、入力商標が「藤原養蜂場」のところ、蜂蜜の生産や蜂に関連する事業といったカテゴリを推定して出力された。

表 3.3.3-13 業務課題 4 の手順 4-④において構成が類似するデータを出力できた例

例 1	例 2
入力商標：仕出しの佐久間	入力商標：藤原養蜂場
生成データ	生成データ
弁当の青山	佐藤蜜園
美食の田中	高木養蜂所
給食の伊藤	田中蜂業場
宅配の山本	山田ハニーファーム
ご馳走の中村	中村ビーランド
お膳の小林	小林アピアリー
和食の高橋	伊藤ハニーガーデン
懐石の渡辺	渡辺蜂の里
お弁当の斉藤	森田蜜蜂園
宅食の鈴木	鈴木ミツバチファーム

構成が類似するデータを出力できなかった例を表 3.3.3-14 に示す。

例 1 では、入力商標が「フィレンツェ」のところ、「フィレンツェ」を実在する地名と推定されず、地名を元にした造語が生成された。

例 2 では、入力商標が「オーヨシ」のところ、「オーヨシ」が氏ではなく名称として推定され、その結果、氏ではない結果が生成された。

表 3.3.3-14 業務課題 4 の手順 4-④において構成が類似するデータを出力できなかった例

例 1	例 2
入力商標：フィレンツェ	入力商標：オーヨシ
生成データ	生成データ
ミラノーバ	エーケイ
ロマネスク	オーマツ
ヴェネチアン	カイトヨ
シチリアーノ	ソーシキ
ナポリータ	テーユカ
パルマーレ	ネーリン
ジェノヴィア	ハーサイ
トリノーリ	ミーカナ
カプリエッセ	ヨータテ
ボローニャン	ルーヨシ

3.3.4. 生成 AI を用いた追加学習データ作成実施結果のまとめ

今回検証したプロンプトを用いることで、学習データとして有用なデータを生成可能であることが明らかになった。また全体の傾向として、以下に示す。

- 「意味の類似」、「カテゴリの類似」といった類似の種類はプロンプトである程度指定可能であると考えられる。
- 商標のみを入力としたときには、意図する意味合い・カテゴリとして扱われない可能性があるため、より厳密に出力を限定したい場合はカテゴリ・類似させたい商標の両方を入力することが望ましい。
- カテゴリのみを入力としたときには、より幅広い種類の単語を含む出力になる傾向がある。
- 「商標を作成してください」という指示文では、造語が作成される可能性がある。

3.4. 学習データ等の整備・加工

3.4.1. 実施方針と手順

3.2 及び 3.3 で作成した学習データについて、1 回目のモデル学習及び精度評価において活用するため、データの構造化や学習データ・評価データへの分割をする等の整備・加工を実施した。以下に業務課題ごとの学習データ等の整備・加工手順を示す。

【業務課題 1】

業務課題 1 については、以下の表 3.4.1-1 の手順でデータの整備・加工を実施した。

表 3.4.1-1 業務課題 1 のデータの整備・加工手順

#	手順	手順詳細	手順概要
1-1	データ前処理	スクリプトを実行して受領データを学習可能な形式に変換する。 スクリプトには以下の機能がある。 ・複数ファイルに分かれている受領データの内容を一つにまとめる ・正解フラグが付いていて、称呼が完全一致している以外の組を抽出する ・重複する組を除去する ・訓令式のローマ字に変換（促音、長音は記号に変換）した称呼を作成し、連番を付与する ・データを指定した比率で学習データ、評価データにランダムに分割する ・学習データ、評価データの各類似組データにおける全称呼の一覧データを作成する	入力：称呼の類似組データのリスト(受領データ) - 類似組データのみを利用するため、類似フラグがついているデータのみを抽出する - 類似組として重複を除く - モデルへの入力形式である、読みを「ローマ字 + 記号」で表現した形式に変換する - 学習用、評価用のデータに、以下の①,②の手順でランダムに分割する。 ①評価データを各基準 100 件ずつ選定する ②①の残りのデータについて 9:1 の比率で学習用、評価用に分割する

1-2	ランキング学習／特徴量の作成	スクリプトを実行し、学習データ、評価データのそれぞれのデータについて各称呼の特徴量を作成する。 実行後、設定したディレクトリに特徴量を記載した csv ファイルが出力される。	入力：1-1 での作成データ - 1-1 で変換した読みの情報をもとに、特徴量に変換する。特徴量の設計は表 3.4.1-2 に示す。
1-3	ランキング学習／差分特徴量の作成及びランクの付与	スクリプトを実行し、各称呼とランク付けの対象となる称呼の差分特徴量の作成及びランクの付与を実施する。差分特徴量には、1-2 のスクリプトで作成した特徴量の差分だけでなく、称呼をローマ字表記した際のレーベンシュタイン距離[3]を加えている。	- ランキング学習の学習には、クエリと正解の間の特徴量の差分と、ランク付け情報(検索時に正解を何位としたいのかの情報。クエリと正解であれば基本的に 1 と設定する)が必要となるため、類似組(クエリと正解)における、特徴量の差分の計算と、ランク付け情報を付与する
1-4	ランキング学習／モデル入力用の学習データ作成	スクリプトを実行し、モデル入力用の学習データを作成します。	- 1-3 の処理結果をもとに、ランキング学習用の入力形式に合わせた学習データファイルを作成する
1-5	ランキング学習／評価データ作成	スクリプトを実行し、評価データの各称呼対全称呼の差分特徴量(+称呼及びローマ字に変換した称呼のレーベンシュタイン距離)を作成する。	- 評価作業に向けて、1-2～1-4 の処理を、評価用データに対しても実施する
1-6	Attention×距離学習／学習・検証用データ作成	スクリプトを実行し、アンカー、類似、非類似の三つ組データを作成する。	入力：1-1 での作成データ - 学習には類似組(anchor × positive)に加え、非類似組(anchor × negative)の三つ組データ(anchor × positive × negative)のデータが必要となる。そのため、1-1 での作成データである類似組に加え、別データからランダムに非類似(negative)を選定し、三つ組データを作成する

[3] 同じ表記にする際に、文字単位で追加、削除、更新の操作が必要となる回数により、類似性を距離として測る

1-7	Attention × 距離学習（推論用）／トークナイズ処理	スクリプトを実行し、評価データの全称呼にトークナイズ処理を実施する。実行後、設定したパスに評価データ用のデータセットクラスのバイナリファイルが出力される。	- 距離学習時には、入力を token と呼ばれる入力形式に変換する必要があるため、評価データに対して token に変換するトークナイズ処理を実施する

ランキング学習における特徴量の設計について表 3.4.1-2 に示す。現行の称呼基準において判断に使用されている音の差分や、表記上の差分が特徴量として扱えるように設計した。

表 3.4.1-2 ランキング学習において設計した特徴量

#	特徴量	具体イメージ
1	称呼をローマ字表記に変換した上での、各アルファベット、記号の出現数の差分	【事例】 称呼 A: スイカ ⇒ suika 称呼 B: サイカ ⇒ saika ○特徴量： s:0, u:1, a:1, i:0, k:0
2	拗音、濁拗音、半濁拗音の出現数の差分	【事例】 称呼 A: スイカ 称呼 B: スイキャ ○特徴量： 拗音:1, 濁拗音:0,半濁拗音:1
3	特殊音（22 音）の出現数の差分	【事例】 称呼 A: ケーキ 称呼 B: クェーキ ○特徴量： 特殊音:1
4	現行の商標検索システムにおけるルールベースの検索エンジン構築のための参考資料に記載されている 2 音の組み合わせの出現数の差分	【事例】 称呼 A: ゴリラ 称呼 B: ゴルイラ ○特徴量：（ルイの出現数):1
5	称呼をローマ字表記した際の文字数の差分	【事例】 称呼 A: ゴリラ ⇒ ko:rira 称呼 B: ゴルイラ ⇒ ko:ruira ○特徴量： 差分:1

6	称呼の文字数の差分	【事例】 称呼 A: ゴリラ 称呼 B: ゴルイラ ○特徴量： 差分:1
7	称呼をローマ字表記した際のレーベンシュタイン距離	【事例】 称呼 A: ゴリラ ⇒ ko:rira 称呼 B: ゴルイラ ⇒ ko:ruira ○特徴量： レーベンシュタイン距離:1
8	称呼のレーベンシュタイン距離	【事例】 称呼 A: ゴリラ 称呼 B: ゴルイラ ○特徴量： 差分:1

【業務課題２】

業務課題２については、学習作業を行わないため、貸与データをすべて評価データとした。

【業務課題３】

業務課題３については、NER による前処理と Tokenizer による前処理の２つの手法でデータの整備・加工を実施した。

NER を活用したデータの整備・加工に向けて NER の学習を実施した。NER の学習データの生成は、図 3.4-1 の手順で知識グラフの情報資源活用により実施した。

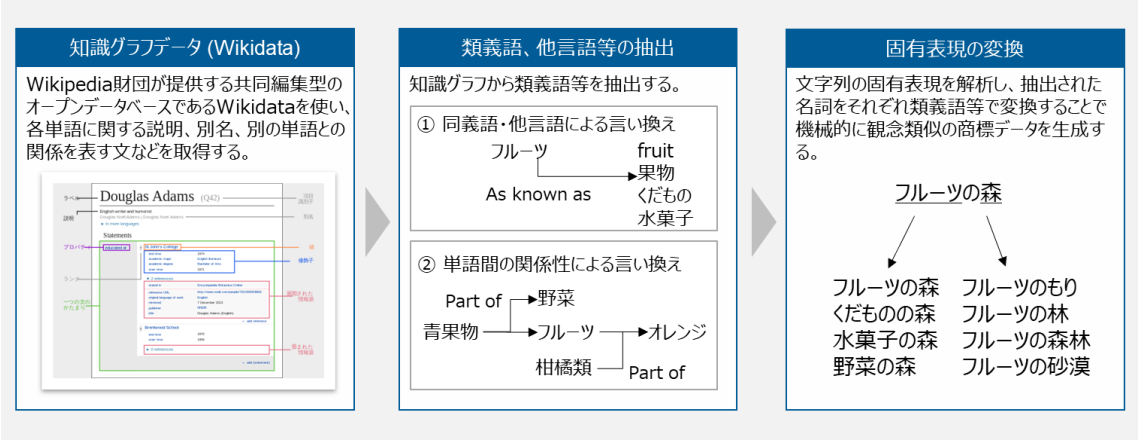


図 3.4-1 知識グラフの情報資源活用によるデータ生成

また Tokenizer を活用し、以下の表 3.4.1-3 の手順でデータの整備・加工を実施した。

表 3.4.1-3 業務課題 3 において Tokenizer を活用したデータの整備・加工手順

#	手順	手順詳細	手順概要
3-1	データ前処理	スクリプトを実行し、データ前処理を実施する。 スクリプトには以下の機能がある。 ・データを指定した比率で学習データ、評価データにランダムに分割する ・学習データ、評価データの各類似組データにおける全商標の一覧データを作成する	入力：表示用商標[4]の類似組データのリスト(受領データ) - 学習データ、評価データに指定した比率でランダムに分割する

[4] 商標の構成態様を文字と記号とを用いて表したデータ

3-2	triplet／学習・検証用データ作成	スクリプトを実行し、学習データ、評価データのそれぞれのデータについてアンカー、類似、非類似の三つ組データを作成する。 実行後、設定したパスにデータセットクラスのバイナリファイルが出力される。	入力：3-1 での作成データ - 学習には類似組(anchor × positive)に加え、非類似組(anchor × negative)の三つ組データ(anchor × positive × negative)のデータが必要となる。そのため、3-1 での作成データである類似組に加え、別データからランダムに非類似(negative)を選定し、三つ組データを作成する
3-3	トークナイズ処理	スクリプトを実行し、評価データの全商標にトークナイズ処理を実施する。 Triplet の場合は実行後、triplet の場合は設定したパスに評価データ用のデータセットクラスのバイナリファイルが出力される。 fasttext の場合は設定したパスに評価データを変換した tsv ファイルが出力される。	- 距離学習時には、入力を token と呼ばれる入力形式に変換する必要があるため、評価データに対して token に変換するトークナイズ処理を実施する

【業務課題 4 】

業務課題 4 については、NER による前処理と Tokenizer による前処理の 2 つの手法でデータの整備・加工を実施した。

NER を活用したデータの整備・加工に向けて、業務課題 3 と同様の手順で NER の学習を実施し、NER による学習データの整備・加工を実施した。

また Tokenizer を活用し、以下の表 3.4.1-4 手順でデータの整備・加工を実施した。

表 3.4.1-4 業務課題 4 において Tokenizer を活用したデータの整備・加工手順

#	手順	手順詳細	手順概要
4-1	データ前処理	スクリプトを実行し、データ前処理を実施する。 スクリプトには以下の機能がある。 ・データを指定した比率で学習データ、評価データにランダムに分割する	入力：検索用商標の類似組データのリスト(受領データ) - 受領データを、学習データと評価データに、以下の①, ②の手順でランダムに分割する。 ①評価データを各ラベル 100 件ずつ選定する

		・学習データ、評価データのデータに対し、総当たりで類似組を作成する ・学習データ、評価データの各類似組データにおける全商標の一覧データを作成する	②①の残りのデータについて 9:1 の比率で学習用，評価用に分割する - 類似組を作るため、受領データに含まれる構成要素情報にあたるラベル情報をもとに、同じラベルが付与されているデータの総当たりで類似組を作成する
4-2	triplet／学習・検証用データ作成	スクリプトを実行し、学習データ、評価データのそれぞれのデータについてアンカー、類似、非類似の三つ組データを作成する。 実行後、設定したパスにデータセットクラスのバイナリファイルが出力される。	入力：4-1 での作成データ - 学習には類似組(anchor × positive)に加え、非類似組(anchor × negative)の三つ組データ(anchor × positive × negative)のデータが必要となる。そのため、4-1 での作成データである類似組に加え、別データからランダムに非類似(negative)を選定し、三つ組データを作成する
4-3	トークナイズ処理	スクリプトを実行し、評価データの全商標にトークナイズ処理を実施する。 Triplet の場合は実行後、triplet の場合は設定したパスに評価データ用のデータセットクラスのバイナリファイルが出力される。 fasttext の場合は設定したパスに評価データを変換した tsv ファイルが出力される。	- 距離学習時には、入力を token と呼ばれる入力形式に変換する必要があるため、評価データに対して token に変換するトークナイズ処理を実施する

3.4.2. 整備・加工結果

3.4.1 に記載の手順で学習データ等の整備・加工を実施した結果を以下に示す。

【業務課題 1】

業務課題 1 について、学習データ及び評価データの件数は表 3.4.2-1 に示す。

表 3.4.2-1 業務課題 1 における学習データと評価データの件数

	学習データ件数	評価データ件数
全件	**18,145 件**	**1,046 件**
称呼基準 1	11,723 件	100 件
称呼基準 2	520 件	100 件
称呼基準 3	424 件	100 件
称呼基準 4	1,052 件	100 件
称呼基準 5	854 件	100 件
称呼基準 6	663 件	100 件
称呼基準 7	691 件	100 件
称呼基準 8	1,723 件	100 件
称呼基準 9	170 件	100 件
称呼基準 10	検証対象外	
称呼基準 11	225 件	100 件
称呼基準 12	100 件	46 件
称呼基準 13	検証対象外	
称呼基準 14	検証対象外	
称呼基準 15	検証対象外	

【業務課題 2】

業務課題 2 について、学習データ及び評価データの件数は表 3.4.2-2 に示す。業務課題 2 についてはルール処理の手法のため、学習を必要としないことから、学習データは 0 件である。

表 3.4.2-2 業務課題 2 における学習データと評価データの件数

	学習データ件数	評価データ件数
全件	**0 件**	**160 件**

【業務課題３】

業務課題３について、学習データ及び評価データの件数は表 3.4.2-3 に示す。

表 3.4.2-3 業務課題３における学習データと評価データの件数

	学習データ件数	評価データ件数
全件	**595 件**	**105 件**

【業務課題４】

業務課題４について、学習データ及び評価データの件数は表 3.4.2-4 に示す。

表 3.4.2-4 業務課題４における学習データと評価データの件数

	学習データ件数	評価データ件数
全件	**17,174 件**	**664 件**

以上のデータ件数をもって、１回目の AI モデルの構築及び精度検証を実施する。

4. 文字商標調査における人工知能等技術の活用のためのシステムの設計及び構築

本事業のシステムは、特許庁アジャイルシステムの担当者と相談して、システム設計や構築、及び AI エンジンの設計や構築、システム全体の動作確認のテストを実施した。

5. システムの精度評価

5.1. AI モデルの構築と精度評価の準備

5.1.1. 検証対象とする AI モデル

検証対象とする AI モデルの各手法を表 5.1.1-1 に示す。

表 5.1.1-1 検証対象の AI モデルの各手法

対象課題	モデルの手法	前処理・特徴量設計の方法	モデル内部で使用する技術	学習処理
業務課題1	1-(1)	人手により設計・処理構築	Ranking SVM	Ranking SVM
	1-(2)	ローマ字表記に変換する前処理のみ(深層学習手法のため、特徴量設計は不要)	Attention	deep metric learning
業務課題2	2-(3)	(ルール処理のため前処理なし)	共通部分文字列の検出	(ルール処理のため学習なし)
業務課題3	3-(1)	形態素解析または NER	事前知識を活用したルール処理	(ルール処理のため学習なし)
	3-(2)	NER*	fastText	なし
	3-(3)	Tokenizer(言語モデル(BERT)のツールの中で実装されている処理を使用)	BERT	deep metric learning
	3-(4)	Tokenizer(言語モデル(DeBERTa)のツールの中で実装されている処理を使用)	DeBERTa	deep metric learning

業務課題4	4-(1)	NER または NER-BERT	事前知識を活用したルール処理	(ルール処理のため学習なし)
	4-(2)	NER*	fastText	なし
	4-(3)	Tokenizer(言語モデル(BERT)のツールの中で実装されている処理を使用)	BERT	deep metric learning
	4-(4)	Tokenizer(言語モデル(DeBERTa)のツールの中で実装されている処理を使用)	DeBERTa	deep metric learning

*5.3 と 5.4 でのモデル精度検証の結果、NER が業務課題に対して適切な技術要素ではなかったため、NER-BERT は実施しないこととした。

5.1.2. 精度評価の方法

業務課題 1、2、3 については、正解がどれだけ上位に検索出来ているのか評価するために、精度評価は Recall@k を用いて定量値を算出する。Recall@k の評価概要及び評価方法については表 5.1.2-1 に示す。

表 5.1.2-1 Recall@k の評価指標の概要及び評価方法

評価指標の概要	評価方法
正解がどれだけ上位に検索出来ているのかを評価する。	一般的な Recall@k は、各クエリに対して上位 k 個の予測に含まれる正解数が、総正解数のうちどの程度の割合含まれているかを計算し、最後にクエリ全体での平均をとる。本検証では、正解は基本的に 1 件となるため、各クエリについて上位 k 個の予測に正解が含まれていれば 1、含まれていなければ 0 として計算し、最後にクエリ全体での平均をとる。

業務課題 4 については、正解以外にも同じ種別に属すると思われる文字要素の組合せが多くあるため、Recall@k の指標を用いた精度評価ではモデルの業務利用可能性を評価することが難しいことが想定される。

そのため、Recall@k に加え、検索結果の上位 100 件について目視確認を実施する。目視確認の実施方法として、文字要素の種別ごとに 2 件ランダムに評価データ

を選定し、それらの検索結果の上位 100 件についてクエリと同じ種別に属する文字要素の組み合わせとなっているかを確認して評価する。

5.1.3. 先行文字商標のデータ特性を考慮した評価方法

先行文字商標には評価データのクエリや正解と同じ表記の先行文字商標が複数含まれていることにより、定量評価の数値が低くなる場合があるため、図 1.5-1 のように、2 つの観点による評価方法を検討した。

(1) 検索システムの観点

検索システムとして正解が最高何位に表示可能なのかを評価することを目的として、正解と同様の表記の先行文字商標の中で、最上位に正解が検索されたものとみなして評価する。

(2) モデル性能の観点

先行文字商標の重複に関係なくモデルの性能として正解が何位に表示されるのかを評価することを目的として、表記が同一の先行文字商標はマージした上で、正解が何位に検索されるのかを評価する。

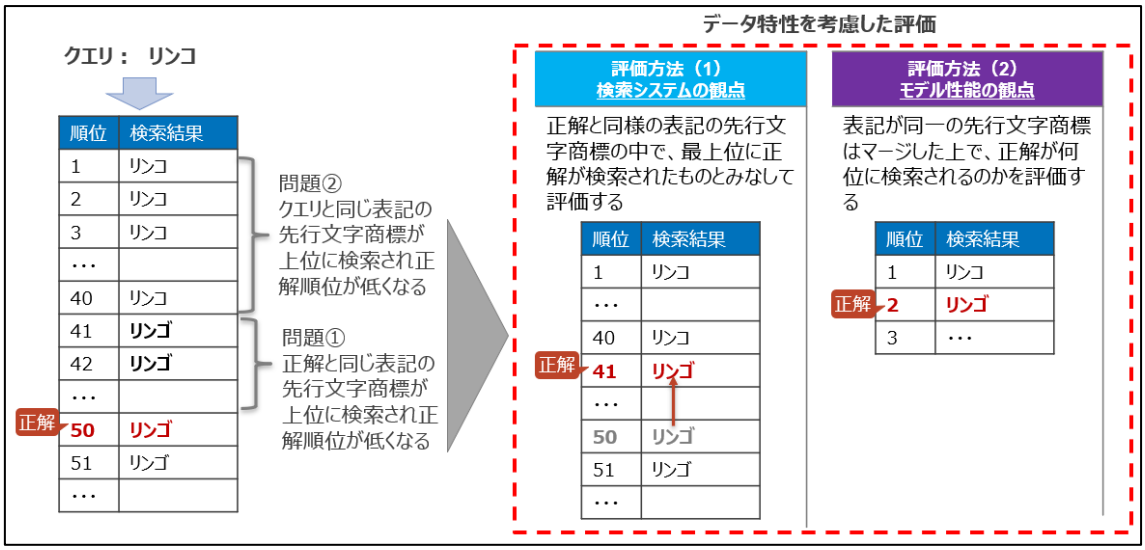


図 5.1-1 データ特性を考慮した評価方法

検索システムの観点とモデル性能の観点の 2 つの観点について、手法 1-(2)距離学習の AI モデルにて精度結果を比較し、同じ表記の先行文字商標が精度に影響しているのかを確認した（表 5.1.3-1、表 5.1.3-2）。検索システムの観点については、全体として Recall@k=100 では 63.9%の精度に対して、モデル性能の観点では 82.3%となっていることから、同じ表記の先行文字商標が含まれていることにより、定量評価する上での精度に影響していることを確認した。

ここではモデル本来の性能を評価・分析するため、以降の精度評価ではモデル性能の観点のみを確認した。

表 5.1.3-1 手法 1-(2)の検索システムの観点

@k	称呼基準											
	1	2	3	4	5	6	7	8	9	11	12	全体
1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
~3	8.0%	16.0%	5.5%	6.0%	6.0%	6.5%	7.5%	5.0%	4.0%	14.5%	0.0%	7.6%
~5	16.5%	39.0%	11.0%	17.0%	18.0%	21.5%	16.5%	13.5%	10.5%	32.5%	0.0%	18.7%
~10	26.0%	62.5%	19.5%	28.5%	27.0%	40.0%	25.0%	22.0%	22.5%	55.5%	2.2%	31.5%
~50	60.5%	84.5%	43.5%	60.5%	57.5%	59.0%	40.5%	41.5%	42.0%	77.5%	3.3%	54.3%
~100	76.5%	91.5%	51.5%	69.0%	69.5%	68.5%	54.5%	53.5%	48.0%	83.5%	5.4%	63.9%
~1000	98.5%	95.5%	84.5%	91.5%	94.5%	90.5%	83.0%	81.0%	72.5%	92.0%	25.0%	85.6%

表 5.1.3-2 手法 1-(2)のモデル性能の観点

@k	称呼基準											
	1	2	3	4	5	6	7	8	9	11	12	全体
1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
~3	59.0%	57.5%	20.0%	34.5%	32.0%	44.0%	23.5%	21.5%	21.5%	43.5%	2.2%	34.2%
~5	79.5%	71.0%	30.5%	54.0%	48.0%	57.5%	34.0%	31.5%	34.5%	59.0%	2.2%	47.8%
~10	93.0%	79.5%	37.0%	74.0%	72.0%	70.0%	43.5%	41.5%	44.5%	70.5%	2.2%	59.9%
~50	97.0%	91.5%	64.0%	95.0%	94.0%	87.0%	73.0%	66.0%	57.5%	86.0%	12.0%	78.1%
~100	97.5%	92.5%	73.5%	97.0%	97.0%	94.0%	81.0%	72.5%	62.5%	86.5%	14.1%	82.3%
~1000	100.0%	97.0%	96.0%	100.0%	100.0%	98.0%	98.0%	93.5%	86.0%	94.0%	41.3%	93.8%

5.1.4. 検証用 DB（データベース）の作成

検索実行時の検索対象となる検証用 DB については、各課題において正解となるデータが含まれている必要であるが、正解となるデータの一部は先行文字商標ではないものも含まれている。

そのため、図 5.1-2 のように、特許庁アジャイルシステム DB から取得した先行文字商標に、先行文字商標ではない正解データを追加することで、検証用 DB を作成する。

各課題に対する学習データや評価データについては 3.4.2 を参照。

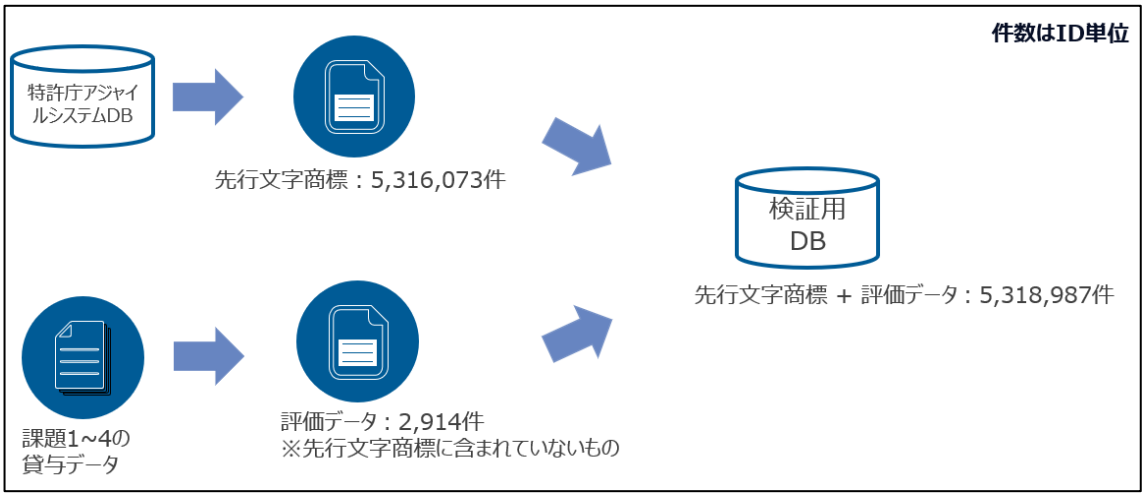


図 5.1-2 検証用 DB の作成

5.2. 「業務課題 1」のモデル精度検証

5.2.1. 手法 1-(1)、手法 1-(2)の精度評価

(1) 精度評価における手法 1-(1)の AI モデルの使用方法の変更

手法 1-(1)ランキング学習の AI モデルでは、検証用 DB 全件に対して類似度の算出を実施すると計算処理が本事業の仕様で求められる 8 秒以内で完了しないことが判明した。

そこで、図 5.2-1 のように、手法 1-(2)距離学習の AI モデルによって検証用 DB 内のレコード全件に対して類似度を算出してソートした上で、上位 1 万件を対して手法 1-(1) ランキング学習の AI モデルでリランキング（類似度の算出）を実施する方法で、検索結果を取得することとした（上位 1 万件に対するリランキングの処理時間については 1～2 秒程度となった）。

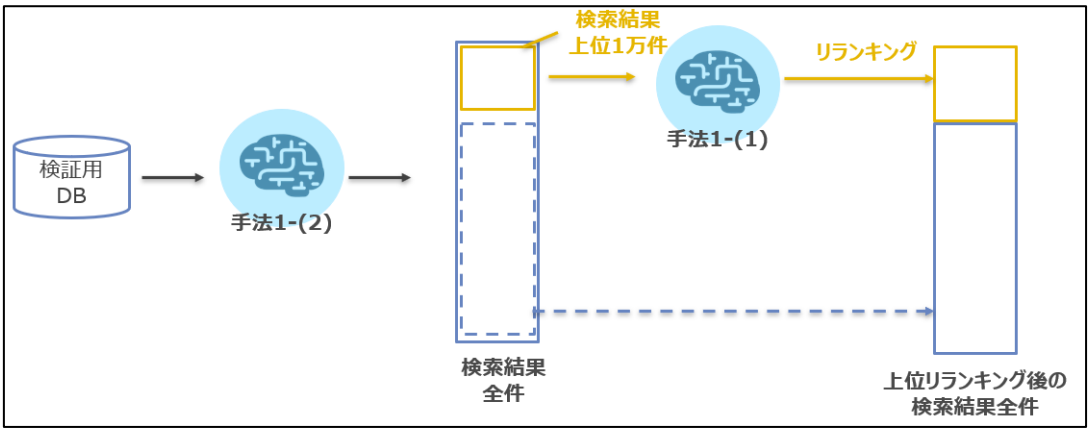


図 5.2-1 精度評価における手法 1-(1)の AI モデルの使用方法

(2) 1 回目の精度評価結果

手法 1-(2)距離学習の AI モデルの精度結果については、表 5.1.3-2 で示した通り、Recall@k=100 において 82.3%の精度となった。

一方で、手法 1-(1)ランキング学習の AI モデルの精度結果について、Recall@k=100 において 82.6%の精度となった（表 5.2.1-1)。

表 5.2.1-1 手法 1-(1)のモデル性能の観点

@k	称呼基準											
	1	2	3	4	5	6	7	8	9	11	12	全体
1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%	1.0%	1.1%	0.2%
~3	20.5%	55.5%	18.0%	14.5%	17.5%	22.0%	15.0%	24.5%	15.5%	61.5%	5.4%	25.5%
~5	30.0%	74.5%	30.5%	19.0%	30.5%	32.5%	22.5%	34.0%	21.0%	79.0%	16.3%	36.4%
~10	41.0%	85.5%	45.5%	34.0%	41.5%	49.0%	36.5%	49.5%	31.5%	86.0%	23.9%	48.9%
~50	78.0%	95.0%	72.5%	72.0%	67.5%	78.5%	66.0%	78.0%	63.0%	91.5%	51.1%	75.1%
~100	90.0%	95.5%	84.5%	81.0%	75.5%	88.5%	77.5%	82.5%	71.5%	93.0%	54.3%	82.6%
~1000	100.0%	95.5%	97.0%	97.5%	94.0%	98.5%	99.5%	96.0%	97.0%	97.0%	64.1%	95.7%

リランキングする前の手法 1-(2) 距離学習の AI モデルの結果と比較をすると、Recall@k=10 においては 11pt の精度低下となっている一方で、Recall@k=1000 では 1.9pt の精度向上となっており、リランキングにより 10 位以内は低下したが、1,000 位以内は向上していることを確認した（表 5.2.1-2)。

表 5.2.1-2 リランキング前（手法 1-(2)による結果)との比較

@k	称呼基準											
	1	2	3	4	5	6	7	8	9	11	12	全体
1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%	1.0%	1.1%	0.2%
~3	-38.5%	-2.0%	-2.0%	-20.0%	-14.5%	-22.0%	-8.5%	3.0%	-6.0%	18.0%	3.3%	-8.7%
~5	-49.5%	3.5%	0.0%	-35.0%	-17.5%	-25.0%	-11.5%	2.5%	-13.5%	20.0%	14.1%	-11.4%
~10	-52.0%	6.0%	8.5%	-40.0%	-30.5%	-21.0%	-7.0%	8.0%	-13.0%	15.5%	21.7%	-11.0%
~50	-19.0%	3.5%	8.5%	-23.0%	-26.5%	-8.5%	-7.0%	12.0%	5.5%	5.5%	39.1%	-3.0%
~100	-7.5%	3.0%	11.0%	-16.0%	-21.5%	-5.5%	-3.5%	10.0%	9.0%	6.5%	40.2%	0.4%
~1000	0.0%	-1.5%	1.0%	-2.5%	-6.0%	0.5%	1.5%	2.5%	11.0%	3.0%	22.8%	1.9%

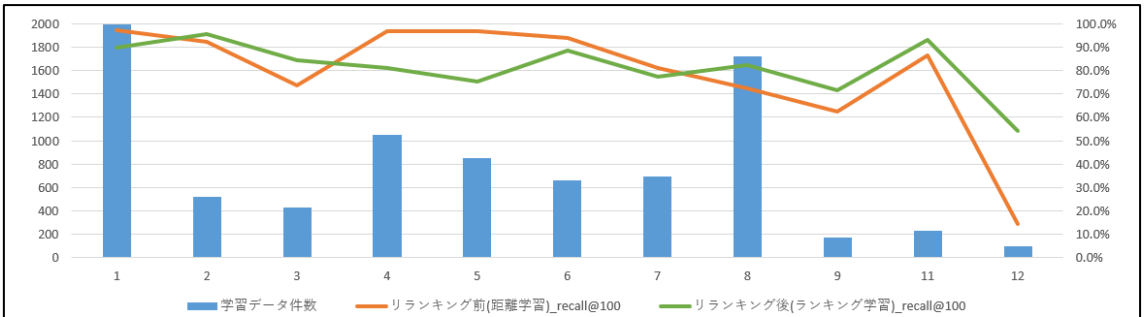
称呼基準毎の学習データの件数と精度の関係を分析した結果、リランキン

グ前である手法 1-(2)距離学習の AI モデルでは、学習データ件数に応じて精度にばらつきがあった一方で、リランキング後である手法 1-(1)ランキング学習の AI モデルでは、手法 1-(2) 距離学習の AI モデルと比較して、精度のばらつきが減っていることが確認できた（図 5.2-2)。

この要因として、手法 1-(1) ランキング学習の AI モデルでは、特徴量として出現する文字の差分や文字数の差分等の、称呼基準に依存しない特徴量を使用しているため、各基準の学習データ件数の偏りに影響されないモデルが構築できていると考えられる。

手法 1-(2) 距離学習の AI モデルに焦点をあてて、称呼基準毎の学習データの件数と精度を確認すると、基本的に学習データ件数が多いものは精度が高い傾向にあり、称呼基準 3,9, 12 の学習データ件数が少ないものは精度が低い傾向であった。

しかし、称呼基準 8 は学習データ件数が多いが精度は低い傾向にあり、称呼基準 11 は学習データ件数が少ないが精度が高い傾向であった。これらは学習データのパターンが不足、または充足していることが影響していると考えられるため、詳細分析を実施する。



	称呼基準										
	1	2	3	4	5	6	7	8	9	11	12
学習データ件数	2000	520	424	1052	854	663	691	1723	170	225	100
リランキング前 recall@100	97.5%	92.5%	73.5%	97.0%	97.0%	94.0%	81.0%	72.5%	62.5%	86.5%	14.1%
リランキング後 recall@100	90.0%	95.5%	84.5%	81.0%	75.5%	88.5%	77.5%	82.5%	71.5%	93.0%	54.3%

図 5.2-2 学習データ件数と精度の関係

(1) 1 回目の精度評価結果の詳細分析

正解の順位が低い事例は、称呼の音の違いのパターンよって正解順位が下がる場合があり、学習データにおいて該当のパターンの件数が少ないことが想定された。

称呼基準 1 に焦点をあてて、正解順位が 10 位以下の事例を分析した結果を表 5.2.1-3 に示す。

正解が低くなる要因として、正解順位が低い事例は称呼の差分が「ディ⇔ジ」、「ティ⇔チ」等、学習データにおける該当パターンの件数が少ないことが考えられる。

また、長音や促音の有無によっても順位が下がる傾向があり、長音や促音の学習データが少ないことから、長音や促音の有無のパターンについても学習データ量を増やすことで、精度の向上に有効と考えられる。

表 5.2.1-3 称呼基準 1 において正解順位が低い例の詳細分析

#	クエリ	正解	正解順位	クエリと正解の差分
1	エディソンクラブ	エジソンクラブ	259	ディ⇔ジ
2	バースティ	バスチー	162	ティ⇔チ、長音の有無
3	デュラメタル	ジュラメタル	135	デュ⇔ジュ
4	アンティドット	アンチドート	66	ティ⇔チ、長音・促音の有無
5	プチルウ	プティルウ	32	ティ⇔チ
6	ビードロ	ビイードロ	26	ビイ⇔ビ
7	ドーベル	ドベル	14	長音の有無
8	ラクショー	ラクショウ	12	ショー⇔ショウ
9	ヨーテイ	ヨーテー	10	テイ⇔テー
10	ジュジュ	ジュージュー	10	長音の有無

大きく順位が低下した称呼の差分を調査した結果、正解の順位が低下する場合の傾向は以下の通りとなった。

- **学習データにおける該当パターンが少ないと考えられる音の違い。**
- **商標を構成する称呼の音数を比較して、特定の 1 音の差があるとき。**
- **称呼基準 3 を中心に、2 音が違うとき。**
- **上記までの差分や、長音や促音の有無等、称呼の事前処理パターンに関する音の違いの組み合わせ。**

(2) 業務課題 1 の対策・改善方針

以上の分析結果から、業務課題 1 における対策・改善方針を表 5.2.1-4 に示す。

また、現行の商標検索システムでは称呼の音の違いに関する事前処理（例として、ドーベルをドベル、ドオベル、ドウベル等に展開する処理）を実装しているため、その事前処理のパターンを参考にした後処理も実装する。

表 5.2.1-4 業務課題 1 における対策・改善方針

#	課題	対策・改善方針
1-1	学習データにおける該当パターンが少ないと考えられる音の違い。 商標を構成する称呼の音数を比較して、特定の 1 音の差があるとき。 称呼基準 3 を中心に、2 音が違うとき。 上記までの差分や、長音や促音の有無等、称呼の事前処理に関する音の違いの組み合わせ。	条件に該当するデータを自動生成し、学習データの件数を増やすことで、該当の条件における類似スコアが高くなるように学習する。
1-2	現行の商標検索システムで実装している称呼の音の違いに関する事前処理のパターンにマッチしている場合は類似スコアを高くする。	AI モデルの結果の上位について、称呼の音の違いに関する事前処理のパターンとマッチするデータについては、類似スコアを高くする後処理を実装する。

5.2.2. 手法 1-(1)、手法 1-(2)の改善と再精度評価

(1) 対策 1-1 の結果

対策 1-1 について、学習データにおいて不足している類似音の関係性にあるデータを補うため、先行商標を拡張元のデータとして、特許庁から提供された音の関係性の情報をもとに、該当の 1 か所を変換したデータを、各音の組み合わせごとに同じ件数を作成した。

なお、実際の称呼基準においては、特定の条件において 2 か所の音の違いも類似とする基準となっているが、データ拡張においては 1 か所のみ異なるデータを作成した。これの理由として、2 か所に違いのあるデータも作成すると拡張後のデータ件数が多すぎてしまうこと、1 か所を変換したデータを作成すれば音の組み合わせとしては学習データとして網羅できることから、本検証では 1 か所のみ異なるデータを作成した。学習データの拡張結果については表 5.2.2-1 に示す。

表 5.2.2-1 学習データの拡張結果

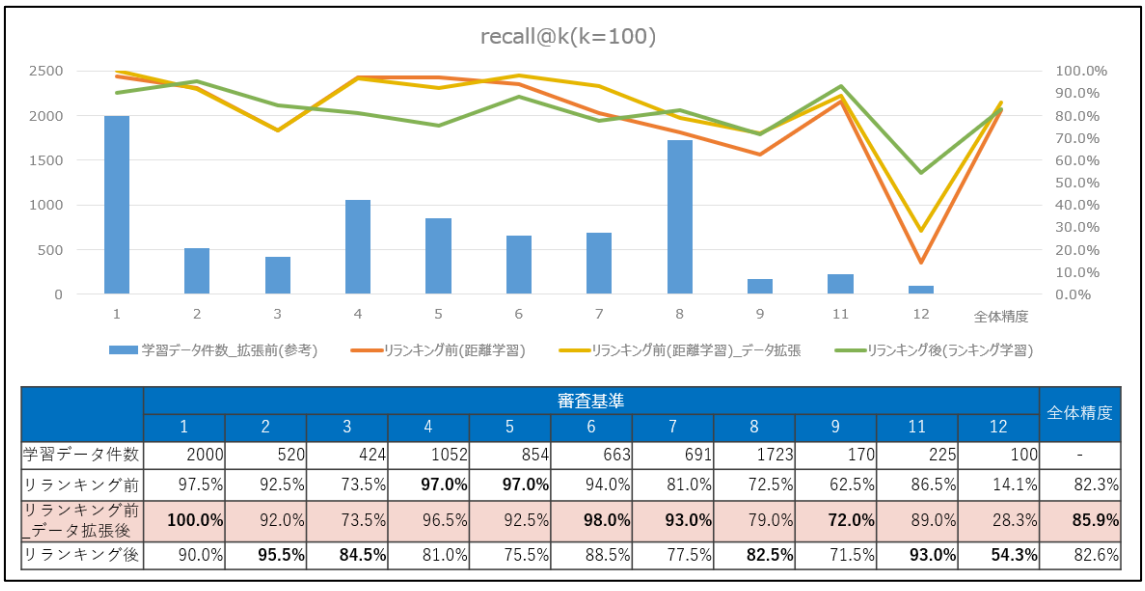
拡張前(1)	拡張した件数(2)	拡張後(1)+(2)
18,145 件	**17,974 件**	**36,119 件**

学習データの拡張による対策 1-1 の結果、手法 1-(2)距離学習の AI モデルでは、モデル性能の観点の Recall@k=100 で全体精度が 82.3%から 85.9%と**3pt 以上向上する結果となり、対策の有効性を確認することができた**（表 5.2.2-2）。

拡張前の手法 1-(2)距離学習の AI モデルと称呼基準別で比較すると、学習データ件数が比較的少ない称呼基準 6 以降で精度の改善が確認できた一方で、称呼基準 2、3 については、精度の向上は確認できなかった（図 5.2-3）。

表 5.2.2-2 対策 1-1 実施後のモデル性能の観点

@k	称呼基準											
	1	2	3	4	5	6	7	8	9	11	12	全体
~3	60.0%	49.0%	18.5%	34.0%	33.5%	51.0%	28.5%	23.0%	21.0%	46.0%	1.1%	34.9%
~5	76.5%	66.0%	30.0%	54.0%	51.0%	63.0%	40.5%	31.5%	30.0%	63.0%	3.3%	48.5%
~10	89.5%	80.0%	45.0%	70.0%	71.5%	78.0%	60.5%	46.0%	39.0%	75.5%	5.4%	62.9%
~50	97.5%	91.5%	69.5%	91.0%	89.0%	94.5%	89.5%	70.5%	60.5%	84.0%	20.7%	81.0%
~100	100.0%	92.0%	73.5%	96.5%	92.5%	98.0%	93.0%	79.0%	72.0%	89.0%	28.3%	85.9%
~1000	100.0%	97.5%	92.5%	100.0%	99.5%	99.0%	98.0%	96.5%	97.0%	96.5%	69.6%	96.4%



	審査基準											全体精度
	1	2	3	4	5	6	7	8	9	11	12	
学習データ件数	2000	520	424	1052	854	663	691	1723	170	225	100	-
リランキング前	97.5%	92.5%	73.5%	**97.0%**	**97.0%**	94.0%	81.0%	72.5%	62.5%	86.5%	14.1%	82.3%
リランキング前_データ拡張後	**100.0%**	92.0%	73.5%	96.5%	92.5%	**98.0%**	**93.0%**	79.0%	**72.0%**	89.0%	28.3%	**85.9%**
リランキング後	90.0%	**95.5%**	**84.5%**	81.0%	75.5%	88.5%	77.5%	**82.5%**	71.5%	**93.0%**	**54.3%**	82.6%

図 5.2-3 評価結果と学習データ件数

対策実施後に正解順位が向上した事例を確認したところ、学習データを拡張することで、拡張前は学習が不足していた異なる音の組み合わせを学習し、正解順位が想定通り改善したと考えられた。

一方で、対策実施後に正解順位が低下した事例を確認したところ、対策により正解以外に称呼が類似するデータが上位に検索されるようになり、結果として正解の順位が下がる結果となっていた。

以上の結果から、**対策 1-1 については、全体で 3pt 以上の精度向上となり、対策の有効性を確認した。**

(2) 対策 1-2 の結果

称呼の音の違いに関する事前処理のパターンを参考に、手法 1-(2)距離学習の AI モデルによる検索結果上位 1,000 件の内、事前処理のパターンと一致するデータについては上位にリランキングする対策を実施した。リランキングする際には、入力テキストと事前処理せずに一致するデータの次の順位になるように、リランキング処理を実施した（図 5.2-4）。

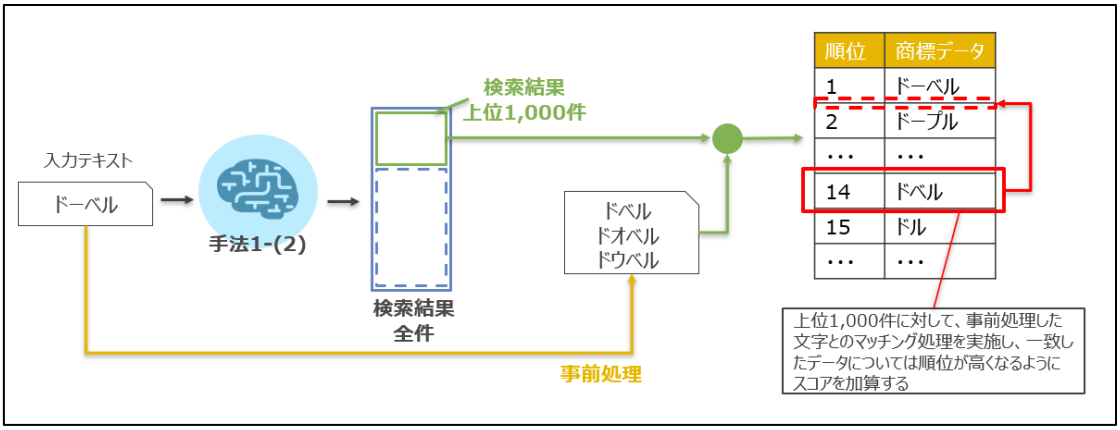


図 5.2-4 リランキングの方法

対策 1-2 では、事前処理パターンのタイプの内、対策時の有効性や検証期間を考慮し、本検証で対策の対象とするものを選定した。

小文字部分を大文字化する処理については、称呼をモデルに入力するためにカタカナから音をあらわすローマ字表記に変換した際に、小文字、大文字部分の子音と母音については同じローマ字が含まれるため、対策を実施しなくても一定の精度が見込めるため、それ以外のタイプを優先的に対策することとして選定した（表 5.2.2-3）。

表 5.2.2-3 対象とした事前処理

タイプ	内　容	内容の説明	対策の対象
1	非展開	元の称呼に記入された称呼から長音部及び促音部をとる。	○
2	長音部を前音の母音に置換	元の称呼から促音部をとり、長音部を前音の母音に置換する。	○
3	長音部を前音の母音に置換し、小文字部分を大文字化	タイプ２．の処置をしたのち拗音部を大文字化する。	
4	前音の母音と同一な母音を削除	元の称呼から長音部及び促音部をとり、前音の母音と同じ母音が連続している場合、この母音を削除する。	○
5	前音の母音と同一な母音を削除し、小文字部分を大文字化	タイプ４．の処置をしたのち拗音部を大文字化する。	
6	前音の母音と同一な母音を削除し、小文字部分を大文字化し、３音目以降削除	タイプ５．の処置をしたのち、３音目以降を削除する。	
7	前音の母音と同一な母音を削除し、３音目以降削除	タイプ４．の処置をしたのち、３音目以降を削除する。	○
8	小文字部分の大文字化	元の称呼から長音部をとり、拗音部及び促音部を大文字化する。さらに促音を除いて拗音部だけを大文字化	
9	小文字部分の大文字化し、３音目以降削除	タイプ８．の処置をしたのち、３音目以降を削除する。	
10	３音目以降削除	３音構成以上の称呼で、２音目以降が全く同一のものは、３音目以降を削除する。	○

対策 1-2 の実施後の結果を表 5.2.2-4、対策実施前との精度の比較結果を表 5.2.2-5 に示す。

対策実施により、称呼基準 1 は改善した一方で、それ以外の称呼基準については精度が低下する結果となった。ただし、精度が低下した称呼基準については、上位 10 位以内での正解順位の精度低下にとどまっているため、上位 50 位や 100 位までの精度を比較したときには、精度が低下しているとはいえない。

表 5.2.2-4 対策 1-2 実施後のモデル性能の観点

@k	称呼基準											
	1	2	3	4	5	6	7	8	9	11	12	全体
~3	74.0%	57.0%	18.0%	30.0%	29.0%	39.0%	22.0%	20.0%	20.5%	43.5%	2.2%	33.8%
~5	92.0%	71.0%	29.5%	49.5%	47.0%	53.0%	31.5%	31.0%	33.0%	59.0%	2.2%	47.6%
~10	94.5%	79.5%	37.0%	73.0%	71.0%	68.5%	42.5%	41.5%	44.5%	70.5%	2.2%	59.6%
~50	97.0%	91.5%	64.0%	95.0%	94.0%	87.0%	73.0%	66.0%	57.5%	86.0%	12.0%	78.1%
~100	97.5%	92.5%	73.5%	97.0%	97.0%	94.0%	81.0%	72.5%	62.5%	86.5%	14.1%	82.3%
~1000	100.0%	97.0%	96.0%	100.0%	100.0%	98.0%	98.0%	93.5%	86.0%	94.0%	41.3%	93.8%

表 5.2.2-5 対策 1-2 実施前後の精度比較

@k	称呼基準											
	1	2	3	4	5	6	7	8	9	11	12	全体
~3	15.0%	-0.5%	-2.0%	-4.5%	-3.0%	-5.0%	-1.5%	-1.5%	-1.0%	0.0%	0.0%	-0.4%
~5	12.5%	0.0%	-1.0%	-4.5%	-1.0%	-4.5%	-2.5%	-0.5%	-1.5%	0.0%	0.0%	-0.3%
~10	1.5%	0.0%	0.0%	-1.0%	-1.0%	-1.5%	-1.0%	0.0%	0.0%	0.0%	0.0%	-0.3%
~50	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
~100	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
~1000	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

青字：精度向上した箇所

赤字：精度低下した箇所

対策 1-2 によって正解順位が向上した事例では長音の違いがある検索結果を想定通り上位に検索することができ、順位が向上していた。

一方で、正解順位が低下した事例では、正解以外で事前処理により一致する事例が正解よりも上位に検索されることで、正解の順位が低下していた。

以上の結果から、**対策 1-2 については、全体精度としては 10 位以内における正解の順位が低下したものの、事前処理により一致する事例が想定通り上位に検索できるようになり、有効性を確認した。**

5.3. 「業務課題２」のモデル精度検証

5.3.1. 手法 2-(3)の精度評価

(1) 1 回目の精度評価

業務課題 2 の評価に当たり、クエリと正解の文字種別の組み合わせが精度に大きく影響することが想定されるため、文字種別の組み合わせごとに精度を算

出した結果を表 5.3.1-1 に示す。

クエリと正解の文字種別が同じ場合には、Recall@k=1 で 80%以上の精度となった。

一方で、クエリと正解の文字種別が異なる場合には精度が大きく低下しており、Recall@k=1,000 で 0%の場合もあった。

表 5.3.1-1 手法 2-(3)のモデル性能の観点

@k	文字種別の組み合わせ(クエリ × 正解)												
	アルファベット × アルファベット	カタカナ × カタカナ	漢字 × 漢字	複数の文字の組み合わせ × 複数の文字の組み合わせ	アルファベット × 複数の文字の組み合わせ	漢字 × 複数の文字の組み合わせ	アルファベット × カタカナ	アルファベット × 漢字	カタカナ × 複数の文字の組み合わせ	漢字 × カタカナ	ひらがな × 複数の文字の組み合わせ	漢字 × ひらがな	全体
1	88.7%	81.8%	91.7%	83.3%	25.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	54.1%
~3	93.0%	81.8%	94.4%	83.3%	25.0%	37.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	55.6%
~5	95.8%	81.8%	94.4%	83.3%	50.0%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	56.9%
~10	95.8%	81.8%	97.2%	83.3%	50.0%	75.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	57.8%
~50	98.6%	90.9%	100.0%	83.3%	50.0%	75.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	60.6%
~100	98.6%	90.9%	100.0%	87.5%	50.0%	75.0%	0.0%	0.0%	50.0%	0.0%	0.0%	0.0%	61.3%
~1000	98.6%	95.5%	100.0%	95.8%	50.0%	87.5%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	63.4%

(2) 1 回目の精度評価結果の詳細分析

文字種別が異なる場合については、**文字種別が異なることに加えて、入れ替えられる文字要素以外にテキストが付け加えられている場合も、精度が低下**する要因となっていた。

文字種別が異なる場合に精度が低下する理由として、手法 2-(3)のようなルール処理では、文字の表層情報から文字要素の入れ替えが発生しているかを判定して類似スコア化しているため、文字種別が異なる場合には一致する要素がないと判定されてしまい、類似スコアが小さくなることが挙げられる。

テキストが付け加えられている場合にも精度が低下する理由として、要素が入れ替わった上で完全一致している場合よりも、部分一致している場合の方が、類似スコアが低くなるようにするため、入れ替わっている要素以外の文字が多いほど類似スコアが低くなる処理としており、類似スコアが小さくなることが挙げられる。

また、文字種別が同じ場合についても、一部のデータでは正解が下位に検索されるケースがあり、**入れ替えられる文字要素以外にテキストが付け加えら**

れている場合、及び長音符号として使用されている記号が異なる場合（「－」と「～」）にも、上記と同様の理由で精度が低下していた。

(3) 業務課題 2 の対策・改善方針

以上の分析結果から、業務課題 2 における対策・改善方針を表 5.3.1-2 に示す。

表 5.3.1-2 業務課題 2 における対策・改善方針

#	課題	対策・改善方針
2-1	前後の構成文字要素の入れ替えに加え、文字種別が異なる関係にある事例の類似スコアを高くする。 長音の表記に異なる記号が使われている場合にも類似スコアを高くする。	クエリ、検索対象に称呼を用いて検索することで、文字種別を統一した形で類似スコアを算出する。
2-2	前後の構成文字要素の入れ替えに加え、部分一致の関係にある事例の類似スコアを高くする。	前後の構成文字要素の入れ替えの関係にあるうえで部分一致の関係にある場合には、他の文字が含まれていても類似スコアを高くする。

5.3.2. 手法 2-(3)の改善と再精度評価

(1) 対策 2-1 の結果

対策 2-1 では**称呼を用いて検索を実施**することで、文字種別を統一した形で類似スコアを算出する。

称呼を用いた検索による精度評価の結果を表 5.3.2-1、検索用商標を用いた検索結果との精度比較を表 5.3.2-2 に示す。全体における Recall@k=100 で 93.8%となり、検索用商標を用いた検索結果と精度比較して 32.5pt の精度向上となった。

文字種別の組み合わせ別の結果では、検索用商標を用いた検索では文字種別が異なる場合の検索精度が低かった部分の精度が、想定通り向上する結果となった。

表 5.3.2-1 対策 2-1 実施後のモデル性能の観点

@k	文字種別の組み合わせ(クエリ × 正解)												
	アルファベット × アルファベット	カタカナ × カタカナ	漢字 × 漢字	複数の文字の組み合わせ × 複数の文字の組み合わせ	アルファベット × 複数の文字の組み合わせ	漢字 × 複数の文字の組み合わせ	アルファベット × カタカナ	アルファベット × 漢字	カタカナ × 複数の文字の組み合わせ	カタカナ × 漢字	ひらがな × 複数の文字の組み合わせ	ひらがな × 漢字	全体
1	83.3%	90.6%	65.7%	83.3%	0.0%	50.0%	85.4%	100.0%	0.0%	100.0%	50.0%	0.0%	79.7%
~3	86.1%	92.2%	71.4%	83.3%	0.0%	62.5%	89.0%	100.0%	0.0%	100.0%	50.0%	0.0%	82.6%
~5	87.5%	93.8%	71.4%	83.3%	0.0%	75.0%	90.2%	100.0%	0.0%	100.0%	50.0%	0.0%	83.9%
~10	87.5%	93.8%	80.0%	91.7%	25.0%	87.5%	92.7%	100.0%	0.0%	100.0%	75.0%	0.0%	87.2%
~50	93.1%	93.8%	80.0%	100.0%	75.0%	100.0%	96.3%	100.0%	0.0%	100.0%	100.0%	25.0%	91.8%
~100	93.1%	96.9%	82.9%	100.0%	75.0%	100.0%	98.8%	100.0%	0.0%	100.0%	100.0%	50.0%	93.8%
~1000	100.0%	96.9%	88.6%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	100.0%	100.0%	50.0%	96.7%
評価データ件数	72	64	35	24	4	8	82	4	2	2	4	4	305

表 5.3.2-2 対策 2-1 実施前後の精度比較

@k	文字種別の組み合わせ(クエリ × 正解)												
	アルファベット × アルファベット	カタカナ × カタカナ	漢字 × 漢字	複数の文字の組み合わせ × 複数の文字の組み合わせ	アルファベット × 複数の文字の組み合わせ	漢字 × 複数の文字の組み合わせ	アルファベット × カタカナ	アルファベット × 漢字	カタカナ × 複数の文字の組み合わせ	カタカナ × 漢字	ひらがな × 複数の文字の組み合わせ	ひらがな × 漢字	全体
1	-5.4%	8.8%	-26.0%	0.0%	-25.0%	25.0%	85.4%	100.0%	0.0%	100.0%	50.0%	0.0%	25.6%
~3	-6.8%	10.4%	-23.0%	0.0%	-25.0%	25.0%	89.0%	100.0%	0.0%	100.0%	50.0%	0.0%	27.0%
~5	-8.3%	11.9%	-23.0%	0.0%	-50.0%	25.0%	90.2%	100.0%	0.0%	100.0%	50.0%	0.0%	27.1%
~10	-8.3%	11.9%	-17.2%	8.3%	-25.0%	12.5%	92.7%	100.0%	0.0%	100.0%	75.0%	0.0%	29.4%
~50	-5.5%	2.8%	-20.0%	16.7%	25.0%	25.0%	96.3%	100.0%	0.0%	100.0%	100.0%	25.0%	31.2%
~100	-5.5%	6.0%	-17.1%	12.5%	25.0%	25.0%	98.8%	100.0%	-50.0%	100.0%	100.0%	50.0%	32.5%
~1000	1.4%	1.4%	-11.4%	4.2%	50.0%	12.5%	100.0%	100.0%	-100.0%	100.0%	100.0%	50.0%	33.3%

青字：精度向上した箇所

赤字：精度低下した箇所

対策 2-1 によって正解順位が向上した事例ではアルファベットとカタカナで異なる文字種別の場合でも、称呼を用いて検索することで文字種別が統一され

たことにより、要素が入れ替わっているデータを検索された。

一方で低下した事例では、要素の部分的な読みが一致するデータが上位に検索されてしまい、正解の順位が低下する結果となった。

以上の結果から、**対策 2-1 については、全体精度で 32.5pt の向上となり対策の有効性は確認できた。**

しかしながら、前提として業務課題２では、単に文字の位置の入れ替えを踏まえた検索をできるようにすることが目的であるところ、クエリと正解が同じ文字種別の場合に精度が低下している。

業務運用上においては、まずは同じ文字種別の場合に、より精度を高くすることが望ましいことから、当初の検索用商標を用いた検索手法の方が、有効性が高いものと思われる。

(2) 対策 2-2 の結果

対策 2-2 の精度評価の結果を表 5.3.2-3、対策実施前との精度比較を表 5.3.2-4 に示す。

全体における Recall@k=100 で 58.8%となり、実施前と比較して 2.5pt の低下となった。アルファベット同士の組み合わせでの精度が一部向上したものの、それ以外では精度が低下する結果となった。

表 5.3.2-3 対策 2-2 実施後のモデル性能の観点

@k	文字種別の組み合わせ(クエリ × 正解)												
	アルファベット × アルファベット	カタカナ × カタカナ	漢字 × 漢字	複数の文字の組み合わせ × 複数の文字の組み合わせ	複数の文字の組み合わせ × アルファベット	複数の文字の組み合わせ × 漢字	アルファベット × カタカナ	アルファベット × 漢字	複数の文字の組み合わせ × カタカナ	カタカナ × 漢字	複数の文字の組み合わせ × ひらがな	ひらがな × 漢字	全体
1	90.1%	81.8%	91.7%	83.3%	25.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	54.4%
~3	91.5%	81.8%	94.4%	83.3%	25.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	55.0%
~5	93.0%	81.8%	94.4%	83.3%	50.0%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	56.3%
~10	94.4%	81.8%	97.2%	83.3%	50.0%	75.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	57.5%
~50	98.6%	81.8%	100.0%	83.3%	50.0%	75.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	58.8%
~100	98.6%	81.8%	100.0%	83.3%	50.0%	75.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	58.8%
~1000	98.6%	90.9%	100.0%	83.3%	50.0%	87.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	60.9%
評価データ件数	72	64	35	24	4	8	82	4	2	2	4	4	305

表 5.3.2-4 対策 2-2 実施前後の精度比較

@k	文字種別の組み合わせ(クエリ × 正解)												
	アルファベット × アルファベット	カタカナ × カタカナ	漢字 × 漢字	複数の文字の組み合わせ × 複数の文字の組み合わせ	複数の文字の組み合わせ × アルファベット	複数の文字の組み合わせ × 漢字	アルファベット × カタカナ	漢字 × アルファベット	複数の文字の組み合わせ × カタカナ	カタカナ × 漢字	複数の文字の組み合わせ × ひらがな	ひらがな × 漢字	全体
1	1.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%
~3	-1.4%	0.0%	0.0%	0.0%	0.0%	-12.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-0.6%
~5	-2.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-0.6%
~10	-1.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-0.3%
~50	0.0%	-9.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-1.9%
~100	0.0%	-9.1%	0.0%	-4.2%	0.0%	0.0%	0.0%	0.0%	-50.0%	0.0%	0.0%	0.0%	-2.5%
~1000	0.0%	-4.5%	0.0%	-12.5%	0.0%	0.0%	0.0%	0.0%	-100.0%	0.0%	0.0%	0.0%	-2.5%

青字：精度向上した箇所

赤字：精度低下した箇所

対策 2-2 によって正解順位が低下した事例について、対策により部分一致の事例のスコアを高くした結果、要素としては一致していないが文字単位では同じ文字が含まれているデータが上位に検索されるようになり、結果として正解順位が低下する結果となった。

以上の結果から、**対策 2-2 については、全体精度で 2.5pt の低下という結果から、有効性を確認できなかった。**

5.4. 「業務課題３」のモデル精度検証

手法 3-(1)と手法 3-(2)は同じ NER を利用した手法のため、ここでは手法 3-(1)よりも高い精度が期待できる fastText も利用した手法 3-(2)の検証を優先し、仮に手法 3-(2)の精度が高かった場合に、手法 3-(1)とのさらなる精度比較を行うこととした。

NER と fastText を利用した手法 3-(2)、BERT を利用した手法 3-(3)、BERT の改良版である DeBERTa を利用した手法 3-(4)の 3 つの手法で精度を比較し、課題に対して有効性が高いと考えられる技術要素を確認し、その技術要素が利用されている手法に対して改善と再精度評価を実施する。

5.4.1. 手法 3-(2)、手法 3-(3)、手法 3-(4)の精度評価

(1) 1 回目の精度評価

精度評価にあたり、クエリと正解の文字種別の組み合わせが精度に大きく影

響することが想定されるため、文字種別の組み合わせごとに精度を算出した。

手法 3-(2)の精度結果について、同じ文字種別の組み合わせごとに精度を算出した結果を表 5.4.1-1、異なる文字種別の組み合わせごとに精度を算出した結果を表 5.4.1-2、全体の精度結果を表 5.4.1-3 に示す。

同じ文字種別の組み合わせの場合は Recall@k=100 で 46.6%、異なる文字種別の組み合わせの場合は Recall@k=100 で 6.5%となり、同じ文字種別の組み合わせの場合と比べて、異なる文字種別の組み合わせの場合は精度が低い結果となった。

また、全体精度では Recall@k=100 で 23.2%という精度となった。

表 5.4.1-1 手法 3-(2)のモデル性能の観点（同じ文字種別の組み合わせ）

@k	文字種別の組み合わせ(クエリ × 正解)					
	漢字 × 漢字	ひらがな × ひらがな	カタカナ × カタカナ	外国語 × 外国語	混合 × 混合	全体
1	10.0%	0.0%	11.1%	2.5%	5.6%	5.7%
~3	40.0%	100.0%	50.0%	7.5%	44.4%	29.5%
~5	50.0%	100.0%	50.0%	7.5%	55.6%	33.0%
~10	50.0%	100.0%	55.6%	7.5%	61.1%	35.2%
~50	50.0%	100.0%	66.7%	12.5%	61.1%	39.8%
~100	70.0%	100.0%	72.2%	15.0%	72.2%	46.6%
~1000	80.0%	100.0%	72.2%	25.0%	88.9%	55.7%

表 5.4.1-2 手法 3-(2)のモデル性能の観点（異なる文字種別の組み合わせ）

@k	文字種別の組み合わせ(クエリ × 正解)													
	ひらがな × 漢字	カタカナ × 漢字	外国語 × 漢字	混合 × 漢字	カタカナ × ひらがな	ローマ字 × ひらがな	外国語 × ひらがな	ローマ字 × カタカナ	外国語 × カタカナ	混合 × カタカナ	外国語 × ローマ字	混合 × ローマ字	混合 × 外国語	全体
1	0.0%	0.0%	0.0%	10.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.9%
~3	0.0%	0.0%	0.0%	20.0%	0.0%	0.0%	0.0%	0.0%	0.0%	25.0%	0.0%	0.0%	0.0%	3.7%
~5	0.0%	0.0%	0.0%	20.0%	0.0%	0.0%	0.0%	0.0%	0.0%	25.0%	0.0%	0.0%	0.0%	3.7%
~10	0.0%	0.0%	0.0%	20.0%	0.0%	0.0%	0.0%	0.0%	0.0%	25.0%	0.0%	0.0%	0.0%	3.7%
~50	0.0%	0.0%	0.0%	20.0%	0.0%	0.0%	0.0%	0.0%	0.0%	25.0%	0.0%	0.0%	0.0%	3.7%
~100	0.0%	0.0%	0.0%	30.0%	0.0%	0.0%	0.0%	0.0%	0.0%	50.0%	0.0%	0.0%	0.0%	6.5%
~1000	16.7%	0.0%	0.0%	60.0%	0.0%	0.0%	0.0%	0.0%	0.0%	50.0%	0.0%	0.0%	16.7%	11.2%

表 5.4.1-3 手法 3-(2)のモデル性能の観点（全体）

@k	全体
1	2.9%
~3	14.3%
~5	15.8%
~10	16.8%
~50	18.8%
~100	23.2%
~1000	29.7%

次に、手法 3-(3)の精度結果について、同じ文字種別の組み合わせごとに精度を算出した結果を表 5.4.1-4、異なる文字種別の組み合わせごとに精度を算出した結果を表 5.4.1-5、全体の精度結果を表 5.4.1-6 に示す。

同じ文字種別の組み合わせの場合は Recall@k=100 で 80.9%、異なる文字種別の組み合わせの場合は Recall@k=100 で 11.2%となり、手法 3-(2)と同様に、同じ文字種別の組み合わせの場合と比べて、異なる文字種別の組み合わせの場合は精度が低い結果となった。

また、全体精度では Recall@k=100 で 42.4%という精度となった。

表 5.4.1-4 手法 3-(3)のモデル性能の観点（同じ文字種別の組み合わせ）

@k	文字種別の組み合わせ(クエリ × 正解)					
	漢字×漢字	ひらがな×ひらがな	カタカナ×カタカナ	外国語×外国語	混合×混合	全体
1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
~3	45.8%	50.0%	66.7%	20.0%	60.0%	40.4%
~5	50.0%	50.0%	66.7%	32.5%	70.0%	47.9%
~10	70.8%	100.0%	66.7%	47.5%	80.0%	61.7%
~50	83.3%	100.0%	83.3%	62.5%	90.0%	75.5%
~100	87.5%	100.0%	83.3%	70.0%	100.0%	80.9%
~1000	91.7%	100.0%	88.9%	80.0%	100.0%	87.2%

表 5.4.1-5 手法 3-(3)のモデル性能の観点（異なる文字種別の組み合わせ）

@k	文字種別の組み合わせ(クエリ × 正解)													
	ひらがな × 漢字	カタカナ × 漢字	外国語 × 漢字	混合 × 漢字	カタカナ × ひらがな	ローマ字 × ひらがな	外国語 × ひらがな	ローマ字 × カタカナ	外国語 × カタカナ	混合 × カタカナ	外国語 × ローマ字	混合 × ローマ字	混合 × 外国語	全体
1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
~3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
~5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	12.5%	0.0%	0.0%	0.0%	0.9%
~10	0.0%	12.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	37.5%	0.0%	0.0%	0.0%	3.4%
~50	0.0%	25.0%	0.0%	16.7%	0.0%	0.0%	0.0%	0.0%	2.3%	37.5%	0.0%	0.0%	0.0%	6.0%
~100	25.0%	37.5%	0.0%	33.3%	0.0%	0.0%	0.0%	0.0%	4.5%	37.5%	0.0%	0.0%	16.7%	11.2%
~1000	62.5%	50.0%	0.0%	100.0%	0.0%	0.0%	25.0%	0.0%	43.2%	62.5%	0.0%	0.0%	50.0%	37.1%

表 5.4.1-6 手法 3-(3)のモデル性能の観点（全体）

@k	全体
1	0.0%
~3	18.1%
~5	21.9%
~10	29.5%
~50	37.1%
~100	42.4%
~1000	59.5%

次に、手法 3-(4)の精度結果について、同じ文字種別の組み合わせごとに精度を算出した結果を表 5.4.1-7、異なる文字種別の組み合わせごとに精度を算出した結果を表 5.4.1-8、全体の精度結果を表 5.4.1-9 に示す。

同じ文字種別の組み合わせの場合は Recall@k=100 で 50.0%、異なる文字種別の組み合わせの場合は Recall@k=100 で 13.8%となり、手法 3-(2)と手法 3-(3)と同様に、同じ文字種別の組み合わせの場合と比べて、異なる文字種別の組み合わせの場合は精度が低い結果となった。

また、全体精度では Recall@k=100 で 32.9%という精度となった。

表 5.4.1-7 手法 3-(4)のモデル性能の観点（同じ文字種別の組み合わせ）

@k	文字種別の組み合わせ(クエリ × 正解)					
	漢字 × 漢字	ひらがな × ひらがな	カタカナ × カタカナ	外国語 × 外国語	混合 × 混合	全体
1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
~3	37.5%	0.0%	5.6%	15.0%	40.0%	37.5%
~5	41.7%	0.0%	11.1%	22.5%	40.0%	41.7%
~10	41.7%	0.0%	27.8%	32.5%	40.0%	41.7%
~50	50.0%	0.0%	66.7%	45.0%	40.0%	50.0%
~100	50.0%	50.0%	72.2%	55.0%	50.0%	50.0%
~1000	83.3%	50.0%	100.0%	80.0%	70.0%	83.3%

表 5.4.1-8 手法 3-(4)のモデル性能の観点（異なる文字種別の組み合わせ）

@k	文字種別の組み合わせ(クエリ × 正解)													
	ひらがな × 漢字	カタカナ × 漢字	外国語 × 漢字	混合 × 漢字	カタカナ × ひらがな	ローマ字 × ひらがな	外国語 × ひらがな	ローマ字 × カタカナ	外国語 × カタカナ	混合 × カタカナ	外国語 × ローマ字	混合 × ローマ字	混合 × 外国語	全体
1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
~3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
~5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
~10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.5%	0.0%	0.0%	0.0%	0.0%	1.7%
~50	0.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	6.8%	37.5%	0.0%	0.0%	0.0%	6.9%
~100	0.0%	37.5%	0.0%	50.0%	0.0%	0.0%	0.0%	0.0%	13.6%	50.0%	0.0%	0.0%	0.0%	13.8%
~1000	12.5%	62.5%	0.0%	100.0%	0.0%	50.0%	0.0%	0.0%	31.8%	75.0%	0.0%	0.0%	16.7%	29.3%

表 5.4.1-9 手法 3-(4)のモデル性能の観点（全体）

@k	全体
1	0.0%
~3	9.5%
~5	11.9%
~10	16.2%
~50	25.7%
~100	32.9%
~1000	53.3%

手法 3-(2)、手法 3-(3)、手法 3-(4)の 3 つの手法の精度結果を比較したものを表 5.4.1-10 に示す。最も全体精度が高かったのは BERT を利用した手法 3-(3)という結果となった。なお、NER と fastText を利用した手法 3-(2)は、手法 3-(3)と手法 3-(4)と比較して全体精度が低い結果となったため、手法 3-(2)と同じく NER を利用した手法 3-(1)は精度評価を実施しなかった。

表 5.4.1-10 3 つの手法の全体精度比較（モデル性能の観点）

@k	手法 3-(2)	手法 3-(3)	手法 3-(4)
1	2.9%	0.0%	0.0%
~3	14.3%	18.1%	9.5%
~5	15.8%	21.9%	11.9%
~10	16.8%	29.5%	16.2%
~50	18.8%	37.1%	25.7%
~100	23.2%	42.4%	32.9%
~1000	29.7%	59.5%	53.3%

(2) 1 回目の精度評価結果の詳細分析

最も精度が高かった手法 3-(3)について、正解が上位に検索できた事例を確認したところ、単語の意味的な類似性という観点においては、想定したデータを上位に検索出来ていることが確認できた。

手法 3-(3)にて、正解の順位が 100 位以下となった事例について分析し、クエリと正解の主な関係性を整理した結果を表 5.4.1-11 に示す。

言語的な特性から整理すると、**他言語間での翻訳関係や、日本語のローマ字表記、英単語のカタカナ読み等、いくつかの種類の関係性が混ざっている**ことが確認できた。

BERT は意味の関係性を捉える手法であるが、他言語間での翻訳関係や、日本語のローマ字表記、英単語のカタカナ読み等の関係性については対応が難しいことが確認できた。

表 5.4.1-11 クエリと正解の主な関係性の事例

#	関係性	内容	事例
1	翻訳関係	日本語と他言語間での翻訳関係	・トモダチ ⇔ FRIEND ・蝶 ⇔ PAPILLON
2	ローマ字表記	日本語と、その読みのローマ字表記の関係	・長閑 ⇔ NODOKA ・中日 ⇔ Chunichi
3	漢字の読みの違い	漢字の読みの違いの関係	・ほうざん ⇔ たからやま
4	省略表記	一方の省略形表記の関係	・ROBONavigator ⇔ ロボナビ
5	英単語のカタカナ読み	英単語と、そのカタカナ読みの関係	・スオード ⇔ THE SWORD ・ライオンクラブ ⇔ LIONSCLUB
6	漢字と読み	漢字と、その読みの関係	・朝・昼・夜 ⇔ あさ ひる ばん

(3) 課題 3 の対策・改善方針

以上の分析結果から、業務課題 3 における対策・改善方針を表 5.4.1-12 に示す。

表 5.4.1-12 業務課題 3 における対策・改善方針

#	課題	対策・改善方針	備考
3-1	日本語と他言語間での翻訳関係にある場合に類似スコアを高くする。	日英の単語辞書を使ってマッチングする。	
		他言語翻訳用の言語モデルを使って類似スコアを算出する。	
3-2	日本語と、その読みのローマ字表記の関係にある場合に類似スコアを高くする。	日本語の読みをローマ字表記に変換してマッチングする。	
3-3	漢字の読みの違いの関係にある場合に類似スコアを高くする。	形態素解析を使って漢字の読みを取得してマッチングする。	異なる読みから漢字に変換する方式は対策 3-6 と共通であるため、対策の有効性は対策 3-6 で確認する。

3-4	一方の省略形表記の関係にある場合に類似スコアを高くする。	読みを取得した上で、部分一致でマッチングする。	省略形式には複数のパターンが考えられ、対策を実施しても部分的な対応になることが想定され、有効性は低いことから検証の対象外とする。
3-5	英単語と、そのカタカナ読みの関係にある場合に類似スコアを高くする。	英単語のカタカナ読みを辞書から取得し、マッチングする。	
3-6	漢字と、その読みの関係にある場合に類似スコアを高くする。	かな漢字変換ソフトを使ってひらがなを漢字に変換する前処理を加えて類似スコアを算出する。	

5.4.2. 手法 3-(3)の改善と再精度評価

(1) 対策 3-1 の結果

対象とする単語については、一般名詞が多く、通常の辞書に登録されている単語が多いことが想定されるため、単語辞書を使用する方式を選定した。対策 3-1 では、「jamdict[5]」という大規模な多言語日本語翻訳辞書ツールを使用し、英語の商標を日本語に翻訳し、日本語に統一した上で類似度を算出した。

対策 3-1 を実施後の精度評価結果を表 5.4.2-1 に示す。Recall@k=100 で 43.33%となり 0.95pt の向上となり、文字種別の組み合わせについては精度が向上したが、部分的な精度改善に留まっている。

表 5.4.2-1 対策 3-1 実施後のモデル性能の観点

@k	改善対象とする文字種別の組み合わせ(クエリ × 正解) 日本語×英語(34 件)		全体(210 件)	
	対策後精度	対策前との比較	対策後精度	対策前との比較
1	2.94%	+2.94	0.95%	+0.95
~3	8.82%	+8.82	20.95%	+2.86
~5	8.82%	+8.82	25.24%	+3.33

[5] jamdict(https://github.com/neocl/jamdict)

~10	11.76%	+11.76	33.33%	+3.81
~50	14.71%	+14.71	39.52%	+2.38
~100	17.65%	+14.71	43.33%	+0.95
~1000	29.41%	+17.65	57.62%	-1.91%

青字：精度向上した箇所
赤字：精度低下した箇所

対策 3-1 によって、英語から適切に日本語に変換できた事例では正解順位が向上したが、日本語変換自体がうまくいかなかった事例も多く、正解順位が低下するという結果となった。

以上の結果から、**対策 3-1 については、Recall@k=100 で 0.95pt の精度向上となり、有効性は確認できたが、部分的な精度向上に留まってしまった。**

(2) 対策 3-2 の結果

対策 3-2 では、「KAKASI[6]」というツールを使用し、日本語の商標をローマ字に変換し、ローマ字表記におけるマッチングを実施した。

対策 3-2 を実施後の精度評価結果を表 5.4.2-2 に示す。Recall@k=100 で 28.1%となり 14.29pt の低下となった。改善対象とする文字種別の組み合わせについては精度が向上したものの、改善対象外のデータに対して悪影響があり、全体の精度が低下した。

表 5.4.2-2 対策 3-2 実施後のモデル性能の観点

@k	改善対象とする文字種別の組み合わせ(クエリ × 正解) ローマ字×漢字・ひらがな(6 件)		全体(210 件)	
	対策後精度	対策前との比較	対策後精度	対策前との比較
1	0.00%	0.00	0.00%	0.00
~3	16.67%	+16.67	10.00%	-8.10
~5	16.67%	+16.67	14.29%	-7.62
~10	33.33%	+33.33	18.57%	-10.95

[6] KAKASI(http://kakasi.namazu.org/index.html.ja)

~50	66.67%	+66.67	25.24%	-11.90
~100	66.67%	+66.67	28.10%	-14.29
~1000	66.67%	+66.67	37.14%	-22.38

青字：精度向上した箇所
赤字：精度低下した箇所

対策 3-2 によって、ひらがなや既知の単語を適切にローマ字変換するができたため正解順位が向上した事例がある一方で、漢字で書かれた未知の単語については適切な読みに変換することができず、順位が低下するという結果となった。

以上の結果から、**対策 3-2 については、対象とする事例においては精度向上が確認できたが、全体精度は-14.29pt の低下となり、有効性は確認できなかった。**

(3) 対策 3-5 の結果

対策 3-5 では、「CMUdict[7]」という英語の発音辞書ツールを使用し、英単語の発音を取得して、ルールベースでカタカナの読みの変換を実施した。

対策 3-5 を実施後の精度評価結果を表 5.4.2-3 に示す。Recall@k=100 で 41.9%となり 0.48pt の低下となった。対策 3-5 も改善対象とする文字種別の組み合わせについては精度が向上したものの、改善対象外のデータに対して悪影響があり、全体の精度が低下した。

表 5.4.2-3 対策 3-5 実施後のモデル性能の観点

@k	改善対象とする文字種別の組み合わせ(クエリ × 正解) 英語×カタカナ(24 件)		全体(210 件)	
	対策後精度	対策前との比較	対策後精度	対策前との比較
1	4.17%	+4.17	0.48%	+0.48
~3	8.33%	+8.33	21.43%	+3.33
~5	8.33%	+8.33	22.86%	+0.95
~10	8.33%	+8.33	29.05%	-0.48

[7] CMUdict（http://www.speech.cs.cmu.edu/cgi-bin/cmudict）

~50	16.67%	+12.50	34.76%	-2.38
~100	20.83%	+16.67	41.90%	-0.48
~1000	41.67%	-8.33	57.62%	-1.91

青字：精度向上した箇所
赤字：精度低下した箇所

対策 3-5 によって、対策の対象となる単語が辞書に登録されている場合はカタカナ読みに変換できるため想定通り正解順位が向上した。

しかし、辞書に登録されている場合においても、辞書に登録されている読みと、商標におけるカタカナ表記の読みの差異が大きい場合、類似とされず順位が低下するという結果となった。

以上の結果から、**対策 3-5 についても、対象とする事例においては精度向上が確認できたが、全体精度は-0.48pt の低下となり、有効性は確認できなかった。**

(4) 対策 3-6 の結果

対策 3-6 では、かな漢字変換ができる「mozc[8]」というツールを使用し、ひらがなの商標を漢字に変換した。

対策 3-6 を実施後の精度評価結果を表 5.4.2-4 に示す。Recall@k=100 で 39.53%となり 2.86pt の低下となった。対策 3-6 も改善対象とする文字種別の組み合わせについては精度が向上したものの、改善対象外のデータに対して悪影響があり、全体の精度が低下した。

表 5.4.2-4 対策 3-6 実施後のモデル性能の観点

@k	改善対象とする文字種別の組み合わせ(クエリ × 正解)		全体(210 件)	
	漢字×ひらがな(6 件)			
	対策後精度	対策前との比較	対策後精度	対策前との比較
1	0.00%	0.00%	0.00%	0.00
~3	33.33%	+33.33	16.67%	-1.43
~5	33.33%	+33.33	21.91%	0.00

[8] mozc(https://github.com/google/mozc)

~10	33.33%	+33.33	28.57%	-0.95
~50	33.33%	+33.33	35.72%	-1.43
~100	50.00%	+16.67	39.53%	-2.86
~1000	83.33%	0.00	58.57%	-0.95

青字：精度向上した箇所
赤字：精度低下した箇所

対策 3-6 によって、ひらがなを適切な漢字に変換することで正解順位が向上した事例もあった。

一方で、「おむすびころりん」を「おむすびコロリン」にといった、変換が不要なデータについても変換してしまうことで、順位が低下する事例もあるという結果になった。

以上の結果から、**対策 3-6 についても、対象とする事例においては精度向上が確認できたが、全体精度は-2.86pt の低下となり、有効性は確認できなかった。**

5.5. 「業務課題４」のモデル精度検証

業務課題 3 と同様に、手法 4-(1)と手法 4-(2)は同じ NER を利用した手法のため、ここでは手法 4-(1)よりも高い精度が期待できる fastText も利用した手法 4-(2)の検証を優先し、仮に手法 4-(2)の精度が高かった場合に、手法 4-(1)とのさらなる精度比較を行うこととした。

NER と fastText を利用した手法 4-(2)、BERT を利用した手法 4-(3)、BERT の改良版である DeBERTa を利用した手法 4-(4)の 3 つの手法で精度を比較し、課題に対して有効性が高いと考えられる技術要素を確認し、その技術要素が利用されている手法に対して改善と再精度評価を実施する。

5.5.1. 手法 4-(2)、手法 4-(3)、手法 4-(4)の精度評価

(1) 1 回目の精度評価

課題 4 を Recall@k で精度評価した結果、手法 4-(2)は Recall@k=1000 で 10%未満、手法 4-(3)と手法 4-(4)は Recall@k=1000 で 10%台と、他の業務課題と比較して低い精度となった。

5.1.2 に示した通り、正解以外にも、クエリと同じ種別の文字要素の組合せが多くあるため、Recall@k の指標を用いた精度評価ではモデルの業務利用可能性を評価することが難しいことから、検索結果の上位 100 件についてクエリと同

じ種別の文字要素の組み合わせとなっているか目視確認を実施した。

目視確認による評価結果について、NER と fastText を利用した手法 4-(2)の結果を表 5.5.1-1、BERT を利用した手法 4-(3)の結果を表 5.5.1-2 に示す。DeBERTa を利用した手法 4-(4)については、BERT を利用した手法 4-(3)と Recall@k の評価値について大きな差分がないことから、手法 4-(3)の検証のみを取り上げる。

手法 4-(2)については、14 個の事例のうち、9 個の事例については検索結果上位 100 件以内にクエリと同じ種別の文字要素の組み合わせが多く含まれており、想定した結果が得られていることを確認した。一方で 5 個の事例については、クエリとは異なる種別の文字要素の組み合わせが多く含まれていた。

手法 4-(3)については、14 個の事例のうち、10 個の事例については検索結果上位 100 件以内にクエリと同じ種別の文字要素の組み合わせが多く含まれており、想定した結果が得られていることを確認した。一方で 4 個の事例については、クエリとは異なる種別の文字要素の組み合わせが多く含まれていた。

表 5.5.1-1 手法 4-(2)の目視確認による評価結果

構成要素の種別		商品役務の一般名称		地名		地名＋商品役務の一般名称		地名＋業種名		氏		氏＋商品役務の一般名称		氏＋業種名	
		事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2
構成要素一致の判定結果	一致	100	100	74	82	44	24	86	92	1	16	71	7	98	85
	不一致	0	0	25	18	55	76	14	8	99	83	29	93	2	13
	判断困難	0	0	1	0	1	0	0	0	0	1	0	0	0	2

青セル：文字要素の組み合わせが同じデータが上位に検索できている事例

赤セル：文字要素の組み合わせが異なるデータが上位に含まれている事例

表 5.5.1-2 手法 4-(3)の目視確認による評価結果

構成要素の種別		商品役務の一般名称		地名		地名＋商品役務の一般名称		地名＋業種名		氏		氏＋商品役務の一般名称		氏＋業種名	
		事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2
構成要素一致	一致	100	100	99	98	52	33	99	99	1	4	98	97	84	93
	不一致	0	0	1	2	43	62	1	1	99	85	1	3	0	0

の判定結果	判断困難	0	0	0	0	5	5	0	0	0	11	1	0	16	7

青セル：文字要素の組み合わせが同じデータが上位に検索できている事例

赤セル：文字要素の組み合わせが異なるデータが上位に含まれている事例

NER と fastText を利用した手法 4-(2)と BERT を利用した手法 4-(3)を比較すると、BERT を利用した手法 4-(3)の方が文字要素の組み合わせが同じデータが上位に検索できていることが確認できた。

(2) 1 回目の精度評価結果の詳細分析

手法 4-(3)について、上位に異なる種別の文字要素の組み合わせが検索された事例については、クエリが、**通常は漢字で記載される表記がひらがなやカタカナで記載されることで、語句の文字要素種別ではなく表記の情報に引きずられて、誤ったデータが上位に検索されている**ことを確認した。

また、NER と fastText を利用した手法 4-(2)について、上位に検索できた事例は BERT を利用した手法 4-(3)でも上位に検索できており、手法 4-(2)の有効性は確認できなかった。そのため、課題 3 と同様に、手法 4-(2)と同じく NER を利用した手法 4-(1)は精度評価を実施しなかった。

(3) 課題 4 の対策・改善方針

以上の分析結果から、業務課題 4 における対策・改善方針を表 5.5.1-3 に示す。

対策 4-1 では、かな漢字変換ソフト「mozc」を使用するが、「mozc」はカタカナから漢字に変換することはできないため、事前にカタカナからひらがなに変換する必要がある。

その際に、本来は変換が不要なカタカナ表記のデータもひらがなや漢字に変換してしまい、精度が低下する要因になることが想定されたため、今回の対策・改善方針ではひらがなのみを対象とした。

表 5.5.1-3 業務課題 4 における対策・改善方針

#	課題	対策・改善方針
4-1	通常は漢字で表記されるが、ひらがな表記されている場合にも、意味を捉えて類似スコアを算出する。	かな漢字変換ソフトを使い、ひらがなから漢字表記に変換する前処理を加えた上で、類似スコアを算出する。

5.5.2. 手法 4-(3)の改善と再精度評価

(1) 対策 4-1 の結果

対策 4-1 では、対策 3-6 と同様に、かな漢字変換ができる「mozc」というツールを使用し、ひらがなの商標を漢字に変換した。

対策の結果について、対策 4-1 実施後の評価結果を表 5.5.2-1、対策 4-1 実施前との比較を表 5.5.2-2 に示す。

対策 4-1 を実施した結果、地名＋商品役務の一般名称の事例 1 においては大幅に改善が見られたが、地名の事例 1 及び事例 2 では精度が下がるという結果となった。

表 5.5.2-1 対策 4-1 実施後の評価結果

構成要素の種別		商品役務の一般名称		地名		地名＋商品役務の一般名称		地名＋業種名		氏		氏＋商品役務の一般名称		氏＋業種名	
		事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2
構成要素一致の判定結果	一致	100	100	84	79	100	35	99	99	1	4	96	92	85	94
	不一致	0	0	11	21	0	63	1	1	99	85	4	5	0	0
	判断困難	0	0	5	0	0	2	0	0	0	11	0	3	15	6

表 5.5.2-2 対策 4-1 実施前との比較

構成要素の種別		商品役務の一般名称		地名		地名＋商品役務の一般名称		地名＋業種名		氏		氏＋商品役務の一般名称		氏＋業種名	
		事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2	事例 1	事例 2
構成要素一致の判定結果	一致	0	0	-15	-19	48	2	0	0	0	0	-2	-5	1	1
	不一致	0	0	10	19	-43	1	0	0	0	0	3	2	0	0
	判断困難	0	0	5	0	-5	-3	0	0	0	0	-1	3	-1	-1

青字：精度向上した事例

赤字：精度低下した事例

対策 4-1 によって改善した事例では、かな漢字変換が想定通り変換され、モデルにより想定したスコア算出ができるようになり、精度改善に至った。

精度が低下した事例では、**本来とは別の漢字に変換され、クエリとは異なる**

構成の事例が上位に検索されることが確認できた。

変化がなかった事例は、**連続する母音が長音で記載される場合（「オオ」と「オー」）等において、想定したかな漢字変換ができておらず、対策により精度改善ができない結果となった。**

以上の結果から、**対策 4-1 については、対象とする事例においては精度向上が確認できたが、対策により精度が下がった事例もあったことから、有効性は確認できなかった。**

5.6. システム検証

業務課題毎に、構築したシステムのフロントエンドから入力したクエリに対して、AI モデルが算出した結果がフロントエンドに返ってくるまでの検索時間について、下記の条件で計 10 回検索を実施し、平均時間を表 5.6-1 に示す。

なお、本事業で構築した AI モデルやシステムについて、同時に操作するユーザ数やクエリの数、文字種別等によっては、検索時間に影響が及ぶため、以下の条件を前提として計測を実施した。

- 操作するユーザ数は 1 人。
- 各クエリは日本語を 1 単語入力。
- 各クエリの単語は 10 回とも異なる。
- 図形要素を有する商標を検索結果から除外はしない。
- 類似群での検索絞り込みはしない。
- 画面上の検索ボタンを押下して、画面に検索結果が表示されるまでの測定。

※サーバスペックなどインフラ条件の記載は割愛。

表 5.6-1 業務課題毎の検索時間

対象課題	検索平均時間
業務課題 1	4.35 秒
業務課題 2	6.67 秒
業務課題 3	4.77 秒
業務課題 4	4.43 秒

本事業で構築したシステムでは、検索時間や AI エンジンの API 動作時のハードウェアメモリの利用容量削減等の、システムにおけるリソースの効率性を改善するための対策を実施した。具体的には、業務課題 1、3、4 についてはデータをベクトルに変換した後に、主成分分析を用いた次元削減を実施することで、検索時のメモリ使用量の削減や

計算量の削減を実施した。業務課題 2 については、最初に n-gram の絞込処理により、文字が N 個以上一致しているデータに検索対象を絞り込んだ上でスコア計算を実施することで、検索時の計算量の削減を実施した。

これらの対策の実施については、AI エンジン等の精度に影響を及ぼすおそれがあるところ、表 5.6-2 から表 5.6-5 に、対策実施前と実施後の精度評価結果を示す。Recall@k を用いて評価を実施した業務課題 1、2、3 については、Recall@k=100 において 0.2pt から 2.2pt の精度低下に留まっており、大きく精度低下しないことを確認した。また、業務課題 4 について、いくつかのクエリにおいて上位の結果をサンプルして確認したところ、検索結果の順序に変動はあるが同じ種別の文字要素の組合せが対策後も検索できていることを確認した。

表 5.6-2 業務課題 1 におけるリソース効率化のための対策実施前後の精度（手法 1-(2)）

@k	全体精度		対策実施前後の比較
	対策実施前*1	対策実施後(次元削減)*2	
1	0.0%	0.0%	0.0%
~3	34.9%	32.9%	-2.0%
~5	48.5%	47.2%	-1.3%
~10	62.9%	62.5%	-0.4%
~50	81.0%	80.8%	-0.2%
~100	85.9%	85.7%	-0.2%
~1000	96.4%	95.6%	-0.8%

*1:対策 1-1 を実施した結果を記載

*2:対策 1-1、対策 1-2 を実施した上で次元削減を実施した場合の精度を記載

表 5.6-3 業務課題 2 におけるリソース効率化のための対策実施前後の精度（手法 2-(3)）

@k	全体精度		対策実施前後の比較
	対策実施前	対策実施後(絞込処理)	
1	54.1%	50.9%	-3.2%
~3	55.6%	53.4%	-2.2%
~5	56.9%	54.4%	-2.5%
~10	57.8%	55.6%	-2.2%
~50	60.6%	58.8%	-1.8%
~100	61.3%	59.1%	-2.2%
~1000	63.4%	60.0%	-3.4%

表 5.6-4 業務課題 3 におけるリソース効率化のための対策実施前後の精度（手法 3-(3)）

@k	全体精度		対策実施前後の比較
	対策実施前	対策実施後(次元削減)	
1	0.0%	0.0%	0.0%
~3	18.1%	18.6%	-0.5%
~5	21.9%	22.4%	0.5%
~10	29.5%	29.0%	-0.5%
~50	37.1%	37.1%	0.0%
~100	42.4%	41.9%	-0.5%
~1000	59.5%	59.0%	-0.5%

表 5.6-5 業務課題 4 におけるリソース効率化のための対策実施前後の検索結果（手法4-(3)）

事例 1			事例 2		
クエリ：金沢市 文字要素：地名			クエリ：木下のパーキング 文字要素：氏＋商品役務の一般名称		
順位	検索結果		順位	検索結果	
	対策実施前	対策実施後（次元削減）		対策実施前	対策実施後（次元削減）
1	金沢市	金沢市	1	木下のパーキング	木下のパーキング
2	福島市	福島市	2	三井のリパーク	前田のチヨコレートレーズン
3	沼津市	沼津市	3	前田のコラーゲン	前田のコラーゲン
4	徳島市	山梨市	4	前田のチヨコレートレーズン	西川の安心マーク
5	山梨市	徳島市	5	三井のパーソナルプラン	三井のパーソナルプラン
6	仙台市	仙台市	6	西川の安心マーク	前田のクラッカー
7	諫早市	諫早市	7	前田のクラッカー	三井のリパーク

8	さいたま市	土佐市	8	あたり前田のクラッカー	野村のレベルフィー
9	静岡市	さいたま市	9	酒井の茶釜	あたり前田のクラッカー
10	土佐市	小野市	10	野村のレベルフィー	貴石のフーガ

また、特許庁職員によるシステム操作にあたり、ユーザビリティの観点において画面のレイアウトや検索結果の表示の仕方、特許庁職員の同時アクセス人数による性能等における評価や改善点等のフィードバックがあったため、システム自体をさらに改善できる点がある。

6. 総合分析

本事業の目的である、商標審査業務における先行文字商標の検索業務を実施する際の業務課題を解決するための、最適な AI 等の手法を検討することについて、総合分析では本事業の調査・検証結果を踏まえ、業務適用に向けた有効性、及び今後の検討課題等について整理する。

総合分析の観点については表 6.1-1 に示す。

表 6.1-1 総合分析の観点

#	観点
1	本事業で構築・検証した各業務課題に対応する AI モデルの有効性
2	本事業で構築・検証した AI モデルを搭載したシステムの有効性
3	実用化に向けて今後解消していく必要がある課題

6.1. 本事業で構築・検証した各業務課題に対応する AI モデルの有効性

(1) 業務課題 1 の AI モデルの有効性

業務課題 1 では 2 つの手法で評価・改善を実施した。手法 1-(1)のランキング学習（Ranking SVM を利用した）の AI モデルでは検証用 DB 全件に対して類似度を算出すると計算処理時間が 8 秒以内に収まらないことから、手法 1-(2)の距離学習（Attention と deep metric learning を利用した）の AI モデルによって検証用 DB 内のレコード全件に対して類似度を算出し、その上で上位 1 万件に対してランキング学習の AI モデルでリランキング（類似度を算出）を実施する方法に変更して、精度を確認した。

リランキング実施前後とも、検索クエリに対して検索結果上位 100 件以内に、類似性を判定したいデータの 8 割以上が検索結果として表示されて同等の精度となったが、上位 10 件以内の精度は手法 1-(2)の方が精度は高く、上位 1000 件以内の精度は手法 1-(1)の方が精度は高かった。

改善として、音の違いや組み合わせ等の特定の条件に対して学習が不足していた点から、条件に該当するデータを生成して学習データを増やす改善方法を実施し、精度向上が確認できた。そのため、今後も不足する条件の学習データを中心に増やしながら継続した学習を実施していくことで更なる精度向上が期待できると考える。

また、現行の商標検索システムでは称呼の音の違いに関する事前処理を実装しているため、その事前処理のパターンを参考にした後処理も実装した結果、こちらも精度が向上したことが確認できた。

処理速度については、手法 1-(1)は類似度の算出に長い時間を要することに対して、手法 1-(2)の検索平均時間が 4.45 秒という結果から、手法 1-(2)は業務運用でも利用できる範囲の速度であることを確認できたため、手法 1-(2)の有効性を確認できた。

学習については、各称呼基準で学習データ量に偏りがあるものの、各称呼基準において類似とする音の組合せのパターンは一定量含まれているため、業務運用的に有効と確認できた。学習については、各称呼基準で学習データ量に偏りがあるものの、学習データを増やす改善を実施したこともあり、各称呼基準において類似とする音の組合せの条件は多くを網羅できたことから、業務運用的に有効と確認できた。

この結果、業務課題 1 においては、一部検証対象外とした称呼基準が有るものの、本事業で構築・検証した AI モデルの業務における有効性を確認できた。

(2) 業務課題 2 の AI モデルの有効性

業務課題 2 では、当初 3 つの手法で検証を予定していたが、貸与データを分析した結果、文字列の構成要素に意味を持たない単一の文字や略語であるデータが含まれていた。文字列の各構成要素が意味を持つ日本語や英語の単語等、定義に基づいた構成の上で、文字要素に分割して類似性を算出する NER では、分割が難しいことが想定されたという理由により、NER を利用した 2 つの手法の有効性が低いと判断したことから、ルール処理である共通部分文字列の検出を利用した手法のみで評価・改善を実施した。

クエリと正解の間で、文字種別が異なる関係や長音の表記の違い等が、全体精度に影響を与えてしまう点から、改善の方法として、検索用商標を用いた検索から称呼を用いた検索に変更し、文字種別を統一する方法を実施したところ、全体精度は大きく向上した。

しかし、改善の実施前と実施後の精度を比較したところ、クエリと正解が同じ文字種別の場合に精度が低下していることがわかった。

業務課題 2 では、単に文字の位置の入れ替えを踏まえた検索をできるようにすることが目的であるため、また、業務運用上においては、まずはクエリと正解が同じ文字種別の場合に、より精度を高くした方がよいことから、改善前の検索用商標を用いた検索手法の方が、有効性が高いものと思われる。

処理速度については、検索平均時間が 6.67 秒という結果から、業務運用でも利用できる範囲の速度であることを確認できた。

学習については、ルール処理の手法であることから、学習が不要のため、業務運用的に有効と確認できた。

この結果、業務課題 2 においては、同じ文字種別同士の検索では、本事業で構築・検証した AI モデルの業務における有効性を確認できたが、異なる文字種別同士では、現時点で業務における有効性の判断は難しく、改善の方法を継続して検討する必要

がある。

(3) 業務課題 3 の AI モデルの有効性

業務課題 3 では BERT を利用した手法 3-(3)が最も精度が良かった。また、同手法では、クエリと正解が同じ文字種別同士での精度は高かったものの、クエリと正解が異なる文字種別同士での精度は低い結果となった。

文脈も考慮した上で学習して、単語間の類似度を算出する BERT は、自然言語処理においては代表的な事前学習モデルではあるが、本事業の検証においては、他言語間での翻訳関係や、日本語のローマ字表記、英単語のカタカナ読み等、いくつかの関係性に対して「意味の関係性がある」と捉えきれていないことが分かったため、それぞれの関係性のパターンに対応した辞書ツールを用いてスコアを算出する改善の方法を実施した。

しかし、全ての関係性に対して 1 つのモデルで対応することは難しく、部分的な改善のみや全体精度を低下させる結果となったため、業務課題 3 においての観念の類似を判断したいパターンの整理を実施していくことが必要と考える。

処理速度については、検索平均時間が 4.47 秒という結果から、業務運用でも利用できる範囲の速度であることを確認できた。

学習については、BERT は事前学習済モデルであることから、多くの学習データを要する必要がないため、業務運用的に有効と確認できた。

この結果、業務課題 3 においては、同じ文字種別同士の検索では、本事業で構築・検証した AI モデルの業務における有効性を確認できたが、異なる文字種別同士においては、現時点で業務における有効性の判断は難しく、改善の方法を継続して検討する必要がある。

(4) 業務課題 4 の AI モデルの有効性

業務課題 4 では、いずれの手法によっても Recall@k での精度評価が難しいことから、検索結果について目視確認による追加評価を実施して精度を確認した。

こちらも BERT を利用した手法を中心に評価・改善を実施したところ、14 件の構成要素の種別事例のうち、10 件がクエリと同じ種別の文字要素の組み合わせを上位に検索できており、4 件が上位に検索できていない結果となった。また、課題 3 と同様に、表記の違いが精度を低下させる原因と捉え、辞書ツールを用いてスコアを算出する改善の方法を実施したが、部分的な精度向上の一方で、精度を低下させてしまう事例もあった。

処理速度については、検索平均時間が 4.43 秒という結果から、業務運用でも利用できる範囲の速度であることを確認できた。

学習については、課題 3 と同様に BERT は事前学習済モデルであることから、多く

の学習データを要する必要がないため、業務運用的に有効と確認できた。

この結果、業務課題 4 においては、漢字で表記されるものは漢字、ひらがなで表記されるものはひらがなで表記される等、想定していた表記の事例においては、本事業で構築・検証した AI モデルの業務における有効性を確認できたが、通常は漢字で表記されているものが、ひらがなやカタカナで表記される等の想定外の表記の事例については、検索時の表記の統一の仕方やデータの持ち方等も踏まえて、改善方法を継続して検討する必要がある。

6.2. 本事業で構築・検証した AI モデルを搭載したシステムの有効性

本事業の検証において、AI エンジンの API 動作時のハードウェアのメモリが 66GB 程度となったため、現行の特許庁アジャイルシステムの動作環境に合わせるために、容量の削減の対策として、AI モデルに対して主成分分析及び次元削減を実施した。実施結果としては、AI モデルの精度を大幅に低下させることなく動作させることに成功しており、各業務課題の検証結果を踏まえると、AI モデルを搭載したシステムとしての有効性を確認できた。

ユーザビリティの観点においては、画面のレイアウトや検索結果の表示の仕方等、特許庁職員の操作によるフィードバックもあり、業務で使いやすいシステムを目指していくにあたり、継続的な改善をしていくことで有効性を高めることができると考える。

しかし、AI モデルの精度を高めていくと、AI モデルを動かしている AI エンジンに必要となるリソースも高まる可能性があることから、システム全体の性能に影響がある。

そのため、今後も AI モデルに対して継続的な検証と改善を進めていくにあたり、システム全体のリソースの監視も必要と考える。改善された AI モデルの精度をなるべく落とさずにシステムに搭載したい、同時に操作するユーザ数を増やしたい等、AI モデルを搭載したシステムの有効性を高めていくには、AI モデルの精度だけでなくプロダクト全体として安定した稼働を意識しながら、どのような機能を実装し、利用者に価値を提供するかが重要となる。

今回の検証結果をもとに、AI モデルの性能を把握した上でユーザに提供する機能を検討し、ユーザからのフィードバックを得ながら継続的に改善していくことが望ましい。

6.3. 実用化に向けて今後解消していく必要がある課題

今後の実運用に向けて、3 つの観点で以下に解消すべき課題を整理する。

まず 1 つ目は、AI モデルを運用していく上での体制に関する課題である。

AI モデルの精度を保つ、もしくは上げるために学習や改善を実施していくことが必要であり、運用の中で継続的に学習や改善が可能なシステムや体制等の整備が求められると考える。

継続的な AI モデルの学習や改善、システムの整備を行うためには、専門的なスキル

を持った人材を配置することが肝要であり、例えば、データサイエンティストや AI 技術を活用できるソフトウェア開発者及びソフトウェア運用者、技術監修者（アドバイザー）の整備は、実用化に向けての課題として考えられる。これらの必要なスキルを持った人材を配置した後に、AI モデルの学習等に利用するデータの整備や分析、システム構成の見直しや AI モデルの軽量化などに取り組めるようになるものと考えている。

2 つ目として、AI モデルによる抽出結果と業務上求められる抽出結果との乖離に関する課題である。

例えば、データの整備や分析において、同じ表記の先行文字商標に対して、業務に適した順位の出し方やデータの持ち方等の工夫も検討し、同じ表記の先行文字商標でも、業務や検索のパターンによっては類似スコアに差をつけることも考慮する必要がある。

なお、本事業で構築したシステムでは、業務課題毎に AI モデルが算出した結果に対して、ユーザによる評価を登録するフィードバック機能を構築したため、AI モデルの継続的な精度改善を実施するために、AI モデルの改善ができる程度のユーザによる評価データ量を蓄積していくことが必要である。

そのためには、検索を実行した特許庁職員が、検索結果画面においてフィードバック機能を日常的に利用して、なるべく多くの評価を登録して欲しいということを、システム利用者全体に周知することが必要と考える。

3 つ目として、システム全体としての利便性向上に向けた課題である。

AI モデルの精度に対するフィードバック機能のみに限らず、システム全体の利用しやすさという点から、システム利用者の要望や意見を収集していくことも必要と考える。

本事業の中で挙がった代表的な意見として、特に業務課題 1 について、1 つの商標には複数の検索キーが付与されることが多いところ、どの検索キーを重要視するのか重み付けを実施でき、その重みによって検索結果を変えることを実現することで、より特許庁職員が得たい情報を手に入れられるのではないかという意見があった。本事業の検証期間内では対応できなかったが、今後の実運用に向けて貴重な課題となった。

また、本事業では検証できなかったが、LLM を用いて、検索キーワードの提案（例として「りんご」の検索キーワードに対して、「フルーツ」を検索キーワードとして拡張し、ユーザに提案する）や、検索結果に「どの要素で類似したのか」の判断根拠を付与すること等を検証できれば、より効率的な検索が実現できると考える。

その他にも様々なケースに対してシステム利用者からの要望や意見を汲み上げていき、検討と構築を繰り返すことで、実運用で利用しやすいシステムを構築できると考える。

7. おわりに

7.1. 本事業で得られた知見

本事業を通して、先願に係る他人の登録商標等の調査と審査判断の均質性を検討するための調査において、これまでルールベースで実施していたところを、AI モデルを構築することで調査を高度化していくことの知見を得ることができた。

AI モデルに限らず、構築した AI モデルを実際の業務システムに組み込む際の知見や、限られたシステムのリソースの中で探索（推論）を実行するための手法等の知見も得ることができた。

また、本事業の内容に限らず、自然言語処理の開発・運用におけるプロセスや手法、自然言語処理における代表的な技術等について、報告会や勉強会を通して、理解を深めることができた。