

言及観点でファインチューニングしたBERTを用いた 聞き返しによるゲームのうろ覚え検索

楊 子正[†] 莊司 慶行[†] Martin J. D'ur[†]

[†] 青山学院大学理工学部情報テクノロジー学科 〒252-5258 神奈川県相模原市中央区淵野辺 5-10-1

E-mail: †jqwsxcfgS@sw.it.aoyama.ac.jp, ††{shoji,duerst}@it.aoyama.ac.jp

あらまし 本研究では、詳細を思い出せないゲームを、キーワードクエリと、聞き返してくるインタフェースを用いて再発見可能にする検索アルゴリズムを提案する。昔プレイしたゲームに対するあいまいな記憶は、自分で言葉にして検索することが困難である。一方で、「音楽の印象は華やかだったか」などの質問に答えることはできる。そこで、ゲームに対するレビューを収集し、レビュー中の文が「物語」「キャラクター」など、どの観点に言及しているか分類した。そのために、Wikipedia から収集したゲームに関する構造化された記事を用いて、BERT を言及している観点到に敏感なベクトル化を行えるようにファインチューニングした。そして、観点到ごとにレビュー文を提示し、当てはまるかを答えることでゲームを絞り込んでいく検索インタフェースを作成した。実際のオンラインレビューサイトから収集した 30 年分のゲームに対して、実際にうろ覚えのゲームを探させる被験者実験を行った。実験結果から、検索精度と実用性について議論する予定である。

キーワード 情報検索, ソーシャルメディア, レビュー, 機械学習

1 はじめに

ビデオゲームが誕生してから、すでに 40 年が経過した。ゲームは我々の生活にすでに溶け込みつつあり、物心つく前からゲームをプレイしていた人たちが、いまでは大人になっている。人々の幼少期の記憶に、感動的なゲームや、燃えるゲームや難しいと思ったゲームなどが、深く残るようになっていく。こういって、幼少期に体験したゲームを再度遊びたいという要求は、一般的である。

たとえば、過去、「ポケットモンスター」というゲームに熱中していた人は、豊富なモンスターの種類が一番印象強いであろう。しかし、単純に「モンスター」をキーワードとして検索すると、現状の検索エンジンでは、近年のゲームしか結果に登場しない。もしはっきりモンスターの名前など特有な情報を覚えていたなら、その名前でポケモンというゲームが検索できる場合もある。一方で、ゲーム内の小さなサブイベントについてしか覚えていない人の場合を考える。例えば、過去のポケットモンスターにおいて、週に 1 回、ロケットが飛んだことを回数として教える、ゲームの本筋に関係のない小さなイベントが存在した。ところが、このエピソードだけ覚えていた人が、「ゲームロケット 毎週」などとキーワードを入力しても、「ポケットモンスター」は検索結果に登場しない。ロケットが直接登場する宇宙戦争のゲームや、タイトルにロケットという単語を含むゲームが多く発見される。

このように、人々のゲーム体験が個人化している中で、個人の体験や印象を基に自分が探したいゲームのうろ覚え検索をすることは困難である。現在の検索は主に、クローリングによって作られたインデックスから、入力されたクエリとマッチング

して最適な検索結果を出している¹。通常の検索アルゴリズムは、うろ覚え検索のこのような状況に、適していない。また、検索対象のデータとして、ゲームに対する印象は、一般のウェブページから作成したインデックスの中には含まれない場合が多い。例えば存在するとしても、ゲームは映画など受動的なメディアと違い、個人の体験によって違いが生み、前書したロケットの例のように、同じところでも、まったく異なるインデックスに含まれることもある。

加えて、ゲームという分野は、人によって検索対象に対する体験や印象が大きく異なる。異なるゲームでも、ジャンルが近いなら似たような体験ができるが、中身の細かく違いによって異なる体験ができることもある。例えば、シューティングゲームは、銃を発砲して敵を倒すことがすごく爽快だという体験があるが、細かいところが異なると、Doom の場合、地獄の怪物を倒したのがすごく爽快で、それが CS の場合、相手をヘッドショットで一発で倒したところが爽快である。このように、入力時ただ「敵を倒すと爽快感がある」と検索すると、従来の方法では当てているものをすべて表示し、ユーザに選べるしかできない。検索用のクエリがもっと曖昧な時、ユーザはそれを区別できなくなる。

そこで本研究では、観点到に敏感な BERT モデルと、聞き返しによってゲームを絞り込める検索インタフェースによって、ゲームのうろ覚え検索を行える手法を提案する。人々がうろ覚えしている際の体験や印象は、「ロケットを飛ばすところが好き」、「ボス戦が難しい」という表現が多い。このような文書は、一般的なゲームに関する文書よりも、むしろレビューの内容に

¹ : Google 「検索の仕組みの基本」:<https://developers.google.com/search/docs/basics/how-search-works?hl=ja>

近い。その上、似たような体験とレビューをマッチすれば、レビューされたゲームがユーザが検索したいゲームと言える。

そこで、自然言語処理モデルである BERT を使って、検索者の入力した体験を表す短文と、ゲームに対するレビュー中の文の類似度を計算可能にする。この際に、本手法では、BERT の言語モデルを、観点に敏感になるようにファインチューニングした。これは、レビュー文に表れるトピックや単語の近いレビューを見せられるよりも、むしろ「音楽に言及したレビュー文」や「プレイ感覚に言及したレビュー文」のように、入力した文と近い観点に言及するレビューを読んだほうが、自分の記憶と一致しているかの判断がしやすいと考えたためである。

加えて、あるゲームの特徴を、レビュー文を見ることで思い出しながら、対象となるゲームを絞り込んでいくような、聞き返しに基づく対話型検索インタフェースを提案した。この検索インタフェースでは、クエリを 1 回入力した後、検索結果を見ながら、さらにクエリを追加入力したり、一覧として表示されたレビューから当てはまるレビューを選ぶことができる。こうすることで、自分の記憶に当てはまらないゲームを、除外していくことができる。これは、自分自身では、うろ覚えの体験を言葉で表せなくても、他人のレビュー文に対して「当てはまる」か「当てはまらない」かを判断することは可能だと考えたためである。このようなインタフェースを持つ検索システム上で、検索者はクエリを入力し、ゲームを絞り込んでいける。このように、観点に敏感な言語モデルを用いて、聞き返しによって絞り込むことで、思い出せる記憶が少ない状態からでも、昔プレイしたゲームを再発見できる。

聞き返しができるシステム、すなわち対話型システムを検索に使う場合、結果を表示するには、従来の手法と違い、クエリがあたる部分は埋めるずつあると考えられる。しかし、普通の BERT ならトピックに敏感であり、そのまま使うと同じところに埋めることが発生すつことが考えられる。そこで我々は BERT を観点敏感にファインチューニングする。観点に敏感なら、埋めるときちゃんとクエリを分けて埋めることができ、対話型システムの精度も上がる。

このような対話的インタフェースを有するうろ覚え検索システムの有用性を確認するために、実際に与えられたゲームについて、固有名詞などを使わずに検索する被験者実験を行った。そのために、実際に絞り込み検索システムを、ブラウザからアクセス可能なウェブアプリケーションとして実装した。評価実験から、BERT をファインチューニングすることで最初に入力したクエリとレビューのマッチングにおいて、より正しいゲームが検索結果上位に出てきやすくなったことが確認された。また、聞き返しによる絞り込みも、ある程度正しくゲームを絞り込めることが分かった。

本論文の構成について述べる。本章では研究の概要と背景について説明した。第 2 章では本研究の関連研究について、ゲームの言語構成などについての研究を紹介する。第 3 章では、BERT の観点に敏感なファインチューニングと聞き返しができるシステムの構成について説明する。第 4 章では提案手法と比較と有用性実験とその結果について説明すし、第 5 章で実験結

果をもとに本研究の貢献を議論する。第 6 章で結論と今後のま

とめについて述べ、まとめる。

2 関連研究

本研究は、BERT による検索と対話型システムの組み合わせでうろ覚え検索を実現する。そのための既存の関連研究をここに説明し、本研究の位置付けを示す。

2.1 BERT による検索

BERT (Bidirectional Encoder Representations from Transformers) は 2018 年に Delvin ら [1] によって開発された言語モデルである。この言語モデルは、言語を理解するタスク、すなわち自然言語処理において高い精度であることが知られている。BERT は一般的な自然言語処理タスク用に開発されたが、情報検索分野でも利用可能である。

一般的な BERT は汎用的な言語モデルであり、用途に合わせて調整することができる。このような調整をファインチューニングと呼び、例として Sun ら [2] BERT を文章分類に適してファインチューニングする方法を発表している。本研究では、こういったファインチューニングの例に倣い、ゲームに関する文書の見出しを推定するモデルで BERT をファインチューニングし、観点に敏感なベクトル化を可能にした。

BERT はベクトル空間モデルに基づく手法であり、文書や単語を分散表現と呼ばれるベクトルで理解する。分散表現によるモデルは、特に、ニューラルネットワークと組み合わせで用いられる。ニューラルネットワークによる分散表現理解は、山田ら [3] が提案した。これによって、分散表現を比較をすれば、情報の検索はできることがわかった。そこで、Yilmaz ら [4] は、BERT が持つ高い言語への理解力を使って、Birch というシステムに組み込んで、類似文書の検索ができるようにしている。本研究では BERT を使って、ユーザの入力したクエリと文書をベクトル化し、コサイン類似度に基づいて絞り込み検索を可能にしている。

2.2 対話型情報検索システム

対話型システムは、大まかに 2 つの部分に分けられる。一つはユーザの入力を理解する部分、一つは出力を入力によって自動生成する。入力の部分は従来、ラベル付けにより入力の認識はできるが、ラベル付けのコストが高い。そこで、機械学習を導入することで、より早く対話型システムを開発することができる。Henderson [5] は、機械学習における対話型システム、主に対話能力におけるトラックと評価を分析した。これをもとに、Wen ら [6] は、対話型システムを簡単に作れるアルゴリズムを提案している。このアルゴリズムでは、ニューラルネットワークを用いることで、人手でのラベル付けを排している。こうして作られた対話式検索システムの有用性をレストランの発見タスクによって示している。より深いニューラルネットワークを使った対話システム作成も一般的になりつつある。例として、Li ら [7] は、ディープラーニングを使うことで、より長い対話を可能な対話型システムを提案している。本研究もこうし

た対話的システムと同様に、ユーザの入力に合わせながら選択肢を提示して、絞り込みを行う。

Gao ら [8] は、対話型システムを主に 3 つの使い方を分類した。ここで言う 3 つの使い方とは、すなわち、質問への回答をマシンにやらせる、特定のタスクに向ける対話をマシンにやらせる、チャット用のロボットである。これらの学習方法をすべて評価することで、現在の対話型システムの発展とこれからの課題を分析した。これは本研究においてシステムの開発に大きい影響を与えた。

2.3 うろ覚え検索 (Known-item Search)

うろ覚え検索、すなわちすでに知っているものの再発見は、「Known-item search」などと呼ばれ、情報検索分野において重要な課題の一つである。幼少時にプレイした、詳細の思い出せないゲームの再発見も、うろ覚え検索の一種である。このような分野で再発見を助けるためには、ゲームについて話している文章が、どのような語彙と文脈を有しているか、システムが理解している必要がある。象徴的な例として、Aarseth ら [9] はゲームの語彙の類型を研究し、ゲームの言葉が次元、すなわち観点に深く言及していることを示唆している。本研究はこの研究を参考に、観点に基づく検索アルゴリズムを提案している。

従来、ゲームを探すこと自体が難しい、他のものと違い、万能だと言えるゲームの検索アルゴリズムは存在していない。Bogers ら [10] は Reddit からゲームの複雑なリクエストをもらい分類し、そのうち 5 つのところがゲームの特徴的なものと判明した。

ゲームの再発見には、ゲームを探している人たちがどんな情報をもって探しているのをわかる必要がある。Zhang ら [11] は、ゲームの中のあるシーンを人の言葉で検索できるようにした。我々の研究において、すでに知っていることを検索するではなく、もっと曖昧なものについて検索ができるのは課題である。Anderson ら [12] は、ゲームの各領域の人がゲームに対する話を分析し、ゲームの検索用のクエリの特徴をまとめた。本研究では、レビューした人も様々な職業を持っているため、この論文も参考になった。

以上の研究を踏まえて、Bogers ら [13] は、ゲームの再発見に関して、Reddit のあるチャンネルのデータを分析し、ゲームの再発見に関する主なクエリをまとめた Milier ら [14] は、Reddit を分析し、今の検索エンジンに何を登録するともっと特定分野のものを検索できるようになるという結果をまとめた。本研究に関する観点の部分、これらを参考にした。

3 提案手法

本研究では、BERT を観点敏感にファインチューニングし、聞き返しができるシステムと組み合わせてゲームのうろ覚え検索のアルゴリズムを提案する。実際の手順としては、

- (1) BERT のファインチューニング,
- (2) BERT によるゲーム特徴の作成,
- (3) BERT による対話型システムの作成

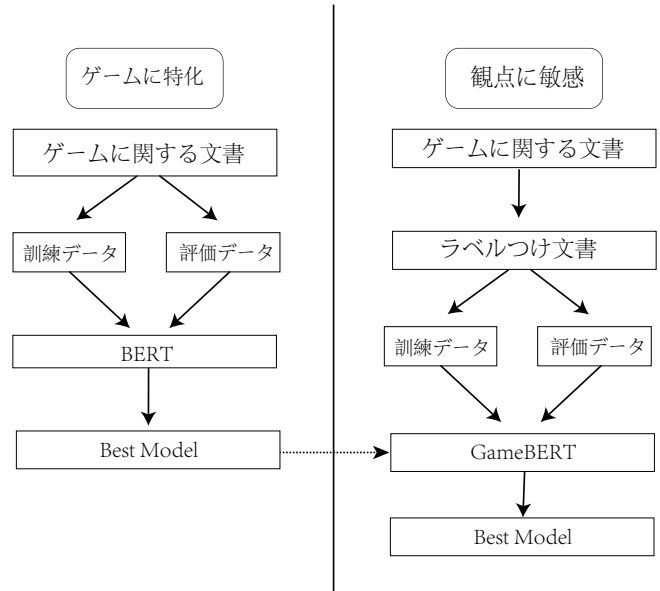


図 1 BERT ファインチューニングのアプローチ

という主に三つのところから構成されている。この際、ユーザの入力の相関度はすべてコサイン類似度により計算される。

以下は今回のアルゴリズムについて詳しく説明する。

3.1 BERT のファインチューニング

BERT のファインチューニングでは、Google が開発された BERT をベースに、ゲームに特化し、観点に敏感するようにファインチューニングする。本研究では、ゲームに特化するにはモデルを新しい単語を辞書に追加するファインチューニングタスクを使い、観点に敏感するにはマルチラベル分類タスクを使う。図 1 が示すように、BERT のファインチューニングは二つの段階に分けている。まずは基の BERT をゲームの表現をわかるようにゲームに特化する。次は特化した BERT を更に観点に敏感するようにファインチューニングする。

BERT をゲームに特化するには、BERT がゲームの表現を辞書リストに追加すればいい。ゲームに関する表現を予め収集して、それを文ごと分割する、BERT では、入力の長さは最大 512 しかないため、512 を超えた文書を全部捨てる。これを訓練データと評価データに分けてモデルに与えてファインチューニングする。BERT のようなあらかじめトレーニング済みのモデルは、特定分野に適する場合、その分野の言葉をファインチューニングされていれば高精度がでる。SciBERT [15] はこのファインチューニング手法の根拠になっている。そこで、評価の精度の一番いいモデル、Best Model だけ残し、追加学習完了したモデル、GameBERT とする。

GameBERT を用いて観点に敏感するには、ラベル付きのデータを使ってファインチューニングする。同じデータなら過学習することが起こりえるため、追加学習で使うデータと異なるデータを使いラベルをつける。人工的につけることもできるが、今回はコストの原因でセクション付きの Wikipedia データを使用した。そこで、同じように訓練データと評価データに分け、GameBERT にトレーニングされる。今回使用するタスク

はマルチラベル分類タスクであり、文書一つにつき、多数のラベルに属することができる。ゲームの文書は常に一つの観点について話しているのではなく、また同じような文書でも、違う観点に属することもある。多くの観点をモデルをファインチューニングすると、自然にそのような状況が対応できると研究で発見され、このようなファインチューニング手法を使う根拠になった。評価データによる精度評価にあと、追加学習のように、BestModel だけ残し、観点に敏感な GameBERT とする。

3.2 BERT によるゲーム特徴の作成

BERT によるゲームの特徴の作成では、前書した観点に敏感な GameBERT を使って、ゲームのレビューの分散表現を得る。観点ごと違うレビューは一つのゲームに対してとても多く、それをまとめる必要がある。

ここで、GameBERT の結果によってまとめるではなく、クラスタリングを使用する。ファインチューニング用の観点ラベルはすべての観点に比べたら遥かに少ない可能性ある、分類されていないレビューも、GameBERT はすでにその観点の違いが分かって、ただ今持っているラベルでは定義できない。そこで、クラスタリングを使って、このようなレビューも分けることができる。今回は、K-Medoids というクラスタリング手法を使用した。K-Medoids は K-Means と似ていて、空間上距離が近いものを一つのグループにする。しかし、K-Means の中心は必ずしもレビューそのものではない、本研究においてゲームの特徴は、すべて実在するレビューを使うため、グループの中心位置が実在するレビューを使う K-Medoids が適していると考えられる。またノイズになるレビューに対して、K-Medoids が K-Means よりロバスト性がいいため、本研究では K-Medoids を使用した。K-Medoids を用いて、GameBERT が出力したレビューの分散表現をクラスタリングし、クラスタの中心をゲームの特徴とする。これを事前に取得したゲームのメタデータとともに、ゲームの検索用のインデックスが作られる。

3.3 BERT による対話型システムの作成

BERT による対話型システムの作成では、ファインチューニングした GameBERT を導入して対話型検索システムを作成する。システムをクライアント側とサーバ側に分けられ、クライアント側はユーザのクエリの入力を受け、それをサーバに転送して、サーバ側はそれに基づいて、ゲームの特徴を提示する上、ゲームの関連ポイントをもとに、ランキングを出す。

図2のように、検索が始まる前に、クライアント側はゲームのメタデータを表示する。メタデータはゲームの再発見において一番の多く使われている情報だと Bogers ら [13] が研究で判明した。ユーザは自分が探したいゲームの情報をメタデータとともに入力し、それをサーバ側に転送する。サーバ側はそれをキャッチしたら、まずメタデータをもとに、各ゲームとの関連ポイントを計算する。メタデータはかなり明白な情報なので、GameBERT を使用する必要がなく、ユーザが入力しているメタデータが各ゲームに含まれる数が多くほど、ゲームの関連ポイントが高くなる。今回使用しているメタデータは三つある：

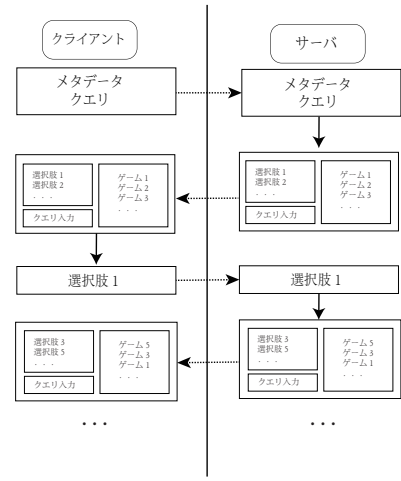


図2 対話型システムのアプローチ

- (1) ゲームのプラットフォーム, platform とする,
- (2) ゲームの発売時期, release date とする,
- (3) ゲームのジャンル, genre とする.

ここで、メタデータによる関連ポイントの算出式は

$$S_{Meta} = S_{platform} + S_{releasedate} + S_{genre} \quad (1)$$

であらわされる。そこで、ユーザが入力したクエリをすべてのゲームの特徴と比較、ここではファインチューニングした GameBERT を使用して、クエリ q の分散表現を得る。これをどのゲームのどの観点に近いかを判定する。ここでは、コサイン類似度

$$\cosim(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (2)$$

を使って観点 v の近さを算出し、この類似度は

$$S_{query} = \cosim(q, v) \quad (3)$$

であり、これを各ゲームの関連ポイントとする。この二つの関連ポイントをかけると、最初の検索で各ゲームがユーザが探したいゲームがどれくらい似ているかの得点

$$S_{game} = S_{Meta} \cdot S_{query} \quad (4)$$

がわかる。これをもとに、サーバ側はゲームのランキングも計算でき、得点が高い順でゲームをランキングする。また、ランキング順に基づき、各ゲームの他の特徴をクライアント側に選択肢として表示する。ここで表示する特徴は、各ゲームの関連ポイントの最大値を示す特徴以外の特徴をランダムに選んでいものである。

以上のサーバ側の情報を用いて、クライアント側がゲームのランキングと選択肢と追加クエリ入力用のところも表示する。もしユーザの記憶に選択肢にない体験が存在したら、それを入力することで追加で検索でき、また選択肢を選んでも、追加検索もできる。追加検索用のクエリをもらったら、サーバ側は同じように、コサイン類似度を使って、各ゲームの新しい関連ポイントを算出、ここで、新しい関連ポイントをもとの関連ポイントをかけることで、新しい得点

$$S_{new} = S_{game} \cdot S_{newgame} \quad (5)$$

を計算する。これを用いて、ゲームのランキングを変えることで、またクライアント側に表示する。

これを繰り返すことで、最終的にユーザが探したいゲームと似たいようなゲームはランキングの上に来て、ユーザはゲームのタイトルを自分で確認することで、自分が探したいゲームが見つかることが可能とする。

4 評価実験

本節では、被験者実験を通じて、提案するうろ覚え検索手法の有用性を示すための実験結果と実験用のデータセットについて述べる。次節以降は、評価実験の詳細について述べる。第 4.1 節では、実験に使ったデータセットについて述べる。第 4.3 節では、実験用のシステムの実装について述べる。第 4.7 節では、実験の手順について述べる。第 4.10 節では、実験結果について述べる。

4.1 データセット

本実験を行うにあたって、実際にゲームをうろ覚え検索可能なウェブアプリケーションを作成した。本評価実験において、ゲームに関するレビュー情報として、レビューサイト Game Spot から情報を利用した。ゲームに関する観点情報を学習するために、Wikipedia からゲームカテゴリに含まれる記事を利用した。

レビュー情報として、Game Spot に約 100,00 本のゲームの情報と、それらの約 10,000,000 件のレビューを収集した。ゲームからは必要のメタデータを収集した。このゲームのレビューは、評点がついてない、ゲームのレビューを含む投稿と示す。

観点情報として、Wikipedia のゲームカテゴリから年代順で約 15,000 本ゲームのページから約 20,000,000 万件の記述を収集した。

4.2 学 習

本研究では、BERT を観点に敏感するため、wikipedia のセクションの下の記事を用いて、記述、セクションのような組み合わせで学習した。対象になるデータは、図 3 に示したように、wikipeida 上のセクション内の記述を学習用のデータとした、ここで、各記述の観点はサブセクションとセクションの組み合わせにした。

はじめに、サブセクションとセクションがついていないデータを BERT に学習した。この学習で得られた一番精度が高いモデルを GameBERT とした。次に、サブセクションとセクションを記述につけて、観点付きデータとして BERT をファインチューニングした。精度が一番いいモデル観点に敏感な GameBERT とした。

4.3 アプリケーションの構成

本実験を行うにあたって、実際にゲームをうろ覚え検索可能なウェブアプリケーションを作成した。次に、アプリケーションの構成を述べる。



図 3 Wikipeida のデータの説明図

4.4 アプリケーション用のデータセット

本研究では、検索の高速化の研究ではないため、検索時の速度を考慮していない。被験者が検索するとき長い時間かかっていたらやる気が減るとを考慮して、収集された 10,000 件のゲームのうち 3,000 件を使用した。使用したゲームのレビューを、観点に敏感な GameBERT でベクトル化した。レビューのベクトルを、K-Medoids でクラスタリングし、ゲームごと 20 個のクラスタを生成した。生成したクラスタの中心となるレビューを取り出し、ゲーム検索時の比較用レビューとした。レビュー数が 20 以下の時、レビューをそのまま使った。

4.5 比較手法

本研究の技術的課題点として、観点に敏感な BERT ベクトライザを用いる点、対話的インタフェースを通じてクエリを質問への回答として入力する点がある。そこで、これら 2 つの工夫のうち、どちらがどの程度検索に貢献したかを確認するために、4 つの手法とひとつのベースライン手法を用意した。しなわち、

- **提案手法** ファインチューニングと対話型 UI の両方を含む有名詞などを用いずに検索する手法、
- **FT ありクエリ** ファインチューニングのみを含む手法、
- **FT なし対話** 学習済み BERT でクエリと対話型 UI で検索する手法、
- **TFIDF** ベースラインとして、単純なベクトル類似度、の 4 手法である。

4.6 検索結果の表示

アプリケーションの入力はクエリとして、第 3.3 節で述べた方法で、ゲームのランキングと質問を出した。一回の検索後の結果は図 4 で示す。左側に手法ごと出した最高点のゲームタイトルが出て、次に選択肢を出す。ここに出している選択肢は、クエリと異なる観点から出すように、各ゲームのレビューがクエリとの類似度一番低いレビューにした。出す順は、ゲームのランキングと同じ順にした。もし選択肢がどれも当てはまっていない場合、他の選択肢を表示し、または自分で新しいクエリを入力することもできた。右側は、ゲームのランキングの上位

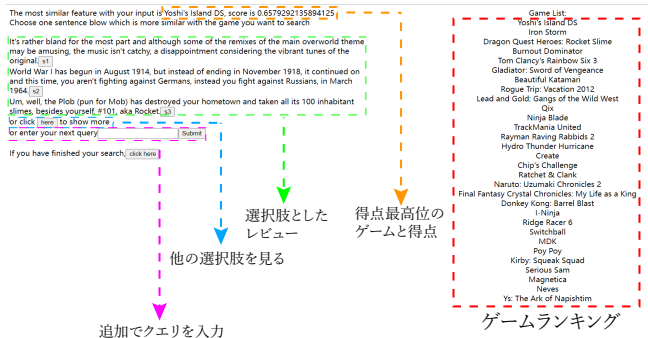


図 4 アプリケーションの結果の詳細

表 1 実験用ゲームタイトルと販売本数

ゲーム	販売本数
Call of Duty: Black Ops	26,200,000
Super Mario World	26,662,500
Pokémon Gold / Silver / Crystal	29,490,000
Grand Theft Auto: San Andreas	27,500,000
Half-Minute Hero	不明
SPQR: The Empire's Darkest Hour	不明

30 件を示した。ここに検索したいゲームがあるなら、左下のボタンを押して検索を終了した。

4.7 実験タスク

本研究では、被験者を集め、比較手法をローテーションして実験した。

4.8 実験用ゲーム

本研究では、各被験者に比較手法を順番いで同じゲームを検索するように実験した。検索用のゲームは表 1 に示す。Wikipedia から累計販売本数 2,500,000 万件以上のゲームを見て、アプリケーションのデータセットに存在するゲームを使った。そこで、マイナーなゲームやレビュー数が少ないゲームも存在するので、提案手法が検索できるかの実験もした。マイナーなゲームは販売本数が登録されていないゲームを 1 本を使った。レビューが少ないゲームがアプリケーションのデータセットから登録されているレビューが 10 件以下のゲーム 1 本を使った。

4.9 実験手順

本節では、実験の手順を述べる。本研究では、メタデータを含む実験する予定があった。しかし、予備実験ではメタデータを使うとゲームがかなり絞られてほぼすべての手法が検索できるようになって、手法の性能が評価できなかった。本番の実験では、メタデータをはじめて実験した。

表 1 で示したゲームを基いて、第 4.3 節で述べたアプリケーションを用いて実験プランを作った。表 2 に実験の概要を示す。被験者にあらかじめ実験用の ID を用意した。アプリケーションでは、ID で表 2 の被験者の順番で比較手法の順番を実装した。被験者が ID を入力した後、メジャーゲーム 1 から 4、マイナーゲーム、レビューが少ないゲームの順にゲームタイトルを提示した。第 4.3 節で述べたアプリケーションを用いて、検索

した。我々は、各手法ごとに見えてきたかどうか、発見までにかかった検索回数、入力したクエリと選択を用いて、各手法の評価をした。

4.10 実験結果

本節では、各手法ごとに発見までにかかった検索回数、実際に入力されたクエリや選択されたレビュー文について述べる。

4.11 各手法の検索成功した回数と検索にかかった回数

被験者に対して検索できた結果について説明する。表 3、表 4 は、各手法でゲームを検索できた結果と検索にかかった回数の結果を示す。ここで、被験者が各手法において、どれぐらいゲームを検索できたかを示す。各手法において、メジャーゲームがそれぞれ 4 回検索した。マイナーゲームとレビューが少ないゲームは、提案手法だけで 4 回検索した。被験者が入力できるクエリと回数に制限がなかった。ここで、メジャーゲームにおいて、提案手法は最高の 3 回まで検索できた、マイナーゲームも一回検索できた。検索が成功時、メジャーゲームにおいて、提案手法が平均 1.7 回検索した。すべてのゲームにおいて、提案手法は平均 1.4 回検索した。検索が失敗時、メジャーゲームにおいて、提案手法は平均 3 回検索した。すべてのゲームにおいて、提案手法は平均 2.2 回検索した。

4.12 実際に入力したクエリと選択されたクエリ

本節では、実験結果において、被験者が入力したクエリと選択されたクエリを述べる。表 6 と表 8 と表 7 はそれぞれ成功時と失敗時のクエリとゲームのランキングを示す。ここで、ゲームのランキングと各手法において、得点がもっとも高いレビューを上位 30 件の順で表す。実験時、被験者が見た検索結果と同じような結果をした。

5 考察

本節では、実験結果をもとに、提案手法の有用性や性質について議論する。はじめに、検索の成功回数とにかかった回数について、考察する。第 4.11 節に示すように、今回の実験において、全体的検索する回数が少なかった。特に検索失敗したとき、メジャーゲームにおいて、平均 3 回ぐらいしか検索していなかった。被験者が今回、すべてのメジャーゲームに対し、シングルモードをクリアした前提で募集されていた。

このような少ない検索回数になった原因の一つに、今回のインターフェースがあまり利用者にとって使いやすくなかったことが考えられる。本研究において改善すべきと考えられる。また、被験者から感想を聞いた結果、一回検索するために、およそ 5 分程度の時間を要することが判明した。特に TFIDF の手法は、一回の検索につき 10 分程度を要した。

本研究では、特に検索の高速化を考慮していないが、検索の時間も被験者の検索する熱意に影響したことが考えられる。提案手法を実用化する前に、検索の速度を上げる研究もする必要がある。

検索対象となるゲームについて考察すると、マイナーなゲー

表 2 実験手順

	メジャーゲーム 1 Call of Duty: Black Ops	メジャーゲーム 2 Super Mario World	メジャーゲーム 3 Pokémon Gold / Silver / Crystal	メジャーゲーム 4 Grand Theft Auto: San Andreas	マイナーゲーム Half-Minute Hero	レビューが少ないゲーム SPQR: The Empire's Darkest Hour
被験者 1	提案手法	FT ありクエリ	FT なし対話	TFIDF	提案手法	提案手法
被験者 2	TFIDF	FT なし対話	提案手法	FT ありクエリ	提案手法	提案手法
被験者 3	FT ありクエリ	提案手法	TFIDF	FT なし対話	提案手法	提案手法
被験者 4	FT なし対話	TFIDF	FT ありクエリ	提案手法	提案手法	提案手法

表 3 各手法でゲームの検索が成功した際の回数

	提案手法	FT ありクエリ	FT なし対話	TFIDF
メジャーゲーム	3	2	1	0
マイナーゲーム	1	-	-	-
レビューが少ないゲーム	0	-	-	-
成功率	0.67	0.50	0.25	0.00

表 4 検索成功時各手法にかかった平均回数

ゲーム	提案手法	FT ありクエリ	FT なし対話	TFIDF
メジャーゲーム	1	2	3	-
マイナーゲーム	1	-	-	-
レビューが少ないゲーム	0	-	-	-
平均	1	2	3	-

表 5 検索失敗時各手法にかかった平均回数

ゲーム	提案手法	FT ありクエリ	FT なし対話	TFIDF
メジャーゲーム	3	2.5	3.3	2
マイナーゲーム	2.2	-	-	-
レビューが少ないゲーム	1.5	-	-	-
平均	2.2	2.5	3.3	2

表 6 成功した検索タスクにおけるクエリ “The story line is scripted and explained.” に対するゲーム対象となったレビュー一文 (Call of Duty: Black Ops が正解のゲーム)

順位	ゲームタイトル	レビュー	スコア
1	Call of Duty: Black Ops	The story line is greatly scripted and well explained.	0.98
2	Star Wars: Masters of Teras Kasi	The story line is kept rather short and simple.	0.94
3	Tactics Ogre: Let Us Cling Together	FFT is linear and therefore very involving in it's story line.	0.92
4	DOOM 3	Progress in the main plotline is portrayed through cutscenes.	0.92
5	Aura: Fate of the Ages	The story line is paper thin.	0.91
6	Indigo Prophecy	The narrative then unravels from various points of view.	0.91
...
3000	Sphinx and the Cursed Mummy		

ムや、レビューが少ないゲームは、ほぼ検索できなかった。レビューが少ないゲームは持っている情報が少ないため、今回のスコア計算において、最初から類似する文がないため、ランキングの最下位にまとめて配置されていた傾向が見られた。

次に、成功したケースと失敗時のケースについて、具体例を挙げながら考察する。すべての結果において、システム側が受け取ったクエリは、ほぼ、検索者が人手で入力したクエリであった。実験において出した選択肢は、生なレビューなので、特徴として使うときかなりの情報量がある。例えば、メジャーゲームの「Grand Thetf Auto: San Andreas」の中に,” What’s Not-So-Good - Out of the 3, its relatively short - Gameplay wise, it doesn’t add much to the GTA IV experience - The difficulty range of some missions compared to others is uneven What’s Bad: - Lack depth In all, LATD is a good DLC and should be a fun experience for fans as long as they keep that in mind.” というかなり長いレビューが存在した。これは、細かい観点が存在するレビューからできているが、検索時ここまで細かく覚えることが存在しないため、こういうレビューが最初からはじくべきと考えられる。

表 7 失敗した検索タスクにおけるクエリ “Every area I reach has its novelty.”, “I win some momey when I beat a opponent”, “the sound track is very specail to me.” に対するゲーム対象となったレビュー一文 (Pokemon Gold が正解のゲーム)

	クエリ	順位	ゲームタイトル	レビュー	スコア
1 回目	Every area I reach has its novelty	1	Mystery Dungeon: Shiren the Wanderer	The challenge Shiren presents adds to its value.	0.95
		2	Halo 2	Well, this is one area where Halo 2 misses the bar set by its predecessor.	0.94
		3	PixelJunk Eden	Once a Spectra is collected, it goes straight to your Eden and starts growing.	0.94
		4	Freedom Force vs The 3rd Reich	Bottom Line FFFVTR lives up to its predecessor and surpasses it in many ways.	0.94
	
		2530	Pokemon Gold	If you beat a trainer, you win some money and continue on your journey.	0.88
	
		3000	Sphinx and the Cursed Mummy
	
	
2 回目	I win some money when I beat a opponent	1	MLB 07: The Show	You start off by making your own player and customizing him.	0.85
		2	Digimon World Championship	you must fight through title matches to digimon championship and it's not easy.	0.85
		3	Bubble Bobble	You play as one of those dragons and if you play a 2-player game your pal will control the other one.	0.85
		4	WCP 2: Featuring Howard Lederer	So even if everything else in the game is horrible you still have the fact that you can fire up the game and play several other artificially generated opponents for some poker fun.	0.85
	
		You beat him and the game IS officially completed but of course you can still play in order to catch them all and of course that's not going to be easy seeing how there's now at least 250 Pokemon.	0.80
		438	Pokemon Gold	...	0.80
	
		3000	Sphinx and the Cursed Mummy
	
3 回目	the sound track is very specail to me.	1	Rocky Balboa	The sound is good for a fighting game.	0.75
		2	Harry Potter and the Half-Blood Prince	There is a golden aura around the screen, some jazz music in the background and a running Harry.	0.75
		3	Robotron: 2084	As well as the sound which is the exact same as when the game released in 1982.	0.74
		4	Split/Second	On the menu there is some great action-like soundtrack which clearly represents a racing game, and is done to good effect off track and on track	0.74
	
		...	Pokemon Gold	It's the small things that make Gold a more refined game than Red.	0.54
		2055
		3000	Sphinx and the Cursed Mummy
	
	

その上、入力するクエリがもし一部の情報しかない、第 3.3 節で述べた類似度の得点が低いということも判明した。これも上記と同じで、すなわち少ない情報から多くの情報に当てはまっていると類似度が低くなる。これを改善するには、BERT をより細かい観点でファインチューニングする必要がある。また Wikipedia の情報量は少ないため、今後学習データを作るために、手作業で正解データを作ることも考えられる。

6 まとめと今後の課題

本研究では、うろ覚えのゲームを検索可能なアルゴリズムを提案した。このアルゴリズムでは、初めにユーザがクエリを入力すると、それとは異なる観点に言及したレビュー文を提示してくる。ユーザはそれらのレビューについて、自分の思い浮かべているゲームが該当するかしないかを選択していく。こうすることで、キーワードを思いつけないよううろ覚えの状態でも、ゲームを検索化可能になる。このような検索を可能にするために、観点到に敏感に文をベクトル化する言語モデルを学習した。この言語モデルは、Wikipedia から抽出した 1 つの文に対して、その直上の見出しを推定するタスクで BERT をファインチューニングすることで作成した。実際の被験者実験を伴う

表 8 成功した検索タスクにおけるクエリ “I used many weapons.”, “I challed the game but infinity enemies is too bad.”, “the sound track is very specail to me.” に対するゲーム対象となったレビュー一文 (Call of Duty: Black Ops が正解のゲーム)

	クエリ	順位	ゲームタイトル	レビュー	スコア
1 回目	I used many weapons.	1	Warhammer 40,000: Dawn of War	It's 4 races were all unique and it had a great depth. Gears of War, Uncharted, and Kane and Lynch have all been ambassadors for the genre, opening it up and letting people know the action that it brings.	0.37
		2	Army of Two	I found the game	0.37
		3	Scoby-Doo! Mystery Mayhem	captured the essence of Scoby Doo, which was the most important thing.	0.36
		4	Too Human	The story is well written and kept me wanting to play more.	0.36
	
		1247	Call of Duty: Black Ops	Kill Streaks are over powered.	0.28
		3000	Harry Potter: Quidditch World Cup
2 回目	I challed the game but infinity enemies is too bad.	1	Too Human	The story is well written and kept me wanting to play more.	0.15
		2	GG XX Accent Core	There is no doubt that GG XX AC is a game that any Guilty Gear fan is going to want, and once they play it, they will realise it's the best version of the game thus far. When it hit the other consoles last year, I thought it was a pretty decent game with good fan service.	0.15
		3	The Godfather	Game Looks Great with Good Cinematics and Sounds it feels like a game of 2008/09 and it can be maxed out with a good Quad and GPU. Actually	0.14
		4	Necrovision	Really enjoyed this his game for a Good Laugh.	0.14
	
		144	Call of Duty: Black Ops	When call of duty black ops was confirmed to have a port for the wii, people looked back at the previous call of duties released for the wii.	0.12
		3000	Harry Potter: Quidditch World Cup
3 回目	Kill Streaks are over powered.	1	Call of Duty: Black Ops	Kill Streaks are over powered.	0.06
		2	GG XX Accent Core	Finally, the balance in the game has been honed from "Great" in previous games to "Perfect" in GGXXAC.	0.06
		3	Crash Bandicoot 2: N-Tranced	Thank god this on e changed from the first one.	0.05
		4	Marvel: Ultimate Alliance	One flaw for the PS3 version tho is the Incredible Hulk is not in the game.	0.05
		3000	Harry Potter: Quidditch World Cup

評価実験を通じて、精度と有効性について議論する予定である。

今後の課題として、よりアプリケーションとしての使い勝手を向上させることを考えている。現時点では、質問項目はただ現在の入力クエリと遠いものを出すようになっている。今後、ランキングモデルを導入して、よりゲームの特定に有効な質問を優先して出題する必要がある。

謝 辞

本研究はJSPS 科研費 18K18161 (代表: 莊司慶行), 21H03775 (代表: 大島裕明) の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] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [2] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pp. 194–206. Springer, 2019.
- [3] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Learning Distributed Representations of Texts and Entities from Knowledge Base. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 397–411, 11 2017.
- [4] Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and Jimmy Lin. Applying BERT to document retrieval with birch. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pp. 19–24, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [5] Matthew Henderson. Machine learning for dialog state tracking: A review. 2015.
- [6] Tsung-Hsien Wen, David Vandyke, Nikola Mrksić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 438–449, Valencia, Spain, April 2017. Association for Computational Linguistics.
- [7] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, Austin, Texas, November 2016. Association for Computational Linguistics.
- [8] Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1371–1374, 2018.
- [9] Espen Aarseth, Solveig Marie Smedstad, and Lise Sunnanå. A multidimensional typology of games. In *DiGRA Conference*, 2003.
- [10] Toine Bogers, Maria Gäde, Marijn Koolen, Vivien Petras, and Mette Skov. “looking for an amazing game i can relax and sink hours into...”: A study of relevance aspects in video game discovery. In *International Conference on Information*, pp. 503–515. Springer, 2019.
- [11] Xiaoxuan Zhang and Adam M. Smith. Retrieving videogame moments with natural language queries. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, FDG '19, New York, NY, USA, 2019. Association for Computing Machinery.
- [12] Barrett R. Anderson and Adam M. Smith. Understanding user needs in videogame moment retrieval. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, FDG '19, New York, NY, USA, 2019. Association for Computing Machinery.
- [13] Ida Kathrine Hammeleff Jørgensen and Toine Bogers. “kinda like the sims... but with ghosts?”: A qualitative analysis of video game re-finding requests on reddit. In *International Conference on the Foundations of Digital Games*, pp. 1–4, 2020.
- [14] Florian Meier, Toine Bogers, Maria Gäde, and Line Ebdstrup Thomsen. Towards understanding complex known-item requests on reddit. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pp. 143–154, 2021.
- [15] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pre-trained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.