

商品検索におけるゼロマッチ解消のためのデータセット構築の検討

雨宮 佑基[†] 真鍋 知博^{††} 藤田 澄男^{††} 酒井 哲也[†]

[†] 早稲田大学基幹理工学研究科情報理工・情報通信専攻 〒169-8555 東京都新宿区大久保 3-4-1

^{††} ヤフー株式会社 〒102-8282 東京都千代田区紀尾井町 1-3 東京ガーデンテラス紀尾井町 紀尾井タワー

E-mail: yukiamemiya@fuji.waseda.jp, {tomanabe, sufujita}@yahoo-corp.jp, tetsuyasakai@acm.org

あらまし Yahoo!ショッピングなどのEコマースサイトの普及により、誰もがオンラインで興味のある商品や購入したい商品を検索することができるようになった。しかし、ユーザーが商品検索エンジンに適切なクエリを入力できずヒットした商品数がゼロの状況(ゼロマッチ)が起ってしまうと、ユーザーの購買意欲減退、サイトの販売機会損失につながる。そこで、ゼロマッチを回避するために、商品検索エンジン側でゼロマッチクエリを適切なクエリに書き換えることが望ましいが、ゼロマッチの原因は様々であり、適切な書き換え方法もこれに依存する。そこで本研究では、まず実際のゼロマッチクエリの分析によりこれらを適切に書き換える方法を(1)単語削除、(2)単語置換、(3)完全な書き換え、(4)誤り修正、(5)その他、の5つに類型化した。さらに、クラウドソーシングにより各ゼロマッチクエリに適した書き換え方法の類型を付与し、ゼロマッチクエリと書き換え方法の類型とのペアからなるデータセットを構築した。最後に、構築したデータセットと機械学習モデルに基づく分類器を用いてゼロマッチクエリを分類を行った。キーワード 情報検索, Eコマース, ゼロマッチクエリ, データセット

1. はじめに

Yahoo!ショッピングやAmazon, 楽天市場といったEコマースサイトを利用するユーザーの数は年々増加傾向にあり、世界のユーザーの67%が月に一回以上オンラインで商品を購入している。このことは、2019年に世界的なコンサルティングファームであるプライスウォーターハウスクーパース(PwC)の調査が示す通りである。^(注1) ユーザーがEコマースサイトで商品を検索するときは、まず商品検索ページで検索クエリ、つまり購入意図を反映した検索キーワードを入力する。しかし、入力したクエリが適切でないと、商品の検索が不成功に終わり、検索結果商品数がゼロの状況(ゼロマッチ)に陥ってしまう。これは「検索クエリ中の単語をスペルミスしてしまう」や「色や形のような条件を絞り込みすぎてしまう」といったことが原因であることが多い。例えば、Yahoo!ショッピングサイトで条件を絞り込みすぎたクエリである「テレビ台 西洋アンティーク 中古」を検索すると、図1で示すように、「テレビ台 西洋アンティーク 中古」に一致する商品は見つかりませんでした」と表示される。多くのユーザーは、こういった状況に遭遇すると、購買意欲を失い、そのサイトから離脱してしまう。そしてこのことは、Eコマースサイト側の販売機会の損失につながってしまう。そこで、ゼロマッチを回避するためには、1つでも商品検索結果が返ってくるようにユーザーによって入力された検索クエリを検索エンジン側で適切に書き換え、その書き換えられたクエリで再度商品を検索するシステムの構築が必要である。

ゼロマッチクエリを適切に書き換える方法は、ゼロマッチに

なってしまう原因に依存するため、様々である。例えば、条件を絞り込みすぎたクエリに対しては、その絞り込みを緩和するために、検索クエリ中の不要な単語を削除する手法が適している。一方、入力誤りをしてしまったクエリに対しては、スペル修正をおこなう手法が効果的である。このように、どのようなゼロマッチクエリに対してどういった手法を適用するのが最適なのか、すなわち類型化された書き換え方法の中から最も適している方法を自動判別するシステムを構築することは、Eコマース検索エンジンの精度向上のために不可欠である。しかし、そのような研究は我々の知る限りまだ存在しない。そこで本研究では、まず実際のゼロマッチクエリの分析により、これらを適切に書き換える方法を類型化する。さらに、あるゼロマッチクエリが与えられたときに、このような自動判別システムを構築する上で必要な、ゼロマッチクエリと書き換え方法の類型とのペアからなるデータセットを構築する。書き換え方法の類型は、(1)単語削除、(2)単語置換、(3)完全な書き換え、(4)誤り修正、(5)その他、とし、クラウドソーシングにより各ゼロマッチクエリに付与する。最後に、構築したデータセットと機械学習モデルに基づく分類器を用いてゼロマッチクエリを分類する。



図1 商品検索におけるゼロマッチクエリの例

(注1) : <https://www.pwc.com/gx/en/consumer-markets/consumer-insights-survey/2019/report.pdf>

2. 関連研究

2.1 ゼロマッチクエリと書き換え方法

Singh ら [1] は、E コマース検索におけるゼロマッチという問題に焦点を当て、ゼロマッチクエリの特徴やゼロマッチが起きる原因、また実際にゼロマッチが起きたときにユーザーがどのように行動するかなどの分析を行なっている。彼らはゼロマッチの原因として、「ゼロマッチクエリは非ゼロマッチクエリに比べて冗長であること」や「ユーザーと E コマースサイトが使用する語彙が違うこと」などを挙げている。

上記のようなゼロマッチを解消するために、ゼロマッチクエリを書き換えに取り組んでいる先行研究も存在する。例えば Parikh ら [2] は、オンラインでの検索行動や商品購入活動を含む高品質のトランザクションデータを使用して、クエリ間のセマンティックな関係を推測する Semantic Query Network を構築した。彼らはこのネットワークをクエリを書き換えに利用することでゼロマッチの解消を試みている。一方で Singh ら [3] は、ユーザーによって入力されたクエリ中の単語を削除する Query Relaxation と、書き換えの正確性を向上させるために時間ベースの適合性フィードバックを利用したクエリ書き換えシステムを構築した。さらに Tan ら [4] は、ユーザーが入力した検索クエリ中の冗長な単語を削除する単語削除アルゴリズムと、ユーザーの行動ログと言語モデルによって適切な単語を生成する単語置換アルゴリズムを利用することで、ユーザーが持つ本来の検索意図を変更することなく、検索結果としての商品数を増やすクエリ書き換えシステムを提案した。

E コマース上でゼロマッチが起こる原因の一つである、ユーザーと E コマースサイトの間で使用する語彙が違う (Vocabulary Gap) という問題の解決に向けた研究も行われている。Maji ら [5] の研究では、Vocabulary Gap につながるクエリを、それと意味的に類似し、頻度とクリックスルー率が高い *Well Performing* クエリに書き換える手法を提案した。また Manchanda ら [6] は、クエリに含まれる、検索のために重要な用語を特定し、より検索意図を表現する単語を追加することで、ユーザーの商品検索エクスペリエンスを改善した。

しかしこれらの先行研究は、ゼロマッチが生じる原因を 1 つあるいは 2 つ挙げ、その原因に対抗する解決策の提案を行っているのみである。そしてゼロマッチクエリの原因にはどのような種類があり、その原因に対してどのようなクエリ書き換え方法が適しているかについては論じていない。しかし本研究では、ゼロマッチクエリを分析し、クエリを書き換え方法を類型化することで、網羅的にゼロマッチを解消することを試みる。

2.2 その他のクエリ書き換え方法

ゼロマッチ解消を目的とはしていないが、E コマース検索エンジンの検索精度を向上させるために、多くの研究者によってクエリ書き換え方法の研究が行われている。

Yang ら [7] によれば、単語の言語的特徴・削除履歴・クエリの文脈といった 3 種類の特徴によって検索クエリ内の単語を表現し、ユーザーの検索ニーズに最も適合していない単語を削除することで、ユーザーの検索成功率の向上につながったという。

一方、Jones ら [8] は、ユーザーが入力した検索クエリをフレーズごとに分割し、各フレーズを置換する手法を提案している。また彼らは、クエリを書き換えは機械翻訳問題と見なすことができ、彼らのクエリ置換アルゴリズムを改善するには機械翻訳技術からインスピレーションを得ることが必要であると述べている。そのため、機械翻訳技術を応用してスペルミスを修正する研究が Hasan ら [9] や Zhou ら [10] によって行われている。

本研究では、これらのクエリ書き換え方法に関する先行研究から得た知見を活用し、ゼロマッチクエリを書き換える方法を類型化する。

2.3 クエリの類型化・クエリ遷移の類型化

ウェブ検索におけるクエリあるいはクエリ遷移の類型化に関する先行研究には、以下のようなものがある。Broder [11] は、ウェブ検索を行うユーザーの検索意図に基づいて、ウェブ検索におけるクエリは *Navigational*, *Informational*, *Transactional* の 3 種類に分類できると述べている。また Huang ら [12] は、ウェブ検索におけるクエリ書き換え方法に関する先行研究の結果をもとに、独自のクエリ類型を構築した。そのクエリ類型の中には、*remove words* や *word substitution*, *spelling correction* など 13 種類のクエリ書き換え方法が含まれている。

一方で、E コマース検索における検索タスク・検索行動に関する研究があまり行われていないという課題に挑戦するため、Sondhi ら [13] は E コマース検索エンジンの検索ログを分析し、ユーザーの検索行動に基づいて検索クエリを 5 つのカテゴリに分類した。このカテゴリ分けにより、E コマース検索エンジンを利用するユーザーの検索行動を把握するだけでなく、個人に関連性の高い検索結果を提供することで、ユーザーの検索を最適化することにつながるという。また、Manchanda ら [6] によって、E コマース検索におけるクエリ遷移パターンは、5 つのカテゴリに分けられることが示されている。

しかし、ゼロマッチクエリを類型化することを目的とした研究は、我々の知る限りまだ存在しない。そのため本研究は、ゼロマッチクエリを書き換え方法を類型化する初めての試みである。

3. ゼロマッチクエリ書き換え方法の類型化

3.1 ゼロマッチクエリの分析

ゼロマッチクエリを適切に書き換える方法を類型化するために、まず Yahoo!ショッピング^(注2)における 2ヶ月分の検索クエリログの一部を使用して、ゼロマッチクエリの分析を行った。このとき、ゼロマッチクエリの分析に利用した検索クエリログは、検索エンジンによるクエリサジェクションをユーザーがクリックした場合や機械的に判断されるアクセスがあった場合などを除き、ユーザーによって手入力されたものに限定した。

使用した検索クエリログに含まれる全クエリのうち、ゼロマッチクエリが占める割合は 6.7%であった。一方、商品検索

(注2) : <https://shopping.yahoo.co.jp/>

表 1 クエリ書き換え方法のタイプ分類

本研究	Parikh [2]	Singh [3], Yang [7]	Tan [4]	Jones [8]	Hasan [9], Zhou [10]	Huang [12]
単語削除	-	Query Relaxation Query Term Deletion	Query Term Dropping	-	-	Remove words
単語置換	Semantic Query Network	-	Query Term Replacement	Phrase Substitutions	-	Word Substitution
完全な書き換え	-	-	-		-	
誤り修正	-	-	-	-	Spelling Correction	Spelling Correction
その他	-	-	-	-	-	Word Reorder Add Words URL Stripping

を行った全ユーザー数のうち、ゼロマッチを経験したユーザーの割合は 22.3%であった。このことから、多くのユーザーが少なくとも 1 回はゼロマッチを経験していることが分かる。

ユーザー 1 人当たりが発するクエリ数を比較すると、ゼロマッチを経験したユーザーは、ゼロマッチを経験しなかったユーザーに比べて、平均で約 1/4 のクエリしか発していなかった。これは、ユーザーは一度ゼロマッチを経験すると、そのゼロマッチクエリを書き換えることを諦め、サイトから離脱してしまうことを示している。そしてその結果、ゼロマッチを経験したユーザーは商品の購入に至らないことが多い。このことは、ゼロマッチクエリを含まないセッション全体のうち、ゼロマッチを経験しなかったユーザーが最終的に商品を購入した割合と、ゼロマッチクエリを含むセッション全体のうち、ゼロマッチを経験したユーザーが最終的に商品を購入した割合を比較すると、前者の方が約 1.4 倍高かったことから明らかである。ここで、セッションとは検索クエリログをユーザ毎に分割し、同一ユーザーによって入力されたクエリで構成されるものを意味する。

次に、クエリに含まれる文字数と、単語の出現頻度という観点から、ゼロマッチクエリと非ゼロマッチクエリを比較する。ゼロマッチクエリの文字数の平均は、非ゼロマッチクエリの文字数の平均の約 1.3 倍であった。一方、ゼロマッチクエリに含まれる単語の出現頻度の平均は、非ゼロマッチクエリに含まれる単語の出現頻度の平均の 0.85 倍であった。つまり、ゼロマッチクエリに含まれる単語は非ゼロマッチクエリに含まれる単語よりも出現頻度が低い。これらの分析結果から、ゼロマッチクエリに対して、クエリに含まれる冗長な部分を削除する方法や、ゼロマッチクエリに含まれる出現頻度の低い単語を非ゼロマッチクエリに含まれる出現頻度の高い単語で置き換える方法が有効であると考えられる。

さらに、Yahoo!ショッピングを利用するユーザーの利用経験と、そのユーザーが発するクエリがゼロマッチとなる傾向についても調査した。ここで分析のため、過去 1 年間にわたり Yahoo!ショッピングで商品を購入していないユーザーを、新規ユーザーと定義する。ゼロマッチを経験したユーザーの利用経験を調べたところ、新規ユーザーによって入力されたクエリがゼロマッチクエリとなる割合は、全ユーザーによって入力されたクエリがゼロマッチクエリとなる割合の約 1.3 倍であった。このことから、新規ユーザーが入力するクエリは、ゼロマッチクエリになりやすいということが分かる。そして Singh ら [1] の調査によれば、新規ユーザーのクエリがゼロマッチとなってしまうのは、スペルミスに起因していることが多いという。そ

のため、ゼロマッチという問題を解決するためには、ユーザーによって入力されたクエリに含まれる誤りを修正する方法が必要である。

3.2 クエリ書き換え方法の類型化

ここで、ゼロマッチクエリの書き換え方法を類型化するため、ゼロマッチを回避するために書き換えたクエリ、すなわち書き換えクエリへのゼロマッチクエリの遷移を以下のように定義する。あるゼロマッチクエリ q が与えられ、同ユーザーセッション内で書き換えクエリ q' が生成されたとき、そのクエリ遷移は $q \mapsto q'$ と表すことができる。ここで、例えば英語で書かれたクエリは、スペースによって区切ることができるため、単語として見分けることが容易である。しかし、日本語で書かれたクエリに含まれる単語をどう定義するかは、研究者によって異なる。そこで本研究では、形態素解析によってクエリを分割することでできる形態素のことを「単語」と定義する。

以下のように本研究では、3.1 節の分析結果と、ウェブ検索あるいは E コマース検索におけるクエリ書き換え方法の先行研究 [3, 4, 6–10, 12, 14] をもとに、ゼロマッチクエリの書き換え方法の類型を提案する。また、我々が提案する書き換え方法の類型と先行研究における書き換え方法の対応を示したものを表 1 に示す。

(1) 単語削除

書き換えクエリ q' は、ゼロマッチクエリ q から任意の数の単語を削除したものである。すなわち、書き換えクエリに含まれる単語は、ゼロマッチクエリに含まれる単語の部分集合である。例：「スマホグリップ ひよこ」 \mapsto 「スマホグリップ」

(2) 単語置換

ゼロマッチクエリ q に含まれる単語を別の単語で置換すると、書き換えクエリ q' が生成される。ただし、 q と q' の間に少なくとも 1 つは単語の重なりが見られる。

例：「カシオ キーボード 脚」 \mapsto 「カシオ キーボード 台」

(3) 完全な書き換え

ゼロマッチクエリ q に含まれる単語を全て別の単語で置換することで、書き換えクエリ q' が生成される。(2) の「単語置換」と類似しているが、「完全な書き換え」の場合、 q と q' の間に単語の重なりが 1 つも見られない。

例：「enekeep」 \mapsto 「乾電池 式 モバイルバッテリー」

(4) 誤り修正

以下のように、ゼロマッチクエリ q に明らかな入力ミスあるいは不完全な単語が含まれている場合、誤り修正がおこなわれ、

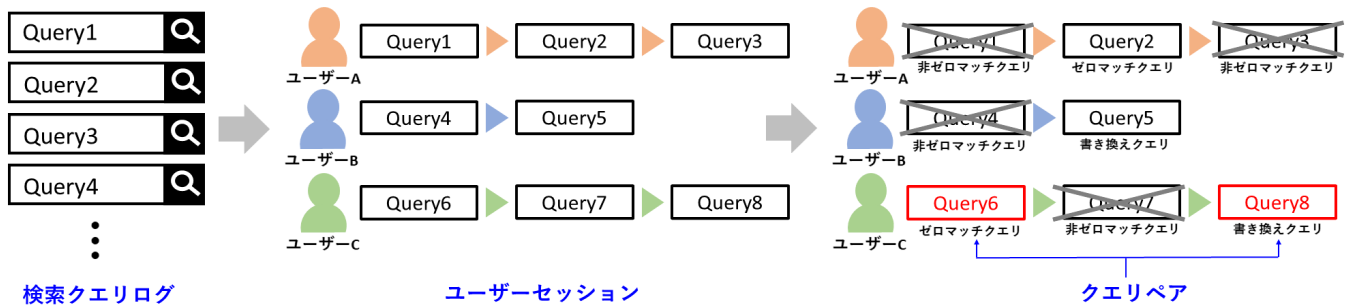


図 2 クエリペア作成の流れ

その結果書き換えクエリ q' が生成される。

- スペルミス, タイプミス
- 単語間のスペース入れ忘れ
- 入力途中で検索ボタンを押してしまった

例: 「サンフローン 除草剤」 \mapsto 「サンフーロン 除草剤」

例: 「edfir」 \mapsto 「edf ir」

例: 「dr.va」 \mapsto 「df.vape」

(5) その他

ゼロマッチクエリ q から書き換えクエリ q' への書き換え方法が, 上記 4 つの書き換え方法のいずれにも該当しない。例えば, ゼロマッチクエリ q に任意の数の単語を追加することで書き換えクエリ q' になる「単語付加」や, ゼロマッチクエリ q に含まれている単語の順番を変えることで書き換えクエリ q' が生成される「単語の順序変更」が挙げられる。

4. 評価実験

4.1 クエリペアの作成

ゼロマッチクエリと書き換え方法の類型とのペアからなるデータセットを構築する際, 3.1 節と同様に, Yahoo!ショッピングにおける 2ヶ月分の検索クエリログを用いる。

まず, データセットを構築する上で必要な, ゼロマッチクエリ q と書き換えクエリ q' が要素となるクエリペアを作成する。このクエリペアを作成する過程を図 2 に示す。クエリペア作成の第一段階として, 検索クエリログをユーザー毎に分割し, 同一ユーザーによって入力されたクエリで構成されるユーザーセッションを作成する。ここで, 長時間に渡るユーザーセッションでは検索意図が変わる可能性があるため, Liu ら [15] や Madvariya ら [16] の先行研究を参考に, ユーザーセッション内の最初のクエリが発せられてから 30 分経過したら, ユーザーセッションを打ち切ることにした。すなわち, 同一ユーザーによって発せられたクエリであっても, 30 分を超えて発せられたクエリは別のユーザーセッションとして扱う。

続いて, 作成したユーザーセッションの中から, ゼロマッチクエリ q と書き換えクエリ q' の両方を含むユーザーセッションを抽出する。ここで, 書き換えクエリ q' は, ゼロマッチが生じたユーザーセッション内で, ユーザーが商品を購入したときに発したクエリと定義する。ただし, 本研究ではゼロマッチクエリの書き換えに焦点を当てているため, 書き換えクエリ $q' \mapsto$

ゼロマッチクエリ q の順で発せられているユーザーセッションは除いている。

さらに本研究では, 一時的に商品の在庫がないことによるゼロマッチを考慮しないため, 抽出されたユーザーセッションに含まれるゼロマッチクエリ q は, 利用する検索クエリログの期間内で常にゼロマッチになるという条件を課す。また, クエリを発したユーザーの個人情報を保護するため, ゼロマッチクエリ q と書き換えクエリ q' は, ともに利用する検索クエリログの期間内で少なくとも 2 名の異なるユーザーから発せられるクエリでなければならないという条件を課す。

このようにして作成されたユーザーセッションの中から, ゼロマッチクエリ q と書き換えクエリ q' のみを抽出し, クエリペアを作成する。ここで, ユーザーセッション内でゼロマッチクエリが複数回発生した場合, セッション内で最初にゼロマッチとなったクエリをゼロマッチクエリ q として扱う。同様に, 書き換えクエリも複数回発生した場合, ゼロマッチクエリに最も近い時間に生成されたものを書き換えクエリ q' とする。

4.2 検索意図の判定

本研究では, ゼロマッチクエリ q と書き換えクエリ q' の検索意図が同じであるクエリペアのみを扱う。これは, 例えば「無洗米 あきたこまち \mapsto ボディソープ」のように, 完全に検索意図が変化している場合, ゼロマッチクエリを書き換えることは極めて難しいからである。一方, ゼロマッチクエリ q と書き換えクエリ q' の検索意図が同じかどうかを判定するために, Yahoo!クラウドソーシング^(注3)を使用し, 作業者に判定を依頼した。Yahoo!クラウドソーシングを用いて検索意図の判定作業を行う際に, 作業者に表示した設問画面を図 3 に示す。

検索意図が同じであるか否かを判定する基準を統一するため, 本研究では, ゼロマッチクエリ q により商品検索結果ページ 1 が取得され, 書き換えクエリ q' により商品検索結果ページ 2 が取得された場合の想定を作業者に依頼し, 「商品検索結果ページ 1 および 2 に共通の商品が含まれて欲しい」と作業者が回答した場合, クエリペアの検索意図は同じであると見なす一方, 「商品検索結果ページ 1 および 2 に, 共通の商品が含まれて欲しくない」と作業者が回答した場合, クエリペアの検索意図は異なると見なす。本実験では, 1 クエリペアあたり 3 名の作業者に判定を依頼し, 2 名以上が一致した意見を採用して, クエリペア

(注3) : <https://crowdsourcing.yahoo.co.jp/>

ショッピングサイトにおいて、検索キーワード列1により商品検索結果ページ1が取得され、検索キーワード列2により商品検索結果ページ2が取得されることを想定してください。そこで、商品検索結果ページ1, 2 に共通の商品が含まれてほしいか否かに応じて、上記ボタンのいずれかをクリックしてください。

キーワード列1	ピアニカ ホース
キーワード列2	カップラーメン

共通の商品が含まれて欲しい

共通の商品が含まれて欲しくない

【回答例】

(例1)
文字列1: 「マイティアCL業務用」
文字列2: 「マイティアCLセット」
→共通の商品が含まれて欲しい

(例2)
文字列1: 「テレビ台西洋アンティーク中古」
文字列2: 「ソニーワイヤレスイヤホン」
→共通の商品が含まれて欲しくない

また、以下のような場合は「共通の商品が含まれて欲しい」と回答してください。
・商品に関する制約(価格、量、素材)などが含まれている場合

(例)
文字列1: 「富士山の天然水24本最安値」
文字列2: 「富士山の天然水」

一方で、以下のような場合は「共通の商品が含まれて欲しくない」と回答してください。
・商品カテゴリ(例えば、「カバン」や「財布」)やブランド名(例えば、「ルイヴィトン」や「アディダス」)が同じであっても、文字列1と文字列2の商品検索意図が異なる場合

(例)
文字列1: 「アディダスリュック」
文字列2: 「ナイキリュック」

図3 検索意図判定作業の設問画面

アに含まれるゼロマッチクエリ q と書き換えクエリ q' の検索意図が同じであるか否かを決定する。ここで、単語区切りの解釈の違いにより、作業員間で意見が分かれてしまうことを避けるため、MeCab^(注4)を用いてゼロマッチクエリ q と書き換えクエリ q' をそれぞれ形態素解析し、スペースで区切ったものを作業員に提示した。辞書データにはmecab-ipadic-NEologd^(注5)を利用した。

3,438件のクエリペアに対しクラウドソーシングによる判定を行ったところ、ゼロマッチクエリ q とそれを書き換えたクエリ q' の検索意図が同じであったものは1,922件であった。

4.3 クエリ書き換え方法の判定

次に、4.2節で同じ検索意図を持つと判定されたクエリペアについて、3.2節において定義された種類のうち、どの種類に該当するかの判定を、再びYahoo!クラウドソーシングを用いて作業員に依頼した。このクエリ書き換え方法の種類を判定する際に、作業員に提示した設問画面を図4に示す。図4に示すように、4.2節と同様に単語の区切りを明示するため、ゼロマッチクエリ q と書き換えクエリ q' をそれぞれMeCabによって形態素解析し、スペースで区切ったものを作業員に表示した。

架空のショッピングサイトで、以下の2つの文字列を利用して商品を検索することを想定してください。

まず、単語の区切りを表すためにスペースで区切られた、文字列1と文字列2の2つの文字列が表示されます。そこで、「文字列1をどのように書き換えたら文字列2になるか」を考え、【2.手法の定義】に示されているような「単語削除」「単語置換」「完全な書き換え」「誤り修正」「その他」の5クラスの中から当てはまる手法を1つ選択してください。あくまで自分の主観でOKです。

キーワード1	Anker PowerCore Fusion 5000
キーワード2	モバイルバッテリー アンカー

単語削除

単語置換

完全な書き換え

誤り修正

その他

【回答例】

(例1)
文字列1: 「マイティア CL 業務用」
文字列2: 「マイティア CL セット」
→単語置換

(例2)
文字列1: 「十六 穀 ごはん」
文字列2: 「雑穀 米」
→完全な書き換え

【2.手法の定義】
<https://yahoo.jp/d9nZFY>

図4 ゼロマッチクエリ書き換え方法の判定作業の設問画面

作業員にはゼロマッチクエリ q と書き換えクエリ q' の両方を見せ、「 q をどのように書き換えたら q' になるか」という観点から、3.2節で定義された5つの類型うち、最もその類型の特徴を示していると考えられるものを1つ選択することを依頼した。その際、1クエリペアあたり5名の作業員に判定を依頼し、3名以上が一致した意見を採用して、クエリペアに含まれるゼロマッチクエリ q の書き換え方法の種類を決定した。

本研究では、4.2節でゼロマッチクエリ q とそれを書き換えたクエリ q' の検索意図が同じであると判定された1,922件のクエリペアから、作業員の3名以上の意見が一致しなかった392件のクエリペアを除いた1,530件を研究対象とした。そして、これらのクエリペアに含まれる各ゼロマッチクエリに書き換え方法の種類を付与し、ゼロマッチクエリと書き換え方法の種類とのペアからなるデータセットを構築した。このデータセットの内訳を表2に示す。

類型	ゼロマッチクエリ (件)	割合 (%)
単語削除	405	26.5
単語置換	438	28.6
完全な書き換え	261	17.1
誤り修正	404	26.4
その他	22	1.4

(注4) : <http://taku910.github.io/mecab/>

(注5) : <https://github.com/neologd/mecab-ipadic-neologd>

5. 考 察

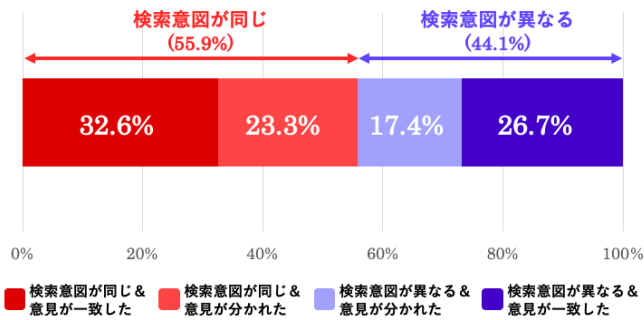


図5 検索意図判定タスクの実験結果

ここでは Yahoo!クラウドソーシングを用いた検索意図判定タスクの実験結果を考察する。このタスクで作業者が判定した結果を、「検索意図が同じか否か」と「作業者間で意見が一致したか否か」という2つの観点から分類した結果を図5に示す。図5が示すように、全クエリペアのうち44.1%のクエリペアが異なる検索意図を持つと判定された。4.1節で述べたように、本研究では検索意図の変化に備え、予めユーザーセッションを30分で打ち切った。それにもかかわらず、約半数のユーザーがゼロマッチ発生後に検索意図を変更していた。このことは、ゼロマッチを経験後、たとえサイトから離脱しなかったとしても、約半数のユーザーは今まで検索していた商品を探すことを諦め、別の商品検索に移ってしまうということを示唆している。

さらに図5から、「検索意図が同じである」あるいは「検索意図は異なる」といういずれの判定になった場合であっても、1クエリペアあたり3人の作業者の間で検索意図の判定について見解が異なることが多いことがわかる。そのため、どのような場合に作業者間で意見が分かれたかについて、検索意図が同じであると判定された場合とそうでない場合というそれぞれ2つの状況において調査を行った。まず、以下のような場合には、複数の作業者が「検索意図は同じである」と判定したが、「検索意図は異なる」と判定する作業者もいることがわかった。

- 入力誤りや変換誤りが含まれている場合
例：「セザム コンシーラー」→「ザセム コンシーラー」
例：「なちゅリズム」→「ナチュリズム」
- より幅広い情報を得るために単語を削除する場合
例：「iface adidas」→「iface」
例：「業務用スーパー トッポギ」→「トッポギ」

一方、以下のような場合には、複数の作業者が「検索意図が異なる」と判定しても「検索意図が同じである」という意見を持つ作業者がいることがわかった。

- 検索対象は変化しているが、同じ商品カテゴリである場合
例：「bmb-102」→「アンカーモバイルバッテリー」
例：「surisuri 化粧水」→「emaked まつげ美容液」
- ブランド名(店名)→商品 となっている場合
例：「のらぎや」→「もんぺ」
例：「villa lodola」→「オイル ヘアワックス」

以上のような検索意図判定タスクの結果に基づき、ゼロマッチクエリ q をどのように書き換えることで、ゼロマッチクエリとならない書き換えクエリ q' が生成されるかを理解するため、再び Yahoo!クラウドソーシングを用いて、作業者に4.3節で述べたようなクエリ書き換え方法の判定タスクを依頼した。

ここで、クエリ書き換え方法判定タスクの実験結果についての考察を行う。表2に示すように、「単語削除」「単語置換」「完全な書き換え」「誤り修正」の4つのタイプの割合が同程度であり、「その他」の割合が1.4%と極めて小さい。このことは、ゼロマッチクエリの書き換え方法として、上記4つのタイプが妥当であるということを示唆している。ここで、上記4つのタイプ以外に主要なゼロマッチクエリ書き換え方法があるかどうかを確認するために、「その他」と判定されたクエリペアについて、作業者によって記述された回答を調査した。その結果、「その他」のクエリ書き換え方法として、以下のような2種類のクエリ書き換え方法を回答している作業者が多かった。

● 単語追加・単語付加

例：「なむ あみ」→「なむ あみだ 仏」

例：「llbeen」→「llbean トートバッグ」

これらの例のように、ゼロマッチクエリ q に単語を付加したものが書き換えクエリ q' であるというクエリペアが観測された。しかし、これらのゼロマッチクエリは、入力途中であるにも関わらず検索ボタンを押下してしまったことやブランド名のスペルミスなどが原因である可能性がある。そのため、ゼロマッチクエリに単語を追加することでゼロマッチを解消することができたと考えるよりも、「誤り修正」がゼロマッチの解消につながったと考えるほうが妥当である。

● 書き換えなし・同一クエリ

例：「ハズキルーベ 最安値」→「ハズキルーベ 最安値」

これらの例のように、ゼロマッチクエリ q と書き換えクエリ q' が全く同じ文字列であるクエリペアが観測された。しかし、実際にユーザーによって入力されたゼロマッチクエリ q は「ハズキルーベ最安値」のようにスペースが挿入されていなかった。本実験では、クラウドソーシングで作業者にクエリペアを提示する際、ゼロマッチクエリ q と書き換えクエリ q' をそれぞれ MeCab によって形態素解析し、スペースで区切った。そのため、ユーザーによって入力されるスペースの有無の区別がつかず、2つのクエリが同一であり、書き換えが行われていないと判定作業者に判断されたと考えられる。

また表2から、「単語削除」「単語置換」「誤り修正」に比べ、「完全な書き換え」の割合が少ないことが分かる。このことから、ユーザーはゼロマッチを経験すると、元のクエリを保持しながらクエリを書き換える傾向があると考えられる。一方で、「cnr mj-108850」や「つるりんマスター」のようにゼロマッチクエリが商品の型番や商品名である場合、「完全な書き換え」によってクエリを書き換えるユーザーが多かった。さらに、「グラズレメディ」→「gran's remedy」のように文字種を変更し、ゼロマッチクエリ全体を書き換えることでゼロマッチを解消することができる場合があることも観測された。

4.3節で述べたように、本研究では、1クエリペアあたり5

名の作業者のうち3名以上が一致した意見を採用し、クエリペアに含まれるゼロマッチクエリ q の書き換え方法の類型を決定した。しかし、5名の作業者間の意見が2:1:1:1に分かれた場合、例えば2名が「単語削除」、別の1名が「単語置換」、もう1名が「完全な書き換え」、さらにもう1名が「誤り修正」と判定した場合についても多数決で類型を決定することができる。このように作業者間の意見が2:1:1:1に分かれたクエリペアは139件であり、これらをデータセットに含める場合と含めない場合にはそれぞれに利点がある。すなわち、これらのクエリペアをデータセットに含める場合、データセット内のデータ数が増えるため、機械学習を用いてゼロマッチクエリを分類する際に有効である。一方、これらのクエリペアを含めない場合、データセット内に誤ったラベルが付与されたクエリペアを含む可能性が低くなるため、ノイズが少ない、つまり信頼性の高いデータセットを構築することができる。しかし、件数が139件と少ないため、データセットに与える影響はそれほど大きくないと考えられる。

6. 追加実験

4.3節で構築したデータセットを用いてゼロマッチクエリの分類を行う。ただし、本実験では「その他」の類型を除いた4つの類型の分類を目的とする。

6.1 前処理

まずデータセットに含まれるゼロマッチクエリを2種類の手法を用いてベクトル化した。一つ目の手法では、ゼロマッチクエリが持つ言語的特徴量とゼロマッチクエリを発するユーザーの属性から特徴ベクトルの生成を行った。さらにもう一つの手法では、ラベル未付与の大規模コーパスから事前学習した汎用言語モデルであるBERT [17]、特に公開されている日本語学習済みBERTモデル^(注6)を用いてゼロマッチクエリの分散表現を獲得した。そして、それぞれの手法によってベクトル化されたゼロマッチクエリと書き換え方法の類型とのペア1,530件を、トレーニングデータとテストデータに9:1の割合で分割する。なお、汎用性能の高い分類器を得るために、層化10分割交差検証を行った。

6.2 特徴量

ここでは6.1節でベクトル化の際に利用した、ゼロマッチクエリが持つ言語的特徴量とゼロマッチクエリを発するユーザーの属性の一覧を表3に示す。

6.3 ゼロマッチクエリの分類

10種類の分類器を用いて、6.1節でベクトル化したゼロマッチクエリを「単語削除」「単語置換」「完全な書き換え」「誤り修正」の4つの類型に分類した。本研究では、機械学習を用いない分類器として、全てのテストデータに対して最も学習データ数の多いクラス、つまり「単語置換」と予測する分類器（以降「最多クラス」と表記する）と、4つの類型のうち無作為に1つの類型を選択し予測する分類器（以降「ランダム」と表記する）の2種類を利用する。さらに、機械学習ライブラ

表3 ベクトル化に利用した特徴量リスト

言語的特徴量	
単語数 [18]	ゼロマッチクエリに含まれる単語の総数
文字数 (合計) [18]	ゼロマッチクエリに含まれる文字数の総数
文字数 (平均) [18]	ゼロマッチクエリに含まれる単語の文字数の平均
文字数 (最初の単語)	ゼロマッチクエリの1単語目の文字数
単語の出現頻度 (平均, 最初の単語)	2ヶ月分の検索クエリログ内における単語の出現頻度
クエリの頻度 [18]	2ヶ月分の検索クエリログ内におけるクエリの出現頻度
品詞の数 [19]	ゼロマッチクエリに含まれる名詞の数, 形容詞の数
組織名の数 [19]	ゼロマッチクエリに含まれる組織名・ブランド名の数
文字種の割合	ゼロマッチクエリ全体におけるそれぞれの文字種の割合 文字種は、漢字, ひらがな, カタカナ, アルファベット, 数字の5種類
文字種の遷移	ある文字種からある文字種への遷移の回数 (例) 漢字 → カタカナ
ユーザーの属性	
新規ユーザーの割合	2ヶ月分の検索クエリログ内で、あるゼロマッチを発したユーザー全体における新規ユーザーの割合

リscikit-learn^(注7)を利用して実装したSVM (SVC), ロジスティック回帰, k近傍法, ナイブベイズ, 決定木, ランダムフォレスト, XGBoost, LightGBMの8つの機械学習モデルを用いてゼロマッチクエリを分類した。その分類結果を正解率 (Accuracy) で評価した値と、「ランダム」分類器と各分類器のペアに対してランダム化Tukey HSD検定 [20]を行った際の p 値を図4に示す。

表4に示す通り、機械学習を用いない場合、分類器の正解率は0.2933にすぎないが、機械学習モデルを用いた最も精度の高い分類器は0.4721の正解率を示した。さらに、ランダム化Tukey HSD検定を行った結果、ナイブベイズ以外の機械学習モデルを用いた分類器と「ランダム」分類器の差は、有意水準 $\alpha = 0.05$ において統計的に有意であることが確認された。このことから、ゼロマッチクエリの分類タスクに機械学習モデルを適用することは有効であるといえる。しかし、機械学習モデルを用いた分類器であっても分類正解率は50%に満たなかった。これは、この分類タスクの難易度が極めて高いということを示唆している。

さらに、最も正解率が高かった分類器、すなわち特徴量を用いてベクトル化し、LightGBMモデルを用いた分類器を対象に、各類型の分類難易度と特徴量の重要度について調査した。まず分類結果に基づいて各類型のF値 (F-score) を計算すると、「単語削除」は0.4882, 「単語置換」は0.4782, 「完全な書き換え」は0.2648, 「誤り修正」は0.5325であった。このことから、類型が「完全な書き換え」であるゼロマッチクエリを正しく分類することは、他の3類型に比べて難しいということが分かる。続いて、6.2節で示した各特徴量が分類の際にどれだけ寄与したかを調べた。その結果、Gini係数に基づく特徴量重要度が高い上位3つの特徴量は、ゼロマッチクエリの1単語目の出現頻度を表す「単語の出現頻度 (最初の単語)」, ゼロマッチクエリに含まれる全単語の出現頻度の平均を表す「単語の出現頻度 (平均)」, ゼロマッチクエリに含まれる文字数の総数を表

(注6) : <https://github.com/huggingface/transformers>

(注7) : <https://scikit-learn.org/stable/>

表 4 ゼロマッチクエリのカテゴリ結果

分類器	Accuracy	
	特徴量でベクトル化	日本語 BERT でベクトル化
最多クラス ランダム	0.2933 ($p = 0.712$) 0.2400	– –
SVM (SVC)	0.4436 ($p = 0.001$)	0.3866 ($p = 0.827$)
ロジスティック回帰	0.4237 ($p = 0.001$)	0.4065 ($p \approx 0$)
k 近傍法	0.3846 ($p = 0.003$)	0.4111 ($p = 0.049$)
ナイーブベイズ	0.3222 ($p = 0.987$)	0.3926 ($p = 0.028$)
決定木	0.4409 ($p = 0.001$)	0.3548 ($p = 0.003$)
ランダムフォレスト	0.4629 ($p \approx 0$)	0.4416 ($p = 0.001$)
XGBoost	0.4708 ($p \approx 0$)	0.4377 ($p \approx 0$)
LightGBM	0.4721 ($p \approx 0$)	0.4496 ($p \approx 0$)

す「文字数 (合計)」であった。

また、ゼロマッチクエリを 6.2 節で提示した特徴量によってベクトル化した場合と、日本語学習済み BERT モデルによってベクトル化した場合の分類精度を比較した。その結果、機械学習モデルを用いた分類器の過半数が、特徴量によってベクトル化したときに高い分類精度を示した。

7. 結論と今後の課題

本研究では、ユーザーが商品検索エンジンに適切なクエリを入力できず、ヒットした商品数がゼロの状況、いわゆるゼロマッチが生じたときのクエリを分析し、これらゼロマッチクエリを適切なクエリに書き換える方法を類型化した。さらに、ゼロマッチを回避するシステムを構築する上で必要となる、ゼロマッチクエリおよび書き換え方法の類型とのペアからなるデータセットをクラウドソーシングを利用して構築した。そして、構築したデータセットと機械学習モデルに基づく分類器を用いて、ゼロマッチクエリのカテゴリを行った。その結果、ゼロマッチクエリのカテゴリタスクにはゼロマッチクエリに含まれる文字数や単語の出現頻度といった特徴量が有効であるという知見が得られた。しかし、機械学習モデルを用いた分類器であっても分類正解率は 50% に満たなかった。それは本研究で利用したデータセットに含まれるデータ数が 1,508 件と少ないことに起因している可能性があるため、今後はさらに分類精度を向上させるためにデータ数を増やすことが不可欠である。その方法の一つとして、今後はユーザーが商品の購入に至ったクエリだけでなく、ユーザーがゼロマッチ経験後、目的とする商品の閲覧および商品をカートに入れた際に発したクエリもゼロマッチの解消に成功したクエリとして定義することが有効であると考えられる。さらに、この分類器を検索エンジンに導入し、自動的にゼロマッチクエリが書き換えられるようにすることで、ゼロマッチを回避し、ユーザーの検索行動や満足度の向上、さらに最終的には E コマースサイトにおける販売機会の損失の阻止につながる事が課題である。

文 献

- [1] Gyanit Singh, Nish Parikh, and Neel Sundaresan. User behavior in zero-recall ecommerce queries. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 75–84. ACM, 2011.
- [2] Nish Parikh and Neel Sundaresan. Inferring semantic query relations from collective user behavior. In *Proceedings of*

the 17th ACM conference on Information and knowledge management, pp. 349–358. ACM, 2008.

- [3] Gyanit Singh, Nish Parikh, and Neel Sundaresan. Rewriting null e-commerce queries to recommend products. In *Proceedings of the 21st International Conference on World Wide Web*, pp. 73–82. ACM, 2012.
- [4] Zehong Tan, Canran Xu, Mengjie Jiang, Hua Yang, and Xiaoyuan Wu. Query rewrite for null and low search results in ecommerce. In *Proceedings of the SIGIR 2017 Workshop on eCommerce (ECOM 17)*, 2017.
- [5] Subhadeep Maji, Rohan Kumar, Manish Bansal, Kalyani Roy, Mohit Kumar, and Pawan Goyal. Addressing vocabulary gap in e-commerce search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1073–1076, 2019.
- [6] Saurav Manchanda, Mohit Sharma, and George Karypis. Intent term selection and refinement in e-commerce queries. *arXiv preprint arXiv:1908.08564*, 2019.
- [7] Bishan Yang, Nish Parikh, Gyanit Singh, and Neel Sundaresan. A study of query term deletion using large-scale e-commerce search logs. In *European Conference on Information Retrieval*, pp. 235–246. Springer, 2014.
- [8] Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*, pp. 387–396. ACM, 2006.
- [9] Saša Hasan, Carmen Heger, and Saab Mansour. Spelling correction of user search queries through statistical machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 451–460, 2015.
- [10] Yingbo Zhou, Utkarsh Porwal, and Roberto Konow. Spelling correction as a foreign language. In *Proceedings of the SIGIR 2019 Workshop on eCommerce (ECOM 19)*, 2019.
- [11] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36, pp. 3–10. ACM, 2002.
- [12] Jeff Huang and Efthimis N Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 77–86. ACM, 2009.
- [13] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. A taxonomy of queries for e-commerce search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1245–1248. ACM, 2018.
- [14] Ahmed Hassan, Ryan W White, Susan T Dumais, and Yi-Min Wang. Struggling or exploring?: disambiguating long search sessions. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 53–62. ACM, 2014.
- [15] Zitao Liu, Gyanit Singh, Nish Parikh, and Neel Sundaresan. A large scale query logs analysis for assessing personalization opportunities in e-commerce sites. WSCD Workshop at WSDM '14, 2014.
- [16] Amber Madvariya and Sumit Borar. Discovering similar products in fashion e-commerce. In *Proceedings of the SIGIR 2017 Workshop on eCommerce (ECOM 17)*, 2017.
- [17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [18] Rohan Kumar, Mohit Kumar, Neil Shah, and Christos Faloutsos. Did we get it right? predicting query performance in e-commerce search. In *Proceedings of the SIGIR 2018 Workshop on eCommerce (ECOM 18)*, 2018.
- [19] Xiaobing Xue, Samuel Huston, and W Bruce Croft. Improving verbose queries using subset distribution. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 1059–1068, 2010.
- [20] Tetsuya Sakai. *Laboratory Experiments in Information Retrieval - Sample Sizes, Effect Sizes, and Statistical Power*. Springer, 2018.