

researchmap 情報を対象とした
LLM Embedding を用いた論文タイトル類似度に基づく
研究者マッチングの研究

2021190 椎木 太一

令和 6 年度提出

目次

1章 はじめに.....	2
2章 従来研究.....	3
3章 提案手法.....	4
3.1 研究者マッチングにおける基本的な考え方.....	4
3.2 論文タイトルの Embedding.....	4
3.3 所属学校・学科の名称の Embedding.....	5
3.4 本研究における「類似度」の扱い方.....	5
4章 実験設定.....	6
4.1 実験の目的.....	6
4.2 対象データ.....	6
4.3 実験方法.....	7
4.3.1 実験の手順 及び入力データと出力データ.....	7
4.3.2 実験データの選び方.....	8
4.4 評価方法.....	8
4.4.1 評価対象の論文.....	8
4.4.2 評価における「正解」の定義.....	9
4.4.3 3つの評価軸による評価.....	9
4.4.4 正解率について.....	10
5章 実験結果.....	11
5.1 結果.....	11
5.2 考察.....	14
6章 まとめ.....	17
参考文献.....	17
付録.....	18

1章 はじめに

学術の進展には、個々の研究者が単独で成果を生み出すだけでなく、研究者同士の協力やネットワーク形成が必要不可欠である。研究者間の連携を促進するためのマッチングには、オフラインとオンラインの 2 つのアプローチが存在する。オフラインでは、学会やシンポジウムなどで直接顔を合わせた交流と、共通の知人を介した紹介という 2 つが代表的である。一方、オンラインでは研究者情報を活用し、研究分野や論文のタイトルなどの情報をもとに研究者を探す方法が知られている。例えば、研究者情報を提供している Web サイトの 1 つである「researchmap¹」は、全国の大学の研究者情報を一元的に参照できる。しかしながら、膨大なデータの中から手作業で必要なデータを探さなければならない問題がある。

この問題の解決策として、人手をかけずに、オンラインで研究者同士をマッチングさせる手法が提案されている。廣吉の研究では、研究者の論文タイトルを分析して、そこから類似する研究を行う研究者を複数人出力するといった手法が提案されている[1]。具体的には、自然言語処理の技術である「Sentence-BERT(以下「SBERT」とする)」を用いて論文タイトルの類似度に基づきマッチングの可能性がある研究者を提示している。この研究では、「大学や研究分野の垣根を超えた」研究者同士のマッチングの必要性が提示されていたものの、使用したモデルが日本語のみであり、英語と日本語の論文を比較できないという問題があった。

そこで、本研究では多言語対応を可能とする新たなモデルとして、2022 年に OpenAI 社が開発した OpenAI Embedding² (以下「OpenAI Emb」とする) に着目し、異なる言語で記述された論文タイトルに対しても類似度を測定することで、従来の SBERT と比較し、双方の違いを考察する。

本研究の目的は「**大学や研究分野の垣根を超えて、多言語のタイトルに対応しつつ、類似する研究を行う研究者を見つける**」ことを目的とする。対象となる研究者は、北海道国立大学機構の 3 つの大学である、小樽商科大学・帯広畜産大学・北見工業大学に所属する研究者から実施する。そして論文の類似度を比較して、どのような結果が出るのかを分析する。

¹ <https://researchmap.jp/>

² <https://openai.com/index/new-and-improved-embedding-model/>

2章 従来研究

馬、松田の研究[2]では、SBERT と OpenAI Emb の ada-002 のモデルを用いて、宿の検索に関する研究を行った。ここでは結果として、ada-002 は「クエリの核心的要望や打ち消し表現を無視し、修飾・理由・無意味な単語にミスリードされる傾向がある（[2]、p2474 より）」というものと「特定分野の知識が備えていないため、クエリにある特定分野の専門用語を正確に解釈できないことがある（[2]、p2474 より）」という 2 点を示した。つまり視点と知識量の 2 点で SBERT よりも劣ると結論付けた。また、旅行関係の専門用語に ada-002 は対応できず、精度が低い検索を行うと論じられていた。一方優位に立っている点として「古文・外国語・絵文字といった一般的ではない表現を正確に解釈できている（[2]、p2474 より）」とした。しかし著者は宿検索で重要な点は、視点と知識量であるとし、その点から SBERT の方が検索に優れていると結論付けた。

しかし[2]の研究では、宿や観光などの分野における Embedding や類似度検索の有用性のみに言及しているため、様々なテーマが含まれる論文タイトルを題材としたときに、どのような結果や違いが表れるのかを明らかにしていきたい。

小林らの研究[3]では、研究者の論文調査や執筆を支援する、論文推薦システムに関して類似性の高い論文を提示することを目的として、OpenAI Emb などを用いて比較と分析を行った。結果として、OpenAI Emb で最も高い性能を示したと述べられ、「このモデルは学術論文を扱うタスクに特化している訳ではなく、他の IR タスクでも高い性能を発揮していることから、分野を問わない高い汎用性を持つと考えられる（[3]、p382 より）」とした。

しかし[3]の研究では、どのような学術論文が対象となったかの具体的な記述がなく、どの分野や領域を取り扱ったのかが不明瞭である。そのため、分野を問わないという記述の真偽を改めて明確にする必要性が高いのではないかと考えた。

上記の全ての研究[1][2][3]において、OpenAI Emb と SBERT との比較が行われており、それぞれが別々の特徴を持っており、論文に関する分析は OpenAI Emb の方が優れているとの見立てが示されている。一方詳細に関しては言及が少なく、どのような文章の分析が優れているのかなどの条件等が明記されておらず、OpenAI Emb の方が優れているという明確な理由を把握できなかった。本研究では、従来研究を調査した上で感じた疑問等を明らかにするという目的でも分析を行う。

3 章 提案手法

3.1 研究者マッチングにおける基本的な考え方

マッチングにおける基本的な考え方として、以下の3点を前提とする。

1. 同じ人物とのマッチングを避けること
2. 意外性を有するマッチングであること
3. 最終的にマッチングする相手を判断するのは人間であること

一つ目については、同一著者を除外しなければ、自分とマッチングするという意味のない結果になってしまう。そのため本研究は分析対象となる論文の著者と同じ人物が執筆した論文は、分析を行わない。

二つ目の考え方が最も重要なものであり、本研究では常にこれを意識した考察を行う。まず意外性あるマッチングとはどういった事かについて定義する。これは言い換えれば「対象論文と結果として出力された論文のタイトルは似ているが、互いの研究者の研究分野は離れている」ということを意味する。1章で研究者はオフラインで関係を構築するには、学会などのイベントに参加・交流すると説明した。しかしこのような場では自分と似た研究分野を取り扱う研究者が集まりがちである。オンラインでも、自分と親和性の高い研究者を探すのならば、自分の研究分野から参照して探す人がほとんどになるであろう。このように従来手法では意外性あるマッチングは実現が難しい。そのため本研究では意外性あるマッチングにフォーカスを当てて、研究と考察を行う。

三つ目について、本研究では類似する論文を出力するのは機械やツールを用いて行うが、最終的には人の目によってその結果が見られ、マッチングする相手が判断されるということである。これは恋愛を目的としたマッチングアプリやサービスも同様である。ユーザーのプロフィールや自己紹介を収集・掲載しているのは機械やシステムであるが、誰とマッチングするかを判断する際には必ず人の目を通すことになる。本研究ではこの「人の目で判断すること」を重視して考察を行う。

3.2 論文タイトルの Embedding

前提として、Embedding とは文章をベクトル形式で数値化し、分析や類似度の比較などを可能にする技術である。ベクトルの個数を次元数と言い、Embedding のモデルごとに次元数は異なる。また類似度の実数値や基準値なども異なる。

では論文タイトルの Embedding について説明する。マッチングの手法として、類似している論文タイトルから参照するため、この作業は必須である。本研究は Embedding 可能な大規模言語モデルが搭載されている、自然言語処理の技術を用いて類似度を算出する。

3.3 所属学校・学科の名称の Embedding

本研究では、[1]と差別化をしている点として、論文タイトルの類似度に注目するだけでなく、所属学校・学部学科等の所属に関しても着目を行う。これを行う理由は「**異分野同士のマッチングが出来たかどうかを確認する指標として用いるため**」である。Embedding を行うことで、どの学部とどの学部が近く判断されたり、遠く判断されたりするのかを可視化することができる。本研究にて取り扱う学部名称には、専門用語やその名称に関連する知識を持たない人には、類似しているかの判断が難しい学部名も存在する。例えば、「グローバルアグロメディシン研究センター」や「原虫病研究センター」などである。このような名称の所属の研究者を対象とした際に、離れているか否かを判断する助けとして、Embedding による類似度を参考として考察を行う。

なお、所属学校・学科の名称の Embedding は異分野の研究者同士でのマッチングが出来たかどうかを測るために、あくまで参考のデータとする。Embedding によって算出された類似度をそのまま用いることはしない。

3.4 本研究における「類似度」の扱い方

本研究では便宜上「類似度」という単語を用いて説明を行っているが、Embedding によって算出された類似度の数値をそのまま用いることはしない。理由としては 3.1 で説明した通りではあるが、本研究が「**類似している論文かどうかは人が判断することを重要視しているから**」である。本研究では実験に当たって多数のデータを閲覧したが、類似論文として機械やツールが出力したデータであっても、人間がその結果を見た場合、視点や論点が異なると考えられる論文が出力されるケースが複数個あり、Embedding の技術における類似度を鵜呑みにしてはならないと考えたからである。また数値も小数点第 5 位まで詳細に出力されるため、データを表として可視化したり、説明を行ったりする際に、視認性や理解度の観点で複雑化してしまい理解しにくい論文となってしまう。誰もが理解できるような論文の執筆に配慮を行うために、類似度をそのまま用いることはしない。また、最終的に誰とマッチングするかを判断するのは人であるという点も考慮し、類似度の数値をそのまま用いない。

もし仮に類似度をそのまま用いた場合に懸念されることを説明する。前述の通り、説明時の視認性や理解度が低下してしまうという点も挙げられる。また、OpenAI Emb と SBERT で類似度に差が表れてしまい、値をそろえる必要がある点がある。これは Embedding モデルの仕様に関わる話であるが、OpenAI Emb によって算出された類似度は SBERT で算出された類似度よりも 0.1~0.15 ほど高く出力される。この値を均一化する際、手間がかかるという点と、本来の類似度と乖離してしまい、正しく分析できなくなってしまう懸念がある。したがって本研究では、論文と所属の両方で、類似度をそのまま用いることはしない。

4 章 実験設定

4.1 実験の目的

本研究および実験の目的は三つある。以下に、それぞれについて順に説明する。

第一に、「**大学や研究分野の垣根を超えること**」である。これは 3.1 で説明した通り、所属学科の名称から、論文著者の研究分野が互いに離れていると判断できるという結果が多数みられる場合、達成されたと言える。この目的の設定理由は、大学が同じ研究者や、同じもしくは似た研究分野の研究者同士がマッチングしても、目新しさがない点やすでに関係を構築していたり、共同で研究を行ったりしている可能性があるという懸念があるためである。従来の研究者同士の関係構築の手法と、本研究による手法が差別化できる手法であり、かつ後者が有用な手法であるかどうかを明らかにするために、目標を設定した。

第二に、「**多言語のタイトルに対応する**」である。[1]では日本語のみの分析にとどまったため、対象となるデータ数が少なかった。そこで本研究では多言語に対応できる Embedding の手法を二種類用いて、それぞれの比較を行う。一つ目は OpenAI Emb のモデルの 1 つである、「text-embedding-ada-002」を用いる。次元数は 1536 次元となる。二つ目は SBERT のモデルの 1 つである、「paraphrase-multilingual-MiniLM-L12-v2」を用いる。次元数は 384 次元である。

二つの手法の比較は[2]でも行われていたが、これは宿検索における文章の分析のみを分析している。本研究では宿や観光に関わる言葉以外にも、様々な言語・単語・専門用語が用いられる文章である、論文タイトルを分析する。ここでどちらの手法がより有用な結果が出力されるのかについて検証を行う。また多言語の分析でも有効な結果となるかを調査する。

第三に「**人の目で見ても、類似する研究を行っていると判断できる研究者を見つける**」である。これは 3 章でも説明した通り、最終的にマッチングする相手を判断するのは人間である。人の目で判断しても、対象論文と結果として出力された論文は似ていると判断できれば、目的は達成され则认为る。本研究では OpenAI Emb と SBERT においてどちらが、類似論文の出力に優れているかを検証する。

4.2 対象データ

対象となる研究者は researchmap 上に論文タイトルを 1 つ以上掲載しており、なおかつ小樽商科大学・帯広畜産大学・北見工業大学のいずれかに所属している研究者とする。したがって、小樽商科大学 83 人、帯広畜産大学 99 人、北見工業大学 81 人となり、論文本数は合計 11,495 本となった。

三大学の研究者のデータは researchmap 上にあり、研究者個人のページに遷移すると論文という項目があり、ここに研究者の論文一覧が掲載されている。URL は以下となる。

https://researchmap.jp/〇〇/published_papers?limit=100

ここの“〇〇”の部分は研究者ごとに異なる。本実験はこの部分を抽出し、csv に一覧化する。そして一覧から 1 名ずつ〇〇の箇所に挿入して、研究者個人の JSON ファイルを取得する。実際の JSON ファイルの中身はこうになっている。

```
"paper_title": {"ja": <中略>
                  "en": "Construction of a Dataset for Extracting
                        the Relationship between Text and Tables for Securities Reports"},
"authors": {"ja": [<中略>],
             "en": [{"name": "Yasutomo Kimura"}]}
]
```

paper_title というラベルの中に論文タイトルが格納されている。その後に、authors というラベルがある。そこに著者名が格納されている。これが、その研究者が researchmap 上に掲載している論文の件数分、記載されている。日本語と英語の片方もしくはその両方に記載がある。実際は researchmap に掲載された年月日や URL などの情報も詳細に記載されているが、それらの情報は本研究で直接使用することや、プログラムによる解析を行うことはしない。データとして使用する部分は“paper_title”と“authors”に含まれる部分である。

“paper_title”内の論文タイトルは、Embedding による類似度計算と Web サイトへの記載に用いる。“authors”内の著者名は、Embedding による類似度計算における、同一著者との類似度計算を除外するために用いる。また Web サイトへの記載に用いる。

4.3 実験方法

4.3.1 実験の手順 及び入力データと出力データ

本節では実験の手順について説明する。4.2 節で説明したように、抽出した JSON ファイルに論文タイトルが格納されている箇所“paper_title”がある。そこに対して Embedding を行う。同時にベクトルを格納する部分を作り、そこにベクトルを挿入する。まとめると、researchmap から取得した研究者のオリジナル JSON ファイルに OpenAI Emb と SBERT で生成したベクトルを挿入した JSON ファイルを 1 つずつ生成する。研究者 1 人当たり 2 つの Embedding 済みの JSON ファイルが存在することになる。

次に、論文タイトルの類似度計算を行う。OpenAI Emb と SBERT はそれぞれ別々で実施する。これはすべての論文に対して実行する。入力データは、無作為に選ばれた研究者の最新の論文 1 件である。対象となる論文を選択し、「その著者が執筆した論文を全て計算対象から除外」したうえで、その他全ての論文から類似度を計算する。これが出力データに当た

る。対象論文の著者と同一著者の論文を除外する理由は、3.1 で説明した通り、研究者マッチングの意義を成立させるため、自分とマッチングするという無意味な結果を出力することを防ぐためである。そして上位 10 件を Web サイト上に表示する。これは結果を可視化して、分析を容易にするために行う。

改めて説明すると、入力データは、無作為に選ばれた研究者の最新の論文 1 件である。この論文を対象として類似度計算を行い、Web サイトに出力された類似度上位 10 件の論文が出力データである。

4.3.2 実験データの選び方

研究者の選び方と論文の選び方を説明する。

まず研究者の選び方は、4.2 で説明した通り、researchmap 上に論文タイトルを 1 つ以上掲載しており、なおかつ小樽商科大学・帯広畜産大学・北見工業大学のいずれかに所属している研究者を対象とする。そして各大学から 10 人ずつ、合計で 30 人選出する。

本研究では下記の研究者 30 人分の論文を用いる。

表 1：対象論文著者 一覧表

小樽商科大学	帯広畜産大学	北見工業大学
木村泰知	鈴木宏志	小原伸哉
佐山公一	谷昌幸	川口貴之
國武英生	加藤清明	森田慎一
沼田ゆかり	宮竹史仁	裡しゃりふ
鈴木和宏	橋本直人	林田和宏
近藤公彦	木村賢人	安井崇
片桐由喜	斉藤準	野田由美意
猪口純路	上村暁子	早川博
竹村壮太郎	齊藤朋子	本間雄二郎
須永将史	平舘善明	早川吉彦

論文の選び方は、無作為に選択した研究者の最新の論文 1 件を対象とする。理由は 2 点ある。まず古い論文を対象とすると、学問や技術の発展によって、研究内容や用いている手法や技術が最新でなくなっている可能性があるためである。次に、研究者の研究テーマや専門分野が現在と異なる可能性があり、マッチングが無意味になってしまう懸念がある点である。この 2 点の発生を防ぐために、最新の論文を用いることとする。実験データは、三大学に所属する研究者各 10 人（計 30 人）の最新論文 1 件（計 30 件）を用いることとする。

4.4 評価方法

4.4.1 評価対象の論文

本研究では、Web サイト上に表示された上位 10 件の類似論文のうち、上位 3 件の論文に着目する。理由としては、上位 4 件以降の論文は、明らかに似ていないとすぐに判断できる論文が多くなってしまふ点と、対象論文 1 件当たり 10 件も分析を行うと、時間と手間が多数かかってしまうという点が挙げられるためである。

4.4.2 評価における「正解」の定義

本研究では、出力結果に対する「正解」を定義し、2 つの手法でどちらが優れているのかを検証する。正解とは「対象論文と結果として出力された論文のタイトルは似ているが、互いの研究者の研究分野は離れている論文」を指す。下記に例を示す。

表 1：類似論文の出力結果における「正解」例（OpenAI Emb）

	タイトル	著者	類似度	所属
対象論文	北海道における新エネルギーの成長可能性	江頭進氏	—	小樽商科大学 商学部経済学科
最も類似	再生可能エネルギーの広域連系による平準効果を考慮した北海道の電力・熱供給ネットワーク	小原伸哉氏	0.90993	北見工業大学 工学部 エネルギー総合工学コース

この例では、タイトル全体で見た場合、「北海道における新エネルギー」と「再生可能エネルギー」・「北海道」という部分で似ていると考えられる。そして対象論文著者は、小樽商科大学の商学部・経済学科所属の江頭氏であり、本実験で取り上げた論文は北見工業大学の工学部・エネルギー総合工学コースの小原氏である。経済学とエネルギー分野の工学の組み合わせは異分野同士のマッチングで意外性があるものであると評価できる。このような類似論文を正解と定義し、考察を行う。ちなみに、所属名称については各大学の研究者総覧に記載されている所属名に準拠する。その研究者がどこに所属しているかのみを考慮し、どのような授業や研究を担当しているかは考慮しない。例えば、研究者が生物学に関わる授業を担当していたり、生物学に関わる論文を執筆していても、所属が一般教養であれば、一般教養の箇所に注目して考察を行う。本実験の例で説明すると、対象論文と類似論文のテーマはどちらも、北海道のエネルギー事情である。しかし江頭氏は経済学の研究者であり、小原氏は工学の研究者であるため、経済学者と工学者同士のマッチングとして取り扱う。前述したが、この組み合わせは領域が異なることは明らかであるため、異分野同士のマッチングと言えるであろう。

4.4.3 2つの評価軸による評価

正解を判断するうえで、評価軸を明確に設定して、フィーリングなどで評価を行うことを防ぎ正当な評価を行うことが重要である。この評価軸を 2 つ設けて、このルールに従って評価を行って、正解を判断する。

配点についてはそれぞれ 0~2 点とする。評価軸が 2 つあるため、合計点は 0~4 点となる。類似論文として出力された 3 件の論文を採点して、最も高い点数となった論文を採用して考察に用いる。

論文の評価は各手法につき 30 人分行うので、合計点の最大は 120 点となる。120 点満点中何点かを比較し、どちらの手法が優れているかを調査する。

評価軸 1 つ目は「**関連している研究であるかどうか**」である。これは Embedding のモデルの単純な類似度算出の性能を測るために設ける。配点基準と点数は以下の通りである。

- 2 点 かなり関連している
- 1 点 少し関連している
- 0 点 関連していない

4.4.2 で示した例を再び用いると、両方の論文が、北海道における新たなエネルギーに関する研究であると読み取ることが出来る。これはかなり関連した研究と言えるであろう。したがって配点は 2 点とする。評価軸 1 つ目ではこのように配点を行う。

評価軸 2 つ目は「**異分野のマッチングであるかどうか**」である。これは研究目標の 1 つである、垣根を超えた研究者同士のマッチングが実現できているかどうかを検証する。配点の手法は下記の通りとする。

- 2 点 大学も学部学科も異なる
- 1 点 大学と学部学科のどちらかが異なる
- 0 点 大学と学部学科のどちらも同じ

4.4.2 の例では、対象論文の著者の江頭氏は、経済学科の研究者であり経済学に明るい研究者であるとわかる。一方、類似論文の著者の小原氏は、工学部の研究者であり、工学の特にエネルギーの部門に専門性が高い研究者であるとわかる。経済学と工学では領域が明らかに異なる異分野のマッチングであると言える。

4.4.4 正解率について

本研究では、合計点数が 3 点と 4 点の論文を「正解」とし、対象論文 30 件のうち、各 Embedding 手法で何件の正解を出力したかを判断する。3 点と 4 点を正解とした理由は、「**2 つの評価軸の両方を少なくとも満たす結果**」と判断できる点数が 3 点と 4 点だからである。この正解率は、2 つの Embedding 手法のどちらが、論文タイトルを用いた研究者マッチングに向いているかを判断するために用いる。

5 章 実験

5.1 結果

表 2：小樽商科大学 OpenAI Emb の採点結果

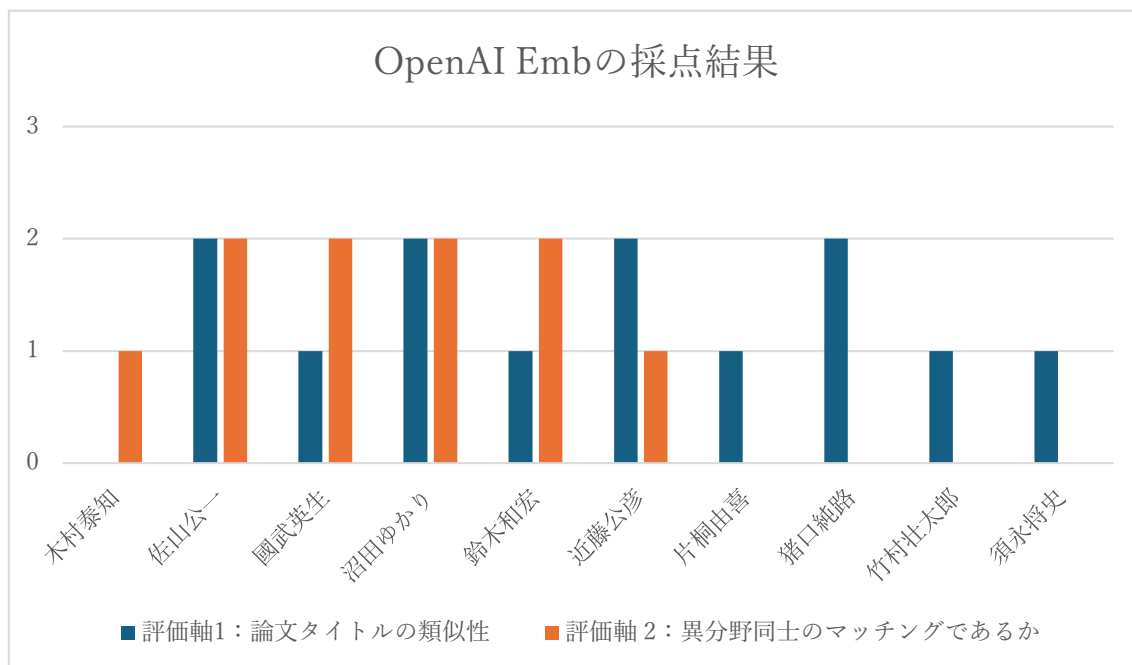
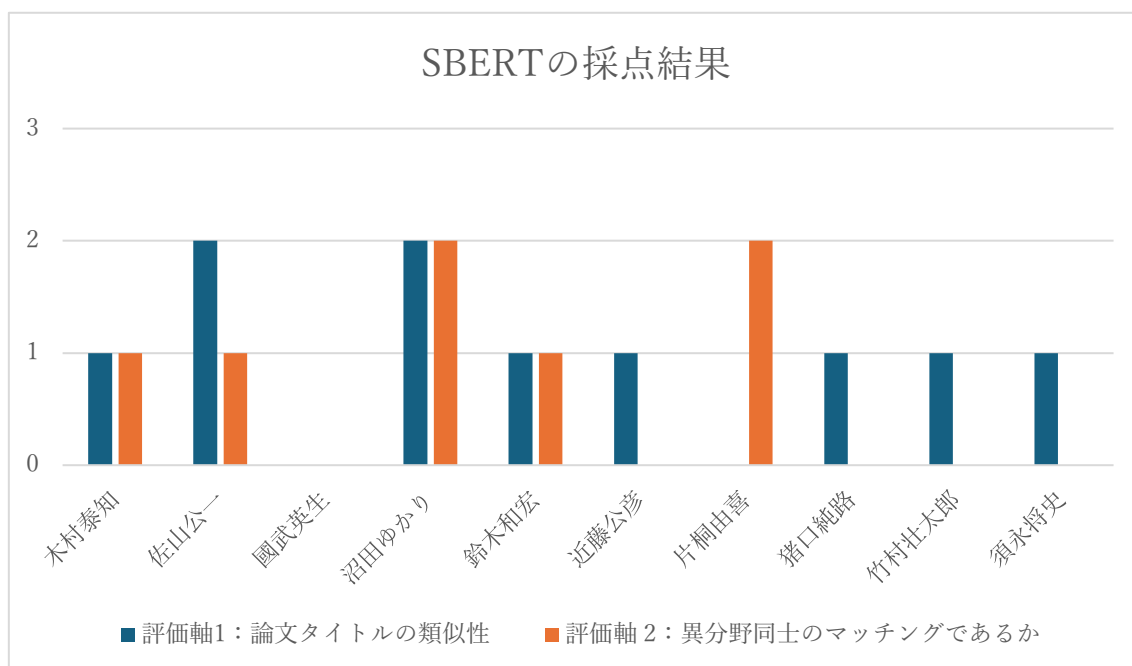


表 3：小樽商科大学 SBERT の採点結果



点数 OpenAI Emb : 23/40 点、正解率 50% (評価軸 1 : 13/20 点、評価軸 2 : 10/20 点)

SBERT : 17/40 点、正解率 20% (評価軸 1 : 10/20 点、評価軸 2 : 7/20 点)

表 4：北見工業大学 OpenAI Emb の採点結果

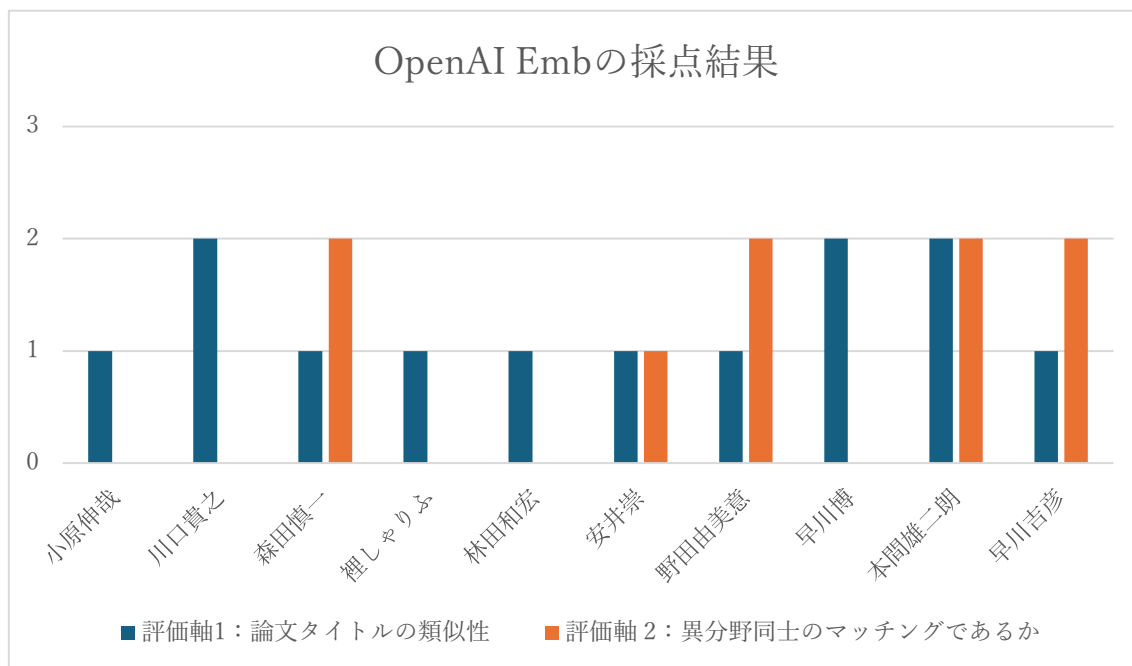
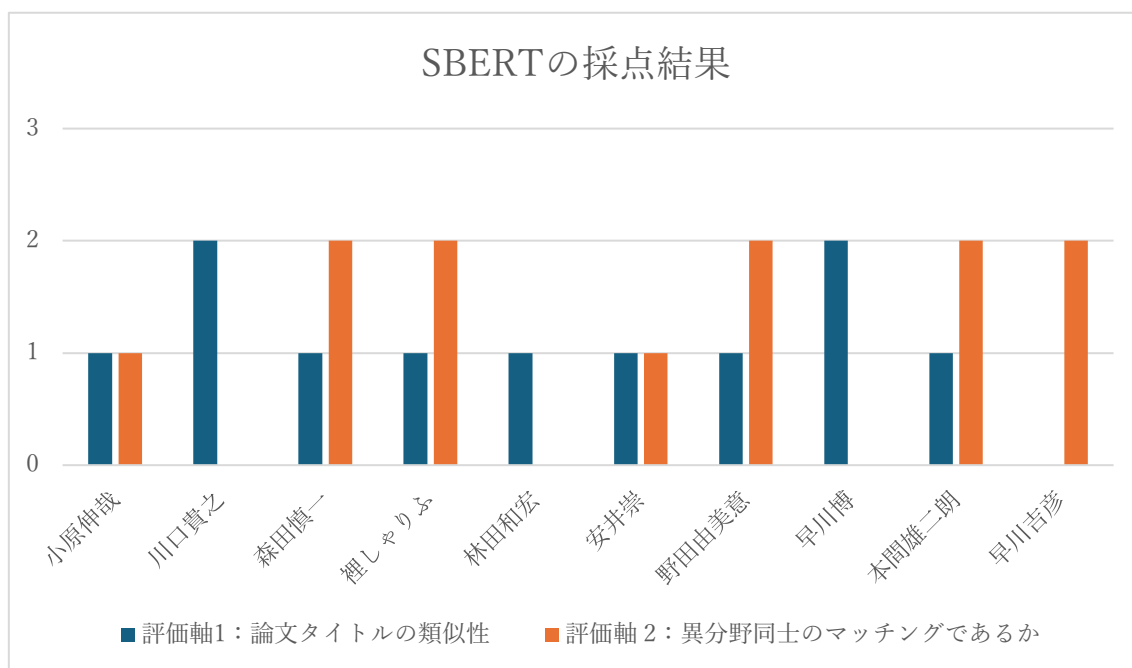


表 5：北見工業大学 SBERT の採点結果



点数 OpenAI Emb：22/40 点、正解率 40%（評価軸1：13/20 点、評価軸2：9/20 点）

SBERT：24/40 点、正解率 40%（評価軸1：12/20 点、評価軸2：12/20 点）

表 6：帯広畜産大学 OpenAI Emb の採点結果

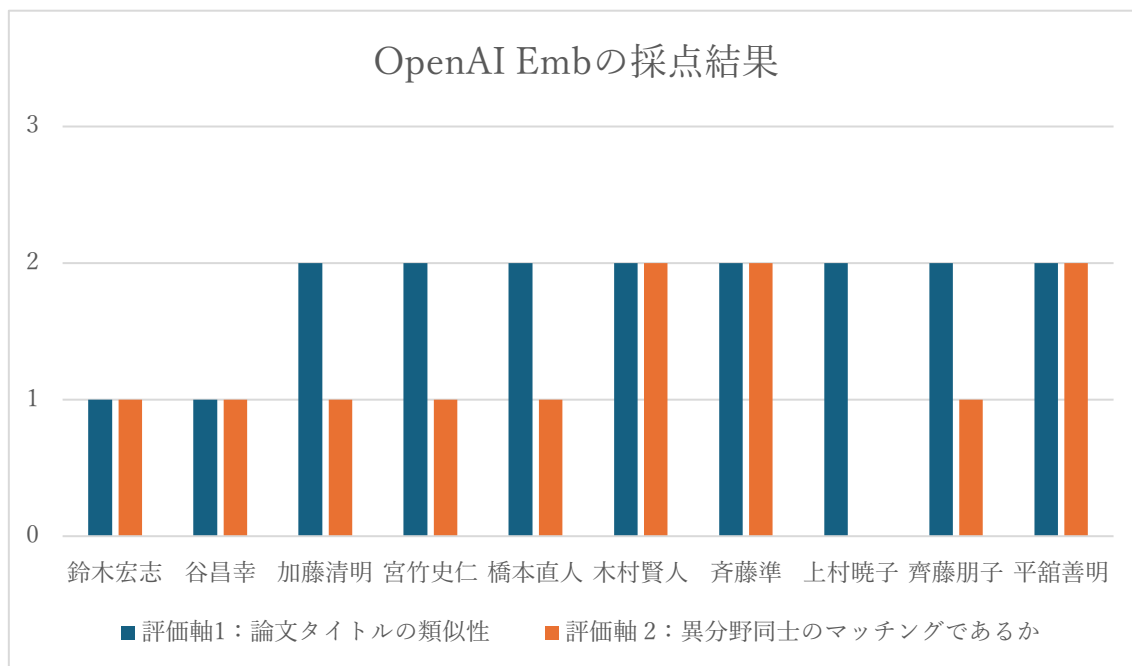
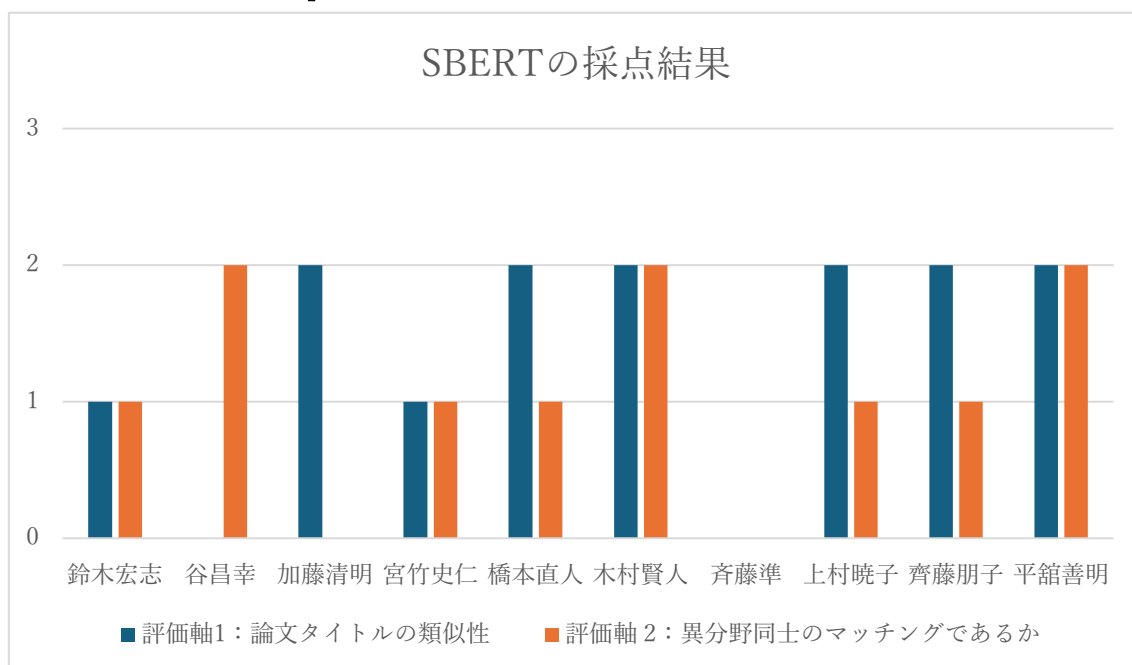


表 7：帯広畜産大学 OpenAI Emb の採点結果



点数 OpenAI Emb：30/40 点、正解率 80%（評価軸 1：18/20 点、評価軸 2：12/20 点）
 SBERT：25/40 点、正解率 50%（評価軸 1：14/20 点、評価軸 2：11/20 点）

5.2 考察

OpenAI Emb は SBERT と比較して正解率で優位性を示した。

本項目では採点した点数、特に正解率に着目した上で考察を行う。考察は「2つの評価軸の各点数」・「対象論文及び類似論文の著者の所属」「論文タイトルに含まれる言葉や、タイトルから想定されるテーマや研究分野」などの観点から行う。そこからなぜ評価点数が高くなったのか、もしくは低くなったのかを理由とともに分析を行う。最後に OpenAI Emb と SBERT を比較して、Embedding における特徴などの違いをまとめていく。

小樽商科大学の正解率は 50% であった。評価軸 2 で 1 点以上が付いた研究者の対象論文は、情報系・商学・経済学・生物学に関わる研究を行う研究者であった。特に情報系は北見工業大学に、「機械知能・生体工学コース」「情報デザイン・コミュニケーション工学コース」が存在し、その研究を行う研究者との親和性が高い可能性がある。また生物学も、畜産や微生物などの研究を行う研究者が多く所属している、帯広畜産大学全体と親和性が高い。一方評価軸 2 が 0 点の研究者の多くは企業法学科に所属している研究者であった。確かに法律の研究を行っていると考えられるところは、三大学のうちで小樽商科大学しか存在しないため、意外性のあるマッチングを開拓できないというネガティブな印象があると考えられるであろう。しかし、逆を言えば三大学の中で唯一法律関係の研究を行う研究部門であるともいえる。したがって三大学連携をより有効なものとして活用するには、企業法学科所属の研究者を誰とマッチングさせるかが重要なポイントとなると考える。

北見工業大学は最も正解率が低かった。考えられる理由としては 2 点ある。まずはエネルギーや防災関係の研究を行っている研究者が他大学にあまり所属していないという点である。次に、同じ所属場所に同じ内容の研究を行う研究者が集中するケースが、他大学のそれよりも多い点である。これは所属名称が他大学よりも詳細になっていることが原因であると考えられる。例えば小樽商科大学の社会情報学科の場合、自然言語処理や量子コンピュータなど、研究者の研究領域が多岐にわたっている。これは学科名が大枠を捉えた名称となっているためであると考えられる。しかし北見工業大学では、工学という限定した領域をさらに「エネルギー総合工学コース」や「社会インフラ工学コース」などとさらに詳細に設定している。したがってこれは同じ研究を行う研究者が、同じ所属に集中している要因であると考えられる。これは類似する研究を行う研究者同士で連携がとりやすいというメリットはあるが、本研究の目標である「意外性のあるマッチング」には適さないというデメリットも存在することが分かった。

帯広畜産大学が最も正解率が高く出力された。私は実験を開始する前は、北見工業大学のように、畜産という限定した領域の研究者が多いため、同様の結果となり正解率が低く出力されるのではないかと考えていた。しかし結果は全く異なった。私は要因を 2 つ考察した。1 つ目は OpenAI Emb の類似度算出の特徴にあると考える。これは「動植物の名称が出た場合、それと同様もしくは近い動植物の名称を含む論文を、類似度が高い論文として出力

する」という特徴である。例えば、橋本氏の対象論文には”Rats”という単語があり、最も評価点数が高い論文にも”Rats”という単語が含まれていた。他にも「牛ふん」や”Rice”などといった特徴的な単語が対象論文と最高評価点数の論文の両方に含まれていた。このような単語が帯広畜産大学の論文には特に多く、正解率に影響したのではないかと考える。2つ目は帯広畜産大学の学部名称にあると考える。帯広畜産大学には様々な学部学科があるが、全ての名称に共通することとして、「どの学部も生物のある特徴を研究する研究者が多い」ということである。どの研究者がどこに所属していても、様々な生物の研究を行っているという点に変わりはない。したがって類似する研究を行っている、所属が異なる場合が多いということである。結果として評価軸2で1点以上が付いたケースが最も多かった。これら2つの要因が正解率に大きく影響したのではないかと考える。

一方で、全体的に SBERT は OpenAI Emb よりも正解率が低かった。

小樽商科大学では最も正解率が低く、それも 20%と極端に低いものとなった。これは OpenAI Emb の方が文章の類似度算出に優れているのではないかと考察をすることが可能である。例えば、片桐氏の「住宅確保における要保障性の所在ーイギリス住宅手当制度を参考にー」という論文では OpenAI Emb の類似論文は「判例研究 売買の目的物である建物の建て替え相当額の賠償請求と居住利益控除の可否 (中略)」となった。しかし SBERT は「部分負荷及び負荷変動を考慮した住宅用コージェネレーションシステムの検討」となった。OpenAI Emb では建築物と法律等に関する題材がテーマとなった論文が出力されていて、対象論文のテーマと少し類似していると判断できた。一方、SBERT は建築物関係にのみ焦点が当たっており、類似していると判断できなかった。これは他の結果でも同様のケースが見られた。そのため、SBERT は重要と判断した単語1つにのみ焦点を当てて類似度を算出しているのではないかと考えられる。そのため先ほど示した例のように、テーマが複数個存在する論文を対象とした場合、どちらかにフォーカスされて十分に類似していない結果が出来るのではないかと考えられる。

北見工業大学では OpenAI Emb と同様の正解率となった。また帯広畜産大学は 50%の正解率であった。この 2 大学でも、先ほど述べた通りの類似論文の算出を行っていると考えられる結果が複数あった。

総合的に本実験の結果は全体で見たときには、OpenAI Emb の方で正解率が高くなった。OpenAI Emb の方では、対象論文に複数個のキーワードがあった場合、それが 2 つ以上含まれる論文を類似論文として出力するケースが多々あった。しかし SBERT は 1 つのキーワードにのみフォーカスしたケースが多く、論文タイトルの全体の意味を分析しきれていない印象を抱いた。また、ほとんどの結果で評価軸 1 の合計点数の方が、評価軸 2 の合計点数よりも高くなるという結果が多くを占めた。ここから考えられることは、「類似論文から意外性のあるマッチングを実現させることは可能であるが、その結果は学校にどのような学部学科があって、どのような研究を行う研究者が、それぞれどこに所属しているかどうかで左右される」ということである。実際、工学という 1 つの大枠で括られている北見工業大

学では、意外性のあるマッチングがあまり実現しなかった。一方、畜産という1つの大枠で括った帯広畜産大学では意外性あるマッチングが多数発見された。本研究では、意外性あるマッチングが実現可能かどうかを考察することが出来ただけではなく、3大学それぞれにどのような学部学科があり、その学部学科それぞれにどのような研究者が所属しているかという点において、その学校の特色を新たに発見出来たと考える。

6章 まとめ

本研究は「大学や研究分野の垣根を超えて、多言語のタイトルに対応しつつ、類似する研究を行う研究者を見つける」ことを目的とした。その結果、大学や研究分野の垣根を超えて、研究分野が異なる研究者同士の組み合わせを発見することができ、日本語だけでなく、英語やドイツ語の論文タイトルも正しく類似度計算をすることを実現した。これにより、本研究の目的は達成できたといえる。

また、新たな知見として、OpenAI Embの方が複数のキーワードに焦点を当てて類似度検索を行う傾向にあることを確認した。そのため、キーワード1つに焦点を当てて分析する傾向のあるSBERTよりも、類似度検索や研究者マッチングに優れている可能性が示唆された。

さらに、本研究では、類似度という数値に加えて、人間の目で判断した類似性を重要視した評価を行うことにより、類似度の高い論文が人間の判断と異なるケースが存在することを明らかにした。

今後の課題としては、人間の目で判断した際に類似性が認められないケースへの対応が必要となる。また、評価軸2で点数が0点になる傾向が高かった、小樽商科大学・商学部・企業法学科や北見工業大学の学部学科全般に所属する研究者が、どのような手法を用いれば意外性あるマッチングが期待できるのかを検討したい。さらにEmbeddingモデルは本研究で用いた手法以外に複数存在するため、新たな手法の活用も進める予定である。

参考文献

- [1] 廣吉和貴 「researchmap の情報を用いた研究者マッチング手法の研究」、小樽商科大学卒業論文、2022 年
- [2] 馬春鵬、松田寛、株式会社リクルート Megagon Labs, Tokyo, Japan 「Ada or Bert : 検索における文埋め込み計算手法の比較研究」、言語処理学会 第 30 回年次大会 発表論文集、2024 年 3 月
- [3] 小林恵大、Qi Yang、成松宏美、南泰浩、電気通信大学、NTT コミュニケーション科学基礎研究所 「推薦理由提示のためのアブストラクトの観点に基づく学術論文推薦」、言語処理学会 第 30 回年次大会 発表論文集、2024 年 3 月

付録 三大学の対象研究者

小樽商科大学

小林敏彦	近藤公彦	竹村壮太郎
西口純代	嘉瀬達男	阿部孝太郎
國武英生	林松国	廣瀬健一
木村泰知	三ツ木真実	杉之原立史
小林友彦	岡部善平	岩本尚禧
西村友幸	山田久就	宮田賢人
深田秀実	石川業	筈井俊輔
佐山公一	片岡駿	小野塚祐紀
中村史	杉山成	岩澤政宗
江頭進	三浦克宜	鈴木和宏
内田純一	長村知幸	石井孝和
小倉一志	藤原健祐	加賀田和弘
田島貴裕	河森計二	高橋優季
加地太一	後藤良彰	佐野博之
多木誠一郎	齋藤健一郎	張博一
齋藤一朗	赤塚広隆	土居直史
小泉大城	醍醐龍馬	山川史
中川喜直	柴山千里	多田伶
沼田ゆかり	西永亮	橋本伸
副島美由紀	市原啓善	石井登
片山昇	權恩熙	上山晋平
石川友和	坂東雄介	羽村貴史
松本朋哉	菅沼真也子	平井進
片桐由喜	中島大輔	須永将史
章天明	王力勇	中村健一
泉貴嗣	水島淳恵	堺昌彦
西出崇	白田康洋	於保淳
猪口純路	沼澤政信	

帯広畜産大学

鈴木宏志
玄学南
横山直明
西川義文
井上昇
押田龍夫
古林与志安
佐々木基樹
河津信一郎
谷昌幸
山内健生
南保泰雄
西田武弘
松井基純
石井利明
平田昌弘
岡村雅史
川村健介
高田兼則
島田謙一郎
加藤清明
白藤梨可
小池正徳
渡邊芳之
韓圭鎬
豊留孝仁
福田健二
宮竹史仁
福本晋也
相内大吾
耕野拓一
橋本直人
五十嵐慎

中馬いづみ
麻田正仁
後藤達彦
手塚雅文
板谷篤司
久保田彰
渡邊謙一
都築直
丸藤祐仁
時岡裕純
保坂和良
室井喜景
茅野光範
柴口順一
近藤大輔
吉田智彦
熊野了州
萩谷功一
得字圭彦
菅沼啓輔
石川透
口田圭吾
伊藤めぐみ
浅利裕伸
斉藤準
宮本明夫
河野洋一
菅原雅之
武田洋平
森田康広
大西一光
川島千帆
吉川琢也

岩本博幸
森正彦
滄木孝弘
松長延吉
仙北谷康
平舘善明
春日純
永田龍次
川合佑典
上川昭博
木下林太郎
峰重隆幸
三上奈々
木下幹朗
福岡直希
三宅俊輔
齊藤朋子
實友玲奈
渡辺純
太田洋輝
秋本正博
上村暁子
中島直久
堀内里紗
落合いづみ
野原香織
森岡昌子
和田大輔
武山暁子
松本高太郎
羽田真悟
木村賢人
窪田さと子

北見工業大学

八久保晶弘
小原伸哉
阿部良夫
星野洋平
川口貴之
小西正朗
駒井克昭
山下聡
川村みどり
森田慎一
柏達也
渡邊康玄
亀田貴雄
井上真澄
大野智也
裡しゃりふ
齋藤徹
前田康成
梶井文人
柴田浩行
木場隆之
館山一孝
金敬鎬
木田真人
杉坂純一郎
陽川憲
久保明彦
富山和也
村田美樹
坂上寛敏
早川吉彦
吉澤真吾
三浦則明
中村大

松田剛
服部和幸
安井崇
曾根宏靖
大津直史
林田和宏
原田康浩
浪越毅
吉田裕
川村武
野田由美意
白川龍生
佐藤和敏
古瀬裕章
早川博
平井慈人
霜鳥慈岳
河野義樹
松田一徳
鈴木正清
春木有亮
奥村貴史
南尚嗣
渡邊達也
兼清泰正
阿曾正浩
中村文彦
平山浩一
邱泰瑛
近藤寛子
松村昌典
小針良仁
本間圭一
鈴木聡一郎

堀彰
蒲谷祐一
岡崎文保
稲葉一輝
佐藤満弘
大野浩
中垣淳
三戸陽一
高橋清
羽二生稔大
黒河賢二
渡邊眞次
本間雄二郎