抽象度が高いクエリによるアンケートデータの設問検索

田中 稔也　熊谷 雄介　藤井 遼
株式会社博報堂 DY ホールディングス
{toshiya.tanaka,yusuke.kumagae,ryo.c.fujii}@hakuhodody-holdings.co.jp

概要

膨大なアンケートデータから，分析目的に合致する設問を検索する作業は，マーケターにとって手間と時間がかかる工程の一つである．この作業工程をユーザのクエリ入力による設問検索で効率化することを提案する．このとき，クエリの意味や内容が抽象的であってもユーザが望む設問を検索することを目指す．本研究では，アンケートデータ設問検索タスクに取り組む．それに伴い，マーケターのアノテーションによる設問検索データセットを構築した．複数の手法を用いた設問検索実験の結果，文から直接埋め込み表現を取得しクエリとの類似度を求める手法が高い精度を示した．また，設問検索特有の課題が明らかとなった．

1　はじめに

マーケティングのためにデータ分析を行うマーケターの頻出業務に，アンケートデータに代表される調査データの分析がある．その際，対象の調査データから分析目的に合致する設問を適切に見つける必要がある．例えば，アンケートデータから「若者の飲酒傾向や割合」を把握したい場合，「飲酒の有無」や「好きなお酒の種類」，「酔っ払いは苦手だ」などの設問を発見する必要がある．

単純な方法として，全設問を確認するという方法があるが，これでは工数がかかりすぎてしまう．次に考えられるのは，キーワードによる文字列検索であるが，適切なキーワードが設定できないとマッチしすぎてしまったり，反対に取りこぼしてしまう．

そこで，本研究は，ユーザが入力したクエリをもとにユーザが望む設問を検索する**設問検索**に取り組む．設問検索は次の二点の違いから全文検索にもとづく文書検索の手法がそのまま適用できない．

まず，文書（設問）に含まれる単語数が設問検索では少ないことである（表 1）．一般的な文書検索タスクは，基本的に文書が長文であり，それ故に文書に含まれる単語数も多いが，設問検索の場合，設問が短文であるため含まれる単語数も少ない．

次に，設問検索のクエリはより抽象度が高いことである．ユーザ行動履歴データなどを用いない，検索エンジンや EC サイトにおける従来の文書検索では，クエリが「白 T シャツ 透けない」[1] のように，文書中に直接現れる単語で構成されることが多く，クエリは非常に具体的なものになりやすい．しかし，設問検索は，マーケターが構想段階の仮説検証や情報収集に質問形式のクエリ（例：「ファッションに関する意識・傾向を教えて」）を入力することを想定しているため，クエリの抽象度が高い．

本研究では設問検索のためのデータセット構築と，その精度検証に取り組んだ．評価データセットは，マーケターによる，実際のアンケートデータに対して考えたクエリと，クエリと設問ごとの対応付けにより構築した．このデータセットに対して代表的な情報検索手法を用い，設問検索の精度や，明らかになった設問検索特有の課題を報告する．

2　関連研究

1 で述べた文書長とクエリの抽象度の高さの二つの観点から，他の日本語文書検索データセットと本タスクのデータセットとの違いを説明する．

さまざまな文書検索ベンチマークデータセット[1)]と本データセットそれぞれにおけるクエリ q_i と文書 d_j の平均単語数[2)]および，関連する q_i, d_j ペアに含まれる共通単語の Simpson 係数を表 1 に示す．

また，本データセット以外の調査データとして，国民生活に関する世論調査 [2]，消費者意識基本調査 [3]，生活意識に関するアンケート調査 [4] における文書（設問）単語数の平均も表 1 に示す．

クエリおよび文書の単語数の比較　本データセットと比較すると，いずれのデータセットも，クエリの単語数は同程度に少なく，文書の単語数は非常に

1）多言語のものは日本語のみを対象とした
2）名詞，動詞，形容詞を単語として扱った

表 1: 各データセットにおけるクエリ，文書の平均単語数と共通単語の Simpson 係数．**太字**は各列で最も低い値を示す．

データセット	クエリ	文書	Simpson
Amazon ESCI （商品検索） [1]	**2.6**	112.3	0.63
NTCIR-15 （統計データ検索） [5]	3.6	58.6	0.57
Mr.TyDi （QA） [6]	4.3	38.4	0.46
JAQKET （QA） [7]	12.1	928.5	0.53
生活定点（本データセット）	5.6	**10.8**	**0.08**
国民生活に関する世論調査	-	14.6	-
消費者意識基本調査	-	15.4	-
生活意識に関するアンケート調査	-	13.6	-

多いことがわかる．一方で，日本語調査データの文書（設問）の平均単語数は，本データセットと同様に少ないことが確認できる．このことから，調査データは一般に設問が短い傾向にあり，本データセットだけが特殊なわけではないことがわかる．

クエリの抽象度の比較　次に，クエリの単語が文書にどれだけ含まれているか，つまり，クエリがどれだけ具体的か（あるいは抽象的か）を測るため，関連するクエリと文書の Simpson 係数の平均を求める．Simpson 係数は

$$\frac{|\mathcal{W}_{q_i} \cap \mathcal{W}_{d_j}|}{\min(|\mathcal{W}_{q_i}|, |\mathcal{W}_{d_j}|)}$$

であり，$\mathcal{W}_{q_i}$ と $\mathcal{W}_{d_j}$ はそれぞれクエリと文書に含まれる単語の集合である．Simpson 係数は 0 から 1 の値を取り，大きいほどクエリと文書に含まれる単語が重複することを意味する．

Simpson 係数を見ても，本データセットはクエリと設問で共通する単語が非常に少ないことが確認できる．これは，本タスクの用途にマーケターの構想段階の仮説検証や情報収集を想定したことにより，多くのクエリの抽象度が高いためである．

3　手法

3.1　設問検索の定義

本研究で取り組む**設問検索**はユーザのクエリを $q_i \in \mathbb{Q}$，検索対象の設問を $d_j \in \mathbb{D}$ として，q_i に関連する順に $d_j \in \mathbb{D}$ を並び替えるタスクである．

本研究においては，q_i と d_j の真の対応関係 $y_{i,j}$ は連続値ではなく，関連がある場合には $y_{i,j} = 1$ を，関連がない場合には $y_{i,j} = 0$ を取るとする．また，並び替え時には真の対応関係 y の情報は用いることができないとする．すなわち，本研究において設問検索は教師なしの設定で取り組む．これは，他の調査データへ転用を行う度に，関連性のアノテーションを行い教師データを構築するのはコストが高いという点と，本研究で用いたデータセットを教師なしの設定でどの程度の精度が得られるか確認できれば，他の調査データでも同等の精度を示すだろうと考えられるためである．

3.2　アプローチ

3.1 で述べたように，設問検索は q_i に関連がある順に d_j を並び替えるタスクである．すなわち，q_i と d_j の類似度 $\mathrm{sim}(q_i, d_j)$ を求めれば良い．類似度の計算には大きく分けて 2 種類のアプローチがある．全文検索にもとづく手法と埋め込み表現にもとづく手法である．

3.2.1　全文検索にもとづく手法

全文検索にもとづく手法では，TF-IDF [8] と BM25 [9] を用いる．TF-IDF と BM25 はどちらも，クエリ q_i と設問 d_j に含まれる単語の出現頻度と出現文書数を用いて q_i と d_j のベクトル化を行うが，TF-IDF が単語の出現頻度と出現文書数のみに着目するのに対し，TF-IDF を拡張した BM25 では，文書長にもとづく正規化や，単語出現頻度に対する影響を調整するハイパーパラメータが導入されている．

3.2.2　埋め込み表現にもとづく手法

埋め込み表現にもとづく手法では，q_i と d_j の文字列そのものを比較するのではなく，それぞれをニューラル言語モデルを用いて，密な埋め込みベクトル $e(q_i), e(d_j)$ で表現した上で類似度 $\mathrm{sim}(e(q_i), e(d_j))$ を求める．

Word2Vec [10] は，単語の埋め込みは可能だが，文を直接埋め込むことはできない．そのため，q_i と d_j それぞれに含まれる単語の埋め込みベクトルを平均プーリングすることで文埋め込みを得る．

Sentence-BERT [11] や大規模言語モデル由来の文埋め込みモデルなどを用いることで，直接文埋め込み表現が得られる．

Sentence-BERT は文の埋め込み表現能力に優れたニューラル言語モデルであり，q_i と d_j を入力した際の [CLS] トークンの出力が $e(q_i)$ と $e(d_j)$ である．

大規模言語モデル由来の文埋め込みモデルは，OpenAI[3] などをはじめとする事業者によって提供されており，彼らの Web API などを用いることで文埋め込み $e(q_i), e(d_j)$ が得られる．

3）https://openai.com/

2 つのアプローチの違い　全文検索にもとづく手法は，q_i と d_j の間に共通する単語が存在しない場合，$\mathrm{sim}(q_i, d_j)$ が 0 になる．2 で述べたように，本タスクで扱うクエリと設問は，単語数が少なく，共通する単語も少ないことから，全文検索にもとづく手法では難しいことが考えられる．一方，埋め込み表現にもとづく手法では，クエリと設問それぞれの内容や意味を考慮した埋め込みを行った上で類似度を求めるため，共通する単語が存在しない場合であっても類似度を求めることができる．そのため，本タスクへのアプローチとしては，全文検索にもとづく手法よりも埋め込み表現にもとづく手法の方が機能すると考えられる．

4　実験

4.1　検証用データセット

本研究では，**生活定点**[4]を対象にマーケターによるアノテーションを行い，検証用データセットを構築した．

4.1.1　生活定点

生活定点は，博報堂生活総合研究所が 1992 年から隔年で実施する時系列観測調査であり，日頃の感情，生活行動や消費態度，社会観など，多角的な設問から，生活者の意識と欲求の推移を分析することを目的としている．本研究では，生活定点の多岐にわたる設問に対して検索を行う．生活定点は質問項目（例：「恋愛・結婚全般の意識・行動」）と詳細（例：「デートの時の勘定は男が払うべきだと思う」）から構成されている．今回は

「#{ 質問項目 }」に関する質問に対して「#{ 詳細 }」と回答

のように質問項目と詳細を連結したものを**設問** d_j とした．総設問数 $|\mathcal{D}|$ は 1,410 件である．

4.1.2　アノテーション

本データセットは，実際のアンケートデータに対する（1）マーケターによるクエリ作成（2）クエリ作成者とは異なるマーケターによる，関連すると思われる設問の対応付け，の 2 ステップで構築した．（1）では，クエリを 108 個作成した．（2）では q_i と d_j の対応関係を $y_{i,j} \in \{0, 1\}$ とし，$y_{i,j}$ は 2 人のマーケターのいずれかが q_i と d_j は関連すると判定した際に，$y_{i,j} = 1$ を，関連しないと判定した場合に $y_{i,j} = 0$ を取るとした．各クエリ q_i において $y_{i,j} = 1$ を取る d_j の数の平均は 40.6，標準偏差は 40.3，最小 1，最大 249 であった．データ例を表 2 に示す．

表 2: 生活定点データの例．クエリ q と関連性 y はマーケターによるアノテーションによって構築した．

クエリ q	設問 d	関連性 y
タイパ意識に関するデータはありますか？	「生活速度/欲求」に関する質問に対して「（どちらかといえば）手早くやりたい方」と回答	1
タイパ意識に関するデータはありますか？	「生活価値観」に関する質問に対して「不正直なことは，どんな状況でもしない」と回答	0
⋮	⋮	⋮
子育てに関する不安や悩みはどんなものがありますか？	「現在お金をかけているもの」に関する質問に対して「子供のための教養・勉強にかけるお金」と回答	1

4.2　ベースライン

3 で述べたように，本研究では，q_i と d_j の類似度 $\mathrm{sim}(q_i, d_j)$ を求めるため複数の手法を使用した．全文検索にもとづく手法では MeCab [12] と mecab-ipadic-NEologd [13] を用いて得られた名詞，動詞，形容詞にもとづく単語 Unigram を用いた．

埋め込み表現にもとづく手法では，ニューラル言語モデルの Word2Vec, Sentence-BERT, 大規模言語モデル由来の文埋め込みモデルを用いた．Word2Vec は，Word2vec/wikipedia2vec_jawiki_20180420_300d[5] を用いて得た 300 次元の単語埋め込みから平均プーリングを用いて文埋め込みを得た．Sentence-BERT は，sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2（384 次元）[6] と，cl-nagoya/ruri-large（1,024 次元）[14] を用いて文埋め込みを得た．大規模言語モデルにもとづく文埋め込みでは，OpenAI の text-embedding-ada-002 [15]（1,536 次元）を用いた．また，$\mathrm{sim}(e(q_i), e(d_j))$ 算出には，全てコサイン類似度を用いた．

同時に，q_i によらず d_j をランダムに並び替えて出力するベースライン（Random）も用意した．

4）https://seikatsusoken.jp/teiten/

5）https://huggingface.co/Word2vec/wikipedia2vec_jawiki_20180420_300d

6）https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

表 3: 実験結果．すべての指標は高いほど予測精度が優れていることを意味する．**太字**はそれぞれの指標の最高値．

手法	MAP@30	ROC-AUC
Random	0.003 ± 0.007	0.500
TF-IDF	0.118 ± 0.190	0.615 ± 0.147
BM25	0.109 ± 0.178	0.616 ± 0.147
Word2Vec	0.058 ± 0.127	0.699 ± 0.151
multilingual-MiniLM	0.225 ± 0.210	0.854 ± 0.146
ruri-large	**0.258** ± **0.214**	**0.882** ± **0.109**
text-embedding-ada-002	0.228 ± 0.211	0.848 ± 0.140

4.3 評価指標

設問検索の評価指標には，関連度の上位 30 件に対する当てはまりを測る MAP@30 と，すべての関連度の良さを測る ROC-AUC を用いた．

Average Precision@30（AP@30） [16] は q_i における類似度の上位 30 件にどれだけ真に類似する d_j が含まれているかを意味する指標であり，全 q_i に対する AP@30 の平均が Mean Average Precision@30（MAP@30）である．AP@30 および MAP@30 は 0 から 1 の値を取り，高いほど類似度の上位 30 件が正確であることを意味する．本論文では MAP@30 と AP@30 の標準偏差を報告する．

ROC-AUC（Receiver Operating Characteristic curve - Area Under the Curve） [17] は，関連があるものをより上位に並び替えられた度合いを意味する．ROC-AUC は 0 から 1 の値を取り，大きいほど予測が正しいことを意味する．また，ランダムな予測に対する ROC-AUC は 0.5 である．本論文では各クエリの ROC-AUC の平均と標準偏差を報告する．

4.4 結果

各手法の MAP@30 と AP@30 の標準偏差，ROC-AUC の平均と標準偏差を表 3 に示す．全ての手法が，ランダムベースラインよりも高い MAP@30, ROC-AUC を示した．また，ruri-large が最も高い MAP@30 と ROC-AUC を示した．これは ruri-large が，今回用いたアプローチの中で最も関連性の高い設問を上位に並べ替える能力が高く，また，全体の並べ替え能力も高いことを示している．

4.5 考察

まず，本タスクは q および d それぞれの単語数が少なく，それに伴い，共通の単語も発生しにくいために，全文検索にもとづく手法と Word2Vec の精度が低くなったと考えられる．一方で，文を単語単位で扱わず，文そのものを埋め込み表現に変換して類似検索するアプローチは精度が高く，本タスクにより適していることが実験から明らかになった．

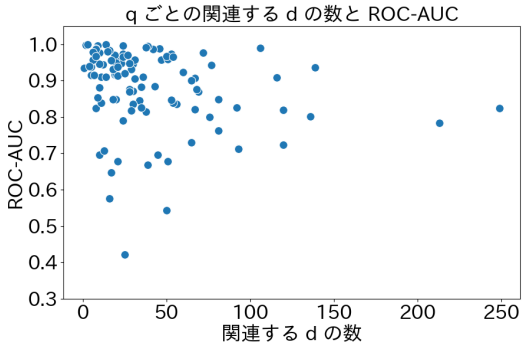

図 1: 関連する d の数別 各クエリの ROC-AUC 分布．縦軸が ROC-AUC，横軸は各 q の関連する d の数を表す．

次に，各 q の正例数（関連する d の数）と ROC-AUC の関係について分析を行う．図 1 は最も高い精度を示した ruri-large における，各 q の関連する d の数と ROC-AUC を表す．図 1 から（1）関連する d の数は ROC-AUC に関係しない（2）ruri-large において ROC-AUC が 0.5 を下回った q は一件のみだった，の二つがわかる．

ruri-large において，ROC-AUC が 0.5 を下回ったクエリは「リスキリングに取り組んでいる人はどれくらいいますか？」（0.421）のみであり，そのクエリに対して類似度が高かった設問上位三つは全て「飲酒の頻度」に関するもので「リスキリング」とは無関係だった．このことから，ruri-large は「リスキリング」という単語を知らないと考えられる．

マーケターは職業柄最新の概念を調べることや，仮に新語そのものが設問群に含まれていなかったとしても，「リスキリング」の例では「学び直し」「生涯学習」のように近い意味の設問を発見したいといった理由から，このような新語対応は必須であり，設問検索においては特に致命的であると言える．

5 おわりに

本論文では，アンケートデータの設問検索に取り組んだ．タスクへの取り組みにあたり，複数の文類似度算出手法を用いた精度検証を行った．その結果，文を直接埋め込み表現する手法が本タスクに適していることが明らかになった．今後の本研究の発展として「国民生活に関する世論調査」などの公的機関によって実施された調査データに対するマーケターのアノテーションと，そのデータ公開による設問検索タスクの共有が挙げられる．

謝辞

本論文の作成にあたり，コメントを下さった，株式会社博報堂 DY ホールディングスの牛久雅崇氏，岩井皓暉氏に感謝申し上げます．

参考文献

[1] Chandan K Reddy, et al. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. **arXiv [cs.IR]**, 2022.
[2] 内閣府. 国民生活に関する世論調査, 2021.
[3] 消費者庁. 消費者意識基本調査, 2021.
[4] 日本銀行. 生活意識に関するアンケート調査, 2022.
[5] Kato P Makoto, et al. Overview of the NTCIR-15 data search task. In **NTCIR**, 2020.
[6] Xinyu Zhang, et al. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In **MRL**, 2021.
[7] 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. Jaqket: クイズを題材にした日本語 qa データセットの構築. 言語処理学会第 26 回年次大会, 2020.
[8] Ho Chung Wu, et al. Interpreting TF-IDF term weights as making relevance decisions. **ACM Trans. Inf. Syst.**, Vol. 26, No. 3, 2008.
[9] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. **Found. Trends® Inf. Retr.**, Vol. 3, No. 4, 2009.
[10] Tomas Mikolov, et al. Distributed representations of words and phrases and their compositionality. In **NIPS**, 2013.
[11] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In **EMNLP**, 2019.
[12] 工藤拓. Mecab: Yet another part-of-speech and morphological analyzer, 2013.
[13] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会, 2017.
[14] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese general text embeddings. **arXiv [cs.CL]**, 2024.
[15] Tom B Brown, et al. Language models are few-shot learners. In **NIPS**, 2020.
[16] Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In **SIGIR**, 2000.
[17] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern Recognition**, Vol. 30, No. 7, 1997.