情報処理学会論文誌　デジタルプラクティス　Vol.6　No.1　1–11 (Jan. 2025)

用例ベースと生成 AI を併用したハイブリッド対話システム

関 陽介[1,a)]

受付日 2024年5月13日, 採録日 2024年8月27日

概要：用例ベースの対話システムを運用する場合，不十分な用例の登録により誤回答や回答不可が頻発して，システムへの信頼が低下する可能性がある．そのため，用例の継続的な追加登録により対話システムの正答率を高める必要があるが，登録作業は膨大な時間や労力が求められる．そこで，本研究では管理者の応答範囲を拡大する作業の負担軽減を目指して，用例ベースによる対話機能に加えて生成 AI を用いた応答機能を備えるハイブリッド対話システムを開発する．具体的には，用例ベースの応答文とは別に RAG を用いて発話文の関連情報から応答文を生成する機能を徳島大学で稼働中の対話システムに実装する．徳島大学に開発したシステムを導入した結果，RAG により用例ベースが誤回答した 19.49% の発話文に対して正しく回答でき，履修関係に関する 8.47% の発話文に対してより詳しく説明できた．また，対話システムとしての正答率の向上や応答文の生成時間の有効活用を実現でき，応答範囲の拡大作業の負担を軽減することが可能になった．

キーワード：対話システム，RAG，システム開発

Dialogue System with Generation AI and Example-Based

Yosuke Seki[1,a)]

Received: May 13, 2024, Accepted: August 27, 2024

Abstract: When operating a example-based dialogue system, frequent occurrences of incorrect responses and inability to respond due to insufficient example entries can lead to decreased trust in the system. Therefore, it is crucial to enhance the accuracy of the dialogue system by continuously adding new examples. However, this registration process requires significant time and effort. This study aims to reduce the workload of expanding the administrator's response range by developing a hybrid dialogue system equipped with a response generation feature using generative AI, in addition to the example-based dialogue function. Specifically, this study implemented a feature in the dialogue system at Tokushima University that generates response sentences from related information of utterance sentences using the Retrieval-Augmented Generation (RAG) model. This system allows users to choose responses either based on example-based alone or combination with example-based and RAG. As a result of introducing the system developed at Tokushima University, RAG was able to correctly answer 19.49% of the utterances that were incorrectly answered by example-based, and provide more detailed explanations for 8.47% of the utterances related to course registration. The combination of example-based and RAG also improved the overall accuracy of the dialogue system and allowed for effective utilization of response generation time. Additionally, using RAG reduced the burden of expanding the response range.

Keywords: dialogue system, RAG, system development

1. はじめに

問い合わせ対応の自動化や業務負担の軽減等を目的として，行政機関や教育機関等では対話システムの導入が拡大している．たとえば，ゴミの分別案内サービス[*1]や観光案内サービス[*2]，進学希望者向けサービス[*3,*4]等があり，このようなシステムを公開することで，時間や場所を問わず対話的な情報発信が可能になる．

徳島大学では対話システムが独自に開発されており，進学希望者の情報収集の支援を目的として，大学の特色や学

[1] 徳島大学高等教育研究センター
Tokushima University, Tokushima 770–8501, Japan
[a)] seki@tokushima-u.ac.jp

[*1] イーオのごみ分別案内 https://www.city.yokohama.lg.jp
[*2] 阿波ナビ https://www.awanavi.jp/
[*3] AI 対話システム https://sdcunivchat.qabot.jp/
[*4] Q&A ボット https://chatbot.ac.tsukuba.ecomas.io/

部・学科の特徴，入学者選抜，徳島県での生活に関して応答可能な受験生向け対話システム [1] が 2019 年 4 月に導入されている．また，2020 年 4 月からは大学生活の支援を目的として，施設案内や各種手続き，健康・相談，就職支援等に関して応答可能な徳島大学の学生向け対話システム [2]（以下，学内版とくぽん talk）が稼働している．これらは特定のタスクを解決するという明確な目標を持つタスク指向型であり，事前登録した発話・応答文を用いて対話が行われる用例ベース[*5] の対話システムである．たとえば，学内版とくぽん talk には無線 LAN に関する用例が登録されており，“学内の無線 LAN の利用方法が知りたい”という発話文に対して，“キャンパス内で無線 LAN を利用することが可能です．申請方法や利用マニュアル，無線 LAN 提供エリア等の詳細はこちらをご覧ください．”という応答文が出力される．

手作業で用例を作成することで，登録された範囲内でより品質が高い応答が可能になる．ただし，履修登録の方法や成績の閲覧方法，必修科目等，特に入学後の履修に関する情報は多岐にわたる．学内版とくぽん talk では，対話で用いられる可能性が高い用例が優先的に登録されるが，不十分な用例の登録により誤回答や回答不可が頻発して，対話システムへの信頼が低下する可能性がある．そのため，用例の継続的な追加登録により対話システムの正答率を高める必要があるが，登録作業は膨大な時間や労力が求められる．

そこで，本研究では管理者の応答範囲を拡大する作業の負担軽減を目指して，用例ベースによる対話機能に加えて生成 AI を用いた応答機能を備えるハイブリッド対話システムを開発する．具体的には，用例ベースの応答文とは別に Retrieval Augmented Generation（以下，RAG）を用いて発話文の関連情報から応答文を生成する機能を学内版とくぽん talk に実装して，徳島大学に試行的に導入する．この仕組みにより，管理者の作業負担の軽減に加えて，対話システムとしての正答率の向上や応答文の生成時間の有効活用が期待できる．

本研究の貢献は以下になる．

- 用例ベースと RAG を併用したハイブリッド対話システムを開発したこと．
- 徳島大学に開発した対話システムを導入して，対話システムとしての正答率の向上や応答文の生成時間の有効活用，応答範囲の拡大作業の負担軽減を実現したこと．

[*5] 用例ベースではユーザの発話文との類似度が最も高い用例の発話文を取得して，この用例の発話文に対応する応答文が出力される．

2. 方向性の検討

2.1 学内版とくぽん talk とその課題

学内版とくぽん talk では，徳島大学のマスコットキャラクターである「とくぽん[*6]」が応答キャラクターとして採用されている．この対話システムは通常の対話機能に加えて，ユーザの発話内容に応じた感情や動作をとくぽんが表現することで，ユーザとシステム間の親和な関係の成立を試みている．具体的には，喜びや別れの挨拶等をとくぽんが表現するアニメーション画像を作成して，感情や動作を選択するための分類モデルを用いて，発話内容に対する適切な画像を応答文と合わせて出力している．

学内版とくぽん talk の参考画面を図 1[*7] に示す．対話は下から上に流れており，発話文が“パスワードの有効期限が切れた場合は？”の場合，応答文として“～所属学部の学務係・教務係へ問い合わせてください～[*8]”と通常の場合のアニメーション画像が，“ありがとう！”の場合，“どういたしまして！”と喜びを表現した画像が出力する．

図 1　学内版とくぽん talk の参考画面
Fig. 1　A screenshot of Tokupon talk for students.

[*6] とくぽん https://www.tokushima-u.ac.jp/about/profile/univ_mascot/
[*7] 視認性を高めるために，キャラクター画像や発話・応答文，入力フォーム等を拡大して貼り付けている．
[*8] “～”は省略を意味する．

表 1　学内版とくぽん talk の用例として用いている発話文（例）と応答文の種類
Table 1　Utterance sentences and types of response sentences in examples of Tokupon-talk for students.

項目	発話文（例）	応答文の種類
キャンパス情報	理工学部の学務関係の窓口はどこ？，学内に ATM はある？，情報センターはどこ	16
各種手続き方法	学生証をなくしたときは？，奨学金を受けたいときは？，授業料の納付方法を教えて	17
履修関係	履修登録期間は？，シラバスの利用方法は？，時間割はどこに公開されている？	1
システム操作	Zoom の使い方，学内無線 LAN の利用方法は？，パスワードの有効期限が切れた場合は？	15
健康・相談	体調が悪くなったときは？，健康診断の日程は？，生活費が不足したのでお金を借りたい	7
課外活動	どんなサークルがあるの？，課外活動中にケガをしたときは？，物品の貸し出し	2
就職支援	求人票について，推薦書はどこでもらえる？，就職ガイダンスの予約方法は？	4
国際交流	海外留学の奨学金制度が知りたい，留学について知りたい，留学生と交流したい	3
施設案内	パソコンが自由に使える部屋はある？，寮に入りたいときは？，飲食可能な部屋はある？	5
規則	授業料免除について，警報が発表された場合は？，試験における不正行為について	5
その他	（とくぽんの）名前を教えて，GPA とは何ですか？，徳島大学ってどんな大学？	70

応答文は fastText[*9] を用いて選択しており，2024 年 5 月時点で用例は 2,478 件で応答文は 145 種類が登録されている．なお，単語の揺れ等を考慮して複数の発話文に対して 1 種類の応答文が紐づいている．本システムは統合認証サービス [3] を用いており，徳島大学の学生を対象に提供している．

表 1 に学内版とくぽん talk の用例として用いている発話文（例）と応答文の種類を項目別に示す．たとえば，キャンパス情報に関する項目には，“理工学部の学務関係の窓口はどこ？”等の発話文が登録されており，応答文は 16 種類になる．用例は徳島大学で発行される学生生活の手引きや Web サイト，事務担当者へのヒアリング結果等を参考に登録している．ただし，履修関係は“総合科学部国際教養コースの進級要件は？”や“歯学部歯学科の 1 年次の必修科目は？”等の多様な内容にわたるため，すべての情報を用例に反映するのは困難である．そのため，履修関係の質問には具体的な回答はしておらず，学部・学科が発行する履修の手引きの公開場所や目的の情報の検索方法，担当者に相談ができる学務窓口の場所のみを提供している．

2020 年 4 月から 2022 年 3 月までを対象にした学内版とくぽん talk の分析結果として，最も多く入力された上位 3 件の発話文の内容は，とくぽんへの質問，履修関係，無線 LAN の設定方法であった．履修に関する質問はすべて同じ応答文が出力されるため，ユーザは履修の手引きの確認や学務窓口への問い合わせが必要になる．対話履歴を分析して不足している用例を登録することで，より正確な情報を提供することは可能である．ただし，登録作業は継続的な作業が求められるため，用例に依存せずに応答文が生成される仕組みを導入することが望ましい．

[*9] fastText https://fasttext.cc/

2.2　生成 AI による応答文の生成

生成 AI を用いることで文章作成や校正等が可能になり，たとえばプログラミングのバグ発見 [4] や医学文献の作成補助 [5]，患者へのクリニックレターの作成 [6] 等の手法が報告されている．生成 AI はすでに実用可能な水準に達しており，学生生活を対象にした応答文の生成にも適用できる可能性が高い．そこで，本研究では言語モデルの 1 つである GPT[*10] を用いて応答文の生成手法を検討する．

GPT は Web ページから抽出された大量のデータを学習している [7]．そのため，“徳島大学について教えて”や“徳島大学のアクセス情報は？”等，一般的な質問には正確に応答できる可能性が高い．ただし，公開情報であっても，“徳島大学理工学部知能情報システムコースの進級要件は？”や“徳島大学薬学部に飛び級制度はある？”等の学部・学科に関する手続きや制度等の詳細には応答できない場合がある[*11]．また，2024 年 3 月時点で公開されている gpt-4-0125-preview は，学習データの収集期間が 2023 年 12 月までであり，最新の情報は学習されていない．GPT は徳島大学の学生生活に関して正確に応答できない場合があるため，生成機能を実装するためには新たな知識を追加する必要がある．

GPT に知識を追加する方法としては，独自にデータを収集して事後学習をする方法や RAG がある．RAG は発話文に関連する情報をデータベース等から取得して，発話文や抽出した複数の情報を組み合わせて応答文を生成する仕組みである．事後学習と比べると RAG の実装作業は容易であり，RAG を用いてモデルに知識を追加する事例 [8]-[10] はすでに報告されている．

ただし，RAG は関連情報の検索や収集した情報から応答文を生成する必要があるため，事後学習で作成したモデルと比べて応答時間が長くなる．そこで，まずは事後学習

[*10] GPT https://openai.com/research/overview
[*11] gpt-4-0125-preview で 2024 年 3 月 22 日に確認した結果である．

を対象に知識の追加を試みる．

2.3　事後学習したモデルの評価

事後学習による知識追加により，どの程度の応答が可能になるかを事前調査した．モデルの規模にもよるが，50件から100件程度の学習データで事後学習をすることで，モデルの改善が期待できる[*12]．そこで，学内版とくぽんtalkの用例から無作為に取得した100件の発話・応答文を学習データとして，gpt-3.5-turboを用いて事後学習を行った．学習データから10件の発話文を無作為に抽出して，事後学習したモデルで生成された応答文を筆者と徳島大学職員の2人で評価[*13]した結果，誤回答や情報不足等により，すべての応答文は回答としては不十分であった．**表2**に学習データとして用いた用例と事後学習により生成された応答文の例を示す．誤回答または情報不足となる箇所には下線をひいている．たとえば，発話文が“落とし物をしたときは？”で用例の応答文が“～落とし物は，学務部教育支援課教養教育係で預かっています．～”の場合，生成された応答文は“～キャンパス管理センター3階窓口またはご青葉生協鳴門店1階コンビニエンスコーナーへお問い合わせください．”であり，下線をひいた場所は徳島大学には存在しない．事前調査では事後学習の応答精度は低くなったが，学習データを増やすことで応答内容が改善する可能性がある．ただし，誤回答の場合は特に新入生は応答内容の真偽の判断が困難になるため，本研究ではRAGを用いて応答文の生成を試みる．

2.4　RAGの課題

徳島大学で公開されているドキュメントを検索対象にすることで，履修の登録方法等のドキュメントに関連する応答文をRAGで生成できる．ただし，“駐輪場の場所は？”等のドキュメントに未記載の内容に対する質問には対応できない．また，RAGの応答は時間を要するため，待ち時間が長くなることでユーザの対話意欲が低下する可能性がある．

学内版とくぽんtalkは，用例ベースでありfastTextを用いているため応答時間が早い．実際，表1の各項目の1つ目の発話文11件に対する平均応答時間（秒）を調査した結果，0.82（0.09）[*14]であった．また，継続的に用例を更新しており，2020年4月から2022年3月までの正答率は71.65% [2]である．この用例ベースを利用することで，対話システムとしての正答率の向上やRAGの応答文が出力するまでの待ち時間の有効活用を実現できる可能性がある．

3.　ハイブリッド対話システムの設計と開発

本研究では，用例ベースとRAGを併用したハイブリッド対話システムを開発する．これは2つのモデルを使用することで，発話文に対して用例ベースで応答した後にRAGの応答文を出力する仕組みである．本研究で開発するハイブリッド対話システムの処理の流れを**図2**に示す．主に（A）～（D）が用例ベース[*15]，（E）がRAGの処理になる．まず，用例ベースの処理について説明する．

3.1　用例ベースの処理

用例データベース（図2右上）には，学内版とくぽんtalkの管理者が登録する発話・応答文が保存されている．専用の管理サイトから用例の更新が可能である．

（A）発話文の解析では，ユーザの発話文から形態素解析器のMeCab[*16]で名詞や感嘆詞等の単語を取得する．辞

表2　事後学習に用いた用例と生成された応答文の例

Table 2　Examples used for fine-turning and examples of generated response sentences.

用例の発話文	用例の応答文	事後学習により生成された応答文
落とし物をしたときは？	教養教育棟周辺における落とし物は，学務部教育支援課教養教育係で預かっています．その他の場所での落とし物については，お近くの各事務室で尋ねてください．	落とし物をした場合は，所在不明のお問い合わせは，キャンパス管理センター3階窓口またはご青葉生協鳴門店1階コンビニエンスコーナーへお問い合わせください．
情報センターはどこにありますか？	情報センターは常三島キャンパスの光応用棟の北側にあります．分室は蔵本キャンパス蔵本附属図書館分館内にあります．	徳島キャンパスの中央図書館の2階に情報センターがあります．必要な情報やサポートを提供してくれる場所です．
入試の成績開示について	本学に入学した者は，6月1日から6月30日の間，合格した入試の成績開示を大学設置の証明書発行機で入手することができます．	入試の成績開示についての詳細については，徳島大学の公式ウェブサイトで確認することができます．入試の成績は個人情報に関わるため，適切な手続きが必要です．

[*12] OpenAI Fine-tuning https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset

[*13] 2人の評価・確認結果が異なる場合は両者の話し合いにより決定した．

[*14] 括弧内の数値は標準偏差を示す．以降は省略する．

[*15] （A）～（D）は学内版とくぽんtalkの処理であり，本研究では改修せずに利用する．

[*16] MeCab https://taku910.github.io/mecab/

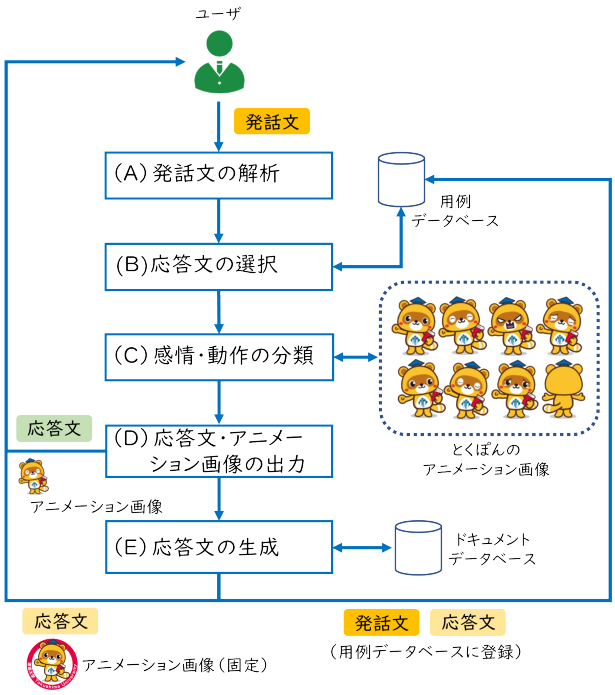

図 2　ハイブリッド対話システムの処理の流れ

Fig. 2　Processing flow of the hybrid dialogue system.

書は NEologd 辞書[*17] を用いた．（B）応答文の選択では，用例データベースに登録された発話・応答文を学習データとして，fastText を用いて作成した分類モデルを用いて応答文が選択される．なお，fastText は Wikipedia 日本語記事全文から作成されたモデルを使用した．（C）感情・動作の分類では，発話内容に応じたとくぽんの感情や動作を決定する．感情や動作の種類は，感情表現として通常，喜び，怒り，悲しみ，驚き，動作表現として挨拶，別れの挨拶，応答不可の計 8 種類になる．発話内容に対する感情や動作を決定するために，発話履歴と日本語評価極性辞書 [11]，単語感情極性対応表 [12] を用いて分類モデルを作成した．（D）応答文・アニメーション画像の出力では，（B）で選択された応答文と（C）で決定されたアニメーション画像を出力する．アニメーション画像は公開されているとくぽんの静止画から Live2D を用いて作成した．アニメーション画像は APNG 形式で幅が 142 px，高さが 166 px，フレームレートが 30 fps，1 回の動作が 5 秒程度，アニメーションの繰り返し設定を有効にしている．

3.2　RAG の処理

（E）応答文の生成の流れを**図 3** に示す．（E1）ドキュメント検索では，RAG で応答文を生成するために発話文の関連情報を検索する．2024 年 3 月時点では，OpenAI は API で用いるプロンプト[*18] の内容をモデルの学習に利用

[*17] NEologd 辞書 https://github.com/neologd/mecab-ipadic-neologd
[*18] プロンプトとはモデルに対する指示や質問になる．

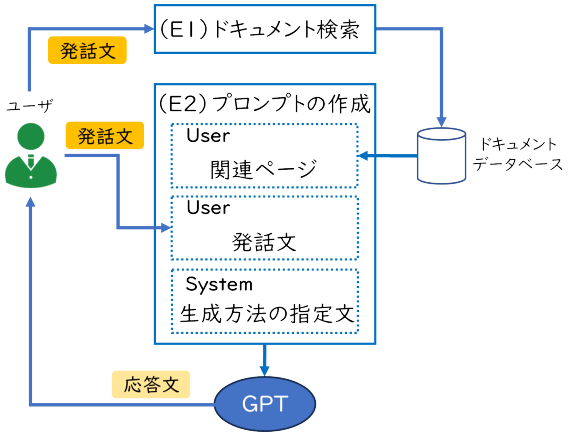

図 3　RAG の処理の流れ

Fig. 3　Processing flow of RAG.

しないと公表[*19] しているが，情報漏洩に繋がる可能性を考慮して，検索対象は学外に公開しているドキュメントに限定する．そこで，諸手続き方法や学務部の窓口等が記載された学生生活の手引き，卒業までに履修する教育プログラムに関する情報が記載されている教養教育・各学部[*20] の履修の手引きの計 8 つを検索対象とした．これらのドキュメントは徳島大学の Web サイトに PDF 形式で公開されている．なお，学生生活の手引きは学内版とくぽん talk の用例を作成する際に参考にしているが，登録漏れを考慮して検索対象に含めている．

ドキュメントの検索環境を構築するために，pdfminer.six[*21] を用いて各 PDF から前処理を行わずに文字を抽出して，自然言語処理ライブラリの GINZA[*22,*23] を用いて分散表現を取得した．そして，ベクトル検索エンジンである Qdrant[*24] のデータベース（図 3 の右上）に取得した分散表現を保存した．毎年求められる更新作業の負担を考慮して，ドキュメントはページ単位でデータベースに保存している．登録したページ数は 1,037 で，平均文字数は 914.83（932.23）[*14] になる．

Qdrant では，id とベクトルデータ，Payload で構成されるレコードが Collections に登録される．id は一意な番号，Payload は付加情報になる．付加情報にはドキュメント名，対象学部，ページ番号を登録した．登録例としては，“履修の手引き，理工学部，112” になる．RAG の応

[*19] OpenAI https://openai.com/blog/introducing-chatgpt-and-whisper-apis
[*20] 徳島大学には総合科学部，医学部，歯学部，薬学部，理工学部，生物資源産業学部の 6 学部がある．
[*21] pdfminer.six https://pdfminersix.readthedocs.io
[*22] GINZA https://megagonlabs.github.io/ginza/
[*23] 図 2 の（B）応答文の選択では，未知語に対応できる fastText を用いたが，検索対象となるドキュメントには大学生が理解できうる単語が多く用いられており，また説明文は文章として記載されているため，文脈を考慮した分散表現を扱える GINZA を用いた．
[*24] Qdrant https://qdrant.tech/

答文は内容が誤っている可能性があるため，検索されたページの情報を応答文に記載するために付加情報を用いる．

プロンプトには，モデルの振る舞いを設定する System，モデルへの質問等を入力する User，モデルの過去の応答文を入力する Assistant を Role として入力できる．ChatGPT[*25] のように対話履歴を利用することで，文脈を考慮した応答が可能になる．ただし，学内版とくぽん talk は一問一答形式の対話が多く，プロンプトの文字数が増加すると応答時間や利用料金が増えるため，対話履歴を用いずに応答文を出力させる．（E2）プロンプトの作成では発話文と取得した関連ページの記載内容を User に，生成方法の指定文を System に入力して，GPT から応答文を取得する．指定文は“与えられた情報を用いて回答してください．回答できない場合は「分かりません.」を出力してください.”になる．関連性が高いページは，分散表現として表現した発話文と，各ページの余弦類似度を算出して取得する．

ユーザによっては，自身の所属学部・学科を入力せずに質問する可能性がある．たとえば，所属名が示されていない“進級要件は？”という発話文に対しては，学部・学科により内容が異なるため正確な応答が困難である．そのため，RAG を利用する場合は，検索対象となるドキュメントを事前にユーザに選択させる仕組みにする．たとえば，理工学部の履修の手引きをユーザが選択することで，理工学部に関する質問に対して誤回答の頻度を軽減できる可能性がある．

RAG の応答文を用例に登録することで，類似する発話文に対して短時間で応答が可能になる．そこで，発話文と RAG の応答文を用例データベースに登録する機能（図 2 の下部）を設けた．ただし，不適切な応答文が生成される場合があるため，学習データとしての使用可否は管理者が判断する．

3.3　ページの取得件数の調査

ページの取得件数を増やすことで，より多くの関連情報が収集されて適切な応答文が生成されやすくなる．ただし，応答時間への影響を考慮して，ページの取得件数を検討する必要がある．そこで，取得件数による応答時間と正答率への影響を事前調査した．調査内容としては，ページの取得件数を 1 件（i）〜3 件（iii）として，10 件の発話文に対する応答時間と応答内容を確認した．ドキュメントを参考に発話文を作成して，応答文として不十分・十分であるかを筆者と徳島大学職員 2 人が評価[*13] した．発話文は“学生証とは何？”や“喫煙してもいい？”，“歯学科の進級要件は？”等になる．調査結果として，平均応答時間（秒）と正確な応答文と判定された件数は，（i）11.64（10.67）[*14] と 7，（ii）11.91（8.12）と 9，（iii）13.55（9.29）と 10 であった．応答文の例としては，発話文が“飛び級制度はある？”の場合，不十分な応答文としては“はい，飛び進級に関する規定があります．留年生が飛び先学年の進級規定単位数を満たしている場合に飛び進級が認められます.”が，適切な応答文としては，“はい，飛び級制度があります．留年生が飛び先学年の進級規定単位数を満たしている場合に，飛び進級が認められます．たとえば，1 年次もしくは 2 年次に留年した学生が，3 年次もしくは 4 年次への進級条件を満たしていれば，1 年次から 3 年次，または 2 年次から 4 年次への飛び進級が可能です.”が出力された．なお，これらの例では具体例が示された後者を適切な回答と判断している．（iii）はやや時間を要しており，（i）より（ii）の応答精度が高いため，本研究ではページの取得件数を 2 件とした．

3.4　システム開発

提案手法に基づく機能を学内版とくぽん talk に実装した．GPT の利用には OpenAI の Chat Completions API を用いた. モデルは gpt-4 を用いたが, 最新モデルが公開されるたびに更新する．**表 3** に開発で用いたソフトウェアを示す．従来の用例ベースは応答時間が短く正答率が 70% 以上であるため，ユーザによっては RAG の生成機能を利用しない可能性がある．そこで，対話画面に RAG の選択項目[*26] を設けた．この項目には以下が登録されている．

(1) 利用しない
(2) 利用する（学生生活の手引き）
(3) 利用する（教養教育履修の手引き）
(4) 利用する（総合科学部履修の手引き）
(5) 利用する（医学部履修の手引き）
(6) 利用する（歯学部学生便覧[*27]）

表 3　開発に用いたソフトウェア
Table 3　A list of used software.

種類	ソフトウェア
仮想化ソフトウェア	VMware ESXi 7.0
OS	CentOS 7.4 64bit
Http サーバ	Apache 2.4.6
DBMS 等	MariaDB 5.5.52, Qdrant
プログラミング言語等	PHP 5.4.16, Python 3.9, jQuery 3.5.1, HTML
その他	fastText 0.9.2, MeCab, Shibboleth 3.2.2, Live2D, gpt-4

[*25] ChatGPT https://chat.openai.com/

[*26] ユーザには RAG という表現は用いずに，生成機能という名称で紹介している．

[*27] 徳島大学では学部によりドキュメントの名称が異なる．

(7) 利用する（薬学部履修の手引き）
(8) 利用する（理工学部履修の手引き）
(9) 利用する（生物資源産業学部履修の手引き）

（1）は用例ベース，（2）～（9）は用例ベースと RAG の併用になり，それぞれ検索対象のドキュメントが記載されている．（2）～（9）が選択された場合，RAG の生成機能は誤回答する可能性がある旨を出力する．

システムの参考画面と（2）を選択した場合の対話例を**図 4**[*7] に示す．発話文が“図書館の開館時間は？”の場合，用例ベースにより“図書館全体に関するご質問は，図書館 HP に掲載しております．HP はこちら～”が，注意文として“生成機能を使って回答文を作成しています．5 秒から 30 秒程お待ちください．”が，RAG により“図書館の開館時間は以下のとおりです：・月曜日～”が順に出力される．RAG で生成された応答文の下部にドキュメント名とページ番号を青字で記載している（図の上部）．生成機能で出力された応答文は，図 2 の（C）で決定されるとくぽんのアニメーション画像は使用せず，徳島大学の名称が入った 1 種類のアニメーション画像（図 4 の左上）を使用している．

管理者の作業負担を軽減するために，検索対象の更新作業の大部分を機械的に処理している．具体的には，管理者が指定の場所に保存したドキュメントが，ドキュメントデータベースに自動で登録される仕組みにしている．

OpenAI のシステム障害により GPT を利用できない場合，RAG の応答文は生成されない．そこで，システム障害が発生した場合は，案内文として“現在，障害が発生しているため本機能は利用できません．復旧までしばらくお待ちください．”を出力する．

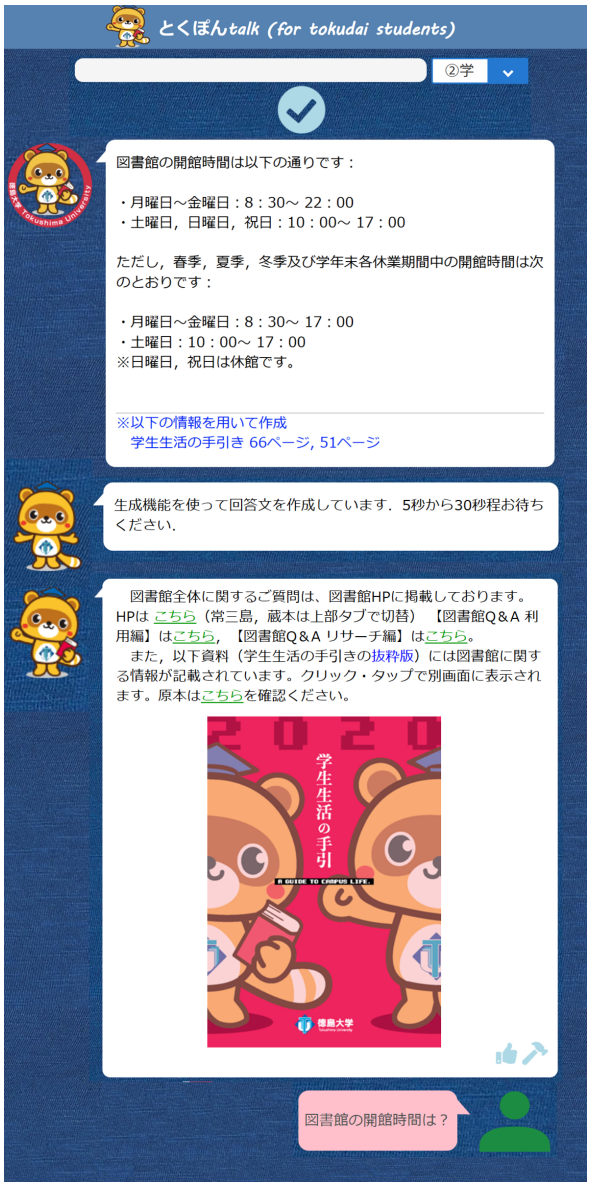

図 4　提案システムの参考画面
Fig. 4　A screenshot of the proposed system.

4.　システム導入

4.1　導入結果

2023 年 10 月 14 日から 2024 年 3 月 31 日までを対象に，提案手法に基づく機能を学内版とくぽん talk に実装して徳島大学に試行的に導入した．RAG の生成機能はシステムの対話画面内でのみ紹介しており，学生にはメールや Web サイト等で本機能を案内していない．

システムの導入結果として，全体，用例ベース，用例ベースと RAG の併用（以降，併用）別に述べる．なお，用例ベースと併用をまとめて全体と表現している．ユーザ数（累計）は全体が 321，用例ベースが 221，併用が 100 であった．発話件数は全体が 947，用例ベース 711，併用が 236 であった．**図 5** に全体，用例ベース，併用別のユーザ数と発話数の推移を示す．棒グラフがユーザ数，折れ線グラフが発話数を示す．12 月頃までは全体のユーザ数と発話数は減少傾向にあるが，2 月頃からは増加傾向がみられた．3 月は進級要件や卒業単位，2024 年度の履修に関する質問が全体の 16.86% を占めており，進級・卒業可否の確認や次年度の準備等のために，ユーザ数と発話数が

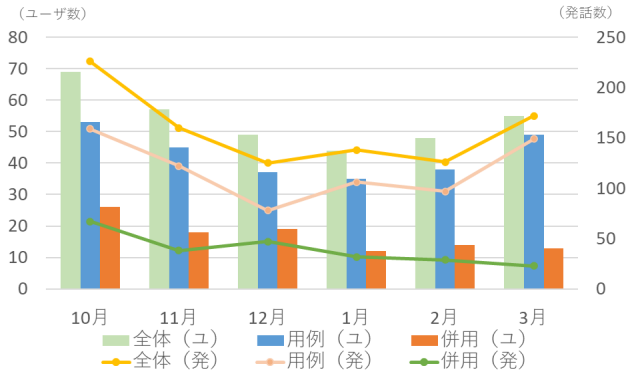

図 5　ユーザ数と発話数
Fig. 5　The number of users and utterances.

表 4　発話文，用例ベースと RAG で生成された応答文の例
Table 4　Examples of utterance and response sentences generated by example-based and RAG.

発話文	用例ベースの応答文	RAG で生成された応答文
飛び級制度はありますか？	勉強中のため回答できません.	はい，飛び級制度が存在します．理工学部昼間コースの学生で，1 年次から 3 年次までの所定の授業科目を優れた成績で修得したと認められる場合，「大学院修士課程の学部 3 年次学生を対象とする特別選抜」に出願することができます．この試験に合格すると学部 3 年次から大学院修士課程に「飛び入学」ができます〜. ※理工学部 履修の手引き 23 ページ，192 ページ
知能情報コースの進級要件は？	履修の手引きはこちらに公開しています〜.	知能情報コースの進級要件は以下の通りです．1 年次から 2 年次への進級要件：卒業に必要な教養教育科目および専門教育科目の合計 35 単位以上．2 年次から 3 年次への進級要件：卒業に必要な教養教育科目および専門教育科目の合計 70 単位以上．3 年次から 4 年次への進級要件：卒業に必要な教養教育科目および専門教育科目の合計 105 単位以上〜. ※理工学部 履修の手引き 127 ページ，152 ページ
他コースを履修したい場合は？	履修の手引きはこちらに公開しています〜.	その場合，コース自由選択科目や他コース選択科目を履修することが可能です．自分が選択した「履修パターン」を参考にして履修計画を立て，選択してください．ただし，視野を広げて自分の専門領域外も俯瞰できるように履修することが推奨されています〜. ※総合科学部 履修の手引き 17 ページ，11 ページ

増えた可能性がある．

RAG で検索されたドキュメントの件数は，（8）理工学部が 111，（5）医学部が 45，（2）学生生活の手引きが 32，（4）総合科学部が 24，（9）生物資源産業学部が 10，（7）薬学部が 6，（3）教養教育が 6，（6）歯学部が 2 であった．理工学部の学生数は学部全体の約 45% を占めるため，本学部の履修の手引きが RAG で最も多く利用されたと考えられる．

試行期間の対話履歴を筆者と徳島大学職員の 2 人で確認[*13] した結果，正答率は全体が 63.04%，用例ベースが 59.07%，併用が 75% であった．併用は用例ベースまたは RAG が正確に応答できた場合に正答と判断している．過去 2 年間の用例ベースの正答率 71.65% と比較すると，試行期間の用例ベースの正答率は 12.58% 低くなった．導入期間は 2023 年 10 月から開始しており，“メールの利用方法は？”や“無線 LAN の設定方法は？”等の，新入生の質問傾向が高く正確に応答できる発話文の入力件数が少なかったため，正答率が低下したと考えられる．なお，用例ベースと比べて併用の正答率は 15.93% 高くなった．これは用例ベースでは対応できなかった発話文に RAG が応答できたことが要因として考えられる．

併用において，用例ベースでは誤回答[*28] で RAG では正答した件数が多い場合に，RAG を導入した効果があると判断できる．併用の内訳（併用の発話文 236 件に対する割合，全体の発話文 947 件に対する割合）として，用例ベースと RAG がともに正答した件数は 53（22.46%，5.60%），用例ベースのみは 78（33.05%，8.24%），RAG のみは 46（19.49%，4.86%），ともに誤回答は 59（25%，6.23%）であった．RAG の導入により，併用では用例ベースが誤回答した 19.49%，全体では 4.86% の発話文に対して正しく回答できたと判断できる．**表 4** に発話文，用例ベースと RAG で生成された応答文の例を示す．たとえば，“飛び級制度はありますか？”に対しては，用例ベースは“勉強中のため回答できません.”が，RAG は“はい，飛び級制度が存在します．理工学部昼間コースの学生で〜”が出力された．用例ベースでは回答できなかった一部の発話文に対して，RAG は履修の手引き等から関連情報を取得することで正確に回答できたと考えられる．全体の正答率が向上した割合は大きくはないが，併用の発話件数は全体の 24.92% であることや，2024 年 3 月時点で 4 年間運用している用例ベースの正答率が高まったことが影響したと考えられる．

用例ベースでは，履修関係の質問には履修の手引きの公開場所等を提供しているが，不明点を解決するための参考にはなりうるため，正答として集計している．ただし，RAG ではより詳しく回答できる場合があるため，履修関係において用例ベースと RAG がともに正答している場合，RAG により応答内容の品質が向上したと判断できる．

[*28] “分かりません.”等の回答不可も含む

併用でともに正答した 53 件の発話文を調べた結果，20 件は履修関係の内容であった．そのため，履修関係に関しては併用では 8.47%，全体では 2.11% の発話文に対して，より詳しく説明できたと判断できる．たとえば，表 4 に示した“知能情報コースの進級要件は？”や“他コースを履修したい場合は？”に対しては，用例ベースでは履修の手引き等の紹介にとどめているが，RAG ではドキュメントの内容に基づき正確な情報を出力できている．

併用では，用例ベースのみが正答した件数が最も多くなった．これは学生生活・履修の手引きには関連しない発話文が，併用の発話文 236 件の 43.98% を占めていたことが要因として考えられる．たとえば，(5) 医学部を対象に“学生証をなくした場合は？”や (8) 理工学部を対象に“けがをしたときの対処方法”が質問されていた．各ドキュメントの記載内容を把握できていないユーザを考慮して，それぞれの目次を紹介する等の対応が必要と考えられる．また，本研究では一問一答形式の対話を前提としているが，“中国語の単位いる？”の後に“それは進級に必要？”等の文脈を考慮した発話や，“明日の天気は？”等の雑談がみられた．ChatGPT 等が普及しているため，過去の発話内容を用いた対話や雑談が本システムにも試みられた可能性がある．ユーザの混乱を防ぐために，雑談には未対応等の生成機能の制限を事前に説明する必要がある．

用例ベースのみが正答した発話文 78 件のうち，履修関係は 22 件の質問がされており，これらは RAG では正しく応答できていなかった．たとえば，“必修何単位取ればいい？”や“2 年への進級に必要な単位数は？”，“卒業条件について”に対しては，RAG では“分かりません．”や“テキストには必修科目の単位数についての具体的な情報は記載されていません．”等の応答がされていた．これらは関連ページの検索に必要な単語や表現が，不十分であることが要因として考えられる．たとえば，“必修科目の単位数は？”や“卒業要件は？”では正確な情報が得られた．また，学科・コース別に進級要件等が定められており，それぞれドキュメントの異なるページに記載されている．このような場合は，発話文にコース名等が示されていないと誤回答になる可能性がある．試行期間では生成機能の説明にとどめていたが，たとえば発話文の作成方法や適切な回答が得られる発話例を事前に紹介することで，RAG の正答率が向上する可能性がある．

RAG の平均応答時間（秒）と応答文の平均文字数，検索で取得された 1 ページ辺りの平均文字数は，誤回答の場合が 10.39（14.06）[*14]，41.11（90.72），1191.34（593.93），正答の場合が 18.89（22.25），170.63（246.71），1056.91（428）であった．誤回答の場合は応答文の文字数が少ないため，正答の場合と比べて平均応答時間が短くなったと考えられる．併用における用例ベースの応答文の平均文字数は 164.38（142.64）であり，応答文に含まれる平均リンク数は 0.58 になる．用例ベースの応答文の文字数にはばらつきがあり，また誤回答の場合もあるが，RAG の応答文が出力されるまでの時間は，ユーザが用例ベースの応答文を閲覧するために利用できたと考えられる．

以上の結果より，用例ベースと RAG の併用で対話システムとしての正答率の向上や応答文の生成時間の有効活用を実現できた．また，RAG の検索対象に履修の手引き等のドキュメントを加えたことで，これらの内容に関する用例を管理者が登録する作業が不要になった．内容にもよるが，これまでの用例の登録作業では，発話文，リンクや画像の挿入等も含めた応答文の作成に 1 件で 10 分程度の時間を要していた．開発したシステムでは，ドキュメントを指定の場所に保存するだけで検索対象が更新されるため，RAG により管理者の応答範囲を拡大する作業負担を軽減できたと判断できる．

4.2　今後の課題

本研究では，RAG の応答時間を有効活用するために用例ベースを活用した．ただし，RAG の応答時間が長い場合，ユーザは応答文が出力されるまで待機する必要があるため，RAG の応答時間を短縮することが望ましい．応答時間を短縮する 1 つの方法として，応答文やプロンプトの文字数の調整が考えられる．参考までに，表 4 の発話文 3 件に対する RAG の平均応答時間（秒）は 13.56（6.23）であったが，応答文の文字数を 50 文字程度に制限した結果，平均応答時間は 5.58（1.13）まで減少した．応答内容への影響を考慮する必要があるが，応答文やプロンプトの文字数を減らして応答時間を短縮する方法を検討したい．

本研究では，RAG で検索対象となるドキュメントをページ単位でデータベースに登録した．ただし，履修上の制限や他大学の授業科目の履修等，同一のページに異なる内容の説明文が記載されており，発話内容に関連しない情報が取得される場合がある．また，説明文がページをまたぐ場合，関連情報が部分的に取得できない場合がある．そのため，ページ単位の検索はプロンプトの文字数が増えることで応答時間が長くなり，不適切な応答文が生成される可能性が高まる．RAG は検索結果が応答内容や応答時間に大きく影響するため，たとえば目次に沿った説明文の登録や，Recall@k や MRR 等による検索精度の調査・改善が必要になる．また，本研究ではドキュメントの取得件数は 2 件としたが，ページにより文字数は異なる．そのため，文字数に応じて取得件数を動的に変更する等，応答時間を考慮してより多くの関連情報を取得する仕組みが求められる．

GPT の利用料金は入力・出力トークン[*29] に基づく従量

[*29] トークンとは GPT が扱うテキストの最小単位であり，プロンプトが入力で応答文が出力に該当する．

課金制である．本研究の導入期間では，利用料金（$）の平均月額が 13.85（18.57）で，最も RAG のユーザ数が少ない 3 月は 0.58 であった．各月の利用料金は少額ではあったが，学内版とくぼん talk のユーザ数は，新入生が学生生活を開始する 4 月が最も多い．そのため，年間を通した運用を想定した場合，前述したがプロンプトの文字数を調整する等，利用料金を減らす方法も検討する必要がある．また，本研究では GPT を使用したが，無償で公開されている Llama[*30] や calm[*31] 等の大規模言語モデルを使用することも可能である．応答精度や応答時間を確認する必要があるが，利用料金を考慮して他のモデルの使用も検討したい．

本研究では PDF から抽出した文字を用いて回答文を生成したが，施設の場所や学年歴等が紹介されている画像には未対応である．応答文には生成に用いたドキュメント名とページ番号を記載しているが，関連する画像を応答文の中に表示することで，ユーザの理解はより深まる可能性がある．そのため，たとえば関連ページへのリンクの貼付や，該当ページに貼られた画像を取得して応答文の中に表示させる等により，視覚的に応答内容の品質を高める方法を検討する必要がある．

5. まとめ

本研究では管理者の応答範囲を拡大する作業の負担軽減を目指して，用例ベースによる対話機能に加えて生成 AI を用いた応答機能を備えるハイブリッド対話システムを開発した．具体的には，用例ベースの応答文とは別に RAG を用いて発話文の関連情報から応答文を生成する機能を学内版とくぼん talk に実装して，2023 年 10 月 14 日から 2024 年 3 月 31 日まで徳島大学に試行的に導入した．導入結果として，併用では用例ベースが誤回答した 19.49%，全体では 4.86% の発話文に対して正しく回答でき，履修関係に関しては併用では 8.47%，全体では 2.11% の発話文に対してより詳しく説明できた．また，対話システムとしての正答率の向上や応答文の生成時間の有効活用を実現でき，応答範囲の拡大作業の負担を軽減できた．

RAG の生成機能の検索対象となるドキュメントは，ページ単位でデータベースに登録している．そのため，プロンプトの文字数の増加や関連情報の部分的な未取得等により，応答時間が長くなり不適切な応答文が生成される傾向が高まる．項目別に説明文の保存や応答文の文字数の調整，検索精度の調査・改善等により，応答時間の短縮や正答率の向上を実現できる可能性がある．対話的に学生の大学生活をより支援するためにも，これらの課題を解決して対話システムの機能改善を行いたい．

本研究では，学生生活・履修の手引きを検索対象として応答文を生成したが，授業で用いる教科書も対象に加えることは可能である．徳島大学のデータサイエンスの講義で用いられる教材を検索対象に加えた結果，“疑似相関とは何ですか？”や“信頼区間の求め方は？”等の質問に正確に応答できた．言語モデルの知識のみを用いて回答する場合も多くみられたが，教育の支援においても十分に RAG を用いた対話システムを活用できる可能性がある．そのため，大学生活や教育の支援を目的とした検索対象の拡充も検討したい．

謝辞　本研究は JSPS 科研費 JP19K14317，JP24K16757 の助成を受けたものです．

[*30] Llama https://llama.meta.com/
[*31] calm https://huggingface.co/cyberagent

参考文献

[1] Seki, Y. and Ueno, Y.: A Recommendation-type Dialogue System Responding to Potential Requests in Consideration of PersonalAttributes, *Information and Technology in Education and Learning*, Vol.3, No.1, pp.Trans–p003 (2023).
[2] 関　陽介：大学の学生生活に関する情報収集を親和的に支援する対話システム，情報処理学会論文誌デジタルプラクティス，Vol.4, No.4, pp.1–10 (2023).
[3] 松浦健二，上田哲史，佐野雅彦：複数認証基盤に対応する複合 SSO 環境でのユーザエクスペリエンス，学術情報処理研究，Vol.16, No.16, pp.138–145 (2012).
[4] Surameery, N. M. S. and Shakor, M. Y.: Use ChatGPT to Solve Programming Bugs, *International Journal of Information technology and Computer Engineering*, Vol.3, No.1, pp.17–22 (2022).
[5] Biswas, S.: ChatGPT and the Future of Medical Writing, *Radiology*, Vol.307, No.2, pp.1–3 (2023).
[6] Ali, S. R., Dobbs, T. D., Hutchings, H. A. and Whitaker, I. S.: Using ChatGPT to write patient clinic letters, *The Lancet Digital Health*, Vol.5, No.4, pp.e179–2181 (2023).
[7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D.: Language Models are Few-Shot Learners, *Advances in Neural Information Processing Systems*, Vol.33, pp.1877–1901 (2020).
[8] Cherumanal, S. P., Tian, L., Abushaqra, F. M., de Paula, A. F. M., Ji, K., Ali, H., Hettiachchi, D., Trippas, J. R., Scholer, F. and Spina, D.: Walert: Putting Conversational Information Seeking Knowledge into Action by Building and Evaluating a Large Language Model-Powered Chatbot, Proc. of the 2024 Conference on Human Information Interaction and Retrieval, pp.401–405 (2024).
[9] Mansurova, A., Nugumanova, A. and Makhambetova, Z.: Development of a Question Ansqering Chatbot for Blockchain Domain, *Scientific Journal of Astana IT University*, Vol.15, pp.27–40 (2023).
[10] Wang, C., Ong, J., Wang, C., Ong, H., Cheng, R. and Ong, D.: Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation, *Annals of Biomedical Engineering*, Vol.52, No.5, pp.1115–1118 (2023).

[11]　東山昌彦，乾健太郎，松本裕治：述語の選択選好性に着目した名詞評価極性の獲得，言語処理学会第 14 回年次大会論文集，pp.584–587 (2008).
[12]　高村大也，乾　孝司，奥村　学：スピンモデルによる単語の感情極性抽出，情報学論，Vol.47, No.2, pp.627–637 (2006).

関 陽介（正会員）

徳島大学高等教育研究センター准教授．2008 年甲南大学理工学部情報システム工学科卒業．2014 年徳島大学大学院博士前期課程修了．2017 年同大学大学院博士後期課程単位修得退学．博士（工学）．2008 年 4 月から IT 系企業での勤務を経て，2011 年 3 月から徳島大学情報センター特任助教．2017 年 5 月より同大学高等教育研究センター特任研究員．特任講師を経て 2022 年 4 月より現職．対話システム，大学入学者選抜等の研究に従事．電子情報通信学会，人工知能学会，日本教育工学会各会員．