含意認識モデルとニューラルトピックモデルを用いた統計指標検索

笹治 拓矢 [1]　加藤 誠 [2]　小山田 昌史 [3]

本論文では，エンティティを特定の基準で良いかどうかを判断するために特定の基準を直接入力し，その程度が推測できるような統計指標を検索する問題とそれに対する手法を提案する．この問題では，特定の基準を潜在指標と呼び，その程度は統計指標によって推測可能とし，統計指標を推測しやすさの高い順に順位付けを行う．提案手法では，潜在指標と内容的なつながりにある統計指標かを推論（含意）の確率で順位付けする「含意認識手法」と，統計指標の数量情報をもとに潜在指標の程度を予測できかで順位付けする「相互情報量手法」を提案する．各手法のランキング結果は，クラウドソーシングを用いて評価し，含意認識手法が最も高い性能を示した．

1　はじめに

近年，市区町村，会社，製品などの実体・概念（以下，「エンティティ」と呼ぶ）を順序付けしたコンテンツが多くの Web サイトで紹介されている．例えば，市区町村の住みやすさランキング，都道府県の魅力度ランキング，働きがいのある企業ランキングなどがある．人々は，そのようなエンティティの順序を意思決定，比較，計画などに利用している．ただし，場合によっては緻密さや信頼性に欠けたコンテンツもあることが指摘されており [18, 19]，不確かな証拠だけで知識を獲得することは偏見やステレオタイプにつながる可能性がある．そのため，エンティティを特定の基準で良いかどうかを判断するために，人の立場や状況によって見方や評価が変化しない，公的統計調査等から指標化された客観的指標（以下，「統計指標」と呼ぶ）の値なども見ることが求められる．例えば，病院数や刑法犯認知件数などが挙げられる．本研究では，このような特定の基準を「潜在指標」と呼び，「直接は観測できないが統計指標により推測可能な指標」と定義する．統計指標の値は，総務省が開発した統計ダッシュボード[*1]に名称を入力することで得られる．

しかし，ユーザがエンティティの潜在指標（例: 住みやすさや魅力など）の程度を推測するのに適した統計指標を知っているとは限らないため，適当な統計指標の値を入手できないことがある．例えば，ユーザが潜在指標の程度を推測できるような統計指標を把握できていない場合，統計ダッシュボードのような統計指標の値が調べられる検索システムに統計指標の名称を入力できず，適当な統計指標の値が入手できない．その問題を解決するため，ユーザの事前知識が不足していたとしても統計指標の値を入手できるように，統計指標の名称を入力とする統計ダッシュボードのような方法ではなく，エンティティの潜在指標を直接入力できる技術が求められる．

そこで本論文では，エンティティの潜在指標を直接入力し，潜在指標の程度が推測しやすくなる統計指標を検索するタスク（以下，「統計指標検索タスク」と呼ぶ）とそれに対する手法を提案する．本研究により，ランキングを閲覧するユーザにとってエンティティの潜在指標の程度を推測するための知識を深めることが期待され，ランキングの作成者にとって少ない労力で魅力的かつ透明性の高いランキングを作成することが可能となる．

エンティティの潜在指標の程度が推測しやすい順に統計指標を順位付けする手法として，「含意認識を用いた手法」と，「相互情報量を用いた手法」の 2 つを提案する．なお，本論文では，それらの手法を以下，含意認識手法，相互情報量手法と呼ぶ．前者の手法は，潜在指標の程度を推論できるような統計指標を探すため，統計指標の相対的な量（大きいや小さいなど）に関する前提文をもとに，潜在指標の程度に関する仮説文が成り立つ確率（含意確率）を求め，より推測するのに適した統計指標を含意確率によって順位付けする．例えば，「ある市は犯罪件数が多い」から「ある市は治安が悪い」かを含意認識すると犯罪件数の含意確率が求められる．ここでいう相対的な量とは，指標の数量を基準からの相対的な量で表したものであり，その量は大小の度合（高いや低いなど）で言語表現できるものとする．後者の手法は，統計指標の値を入手することで潜在指標の予測しにくさがどの程度減るかを調べ，不確かさの減少度でランキングを作成する．例えば，「高齢者数」を入手することで，「高齢化」の不確かさはある程度減少すると考えられる．このような，2 つの確率変数において，片方の値を観測した際にもう一方の値の不確かさがどの程度減るかは相互情報量と呼ばれており，本手法では相互情報量を用いて順位付けを行う．これらの手法は，エンティティの潜在指標が推測しやすくなる統計指標を「内容的なつながり」と「数量情報」をもとに順位付けすることで統計指標検索を実現する．

統計指標検索タスクの実験を行うため，本研究では市区町村のエンティティを対象とし，検索対象となる統計指標（783 種類）のデータセットを構築した．また，後者の手法のエントロピー計算で必要となる確率を得るために，統計指標のデータセットと共に利用する，市区町村名と西暦や和暦のような年の表現を含むテキストコーパスを多言語版テキストコーパス（mC4）を用いて構築した．手法の性能を評価するため，用意した 50 個のクエリ（潜在指標を表す単語）に対するランキングを手法毎に出力した後，クラウドソーシングを用いて，上位 10 件の統計指標が潜在指標を推測するのに適しているかどうかの判定を行った．この判定結果を利用し，2 つの評価指標（P@10 と nDCG@10）でベースライン手法と提案手法の比較を行った．その結果，含意認識を用いた提案手法が両方の評価指標で最も高い性能を示した．

この論文における貢献を以下に示す：(1) 統計指標検索タスクを提案した．これによって，魅力のような潜在指標の程度を推測するのに有効な別の観測可能な代替的な客観的指標（統計指標）

[1] 学生会員　筑波大学大学院 情報学学位プログラム
s1913574@klis.tsukuba.ac.jp
[2] 正会員　筑波大学 図書館情報メディア系
mpkato@acm.org
[3] 正会員　日本電気株式会社 データサイエンス研究所
oyamada@nec.com
[*1] https://dashboard.e-stat.go.jp/

を理解し，エンティティをさまざまな基準で良いかどうかを判断することが可能となる．(2) 実験により，含意認識用にファインチューニングされた事前学習済みのニューラル言語モデルは統計指標の大小の度合を捉えられることが判明した．相対的な量に関するテキスト文を用いた含意認識により高品質なランキングを作成できることを示した．(3) 幅広い潜在指標に対する実験を行い，含意認識を用いた手法の有効性を示した．

本論文の構成は以下の通りである．2 節では関連研究について，3 節では問題設定と提案手法について述べる．4 節では実験について説明し，5 節では本論文の結論と課題について述べる．

2 関連研究

本節では，統計指標検索タスクに類似するエンティティ検索タスクの関連研究と，提案手法で用いる含意認識モデルとニューラルトピックモデルの関連研究について述べる．

■コンセプトによる説明可能モデル 本研究で対象とする問題は，機械学習を使った意思決定支援においても存在する．具体的には，予測結果を出力する機械学習モデルの解釈性の低さ（算出方法の不透明さ）によって結果の信頼性が低くなり，判断根拠を把握する手段が求められていることと似ている．機械学習モデルの個々の予測に対する説明技術の一つである，モデルが専門家が判断に用いる概念を根拠として提示する方法 [7, 9, 13, 16] は本研究と関連している．しかしながら，これらの方法は，客観的根拠は提示できる一方で，数量的に解釈することができない．

■エンティティ検索モデル エンティティ検索タスクとは，統計指標検索タスクとは異なり，エンティティが数値属性で表現される場合に，「幸福度などの特定の基準」を入力としてエンティティのランキングを返すタスクである．統計指標検索タスクは，このタスクと比べて，エンティティランキングの算出方法に関する不透明さに対する不安・不信感を減らし，信頼性が高い客観的データに基づいてエンティティの潜在指標の程度を推測することが可能となることが期待される．既存研究 [5, 6, 8, 14, 17] では，様々な観点でエンティティを順序付けする方法を提案しているが，予測順序となる根拠はユーザ側に提示されない．また，エンティティの数値属性を判断根拠とするならば，その属性値の度合によってもエンティティ順序は変化するため，数値属性の値の度合により順序が決まるようなランキングを出力できる手法が求められる．

■含意認識モデル 含意認識タスクとは，前提文 (P) と仮説文 (H) が与えられたときに，その仮説が成り立つ (含意) か，成り立たない (矛盾) か，判断できない (中立) かを判定するタスクである．なお，与えられる 2 つのテキスト文は前後関係をもち，出力ラベルは含意，矛盾，中立の 3 つとなる．このタスクは，P が正しければ H も正しいと推論できるなら “含意”，P が正しければ H は誤っていると推論できるなら “矛盾”，どちらともいえないのならば “中立” と判定する．近年では，ニューラル言語モデルを用いた含意認識タスクの研究が盛んに行われており，自然言語理解を評価するベンチマークデータセットの SuperGLUE [15] において，人間のスコアを超える性能を示した，ニューラル言語モデルの DeBERTa [4] も登場している．

■ニューラルトピックモデル 近年では，ニューラルトピックモデルの研究が盛んに行なわれており，従来よりもモデルの拡張性が増し，調整も容易になっている [11, 12]．著者や所属のような補助情報を考慮したニューラルトピックモデルの研究も行なわれている．特に，Card らは文書のメタデータとラベルをモデリングに取り込んだニューラルトピックモデル SCHOLAR を提案している [2]．本モデルは，sparse additive generative models (SAGE) [3] と supervised LDA (SLDA) [10] の優れた点を組み合わせて，SAGE と同様に観測された補助情報の特徴表現に対する明示的な偏差，およびトピックとの相互作用に対する効果を取り入れ，SLDA と同様にメタデータをラベルとして利用して，ラベルの予測に関連するトピックの推論を行うことができる．ただし，補助情報の特徴表現は単語分布に影響を及ぼすため，補助情報の特徴表現からトピックを推定するモデルではない．

3 提案手法

本節では，統計指標検索タスクと提案手法の概要について述べた後，含意認識手法と相互情報量手法について述べる．

3.1 問題設定

統計指標検索タスクとは，あるエンティティタイプ（エンティティの種類）に属するエンティティの潜在指標を表すような単語が入力された時に，そのエンティティの潜在指標の程度（高いや低いなど）が推測できるような統計指標（以下，「適合指標」と呼ぶ）を推測しやすさの高い順に出力する問題である．本研究では，エンティティの潜在指標が高いと推測されるのは，統計指標の値が高い場合か低い場合であるかを正負の符号で表すものとする．本論文では，正負の符号を大小表現と呼び，大小表現 (+) は当該指標の値が高い場合，大小表現 (−) は当該指標の値が低い場合を表す．潜在指標の程度は，同じエンティティタイプに属するエンティティとの客観的な比較によって判明するものとする．それを踏まえて，統計指標検索タスクの定式化を説明する．

本問題は，あるエンティティタイプ $c \in C$ の潜在指標 $q \in Q_c$ の名称と統計指標 k の集合 B_c に対して，B_c に含まれる統計指標 b_k を大小表現 $l_k \in \{+, -\}$ 付きでランク付けする問題であり，ランキング $((l_1, b_1), (l_2, b_2), \ldots, (l_k, b_k))$ を返す．ただし，C はエンティティタイプの集合，Q_c はあるエンティティタイプ c の潜在指標の集合，各 k については $b_k \in B_c$ とする．なお，C の要素には「市区町村」，「都道府県」，「企業」などが該当し，Q_c の要素には，「魅力」，「衰退」，「高齢化」，「貧困」などが該当する．Q_c に含まれる単語は，あるエンティティタイプ c に属するエンティティ $e \in E_c$ について言及している Web 文書 $d^{(e)} \in D_e$ 中より観測されるような単語である．ここでいう D_e とはエンティティ e について言及している Web 文書の部分集合，E_c とはエンティティタイプ c に属するエンティティの集合とする．本論文で検索対象とする統計指標 $b_k \in B_c$ は，あるエンティティタイプ c について官公庁などが調査・公開している統計データ（公的統計）より指標化された，エンティティ e の客観的指標を指す．例えば，エンティティタイプ c が「市区町村」の場合，人口や学校数，病院数などが挙げられる．統計指標 b_k の値 $a_{y,k}^{(e)}$ はエンティティ e，調査年 $y \in Y_c$ を一意に定めることで得られ，あ

る調査年 y におけるエンティティ e の統計指標のベクトルは，以下のように表現される．ただし，Y_c とはあるエンティティタイプ c について調査された年の集合とする．

$$\boldsymbol{a}_y^{(e)} = (a_{y,1}^{(e)}, a_{y,2}^{(e)}, \ldots, a_{y,k}^{(e)}, \ldots, a_{y,n_c}^{(e)})^\mathsf{T} \in \mathbb{R}^{n_c} \tag{1}$$

ここで，$\boldsymbol{a}_y^{(e)}$ は n_c 次元の任意の実数ベクトルとし，$a_{y,k}^{(e)}$ はある調査年 y におけるエンティティ e の統計指標 b_k の値とし，$k \in \{1, 2, \ldots, n_c\}$ とする．

本論文では，このランキング問題を，あるエンティティタイプ $c \in C$ の潜在指標 $q \in Q_c$ の名称と検索対象となる統計指標の集合 B_c を入力として，スコアを出力するスコア関数 $s : Q_c \times B_c \to \mathbb{R}$ を設計する問題とみなす．最終的なランキングとしては，大小表現と統計指標のタプル (l_k, b_k) をスコア関数 s の出力するスコアの降順に並べたものとなる．

3.2 含意認識を用いた手法

含意認識手法では，含意認識モデルを用いて，統計指標 b_k の相対的な量に関する前提文 t_k から潜在指標 q の程度に関する仮説文 h_q を含意認識し，仮説文 h_q が成り立つ確率（含意確率）$P_{\mathrm{NLI}}(y = 1 \mid t_k, h_q)$ によってランキングを作成する．つまり，統計指標の大小の度合に関する情報から，潜在指標の程度を推論できるかを含意認識モデルで調べて，その含意確率が高くなるような統計指標を上位になるように順位付けする．なお，含意認識モデルは，その仮説が成り立つ (含意) か，成り立たない (矛盾) か，判断できない (中立) かの 3 クラスから予測するため，含意クラスの予測確率を含意確率として扱う．また，相対的な量や程度を表現するため，程度や数量の大小などを表す形容詞を用いる．本研究では，このような形容詞を「相対的形容詞」と呼び，(相対的に) 上方にあるものに (+)，(相対的に) 下方にあるものに (−) を示すこととする．相対的形容詞の用例としては，「大きい・小さい」，「多い・少ない」，「高い・低い」，「良い・悪い」などが挙げられる．仮説文は，潜在指標のみでは文が完結されないため，「である・している調」や「相対的形容詞 (+)」などを用いる．含意認識の例としては，「ある市は犯罪件数が多い」から「ある市は治安が悪い」かを推論すると犯罪件数の含意確率が求められる．

$$P_{\mathrm{NLI}}(y = 1 \mid t_k, h_q) = \mathrm{Softmax}_y(f_{\mathrm{NLI\_SCORE}}(t_k, h_q)) \tag{2}$$

$$\mathrm{Softmax}_i(\mathbf{x}) = \exp(x_i) / \sum\nolimits_j \exp(x_j) \tag{3}$$

ここで，$y \in \{1, 2, 3\}$ は各出力クラス（含意・矛盾・中立），t_k と h_q はそれぞれ前提文と仮説文，$f_{\mathrm{NLI\_SCORE}}(\cdot) \in \mathbb{R}^3$ は含意認識モデルの出力ベクトルを表す．各クラスの予測確率は $[0, 1]$ とし，3 クラスに対する予測確率の合計は 1 とする．

さらに，統計指標 b_k の大小表現 l_k を定めるため，潜在指標 q の程度に関する仮説文 h_q は固定したまま，統計指標 b_k の相対的な量に関する前提文 $t_k^{(+)}, t_k^{(-)}$ ごとに含意認識し，より仮説文 h_q の含意確率が高くなる前提文の含意確率（=スコア）と大小表現 l_k を採用する．例えば，「この市は過疎化が進む」という仮説文に対して，「この市の空き家数が多い」と「この市の空き家数が少ない」という 2 つの前提文で含意認識を個別に行うと，前者の「空き家数が多い」の方が含意確率が高くなるため，空き家数の大小表現は + となる．

$$s(q, k) = \max_{l \in \{+,-\}} (P_{\mathrm{NLI}}(y = 1 \mid t_k^{(l)}, h_q)) \tag{4}$$

$$l_k = \arg\max_{l \in \{+,-\}} (P_{\mathrm{NLI}}(y = 1 \mid t_k^{(l)}, h_q)) \tag{5}$$

ここで，$t_k^{(+)}$ は統計指標の値が大きい場合の前提文を表し，$t_k^{(-)}$ は統計指標の値が小さい場合の前提文を表すものとする．

最終的には，大小表現 l_k と統計指標 b_k のタプルの集合を含意確率が高い順に並べかえることで，潜在指標に対する統計指標のランキングが完成する．含意認識モデルは，中立も選択肢にあることで，判定が難しい事例に対して，含意または矛盾の予測確率が高くなるのを防ぎ，モデルは潜在指標が推測しやすくなる統計指標を上手く選定できることが期待される．例えば，「犯罪件数が少なくて治安が良い」と「犯罪件数が少なくて魅力度が高い」という事例の場合，後者の方が中立の予測確率が高くなる．この理由として，事前学習時に犯罪件数と治安の文脈が犯罪件数と魅力度の文脈より強いこと，つまり，世間一般の人々が同意できるような関係を捉えられるためである．

3.3 相互情報量を用いた手法

相互情報量手法では，潜在指標 q を表すような単語 v_q の出現有無に関する不確実性が下がるような統計指標 b_k，言い換えれば，潜在指標 q を表すような単語 v_q の出現有無を予測する上で意味がある・情報量がある統計指標 b_k を適合とする．ここでいう予測しにくさはエントロピー $H(\cdot)$ を用いて数値化することとする．具体的には，統計指標 b_k の情報を入手することで潜在指標を表すような単語 v_q の予測しにくさがどの程度減るかを調べて，不確かさの減少度 $I(X_q; Y_k)$ でランキングを作成する．

$$I(X_q; Y_k) = H(X_q) - H(X_q \mid Y_k) \tag{6}$$

$$\begin{aligned} H(X_q) = &-P_v(X_q = 1) \log P_v(X_q = 1) \\ &- (1 - P_v(X_q = 1)) \log(1 - P_v(X_q = 1)) \end{aligned} \tag{7}$$

$$\begin{aligned} &H(X_q \mid Y_k) \\ &= -P_b(Y_k = 1)\{P_c(X_q = 1 \mid Y_k = 1) + (1 - P_c(X_q = 1 \mid Y_k = 1))\} \\ &\quad - (1 - P_b(Y_k = 1))\{P_c(X_q = 1 \mid Y_k = 0) + (1 - P_c(X_q = 1 \mid Y_k = 0))\} \end{aligned} \tag{8}$$

ここで，$H(X_q)$ は単語 v_q の出現有無の不確かさ，$H(X_q \mid Y_k)$ は統計指標 b_k の情報を知ったあとの単語 v_q の出現有無の不確かさを表す．$P_v(X_q = 1)$ は単語 v_q の出現確率，$P_b(Y_k = 1)$ は統計指標 b_k の情報の入手確率，$P_c(X_q = 1 \mid Y_k = 1)$ は統計指標 b_k の情報が与えられたときの単語 v_q の出現確率，$P_c(X_q = 1 \mid Y_k = 0)$ は統計指標 b_k の情報が与えられていないときの単語 v_i の出現確率を表す．また，相互情報量を使った例として，統計指標の一つである「高齢者数」を入手することで，潜在指標を表すような単語「高齢化」の出現有無に関する不確かさがどの程度減るかは，$H($ 高齢化 $) - H($ 高齢化 $\mid$ 高齢者数 $)$ より計算できる．

前述のエントロピーを計算するためには，潜在指標 q を表すような単語 v_q と統計指標 b_k の確率を求める必要がある．その確率を計算できる手法として，共起頻度を用いた手法があるが，表記揺れにより適切な共起関係や統計指標 b_k の時系列の変化を

捉えられないという問題がある．そのため，文書中に現れない潜在的な共起性を扱い，高齢化のような統計指標 b_k の時系列の変化を考慮できる，補助情報ありニューラルトピックモデルを用いる．具体的には，あるエンティティの統計指標のデータをもとにエンティティの内容（潜在指標を表す単語などの集合）に関する文書が生成されるという過程を取り入れたモデルを用いて確率を求める．最終的には，エントロピーを計算し，不確かさの減少度で統計指標 b_k を順位付けする．また，統計指標 b_k の大小表現については，モデルの学習により，統計指標と単語の対応関係が反映された重み行列が獲得できるため，学習後のモデルの重み行列 $\mathbf{W}^{bv}$ の k 行 p 列目の要素 $\mathbf{W}^{bv}_{k,q}$ の符号を用いる．

$$s(q,k) = I(X_q; Y_k) \tag{9}$$

$$l_k = \begin{cases} + & (\mathbf{W}^{bv}_{k,q} \geq 0 \text{ の場合 }) \\ - & (\text{ 上記以外の場合 }) \end{cases} \tag{10}$$

単語 v_q と統計指標 b_k の確率は，学習済みモデルの重み行列にソフトマックス関数を適用することで得ることができる．$P_v(X_q = 1)$ は $\mathrm{Softmax}(\mathbf{W}^{\mathsf{T}}_{vt})$ の q 列目のベクトルの各要素の値の平均とする．ここで，$\mathrm{Softmax}(\mathbf{W}^{\mathsf{T}}_{vt})$ は重み行列 $\mathbf{W}^{\mathsf{T}}_{vt}$ にソフトマックス関数が適用した後の行列を表し，その行列の t 行 q 列目の要素は，トピック t に対する単語 v_q の出現確率となる．$P_b(Y_k = 1)$ は統計指標 b_k の入手の有無に関する確率となるため，1/2 とする．$P_c(X_q = 1 \mid Y_k = 1)$ は $\mathrm{Softmax}(\mathbf{W}^{\mathsf{T}}_{tb}\mathbf{W}^{\mathsf{T}}_{vt})$ の k 行 q 列目の要素の値とし，その値は統計指標 b_k が与えられた時の単語 v_q の出現確率となる．$P_c(X_q = 1 \mid Y_k = 0)$ は $\mathbf{W}^{\mathsf{T}}_{tb}\mathbf{W}^{\mathsf{T}}_{vt}$ から変換された同時確率 $P_j(X_q = 1, Y_k = 1)$ と $P_b(Y_k = 1) = 1 - P_b(Y_k = 1)$ を用いて，次にように計算される．

$$P_c(X_q = 1 \mid Y_k = 0) = P_j(X_q = 1, Y_k = 0) \,/\, P_b(Y_k = 0) \tag{11}$$

$$P_j(X_q = 1, Y_k = 0) = \sum_{y \in \{1,2,3,\ldots,n_c\} \setminus k} P_j(X_q = 1, Y_y = 1) \tag{12}$$

本モデルは，Card らが提案したニューラルトピックモデル SCHOLAR [2] に基づき構成され，トピック分布を生成する際に補助情報を利用できる．この点は従来モデルとは異なる．以下では，文書生成過程と学習方法について述べる．

3.3.1 文書生成過程

文書生成過程では，エンティティ e と西暦・和暦といった年を表す数値表現（以下，「年表現」）y を含んだ文書の集合 $\mathcal{D}$ があり，各文書の単語は確率過程により生成されるものと考える．文書数は D 個，トピック数は T 個，文書 i の単語数は N_i 個，単語の種類数は V 個，文書 i の j 番目の単語は x_{ij} とする．文書 i のトピック分布を $\boldsymbol{\theta}_i \in \Delta^T$，トピック t の単語分布 $\boldsymbol{\phi}_t \in \Delta^V$ とする[*2]．文書 i の補助情報として，調査年 y におけるエンティティ e の統計指標ベクトル $\boldsymbol{a}^{(e)}_y$ が付与されており，$\boldsymbol{a}^{(e)}_y$ の k 次元目の値は統計指標 b_k の値とする．

それらを踏まえて，以下の過程により文書集合が生成されるものとする．まず，潜在変数 $\boldsymbol{r_i}$ はロジスティック分布からサンプリングし，関数 $f_{\mathrm{t}}(\boldsymbol{r}_i, \boldsymbol{a}^{(e)}_{i,y})$ よりトピック分布 $\boldsymbol{\theta}_i \in \Delta^T$ に変換する．ロジスティック分布の平均と共分散の対角項は $\boldsymbol{\mu}_0(\beta)$ と $\boldsymbol{\sigma}^2_0(\beta)$ とし，ハイパーパラメータ β を持つ対称ディリクレ分布を近似するために，$\mu_{0,t}(\beta) = 0$ と $\sigma^2_{0,t} = (T-1)/(\beta T)$ となるようにする．次に，関数 $f_g(\boldsymbol{\theta}_i) = \mathrm{Softmax}(\boldsymbol{d} + \mathbf{W}_{vt}\boldsymbol{\theta}_i)$ より単語分布 $\boldsymbol{\eta}_i \in \Delta^V$ に変換する．ただし，$\boldsymbol{d}$ は V 次元の背景項（全体の単語頻度の対数を表したもの），$\mathbf{W}_{vt} \in \mathbb{R}^{V \times T}$ は重み行列を表す．最後に，$p(x \mid \boldsymbol{\eta}_i)$ より N_i 個の単語をサンプリングする．なお，$p(x \mid \boldsymbol{\eta}_i)$ は多項分布，$\boldsymbol{\eta}_i$ はソフトマックス関数によって出力値を正規化し確率分布として扱えるようにした単語の確率分布ベクトルを表す．

本研究では，トピック分布の関数 $f_{\mathrm{t}}(\cdot)$ を次のような 3 種類の方法により，単年（または複数年）の統計指標ベクトルからトピック分布が得られるものとする．このような手法により，人口に占める高齢者の割合が増加する高齢化のような潜在指標を表す単語に関するトピックの対応も期待される．なお，本節では統計指標の値を「指標値」とし，関数 $f_t(\cdot)$ には 3 年分の統計指標ベクトル（5 年ずつ間隔をあけて取得したベクトル）を入力する．

まず，複数年の指標値の大小関係でトピック分布を予測する関数 $f^{\mathrm{vec}}_t(\cdot)$ は次のように計算される．ただし，$[\cdot;\cdot]$ は行列（ベクトルを含む）の列方向の連結を表し，$\mathbf{W}_{tb} \in \mathbb{R}^{T \times 3n_c}$ は統計指標とトピックの重み行列を表す．n_y は 5 年ずつ間隔をあけるため，5 とする．

$$\begin{aligned} \boldsymbol{\theta}_i &= f^{\mathrm{vec}}_t(\boldsymbol{r}_i, \boldsymbol{a}^{(e)}_{i,y}) \\ &= \mathrm{Softmax}(\boldsymbol{r}_i + \mathbf{W}_{tb}[\boldsymbol{a}^{(e)}_{i,y-n_y}; \boldsymbol{a}^{(e)}_{i,y}; \boldsymbol{a}^{(e)}_{i,y+n_y}]) \end{aligned} \tag{13}$$

次に，カーネルを通して複数年に渡る指標値の傾向を捉えてから指標グループ単位に集約し，指標グループが特定の変動をした際にあるトピックの値が大きくなることを考慮する関数 $f^{\mathrm{cnn}}_t(\cdot)$ は次のように計算される．ただし，$\mathbf{C}_i \in \mathbb{R}^{n_c \times 3}$ は複数年の統計指標ベクトルを横に並べた行列，$\mathbf{W}_{ker} \in \mathbb{R}^{3 \times M}$ は M 個のカーネルの傾向パターンを捉える重み行列，$\mathbf{W}_{gro} \in \mathbb{R}^{L \times n_c}$ は L 個の指標グループを捉える重み行列，$\mathbf{W}_{tm} \in \mathbb{R}^{T \times LM}$ は各トピックに対する（指標グループの）カーネルの反応を捉える重み行列を表す．また，$\mathrm{Reshape} : \mathbb{R}^{n \times m} \to \mathbb{R}^{(n \times m) \times 1}$ は行列の各行を縦方向に並べる関数，$\mathrm{Relu}(u) = \max(0, u)$ は勾配消失問題を回避するのに用いられる正規化線形関数を表す．

$$\boldsymbol{\theta}_i = f^{\mathrm{cnn}}_t(\boldsymbol{r}_i, \boldsymbol{a}^{(e)}_{i,y}) = \mathrm{Softmax}(\boldsymbol{r}_i + \mathbf{W}_{tm}\mathbf{H}_i) \tag{14}$$

$$\mathbf{H}_i = \mathrm{Reshape}(\mathbf{W}_{gro}\,\mathrm{Relu}(\mathbf{C}_i\mathbf{W}_{ker})) \tag{15}$$

$$\mathbf{C}_i = \left(\boldsymbol{a}^{(e)}_{i,y-n_y}, \boldsymbol{a}^{(e)}_{i,y}, \boldsymbol{a}^{(e)}_{i,y+n_y}\right) \tag{16}$$

次に，過去の時点の指標値を考慮してトピック分布を予測する関数 $f^{\mathrm{rnn}}_t(\cdot)$ は次のように計算される．ただし，RNN(Recurrent Neural Network）は再帰的構造（ある時点の入力がそれ以降の出力に影響を及ぼす構造）を持ったニューラルネットワーク，$\mathbf{W}_{tb} \in \mathbb{R}^{T \times n_c}$ は RNN の出力に関する重み行列を表す．

$$\boldsymbol{\theta}_i = f^{\mathrm{rnn}}_t(\boldsymbol{r}_i, \boldsymbol{a}^{(e)}_{i,y}) = \mathrm{Softmax}(\boldsymbol{r}_i + \mathbf{W}_{tb}\mathbf{h}_{i,y}) \tag{17}$$

$$\mathbf{h}_{i,y} = \mathrm{RNN}(\boldsymbol{a}^{(e)}_{i,y}, \mathbf{h}_{i,y-1}) \tag{18}$$

大小表現を定めるのに用いる $\mathbf{W}^{bv}$ については，関数 $f^{\mathrm{vec}}_t(\cdot)$ と関数 $f^{\mathrm{rnn}}_t(\cdot)$ の場合，$\mathbf{W}^{\mathsf{T}}_{tb}$ と $\mathbf{W}^{\mathsf{T}}_{vt}$ の行列積で得られるものとする．関数 $f^{\mathrm{cnn}}_t(\cdot)$ の場合，最も強く反応したカーネルの変動パ

[*2] Δ^K は要素の和が 1 になる任意の K 次元の非負実数のベクトルを表す．

ターンによって大小表現が定まり，右上がりの場合は +，右下がりの場合は – となる．

3.3.2 学習方法

本研究では，VAE でも利用されている reparametarization trick と呼ばれる手法を利用し，勾配の不偏推定を実現し，勾配推定の分散を小さくする．また，ミニバッチで学習することで，各サンプルについて潜在表現 $\boldsymbol{r}_i$ を 1 つだけサンプリングする形で学習を行う．そして，モンテカルロサンプリングによって目的関数は次のようになる．ただし，D_{KL} は 2 つの確率分布間の差異を計る尺度を表す KL ダイバージェンスとする．

$$\sum_{j=1}^{N_i} \log p(x_{ij} \mid \boldsymbol{r}_i^{(s)}, \boldsymbol{a}_{i,y}^{(e)}) - \mathrm{D}_{\mathrm{KL}}[q_{\boldsymbol{\Phi}}(\boldsymbol{r}_i \mid \boldsymbol{x}_i, \boldsymbol{a}_{i,y}^{(e)}) \| p(\boldsymbol{r}_i \mid \beta)] \tag{19}$$

潜在表現 $\boldsymbol{r}_i$ のサンプル近似を行うため，エンコーダ f_e は入力 $\boldsymbol{x}_i, \boldsymbol{a}_{i,y}^{(e)}$ から $\boldsymbol{r}_i$ をサンプルする分布のパラメータを出力する．具体的には，平均ベクトル $\boldsymbol{\mu}_i = f_{\boldsymbol{\mu}}(\boldsymbol{x}_i, \boldsymbol{a}_{i,y}^{(e)})$ と対角の共分散行列 $\boldsymbol{\sigma}_i^2 = f_{\boldsymbol{\sigma}}(\boldsymbol{x}_i, \boldsymbol{a}_{i,y}^{(e)})$ を出力するネットワークとし，$q_{\boldsymbol{\Phi}}(\boldsymbol{r}_i \mid \boldsymbol{x}_i, \boldsymbol{a}_{i,y}^{(e)}) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$ となるようにする．また，対角の共分散行列は，行列ではなく対角成分を要素にもつベクトルとする．なお，$\boldsymbol{x}_i$ は V 次元の単語ベクトル，f_e は多層パーセプトロンを表す．ニューラルネットワークの誤差逆伝搬を行うため，エンコーダ f_e で推定したパラメータとノイズより決定的に潜在表現 $\boldsymbol{r}_i$ を生成し，潜在表現 $\boldsymbol{r}_i$ から観測されたデータを生成できるよう学習することで，対数尤度は最大化され，KL ダイバージェンスも 0 に近づく．

4 実験

本節では，まず統計指標のデータセットとテキストコーパスの概略と構築方法について述べる．その後，ベースライン手法とランキングの適合判定，評価指標などの実験設定について述べる．最後に実験結果について述べる．

4.1 データセット

統計指標のデータセットは，統計ダッシュボード上で公開されている統計指標のデータを収録したものである．本研究では，検索対象とする統計指標を定める必要があるため，統計指標のデータや調査期間，エンティティタイプなどが取得可能な統計ダッシュボードを利用する．本データセット内に含まれる統計指標の種類数は 783 個である．なお，統計指標の定義については，政府が実施する統計調査の結果をオープンデータとして公開している政府統計の総合窓口 (e-Stat) の項目定義に関する Web ページ[*3]の内容に基づく．実験では，補助情報ありニューラルトピックモデルで統計指標と単語の対応関係を学習する際に，値の範囲 (Scale) を平均 0，分散 1 になるように変換したものを使用する．

テキストコーパスは，大規模な Web クロールデータの Common Crawl をフィルタリング・前処理した多言語コーパス mC4(Multilingual C4) [*4]に含まれる日本語コーパスから機械的に抽出した「固有名詞（市区町村）」と「西暦・和暦といった年表現」を含んだテキスト（4,870,000 文書）を収録したものである．なお，テキストには市区町村名と年表現が 1 つずつ含まれており，市区町村と年が 1 対 1 対応しているものとする．本コーパスは次のように構築した．まず，文章を句点・改行・記号（例: ！？）で区切り，区切られた各文に市区町村名が含まれているかを判定する．また，区切られた各文から年表現を抽出する．次に，区切られた各文ごとに市区町村名の位置から（文字数で）最も近い年表現を探索し，市区町村名とペアとなる年を決定したあと，その文の前後 5 文を含めた文章を一つの事例として扱い，本コーパスに加える．なお，ニューラルトピックモデルの学習・評価を行うために用いる，テキストコーパスの学習・評価用データはそれぞれ，3,896,000 個と 974,000 個に分割した．

4.2 実験設定

本研究では，ベースライン手法として，ランダム手法と共起回数手法を用いた．(1) ランダム手法: 大小表現の集合から重複ありでランダムに選択したもの（リスト）と，統計指標名の集合から重複なしでランダムに選択したもの（リスト）を用いて，互いのリストの同じ位置の要素同士のペアとし，ランキングを作成する手法，(2) 共起回数手法: 全部の統計指標名に対して，まず形態素解析で統計指標名から（一つまたは複数個の）名詞を抽出し，次に前節で述べたテキストコーパスを用いて，それぞれの名詞が潜在指標を表すような単語と共に出現する回数を計算し，それらの平均値を出す．最後に，統計指標名を平均値が高い順に並べてランキングを作成する手法．なお，大小表現は手法 (1) と同様にランダムに選択したものをペアとする．

含意認識モデルは，多言語コーパス CC100 で事前学習済みの mDeBERTaV3 を含意認識タスク用にファインチューニングされたモデル mDeBERTa-v3-base-mnli-xnli [*5]を用いる．ファインチューニングで用いられたデータセットは，15 言語の含意認識タスクの事例を含む XNLI データセットと英語の MNLI データセットである．ただし，MNLI データセットは機械翻訳の品質上の問題から英語の訓練データのみを用いている．前提文と仮説文内の「相対的形容詞」は統計指標毎に事前に人手で付与した．仮説文の「補助テキスト」も同様に人手で付与した．

補助情報ありニューラルトピックモデルのパラメータチューニングでは，学習データに含まれる各文書内で検証用 (20%) に配分された単語を用いて，パープレキシティが最も低くなる組み合わせを探索し，最適なパラメータを決定した．

統計指標検索タスクの評価を行うための前段階として，ランキングの出力までの手順について述べる．まず，手法に入力する単語（以下,「クエリ」と呼ぶ）を用意するため，今回は市区町村の潜在指標を表すような単語として，社会課題に関連する単語（例: 魅力，過疎，治安）を 50 個選んだ．ただし，クエリとなる単語は事前に定義した語彙集合に含まれるものとする．次に，ベースライン手法も含めた 10 手法を用いて，クエリごとに上位 10 件のランキングを出力した．なお，相互情報量を用いた手法では，トピック分布の関数 $f_t(\cdot)$ において，3 種類の方法 (線形変

[*3] https://www.e-stat.go.jp/koumoku/koumoku_teigi/A
[*4] https://huggingface.co/datasets/mc4
[*5] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli

換 $f_t^{\text{vec}}(\cdot)$，CNN $f_t^{\text{cnn}}(\cdot)$，RNN $f_t^{\text{rnn}}(\cdot)$) があり，入力特徴は単年（または複数年）の統計指標ベクトル（1 年分，3 年分）があるため，それぞれでランキングを出力して，検索性能に違いが生じるかを調査することとする．また，入力特徴では，対象年（事例に含まれる西暦）の統計指標ベクトルと 10 年前の統計指標ベクトルの差をとったもの（差分ベクトル）も調査する．トピック分布の関数 $f_t(\cdot)$ の有効性を調査するため，関数内で全く処理を行わず，入力ベクトルをそのまま出力して，統計指標ベクトルをもとに単語分布に変換した場合（以下「処理なし」）も調査する．

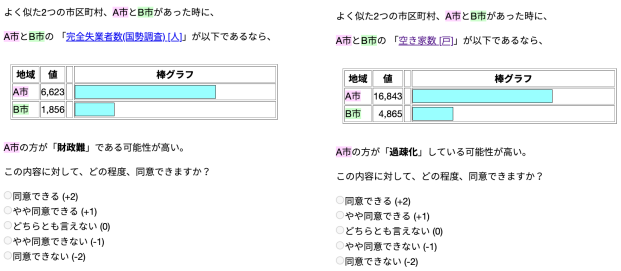
4.2.1 統計指標検索タスクの評価方法

本研究では，クラウドソーシングを用いて，ランキングに含まれている統計指標がクエリの潜在指標を推測するのに適しているかを人手で判断し正解データを作成した．評価方法は，客観性を持たせた評価に基づき，世間一般の人々が同意できるか否かで評価する．本タスクでは，統計指標の数量の大小度合をもとに潜在指標の程度を推測することに主眼を置いているため，潜在指標と擬似相関がある統計指標であっても一般の人々が同意できる場合は適合と判断する．疑似相関である場合，統計指標を主体的に動かしても潜在指標が追従しないため，地域の観光客数を少なくしても魅力が下がるわけではない．そこで，評価では，地域の特定の統計指標の数量が増加・減少した状況ではなく，ある時期の数量をもとに潜在指標の程度を推測できるかを判断する．

クラウドソーシングサービスとしてランサーズ[*6]を使用した．ワーカーは，図 1(a)(b) に示すような質問に，5 段階のリッカート尺度で回答した．質問の内容は，与えられた統計指標名のみでクエリの潜在指標の程度を推測するのに適しているかをワーカーが判断するのは困難であるため，よく似た 2 つの市区町村 (A 市と B 市) を仮定し，与えられた統計指標の値のみが 平均値 ± 標準偏差 ほど差がある状況を想定する．そして，与えられた統計指標の値を棒グラフを示し，A 市の方が潜在指標が高くなるかを聞くことで，ワーカーは客観的な比較に基づき判定が行えるようにした．タスクには 10 問の質問を掲載し，そのうちの 1 問はワーカーの回答に一貫性があるのかを確認するための質問とし，残りの 9 問の回答を評価の際に使用した．なお，本研究では，この回答を「推測的スコア」と呼ぶ．なお，タスクの説明文の中には，ワーカーの回答を拒否する条件として，(1) 適当に入力している場合，(2) 比較をせずに判断している場合，(3) 評価の一貫性がない場合，があることを明記した．ワーカーの作業品質を保つため，2 つ以上のタスクで拒否条件を満たす回答をしたワーカーのタスクは承認せず，別のワーカーに再度依頼した．

評価指標としては，情報検索においてよく用いられる P@10（Precision@10）と nDCG@10 [1] を用いた．なお，P@10 と nDCG@10 は負のスコアが扱えないため，クラウドソーシングで集めた正解データに含まれるマイナスのスコアは 0 で置換し，3 段階の推測的スコアで計算を行なった．

[*6] https://www.lancers.jp/



(a) 財政難に関する質問　　**(b)** 過疎化に関する質問

図 1: クラウドソーシングで依頼したタスクの質問例

4.3 実験結果

4.3.1 統計指標検索タスクの実験結果

統計指標検索タスクにおける 10 手法の P@10 と nDCG@10 の結果を表 1 に示す．なお，P@10 と nDCG@10 の値は，50 個のクエリの平均値とする．本研究で提案した「含意認識手法」は，P@10 と nDCG@10 の両方でベースライン手法やもう一つの提案手法「相互情報量手法」を上回る性能を示しており，本タスクで最も有効であったことが明らかになった．これは，含意認識モデルが，+1 以上の推測的スコアが付与された統計指標に対して高い含意確率を出力できたことを示唆しており，含意認識モデルが大きい・小さいなどの相対的形容詞で表現された統計指標の数量度合と潜在指標の程度の関係を内容的に近いものとして捉えられているからだと考えられる．相互情報量手法では，トピック分布の関数 $f_t(\cdot)$ を 7 種類の場合に分けて検索性能を評価したが，他の手法よりも低い性能を示した．特に，P@10 による比較では，トピック分布の関数 $f_t(\cdot)$ の 7 種類の中で，「処理なし（1 年分）」がその他の場合よりも性能が高く，トピックを介した手法の方が劣る結果となった．また，nDCG@10 による比較では，3 年分（10 年前と 5 年前と対象年）の統計指標ベクトルを扱ったトピック分布の関数 $f_t(\cdot)$ である「線形変換（3 年分）」「CNN（3 年分）」が 1 年分の統計指標ベクトルを扱ったものより高い性能を示した．つまり，潜在指標を表すような単語をクエリとするランキングにおいて，複数年の統計指標ベクトルを扱ったモデルの方が高い性能となることが判明した．ベースライン手法の「共起回数手法」は，「含意認識手法」よりも低い性能であるが，「相互情報量手法」よりは高い性能を示した．これは，補助情報ありニューラルトピックモデルが統計指標と潜在指標を表すような単語との対応関係を捉えきれず，推測的スコアが高くなる潜在指標を検索できていないことを示唆している．この理由として，モデルの入力単語列が Bag-of-Words であるために，潜在指標を表す単語と程度を表す単語は個別に扱われ，統計指標の値によって潜在指標の程度が決まることを学習できていないことが挙げられる．

50 個のクエリにおける 10 手法の性能差の検定として，Two-way ANOVA を行い，手法の効果は統計的に有意であることが明らかになった．なお, nDCG@10 の場合は, $F(9, 441) = 16.5, p < 0.05$，P@10 の場合は $F(9, 441) = 12.5, p < 0.05$) となった．また，Tukey HSD を実施し，有意水準 $\alpha = 0.05$ の下では，「含意認識手法と他の手法の間」および「共起回数手法と相互情報量手法

表 1: 統計指標検索タスクにおける各手法の検索性能

	手法		Precision@10	nDCG@10
ベースライン手法		ランダム手法	0.326	0.221
		共起回数手法	0.394	0.290
提案手法		含意認識手法	**0.594**	**0.476**
	トピック分布の関数 $f_t(\cdot)$ の方法と入力特徴	相互情報量手法		
		線形変換（1 年分）	0.282	0.186
		線形変換（3 年分）	0.368	0.261
		線形変換（差分）	0.366	0.249
		CNN（3 年分）	0.344	0.256
		RNN（3 年分）	0.316	0.210
		処理なし（1 年分）	0.372	0.254
		処理なし（差分）	0.336	0.221

（線形変換 1 年分）の間」に有意差が見られた（p 値が 0.05 より小さい）．この結果は両方の評価指標で同じとなった．

4.3.2 含意認識手法の失敗分析

表 1 より，含意認識手法の Precision@10 は 0.594 であるため，上位 10 件のうち約 4 件は不適合な統計指標（不適合指標）であることが判明した．クエリに対する適合指標が 10 件未満しか存在しない場合，現状のランキングではユーザに多くの不適合指標を提示してしまい，誤解を招く可能性がある．ユーザの誤解を招く事態を防ぐため，閾値を設定し，最低限の条件を満たした統計指標のみを提示することを検討する．本研究では，含意確率よりも中立確率（または矛盾確率）の方が高くなった統計指標が上位 10 件に含まれていたことで手法の性能が低下したかを検証する．具体的な方法としては，個々のクエリに対して手法が出力したランキング内の不適合指標における矛盾確率と中立確率より作成したヒストグラムで傾向を捉える．その後，手法の P@10 を上げるために，偽陽性（FP: False Positive）数を抑えつつ，真陽性（TP: True Positive）数が僅かしか減らないような閾値を発見する．

まず，50 個のクエリに不適合な統計指標に対して，手法が出力した中立確率に関するヒストグラムを図 2 に示す．この結果より，手法が出力した上位 10 件内の不適合指標のうち，中立確率が 0.1 未満であったものが約 28% であり，残りの約 72% は中立確率が 0.1 以上であることが判明した．つまり，不適合と判定された統計指標の約 72% は含意確率が 0.9 未満であるため，含意確率が 0.9 を切るような統計指標は潜在指標の程度を推測するのに適さないことが示唆された．次に，50 個のクエリに不適合な統計指標に対して，手法が出力した矛盾確率に関するヒストグラムを図 3 に示す．この結果より，手法が出力した上位 10 件内の不適合指標のうち，矛盾確率が 0.1 未満であったものが約 81% であり，0.2 未満までを含めれると約 91% になることが判明した．つまり，不適合と判定された統計指標の約 91% は，矛盾確率より含意確率（中立確率）の方が高くなっており，矛盾確率が低くとも不適合指標が存在することが示唆された．

先程の結果を踏まえて，手法の P@10 が上がるとされる，FP 数を抑えつつ，TP 数が僅かしか減らないような閾値を発見するため，中立確率を変化させながら，その時の Precision を計算した結果を図 4 に示す．なお，P@10 は，「中立確率を閾値以下」または「矛盾確率を 0.2 以下」を検索条件として加えた際のスコアとし，矛盾確率は 0.2 以下で固定する．その結果，手法は，「中立確率が 0.1 以下」または「矛盾確率が 0.2 以下」となる統計指標のみ出力することで，P@10 を 0.837 まで向上できることが判明した．これにより，上位 10 件に含まれる不適合指標を約 4 件から約 2 件まで減らすことができ，不適合指標の数量度合に基づき潜在指標の程度を推測できるとユーザが誤解することも防ぐことができると考えられる．

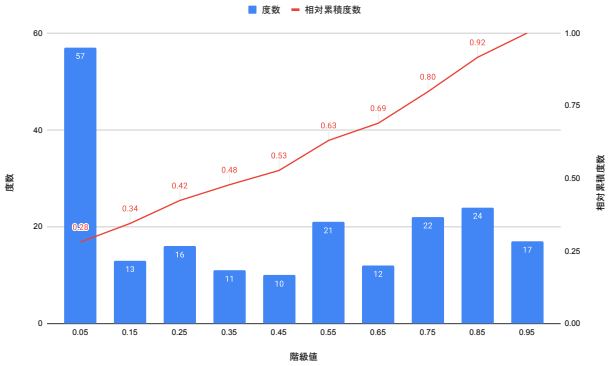


図 2: 不適合な統計指標に対して含意認識手法が出力した中立確率に関するヒストグラム

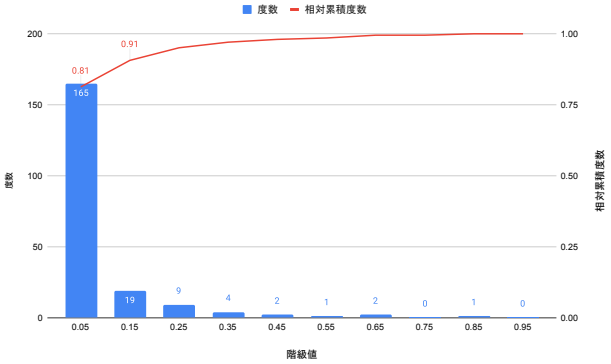


図 3: 不適合な統計指標に対して含意認識手法が出力した矛盾確率に関するヒストグラム

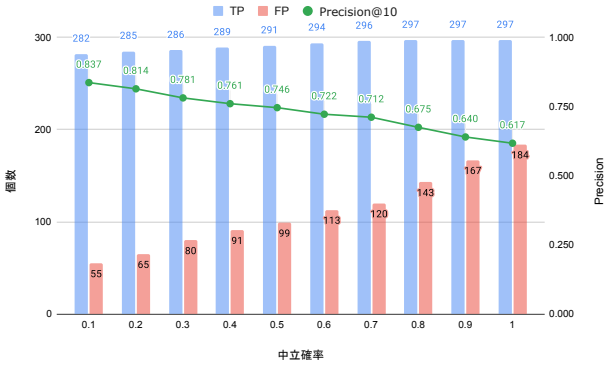


図 4: 「中立確率を閾値以下 」または「矛盾確率を 0.2 以下」を検索条件として加えた際の TP・FP・P@10

表 2: 「魅力」のランキング結果

順位		含意認識手法		相互情報量手法（線形変換 3 年分）		相互情報量手法（CNN 3 年分）		相互情報量手法（処理なし 1 年分）
1	+	付加価値額 (民営) (宿泊業、飲食サービス業)	+	従業者数 (民営) (情報通信業)	+	卸売業年間 商品販売額	−	40〜44 歳人口 (男)
2	+	付加価値額 (民営) (鉱業、採石業、砂利採取業)	+	従業も通学も していない人口	−	無店舗小売店数	−	事業所数 (市区町村)
3	+	付加価値額 (民営) (農林漁業)	−	近隣商業地域面積	−	その他の小売店数	−	売上金額 (民営) (学術研究、 専門・技術サービス業)
4	+	付加価値額 (民営) (電気・ガス・熱供給・水道業)	−	従業者数 (民営) (金融業、保険業)	−	機械器具小売店数	+	事業所数 (民営)
5	+	付加価値額 (民営) (医療、福祉)	−	住居地域面積	−	飲食料品小売店数	+	日本人人口 (総数)

4.3.3 提案手法のランキング結果

提案手法のランキング結果において，統計指標が細分化されていることで，同じような統計指標が上位に順位付けられてしまうという事例が見られた．例えば，表 2 の「付加価値額」のような似た名称をもつ統計指標は，クエリと意味的に類似するため，上位に順位付けられたと考えられる．含意認識手法は，統計指標の名称が十分な情報を含んでいない場合，単語同士が意味的に類似する統計指標，つまり，内容的につながりが弱い統計指標の方を上位に順位付けしてしまう可能性がある．それに対処するため，付加価値額のような十分な情報を含んでいない統計指標を扱う場合，（より多くの情報を含んだ）説明文を用いて含意認識が行えるよう改善すべきだと考えられる．その一方で，相互情報量手法は，トピック分布の関数が異なるとランキング結果も大きく変化しており，共通する統計指標が少ないことが判明した．本来であれば，予測において大きく貢献している統計指標は，一定程度共通すると考えられるが，結果は相反するものであった．その要因として，本研究の実験では，テキスト中に出現した年表現と市区町村名をもとに統計指標ベクトルを用意するアプローチを採用したため，年表現の制約が厳しすぎた可能性がある．

5 まとめ

本論文では，魅力や活気のようなエンティティの潜在指標を直接入力して，潜在指標の程度が推測できるような統計指標を検索する統計指標検索タスクと，それに対する手法を提案した．手法の性能評価では，クラウドソーシングを用いて判定を行い，2 つの評価指標（P@10 と nDCG@10）でベースライン手法と提案手法の比較を行った．その結果，含意認識手法が両方の評価指標で最も高い性能を示した．本研究の限界として，複数の統計指標を組み合わせて潜在指標の程度を説明することはできない．今後の課題は，統計指標のデータを利用した含意認識手法の検討，相互情報量手法の性能改善などを挙げられる．

謝辞

本研究は JSPS 科研費 21H03554，21H03775 の助成を受けたものです．ここに記して謝意を表します．

参考文献

[1] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 89–96, New York, NY, USA, 2005. Association for Computing Machinery.

[2] Dallas Card, Chenhao Tan, and Noah A. Smith. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[3] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1041–1048, Madison, WI, USA, 2011. Omnipress.

[4] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021.

[5] Tatsuya Iwanari, Naoki Yoshinaga, Nobuhiro Kaji, Toshiharu Nishina, Masashi Toyoda, and Masaru Kitsuregawa. Ordering concepts based on common attribute intensity. In *IJCAI*, pages 3747–3753, 2016.

[6] Makoto P. Kato, Wiradee Imrattanatrai, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka. Context-guided learning to rank entities. In *ECIR*, pages 83–96, 2020.

[7] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 13–18 Jul 2020.

[8] Yuliang Li, Aaron Feng, Jinfeng Li, Saran Mumick, Alon Halevy, Vivian Li, and Wang-Chiew Tan. Subjective databases. *Proc. VLDB Endow.*, 12(11):1330–1343, jul 2019.

[9] Diego Marcos, Ruth Fong, Sylvain Lobry, Remi Flamary, Nicolas Courty, and Devis Tuia. Contextual semantic interpretability. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[10] Jon Mcauliffe and David Blei. Supervised topic models. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[11] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2410–2419. JMLR.org, 2017.

[12] Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1727–1736. JMLR.org, 2016.

[13] Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. SELFEXPLAIN: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[14] Shinryo Uchida, Takehiro Yamamoto, Makoto P. Kato, Hiroaki Ohshima, and Katsumi Tanaka. Entity ranking by learning and inferring pairwise preferences from user reviews. In *AIRS*, pages 141–153, 2017.

[15] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[16] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. *ArXiv*, abs/2211.11158, 2022.

[17] Lanbo Zhang, Yi Zhang, and Yunfei Chen. Summarizing highly structured documents for effective search interaction. SIGIR '12, page 145–154, New York, NY, USA, 2012. Association for Computing Machinery.

[18] 山枡慧. "法的措置 発言の波紋 都道府県魅力度ランキング". NHK 政治マガジン (NHK NEWS WEB), 2021. https://www.nhk.or.jp/politics/articles/feature/72908.html, (参照 2022-07-14).

[19] 秋山度 and 島田尚朗. "信じてますか no.1". NHK NEWS WEB, 2022. https://www3.nhk.or.jp/news/html/20220517/k10013629911000.html, (参照 2022-07-15).