

令和3年度戦略的基盤技術高度化支援事業
(AIを活用した採択審査に関する調査)
報告書

VISITS

© VISITS Technologies All rights reserved.

事業の目的と実施の基本方針

AI(自然言語処理技術)等を活用して補助金審査の効率化・高度化に向けた実証を行った

目的(仕様書記載内容)

- 戦略的基盤技術高度化支援事業(以下、「サポイン事業」という。)の採択審査にあたっては、1件の審査に複数人の有識者や職員を当てているところ、毎年数百件の中から百件前後の案件を採択している状況を踏まえても、審査に係るコストも無視できないものとなっている。
- こうした中、近年のAI技術の精度向上を踏まえ、採択審査に係るコストを削減できないかの検討を行うこととする。基本的にはAIのみで審査を行うことは難しいことから、審査員を補助する仕組みとなることが想定されるが、どのような仕組みが最適となりうるかを本事業のAIによる審査の精度を踏まえて提案することとする。
- また、対象としては、教師データとして過去の採択データが豊富にあるサポイン事業を対象とするが、コスト削減という観点では横展開できる仕組みであることも重要であり、他の委託事業、補助事業の採択にも流用できる点がないかを含めての検討も合わせて行うこととする。
- なお、現状として、委託事業や補助事業の申請書においては、紙(PDF)から電子データへの転換の狭間があり、教師データとして過去のデータを有効活用するという観点から、可能であれば紙(PDF)からのデータを抽出することも含めて検討を進める必要がある。

実施方針

- AIを活用した補助金審査の実証
 - 申請書データの機会判読可能な電子データ化
 - AIを活用した審査の実証
 - 教師データ(審査済み申請書データと審査結果)から自然言語処理により、点数の差分を生み出した要因を学習
 - 学習に基づき、各申請書のそれぞれの評価軸に対してAIが点数付けして結果を検証
 - 独自技術であるCI技術を活用した教師データの向上
- 他事業への展開も見据えた課題出し・運用イメージの検討・提案
 - AI審査導入に向けた論点の整理と調査から得られた結論の整理
 - 費用対効果試算
- AIが活用しやすいデータを取得するための申請フォームの在り方検討
 - 申請フォーム変更に向けた示唆

実証結果のサマリー

PDFデータの テキストデータ化

- ツールによるPDFデータの機械読み出しでは主要文章以外の部分(図表説明等)が混在するため、不要部分をフィルタリングする処理を行った上で性能を評価
- フィルタリングを適用することにより、主要部分を人手により抽出したデータとの一致度が向上したため、フィルタリングが適切に機能したと確認
 - 今後、各種補助金事業の蓄積された申請書PDFデータへ適用できる可能性

AI審査

- 人手抽出データと審査結果データから、審査結果(点数)を予測する機械学習モデル(審査AI)を構築・評価し、審査AIの性能が人間の審査員の正解率及び相関係数にほぼ匹敵する精度を確認
- 教師データ数をサンプリングして増減させて審査精度を確認したところ、教師データ数が増えると審査精度も向上する傾向が見られ、さらに教師データを増やすと精度向上の余地があると推定される
- 評価貢献度の高い審査員の審査結果を抽出して教師データとすることで審査AIの性能向上を試みたところ、全量を教師データとした審査AIと比べて、現時点のデータ量では性能向上の確認に至らなかったが教師データを増やすことで性能向上が見込まれた
- 以上から、現在の教師データ量でも人間の審査員と同程度の精度で審査AIが機能することは確認できたが、評価貢献度の高い審査員の教師データを拡充することでさらに性能の高い審査AIが構築できる可能性が示唆された

申請書様式

- 今回比較した3事業は審査の構造が似通っているものの、申請書記載項目と評価軸が若干ずれているため、サポインで構築した審査AIを適用するのはリスクを伴う。しかしながら、同様の学習方法で各事業ごとの審査AIを構築できる可能性は高い
- 今後は、各事業において申請書記載項目と評価軸を揃えることで、中企庁共通の事業性審査の審査AIの構築を目指すのがよいのではないかと

- 
1. 実施概要
 2. 結果概要
 3. 考察・今後に向けた示唆
 4. 付属資料(実施内容・結果の詳細)



1. 実施概要

2. 結果概要

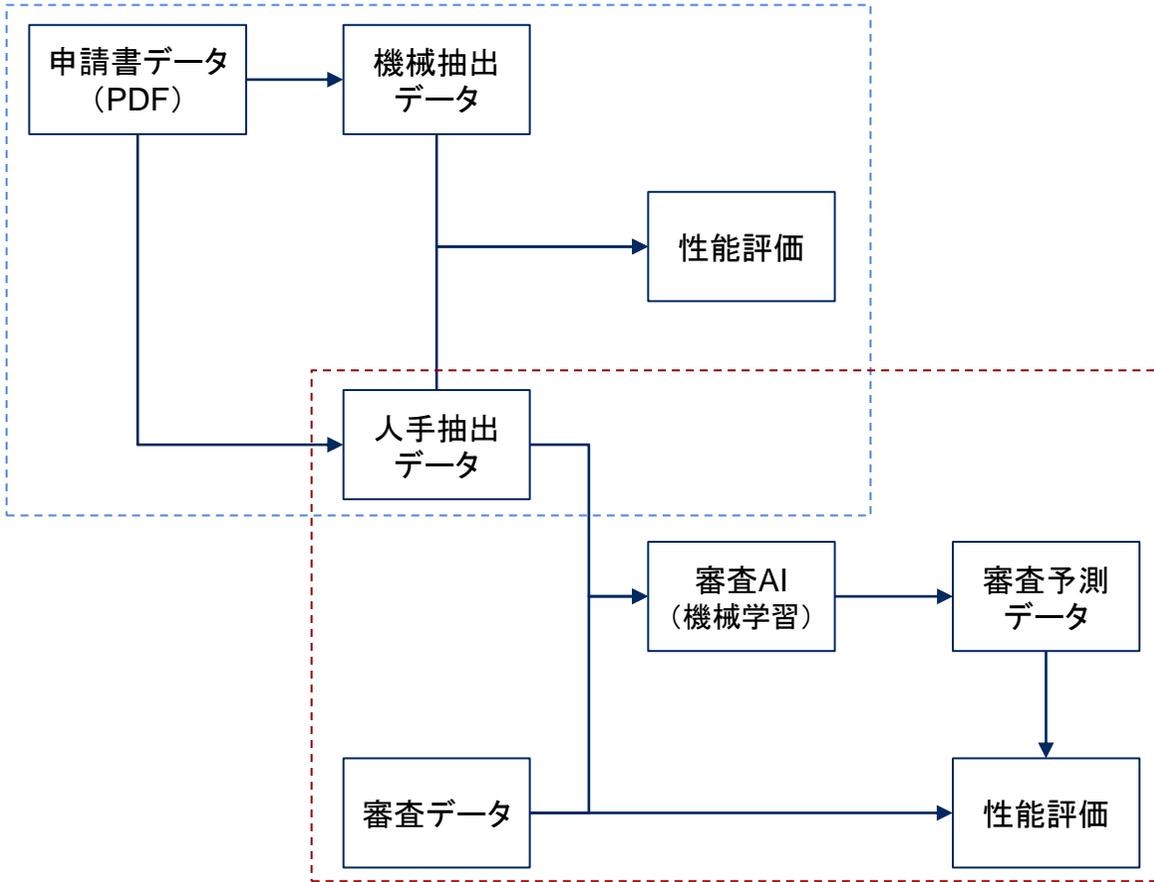
3. 考察・今後に向けた示唆

4. 付属資料(実施内容・結果の詳細)

調査の全体像

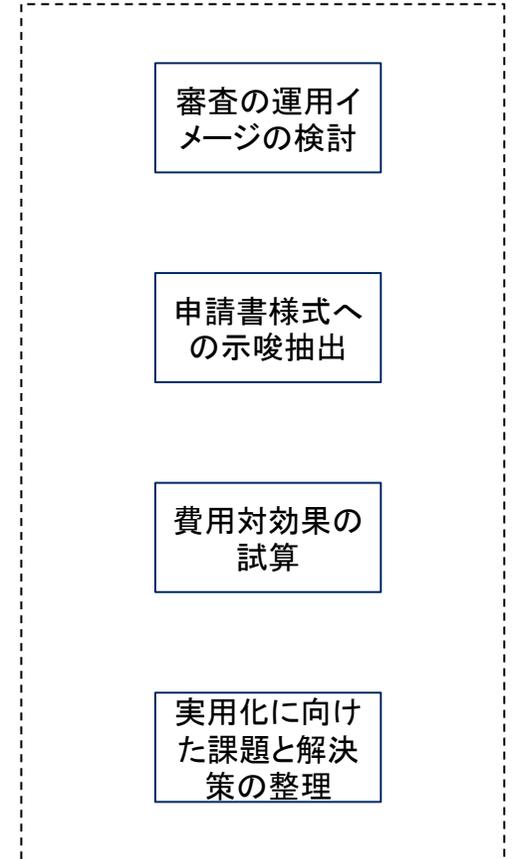
仕様書に記載の業務項目(1)~(4)を以下のように位置付けて調査を実施

(1)過去の申請書データ(PDF、Excel)の機械判読可能な電子データ化



(2)サポイン事業の採択審査へのAI活用の可能性調査 (CIデータを用いた調査を含む)

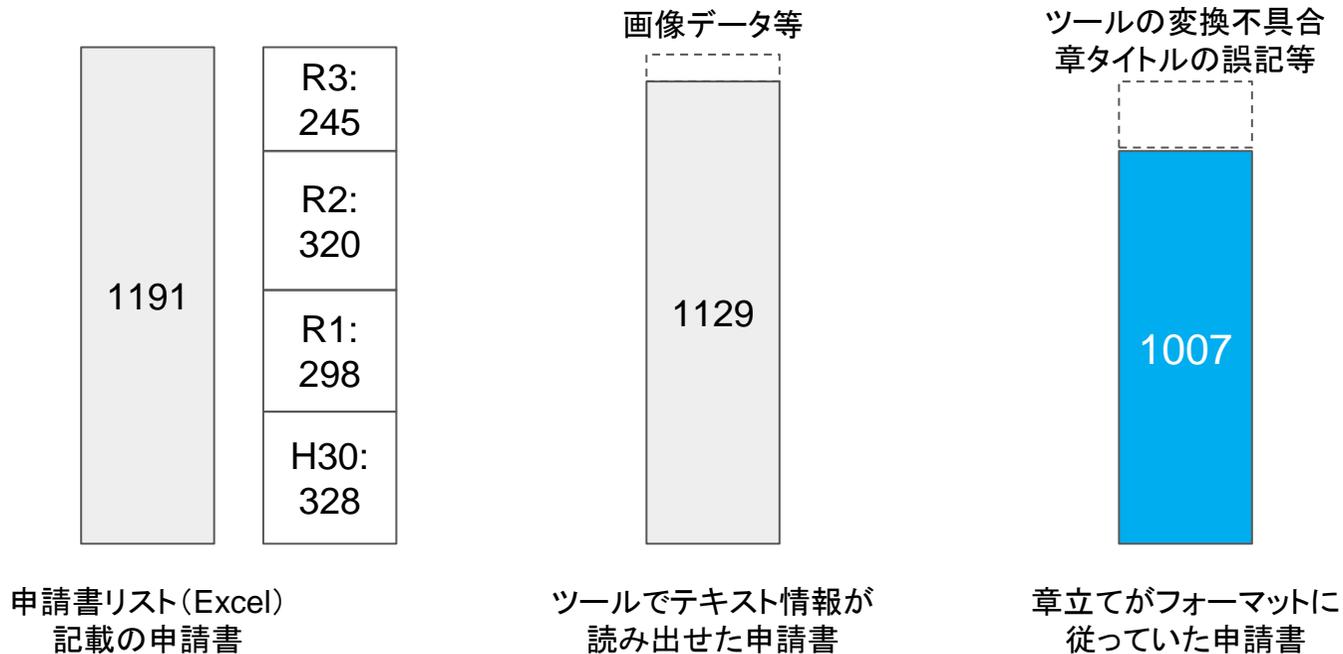
(3)審査制度の分析と実用化に向けた課題の整理 (4)調査報告書の作成等



ご参考) 対象データ

サポイン事業の過去4か年分を対象として調査を行い、1007件をAI処理の対象とした

- PDF一覧データは1191件を受領したが、PDFがツール(pdfTOText)を用いて適切に読み込め、かつ申請書の章立てが一定のフォーマット範囲にあったのは1007件
 - 申請書が画像データ等であった場合、OCRツールを使って文字認識を行う方法があるが、①OCRでは文字認識の精度が下がる点、②対象申請書の割合が小さい点、から本調査の対象外とした。こうした申請者には画像データ(印刷したPDFをスキャンしたもの等)ではなく、文書作成ソフトウェアから直接PDFにエクスポートするように要請することが望ましい



ご参考) 申請書の構成と本調査における分析対象箇所

本調査では自然言語処理AIが適切に機能するテキストデータが分析対象であることから、申請書(様式3)の8項目のうち、以下5項目を対象とした

様式3の項目

AI審査の対象とした提案書の箇所

- | | |
|-----------------------------|-------------------|
| ①研究開発の概要及び背景、当該分野における研究開発動向 | → 対象 |
| ②研究開発の具体的内容 | → 対象 |
| ③研究開発の高度化目標及び技術的目標値 | → 対象 |
| ④研究実施スケジュール | → 線表のため対象外 |
| ⑤研究開発成果及び期待される効果 | → 対象 |
| ⑥事業化計画 | → 対象 |
| ⑦事業化に至るまでのスケジュール | → 線表のため対象外 |
| ⑧専門用語等の解説 | → 採点に直接影響しないため対象外 |



1. 実施概要

2. 結果概要

3. 考察・今後に向けた示唆

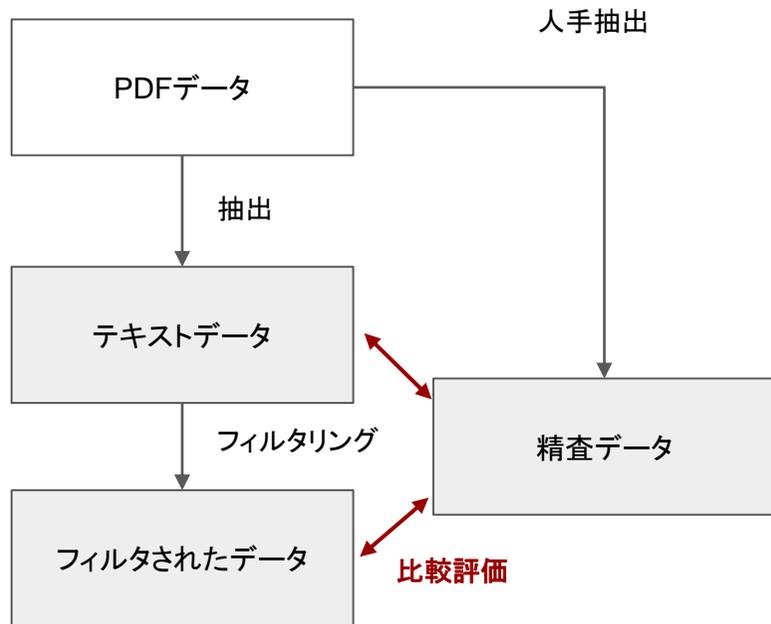
4. 付属資料(実施内容・結果の詳細)

PDFデータのテキストデータ化・フィルタリング

図表説明等を削除するフィルタリングを適用することにより、主要部分を人手により抽出したデータとの一致度が向上したため、フィルタリングが適切に機能したことを確認

性能評価方法

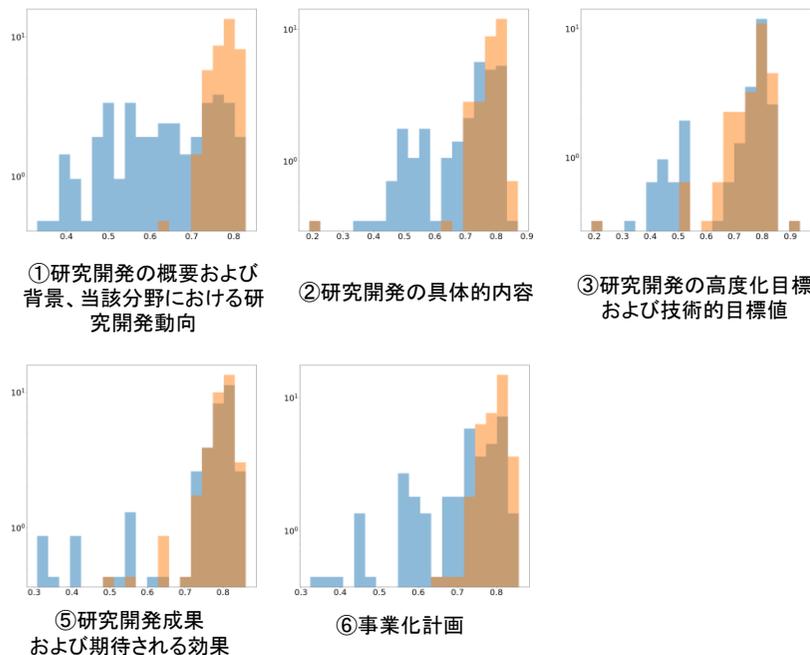
- 機械処理されたデータを、人間が抽出したデータ(精査データ)と比較し、その性能を評価する
- 評価には文字列の一致度比較の指標を用いる
- フィルタリングが適切に機能していれば、評価結果の値が向上することが見込まれる



評価結果 Jaro-Winker距離による一致度

- フィルタリングにより、人手により抽出したデータとの一致度が向上
- フィルタリングが適切に機能
 - 今後、各種補助金事業の蓄積された申請書PDFデータへ適用できる可能性

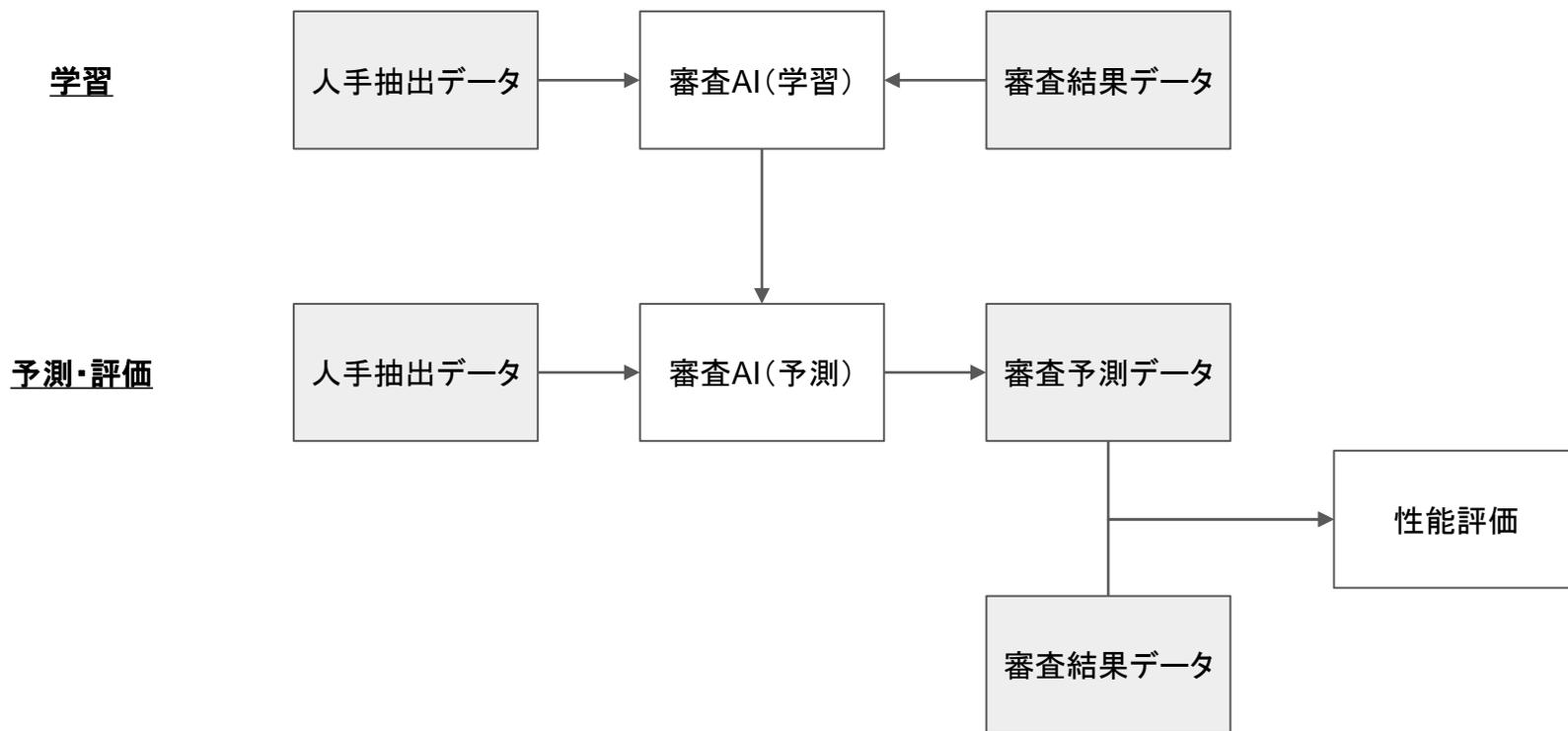
■ テキストデータ - 精査データ
■ フィルタされたデータ - 精査データ



審査AI(採点予測システム)の全容

申請書のテキストデータ(令和2年度までのデータを元に人手により主要部分を抽出したデータ)と審査結果データから審査AIの学習を行い、評価用データ(令和3年度のデータ)を使って審査予測を行い、実際の審査結果と比較することで性能評価を実施

- 人手抽出データと審査結果データから、審査結果(点数)を予測する機械学習モデル(審査AI)を学習
- 予測・評価用のデータを用いて機械学習モデルを性能評価



参考) 性能評価の方法

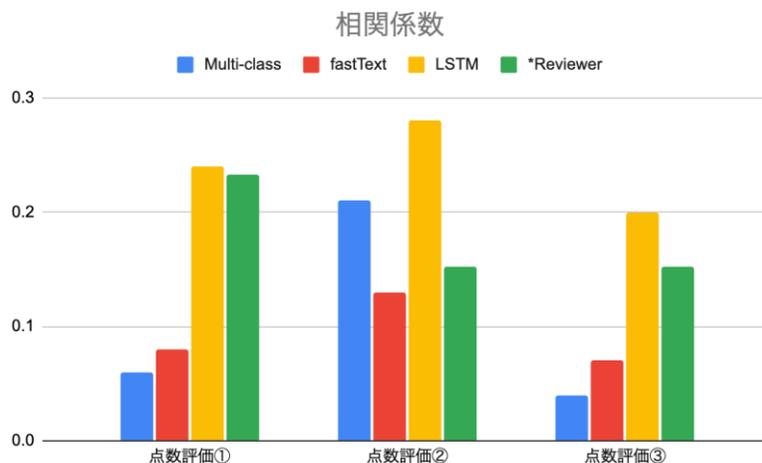
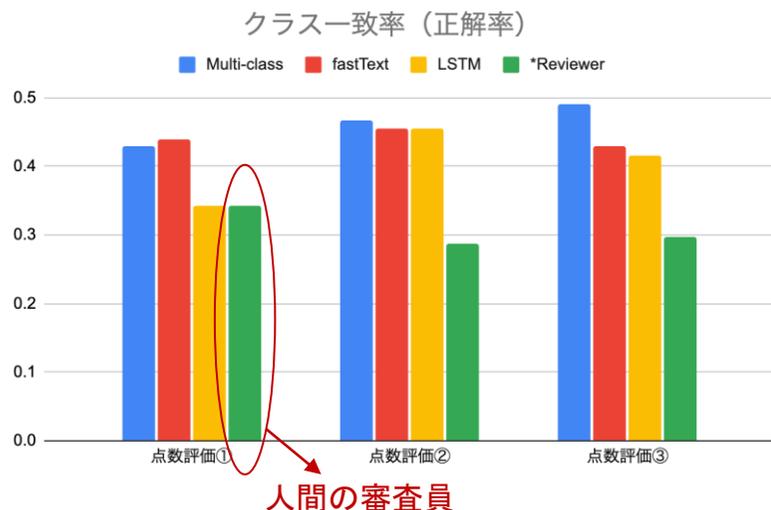
クラス一致率と相関係数の2パターンでの性能評価を実施

性能評価方法	説明	学習と予測・評価のデータ組合せ	
		学習:H30~R1 予測・評価:R2	学習:H30~R2 予測・評価:R3
クラス一致率 (正解率)	<p>機械が予測したクラス(1~5)と正解クラス*の一致率を算出</p> <p>*3審査員の平均値をとり、小数点以下を四捨五入して正解クラスを算出</p>	適用可能	<p>適用不可</p> <p>※R3データとそれ以前のデータのクラスの対応を人手で決めれば適用可能</p>
相関係数	機械が予測した審査点と正解審査点の相関係数を算出(相対評価)	適用可能	適用可能

性能評価の結果(全体像、学習データ=2018,2019、検証データ=2020)

審査AIの性能が人間の審査員の正解率、相関係数にほぼ匹敵する精度を確認

- クラス一致率(正解率)は、Multi-class(SVM)とfastTextを用いた精度が最も高く、相関係数は、LSTM(Long Short-Term Memory)を用いた予測精度が最も高かった
 - これは、前者の2つの手法はクラス(整数スコア)を(A~Eのように)記号とみなして直接予測するため、予測を外した場合に大きく外れる場合があるためである。この点において、後者は連続値のスコアを予測する性質のものであるため、(例えば4.4と4.6の違いのような)微妙な予測誤りがあっても相関係数としては悪くない値をとることによる
- 特にLSTMを用いた場合において、(正解=3名の審査員の総意を完全に当てられないことがあっても)個々の人間の審査員の正解率(*Reviewer)、相関係数にはほぼ匹敵する精度を確認することができた
 - 審査員の精度を超えている場合がある理由は、3名の審査員の集合知を学習しているため



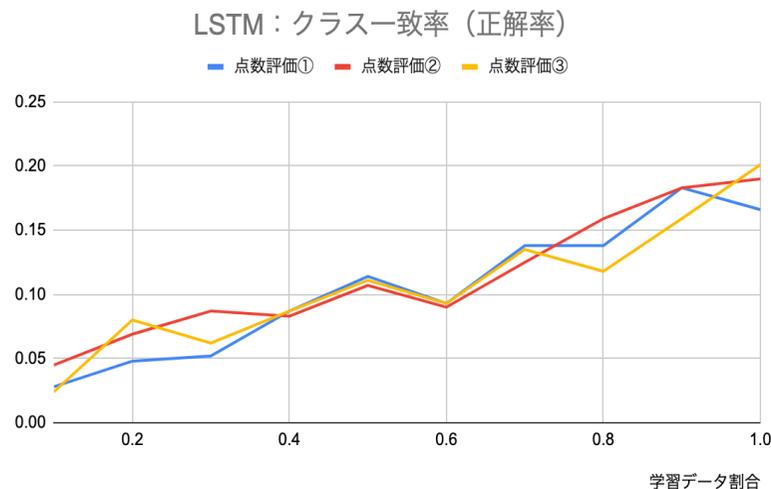
*: Multi-class, fastText, LSTM (Long Short-Term Memory)はそれぞれ機械学習のモデル名
Reviewerは人間の審査員

- ① 目標を達成するための経営的基礎力
- ② 事業化計画の妥当性
- ③ 事業化による経済効果

教師データ数の影響度調査 (学習データ=2018,2019、検証データ=2020)

- 教師データ数をサンプリングして増減させて審査精度を確認したところ、教師データ数が増えると審査精度も向上する傾向が見られ、さらに教師データを増やすと精度向上の余地があると推定される

- LSTMを用い、学習データ(2018, 2019年)を増減させた場合の精度への影響を確認した
- クラス一致率(正解率)、相関係数のどちらの場合も右肩上がりの傾向がみられ、暫くの場合において、学習データを増やすことで更に精度が向上することが見込まれる
- 相関係数は学習サンプルにどのデータが選ばれたか、および初期値の影響で各回ごとに多少の精度ブレがみられる



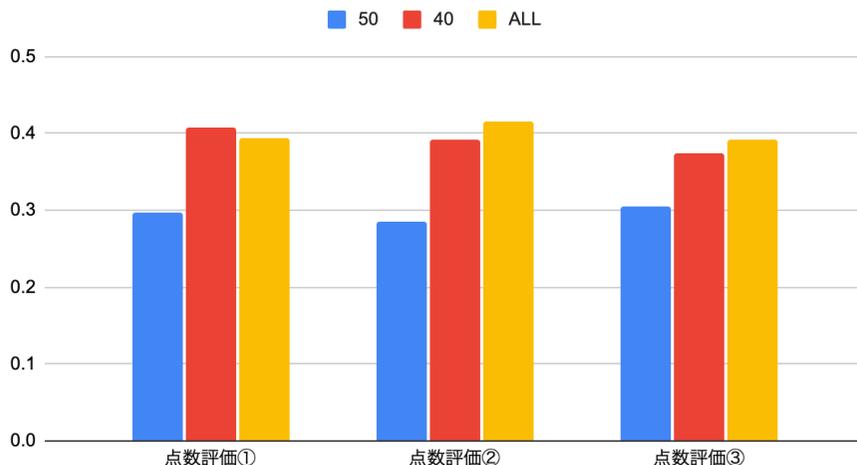
評価貢献度で閾値を設けた結果(学習データ=2018,2019、検証データ=2020)

評価貢献度の高い審査員の審査結果を抽出して教師データとすることで審査AIの性能向上を試みたところ、全量を教師データとした審査AIと比べて、現時点のデータ量では性能向上の確認に至らなかったが教師データを増やすことで性能向上が見込まれる

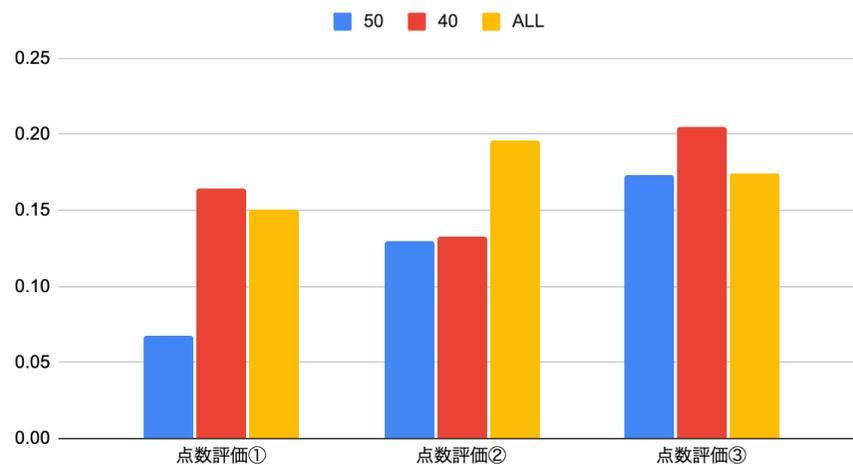
- 偏差値40でフィルタリングした場合:軸によって精度微増
- 偏差値50でフィルタリングした場合:精度低下

- 精度が変化した理由
 - ー 向上要素:フィルタリングによるスコア信頼度(学習データ)の向上
 - ー 低下要素:学習データが減少することによる予測精度の低下
 - 学習データが十分な量に到達すれば(量と精度が比例しなくなる程度)解消されることが予想

クラス一致率 (正解率)



相関係数



各事業の申請書記載項目の比較

各補助金の申請書記載項目は『取り組み内容』『展望』『実施スケジュール』『その他』の4項目に分類され、それらの詳細は共通部分が多い

	サポイン事業	事業再構築補助金	ものづくり・商業・サービス生産性向上促進事業
取り組み内容	<ul style="list-style-type: none">①研究開発の概要及び背景、当該分野における研究開発動向②研究開発の具体的内容③研究開発の高度化目標及び技術的目標値	<ul style="list-style-type: none">①補助事業の具体的取組内容<ul style="list-style-type: none">-事業の状況、強み・弱み、機会・脅威、事業環境、事業再構築の必要性、事業再構築の具体的内容-事業計画-他社との差別化方法-従業員解雇に関する配慮-事業者ごとの取り組み内容・役割	<ul style="list-style-type: none">①補助事業の具体的取組内容<ul style="list-style-type: none">-今までの自社での取組みの経緯・内容-開発内容、目標、達成手段-事業計画-他社との差別化方法
展望	<ul style="list-style-type: none">⑤研究開発成果及び期待される効果⑥事業化計画⑦事業化に至るまでのスケジュール	<ul style="list-style-type: none">②将来の展望(事業化に向けて想定している市場及び期待される効果)<ul style="list-style-type: none">-ユーザー、マーケット、市場規模-優位性・収益性-目標時期、売上規模、価格等	<ul style="list-style-type: none">②将来の展望(事業化に向けて想定している市場及び期待される効果)<ul style="list-style-type: none">-ユーザー、マーケット、市場規模-優位性・収益性-目標時期、売上規模、価格等
実施スケジュール	<ul style="list-style-type: none">④研究実施スケジュール	<ul style="list-style-type: none">④収益計画	<ul style="list-style-type: none">③会社全体の事業計画
その他		<ul style="list-style-type: none">③本事業で取得する主な資産	

申請書への記載が求められる項目

各事業の評価軸(事業化面)の比較

各補助金の事業化面の評価軸は『経営基盤』『計画の妥当性』『補助事業としての妥当性』の3項目に分類され、それらの詳細は共通部分が多い

	サポイン事業	事業再構築補助金	ものづくり・商業・サービス生産性向上促進事業	
評価軸	経営基盤	<p>① 体制・財務状況の適格性</p> <ul style="list-style-type: none">-体制(人材、事務処理能力)や財務状況は十分か-金融機関からの十分な調達が見込めるか	<p>① 体制・財務状況の適格性</p> <ul style="list-style-type: none">-体制(人材、事務処理能力)や財務状況は十分か-金融機関からの十分な調達が見込めるか	
	計画の妥当性	<p>② 事業化計画の妥当性</p> <ul style="list-style-type: none">-製品等が明確か-既存製品への優位性-スケジュールが明確、的確か-売上見込、想定市場が明確、的確か-体制が具体的か-知財ライセンス収入のスケジュールが具体的か	<p>② 市場規模・市場ニーズ</p> <ul style="list-style-type: none">-競合他社の動向、市場ニーズ、市場規模は明確か-ニーズの有無は検証出来ているか <p>③ 優位性・収益性</p> <ul style="list-style-type: none">-价格的・性能的な優位性や収益性はあるか-事業の課題と解決策が明確か	<p>② 市場規模・市場ニーズ</p> <ul style="list-style-type: none">-競合他社の動向、市場ニーズ、市場規模は明確か-ニーズの有無は検証出来ているか <p>③ 優位性・収益性</p> <ul style="list-style-type: none">-价格的・性能的な優位性や収益性はあるか-遂行方法・スケジュールが妥当か
	補助事業としての妥当性	<p>③ 事業化による経済効果</p> <ul style="list-style-type: none">-経済効果を及ぼす産業が具体的か-経済効果は国の補助事業として妥当	<p>④ 補助事業としての費用対効果</p> <ul style="list-style-type: none">-補助事業として費用対効果は高いか-現在の自社の人材、技術・ノウハウ等の強みを活用することや既存事業とのシナジー効果が期待されること等により、効果的な取組か	<p>④ 補助事業としての費用対効果</p> <ul style="list-style-type: none">-補助事業として費用対効果は高いか

※一部の評価軸名称については内容踏まえて弊社にて記載

審査AIを活用するための申請書様式・評価軸への示唆

- 3事業は申請書記載項目と評価軸が若干ずれているため、サポインで構築した審査AIを適用するのはリスクを伴う。しかし、同様の学習方法で各事業ごとの審査AIを構築できる可能性は高い
- 今後は、各事業において申請書記載項目と評価軸を揃えることで、中企庁共通の事業性審査の審査AIの構築を目指すのがよいのではないか

申請書様式・評価軸検討のポイント

- AIは申請書の章立て(申請書記載項目)ごとに学習するため、章立てを明確にわけることが望ましい
- 過去データと今後のデータとで申請書の章立て・評価軸の構造が変わると審査AIが適切に機能するかは分からない
- 各種補助金事業に共通の審査AIを活用するためには、活用対象の補助金事業において、申請書の章立て(申請書記載項目)と評価軸を共通化することが望ましい

参考) 審査AIの他事業への展開

サポインの章立て構造をベースに他2事業の申請書記載項目を整理すると以下のような形に整理可能だが、今後、新たに申請書様式を作成し、新たにデータを学習する場合はこの構造に捕らわれる必要はない

サポイン事業ベースで共通化する例

サポイン事業ベースで共通化する例

申請書への記載が求められる項目

取り組み内容

- **取り組み概要**
 - 当該事業領域における動向及び補助事業による実施概要
- **取り組みの具体的内容**
 - 強み・弱み、機会・脅威
 - 他社との差別化方法
 - 取り組みの詳細
- **取り組みの目標**
 - 研究会開発等の高度化目標及び技術的目標値

展望

- **取り組みの成果及び期待される効果**
 - 競合に対する優位性・収益性
- **事業化計画**
 - ユーザー、マーケット、市場規模
 - 売上規模、価格等
 - 事業化の戦略
- **事業化に至るまでのスケジュール**
 - 目標時期

スケジュール

- **事業実施スケジュール**

その他

事業ごとに設定して個別の学習が必要
(『本事業で取得する主な資産』であれば確認程度であり、審査への影響は軽微か)

経営基盤

目標を達成するための
経営的基礎力

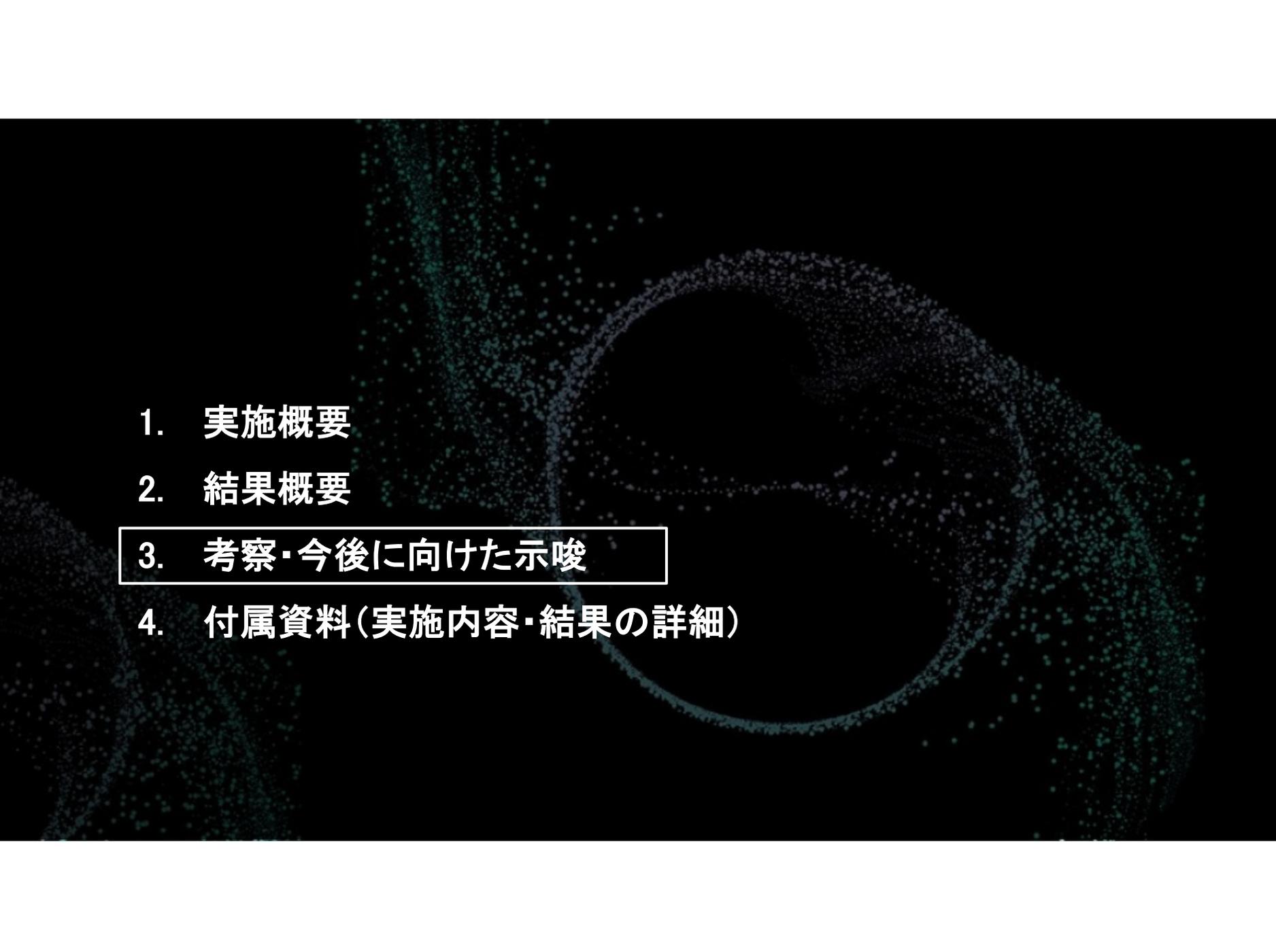
計画の妥当性

事業化計画の妥当性
※本評価軸はサポインは1つのみ、事業再構築補助金は2つに分割されているため、統合可能性については要検討

評価軸

補助事業としての妥当性

事業化による経済効果



1. 実施概要

2. 結果概要

3. 考察・今後に向けた示唆

4. 付属資料(実施内容・結果の詳細)

PDFデータの
テキストデータ化

- ツールによるPDFデータの機械読み出しでは主要文章以外の部分(図表説明等)が混在するため、不要部分をフィルタリングする処理を行った上で性能を評価
- フィルタリングを適用することにより、主要部分を人手により抽出したデータとの一致度が向上したため、フィルタリングが適切に機能したと確認
 - 今後、各種補助金事業の蓄積された申請書PDFデータへ適用できる可能性

AI審査

- 人手抽出データと審査結果データから、審査結果(点数)を予測する機械学習モデル(審査AI)を構築・評価し、審査AIの性能が人間の審査員の正解率及び相関係数にほぼ匹敵する精度を確認
- 教師データ数をサンプリングして増減させて審査精度を確認したところ、教師データ数が増えると審査精度も向上する傾向が見られ、さらに教師データを増やすと精度向上の余地があると推定される
- 評価貢献度の高い審査員の審査結果を抽出して教師データとすることで審査AIの性能向上を試みたところ、全量を教師データとした審査AIと比べて、現時点のデータ量では性能向上の確認に至らなかったが教師データを増やすことで性能向上が見込まれた
- 以上から、現在の教師データ量でも人間の審査員と同程度の精度で審査AIが機能することは確認できたが、評価貢献度の高い審査員の教師データを拡充することでさらに性能の高い審査AIが構築できる可能性が示唆された。

申請書様式

- 今回比較した3事業は審査の構造が似通っているものの、申請書記載項目と評価軸が若干ずれているため、サポインで構築した審査AIを適用するのはリスクを伴うものの、同様の学習方法で各事業ごとの審査AIを構築できる可能性は高い。
- 今後は、各事業において申請書記載項目と評価軸を揃えることで、中企庁共通の事業性審査の審査AIの構築を目指すのがよいのではないかと。

AI審査の運用イメージ(サポイン以外を想定)

審査AIの性能が人間の審査員の正解率、相関係数にほぼ匹敵する精度が確認できたため、3名の審査員のうち、1名を置きかえる運用を検討するのはどうか

これまで

各案件を3名の事業性審査員で審査

AI審査導入後

3名のうち1名をAIに置き換えて、AIと審査員の1票を同等に扱う

本運用の特徴

A案件



B案件



⋮

⋮

⋮

- 後工程で技術点や政策点と合算して採否決定する場合にも整合させやすい
- 既存の審査スケジュールを維持しやすい
- AIが審査員と同様の影響力を持つことへの受容性は要検討ではあるが、実証結果より同等の審査能力を持つ

費用対効果の試算 謝金のコストイメージ

仮にサポイン事業の審査員1名をAIに置き換える場合、100万円強の謝金のコストダウンに留まるが、事業再構築補助金のような大規模補助事業に適用するとインパクトが大きい

これまで

AI審査導入後

試算の前提



サポイン



- 各案件を3名で審査 (AIも1名とみなす)
- 申請案件数: 350件
- 審査単価: 3600円
- 評価者評価については考慮せずに試算

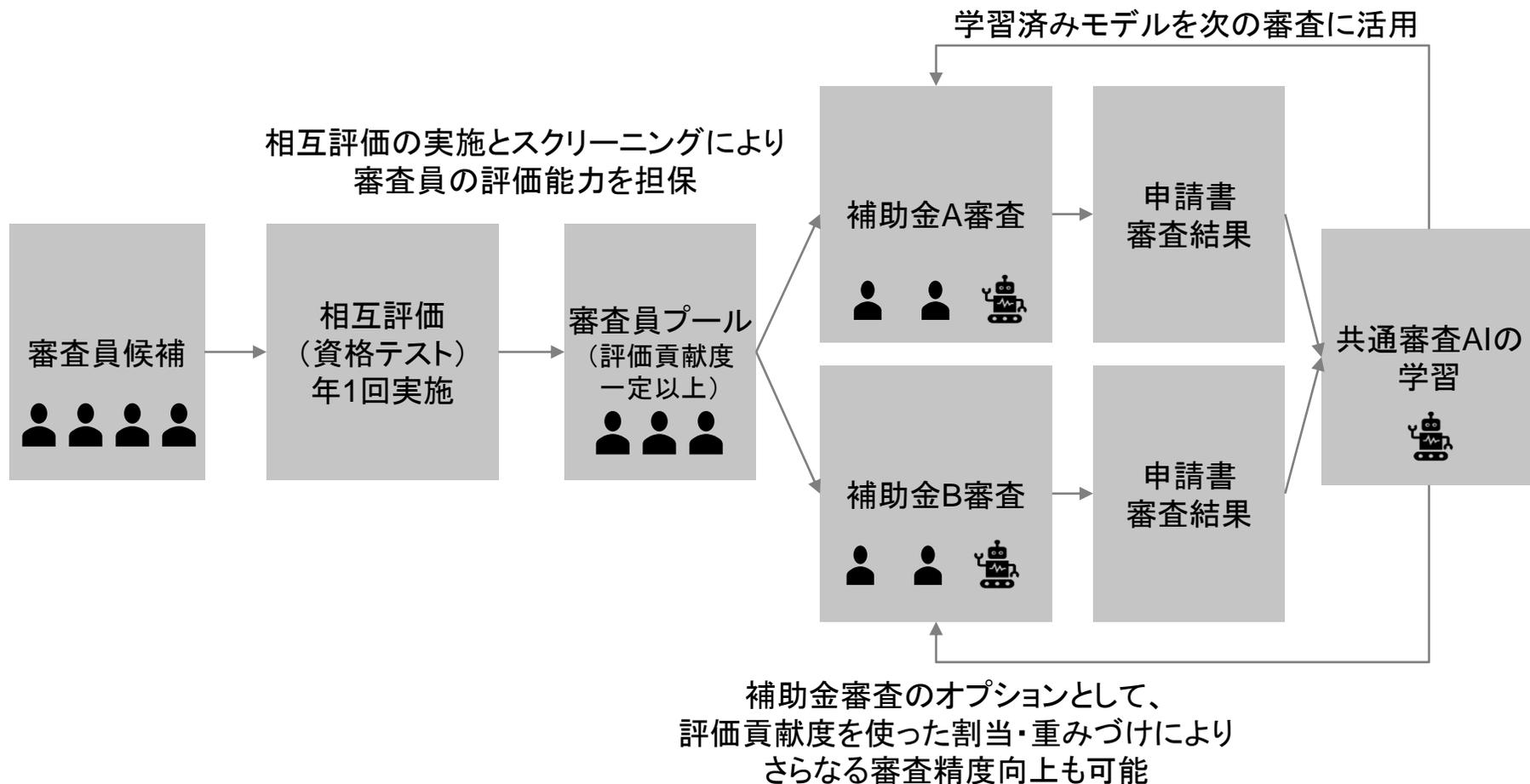
事業再構築補助金規模と想定した場合



- 各案件を3名で審査 (AIも1名とみなす)
- 申請案件数: **2万件**
- 審査単価: 3600円

今後の展望

中企庁の補助金審査の審査員は資格テストとして相互評価でスクリーニングを実施することで質を保つとともに、各補助金共通のAIを審査に参加させる



導入に向けた留意点

AIの精度

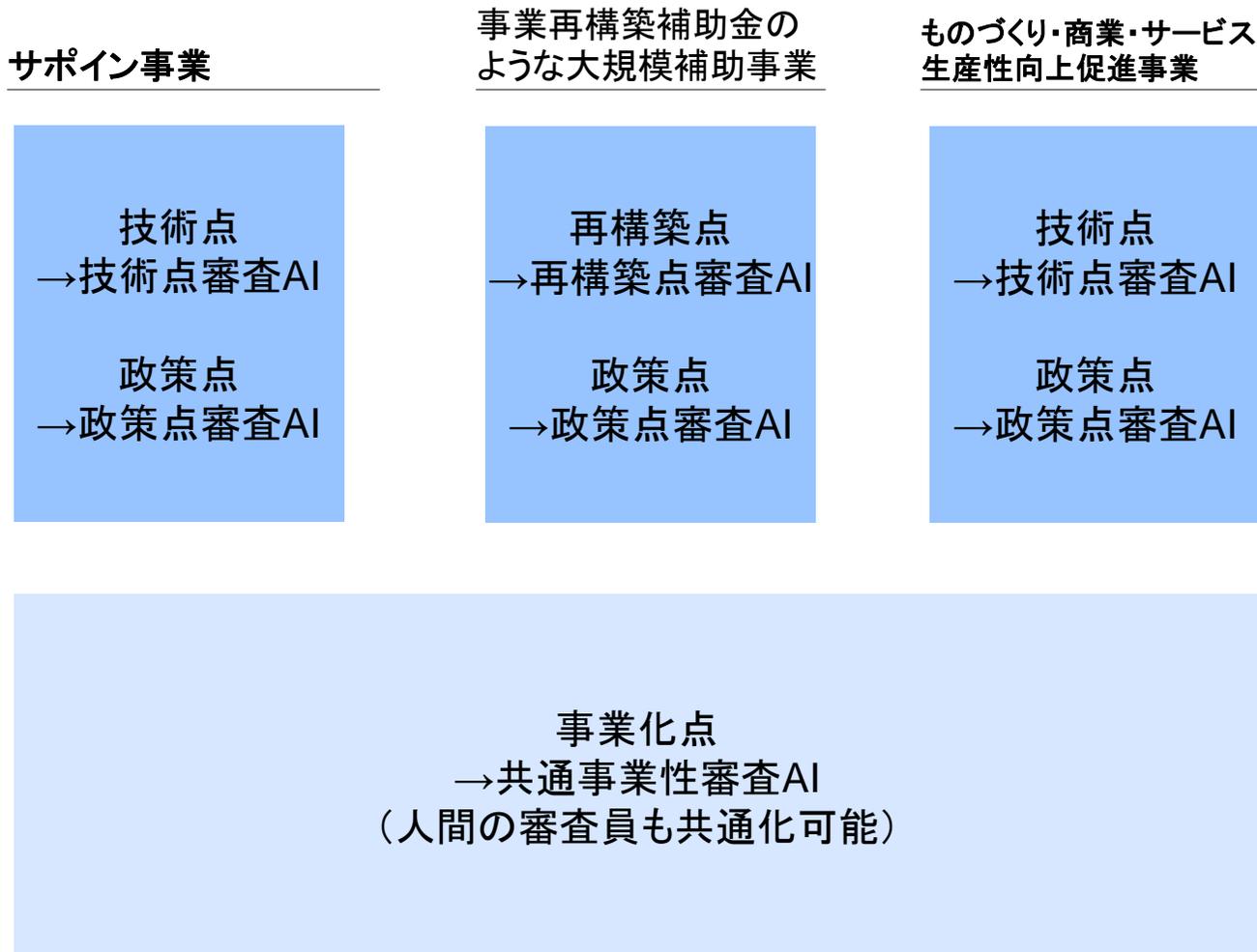
- 現状、AIはテキストデータを学習して審査することから、スケジュール等の図表情報はAIの審査結果には反映されない
 - 現状ではスケジュール等において図表のみ提出している申請書が多く、AIの適用対象外となっている。今後はそのような箇所に関して、重要事項についてはテキストで説明を記載してもらうことでAI適用対象とするのが良いか
- 教師データ数はどの程度あればよいか
 - 今回の実証から600件程度あれば人間の審査員と同等の精度が得られることが分かった。教師データが増加することで更なる精度向上が見込まれる

AIの展開

- 補助金事業ごとに目的が異なり、申請側がアピールすべきポイントや審査側が評価すべきポイントが異なる可能性があるが、申請書項目・評価軸を共通化してもよいか
 - 事業面の審査においては見るべき項目は共通しているため問題はなく、事業特性に応じた部分は事業面の審査以外(技術点、政策点、事業再構築点、等)で考慮するのがよいと思料
- 事業審査以外の審査項目についてAI審査を適用することは可能か
 - 今回の手法をサポイン技術審査データに適用したところ、事業審査に比べてやや精度が低い結果となった。技術審査では、審査員の専門的な背景知識に基づいて申請書のレビューが行われるため、申請書の内容のみからスコアの予測を行うハードルが(事業性審査の場合と比較して)高い可能性がある。審査員プロフィールを入力に含めて学習を行うシステムを考案するなどの改善案が考えられる

審査項目ごとの審査AIイメージ

各種補助事業共通の事業化点については共通の審査AIを活用し、技術点・政策点・再構築点など補助事業ごとに特色のある項目については別途専用の審査AIが必要



補助金審査のDXの全体像

実施概要・効用

評価貢献度を
活用した審査
(サポイン事業に
導入済み)

- 本来、各審査員の評価力等は一人一人異なるにも関わらず、付けられた評価点は同じ重みで扱われているおり、適正な審査になっていない可能性がある
- CI技術*を活用することで、各審査員の評価貢献度を算出し、各評価に対して評価貢献度に応じた補正を行うことで評価精度が向上
 - － 案件評価に携わる審査員間にて、評価コメントに対する相互評価を実施し、独自アルゴリズムにより各評価者の評価貢献度を算出
 - － 評価貢献度は、評価点数の補正と審査員-申請書の割当てにサポインでは利用中
 - － 評価貢献度は、審査の質を担保するためにP24の資格テストとして利用可

AIによる審査員と
申請書のマッチング
(実証済み)

- 審査員と申請者の専門分野の自己申告に基づく審査員-申請書の割振りは、各分野の中でも専門性が多岐にわたるはずだがそれが十分に考慮されておらず、また自己申告そのものが疑わしいことなどもあり、マッチングの精度への悪影響が生じていることが想定される
- この課題に対して、審査員のプロフィール情報と申請書の記載内容に自然言語処理のAIを活用することで、より適切な審査員と申請書の割振りが可能となる

AI審査
(本調査で実証)

- 審査員一人当たりの負担(案件数)が大きくなると審査の質に悪影響があると想定され、審査精度の悪化や審査期間の長期化につながる
- この課題に対して審査にAIを導入することで以下のような価値がある
 - － 審査員の負担を減らすことによる審査精度の向上と審査期間の短縮
 - － 審査員に対する謝金のコストダウン
- 本調査において、現時点での教師データ量でも人間の審査員と同程度の精度で審査AIが機能することは確認できたが、評価貢献度の高い審査員の教師データを拡充することでさらに性能の高い審査AIが構築できる可能性が示唆された

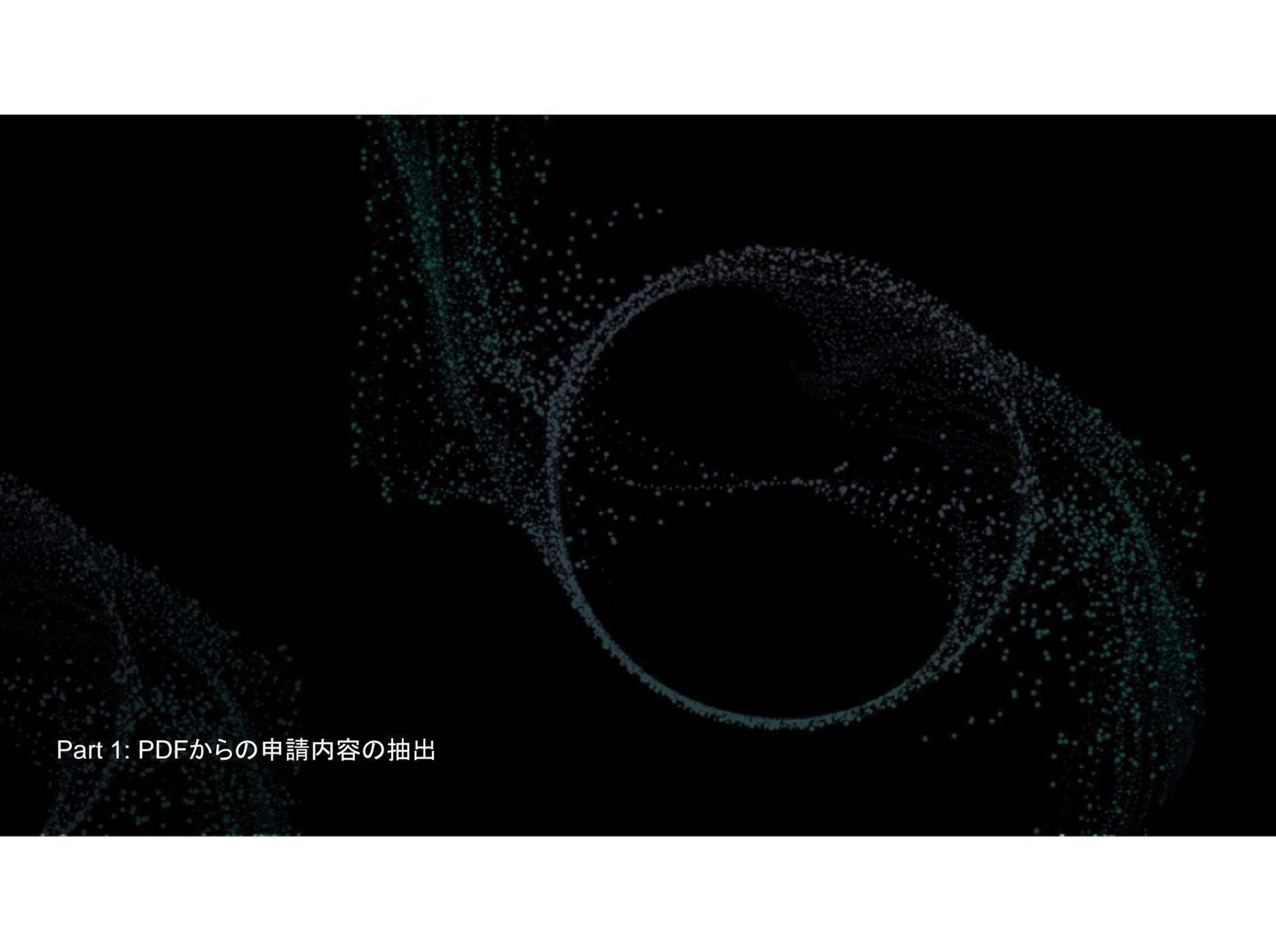


1. 実施概要

2. 結果概要

3. 考察・今後に向けた示唆

4. 付属資料(実施内容・結果の詳細)



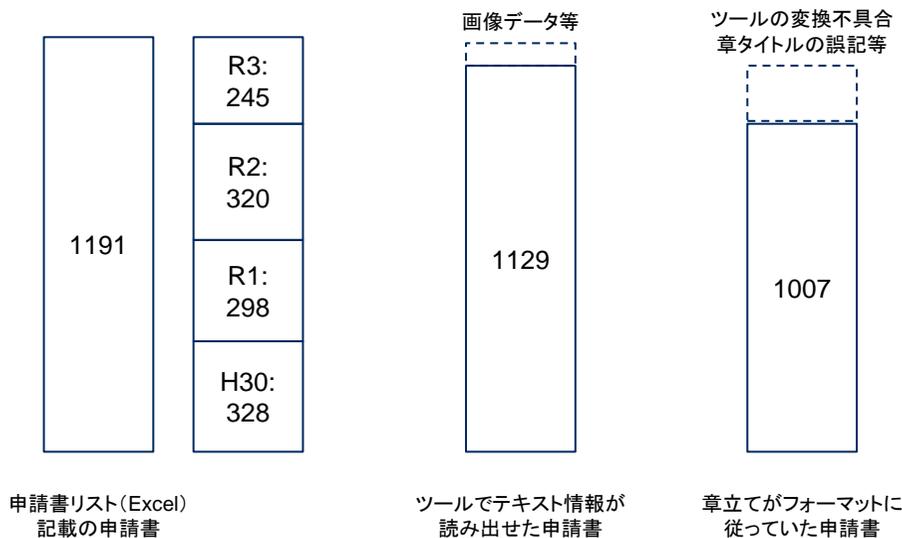
Part 1: PDFからの申請内容の抽出

PDFからの申請内容の抽出(概要)

- 背景
 - PDFデータ: 人間の視認性が高いデータだが、画像で構成されている場合など、機械が読むことが困難な場合がある
 - 機械的に抽出したテキストデータには、主要な文章部分と図表(線表)部分が含まれるため、これを選別したい
- 抽出処理
 - ツールによるテキスト情報の読み出し
 - 申請フォーマットに従っているかの確認
 - 図表のフィルタリング
- 評価
 - 主要な文章部分を人間が抽出したデータとの比較評価

PDFからのテキスト情報の読み出し・申請フォーマットの確認

- PDF一覧データは1191件を受領したが、PDFがツール(pdfotext)を用いて適切に読み込み、かつ申請書の章立てが一定のフォーマット範疇にあったのは1007件だった
 - 申請書が画像データ等であった場合、OCRツールを使って文字認識を行う方法があるが、①OCRでは文字認識の精度が下がる点、②対象申請書の割合が小さい点、から本調査の対象外とした。こうした申請者には画像データ(印刷したPDFをスキャンしたもの)ではなく、文書作成ソフトウェアから直接PDFにエクスポートするように要請することが望ましい
- 本章では、この1007件を対象に申請内容(図表等を除いた文章の集合)をどの程度正確に抽出できるかを調査する



テキスト情報からの図表部分のフィルタリング②

- 主要な文章が多いことが分かっている「①研究開発の概要および背景、当該分野における研究開発動向」と、図表部分が多いことが分かっている「④研究実施スケジュール」のテキスト情報において、それぞれに含まれる行の特徴を統計的に調査した

主要な文章の例

わが国が世界をリードしてきた発酵技術の上に、

$$= 23 / N$$

$$= 23 / 23 = 1.0$$

$$= 0$$

$$= 0 / 23 = 0.0$$

図表部分の例

_____事業年度_____平成_33_年度_____

$$= 19 / N$$

$$= 19 / (19 + 5 + 3) \doteq 0.7$$

$$= 8$$

$$= 7 / 19 \doteq 0.37$$

i. 行両端の空白部分を除いた部分の
文字列長(正規化済み)

ii. 行両端の空白部分を除いた部分の
行全体に対する長さの割合

iii. 行両端の空白部分の長さ

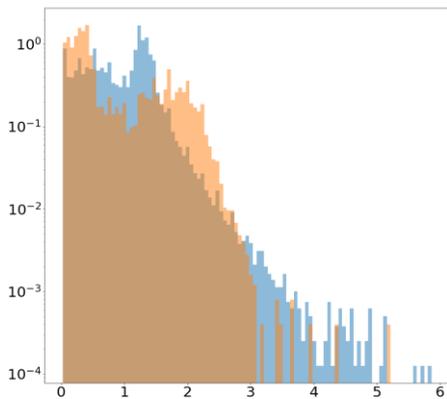
iv. 行両端の空白を除いた部分に
含まれる空白の割合

テキスト情報からの図表部分のフィルタリング③

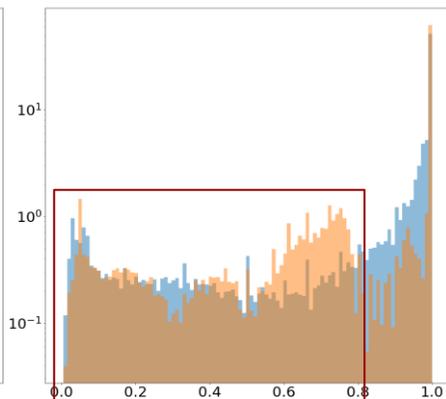
- 主要な文章が多いことが分かっている「①研究開発の概要および背景、当該分野における研究開発動向」と、図表部分が多いことが分かっている「④研究実施スケジュール」のテキスト情報において、それぞれに含まれる行の特徴を統計的に調査した

■ ①研究開発の概要および背景、当該分野における研究開発動向
■ ④研究実施スケジュール

割合(正規化済み)

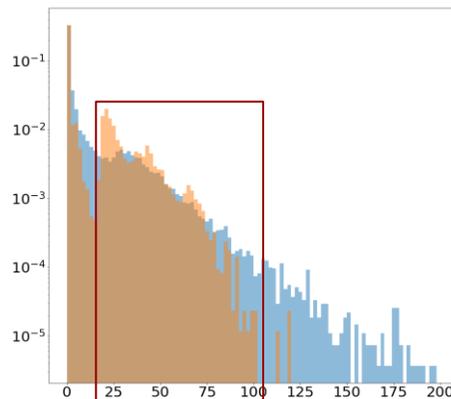


i. 行両端の空白部分を除いた部分の文字列長(正規化済み)



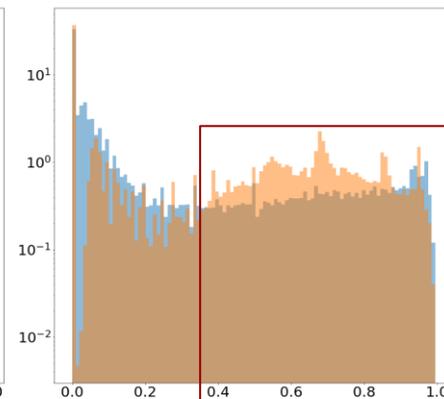
ii. 行両端の空白部分を除いた部分の行全体に対する長さの割合

→全体の長さに対して、文字部分が少ない行は、図表に多く見られる



iii. 行両端の空白部分の長さ

→両端に空白部分が多い行は、図表に多く見られる
 (主要な文章でも極端に空白が多い行が存在しているが、こうした行は無視してよいと考えられる)

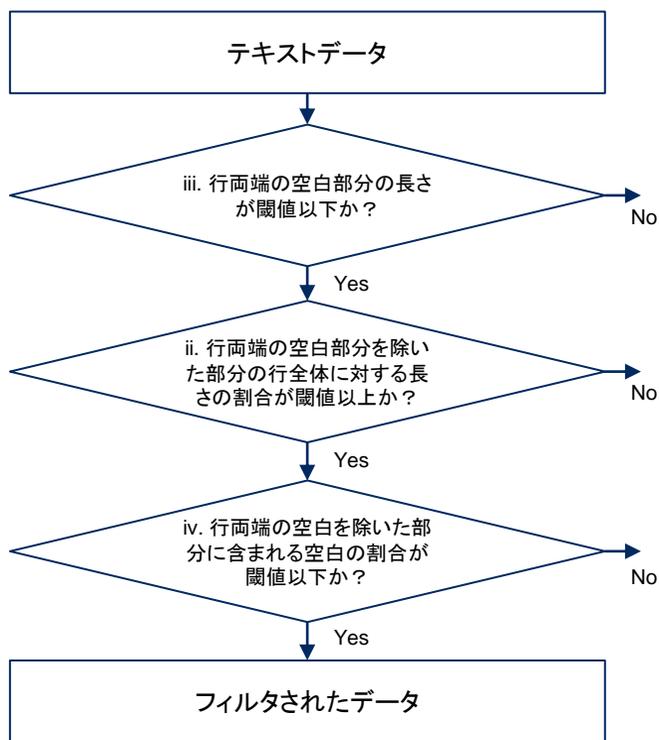


iv. 行両端の空白を除いた部分に含まれる空白の割合

→両端の空白を詰めた上で、更に多くの空白を内包する行は、図表に多く見られる

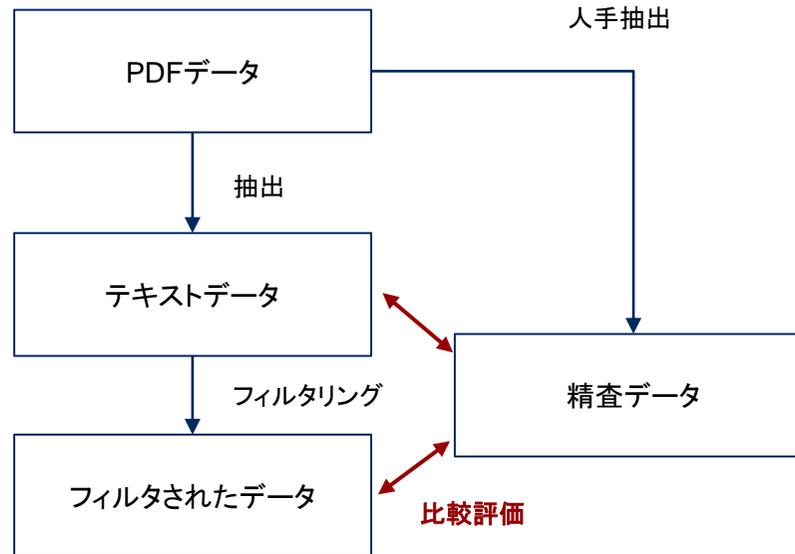
テキスト情報からの図表部分のフィルタリング④

- 前項で得られたグラフをもとに、図表部分をフィルタリングするアルゴリズム(ルールベース)を定義した



主要な文章部分を人間が抽出したデータとの比較評価①

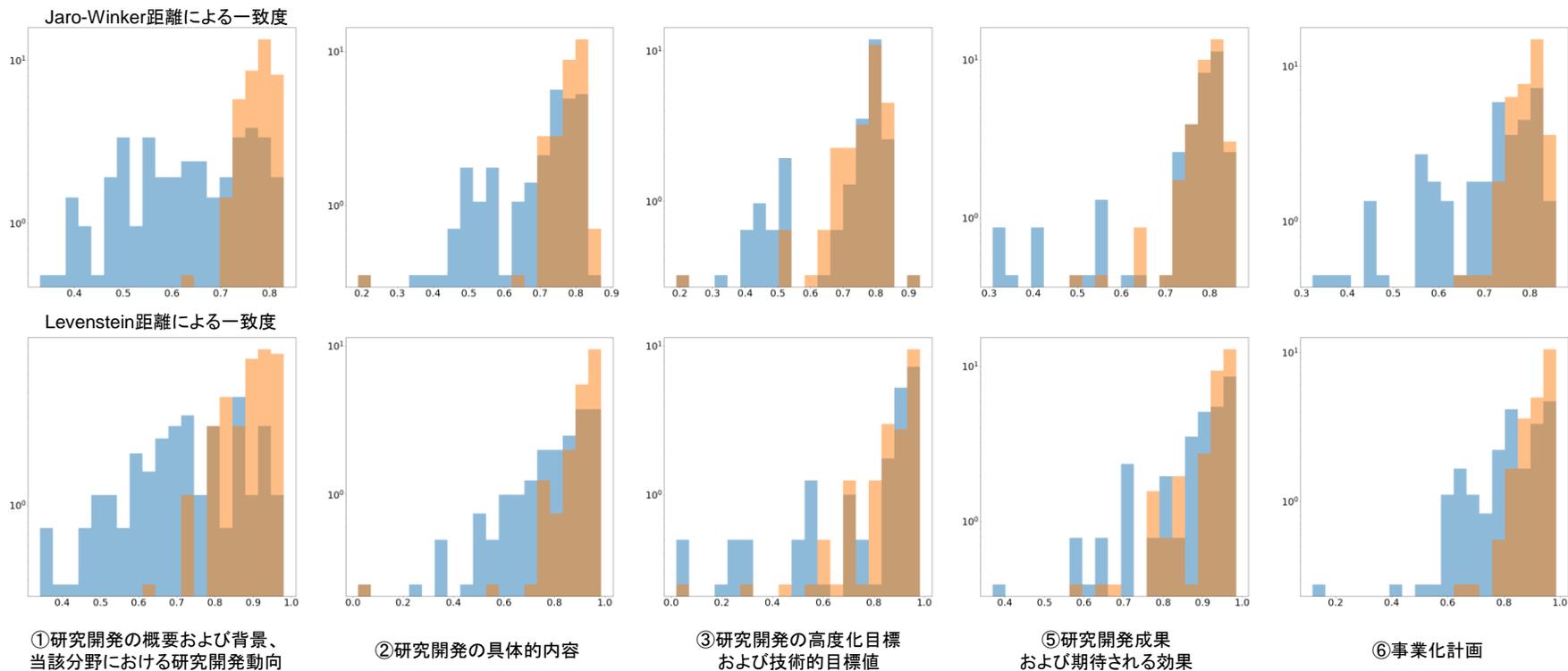
- 機械処理されたデータを、人間が抽出したデータ(精査データ)と比較し、その性能を評価する
- 評価には文字列の一致度比較の指標を用いる
- フィルタリングが適切に機能していれば、評価結果の値が向上することが見込まれる

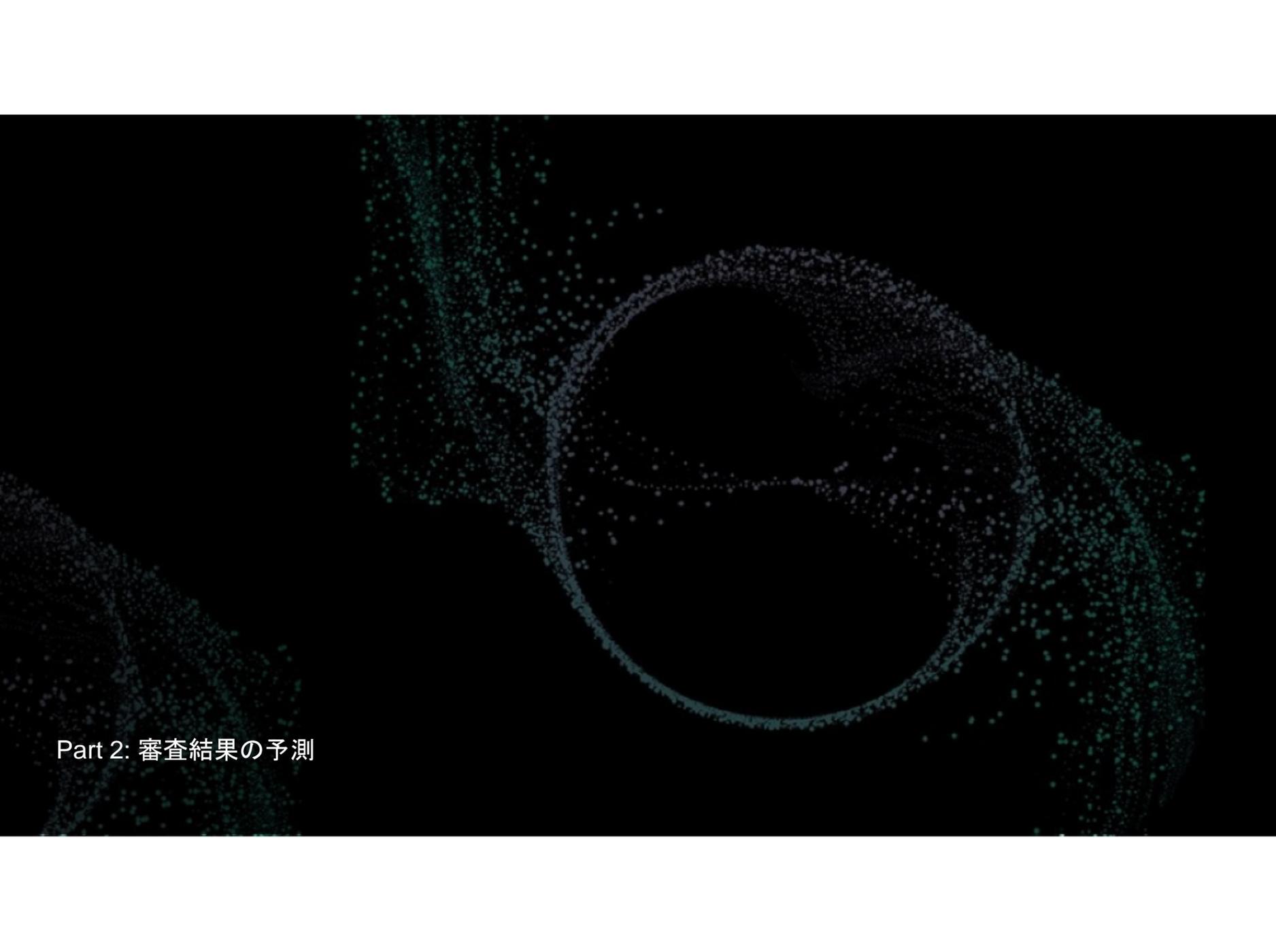


主要な文章部分を人間が抽出したデータとの比較評価②

■ テキストデータ - 精査データ
■ フィルタされたデータ - 精査データ

フィルタをかけることで、人間が抽出したデータとの一致度が全体的に向上している





Part 2: 審査結果の予測

審査結果の予測(概要)

- 背景
 - 審査ではそれぞれの申請書につき、3名(技術審査、事業性審査)の審査員が採点を行っている
 - 本章では主に、事業性審査についての調査結果を報告する
 - 将来的な審査プロセスのアップデートを見据え、AIによる採点予測システムの性能を検証する
- 調査
 - 審査結果データの調査
 - 申請書データとの突合
- 採点予測システム
 - システムの全容と技術概要
- 評価
 - 審査結果データを用いたシステムの性能評価

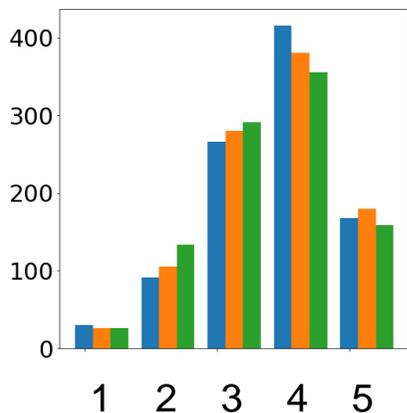
点数結果の分布(事業化評価)

- ①目標を達成するための経営的基礎力
- ②事業化計画の妥当性
- ③事業化による経済効果

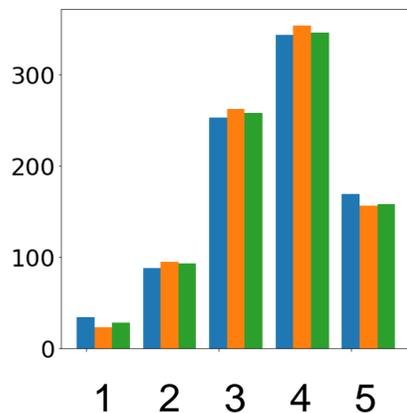
- 各年度ごとに事業化評価(①～③)の点数評価の頻度をカウント
- 2021年度から点数評価が5段階→4段階へと変更されている
- 2018～2020年度については傾向は同様に、スコア4が最頻となるような分布となっている

- 点数評価①
- 点数評価②
- 点数評価③

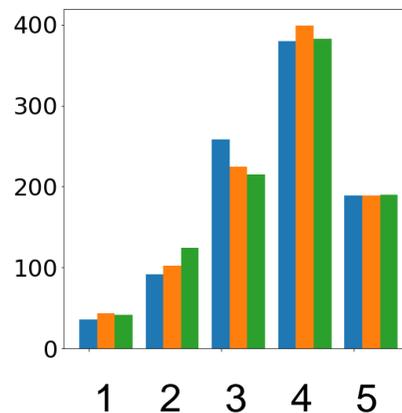
2018年度



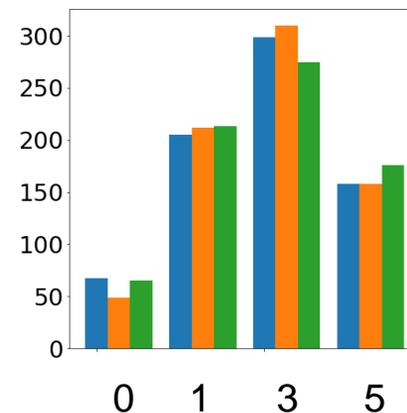
2019年度



2020年度



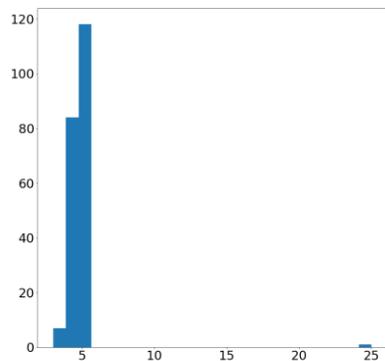
2021年度



審査員の評価案件数の分布

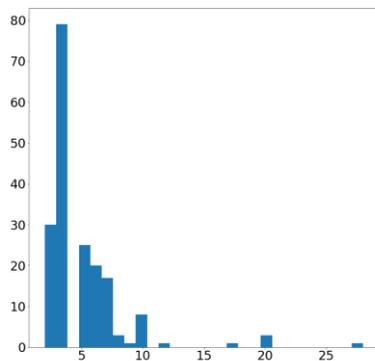
- 年度により1人の審査員が審査した申請書の案件数にばらつきがある
 - 審査員数、割当ロジックの差異

2018年度



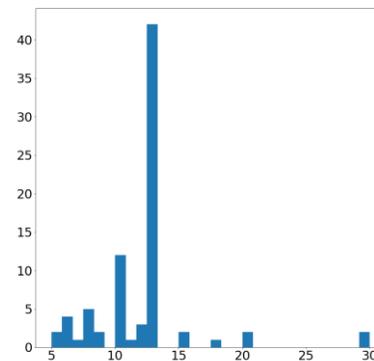
審査員数:210名
申請書数:324案件

2019年度



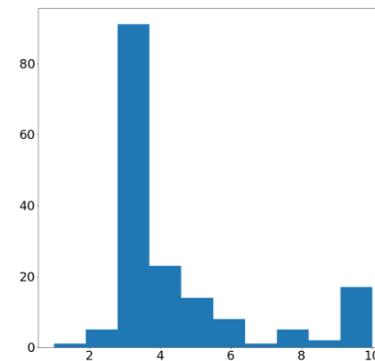
審査員数:189名
申請書数:297案件

2020年度



審査員数:79名
申請書数:320案件

2021年度



審査員数:167名
申請書数:243案件

重複案件を持った審査員ペアの頻度分布

- 審査員の採点一致度(後述)を集計するため、重複案件を持った審査員ペアの頻度を集計
- 例
 - 審査員①: 申請書A、申請書B、申請書C
 - 審査員②: 申請書B、申請書C、申請書D
 - 審査員③: 申請書D、申請書E、申請書F
 - 重複案件数: {①x②: 2, ①x③: 0, ②x③: 1}
- こちらも年度により、審査員数と申請書数の比率、割当ロジックの差異によって傾向の差が生じている
 - 今後も定期的に審査状況を調査する場合、重複案件を避ける割当ロジックを採用すると採点一致度が掴み難くなる可能性がある

	2018年度 (210名)	2019年度 (189名)	2020年度 (79名)	2021年度 (167名)	2018-21年度 ※ (238名)
重複案件数0	21504組	17035組	2315組	13166組	25719組
重複案件数1	117組	636組	606組	663組	1826組
重複案件数2	129組	62組	131組	30組	363組
重複案件数3以上	195組	33組	29組	2組	295組

※ 年度を跨ぐと審査員IDが変わる場合があるため、氏名により集計している

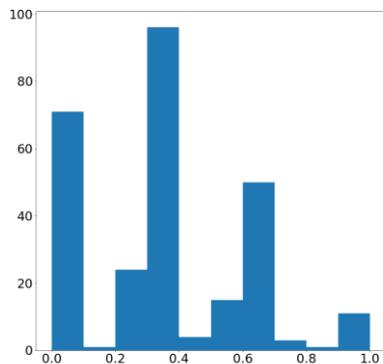
審査員による採点一致度の算出①

- ①目標を達成するための経営的基礎力
- ②事業化計画の妥当性
- ③事業化による経済効果

- 点数評価一致率

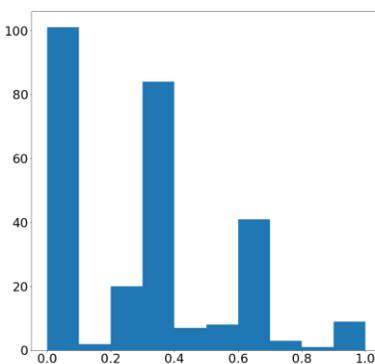
- 審査した申請書に3案件以上の一致があった審査員のペアにおいて、同じ点数評価を与えた申請書の割合
- 例(4案件重複)
 - 評価者①: [申請書A:5点, 申請書B:5点, 申請書C:4点, 申請書D:3点]
 - 評価者②: [申請書A:4点, 申請書B:5点, 申請書C:4点, 申請書D:2点]
- 一致率 = 0.5
- 2018~2020年度の採点を集計したところ、平均的には約3案件に1件が点数一致することが判明

点数評価①



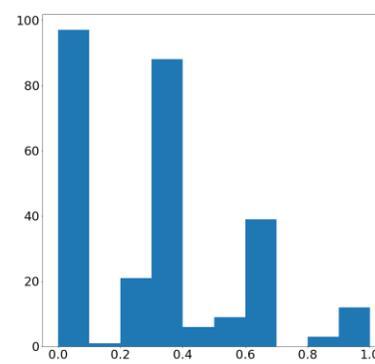
平均値:0.34

点数評価②



平均値:0.29

点数評価③



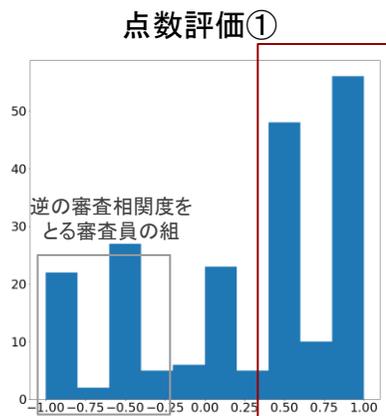
平均値:0.30

審査員による採点一致度の算出②

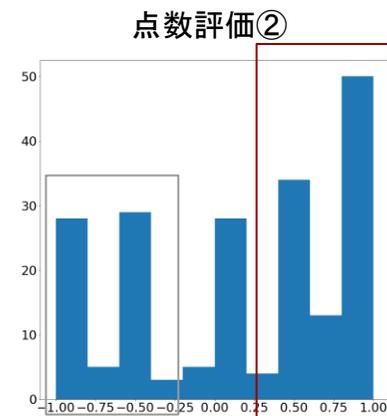
- ①目標を達成するための経営的基礎力
- ②事業化計画の妥当性
- ③事業化による経済効果

- 点数評価相関

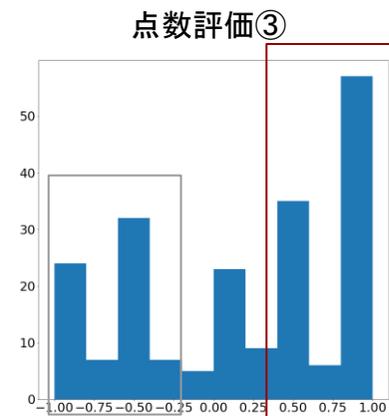
- 一致率→基準点が異なる可能性. 審査した申請書の相対的な一致度を相関で計測(完全一致:1.0, 完全逆順:-1.0)
- 例(4案件重複)
 - 評価者①:[申請書A:5点, 申請書B:5点, 申請書C:4点, 申請書D:3点]
 - 評価者②:[申請書A:4点, 申請書B:5点, 申請書C:4点, 申請書D:2点]
- 相関 = 0.89
- 点数の順序が完全一致した割合が最も高いが、逆となるような場合も少なくない



平均値:0.23



平均値:0.15



平均値:0.15

審査相関度の高い審査員の組

→審査AIは、審査員(人間)の平均的な点数評価一致率、
点数評価相関に到達することを目標とする

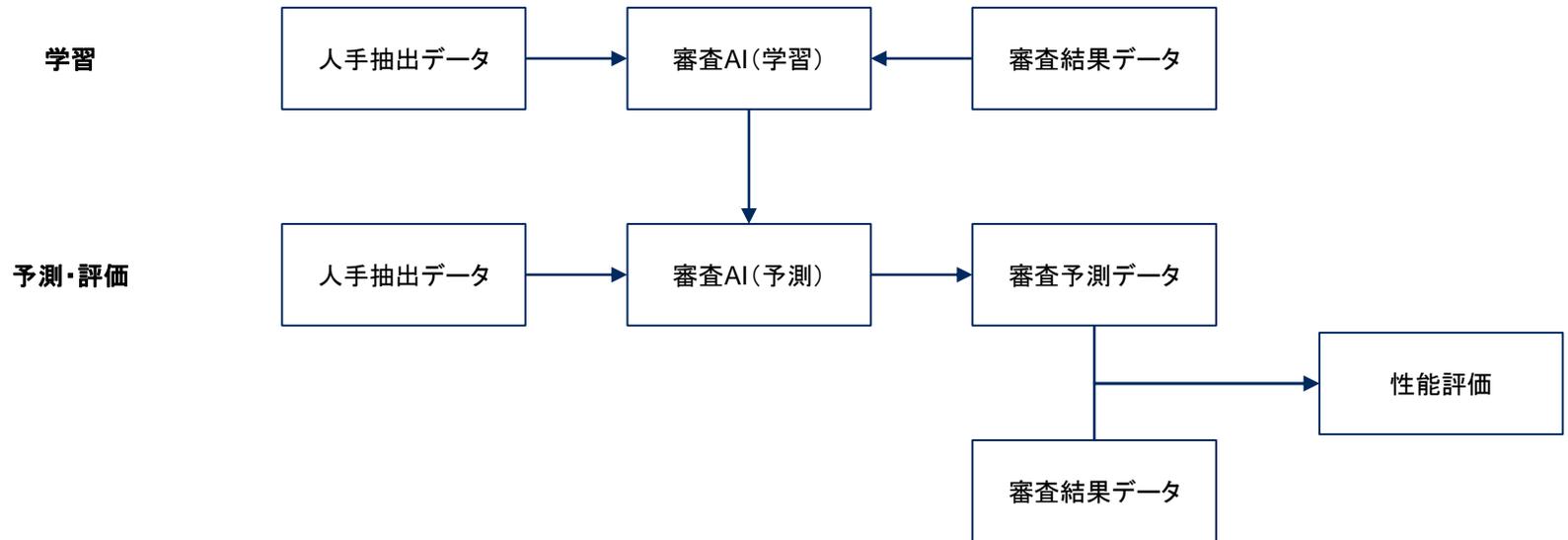
申請書データとの突合

- 背景
 - 審査ではそれぞれの申請書につき、3名(技術審査、事業性審査)の審査員が採点を行っている
 - 将来的な審査プロセスのアップデートを見据え、AIによる採点予測システムの性能を検証する
- 調査
 - 審査結果データの調査
 - 申請書データとの突合
- 採点予測システム
 - システムの全容と技術概要
- 評価
 - 審査結果データを用いたシステムの性能評価

採点予測システム

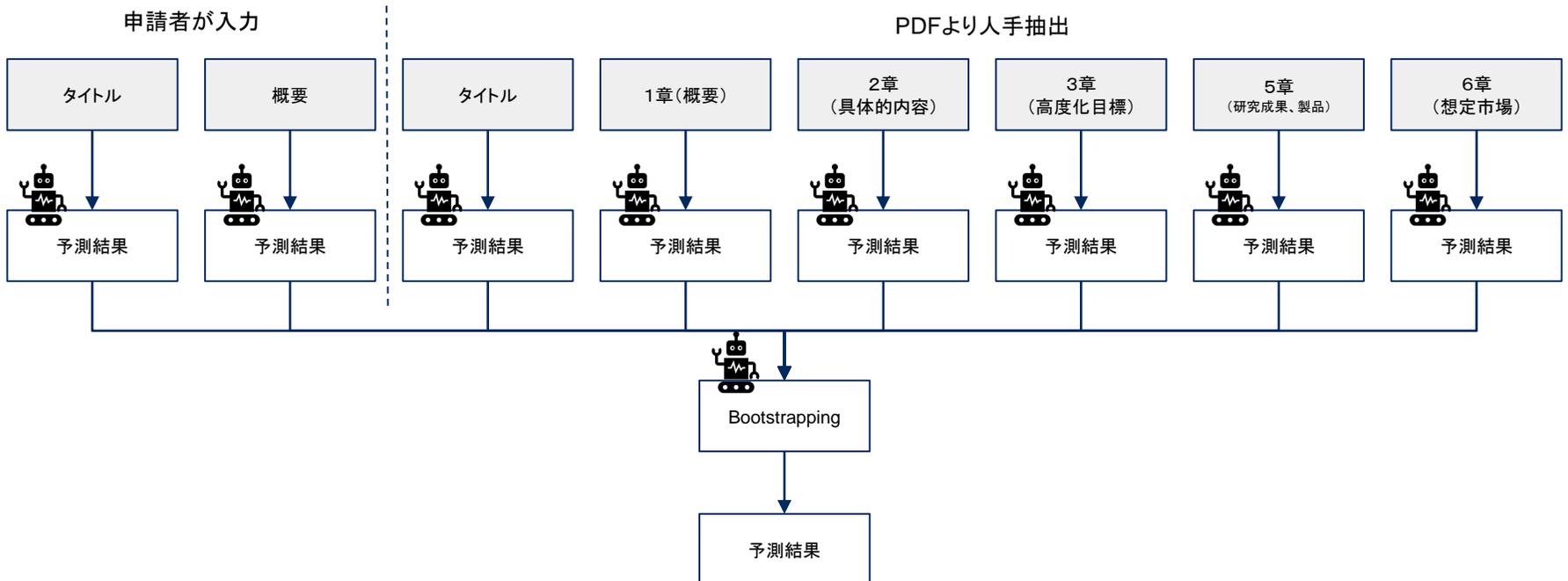
- システムの全容

- 人手抽出データと審査結果データの組から、審査結果(点数)を予測する機械学習モデル(審査AI)を学習
- 予測・評価用のデータを用いて機械学習モデルを性能評価



採点予測システム (Bootstrapping)

- 審査AIシステムは、個々の機械学習機と、Bootstrapping部から構成される
 - 全要素を予測のインプットとする一方で、ノイズに強い構成
- Bootstrapping部では、個々の機械学習機が予測したスコアを重み付けすることで、正解データに近くなるように学習(回帰予測)を行う



機械学習モデルと学習方法の一覧

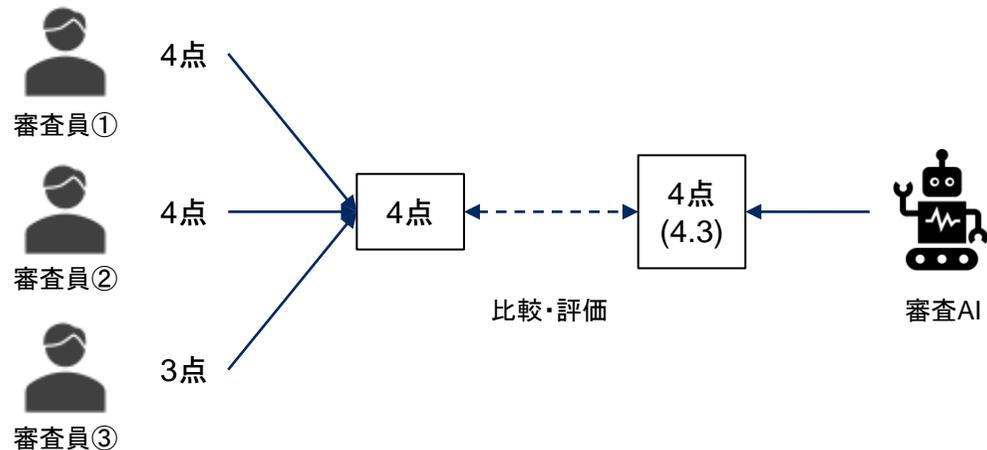
モデル名	学習タイプ	予測結果の例	備考
Multi-class	多クラス分類	1, 3, 2, ...	SVM(Support Vector Machine)
fastText	多クラス分類	1, 3, 2, ...	深層学習
LSTM(Regression)	回帰予測	1.4, 3.5, 2.3, ...	深層学習

性能評価の方法

性能評価方法	学習:H30~R1 予測・評価:R2	学習:H30~R2 予測・評価:R3	備考
クラス一致率(正解率)	適用可能	適用不可 ※R3データとそれ以前のデータのクラスの対応を人手で決めれば適用可能	機械が予測したクラス(1~5)と正解クラスの一致率を算出
相関係数	適用可能	適用可能	機械が予測した審査点と正解審査点の相関係数を算出(相対評価)

前処理・後処理等

- 審査点数
 - 3審査員の平均値をとり、小数点以下を四捨五入して正解クラスを算出
 - 回帰予測で評価をクラス一致率で測る場合は、予測スコアの小数点以下を四捨五入して予測クラスとする
- 入力テキスト
 - 形態素解析 (MeCab) を用いて形態素に分割
 - 名詞句を抽出して学習に利用 (不要な記号、助詞等をフィルタするため)

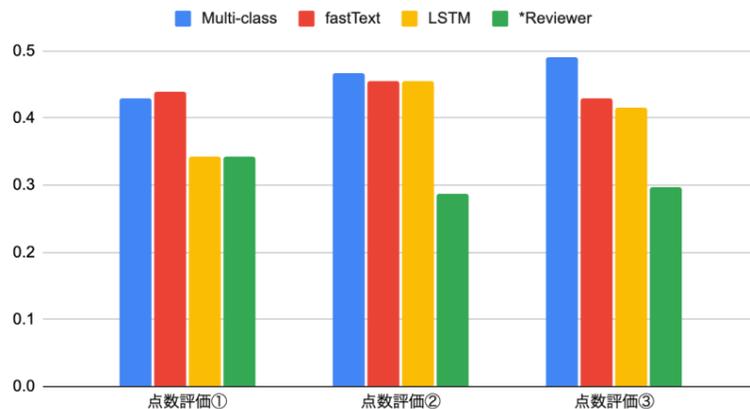


- ①目標を達成するための経営的基礎力
- ②事業化計画の妥当性
- ③事業化による経済効果

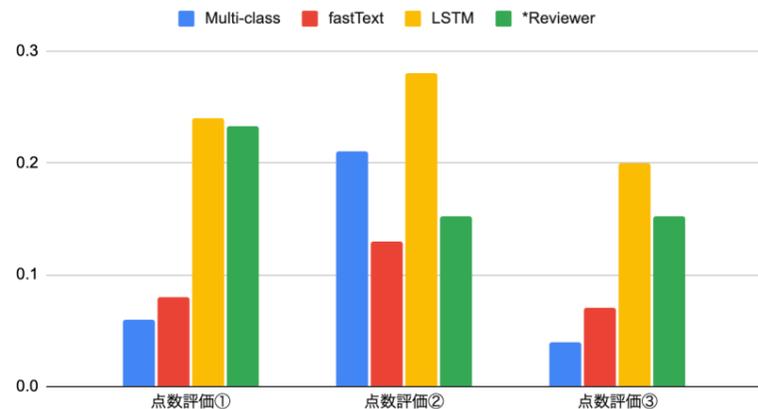
評価結果(全体像、学習データ=2018,2019、検証データ=2020)

- クラス一致率(正解率)は、Multi-class(SVM)とfastTextを用いた精度が最も高かった
- 一方で相関係数は、LSTMを用いた予測精度が最も高かった
- これは、前者の2つの手法はクラス(整数スコア)を(A~Eのように)記号とみなして直接予測するため、予測を外した場合に大きく外れる場合があるためである。この点において、後者は連続値のスコアを予測する性質のものであるため、(例えば4.4と4.6の違いのような)微妙な予測誤りがあっても相関係数としては悪くない値をとることによる
- 特にLSTMを用いた場合において、(正解=3名の審査員の総意を完全に当てられないことがあっても)個々の人間の審査員の正解率(*Reviewer)、相関係数にはほぼ匹敵する精度を確認することができた
 - 審査員の精度を超えている場合がある理由は、3名の審査員の集合知を学習しているため

クラス一致率 (正解率)



相関係数

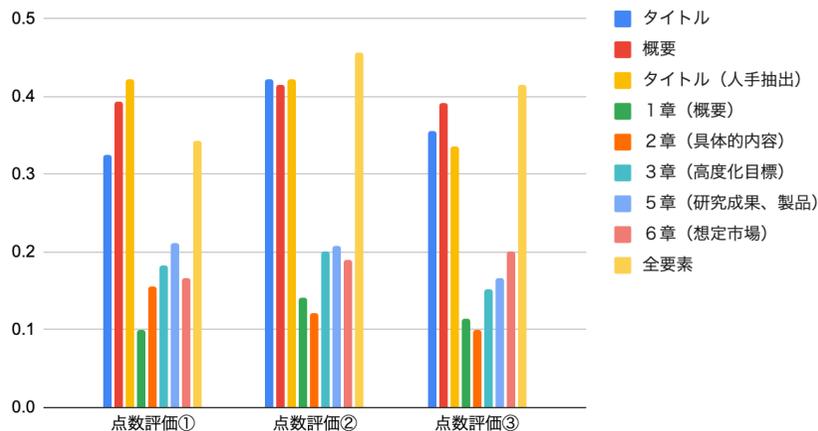


入力で用いるセッションの寄与度調査(学習=2018,2019、検証=2020)

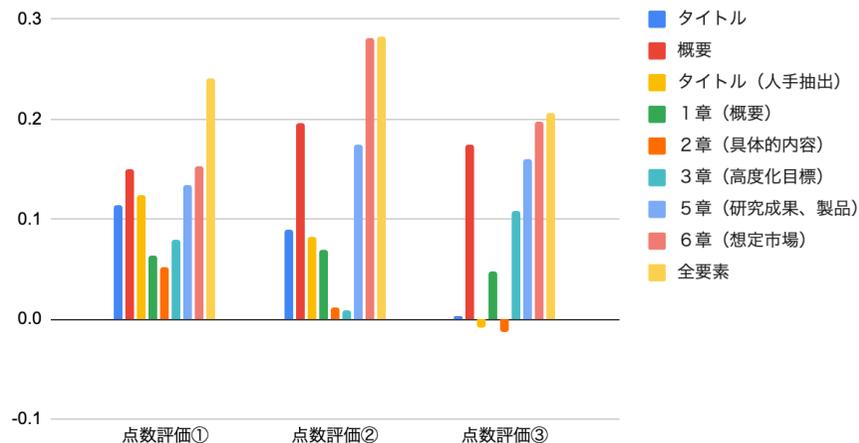
- ①目標を達成するための経営的基礎力
- ②事業化計画の妥当性
- ③事業化による経済効果

- LSTMを用いた場合において、Bootstrappingを用いず、個々の入力セクションのみを用いた場合におけるそれぞれの精度を計算した
- クラス一致率(正解率)に寄与していたのは「タイトル」「概要」であった一方、相関係数に寄与していたのは「6章(想定市場)」および「概要」であった

LSTM：クラス一致率（正解率）



LSTM：相関係数

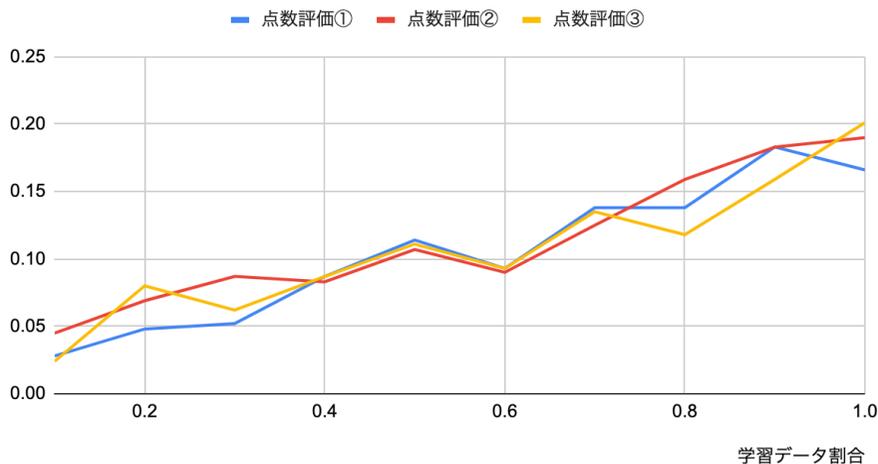


学習データ数の影響度調査(学習データ=2018,2019、検証データ=2020)

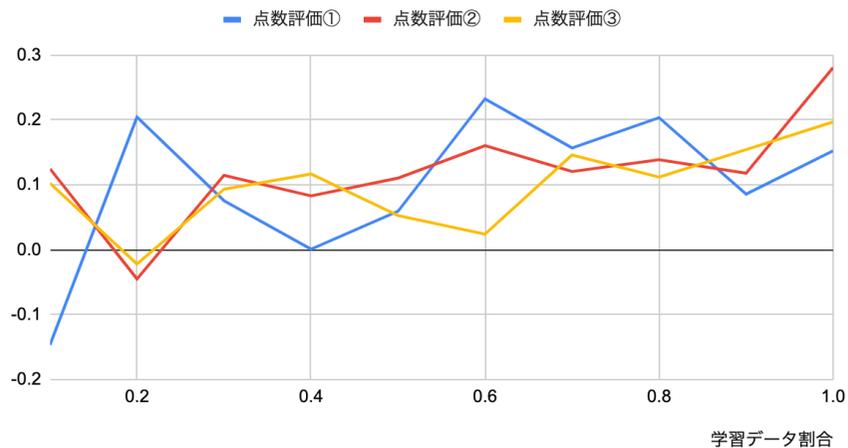
- ①目標を達成するための経営的基礎力
- ②事業化計画の妥当性
- ③事業化による経済効果

- LSTMを用い、入力に「6章(想定市場)」を用いた場合において、学習データ(2018, 2019年)を増減させた場合の精度への影響を確認した
- クラス一致率(正解率)、相関係数のどちらの場合も右肩上がりの傾向がみられ、暫くの場合において、学習データを増やすことで更に精度が向上することが見込まれる
 - 相関係数は学習サンプルにどのデータが選ばれたか、および初期値の影響で毎回ごとに多少の精度ブレがみられる

LSTM：クラス一致率（正解率）



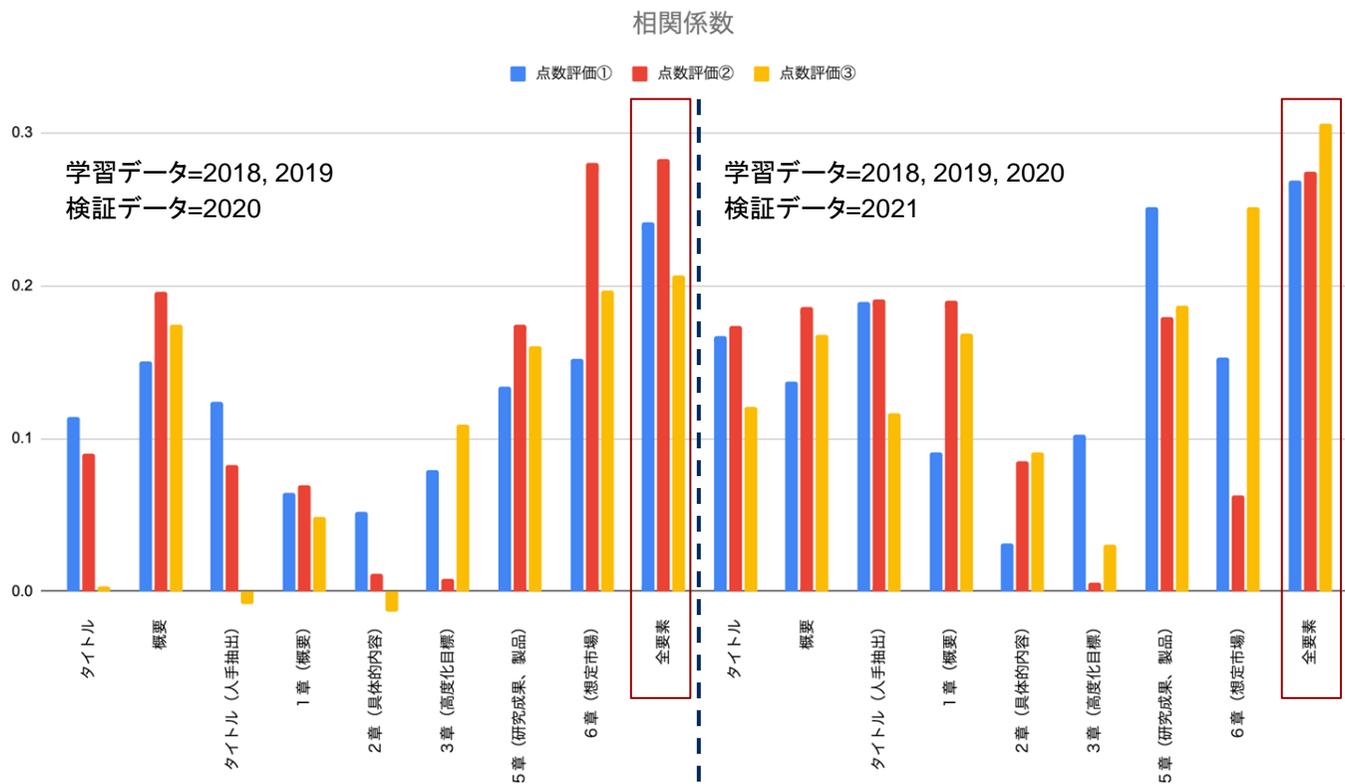
LSTM：相関係数

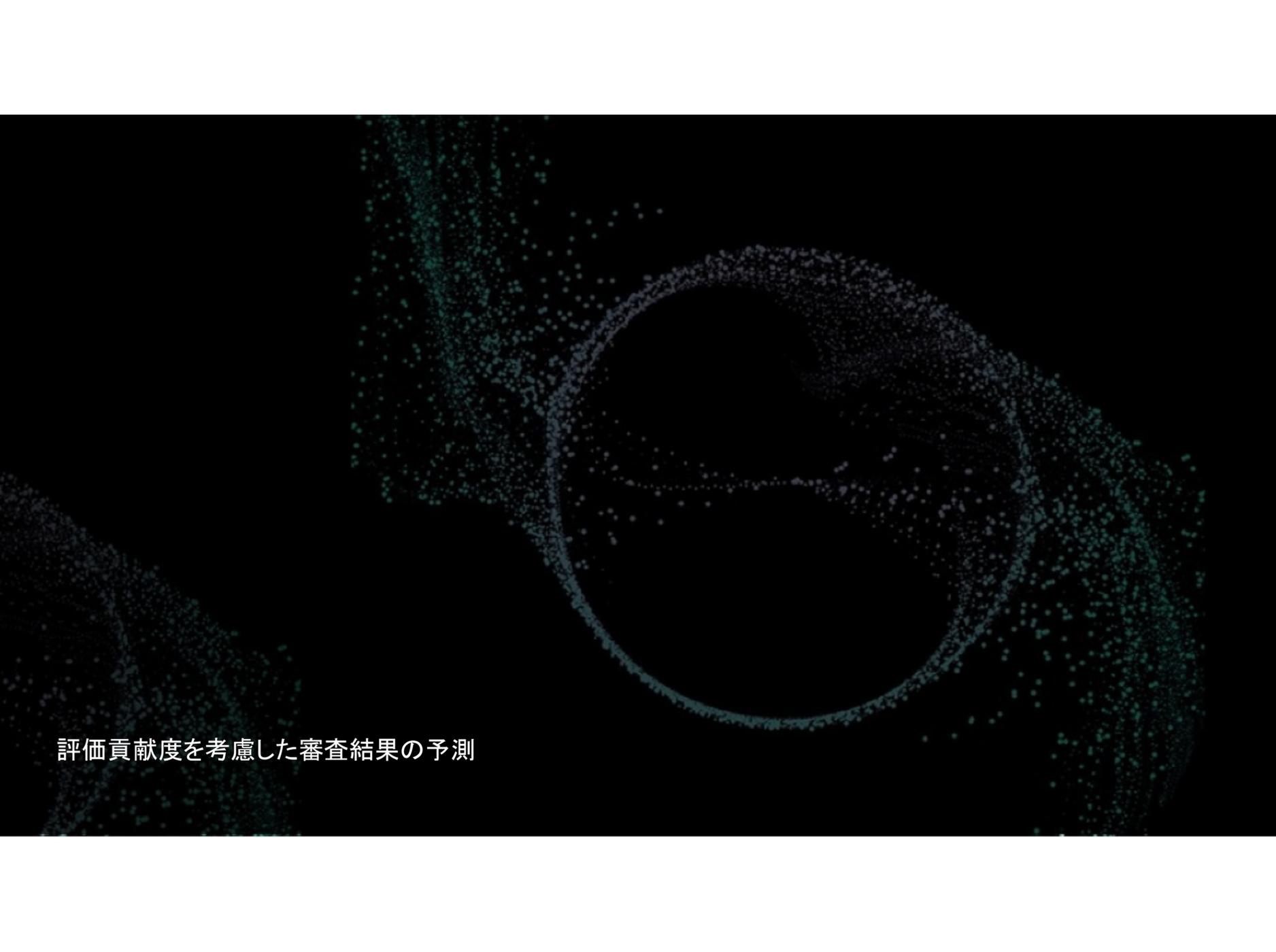


スコアの段階が変更された場合の精度への影響調査(LSTM)

- ①目標を達成するための経営的基礎力
- ②事業化計画の妥当性
- ③事業化による経済効果

- 5段階で学習、4段階で予測(検証)した場合も、スコアの相対的な予測精度(相関係数)は保たれることを確認

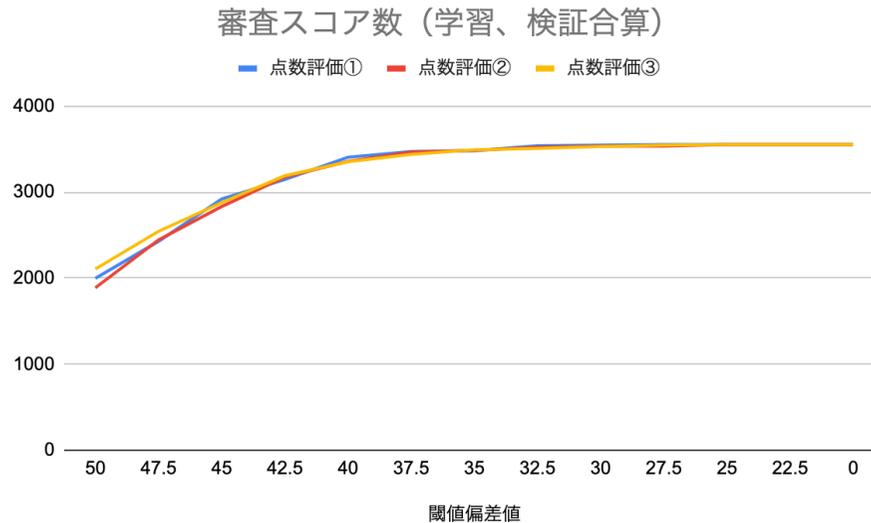




評価貢献度を考慮した審査結果の予測

評価貢献度を考慮した審査結果の予測

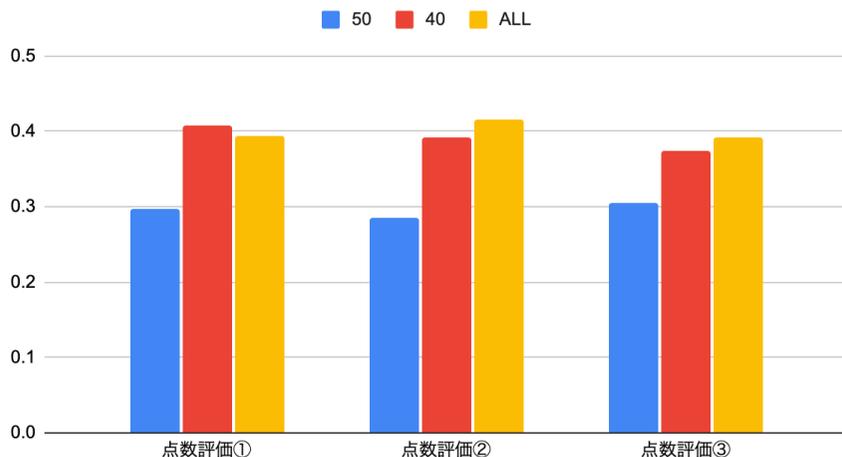
- 評価貢献度
 - 過去に実施した評価貢献度の計測結果を用いる
 - 複数回評価貢献度が算出された審査員については平均値を用いた
- フィルタリング
 - 点数評価の各軸ごとに、評価貢献度が閾値に達していない審査員の審査スコアを除外して集計し、教師・検証データとする
 - 影響の例: 3評価が2評価になるため、平均点が変わる



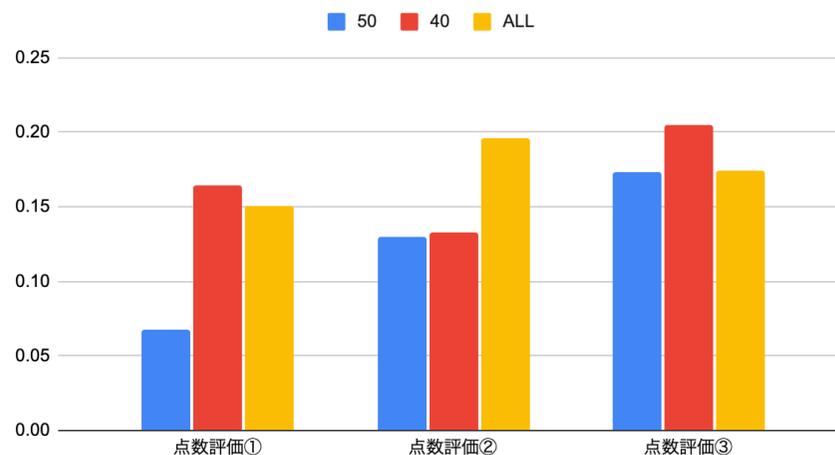
評価貢献度で閾値を設けた結果(学習データ=2018,2019、検証データ=2020、「概要」)

- 評価貢献度(各評価軸ごとの偏差値)で審査員のフィルタリングを行い、残った審査員による評価スコアのみを用いた学習・予測(検証)を実施
- 審査員フィルタリングによって明確に予測精度が向上したとまでは言えない
 - 偏差値40でフィルタリングした場合:軸によって精度微増、偏差値50でフィルタリングした場合:精度低下
 - 向上要素:フィルタリングによるスコア信頼度(学習データ)の向上
 - 低下要素:学習データが減少することによる予測精度の低下
 - →学習データが十分な量に到達すれば(量と精度が比例しなくなる程度)解消されることが予想

クラス一致率 (正解率)



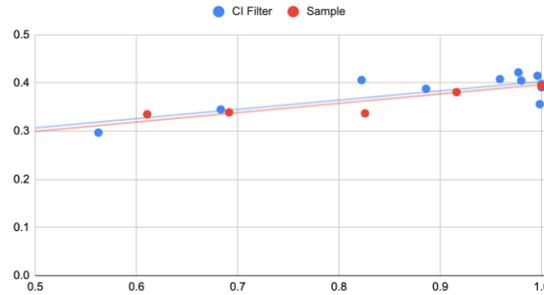
相関係数



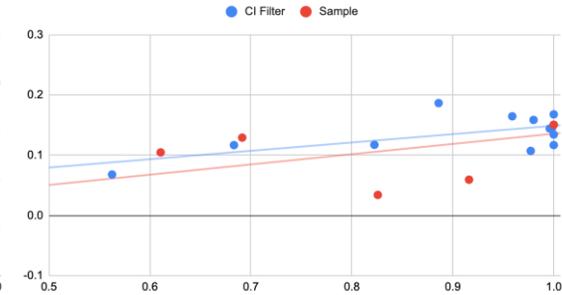
評価貢献度で閾値を設けた結果 (学習データ=2018,2019、 検証データ=2020)

- 審査数をサンプリングした結果 (Sample)と、評価貢献度でフィルタリングした場合の結果(CI Filter)を比較
- 横軸:総審査数を1とした場合のフィルタ (サンプリング)後の審査数
- 同一審査数で比較すると、評価貢献度でフィルタリングした結果が、より高精度の結果(特に相関係数)
- 今後、学習データが増えることにより、評価貢献度の活用がより有効となることが予想される

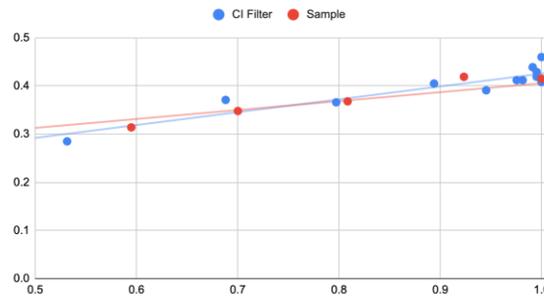
点数評価①: クラス一致率 (正解率、LSTM)



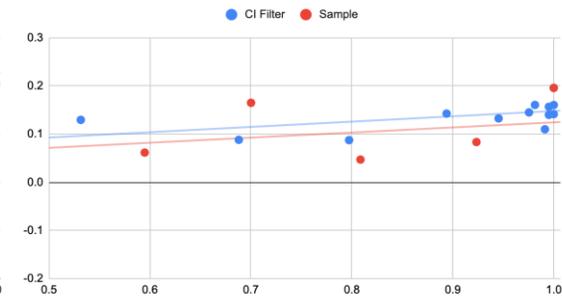
点数評価①: 相関係数 (LSTM)



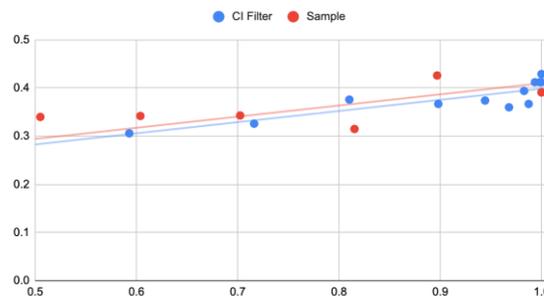
点数評価②: クラス一致率 (正解率、LSTM)



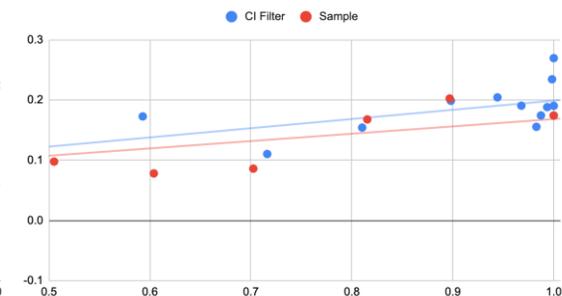
点数評価②: 相関係数 (LSTM)

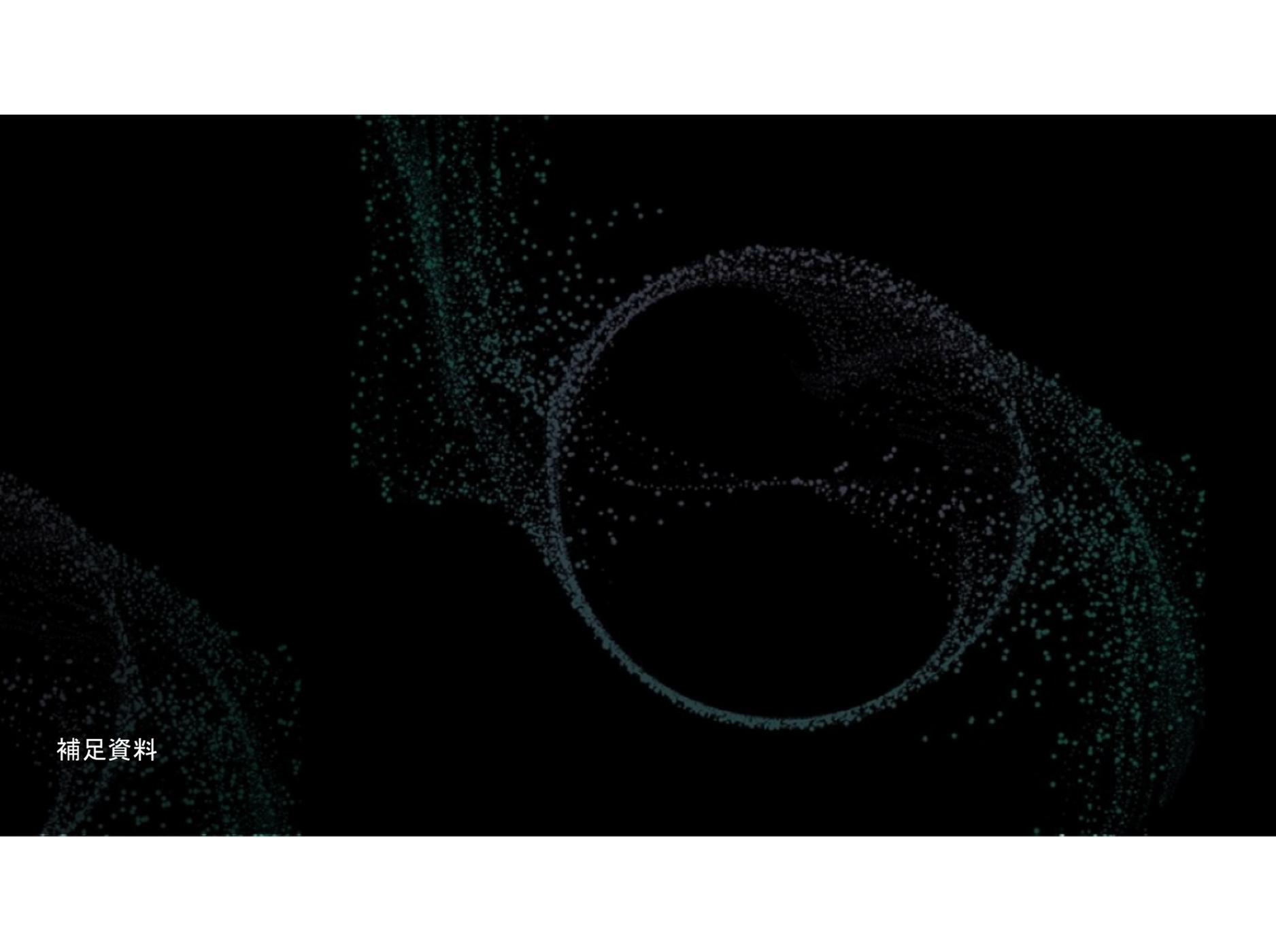


点数評価③: クラス一致率 (正解率、LSTM)



点数評価③: 相関係数 (LSTM)

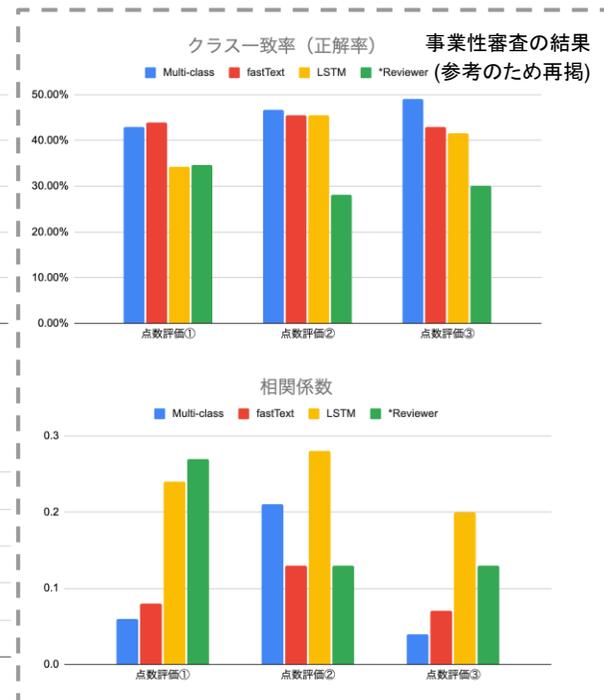
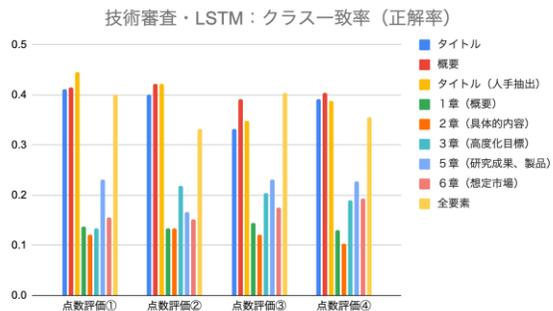
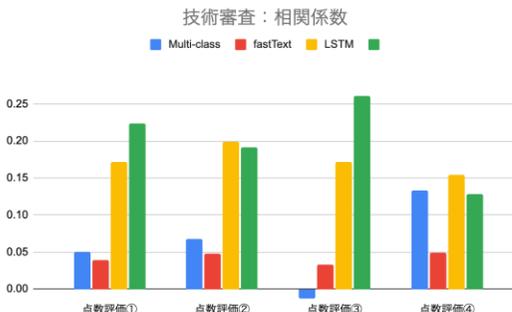
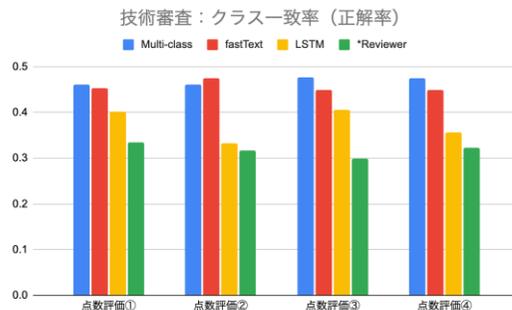




補足資料

技術審査データでの実験結果 (学習データ=2018,2019、 検証データ=2020)

- 今回検証を行った手法を、技術審査データに学習・適用した
- クラス一致率は事業性審査の結果と相当するものとなった一方で、相関係数の精度はやや低下した
- 技術審査では、審査員の専門的な背景知識に基づいて申請書のレビューが行われるため、申請書の内容のみからスコアの予測を行うハードルが(事業性審査の場合と比較して)高い可能性がある
- 審査員プロフィールを入力に含めて学習を行うシステムを考案するなどの改善案が考えられる

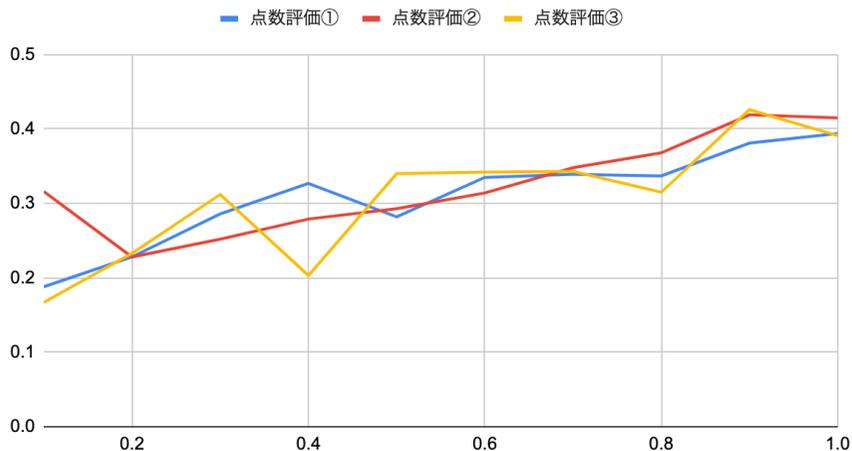


学習データ数の影響度調査(学習データ=2018,2019、検証データ=2020)

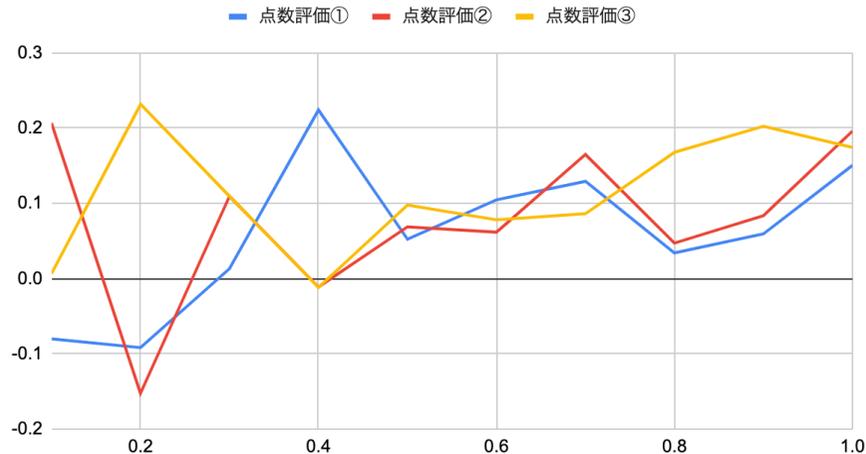
- ①目標を達成するための経営的基礎力
- ②事業化計画の妥当性
- ③事業化による経済効果

- LSTMを用い、入力に「概要」を用いた場合において、学習データ(2018, 2019年)を増減させた場合の精度への影響を確認した
- クラス一致率(正解率)、相関係数のどちらの場合も右肩上がりの傾向がみられ、暫くの場合において、学習データを増やすことで更に精度が向上することが見込まれる
 - 相関係数は学習サンプルにどのデータが選ばれたか、および初期値の影響で毎回ごとに多少の精度ブレがみられる

LSTM：クラス一致率（正解率）



LSTM：相関係数



日本を再びイノベーション大国に

VISITS



