健康情報検索におけるクエリに含まれる暗黙的な仮定の検出

今坂　優太†　　山本　岳洋†,††　　大島　裕明†,††　　加藤　　誠†††　　藤田　澄男††††

† 兵庫県立大学 社会情報科学部　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 兵庫県立大学 大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
††† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
†††† ヤフー株式会社　〒 102–8282 東京都千代田区紀尾井町 1–3
E-mail: †fa19l011@stsis.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp, †††ohshima@ai.u-hyogo.ac.jp,
††††mpkato@acm.org, †††††sufujita@yahoo-corp.jp

あらまし　本研究では，健康情報を検索する際に置かれることのある仮定に注目した．仮定とは検索者が持つ「ある物質や治療法は，ある疾病に対して効果的である」のような考えを指す．健康情報検索における仮定の種類は【効果あり，効果なし，検証】の 3 種類が存在すると考え，クエリからの検出を目指す．本研究では，仮定の種類は一意に定まるものではないと考え【効果あり，効果なし，検証】の各割合を表す分布としての予測を行う．まず，予測モデルの構築に必要な学習データを得るために，クラウドソーシングを用いて仮定が置かれたクエリを収集する．次に，得られたクエリを用いて BERT とロジスティック回帰の 2 種類のモデルを構築し，分布の予測モデルを構築する．この際に，クエリ単体で学習した場合とクエリに加えて検索結果も追加して学習した場合の 2 通りの学習方法を用意した．構築したモデルの有効性を確かめるため，クエリから仮定の分布を予測する評価実験を行った．521 件のクエリをテストデータとして実験を行った結果，クエリ単体で学習した BERT が最も良い精度を示した．

キーワード　情報検索，Web 検索，健康情報，クエリ分析

1　は じ め に

Web へのアクセスが容易になり，Web 上の健康情報を信用できると感じる人が増加している．平成 26 年版厚生労働白書によると，Web を通じて健康情報を得ていると回答する人の割合は高くなっている．また，Web 上の健康情報に対する受け取り方も変容してきており,「Web 上の健康情報を信用できる」と回答した人の割合は 2009 年現在で 43%だったのに対して，2014 年現在で 56%にまで上昇している [18].

Web からの健康情報の入手が拡大する中で，その情報の質については様々である．令和 3 年版情報通信白書によると，検索エンジンに対する信頼度が「信頼できる」「半々くらい」を合計して 85%を超えている一方で，検索エンジンは偽情報を取得したメディアの 1 つとしても挙げられている．テレビやラジオ，新聞といった媒体では，情報に対する裏付けが十分に行われて発信されるケースが多いが，Web においては必ずしもそうとは限らない．

誤った情報を獲得しないためには，複数の情報源を参照することによる情報の精査が必要である．得た情報が正しいものなのか，一方で誤ったものなのかを多様なページから判断することは，特に健康情報を検索する際においては重要である．しかし，検索者が前提となる考えに基づいて検索を行っている場合，こうした態度は中々喚起されないと考えられる．例えば「シナモンは糖尿病に効果がある」という考えを持って「シナモンと糖尿病」について検索を行う場合である．White の研究 [10] によれば，この検索者は検索前に持つ「シナモンは糖尿病に効果がある」という考えに引っ張られてしまい「効果がある」と主張する情報を中心に，偏って閲覧する傾向があることが分かっている．本当に効果的かどうかを確かめるためには,「効果的でない」と主張する文書も参照しようとする検証態度が重要である．このような検証態度は検索者自身が自発的に起こすことは困難であると考えられ，検索者に対するシステム的な支援が必要である．

本研究では前提となる考えに基づいて行われる検索の背後に存在する考えを暗黙的な仮定と定義した上で，クエリやその検索結果から検出することを目指す．検出が可能になれば，検索者を支援する仕組みにつなげることが可能である．例えば，検索者が置いていると考えられる仮定について検証する Web ページを検索結果に含めたり，仮定の真偽を検証を促すようなシステムが実現できる．本研究では，クエリからの暗黙的な仮定検出の可否について，次のようなリサーチクエスチョン（RQ）を設定する．

RQ:　クエリからの暗黙的な仮定の検出は，どの程度の精度で可能なのか．

本研究では，健康情報検索における仮定の種類が【効果あり，効果なし，検証】の 3 種類しか存在しないと考える．さらに，仮定の種類は一意に定まらないと考え【効果あり，効果なし，検証】の各割合を表す分布として予測を行う．分布としての予測を行うモデルを構築することで，暗黙的な仮定の検出を目指す．まず，予測モデルの構築に必要な学習データを得るために，クラウドソーシングを用いて仮定が置かれたクエリを収集する実験を行う．実験では，験参加者に仮定が置かれた検索を

行ってもらうため，治療法と疾病の関係について述べた事前情報を提示する．この事前情報は「ある物質や治療法は，ある疾病に対して効果的である」と主張する情報で，前提となる考えを持ってもらうために提示する．提示後に実験参加者には提示した情報と似た情報を得るためのクエリを作成してもらう．得られたクエリを用いて，BERT とロジスティック回帰の 2 種類のモデルを構築する．この際に，クエリ単体で学習を行う場合と，クエリに追加してクエリの検索結果も用いて学習を行う場合の 2 通りの学習方法を用意した．

構築したモデルの有効性を確かめるために，クエリから仮定の分布予測を行う評価実験を行った．521 件のクエリをテストデータとし，RSS（Rooted Residual Sum of Squares）と JSD（Jensen-Shannon Divergence）の 2 指標を用いてモデルの評価を行なった．その結果，クエリ単体で学習した BERT が最も良い精度を示した．

2 関 連 研 究

2.1 Web 検索行動の分析

情報検索時のユーザの Web 検索行動を明らかにする研究は盛んに行われている [7] [15] [19]．Yamamoto らは，信頼できる情報獲得のために必要な行動に対する心がけを検証態度と定義し，検証態度と実際の Web 検索行動の関係について分析を行っている [12]．検証態度の高い検索者は,「証拠」や「本当」といった事物を検証するために用いられる語含んだクエリを投入し，慎重に検索を行っていることなどが明らかになった．検証態度が高い検索者の特徴がわかった一方で，検証態度の低い検索者をいかにして検証態度の高い検索者にするかについては，議論の余地があるとされている．

こうした研究の中でも特に，情報検索時の事前の知識や考えが検索行動に与える影響について明らかにする研究も行われてきている [20] [1]．White は検索者の事前の考え，そして無意識的なバイアスが Web 検索における意思決定に影響を及ぼすことに注目し，初めて大規模な調査を行った．この調査の結果，検索者は検索行動の結果として事前の考えを変えにくいこと，自身にとって都合の良い情報ばかりを見る傾向にあることを明らかにした．また，検索システム側についてもバイアスの調査を行っており，yes-no 型の問いを入力したときに yes 側の意見が結果として多く表示されることを明らかにした．その結果として検索者は誤った決定を行ってしまう傾向があることを明らかにし，検索者，検索システムの両サイドで発生しているバイアスを改善する取り組みの必要性を主張した [10] [11]．

Pothirattanachaikul らは，文書の意見と信憑性が検索行動および信念の変化に与える影響について分析を行い，2 つのことを明らかにした [8]．1 つ目が，検索前の考えと相反する文書を閲覧した際には検索に労力をかけ，逆の意見に流されやすいこと，2 つ目が，検索前の考えと同じ文書を閲覧した時に，検証を行って別の意見に変えるのではなく，元の意見を保持しやすいことである．これらの結果から，検索者の事前の考えとは相反するような文書を検索結果上位に表示することで慎重な Web 検索を促進することの必要性を主張した．

2.2 Web 検索行動を支援する取り組み

2.1 節で述べたような問題に対するシステム的な検索行動支援のアプローチとして，ランキングを改善する試みがなされている．Zhang らは，健康情報を正しさと信頼性に基づいてランキングする手法を提案した．Web の健康情報が正しいのかを 2 値分類するモデル（Stance Detection Model）と，発信元ドメインが信頼できるかどうか（Trust Model）の 2 つのモデルからなるランキング手法で，TREC 2021 Health Misinformation Track [3] における Compatibility-difference スコア [4] [2] において 0.129 を達成した [16]．このスコアは，TREC 2021 Health Misinformation Track における人手による評価を要しないランキングとしては最も高いものだった．

Yamamoto らは批判的情報検索を促すことで，検索者の検証態度を喚起する技術として，クエリプライミング [13] を提案した．この研究では，先行する刺激が後の行動に影響を与えるプライミング効果に着目した.「検証」や「調査」といったプライム語をクエリ補完，推薦することで，検索者の検証態度を喚起することに成功している．

このように，Web 検索行動を支援する取り組みが多数行われている一方で，検索の背後に存在する仮定を検出する試みはなされてこなかった．事前の考えに基づいた検索を検出することができれば，2.1 節で述べたような，検索者に発生する問題を改善することができると考える．本研究では特に，健康情報検索の分野に絞り，かつ検索前に持つ考えを 3 種類に定めた上で，クエリから検出することを目指した．

3 暗黙的な仮定の定義とデータ収集実験

本節ではまず，暗黙的な仮定の考え方とその定義，データ収集実験について述べる．本実験は，4 節で述べる暗黙的な仮定の検出モデル構築に必要な正解データを収集するために実施する．実験参加者に仮定が置かれた検索を行ってもらうため，治療法と疾病の関係について述べた事前情報を提示する．この事前情報は,「ある物質や治療法は，ある疾病に対して効果的である」と主張する効果あり情報,「ある物質や治療法は，ある疾病に対して効果的でない」，または「ある物質や治療法と疾病の関係は検証する必要がある」と主張する効果なし情報からなる．これらの情報を実験参加者に提示し，仮定を含んだクエリの作成を行ってもらった．

3.1 暗黙的な仮定とその種類

健康情報を検索する際に，仮定を置いて検索する場合がある．例えば,「糖尿病 シナモンサプリ 通販」のようなクエリで検索を行う場合である．このようなクエリの背後には,「シナモンは糖尿病に効果的である」のような考えが存在すると考えられる．本研究では，このようなクエリの背後に存在する考えを暗黙的な仮定と定義する．本研究では，健康情報検索における仮定は次の 3 種類しか存在しないと定義する．

図 1　実験の手順.

（a）　ある物質や治療法は，ある疾病に対して効果的である.
（b）　ある物質や治療法は，ある疾病に対して効果的でない.
（c）　ある物質や治療法と疾病の関係は検証する必要がある.

トピック「シナモンと糖尿病の関係」を例に考える．このトピックに対する仮定は上の定義に従うと (a) の場合は「シナモンは糖尿病に効果的である」,（b）の場合は「シナモンは糖尿病に効果的でない」,（c）の場合は「シナモンと糖尿病の関係は分からないため検証する必要がある」である.

3.2　実験の実施手順

クラウドソーシングで実施したデータ収集実験の手順（図 1）について説明する．本データ収集実験において，実験参加者は実際には検索を行わずにクエリの作成のみを行う，擬似的な検索タスク 3 種類に取り組んでもらう．その際に得られるクエリを集めることを目的としている．収集したいクエリは 3 種類あり， 3.1 節で述べた（a）〜（c）のクエリである．実験参加者が検索時に持つ仮定の種類に応じて得られるクエリの暗黙的な仮定も異なる．下記に示す実験手順は，実験参加者に「シナモンと糖尿病」トピックに関する効果あり情報を与え,（a）,（c）の仮定を含んだクエリを得ることを目的としたものである.

まず，実験参加者には実験概要，収集するデータとその利用目的について説明した．その後，実験内容が学術研究に利用されることに関する説明文と，タスク実施に際しての報酬に関する説明文を提示した．また,「実験の途中辞退は可能だが，その際には報酬を支払わない可能性がある」ことを予め説明した．これらについて同意した者のみが実験タスクに進んだ.

（1）　まず実験参加者にはタスクに関する背景及びタスクの内容に関する説明を提示した．その後，実験参加者には「シナモンは糖尿病に効果的である」と主張する Web ページ（事前に実験者が選定したもの）を 3 件閲覧してもらう．例えば，表 1 に示したタスク 1 に関して効果あり Web ページの提示を行う際には，次に示す説明文を提示した.

> あなたの知人に糖尿病を患っている方がいると仮定し，あなたはその方の症状を改善したいと考えています.「糖尿病を改善するために効果的なものは何か」についてインターネットで調べていたところ，以下に示す Web ページを見つけました.

これらの説明文とともに，表 1 で示す効果あり文章を提示した.

（2）　（1）で Web ページを閲覧してもらった後，1 つの質問文を提示した．この質問文では，タスクとして与えられた治療法と疾病の関係について「効果的と言えるかどうか」を問うた．この質問に対して，4 段階のリッカート尺度（1: 効果的だと思う，2：どちらかと言えば効果的だと思う，3：どちらかと言えば効果的だと思わない，4：効果的だと思わない）で回答してもらった.

（3）　（a）のクエリを収集することを目的として，実験参加者にクエリの作成タスクに取り組んでもらった．次に示す説明文は，そのタスクに取り組んでもらうため提示したものである.

> これらの Web ページを見たあなたは「シナモンは糖尿病に効果的であるため，知人に勧めたい」と考えました．知人にどのように勧めたら良いかを検討したいあなたは，さらに多くの情報を探そうとしています．その際に考えられる検索キーワードを 3 種類作成してください.

（4）　実験参加者にタスク（1）で提示した Web ページと相反する主張の Web ページを提示した．提示後，タスク（2）〜（3）と同様に Web ページの主張を問う質問の提示と,（c）のクエリの作成を行ってもらった．その際に，次に示す説明文を提示した.

> 情報収集を行う中であなたは，ウェブページ：厚生労働省「「統合医療」に係る 情報発信等推進事業」 シナモンを発見しました.
> 〜〜〜〜〜〜〜〜〜〜〜〜質問文〜〜〜〜〜〜〜〜〜〜〜〜
> 「シナモンと糖尿病の関係性は不明」とする考えもあることを知ったあなたは,「シナモンが糖尿病に効果的かどうか」についてより深く調べる必要があると考えました．その際に考えられる検索キーワードを 3 種類作成してください.

（5）　手順（1）〜（4）を以下に示す 2 つのトピックでも実施してもらった.

- がんとお茶の関係
- 喘息とカフェインの関係

（6）　実験参加者のデモグラフィック属性，学歴，検索専門性，e ヘルスリテラシー [6] に関するアンケートに回答してもらい，実験を終了した．なお，検索専門性のアンケートは，Yamamoto らの研究 [14] を参考にし，e ヘルスリテラシーに関するアンケートは Mitsutake らの研究 [17] を参考に作成した.

上記の実験を,（1）の段階で効果あり情報を付与する場合と，効果なし情報を付与する場合とで 2 種類のタスクを用意し，1 つのタスクあたり 150 人に回答してもらった．実験参加者は 1 つのタスクに 1 回しか実施できず，異なるもう一方のタスクの実施は認めないこととした.

3.3　実施方法と実験対象

兵庫県立大学大学院情報科学研究科倫理審査委員会の承認（承認番号：UHIS-EC-2022-005）を得て，データ収集実験を実施した．2022 年 12 月 2 日から 2022 年 12 月 6 日にかけて，ク

表 1　実験で用いたタスク文と情報源.

検索タスク		情報源	
ID	トピック	効果あり文章	効果なし文章
1	シナモンは糖尿病の改善に効果的か.	https://bit.ly/3AwGXW8（食品会社）	https://bit.ly/3Doufdy（厚生労働省：eJIM）
2	お茶はがんの改善に効果的か.	https://bit.ly/3VabCAI（食品会社）	https://bit.ly/3RdtCsQ（厚生労働省：eJIM）
3	カフェインは喘息の改善に効果的か.	https://bit.ly/3ERenBm（厚生労働省：eJIM）	https://bit.ly/3OojRa1（医療機関）

表 2　収集データに対する後処理を終えた後のクエリ件数.

トピック	(a) 効果あり	(b) 効果なし	(c) 検証	合計
糖尿病とシナモン	411	411	558	1,380
がんとお茶	399	384	486	1,269
喘息とカフェイン	324	408	519	1,251

表 3 「糖尿病とシナモン」トピックで得られたクエリの例と分布.

クエリ	(a) 効果あり	(b) 効果なし	(c) 検証
糖尿病 効果 シナモン (N=87)	0.39	0.37	0.24
血糖値 シナモン (N=27)	0.63	0.19	0.19
糖尿病改善 シナモン (N=10)	0.80	0.10	0.10

表 4 「がんとお茶」トピックで得られたクエリの例と分布.

クエリ	(a) 効果あり	(b) 効果なし	(c) 検証
緑茶 効果 がん (N=20)	0.40	0.40	0.20
がん予防 お茶 (N=16)	0.69	0.25	0.06
緑茶 がん予防 (N=15)	0.47	0.20	0.33

ラウドソーシングプラットフォームの 1 つである Lancers[1] を用いて実験参加者を募集し，300 名のワーカが実験に参加した．実験時間として 20 分程度を想定した本実験を最後まで完了した 300 名の実験参加者に対し，1 人あたり 385 円を支払った．また，300 名の実験参加者から得られたクエリから，次のような回答を除いた．

（1） 入力フォームへの回答形式が誤っていたもの
（2） 回答の途中から空欄になっていたもの

以上の操作に加えて，実験の意図通りに意見が変化しなかったワーカのクエリも除いた．具体的には実験手順 3.2 節において，次の条件に当てはまるワーカのクエリである．

（3） 実験手順 3.2 節（1）の効果ありと主張する Web ページを提示した際の質問で「効果的だと思う」あるいは「どちらかと言えば効果的だと思う」と回答しなかったワーカ

（4） 実験手順 3.2 節（4）で効果なしまたは効果の有無が分からないと主張する Web ページを提示した際の質問で,「効果的だと思わない」あるいは「どちらかと言えば効果的だと思わない」と回答しなかったワーカ

これらの操作をトピックごとに分割して実施した．最終的に得られたクエリデータの件数を表 2 に示す．

3.4　収集したデータの概要

実験によって得られたデータについて述べる．表 3，表 4，表 5 は，実験で得られたクエリと件数の割合を表した分布の一例である，各割合は小数点第 2 位で四捨五入しているため，必ずしも割合の総和が 1 にならないことに注意されたい．表 3 のクエリ（糖尿病改善 シナモン）を例に説明する．このクエリは合計 10 件得られ，その分布から 3.1 節で述べた定義に基づき次のように考える．

表 5 「喘息とカフェイン」トピックで得られたクエリの例と分布.

クエリ	(a) 効果あり	(b) 効果なし	(c) 検証
喘息 効果 カフェイン (N=54)	0.46	0.30	0.24
喘息 カフェイン (N=42)	0.52	0.38	0.10
喘息 コーヒー (N=13)	0.46	0.46	0.08

（a） 80%の確率で，効果的である仮定を持つ
（b） 10%の確率で，効果的でない仮定を持つ.
（c） 10%の確率で，糖尿病とシナモンの関係を検証している.

3.5　予測精度向上のためのデータ

4 節で述べるモデルの学習データには，実験で得られるクエリの検索結果も用いることとした．クエリ単体に加えて検索結果も用いた方が，予測精度の向上に寄与すると考えられるためである．Yue らの研究 [15] によると，検索者は検索結果の上位に掲載された情報，特に上位 3 件までを閲覧する傾向が強いことを明らかになっている．よって本研究では，検索結果上位 3 件までのタイトルを予測に用いることにした．この際に，検索結果中に含まれる，中括弧やカンマ，ピリオドなどの記号は除去する処理を行った．なお，検索結果の取得には，Bing Web Search API[2] を用いた．

4　暗黙的な仮定の検出

4.1　問 題 定 義

本研究では 3.1 節で定義した暗黙的な仮定をクエリから検出することを目指す．実験で得られるクエリの集合を Q とする．なお，N は全データの件数を表す.

$$Q = \{q^{(1)}, q^{(2)}, \cdots, q^{(i)}, \cdots, q^{(N)}\}$$

3.1 節で定義したように，クエリ $q^{(i)}$ には 3 種類の仮定が存在する．この仮定を label(l) と表し，次のように定義する.

$$\text{label}(l) = \begin{cases} l_1 & (\text{効果あり}) \\ l_2 & (\text{効果なし}) \\ l_3 & (\text{検証}) \end{cases}$$

1：https://lancers.jp

2：https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

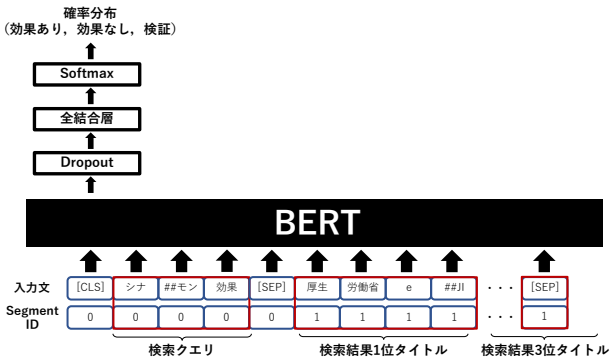

図 2　検出モデルとして利用する BERT モデル図.

実験で得られるクエリは，表 3 のように分布を持つ．この時，クエリ $q^{(i)}$ に対する label の確率を次のように表す.

$$P\left(q^{(i)}\right) = \left(P_{l_1}\left(q^{(i)}\right), P_{l_2}\left(q^{(i)}\right), P_{l_3}\left(q^{(i)}\right)\right) \tag{1}$$

この $P(q^{(i)})$ の分布を予測するのが本研究の目的であり，ラベル分布予測と呼ぶ．ここで，4.2 節で述べる検出モデルの予測値 $\hat{P}(q^{(i)})$ と表す.

$$\hat{P}(q^{(i)}) = \left(\hat{P}_{l_1}\left(q^{(i)}\right), \hat{P}_{l_2}\left(q^{(i)}\right), \hat{P}_{l_3}\left(q^{(i)}\right)\right) \tag{2}$$

表 3 のクエリ（糖尿病改善 シナモン）を例に考える．次の等式が成り立つような予測ができれば，誤差のない予測が行えたと考える.

$$\hat{P}_{l_1}(\text{糖尿病改善シナモン}) = 0.80$$
$$\hat{P}_{l_2}(\text{糖尿病改善シナモン}) = 0.10$$
$$\hat{P}_{l_3}(\text{糖尿病改善シナモン}) = 0.10$$

4.2　暗黙的仮定の検出モデル

本研究では，暗黙的仮定の検出モデルとして BERT [5] を用いる．BERT は Transformer で構成されており，自然言語に対して文脈を考慮した分散表現を獲得できるという特徴がある．BERT は質問応答，固有表現抽出，文分類などの様々な自然言語処理タスクで高い精度を示すことが明らかになっている．BERT の特徴として，特定のタスクに対するファインチューニングを行うことでタスクに適した分散表現の獲得ができる点が挙げられる．本研究では，東北大学乾研究室が公開する，日本語版 Wikipedia のテキストデータで事前学習された BERT モデル [3]を用いたファインチューニングを行うことで，暗黙的仮定の検出モデルを構築する.

図 2 に示すのは，本研究で構築する BERT モデルである．本モデルのファインチューニングには，3 節で述べた実験を通じて得られるクエリデータとその検索結果を用いた．これらのデータをトークン化した上で，クエリの先頭に [CLS] 特殊トークンを付与し，クエリと検索結果の間，そして末尾に [SEP] 特殊トークンを付与した．なお，クエリと検索結果の長さが最大入力長に満たない場合には，[PAD] 特殊トークンで埋める処理を行なった．入力データのトークン化と特殊トークンの付与後，BERT を適用することでベクトル化する．[CLS] 特殊トークンが入力データの特徴を表していると考え，[CLS] 特殊トークンを結果として用いた．この出力結果に対して Softmax 関数を用いることにより，出力結果を確率分布として解釈可能にした．なお，BERT モデルの損失関数には交差エントロピーを用いた.

4.2.1　損 失 関 数

BERT モデルの学習時において，正解分布と予測分布の一致度合いを測る損失関数として，交差エントロピーを用いる．式 1，式 2 を元にした定義を次に示す.

$$-\sum_{i=1}^{N}\sum_{j=1}^{3} P_{l_j}\left(q^{(i)}\right)\log\left(\hat{P}_{l_j}\left(q^{(i)}\right)\right) \tag{3}$$

5　評 価 実 験

4 節で提案した BERT による検出モデルの有効性を検証するため，ロジスティック回帰モデルを比較手法とした上で，2 種類のタスクについて評価実験を行った．本節では，タスクの概要と評価尺度，手法の詳細，実験結果について述べる.

5.1　実験で取り組むタスクと評価指標

本評価実験では，2 つのタスクについて取り組む．1 つ目がマルチクラス分類タスク，2 つ目がラベル分布予測タスクである．まず，それぞれのタスクの概要について説明する．その後，各タスクに対する評価尺度について述べる.

5.1.1　問題設定 1：マルチクラス分類タスクと評価

問題設定の 1 つ目は，4.1 節で述べた，ラベル分布予測タスクを解きやすいより単純なタスクにしたものである．分布予測よりも単純な本タスクにおける分類精度を評価することにより，分布予測の可能性を検証する．表 3，4，5 で示した各トピック，クエリの分布から最頻のクラスを取ったものを正解のクラスとする.「糖尿病とシナモン」トピックのクエリ「糖尿病改善 シナモン」を例に考えると，(a) の割合が最も高いため，(a) を正解クラスと考える．この例のように，本問題ではクエリが 3.1 節で定義した 3 種類のいずれかのクラスに属すると考え，マルチクラス分類タスクとして取り組む．なお，表 3 におけるクエリ「糖尿病改善 シナモン」のように，あるクラスに属する割合が重複するクエリは除去し，各クラスの分布が等しくなるようにアンダーサンプリングを行なっている．最終的に「糖尿病とシナモン」「がんとお茶」「喘息とカフェイン」の 3 トピックのクエリ（$N = 2058$）をモデルの学習と評価に用いた．分類精度の評価指標には 4 指標：正解率，マクロ適合率，マクロ再現率，マクロ F 値を設定した.

5.1.2　問題設定 2：ラベル分布予測タスクと評価

問題設定 2 は，4.1 節で述べたラベル分布の予測タスクである．表 3，4，5 で示した，各トピックとクエリにおける分布：【(a) 効果あり，(b) 効果なし，(c) 検証】を直接予測するのが，本問題で取り組むタスクである.「糖尿病とシナモン」「がんとお茶」「喘息とカフェイン」トピックのクエリ（$N = 2603$）をモデルの学習と評価に用いた．また，分布予測精度の評価に

3：https://github.com/cl-tohoku/bert-japanese

表 6　ファインチューニングに用いたハイパーパラメータ.

ハイパーパラメータ	数値等
バッチサイズ	16
最適化手法	Adam
損失関数	交差エントロピー
最大入力長	128
早期終了	8 patience
ドロップアウト率	0.1
学習率	2.0×10^{-5}

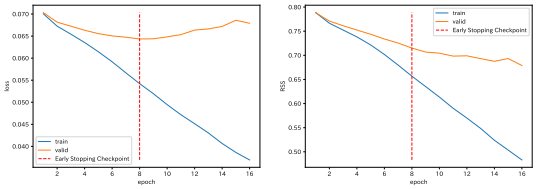

図 3　ラベル分布予測タスクにおける学習曲線.
（赤の破線は早期終了した際の epoch 数を表す）

は，2 指標：RSS（Rooted Residual Sum of Squares），JSD（Jensen-Shannon Divergence）を設定した．タスクは異なるものの，これらの評価指標は Tao らの研究 [9] において分布の一致度合いを測るために用いられている．正解分布と予測分布の一致度合いを評価する必要のある本研究においても用いることにした．なお，どちらの指標も小さいほど高い予測性能を持つことが示唆される．次に定義を示す．

a）　Rooted Residual Sum of Squares（RSS）.

i 番目のクエリ q における RSS は以下のように定義される．

$$\mathrm{RSS} = \sqrt{\sum_{j=1}^{3} (P_{l_j}(q^{(i)}) - \hat{P}_{l_j}(q^{(i)}))^2} \tag{4}$$

b）　Jensen-Shannon Divergence（JSD）.

i 番目のクエリ q における JSD は以下のように定義される．なお，$p_M(i) = \frac{P(q^{(i)}) + \hat{P}(q^{(i)})}{2}$ である．

$$\begin{aligned} &\mathrm{JSD}(P(q^{(i)}), \hat{P}(q^{(i)})) \\ &= \frac{\mathrm{KLD}(P(q^{(i)}) \parallel p_M(i)) + \mathrm{KLD}(p_M(i) \parallel \hat{P}(q^{(i)}))}{2} \end{aligned} \tag{5}$$

$$\text{ここで, } \mathrm{KLD}\left(p_1 \parallel p_2\right) = \sum_{i\ s.t.\ p_1(i)>0} p_1(i) \log_2 \frac{p_1(i)}{p_2(i)} \tag{6}$$

5.2　分類タスクと分布予測タスクに用いる手法

5.1 節で述べたタスクに対し，入力データに対するベクトル化方法を変えることで 2 種類のモデルを構築した．1 種類目が 4 節で述べた，東北大学の事前学習済みモデルをファインチューニングした BERT モデル，2 種類目がベクトル化として TF-IDF を用い，予測にロジスティック回帰を用いたモデルである．まず，それぞれのモデルについて説明する．

5.2.1　手法 1：BERT

手法の 1 つ目は，4 節で述べた東北大学の事前学習済みモデルをファインチューニングした BERT モデルである．本モデルにおいて，学習データ，検証データ，テストデータの比率は 6:2:2 としている．本研究では，ハイパーパラメータのチューニングは行っておらず，既存研究 [5] を参考に決定した．実際に用いたハイパーパラメータを表 6 に示す．また，ラベル分布予測タスクにおける学習時の学習曲線を図 3 に示す．

5.2.2　手法 2：ロジスティック回帰

手法の 2 つ目は，ロジスティック回帰モデルである．このモデルは，入力データのベクトル化として TF-IDF を用いている．本研究ではデータに含まれる「動詞，助動詞，形容詞，名詞」のみを用いてベクトル化を行った．TF-IDF により得られたベクトルをロジスティック回帰モデルで分類する．ロジスティック回帰モデルは入力データから学習した重みから，特定のクラスに属する確率を計算できる．この確率に対して閾値を定めた場合には問題設定 1（マルチクラス分類）に対する予測，閾値を定めない場合には問題設定 2（ラベル分布予測）に対する予測と考える．なお本研究では，ロジスティック回帰モデルのハイパーパラメータのチューニングは行っておらず，学習データとテストデータの比率は 8:2 としている．また，同じ問題設定の中では同一のテストデータを用いた評価を行っている．

5.3　実 験 結 果

本項では 5.1 節で述べたマルチクラス分類タスクとラベル分布予測タスクの 2 つの問題に対して，5.2 節で述べたモデルで分類，または予測を行った結果について述べる．

5.3.1　マルチクラス分類タスクに対する分類結果

マルチクラス分類タスクにおける分類結果について述べる．表 7 に示すのは本タスクにおける 2 手法（BERT，ロジスティック回帰）の分類精度である．2 手法については，クエリ単体のみを学習に用いた場合と，クエリに加えて検索結果も学習に用いた場合とでそれぞれ 2 種類ずつ示している．まず，クエリ単体で学習を行った場合と，クエリに加えて検索結果も学習に用いた場合で比較する．BERT モデルでクエリ単体を学習に用いた場合の正解率は 0.519 であった一方で，クエリに加えて検索結果も学習に用いた場合の正解率は 0.536 であった．ロジスティック回帰モデルでクエリ単体を学習に用いた場合の正解率は 0.510 であった一方で，クエリに加えて検索結果も学習に用いた場合の正解率は 0.517 であった．この結果から，クエリ単体に加えて検索結果も用いた方が分類精度が高くなることが分かった．次に，BERT モデルとロジスティック回帰モデルの比較を行う．クエリ単体を学習に用いた場合の BERT モデルの正解率が 0.519 であった一方で，ロジスティック回帰モデルの正解率は 0.510 であった．また，クエリに加えて検索結果も学習に用いた場合の BERT モデルの正解率は 0.536 であった一方で，ロジスティック回帰モデルの正解率は 0.517 であった．クエリ単体を学習に用いる場合，クエリに加えて検索結果も学習に用いた場合のどちらの場合おいても，BERT モデルの方が分類精度が高くなることが分かった．

5.3.2　ラベル分布予測タスクに対する予測結果

ラベル分布予測タスクにおける分布予測結果について述べる．表 8 に示すのは，本タスクにおける 2 手法（BERT，ロジスティッ

表 7　マルチクラス分類における各手法の分類精度.

モデル	正解率	マクロ適合率	マクロ再現率	マクロ F1 値
BERT（クエリ単体）	0.519	0.510	0.515	0.511
BERT（クエリ ＋ 検索結果）	**0.536**	**0.536**	**0.536**	**0.534**
ロジスティック回帰（クエリ単体）	0.510	0.506	0.507	0.506
ロジスティック回帰（クエリ ＋ 検索結果）	0.517	0.515	0.516	0.515

表 8　ラベル分布予測における各手法の予測精度.

モデル	RSS	JSD
ベースライン（一様分布として予測）	0.785	0.303
ベースライン（各分布の平均値として予測）	0.775	0.298
BERT（クエリ単体）	**0.698**	**0.261**
BERT（クエリ+検索結果）	0.737	0.279
ロジスティック回帰（クエリ単体）	0.704	0.261
ロジスティック回帰（クエリ+検索結果）	0.712	0.266

ク回帰）の予測精度と，比較用に用意したベースライン手法による予測精度である．比較用のベースラインは 2 種類の手法からなる．1 種類目の手法は予測分布を一様分布：【0.333,0.333,0.333】にしたものであり，2 種類目の手法は予測分布をテストデータ中における各分布の平均値：【0.230,0.289,0.481】にしたものである．ベースライン以外の 2 手法についてはクエリ単体のみを学習に用いた場合と，クエリに加えて検索結果も学習に用いた場合とでそれぞれ 2 種類ずつ示している．まず，クエリ単体で学習を行った場合と，クエリに加えて検索結果も学習に用いた場合で比較する．BERT モデルでクエリ単体を学習に用いた場合の RSS は 0.698 であった一方で，クエリに加えて検索結果も学習に用いた場合の RSS は 0.737 であった．ベースライン手法の RSS が 0.78 前後であることを鑑みると，BERT モデルはベースライン手法よりも高い精度を示した．マルチクラス分類の際には，クエリに加えて検索結果も用いた場合の方が精度が向上する傾向が見られたが，本タスクにおいては精度が低下するという結果になった．ロジスティック回帰モデルでクエリ単体を学習に用いた場合の RSS は 0.704 であった一方で，クエリに加えて検索結果も学習に用いた場合の RSS は 0.712 であった．BERT モデルと同様で，ベースライン手法よりも良い予測精度を示した．また，クエリに加えて検索結果も用いた場合の方が予測精度が低下する結果になり，BERT モデルと似た傾向が見られた．次に，BERT モデルとロジスティック回帰モデルの比較を行う．クエリ単体を学習に用いた場合の BERT モデルの RSS は 0.689 であった一方で，ロジスティック回帰モデルの RSS は 0.704 であった．また，クエリに加えて検索結果も学習に用いた場合の BERT モデルの RSS は 0.737 であった一方で，ロジスティック回帰モデルの RSS は 0.712 であった．クエリ単体を学習に用いる場合は BERT モデル，クエリに加えて検索結果も学習に用いた場合はロジスティック回帰モデルの方が高い予測精度を示した．

6　議　　論

本節ではまず，前節で得られた実験結果を元にしてリサーチクエスチョン（RQ）に対する回答と考察を行い，その後に本研究で明らかにすることができなかった限界点や課題点について述べる．本研究では，以下のようなリサーチクエスチョンを明らかにすることに取り組んだ．

RQ：　クエリからの暗黙的な仮定の検出は，どの程度の精度で可能なのか．

RQ について取り組むにあたって本研究では，マルチクラス分類タスクとラベル分布予測タスクの 2 種類のタスクを設定した．ここでは，ラベル分布予測タスクを踏まえた RQ への回答を行う．ラベル分布予測タスクにおいて，最も高い予測精度を示したのは BERT モデル（クエリ単体）であった．このモデルの RSS は約 0.70 であり，2 種類のベースライン手法よりも高い結果になった．この結果から，クエリ単体を入力として【(a) 効果あり, (b) 効果なし, (c) 検証】の分布をベースライン手法よりも精度良く予測できることが分かった．しかしどの手法においても，クエリに検索結果を加えて予測を行った場合には予測精度が悪くなるという結果になった．これらの結果について，表 9 に示す各手法の入出力例を元に考察を行う．ここでは，クエリ「効果なしがんお茶」を入力とする場合の出力値（予測値）に注目する．このクエリを入力とする場合の真値は【0,0.8,0.2】であり，効果なしの仮定（b）の割合が 80%，検証（c）の割合が 20%ということを示している．つまり，このクエリを投入した実験参加者は，80%の割合でがんにお茶が効果的でないと考えており，効果なし（b）の暗黙的仮定を置いていたことになる．逆に，20%の割合でがんにお茶が効果的かどうかを検証しようと考え，効果の有無について根拠を求めたと考えられる．各手法の RSS と JSD に注目すると，検索結果の有無で精度が大きく変化したのは BERT モデルであった．BERT（クエリ+検索結果）の出力値について，BERT（クエリ単体）の場合の出力値と比較すると，効果なし（b）の割合が高くなった上に，検証（c）の割合が低くなったことで真値の分布に近いものになっている．検索結果に「がんを作る」「がんになる」といった表現が含まれたことで,「がんに対してお茶は効果的でない」という割合が高まったと考えられる．

7　まとめと今後の課題

本研究では，健康情報を検索する際に置かれることのある仮定は【効果あり，効果なし，検証】の 3 種類であると定義し，クエリからの検出に取り組んだ．ラベル分布予測タスクの結果，最も高い精度を示したのはクエリ単体で学習を行った BERT モデルだった．予測分布を【0.333,0.333,0.333】のような一様分布とするベースライン手法の RSS は 0.785 であった一方で，本モデルの RSS は 0.698 だった．ラベル分布予測は，ベースラ

表 9　ラベル分布予測における各手法の入出力例.

モデル	入力データ	真値	予測値	RSS	JSD
ベースライン（一様分布として予測）	なし	【0, 0.8, 0.2】	【0.33, 0.33, 0.33】	0.196	0.173
BERT（クエリ単体）	[CLS] 効果なしがんお茶 [SEP]	【0, 0.8, 0.2】	【0.06, 0.57, 0.38】	0.099	0.045
BERT（クエリ+検索結果）	[CLS] 効果なしがんお茶 [SEP] あなたも飲んでいるお茶がガンを作る! 研究から分かった意外お茶・コーヒー類について 末期がんの緩和ケア 川崎市 知らずにみんなが飲んでるガンになるお茶 [SEP]	【0, 0.8, 0.2】	【0.13, 0.62, 0.26】	0.077	0.053
ロジスティック回帰（クエリ単体）	効果, なし, がん, お茶	【0, 0.8, 0.2】	【0.05, 0.62, 0.33】	0.076	0.031
ロジスティック回帰（クエリ+検索結果）	効果, なし, がん, お茶, あなた, 飲ん, いる, お茶, ガン, 作る, 研究, 分かっ, た, 意外, お茶, コーヒー, 類, 末期, がん, 緩和, ケア, 川崎, 市, 知ら, ず, みんな, 飲ん, でる, ガン, なる, お茶	【0, 0.8, 0.2】	【0.07, 0.61, 0.32】	0.079	0.038

イン手法よりも高い精度での予測が可能であることが明らかになった．本研究で取り組むことができなかった，モデルの構築面における課題と限界点について述べる．本研究では，実験参加者が作成したクエリを検出モデルの学習に用いて，暗黙的な仮定の検出に取り組んだ．ここで考えられるのが「暗黙的な仮定を検出するのに必要な情報や特徴が学習データに含まれていたのか」という点である．実際に我々が Web 検索を行う際にはクエリを複数回作成し，様々な Web ページを閲覧する．こうした複数回の Web 検索行動（検索セッション）の中で，仮定を置いた検索が生まれる場合もあると考えられる．検索者が暗黙的な仮定を置いているかどうかを検出するには，1 回の検索で投入されるクエリではなく，検索セッション全体に注目するべきである．本研究における実験では，検索セッションについては全く考慮できておらず，そこで得られる特徴については全て無視している．検索時のセッションについても特徴として捉えた上でモデルの学習を行った場合，本研究の結果より高い精度が期待できると考えられる．

謝辞　本研究は JSPS 科学研究費助成事業 JP21H03774, JP21H03775，による助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Neda Ashrafi-Amiri and Josef Al-Sader. Effects of Confirmation Bias on Web Search Engine Results and a differentiation between Non-assumptive versus Assumptive Search Queries. *Bachelor of Science Thesis in the Software Engineering and Management Programme in University of Gothenburg*, 2016.

[2] Charles L. A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. Offline Evaluation by Maximum Similarity to an Ideal Ranking. In *Proc. of CIKM*, pp. 225–234, 2020.

[3] Charles LA Clarke, Mark D Smucker, and Maria Maistro. Overview of the TREC 2021 Health Misinformation Track. In *Proc. of TREC*, 2021.

[4] Charles LA Clarke, Alexandra Vtyurina, and Mark D Smucker. Assessing Top-k Preferences. *ACM Transactions on Information Systems*, pp. 1–21, 2021.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*, pp. 4920–4928, 2019.

[6] Cameron D Norman and Harvey A Skinner. eHEALS: the eHealth literacy scale. *Journal of medical Internet research*, Vol. 8, No. 4, p. e27, 2006.

[7] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proc. of SIGIR*, pp. 209–216, 2017.

[8] Suppanut Pothirattanachaikul, Takehiro Yamamoto, Yusuke Yamamoto, and Masatoshi Yoshikawa. Analyzing the Effects of Document's Opinion and Credibility on Search Behaviors and Belief Dynamics. In *Proc. of CIKM*, pp. 1653–1662, 2019.

[9] Sijie Tao and Tetsuya Sakai. Overview of the NTCIR-16 Dialogue Evaluation (DialEval-2) Task. *Proc. of NTCIR-16*, pp. 51–65, 2022.

[10] Ryen W. White. Beliefs and biases in web search. In *Proc. of SIGIR*, pp. 3–12, 2013.

[11] Ryen W. White. Belief dynamics in web search. *Journal of the Association for Information Science and Technology*, Vol. 65, No. 11, pp. 2165–2178, 2014.

[12] Takehiro Yamamoto, Yusuke Yamamoto, and Sumio Fujita. Exploring People's Attitudes and Behaviors Toward Careful Information Seeking in Web Search. In *Proc. of CIKM*, pp. 963—972, 2018.

[13] Yusuke Yamamoto and Takehiro Yamamoto. Query Priming for Promoting Critical Thinking in Web Search. In *Proc. of CHIIR*, pp. 12–21, 2018.

[14] Yusuke Yamamoto, Takehiro Yamamoto, Hiroaki Ohshima, and Hiroshi Kawakami. Web Access Literacy Scale to Evaluate How Critically Users Can Browse and Search for Web Information. In *Proc. of WebSci*, pp. 97–106, 2018.

[15] Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proc. of WWW*, pp. 1011–1018, 2010.

[16] Dake Zhang, Amir Vakili Tahami, Mustafa Abualsaud, and Mark D. Smucker. Learning Trustworthy Web Sources to Derive Correct Answers and Reduce Health Misinformation in Search. In *Proc. of SIGIR*, pp. 2099–2104, 2022.

[17] 光武誠吾, 柴田愛, 石井香織, 岡崎勘造, 岡浩一朗. eHealth Literacy Scale (eHEALS) 日本語版の開発. 日本公衆衛生雑誌, pp. 361–371, 2011.

[18] 厚生労働省. 平成 26 年版厚生労働白書～健康・予防元年～. `https://www.mhlw.go.jp/wp/hakusyo/kousei/14/index.html`. 2022 年 10 月 10 日閲覧.

[19] 浜島聡一郎, 山本岳洋, 山本祐輔, 大島裕明. 健康情報検索における信憑性判断と意見の形成に関する調査. 第 14 回データ工学と情報マネジメントに関するフォーラム, A34–5, 2022.

[20] 鈴木雅貴, 鈴木雅貴, 山本祐輔. 確証バイアスとウェブ検索行動の関係分析. 第 12 回データ工学と情報マネジメントに関するフォーラム, D4–3, 2020.