JPX
WORKING
PAPER

JPXワーキング・ペーパー

日本の上場会社における監査上の主要な検討事項の自然言語処理を用いた分析

－監査領域の自動分類・意味的類似性・記載内容管理の実証－（要約版）

2025年11月27日

土井 惟成[1], 信田 裕介[2], 水野 豪[3]

1 株式会社日本取引所グループ 総合企画部 主任研究員,
東京大学大学院工学系研究科 博士後期課程
2 株式会社東京証券取引所 上場部 調査役
3 株式会社JPX総研 クライアントサービス部 調査役

備考

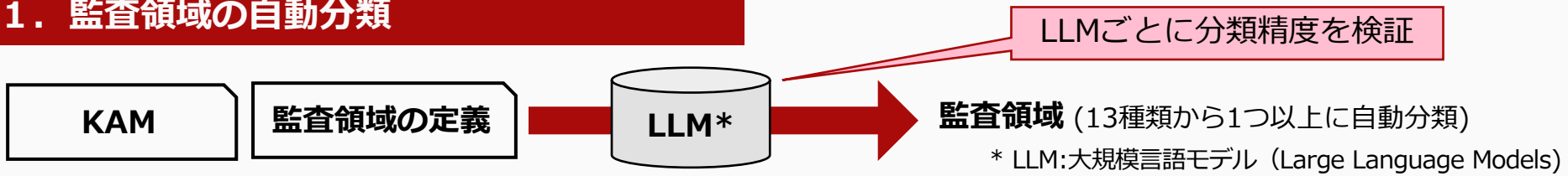
JPXワーキング・ペーパーは、株式会社日本取引所グループ及びその子会社・関連会社（以下「日本取引所グループ等」という。）の役職員及び外部研究者による調査・研究の成果を取りまとめたものであり、学会、研究機関、市場関係者他、関連する方々から幅広くコメントを頂戴することを意図しております。なお、掲載されているペーパーの内容や意見は執筆者個人に属し、日本取引所グループ等の公式見解を示すものではありません。

なお、本稿の作成に当たっては、日本取引所グループ等のスタッフから有益なコメントを頂きました。ここに深く感謝申し上げます。

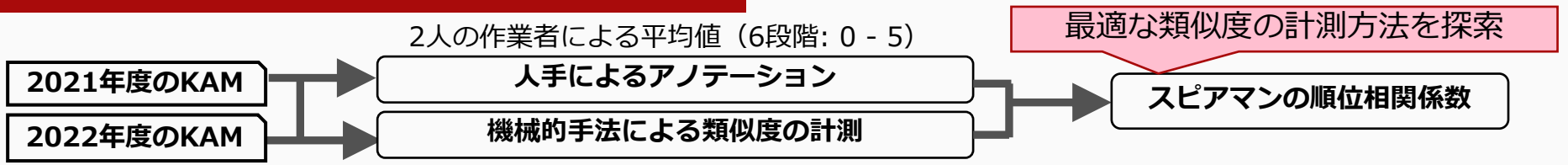
本稿の概要

- **監査上の主要な検討事項（Key Audit Matters: KAM）**は、監査人が財務諸表等の監査において、職業専門家として特に重要だと判断した事項であり、監査報告書を通じて報告される。
- KAMに対する自然言語処理を用いた分析を通じて、次に挙げる３種類の研究を通じて、 **KAMにおける自然言語処理を用いた手法の有用性**を示す。

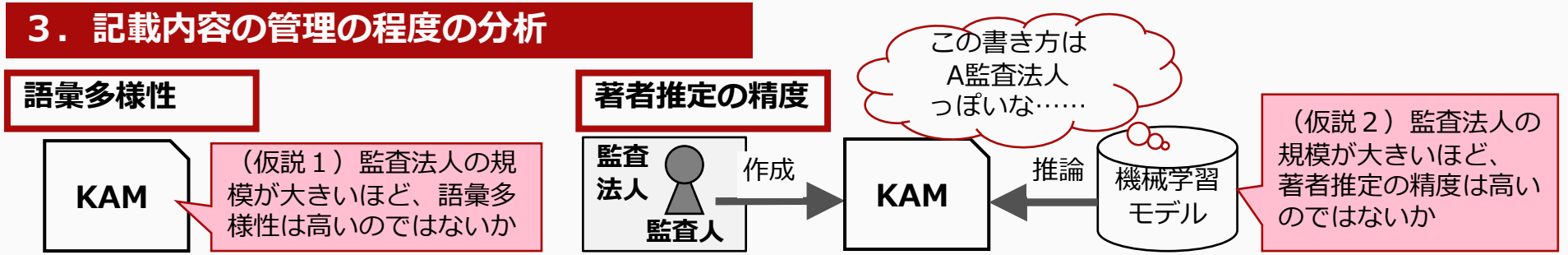
１．監査領域の自動分類



２．意味的類似性の評価指標

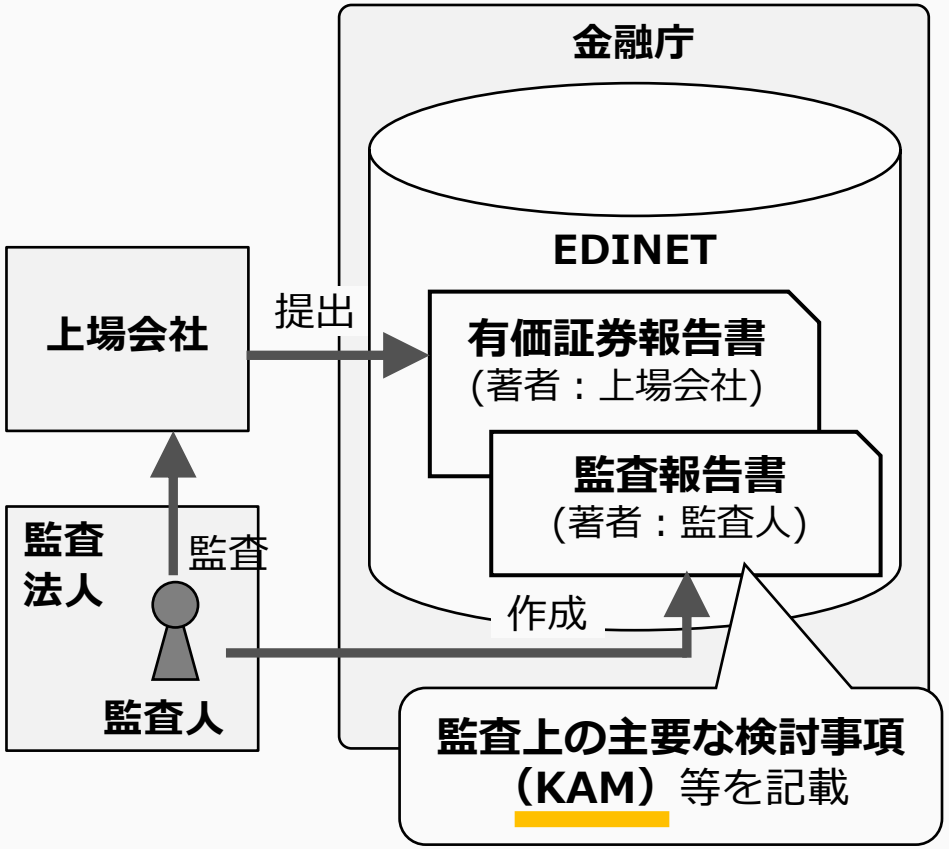


３．記載内容の管理の程度の分析



監査上の主要な検討事項（KAM）の概要

- 上場会社は、企業概況や経理状況を含む有価証券報告書を、事業年度末から3ヶ月以内に、内閣総理大臣に提出する義務がある。
- 有価証券報告書の財務諸表等は、監査人の監査を受けなければならない。
- KAMは、監査人が財務諸表等の監査において、職業専門家として特に重要だと判断した事項であり、監査報告書を通じて報告される。
- 従来のいわゆる短文式の監査報告書は透明性に欠けていたため、監査の透明性向上の観点から、2021年3月期決算から日本の全上場会社は監査報告書へのKAMの記載を義務付けられている。



KAM導入の意義と期待される効果

導入の意義

- 監査プロセスの透明性を向上させること

期待される効果

- 財務諸表利用者に対して監査のプロセスに関する情報が、監査の品質を評価する新たな検討材料として提供されることで、監査の信頼性向上に資すること
- 財務諸表利用者の監査や財務諸表に対する理解が深まるとともに、経営者との対話が促進されること
- 監査人と監査役、監査役会、監査等委員会又は監査委員会（以下「監査役等」という。）の間のコミュニケーションや、監査人と経営者の間の議論を更に充実させることを通じ、コーポレート・ガバナンスの強化や、監査の過程で識別した様々なリスクに関する認識が共有されることによる効果的な監査の実施につながること

出典：企業会計審議会「監査基準の改訂に関する意見書」
https://www.fsa.go.jp/news/30/sonota/20180706/1.pdf

KAMの例：株式会社日本取引所グループ（2025年3月期）（IFRS適用会社）

- 監査報告書には、「当期連結財務諸表に対する監査報告書」と「当期財務諸表に対する監査報告書」の2種類があり、本研究では「当期連結財務諸表に対する監査報告書」を分析の対象として設定
- 1つの監査報告書に複数のKAMが記載される可能性あり
- KAMは「見出し」「内容及び決定理由」「監査上の対応」で構成

見出し	2　ソフトウエア及びソフトウエア仮勘定の評価
内容及び決定理由	連結財務諸表注記「13．のれん及び無形資産」に記載されているとおり、当連結会計年度末において、ソフトウエアが32,556百万円、ソフトウエア仮勘定が1,751百万円計上されている。 …… 以上より、当監査法人は当該事項を監査上の主要な検討事項に相当する事項に該当するものと判断した。
監査上の対応	当監査法人は、IT専門家と連携して、ソフトウエア及びソフトウエア仮勘定の評価に係る内部統制の有効性を評価するとともに、開発中の新システムについて、減損の兆候の有無を検討するため、主として以下の監査手続を実施した。 ・開発中の新システムの活用及び開発方針について、計画時からの重要な変更が生じていないかどうかを検討するため、システム管理者に質問するとともに情報システム部門内における会議体議事録等を閲覧した。 ……

本稿の構成

- 1. LLMを用いたゼロショットテキスト分類によるKAMの監査領域の分類
  P.9-P.16
- 2. KAMの意味的な類似性の測定方法
  P.17-P.27
- 3. 監査法人における監査上の主要な検討事項の記載内容の管理に関する分析
  P.28-P.38

- なお、本研究は、以下の論文の拡張版である。
  - Nobushige Doi, Yusuke Nobuta, and Takeshi Mizuno. Topic Classification of Key Audit Matters in Japanese Audit Reports by Zero-shot Text Classification. In 2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pages 540–545, 2023. doi: 10.1109/IIAI-AAI59060.2023.00108.
  - Nobushige Doi, Yusuke Nobuta, and Takeshi Mizuno. Measuring Semantic Similarity in Japanese Key Audit Matters. In 2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pages 468–475, 2024. doi: 10.1109/IIAI-AAI63651.2024.00091.
  - Nobushige Doi, Yusuke Nobuta, and Takeshi Mizuno. Do Audit Firms in Japan Manage the Content of Their Key Audit Matters? Evidence Based on Lexical Diversity and Author Identification Accuracy. In 2025 18th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 2025. to appear.

各テーマの貢献のまとめ

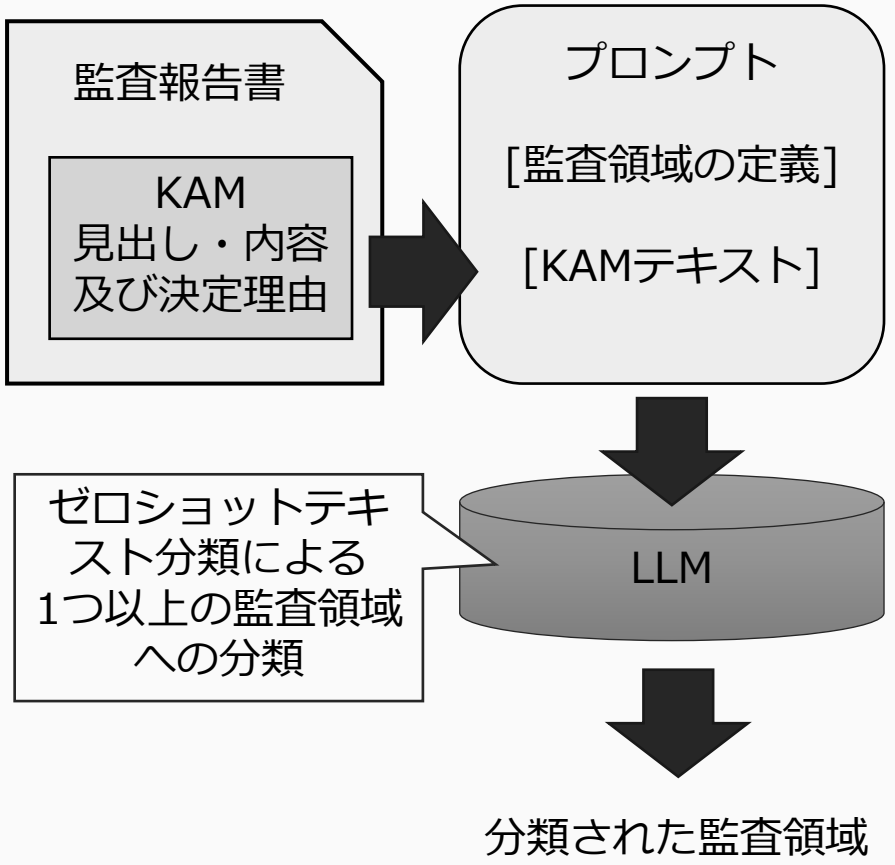
テーマ	内容	結果
1. LLMによる監査領域のゼロショットテキスト分類	KAMの「見出し」「内容及び決定理由」と13の監査領域定義をLLMに提示し、ゼロショットで監査領域を分類。 2022年4月期～2023年3月期における250件のKAMで精度を検証。	完全一致Accuracyの最大は92.8%（GPT-5）、Micro-F1は95.2%。GPT-4oは92.0%、Micro-F1は94.8%。 複数監査領域のKAMはGPT-5が高精度（Accuracy 90.9%）、単一監査領域のKAMはGPT-4oがやや優位（95.4%）。 残差カテゴリ「その他」は両モデルとも難しい。
2. KAMのテキストの意味的な類似性の評価手法	TOPIX Core 30の同一企業2ヵ年KAM 50ペアに対し、6段階（0～5）の人手アノテーションを作成。 金額・時期・割合等の数値表現を[MASK]に置換する前処理を導入。 複数の類似度の測定手法と人手のアノテーション結果の順位相関係数を分析。	数値マスキングにより全指標で人手評価との順位相関係数が一様に改善。 ボイラープレート化の程度の測定にはBLEU等の単語一致系が有効、より精緻な意味的類似性の評価にはBERTScore等の埋め込み表現に基づく手法が有効。
3.記載内容の管理の程度に関する分析	語彙多様性の評価指標と、監査人が所属する監査法人を推定する著者推定の精度を用いて、監査法人の規模別に、KAMの記載内容の管理の程度を分析。	大手監査法人は語彙が豊富でありながら、著者推定の精度が高い傾向があり、規模の小さい監査法人ほど、語彙が少ないにもかかわらず、著者推定の精度は低い傾向。 大手監査法人ほど組織的な品質管理が行われ、KAMの内容が管理されている可能性を示唆。

LLMを用いたゼロショットテキスト分類による KAMの監査領域の分類

本研究の概要

- KAMには監査領域を示すタグが存在しないため、上場会社横断の集計や年次推移の把握を行うには大きな作業コストが発生

- LLMを用いたゼロショットテキスト分類により、教師データの事前準備やモデル学習を伴わない分類手法を提案
  - ゼロショットテキスト分類＝事前にラベル付きデータを用いた学習を行わずに、新たなテキストをいずれかのラベルへ分類する手法
- 250件のKAMを対象とした13種類の監査領域の分類の結果、GPT-5を用いることで**92.8%**の精度を達成



分類対象となるKAMの監査領域の定義

#	監査領域名	定義
1	固定資産の評価	のれんを除く有形固定資産・無形固定資産の評価に係る論点
2	のれんの評価	のれんの評価に係る論点
3	収益認識	売上高を含む収益認識に係る論点（実在性・正確性に係る論点、期間帰属に係る論点、企業会計基準第29号「収益認識に関する会計基準」の適用開始に伴う論点、工事進行基準やソフトウェア開発等の一定期間にわたり履行義務が充足される契約に起因する収益認識に係る論点、等）
4	繰延税金資産の評価	回収可能性や妥当性を始めとする、繰延税金資産の評価に係る論点
5	棚卸資産の評価	在庫として保有している商品、製品、原材料、仕掛品等の、棚卸資産の評価に係る論点
6	債権の評価	売掛金等の営業債権の評価に係る論点及び、営業債権に対する貸倒引当金の見積りに係る論点
7	債務の見積り	引当金をはじめとする債務の見積りに係る論点（ただし、貸倒引当金を除く）
8	組織再編	自社に関する企業結合や分社化を始めとする組織再編に係る論点及び、前記に実施された組織再編における配分手続き等を当期に完了したものに係る論点
9	継続企業の前提	継続企業の前提に重大な疑義を生じさせるような状況が存在しているが、現時点では継続企業の前提に関する重要な不確実性が認められない場合の継続企業の前提に係る論点
10	ITシステムの評価	ITシステムの信頼性や新ITシステムへの移行・稼働をはじめとする、ITシステムの評価に係る論点
11	投融資の評価	非上場会社を始めとする投資有価証券の評価に係る論点
12	不正・不適切な会計処理	不正な財務報告、不適切な取引、内部統制の不備等に起因する、会計処理上の問題に係る論点
13	その他	上記以外の論点

LLMに入力したプロンプト

以下に述べるのは、日本の上場会社の、監査上の主要な検討事項（KAM）と、その監査領域の候補である。

KAMの見出し

…

KAMの内容及び決定理由

…

監査領域の候補の一覧

固定資産の評価, のれんの評価, 収益認識, 繰延税金資産の評価, 棚卸資産の評価, 債権の評価, 債務の見積り, 組織再編, 継続企業の前提, ITシステムの評価, 投融資の評価, 不正・不適切な会計処理

監査領域の候補の定義の一覧

固定資産の評価: のれんを除く有形固定資産・無形固定資産の評価に係る論点

…

タスク

上記の監査領域の候補から、上記のKAMが該当する1つ以上の監査領域のみを出力してください。どれにも該当しない場合は、「その他」と出力してください。

該当する監査領域

実験設定

- 対象：250件のKAM
- 分類対象の監査領域：「その他」含む13種類から1つ以上
- 正解データ：3人の専門家の合議で作成
- 入力するKAMのテキスト：「見出し」「内容及び決定理由」
- LLMに関する設定：
  - 使用モデル：gpt-4o-2024-05-13（GPT-4o）、gpt-5-2025-08-07（GPT-5）、gpt-5-mini-2025-08-07（GPT-5 mini）、gpt-5-nano-2025-08-07（GPT-5 nano）
  - GPT-4oはtemperatureを0に設定（GPT-5、GPT-5mini、GPT-5nanoはtemperatureの設定が不可能）
- 評価指標：厳密一致率（Accuracy）、部分的一致率（Any-hit）、適合率（Precision）、再現率（Recall）、Micro-F1
  - Accuracyについては、単一の監査領域を持つKAMと、複数の監査領域を持つKAMに分けて、それぞれAccuracy(単一領域)とAccuracy (複数領域)を算出

全体的な実験結果と考察

モデル	Accuracy (全体)	Accuracy (単一領域)	Accuracy (複数領域)	Any-hit	Precision	Recall	Micro-F1
GPT-4o	92.0%	95.4%	18.2%	97.6%	95.4%	94.3%	94.8%
GPT-5	92.8%	92.9%	90.9%	98.0%	92.7%	97.7%	95.2%
GPT-5mini	90.0%	87.9%	90.9%	97.6%	87.8%	96.6%	92.0%
GPT-5nano	88.0%	90.8%	72.7%	96.8%	91.0%	96.6%	93.7%

- いずれのモデルもAccuracy（全体）が高く、大半のKAMについて主要な監査領域を推定できている。
- すなわち、KAMの全社的な分析において、LLMを用いたゼロショットテキスト分類に基づく監査領域を暫定的な監査領域として与えることで、効率的に全てのKAMの監査領域を決定することが可能となる。
- 一方で、モデル別の傾向の差異が確認された。
- GPT-5は複数の候補を積極的に併記する挙動を示すのに対し、GPT-4o単一の候補を割り当てるという保守的な挙動が確認された。
- そのため、異なる複数のモデルの出力結果のアンサンブルにより、Accuracyを底上げできる可能性がある。

監査領域別の実験結果と考察

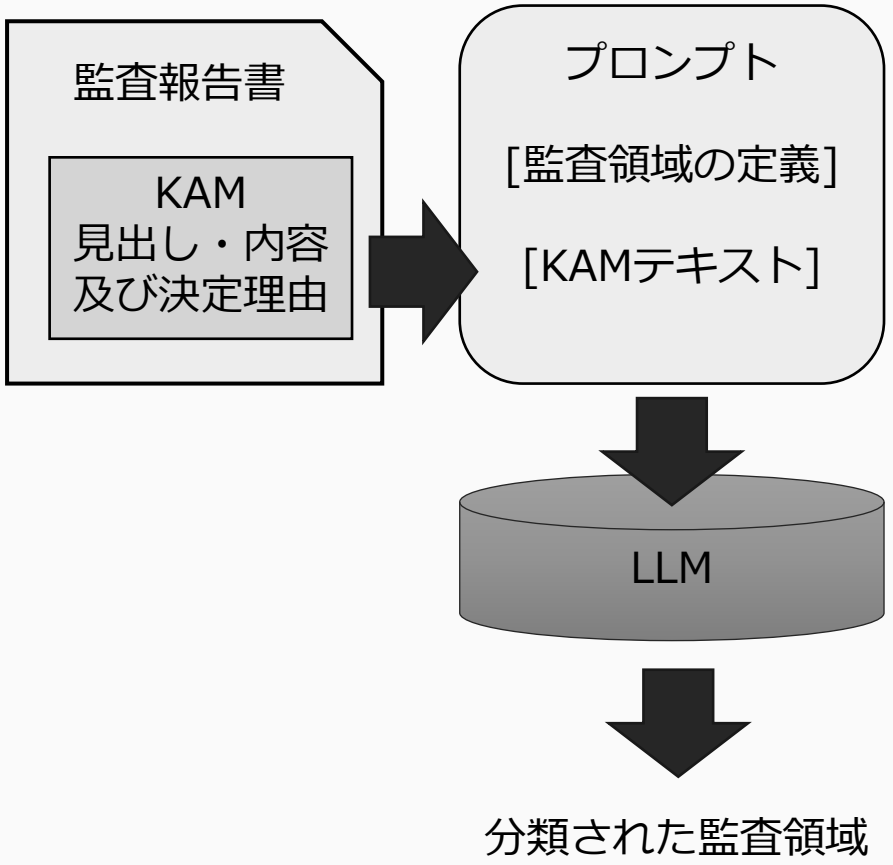
監査領域	サンプル数	GPT-4o			GPT-5		
		Precision	Recall	Micro-F1	Precision	Recall	Micro-F1
固定資産の評価	64	95.2%	92.2%	93.7%	94.0%	98.4%	96.2%
のれんの評価	23	82.1%	100.0%	90.2%	85.2%	100.0%	92.0%
収益認識	69	98.6%	100.0%	99.3%	98.6%	100.0%	99.3%
繰延税金資産の評価	25	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
棚卸資産の評価	28	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
債権の評価	15	100.0%	93.3%	96.6%	100.0%	80.0%	88.9%
債務の見積り	10	90.0%	90.0%	90.0%	76.9%	100.0%	87.0%
組織再編	4	75.0%	75.0%	75.0%	80.0%	100.0%	88.9%
継続企業の前提	5	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
ITシステムの評価	7	100.0%	42.9%	60.0%	70.0%	100.0%	82.4%
投融資の評価	5	83.3%	100.0%	90.9%	62.5%	100.0%	76.9%
不正・不適切な会計処理	3	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
その他	3	0.0%	0.0%	0.0%	50.0%	33.3%	40.0%
全体	250	95.4%	94.3%	94.8%	92.7%	97.7%	95.2%

- 残差カテゴリである「その他」への正確な分類は難しい。
- より細かい監査領域の定義が、分類精度の向上に寄与する可能性がある。

小括

- LLMを用いたゼロショットテキスト分類により、KAMの監査領域の推定を高精度に実現した。
- 一層の精度向上を目指す方法として、より詳細な監査領域の定義や、モデル特性の差異の補完を狙ったアンサンブルが有効である可能性がある。

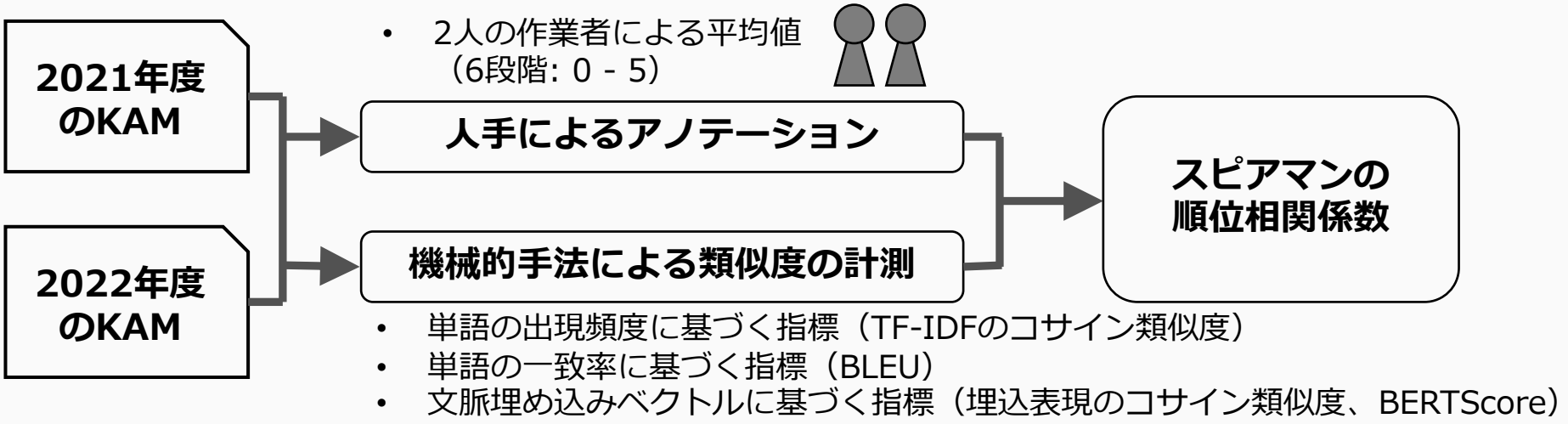
- この結果は、KAMの監査領域の推定において、自然言語処理の手法が有用であることを示唆する。



KAMの意味的な類似性の測定方法

本研究の概要

- KAMの制度上の懸念として、KAMの内容が定型化する「ボイラープレート化」の問題が指摘されている。
- そこで、本研究では自然言語処理技術を利用したKAMの意味的類似性の自動評価を検討した。
- 具体的には、単語の出現頻度に基づく指標、単語の一致率に基づく指標、文脈埋め込みベクトルに基づく指標を用いて、KAM間の類似性を計測し、その結果を人手によるアノテーションと比較した。



KAMのボイラープレート化

- ボイラープレート化： 内容が定型化又は画一化してしまう現象
  - 「テンプレート化」と類義
- 2022年3月に金融庁が主催した「KAMに関する勉強会」では、参加メンバーの主なコメントにおいて、次のとおり指摘

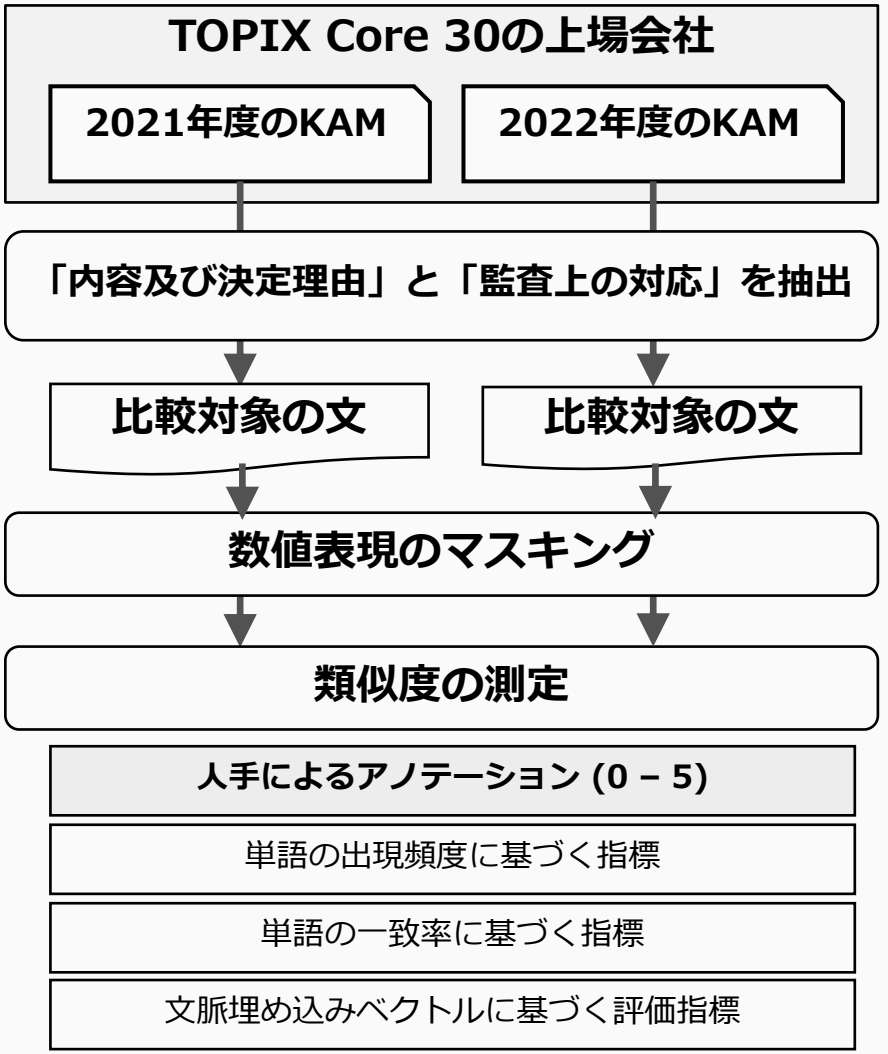
参加メンバーのコメントの一部

KAMを有益なものとするためには、定型化、画一化を避けるべきであり、**前年度の踏襲**や**監査法人横断で記載が似ている**といったような横並びは望ましくなく、各社固有の状況を具体的に記載することが重要である。

出典：金融庁「監査上の主要な検討事項（KAM）の特徴的な事例と記載のポイント」（金融庁、2022年3月4日）
https://www.fsa.go.jp/news/r3/sonota/20220304-2/01.pdf

リサーチ・デザイン

- 本研究では、既存の類似性の自動評価指標を使用することで、KAMの意味的な類似性の自動的な評価手法を検討する。
- まず、TOPIX Core 30に選定されている同一の上場会社の2ヶ年のKAMのデータセットを作成し、後述する0から5までのスコアにより、類似性を人手で評価した。
- その後、様々な自動評価指標により2ヶ年のKAMの類似性を評価し、評価用データセットの内容と比較した。
- 本研究では、これらの手法を通じて、KAMの意味的な類似性の評価手法を提案する。



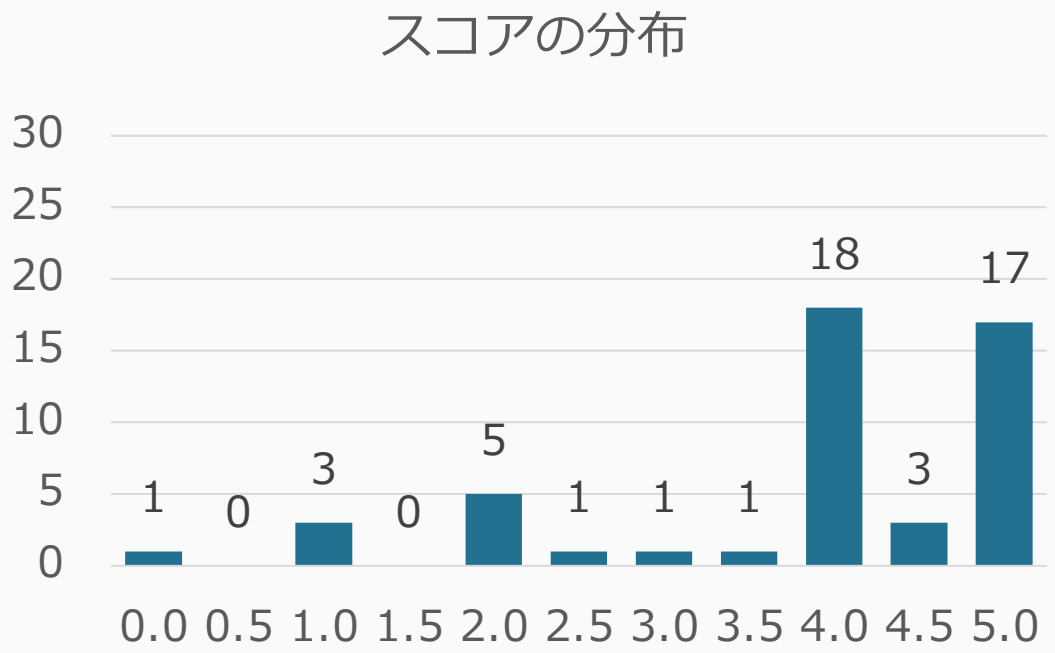
人手によるアノテーション（1/2）

- アノテーションの信頼性を確保するため、表IIのとおり、Semantic Textual Similarity (STS) に倣う形で、6段階のクライテリアを設けた。
- 1から5のスコアは、「会社の状況やリスクに関する前年度からの変化」の評価結果を示しており、0はKAMの対象論点が異なることを示している。

スコア	クライテリア
(5)	2つのKAMは、数値や時期を除けば、完全に同等で、同じ意味を持つ。
(4)	2つのKAMは、数値や時期を除けばほぼ同等だが、重要でない細部が異なる。目安として、意味が異なる段落が1つ追加／削除されている。
(3)	2つのKAMは、数値や時期を除けばほぼ同等だが、重要な情報が異なる／欠けている。これには新会計基準採用による変更も含まれる。目安として、意味の異なる段落が1つ追加／削除されている。
(2)	2つのKAMは同等ではないが、一部の詳細を共有している。目安として、意味の異なる段落が2つ追加／削除されている。
(1)	2つのKAMは同等ではないが、同じトピックについて述べている。
(0)	2つのKAMは異なるトピックについて述べている。

人手によるアノテーション（2/2）

- 各KAMは2名の作業者が独立してレビューした。そして、両者が付与したスコアの平均値をアノテーションとして採用した。なお、作業者は、5年以上の証券又は監査関連の業務経験を有している。
- 本データセットの78%においてアノテーションの結果は4.0から5.0の間となった。このことから、TOPIX Core 30に選定されている上場会社のKAMにおいて、ボイラープレート化の傾向は一定程度認められる。
- また、アノテーションの結果に対して、Shapiro-Wilkテストを使って正規性検定を行ったところ、正規性は認められなかった。



数値表現のマスキング

- KAMのテキストに含まれる数値表現が類似度算出に与える影響を低減することを目的として、数値表現のマスキングを行う。
- 具体的には、財務諸表監査に関する金額、期日、割合、年度などの数値表現を特定し、これらを[MASK]という統一的な文字列に置き換える。

マスキング前	マスキング後
会社は2015年3月期から2018年3月期までの4事業年度につき、……	会社は[MASK]期から[MASK]期までの[MASK]事業年度につき、……
2022年3月末現在の連結財政状態計算書に、のれんを339,904百万円計上しており、総資産の12.7%を占めている。	[MASK]末現在の連結財政状態計算書に、のれんを[MASK]円計上しており、総資産の[MASK]%を占めている。

本研究で採用する評価指標

単語の出現頻度に基づく指標

- KAMのテキストをTF-IDFによるベクトル化を行い、それらのコサイン類似度を使用した。

単語の一致率に基づく指標

- BLEU（BiLingual Evaluation Understudy）を用いた。
- これは、翻訳の質を測るために開発された指標であり、元々は機械翻訳の評価に用いられていたが、本研究ではKAM文書間の類似性を評価するために応用する。

文脈埋め込みベクトルに基づく指標

- KAMのテキストを、BERTモデルとDeBERTaモデルによるベクトル化を行い、それらのコサイン類似度を使用した。
- 併せて、これらのモデルに基づくBERTScoreを使用した。

実験設定

- この実験では、各評価指標によるKAM文書間の類似度スコアを計算し、それらのスコアと人手によるアノテーション結果との相関を分析した。
    - この分析には、アノテーションのスコアに正規性が認められなかったことを踏まえ、スピアマンの順位相関係数を用いて、評価指標の精度を定量的に評価した。

- TF-IDFとBLEUにおけるトークナイズには、MeCabとUniDicを用いた。
- BLEUスコアの算出にはSacreBLEUを使用し、KAMのテキストの短い変化と長い変化の評価バランスを考慮し、デフォルト値であるn=4とした。
- Embeddingsのコサイン類似度とBERTScoreの算出時に使用したモデルには、日本語に最適化されたBERTモデルとDeBERTaモデルを採択した。

実験結果

自動評価指標	モデル	スピアマンの順位相関係数	
		数値表現のマスク前	数値表現のマスク後
TF-IDF	MeCab + UniDic	0.649	0.845
BLEU	SacreBLEU (n=4)	0.787	0.856
BERTScore	BERT	0.794	0.855
	DeBERTa	0.770	0.853
埋め込み表現のコサイン類似度	BERT	0.783	0.853
	DeBERTa	0.759	0.844

- マスキングの前後で全ての指標において相関係数が一様に改善した。
  - マスキングにより類似性の評価に係るノイズが除去されたものと推察。
- BLEU、BERTScore（BERT、DeBERTa）、BERTによる埋め込み表現のコサイン類似度を見ると、順位相関係数は約0.85から約0.86の範囲に収斂した。
- 表層的なボイラープレート化の程度の計測には計算に係る負担が少ないBLEUを、精密な類似性の評価にはBERTScoreやBERTによる埋め込み表現のコサイン類似度を採用するなど、目的に応じた使い分けが考えられる。

小括

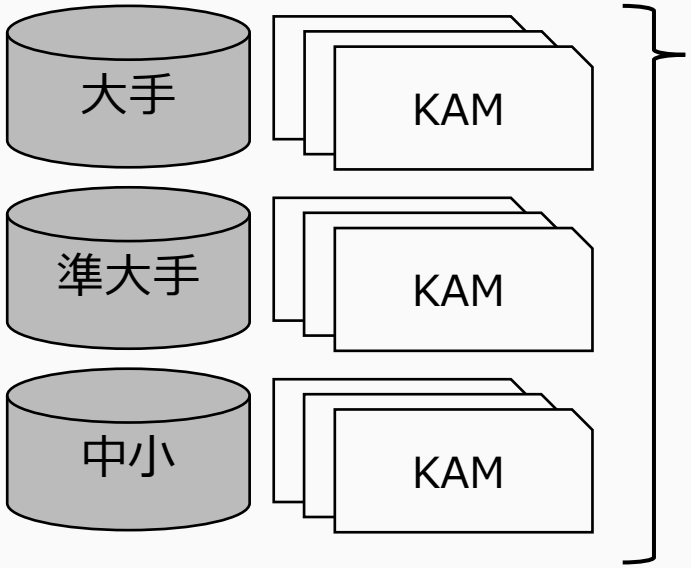
- 全ての評価指標において数値表現のマスキング後に一様な改善が見られたことから、意味的な類似性の評価におけるマスキングの有用性を確認した。
- 単語の一致率に基づく指標がボイラープレート化の程度を測定するのに有効であり、文脈埋め込みベクトルに基づく評価指標がKAMの意味的類似性を捉える上で有効であることが判明した。

- この結果は、KAMの意味的類似性の評価において、自然言語処理の手法が有用であることを示唆する。

監査法人における監査上の主要な検討事項の記載内容の管理に関する分析

本研究の概要

- 本研究は、KAMを対象に、監査法人による品質管理の存在と程度を、自然言語処理の技術を用いて調査したものである。
- 語彙多様性の評価指標と、埋め込みモデルを用いた著者推定の精度に基づく分析により、大手監査法人は語彙が豊富でありながら、著者推定の精度が高い傾向があった。
- これらの結果は、大手監査法人ほど組織的な品質管理が行われ、KAMの内容が管理されている可能性を示唆している。



疑問: 大手監査法人ほど、KAMの内容は管理されているか？

H1: 大手ほど、監査対象の会社の状況に応じた内容を記載しているのではないか
→大手ほど、語彙多様性が高いのではないか

H2: 大手ほど、各監査法人のスタイルに応じた記載になっているのではないか
→大手ほど、著者推定の精度が高いのではないか

大手監査法人におけるKAMの品質管理

- 日本の大手監査法人では、KAMのボイラープレート化を避けるよう、各監査現場を支援する組織的な取り組みを行っていることが伺える。
  - PwCあらた：品質管理本部内にKAM担当チームを設置し、各監査チームの相談対応や文面レビューを積極的に行っている。
  - トーマツ：金融庁や証券アナリスト協会が公表する好事例を分析し、改善ポイントを監査チームにフィードバックするとともに、KAM検索可能なデータベースを構築し、過去の事例分析を可能にしている。
  - あずさ：外部のステークホルダーとの定期的な対話を通じたフィードバックの獲得や、重層的なレビュー体制の整備により、KAMの固定化を防ぎ、情報価値を高める工夫を行っている。

- これらの取り組みはいずれも、KAMの品質を高めるだけでなく、監査チーム内での文面統制や専門家としての判断を適切に反映するための仕組みを整えている点が特徴である。
- こうした仕組みの存在は、各監査法人のKAMにおいて、記載内容が各社に応じたものとなり、文章表現のバリエーションが保持されつつも、一定の統一性や一貫性をもたらすと考えられる。
- **しかしながら、このような監査法人によるKAMの管理による影響について、計量的な測定はこれまでに為されていない。**

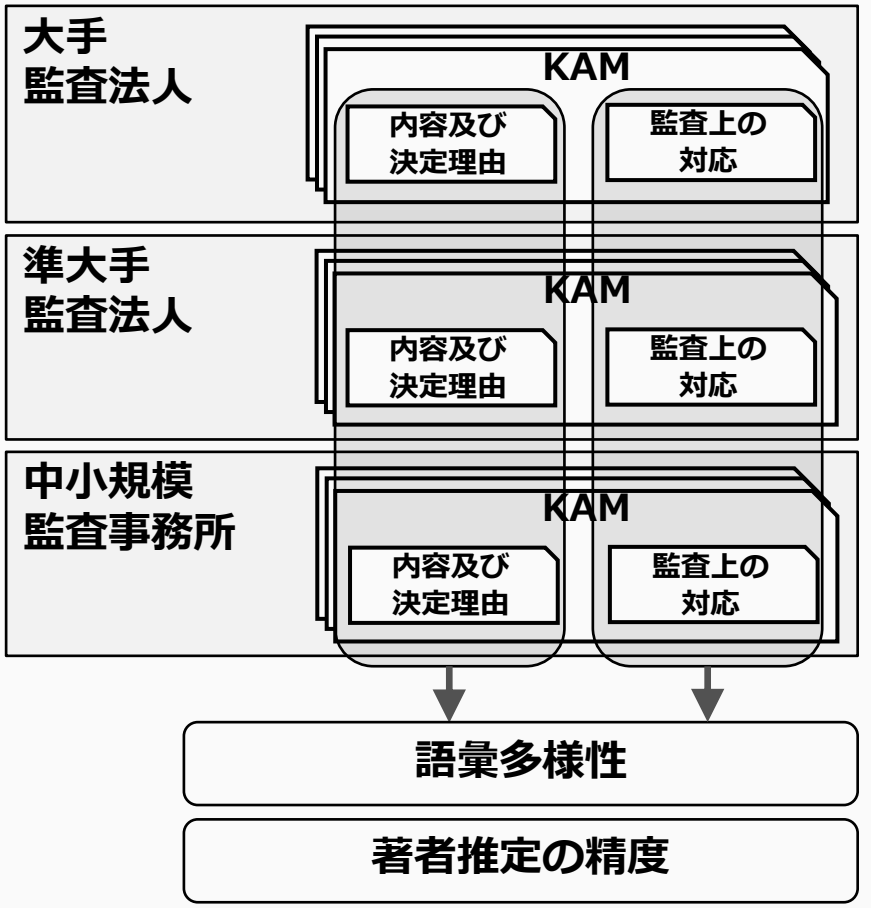
仮説の設定

仮説１：大手監査法人、準大手監査法人、中小規模監査事務所の順で、KAMの語彙多様性が高い。

- 大手監査法人では、KAMの品質管理の結果として、監査対象の各社に応じた内容が記載されると考えられる。
- この結果として、大手監査法人では、KAMの記載に用いられる語彙も幅広くなることが予想される。

仮説２：大手監査法人、準大手監査法人、中小規模監査事務所の順で、KAMに関する著者推定の精度が高い

- 大手監査法人は、組織としてマニュアルや体制を整備し、KAMの記載内容が最終的に法人の標準的な表現やスタイルに修正されている可能性がある。
- 結果的に、大手監査法人ほど、著者推定の精度が高くなると予想される。



使用するデータセット

- 2021年から2025年までの各年の3月期の全上場会社の連結KAM（5年分）
- それぞれのKAMにおける「内容及び決定理由」と「監査上の対応」の両方のテキストについて分析を実施
- 一定数のKAMを記載している監査法人を対象とするため、5ヵ年のそれぞれにおいて10社以上の上場会社の監査を行っている監査法人を対象とした。
- この結果、規模別の監査法人の数の内訳は次のとおりとなった。
    - 大手監査法人：4法人
    - 準大手監査法人：5法人
    - 中小規模監査事務所：8法人
- なお、2023年12月1日、PwCあらた有限責任監査法人はPwC京都監査法人を吸収合併し、PwC Japan有限責任監査法人に改称した。そのため、本研究では、2023年度以前の準大手監査法人は5法人、2024年度以降の準大手監査法人は4法人として取り扱う。

語彙多様性に係る評価指標

- 監査法人別のKAMの語彙多様性を評価するため、同一の監査法人の**10件のKAMを1つのデータセットとして実験**を行った。
  - 大手監査法人では10件×11グループ、準大手監査法人では10件×3グループ、中小規模監査事務所では10件×1グループのKAMをランダムに抽出した。
- 本研究では、10個連結されたKAMのテキストを対象として、次の3つの、テキスト長への依存が少ない語彙多様性の指標を使用した。

手法	概要	解釈	数式
Maas	語彙数と単語数のべき関数	**小さいほど**語彙多様性が高い	$\frac{\log N - \log V}{(\log N)^2}$
HD-D	ある単語数（通常42語）の中に文書中の単語が出現する確率の総和	**大きいほど**語彙多様性が高い	$\frac{1}{N}\sum_{i=1}^{V}\left(1 - \frac{\binom{N-f_i}{n}}{\binom{N}{n}}\right)$
MTLD	古典的な語彙多様性の評価指標がしきい値（通常0.72）に落ちるまでの単語数の平均値	**大きいほど**語彙多様性が高い	$\frac{N}{S + \left(\frac{1 - TTR_{end}}{1 - \theta}\right)}$

N=総単語数、V=異なり語彙数
fi=単語iの出現頻度、
n=サンプルサイズ（通常42）
S=しきい値に達したセグメントの数
Θ=しきい値（通常0.72）

語彙多様性に係る実験結果

内容及び決定理由

評価指標	規模	2021年	2022年	2023年	2024年	2025年	平均
Maas（↓）	大手監査法人	0.0273	0.0272	0.0269	0.0269	0.0270	0.0271
	準大手監査法人	0.0274	0.0271	0.0272	0.0271	0.0275	0.0272
	中小規模監査事務所	0.0287	0.0285	0.0280	0.0284	0.0284	0.0284
HD-D（↑）	大手監査法人	0.8118	0.8111	0.8117	0.8133	0.8126	0.8121
	準大手監査法人	0.8080	0.8102	0.8079	0.8091	0.8084	0.8087
	中小規模監査事務所	0.8026	0.8010	0.8032	0.8038	0.8027	0.8026
MTLD（↑）	大手監査法人	58.4935	58.1984	58.1420	57.6568	57.8547	58.0691
	準大手監査法人	58.2540	58.1833	57.3869	58.0288	57.5774	57.8933
	中小規模監査事務所	57.0173	56.3265	59.2736	57.4718	57.9392	57.6057

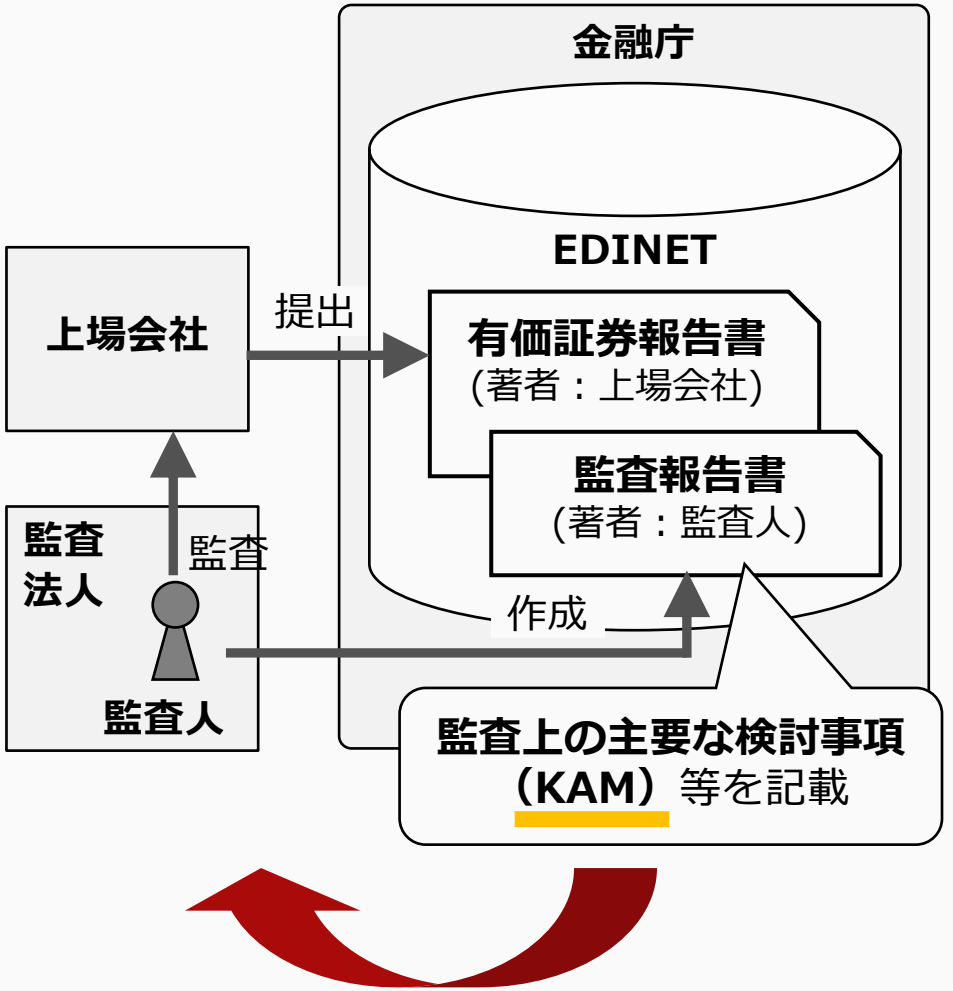
監査上の対応

評価指標	規模	2021年	2022年	2023年	2024年	2025年	平均
Maas（↓）	大手監査法人	0.0300	0.0297	0.0297	0.0298	0.0298	0.0298
	準大手監査法人	0.0305	0.0301	0.0300	0.0299	0.0304	0.0302
	中小規模監査事務所	0.0315	0.0310	0.0309	0.0312	0.0308	0.0311
HD-D（↑）	大手監査法人	0.7646	0.7661	0.7667	0.7669	0.7664	0.7661
	準大手監査法人	0.7599	0.7633	0.7639	0.7651	0.7637	0.7631
	中小規模監査事務所	0.7567	0.7573	0.7547	0.7568	0.7607	0.7572
MTLD（↑）	大手監査法人	49.6956	49.3724	49.3798	49.1045	49.4004	49.3905
	準大手監査法人	48.3184	48.4697	48.9176	49.2503	48.8091	48.7290
	中小規模監査事務所	47.0216	46.9647	46.9794	48.0687	48.3416	47.4752

- いずれの評価指標でも、3つのクラスの中で、大手監査法人の語彙多様性が最も高い傾向を示した。
- HD-Dは、一貫して、大手監査法人＞準大手監査法人＞中小規模監査事務所の順に語彙多様性が高い傾向を示しており、年度差による大きな変動は無い。
- これらの結果は、仮説１を支持する。

本研究における著者推定の概要

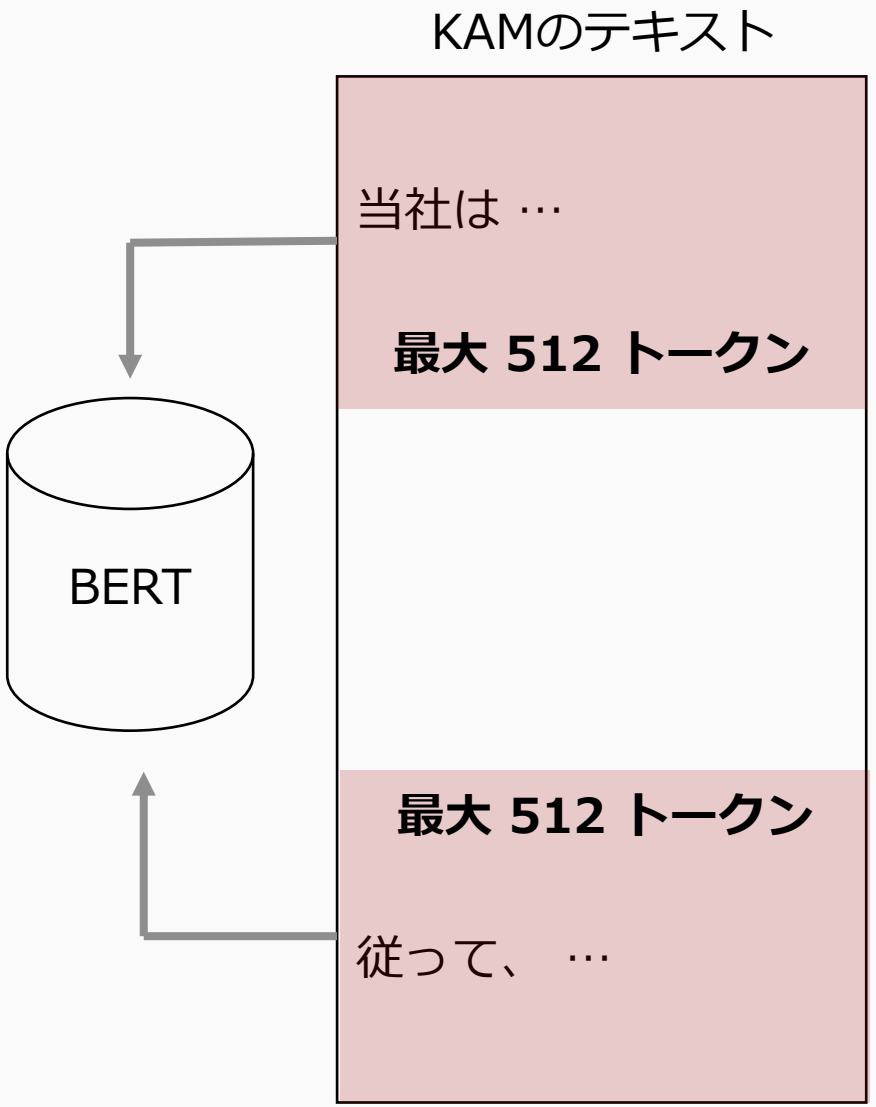
- 本研究のもう一つの分析視点である「著者推定精度」は、KAM文書を「どの監査法人が作成したものか」を判別するタスクとして位置づけることで測定する。
- つまり、各KAMを特徴ベクトルに変換し、監査法人をラベルとした分類問題を設定する。著者推定モデルの精度が高いほど、「監査法人ごとに文書の書きぶり（スタイル）が一貫している」ことを意味すると考えられる。



KAMのテキストから監査法人が推定できるか？

著者推定の手法

- 本研究では、著者推定モデルの学習のため、事前学習済みモデルとして日本語のBERTモデルを利用する。
- ただし、KAMの「内容及び決定理由」と「監査上の対応」のテキストは、BERTの一般的な入力上限である512トークンを超える傾向がある。
- そのため、本研究では、KAMのテキストの先頭近辺と末尾近辺に監査法人ごとの特徴がより強く現れると考え、これらのテキストについて先頭512トークンと末尾512トークンを抽出し、それぞれを同一ラベルとして学習データに取り込む方法を採用した。
- これにより、監査法人ごとのKAMの特徴を捉えながら、著者推定モデルの学習を実現している。
- この手法を用いて5回の交差検証を通じて実施することで、精度を計測した。



著者推定に係る実験結果

内容及び決定理由

規模	2021年	2022年	2023年	2024年	2025年	平均
大手監査法人	69.46%	64.85%	59.91%	59.55%	55.85%	61.92%
準大手監査法人	30.31%	18.93%	19.06%	12.88%	15.97%	19.43%
中小規模監査事務所	20.97%	14.21%	15.30%	7.30%	3.83%	12.32%

監査上の対応

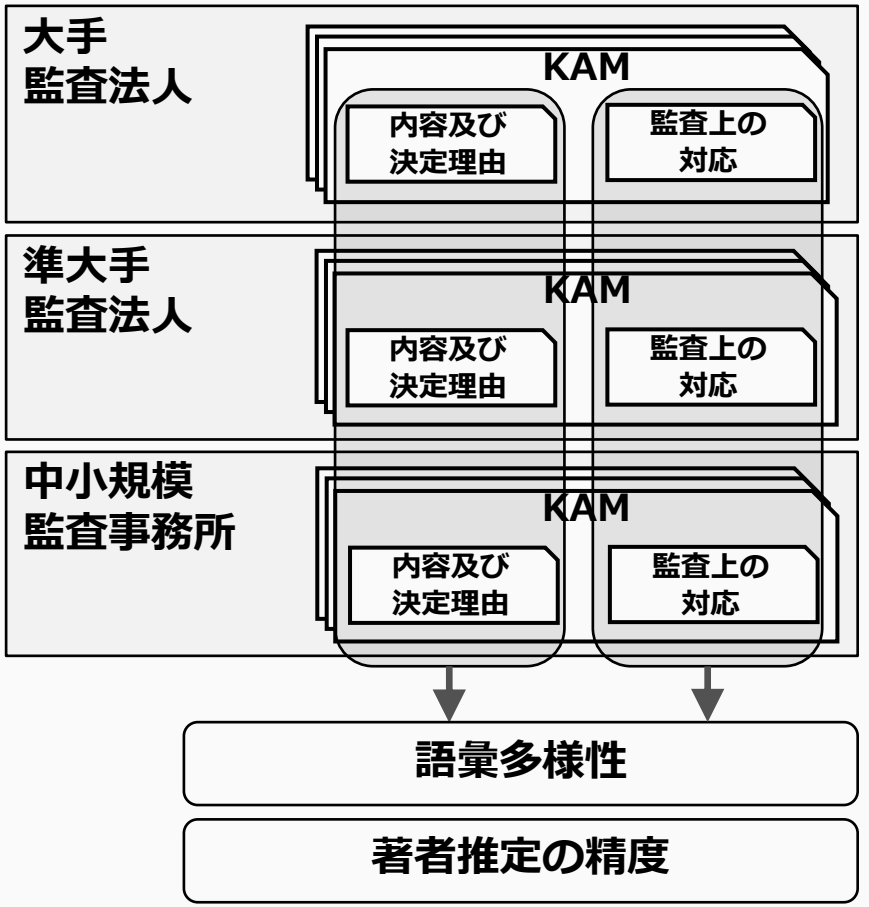
規模	2021年	2022年	2023年	2024年	2025年	平均
大手監査法人	77.52%	78.22%	75.44%	60.68%	65.82%	71.53%
準大手監査法人	34.86%	33.42%	27.45%	17.08%	22.56%	27.52%
中小規模監査事務所	28.62%	16.94%	17.31%	4.64%	8.14%	15.13%

- 「内容及び決定理由」と「監査上の対応」のいずれにおいても、大手監査法人の推定精度が相対的に高い。
- 規模別の傾向として、著者推定の推定精度は大手監査法人＞準大手監査法人＞中小規模監査事務所の順で高い傾向があった。
- これらの結果は、仮説2を支持する。

小括

- 本研究では、日本の上場会社の監査報告書に含まれるKAMについて、語彙多様性と著者推定精度に基づく分析の結果、大手監査法人ほど多様な語彙を用いながらも統一的な文書スタイルを持つ傾向が確認された。
- この結果は、大手監査法人が組織的なレビュー体制やマニュアルを通じてKAMの内容を管理している可能性を示唆している。

- この結果は、監査法人におけるKAMの記載内容の管理に関する分析において、自然言語処理の手法が有用であることを示唆する。
- ただし、本研究の結果は、あくまで語彙多様性や著者推定の観点からの示唆であり、実際の監査品質そのものを直接評価するものではない点については留意が必要。



まとめ

本稿のまとめ

- 本稿では、KAMに対する自然言語処理を用いた分析を通じて、 3つの観点から実験を行い、KAMの分析における自然言語処理の手法の有用性を検証した。
  (1) LLMによる監査領域のゼロショット分類
  (2) KAMのテキストの意味的な類似性の評価手法
  (3) 記載内容の管理の程度に関する分析
- これらの結果は、KAMの分析等において、自然言語処理を用いた手法が現実的な精度で寄与しうること、そして実務上の応用可能性があることを示した。
- 本稿の知見が、KAMの開示実務の高度化と監査の透明性・信頼性の一層の向上、さらにはデータ駆動の監査研究の進展に資する端緒となることを期待する。

今後の課題

1. ゼロショットテキスト分類において、監査領域の階層化と細分化を行い、「その他」への分類を減らすよう、監査領域を再定義すること
2. 意味的類似性の測定において、長文への対応力が高いモデルの活用や分割入力戦略の最適化により、KAM全体の文脈を途切れなく扱うようにすること
3. 著者推定において、文体要素に焦点を当てた手法を検討すること
4. 年度をまたいだ長期パネルを構築し、意味的類似性、語彙多様性、著者推定の精度が、時系列による推移の変化を観察すること
5. 監査領域別に、意味的類似性、語彙多様性、著者推定の精度の傾向を確認すること