# 公式ウェブサイトをベースにしたQAチャットボットの自動構築

坂田 亘 [1,2]　田中 リベカ [2]　黒橋 禎夫 [2,3]
[1]LINE 株式会社　[2] 京都大学　[3] 国立情報学研究所 CRIS
wataru.sakata@linecorp.com {tanaka, kuro}@nlp.ist.i.kyoto-u.ac.jp

## 1　はじめに

企業・自治体などの商品・サービスについて質問したいということは日常的に発生する。マニュアルやウェブページで調べられることもあるが、適切な情報にたどりつくのに時間がかかることも少なくなく、情報弱者であればなおさら困難である。コンタクトセンターに電話をして教えてもらえると良いが、コンタクトセンターの運営には大きなコストがかかっており、往々にしてなかなかつながらない。

そのような問合せに自動的に回答する QA チャットボットがあれば理想的である。しかし、現状、QA チャットボットを構築するためにはシナリオの作り込みを含む膨大な情報管理コストが必要であり [3]、導入は十分に進んでおらず、それほど便利なものも存在しない。

一方、ニューラルネットワークによる対話研究が大きな進展を見せており、たとえば Wikipedia のテキストを用いて質問に答える大規模機械読解システム [4] などには大きな可能性が感じられる。しかし現状では、作り込みシナリオのように複数ターンで徐々にユーザの疑問を絞り込んで質問に答えるようなニューラル対話システムは実現されていない。

情報管理コストという意味では、一定レベルの企業・自治体などの組織には公式ウェブサイトが存在し、商品・サービスや組織そのものに関する情報が、相応の更新頻度で、ある程度詳細に提供されている。そこで、本研究では、公式ウェブサイト上の情報（HTML 構造を含む）を活用し、そこから対話のシナリオ（対話フローチャート）を自動構築し、これをベースに動作する QA チャットボットを提案する。これによって、QA チャットボットのためだけの情報構築・管理や訓練データ構築が不要な QA チャットボットを実現する。

実験により、提案手法によって複数言語・ドメインのウェブサイトから質の良い対話フローチャートを生成できること、構築した QA チャットボットがユーザの意図を捉えた適切な応答を実現することを示す。

## 2　対話フローチャートの構築

本研究では、組織の公式ウェブサイトの各ページを、文書構造を利用して木構造のデータに変換し、QA チャットボットの対話シナリオとして利用する（図 1）。この木構造のデータを対話フローチャートと呼ぶ。対話フローチャートの各ノードは QA チャットボットの 1 発話に対応し、3.1 節で述べる対話管理では木構造を参照することで複数ターンの対話を実現する。

### 2.1　ページのフィルタリング

公式ウェブサイト上の多くのページは利用者にとって有益な情報を含むが、過去の告知等の古い情報や、単なるリンク集のようなページも存在する。これらは QA チャットボットの知識源からは除外することが望ましい。ここでは、以下のページを変換の対象外とした。

- ページ内のリンクテキストの長さが、内容のテキストの長さを上回るもの
- ページのタイトルや URL に特定の日時を含むもの
- 見出しの H タグを含まないページ

### 2.2　木構造への変換

一般に文書は章・節・段落・文といった構造をもつ。公式ウェブサイトでは特に、大量の情報をわかりやすく提示するため見出しがよく活用され、文書構造もより明確に表現される。

この見出しの階層構造を用いて、ウェブ文書を対話フローチャートへ変換する。見出しは H タグで記述され、最も上位レベルの大見出しは h1 から始まり、数字が大きくなるほど見出しは小さくなる。見出しの大小関係をそのまま親子関係として、各ノードが見出し（タイトル）とその直下のテキスト（本文）に対応した木構造に変換する。これにより、元の文書での内容のまとまりや主従関係を保持した対話フローチャートが得られる。

上位レベルの見出し文字列からは内容が読み取れるのに対し、下位にいくほど見出しは細部化され、それ単独では何のことだかわかりづらい場合がある。そのため、注目する見出しの文字列だけでなく、それより

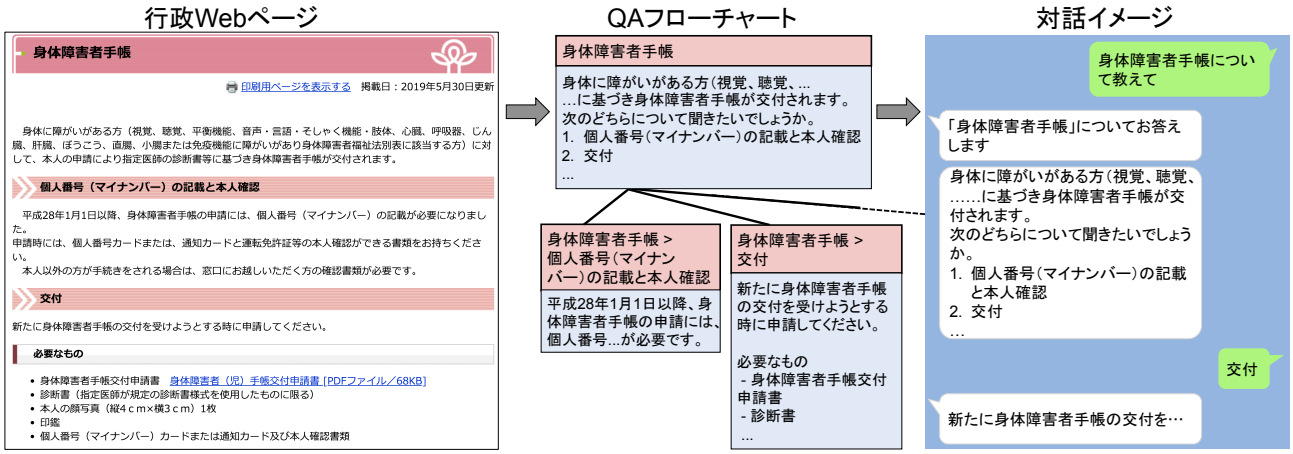


図 1: 提案手法の概要

上のレベルの見出しの文字列も連結したものをタイトル文字列として用いることとした。

### 2.3 タイトルと本文の整形

対話フローチャートを QA チャットボットの応答に利用するには、1 発話に対応する各ノードが長さ・内容の両面で適切な粒度であることが望ましい。そこで、木構造への変換後、QA チャットボットの発話として用いやすいようタイトルと本文を整形する。

タイトルの整形では、表現の冗長性を取り除く。たとえば、以下は丹波市ウェブページの福祉人材確保奨学金返還支援補助金の補助内容に関する記載に注目した場合に得られる項目名であるが、同じ表現が繰り返し現れ、必要以上に項目名が長くなってしまっている（連結文字には ‘>’ を用いる）。

> 福祉人材確保奨学金返還支援補助金について > 丹波市福祉人材確保奨学金返還支援補助金事業について > 1. 補助内容

そこで、隣接する見出しに出現する自立語を比較し、一致率が一定割合以上なら上位の見出し（連結文字列の左側）を消去する。最終的な項目名は以下のようになり、ページ内の記載内容を適切に表現できる。

> 丹波市福祉人材確保奨学金返還支援補助金事業について > 1. 補助内容

本文の整形では、各ノードが適切な長さ・情報量になるように、ノードをマージする。以下のケースで、1 つのノードに複数の見出しの内容を統合した。

- 直接の親子関係にある小見出しが 1 つの場合
- 直接の親子関係にある小見出しが複数あるが、すべての子孫ノードの質問項目・回答内容の本文の文字数の合計が 700 以下の場合

また、直接の子にあたる小見出しが複数存在する場合には、それらの小見出しのリストを回答内容の一部に含める。実際の対話ボットでは、ユーザはこのリストから項目を選択することで、対話的に細部の情報にアクセスする。最後に、回答内容の末尾に元のページの URL を付記する。

## 3 QA チャットボットの構成

### 3.1 対話管理

提案する QA チャットボットは、フローチャートに従って対話状態を管理し、複数ターン対話を行う。

初期状態でユーザ発話 $q$ を受け付けると、システムは以下のように全ノードから最も適切なノード $N_{top}$ を探す。

$$N_{top} = \underset{N_k \in \mathrm{AllNodes}}{\arg\max} \ (Score(q, N_k))$$

ここで上位のノード複数個が同程度のスコアで並んだ場合、$N_{top}$ が決まるまで聞き返しを行う。システムは $N_{top}$ の情報を出力した後、システムの状態を $N_{top}$ に更新する。適切なノードが見つからない場合には「他のワードを試してください」などと別の発話を要求する。

システムの状態があるノード $N_i$ であるとき、$N_i$ が属する対話フローチャート上の各子ノード及び祖先の子ノードの中から検索を行う。もっとも類似度の高いノードが見つかった場合、タイトルと本文の文字列を用いて発話し、そのノードをシステムの状態として更新する。適切なノードが見つからない場合には、システムを初期状態に戻して全ノードから適切なノードの検索を行う。

### 3.2 検索モデル

全対話フローチャートの数千にのぼるノードの中から適切なノードを見つけるには、ユーザのクエリを賢く理解する必要があり、BM25 などの記号マッチングによる手法では不十分である。

柔軟なマッチングを実現する手法として、ニューラルネットワークを利用した検索モデルが数多く提案されている。それらの手法では大量のクエリログと正解文書を用いるのが一般的だが、このようなデータセットを準備するのはコストが大きい。

大規模な訓練データがなくても利用可能な手法としては、ranker-base の手法と content-base の手法がある。ranker-base の手法では、BM25 などの単純な検索モデルで取得できる文書を各クエリの擬似正例とみなし学習を行う [1]。一方、content-base の手法ではパターンなどを用いて ⟨Question, Answer⟩ や ⟨Title, Content⟩ などのペアを収集し、⟨ クエリ, 正解文書 ⟩ ペアとみなす [5, 6]。質の良いペアを収集できれば、後者の精度が前者の精度を上回ることが報告されている。今回収集した各ノードのタイトル・本文ペア ⟨Title, Text⟩ は十分にこの性質を満たしていると考えられることから、content-base の手法を採用する。また、BERT [2] が検索タスクにおいて高い精度を示すことが報告されていることから [6] 本研究でも検索に BERT を用いる。

具体的には、各ノードの ⟨Title, Text⟩ を正例、無作為に抽出した ⟨Title, Text'⟩ を負例として利用し、Relevance(Title, Text)= 1, Relevance(Title, Text')= 0 となるように fine-tuning する。また先行研究 [6] に従い BM25 ベースの検索システム TSUBAKI [7] のスコアとの重み付き平均を計算し、最終的なスコアとして利用する。

## 4 実験

### 4.1 対話フローチャートの評価

提案手法で構築した対話フローチャートを評価するために、丹波市の公式サイト[1]および ハーバード大学[2]の公式サイトからフローチャートの構築を行った。表 1 に取得したフローチャートの統計情報、図 2 に各ノードの平均単語長を示す。図 2 における Base モデルは各ページを 1 ノードとして利用したもの、Simple モデルは H タグを全て 1 ノードとして利用したものである。提案した方法では各ノードが長すぎず短すぎない適切なサイズになっていることがわかる。

[1]https://www.city.tamba.lg.jp
[2]https://www.seas.harvard.edu

| | ハーバード大学 | 丹波市 |
|---|---|---|
| サイト内ページ数 | 7,405 | 5,981 |
| フィルタ後ページ数 | 2,641 | 2,434 |
| フローチャート数 | 2,484 | 2,054 |
| ノード数 | 5,510 | 3,680 |
| タイトル長 | 7.1 | 15.6 |
| 本文長 | 238.8 | 136.8 |

表 1: 各ウェブサイト統計

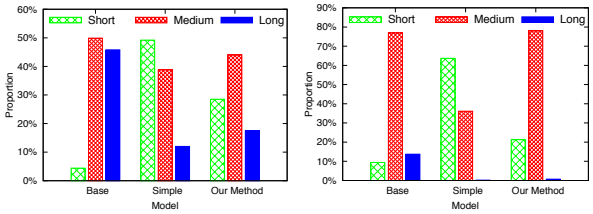


図 2: 構築したフローチャートの平均のノードサイズ（左：丹波市、右：ハーバード大学）。Short は 50 単語以内、Medium は 50 から 500 単語、Long は 500 単語以上を表す。

丹波市公式サイトから構築したフローチャートからランダムに 100 件を選び、ノードのタイトル・本文の適切さを評価した。91 件のノードは応答に利用可能な形式・内容であり、提案手法が知識源の構築に有効であることを確認した。不適切な事例としては、「文化ホール＞バンドフェスタ in Kasuga」というタイトルに対して本文が「平成 24 年度アマチュアアーティスト育成支援事業第 1 弾として...」であり、日付によるフィルタリングが上手く働いていないもの、「医療マップ【氷上地域】」というタイトルに対して本文が「市内で診療をされている医院を地域別に区分して医療マップを掲載しています...」であり、タイトル単体では内容が想像しづらいものがあった。ノード単体で内容が理解できないものや誤情報を含むものなどは見られず、信頼性のある情報源が構築できたと考えられる。

### 4.2 システム応答の精度

ユーザの意図を正しくとらえた検索ができているか確認するため、丹波市のウェブサイトを利用して評価実験を行った。

評価データの構築のため、クラウドソーシングを用いて行政自治体への質問（クエリ）を 1000 件収集した。各クエリに対して TSUBAKI、BERT の検索結果を 5 件出力し、各出力についてクエリとの関係度を 4 段階でラベル付けした。検索結果内に正しい回答が見つからない場合には、キーワードを変えながら検索を行ってクエリに関連するノードを見つけ、上記と同

| Model | SR@1 | SR@5 | NDCG |
|---|---|---|---|
| TSUBAKI | 0.469 | 0.764 | 0.535 |
| BERT | 0.390 | 0.777 | 0.492 |
| T + B | **0.586** | **0.805** | **0.635** |

表 2: 評価結果（T+B は TSUBAKI と BERT の併用）

| | 正解 | 関連 |
|---|---|---|
| TOP1 | 25.7% | 34.3% |
| TOP5 | 36.9% | 31.6% |

表 3: 実クエリを用いた評価

様のラベルを付与した。関連情報が 1 つも見つからないクエリは評価セットから除外した。その結果、785 クエリに対する評価セットが得られた。評価指標には SR@K[3] と NDCG を用いた。

結果を表 2 に示す。TSUBAKI と BERT を併用することで検索結果は大きく改善しており、利用した検索モデルが機能していることがわかる。

また、丹波市・尼崎市で運用されている行政対話ボット [8] に寄せられた実際のクエリをもとに、本システムの有効性の評価を行った。丹波市では、既存 FAQ をベースに人手拡張を行った約 800 件の質問応答ペアを情報源として用いて、行政対話ボットを運用している。この質問応答ペアの中に関連情報がなく回答できなかった実クエリ 73 件を収集し、本システムで回答できるかを確認した。結果を表 3 に示す。ウェブページの情報を利用することで約 70%の質問に回答できており、人手で設計・構築した QA ペアのみを使用した場合と比較して大きな改善が見られた。

表 4 にシステムの出力例を示す。上 2 つは正しく回答できた例である。どちらもユーザの意図に合う情報を提示できていると言える。最後の例は、公式ウェブサイトにも正しい情報がなく、適切な応答ができなかった例である。丹羽市ではイベント情報が観光協会のサイトに集約されており、市の公式サイトには情報が少なかったのが一因であると考えている。複数のウェブサイトの情報を利用することで改善が見込めるのは勿論だが、確信度が低いときは回答を出力しない機能も必要であると考えられる。

| 質問 | | 応答 |
|---|---|---|
| 人権に関する講演会は有りますか | ✓ | 「第 1 回「じんけんセミナー」開催のご案内」についてお答えします。市民一人ひとりの人権が尊重されるまちづくりを進めるため「人権文化をすすめる市民運動」の一環として、「じんけんセミナー」を開催しております。... |
| 女性専用の相談窓口は有りますか | ✓ | 「母と子の健康＞『子育て世代包括支援センター』のご案内」についてお答えします。健やかに安心して出産、子育てをしていただけるよう、氷上保健センター内に... |
| 週末のイベントは、有りますか | × | 「丹波市障がい者スポーツ推進委員会」委員を募集します！＞委員会等日程」についてお答えします。概ね年 3 回程度開催します。（平日昼間に会議を開きます。）10 月開催「ふれあいスポーツの集い」（終日） |

表 4: システムの出力例

## 5 まとめ

本論文では自治体・企業・大学などの公式ウェブサイトを活用して対話フローチャートを自動構築し、これをベースに動作する QA チャットボットを提案した。実験ではウェブサイトから対話フローチャートを生成する提案手法が有効であること、対話時にはユーザの意図を解釈して適切な応答ができていることを示した。

## 謝辞

この研究は国立情報学研究所 (NII) CRIS と LINE 株式会社とが推進する NII CRIS 共同研究の助成を受けています。

### 参考文献

[1] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. Neural ranking models with weak supervision. In *SIGIR2017*, pp. 65–74, Tokyo, Japan, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL2019*, pp. 4171–4186, Minneapolis, Minnesota, 2019.

[3] Asbjørn Følstad and Petter Bae Brandtzæg. Chatbots and the new world of HCI. *Interactions*, Vol. 24, No. 4, pp. 38–42, 2017.

[4] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *ACL2019*, Florence, Italy, 2019.

[5] Sean MacAvaney, Andrew Yates, Kai Hui, and Ophir Frieder. Content-based weak supervision for ad-hoc reranking. In *SIGIR2019*, pp. 993–996, Paris, France, 2019.

[6] Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. FAQ retrieval using query-question similarity and BERT-based query-answer relevance. In *SIGIR2019*, pp. 1113–1116, Paris, France, 2019.

[7] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access methodology. In *IJCNLP2008*, pp. 189–196, Hyderabad, India, 2008.

[8] 田中リベカ, 坂田亘, 柴田知秀, 黒橋禎夫, 橋本泰一. 対話ボットをベースとした行政と市民の新たなコミュニケーションチャネルの構築. 情報処理学会 第 81 回年次大会, pp. 4:415–4:416, 2019.

[3]出力の上位 $k$ 件の中に正解があるものの割合