Constructing Transparent QA Chatbot based on the Official Website Documents

Wataru Sakata
LINE Corporation
Kyoto University
wataru.sakata@linecorp.com

Ribeka Tanaka
Kyoto University
tanaka@nlp.ist.i.kyoto-u.ac.jp

Sadao Kurohashi
Kyoto University
NII, CRIS
kuro@nlp.ist.i.kyoto-u.ac.jp

ABSTRACT

Question answering (QA) chatbots, which answer user questions interactively, are useful for people to get the information they need. When an organization such as a local government, a university, or a company introduces a QA chatbot that can answer inquiries about the organization itself or its services/products, it faces the problem of huge costs of creating and updating question-asnwer pairs and designing dialogue rules.

One may expect that recently-developed machine learning approaches, such as the end-to-end neural dialogue system and the machine reading comprehension system, will alleviate the situation. However, at this point, there are still problems in applying these methods to actual services.

In this paper, we propose a framework to construct a QA chatbot directly from organizations' official websites. Many organizations have official websites with well-maintained information about their services or products. We automatically convert HTML documents on the website to tree-structured data, which we call dialogue flowcharts. Our QA chatbot answers questions based on this dialogue flowcharts. By making dialogue state transition based on the tree structure, the chatbot realizes multi-turn QA.

Since the proposed method operates on only the existing official website as the information source, we can minimize the costs of the introduction and maintenance of QA chatbots. Also, our dialogue flowcharts are in interpretable format, which enables humans to check the information source and detect the cause of errors. Furthermore, we adopt the retrieval method in order to guarantee the correctness of the contents of system outputs. Here we use the efficient retrieval model adopting BERT to capture user intention correctly. In the training stage, we used content-based weak-supervision learning that requires no additionally-constructed training data. The proposed framework realizes a transparent QA chatbot consists of interpretable dialogue flowcharts and a simple retrieval-based architecture, which can be used for actual services that require responsibility for their behaviors.

Our experiments show that (1) our method can construct the dialogue flowcharts of good quality from websites in different languages and domains and (2) the constructed chatbot based on the resultant flowcharts replies with the appropriate information.

CCS CONCEPTS

• **Applied computing → Enterprise applications**.

KEYWORDS

Retrieval-Based Dialogue System, Question Answering, Transparent System

ACM Reference Format:
Wataru Sakata, Ribeka Tanaka, and Sadao Kurohashi. 2020. Constructing Transparent QA Chatbot based on the Official Website Documents. In *Proceedings of AIIS'20.* ACM, New York, NY, USA, 10 pages. https://doi.org/TBA

1 INTRODUCTION

Questions about the products and services of companies or local governments arise on a daily basis. Sometimes the answer is written in a manual or on a web page, but it often takes time to find the right page with a web search and find the information we need. If customer support is active, one solution is to ask them for help via phone or web-chat. However, such cases are limited because running customer center is generally costly.

It is ideal if there is a question answering (QA) chatbot that can automatically respond to such queries. However, at present, building a QA chatbot that handles such specialized contents requires enormous information management costs, including the creation of knowledge sources and scenarios [10, 12]. For this reason, introducing QA chatbots to actual services is not easy.

Under the remarkable progress of the machine learning studies such as end-to-end neural dialogue system and the machine reading comprehension system, one may expect that those modern techniques will alleviate the situation. Building a QA system is, in fact, one of the active research areas: some studies deal with open-domain question answering systems [1, 2, 15], and others aim at multi-turn QA [3, 24]. These studies show that powerful neural methods improve their system performance.

However, there are still some problems in putting those methods to practical use. First, one needs to construct a large dataset. Neural methods require a considerable amount of data for successful training. Since QA chatbots considered here is a kind of low-resource setting, i.e., targeting specific contents, one has to construct a dataset according to their target domain. Even if the target content is relatively generic, utilizing existing large Web corpus or Wikipedia data cannot entirely solve this problem because it may

AIIS'20, July 30, 2020, Virtual Event, China
© 2020 Association for Computing Machinery.
ACM ISBN TBA…$TBA
https://doi.org/TBA

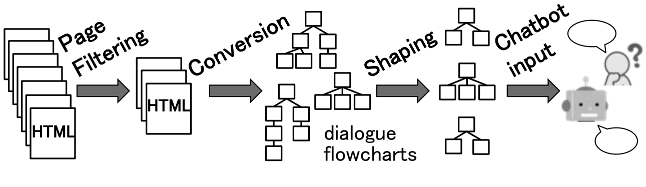


Figure 1: Overview of the framework.

increase the risk of inappropriate utterances (e.g., discriminatory remarks and incorrect information).

Also, neural methods are currently black box-like and not transparent enough. For example, it is still difficult to find a cause of errors and take control of their behaviors. This opacity is often problematic for actual services, which require responsibility for the output information and the behavior of the system [14].

Furthermore, at present, a neural dialogue system that can provide the required information by gradually narrowing down a user's questions in multiple turns, like as in a built-in human-made scenario, has not been realized yet.

In this paper, we propose a simple but practical framework to construct a transparent QA chatbot that answers users' questions based on the information on the organization's official website. In the era of the Internet, organizations generally maintain information about themselves or their products/services in the form of official websites. By making use of the existing content on the official website, we solve the problem of information management costs with regards to introducing QA chatbot.

Figure 1 shows an overview. We automatically convert HTML documents of an official website to tree-structured data, which we call *dialogue flowcharts*. Some previous studies show that the structure of documents is a natural source of the information structure [11, 13]. Our QA chatbot answers the users' questions based on these dialogue flowcharts. Each node of the dialogue flowcharts corresponds to a dialogue system utterance, and, by making dialogue state transition based on the tree structure of dialogue flowcharts, the chatbot realizes multi-turn conversation.

Since the proposed method operates by using only the website as the information source, we can minimize the costs of the introduction and maintenance of chatbots. We show that we can construct the dialogue flowcharts of good quality from organizations' official websites of different languages and domains. Also, our dialogue flowcharts are in an interpretable format, which enables humans to review the information and easily detect the cause of an error appearing in the system answers.

Furthermore, our chatbot is retrieval-based, not generative-based. The retrieval-based approach enhances the safety of the QA chatbot use because it guarantees the correctness of the output content even though it may output the irrelevant information.

Here we use the retrieval model based on BERT [6], which is one of the most effective approaches according to the previous studies [4, 20]. In fine-tuning, we used content-based weak-supervision learning [20, 26], which is reported to show high performance without training data. Our experiment shows that the constructed chatbot replies with the appropriate information, even though we do not have any additional training data of the target domain.

The proposed framework realizes a transparent QA chatbot consists of interpretable dialogue flowchart and simple retrieval-based architecture, which can be used for actual services that require responsibility for their behaviors.

Our contributions can be summarized as follows:

- We propose a way to construct a transparent and smart QA chatbot only from the organization's official websites.
- Our study bridges academic research into real-world solutions by showing a reasonable way to integrate the modern neural methods into QA chatbot services.

The rest of the paper is organized as follows. First, in Section 2, we present a way to convert Web documents to dialogue flowcharts. Then, we explain the design of dialogue management based on the dialogue flowcharts in Section 3. In Section 4, we show experimental results and provide an in-depth analysis. Our experiments show that (1) our method can construct the dialogue flowcharts of good quality from websites in different languages and domains and (2) the constructed chatbot based on the resultant flowcharts replies with the appropriate information. After reviewing related works in Section 5, we give conclusions in Section 6.

2 CONSTRUCTING DIALOGUE FLOWCHART

This paper considers a QA chatbot for an organization such as a local government, a university, or a company, which can talk about the organization itself or its service based on the official website.

Here we have two assumptions on the characteristics of such official website. First, as organizations provide official websites for the users/customers to get the information they want, the information found there is generally detailed enough, up-to-dated, and reliable. Second, to organize the considerable amount of complicated information, pages on official websites typically have relevant document structure and section/subsection titles.

The point of our work is to exploit such document structures to convert the web pages to trees and use them as dialogue scenarios for a dialogue system. We call this tree-structured data *dialogue flowcharts*. Each node of dialogue flowcharts corresponds to a system utterance. We use the tree structure for dialogue management to realize the multi-turn question answering (see, Section 3.1).

2.1 Page Filtering

Some pages in the website contain meaningful information, but others do not. On some pages, for example, there is only a list of page links. The role of such pages is only to gather information. Besides, there are many pages of past information, such as congressional series and city newsletters. The importance of those pages is relatively low for the chatbot users.

Therefore, we filter out pages of the following:

- pages where the length of link text exceeds the length of content text
- pages whose title or URL contains a particular date
- pages that do not contain any heading tags

2.2 Conversion to Trees

Next, HTML pages are converted to the tree-structured data based on their document structure. It is a general view that documents

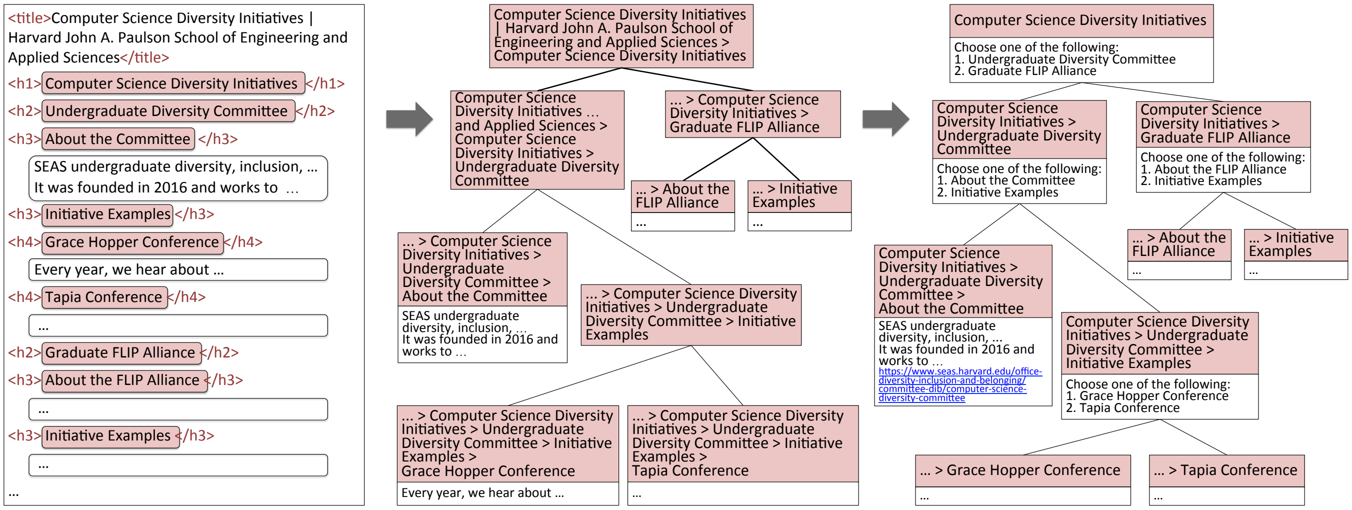

<title>Computer Science Diversity Initiatives | Harvard John A. Paulson School of Engineering and Applied Sciences</title>
<h1>Computer Science Diversity Initiatives</h1>
<h2>Undergraduate Diversity Committee</h2>
<h3>About the Committee</h3>
SEAS undergraduate diversity, inclusion, ... It was founded in 2016 and works to ...
<h3>Initiative Examples</h3>
<h4>Grace Hopper Conference</h4>
Every year, we hear about ...
<h4>Tapia Conference</h4>
...
<h2>Graduate FLIP Alliance</h2>
<h3>About the FLIP Alliance</h3>
...
<h3>Initiative Examples</h3>
...
...
Computer Science Diversity Initiatives | Harvard John A. Paulson School of Engineering and Applied Sciences > Computer Science Diversity Initiatives
Computer Science Diversity Initiatives ... and Applied Sciences > Computer Science Diversity Initiatives > Undergraduate Diversity Committee
... > Computer Science Diversity Initiatives > Graduate FLIP Alliance
... > About the FLIP Alliance
...
... > Initiative Examples
...
... > Computer Science Diversity Initiatives > Undergraduate Diversity Committee > About the Committee
SEAS undergraduate diversity, inclusion, ... It was founded in 2016 and works to ...
... > Computer Science Diversity Initiatives > Undergraduate Diversity Committee > Initiative Examples
... > Computer Science Diversity Initiatives > Undergraduate Diversity Committee > Initiative Examples > Grace Hopper Conference
Every year, we hear about ...
... > Computer Science Diversity Initiatives > Undergraduate Diversity Committee > Initiative Examples > Tapia Conference
...
Computer Science Diversity Initiatives
Choose one of the following: 1. Undergraduate Diversity Committee 2. Graduate FLIP Alliance
Computer Science Diversity Initiatives > Undergraduate Diversity Committee
Choose one of the following: 1. About the Committee 2. Initiative Examples
Computer Science Diversity Initiatives > Graduate FLIP Alliance
Choose one of the following: 1. About the FLIP Alliance 2. Initiative Examples
... > About the FLIP Alliance
...
... > Initiative Examples
...
Computer Science Diversity Initiatives > Undergraduate Diversity Committee > About the Committee
SEAS undergraduate diversity, inclusion, ... It was founded in 2016 and works to ...
https://www.seas.harvard.edu/office-diversity-inclusion-and-belonging/committee-dib/computer-science-diversity-committee
Computer Science Diversity Initiatives > Undergraduate Diversity Committee > Initiative Examples
Choose one of the following: 1. Grace Hopper Conference 2. Tapia Conference
... > Grace Hopper Conference
...
... > Tapia Conference
...

Figure 2: Conversion from an HTML code to a dialogue flowchart.

have a structure of chapters, sections, paragraphs, and sentences [11]. Here we consider a typical style of official websites, where the page contents have relevant document structure that helps to organize the information.

Contents of those HTML pages are generally structured by sections/subsections, each of which has a heading and accompanying description. The heading starts with a larger one, and a smaller heading is placed below it to indicate more detailed content. This heading hierarchy is directly converted into a tree structure, where every node consists of the heading texts and the description below the heading. Figure 2 shows this idea.

We refer to heading tags in the HTML source code to get the hierarchical structure. The heading tag h1 is the biggest one and represents the main content of the document. Thus, the title and accompanying descriptions are associated with the root node. The texts tagged by h2, h3, and so on, are smaller headings and represent the contents of the sections/subsections. Thus, they correspond to descendant nodes. Texts located directly under a heading tag belong to the same node as the heading tag. In this way, each node in the resultant dialogue flowcharts is formed of the title part and the description part.

Note that we do not directly use the original heading text (referred to as *short title*) for the title part. Due to the nature of the section/subsection title in the document, the short title of a root node represents the relevant topic relatively well. However, the short title of a leaf node is too specified to identify the content by itself: for instance, in Figure 2, the short title "About the Committee" is hardly understandable unless one knows a committee of what. Thus, for the title part, we use the strings obtained by conjoining all short titles of the ancestors as well as the page title.

By this conversion, we obtain dialogue flowcharts that preserve the original super-sub relations among contents.

2.3 Shaping Titles and Content Texts

Since each node of dialogue flowcharts corresponds to one system utterance, the node texts should be in an appropriate length, and the contents should be meaningful by itself. Thus, after the conversion of HTML to trees, we modify titles and content texts into a suitable form for the utterance of a QA chatbot.

Titles. The title part of a section/subsection node tends to be redundant if we simply conjoin all the short titles of the ancestors. For example, we obtain the following text, where short titles are conjoined with the symbol ">".

> Computer Science Diversity Initiatives | Harvard John A. Paulson School of Engineering and Applied Sciences > Computer Science Diversity Initiatives > Undergraduate Diversity Committee

Thus, we discard some short titles if the content words highly overlap. Let N_n is our target node, whose path from the root node is represented by $N_0, N_1, ..., N_n$. Let T_n is the short title corresponds to N_n and W_n is a set of content words in T_n. To obtain the label for N_n, we successively compare W_i and W_{i+1} for all $i = 0, 1, ..., n-1$ (i.e., for all adjacent nodes on the path start from the root) and exclude T_i if the following formula holds with an overlap rate r.

$$|W_i \bigcup W_{i+1}| \,/\, min(|W_i|, |W_{i+1}|) \;\geq\; r$$

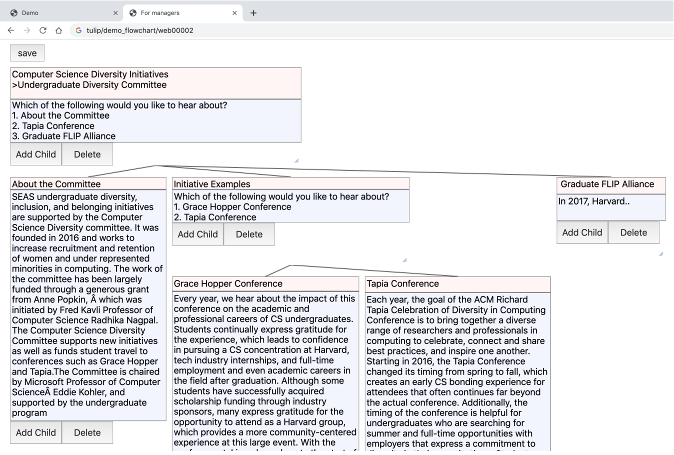
By this operation, we finally obtain the following simplified title.

> Computer Science Diversity Initiatives > Undergraduate Diversity Committee

Content texts. Some pages might contain too small subsections. As each node of our dialogue flowchart corresponds to a system utterance, we merge multiple nodes in the following cases to avoid too short and narrow-scope responses.

- a node has only one child
- the total text length of a node (title and sentences) and its children is less than a length lower limit l

If a node still has children after the merging operation, we insert the string "Choose one of the following" at the end of the original content texts together with the list of children's short titles. For a leaf node, we attach the URL of the original page.



Computer Science Diversity Initiatives
>Undergraduate Diversity Committee

Which of the following would you like to hear about?
1. About the Committee
2. Tapia Conference
3. Graduate FLIP Alliance

About the Committee

SEAS undergraduate diversity, inclusion, and belonging initiatives are supported by the Computer Science Diversity committee. It was founded in 2016 and works to increase recruitment and retention of women and under represented minorities in computing. The work of the committee has been largely funded through a generous grant from Anne Popkin, Â which was initiated by Fred Kavli Professor of Computer Science Radhika Nagpal. The Computer Science Diversity Committee supports new initiatives as well as funds student travel to conferences such as Grace Hopper and Tapia.The Committee is chaired by Microsoft Professor of Computer ScienceÂ Eddie Kohler, and supported by the undergraduate program

Initiative Examples

Which of the following would you like to hear about?
1. Grace Hopper Conference
2. Tapia Conference

Graduate FLIP Alliance

In 2017, Harvard..

Grace Hopper Conference

Every year, we hear about the impact of this conference on the academic and professional careers of CS undergraduates. Students continually express gratitude for the experience, which leads to confidence in pursuing a CS concentration at Harvard, tech industry internships, and full-time employment and even academic careers in the field after graduation. Although some students have successfully acquired scholarship funding through industry sponsors, many express gratitude for the opportunity to attend as a Harvard group, which provides a more community-centered experience at this large event. With the

Tapia Conference

Each year, the goal of the ACM Richard Tapia Celebration of Diversity in Computing Conference is to bring together a diverse range of researchers and professionals in computing to celebrate, connect and share best practices, and inspire one another. Starting in 2016, the Tapia Conference changed its timing from spring to fall, which creates an early CS bonding experience for attendees that often continues far beyond the actual conference. Additionally, the timing of the conference is helpful for undergraduates who are searching for summer and full-time opportunities with employers that express a commitment to

Figure 3: Interface of our dialogue flowchart viewer.

2.4 Visualization Interface

Chatbot designers and engineers may want to review the obtained flowcharts, especially when they find errors in a system utterance. The structure of our dialogue flowcharts is intuitive and straightforward enough for a human to interpret. One can check the content texts as well as the connections among the nodes through our flowchart viewer. Figure 3 shows the screenshot of the interface. This human-readable format makes it easy to keep the system more responsible and transparent in terms of the content of its utterance.

It is also possible to modify dialogue flowcharts using this interface if necessary, although the fundamental solution is to update the official website. One can add/remove nodes and correct titles/texts by simple operations on the screen.

3 PROPOSED DIALOGUE SYSTEM

In this section, we describe how our system operates based on the dialogue flowchart constructed in Section 2. Our system performs dialogue transition using simple rules based on a dialogue flowchart. Since the flowchart is highly interpretable, and the rules are simple, it is easy to control the system and investigate the cause of the error, which makes the overall system to be transparent. We describe the rules of the dialogue transition in detail in Section 3.1.

Also, we adopt an information retrieval system to find an appropriate node for the system utterance. The traditional symbol matching model such as BM25 is difficult to properly respond to a variety of user questions. Therefore, we adopt the weak supervision-based neural ranking model to perform the robust retrieval, by exploiting the constructed flowchart as training data (Section 3.2).

3.1 Dialogue Management

Our proposed chatbot manages dialogue states in accordance with the dialogue flowchart. This enables multi-turn dialogue as shown in the Figure 4.

The system has a state S_t at each time, makes an utterance R_t, and update the state to S_{t+1}. Each state S_t is either initial state ϕ or a certain node N_i in the dialogue flowcharts. When the system receives user's query q in the initial state, the system finds the most suitable node N_{top} from all nodes of all dialogue flowcharts as follows:

$$N_{top} = \mathop{\arg\max}_{N_k \in \text{AllNodes}} (Score(q, N_k))$$

$Score(q, N_k)$ represents the relevance score between the user's query q and a node N_k, which is computed by the retrieval system (see Section 3.2). If multiple nodes are ranked with almost the same score, the system presents them as candidates, and ask the user to select one. If N_{top} is determined, the system outputs the information of N_{top} and updates it's state to N_{top}. If an appropriate node is not found, request another utterance such as "Please try another word".

When the system status is a node N_k, the system searches for the next appropriate node N_{next} from the *adjacent nodes* of N_k, which is mainly composed of children and siblings of N_k. Thereafter, the dialogue makes outputs the information of N_{next}, and the system state is updated to N_{next}. When an appropriate node is not found in the adjacent nodes, the system search an appropriate node again from all nodes as in the initial state. See Algorithm 1 for the details of our dialogue management procedure.

Algorithm 1 Dialogue Algorithms

```
Initialize:
AdjacentNodes : empty list
RetrievalNodes : empty list
AllNodes : All nodes in the dialogue flowcharts
CurrentNode : null
while true do
    Query ⇐ get user's query
    NextNode ⇐ null
    if AdjacentNodes is not empty then
        NextNode ⇐ Search(Query, AdjacentNodes)
    end if
    if NextNode is null and RetrievalNodes is not empty then
        NextNode ⇐ Search(Query, AdjacentNodes)
    end if
    if NextNode then
        AdjacentNodes ⇐ AdjacentNodes $\bigcup$ NextNode.Children
        CurrentNode ⇐ NextNode
        Output(CurrentNode.Info)
        CONTINUE
    end if
    RetrievalNodes ⇐ Search(Query, AllNodes)
    AdjacentNodes ⇐ empty list
    Output(RetrievalNodes)
end while
```

3.2 Retrieval Model

In order to find an appropriate node among thousands of nodes of the dialogue flowcharts, it is necessary to understand the intention of the user's query. We use the neural ranking model, which has recently been reported with high accuracy in search tasks.

Generally the neural ranking model is trained to minimize pairwise loss between training triples consisting of a query q, relevant document d+, and non-relevant document d-. The construction of

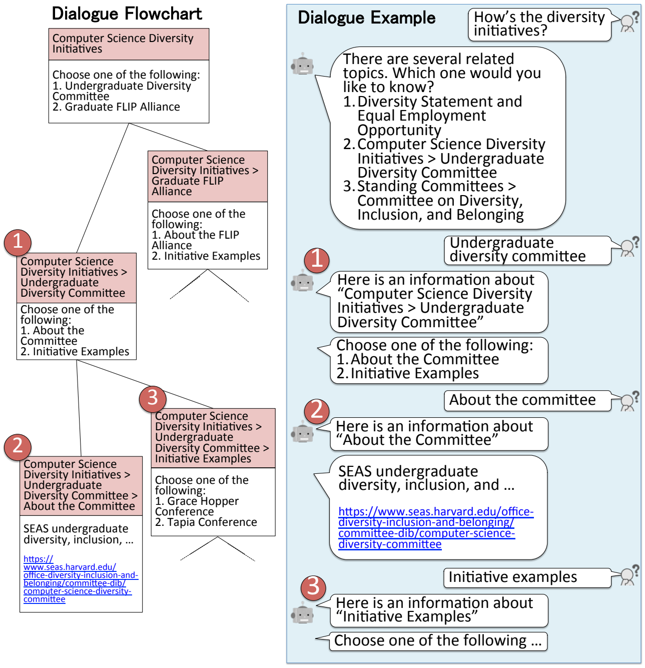

Dialogue Flowchart
Computer Science Diversity Initiatives
Choose one of the following:
1. Undergraduate Diversity Committee
2. Graduate FLIP Alliance
Computer Science Diversity Initiatives > Graduate FLIP Alliance
Choose one of the following:
1. About the FLIP Alliance
2. Initiative Examples
1
Computer Science Diversity Initiatives > Undergraduate Diversity Committee
Choose one of the following:
1. About the Committee
2. Initiative Examples
3
Computer Science Diversity Initiatives > Undergraduate Diversity Committee > Initiative Examples
Choose one of the following:
1. Grace Hopper Conference
2. Tapia Conference
2
Computer Science Diversity Initiatives > Undergraduate Diversity Committee > About the Committee
SEAS undergraduate diversity, inclusion, ...
https://www.seas.harvard.edu/office-diversity-inclusion-and-belonging/committee-dib/computer-science-diversity-committee
Dialogue Example
How's the diversity initiatives?
There are several related topics. Which one would you like to know?
1. Diversity Statement and Equal Employment Opportunity
2. Computer Science Diversity Initiatives > Undergraduate Diversity Committee
3. Standing Committees > Committee on Diversity, Inclusion, and Belonging
Undergraduate diversity committee
1
Here is an information about "Computer Science Diversity Initiatives > Undergraduate Diversity Committee"
Choose one of the following:
1. About the Committee
2. Initiative Examples
About the committee
2
Here is an information about "About the Committee"
SEAS undergraduate diversity, inclusion, and ...
https://www.seas.harvard.edu/office-diversity-inclusion-and-belonging/committee-dib/computer-science-diversity-committee
Initiative examples
3
Here is an information about "Initiative Examples"
Choose one of the following ...

Figure 4: Dialogue flowchart and possible dialogue.

such a data set is highly costly and not suitable for our low-resource and specific-domain purposes.

There are the neural ranking models with weak-supervision. One approach is the ranker-based method [5, 35], that makes use of the query log and simple ranker model (e.g., BM25 [25]) results as a source of training data. The other approach is the content-based model [20, 26], where they collect (Title, Contents) or (Question, Answer) pairs and regarded them as (pseudo-query, relevant-document) pairs, which can be used for training data. It has been pointed out that the accuracy of the content-based method exceeds the accuracy of the ranker-based method if good quality pairs can be collected [19]. Since our collected (Title, Text) pairs of each node are considered to satisfy this property sufficiently, we adopted the content-based method. In addition, we adopt BERT [6], which shows a high score in the retrieval tasks [4, 19].

For each positive example *(title, text)*, we randomly select $\overline{text}$ and produce negative training data *(title,* $\overline{text}$*)*. On this data, we train BERT to solve the two-class classification problem: *Relevance(title, text)* is 1 and *Relevance(title,* $\overline{text}$*)* is 0, where *Relevance(title, text)* stands for the relevance between *title* and *text*. At the search stage, we compute *Relevance(q, text)* for the user's query *q* and *text* in every nodes. Nodes in a higher rank are used as search results. Following the previous research [26], here we adopt the hybrid model, where the weighted sum with the BM25-based model's score is used as the final score.

4 EXPERIMENTS AND EVALUATION

In this section, in order to confirm our proposed QA chatbot is practical, we conduct two experiments and provide an in-depth analysis with real examples. One experiment is the quality evaluation of our constructed dialogue flowcharts (Section 4.1), and the other is the experiment about the accuracy of system responses (Section 4.2).

Table 1: The statistics of each website.

Name	Austin	Harvard	Tamba	Kyoto Univ.
Language	English		Japanese	
Domain	City	Univ.	City	Univ.
#original pages	19,059	7,405	5,981	6,539
#flowcharts	3,682	518	2,054	721
#nodes	8,075	1,662	3,680	2,047
#trees (*height* ≥ 1)	1,000	183	298	173
#trees (*height* ≥ 2)	111	34	48	57
avg title length	11.7	7.2	15.6	35.6
avg text length	489.1	128.1	136.8	333.0

4.1 Dialogue Flowchart Evaluation

To confirm that our system can construct the dialogue flowcharts of good quality from various official websites, we conducted experiments in multi-languages (English and Japanese), and multi-domains (local governments and the universities).

The Table 1 shows four official websites used for our evaluation. For English, we used the websites of city of Austin, which is the capital city of the U.S. state of Texas[1], and Harvard John A. Paulson School of Engineering and Applied Sciences, which is the university in Cambridge, Massachusetts[2]. For Japanese, we used the websites of Tamba city, a city in Japan[3], and Kyoto University, which is a university in Kyoto, Japan[4].

Settings. For each website, we specified an id of the div tag representing the main content and analyzed only the inside of it. We crawled pages of depth ten or less. The page using GET parameter was excluded from the analysis. We used corenlp [21] for English part of speech tagging and juman++ [31] for Japanese. For the title and text shaping, we set the overlap rate to 0.8 and the length lower limit to 700.

Results. See Table 1 for the statistics of each website. "#original pages" shows the number of pages before filtering described in Section 2.1. "#flowcharts" shows the number of the resultant flowcharts, which coincides with the number of the pages after filtering.

We also tested whether each node is in appropriate size, not too long and not too short as a system utterance. For comparison, we introduced two naive baselines, that is, the *Base* method and *Simple* method. The *Base* method simply considers each page as a single node, and the *Simple* method considers every heading tag and following paragraphs as a node, without the shaping process described in Section 2.3. Figure 5 shows the statistics of the word length in each node. We can see that, compared to the naïve methods, more nodes are in medium size, which is more suitable for the response of a QA chatbot.

Table 2 and Table 3 show randomly-selected sample nodes collected from Harvard and Tamba city websites. For each sample, the

[1] http://www.austintexas.gov
[2] https://www.seas.harvard.edu
[3] https://www.city.tamba.lg.jp
[4] http://www.kyoto-u.ac.jp/ja

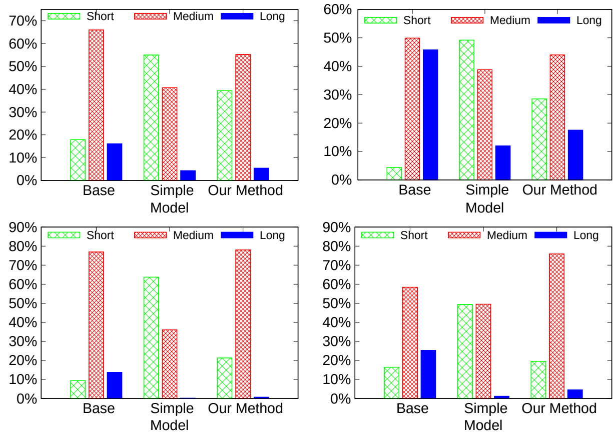


Figure 5: Average node size each model construct from Austin city (Up Left), Harvard (Up Right), Tamba city (Down Left), and Kyoto University website (Down Right). Short stands for less than 50 words, Medium for 50 ~ 500 words, and Long for more than 500 words.

title represents the text content well, and the title-text pairs meaningful by itself. This suggests that our method is applicable enough to construct dialogue flowcharts of good quality from websites.

To evaluate the quality of nodes in our dialogue flowcharts, we sampled 100 nodes at random from the flowcharts constructed from each official website and manually assigned one of the following four categories:

Good The node is informative, namely, the title represents the text content well and the title-text pair is meaningful by itself.

Inappropriate Title The title is redundant or insufficient to represent the outline of the content.

Valueless Information The node does not contain the valuable information for users (e.g., out-dated information)

Error The title-text pair is not understandable.

Table 4 shows the results. We can see most nodes are in good quality and the procedure of dialogue flowchart construction (filtering, conversion, and reshaping) generally works well. Table 5 shows the examples whose label is other than "Good"[5]. In the first example, the title is insufficient because "Committee Info" in the title lacks the information about which committee is in question. In the second example, the information is less valuable for the user, because this is a notification about a meeting in the past, 2019. Although the information is out-dated, it is not successfully removed in the filtering stage. In the third example, the title and text make no sense.

4.2 Evaluation of System Responses

In this section, we evaluate whether our system can respond to user's queries appropriately using the dialogue flowcharts constructed from the official website. Here, we focus on the Tamba-city website introduced in Section 4.1.

In order to test the effectiveness of our retrieval system described in Section 3.2, we perform an evaluation experiment using a traditional search model as a baseline, where we use the evaluation set that includes pairs of a query and the correct answer (a node).

Also, as we discussed in Section 1, the cost of constructing information sources is a significant obstacle in introducing a QA chatbot into real services. To demonstrate that our proposed dialogue flowcharts are useful enough for the information source of real services, we compared our proposed system that operates on automatically-constructed dialogue flowcharts with a system that operates on manually-constructed QA pairs. Here, we take the local government QA chatbot introduced by Tamba city. The Tamba QA chatbot operates on QA pairs constructed manually, like as many other QA chatbots. The number of question-answer pairs it uses is 787.

Note that most of the manually constructed QA pairs and the information obtained from "faq" domain of the Tamba official website are overlapping. We excluded flowcharts obtained from the "faq" domain and use the manually-constructed QA pairs as flowcharts. As a result, the number of dialogue flowchart nodes our system used are 3,912 in total.

Evaluation Data by Crowdsourcing. We first collect 1000 queries to the Tamba local government via crowdsourcing. Examples are shown in Figure 6. For each query, both BM25-based and BERT-based models outputted at most five relevant nodes. Each node was manually assigned one of the following four categories:

A Contain correct information.
B Contain relevant information.
C The topic is same as a query, but do not contain relevant information.
D Contain only irrelevant information.

In general, information retrieval evaluation based on the pooling method has inherently a *biased* problem. To alleviate this problem, when there are no relevant nodes among the outputs by each model, we searched pages in Tamba city with correct information by using different appropriate keywords. If there are no relevant pages found, this query was excluded from our evaluation set. The resultant queries were 795. For 67 queries among the excluded 205 queries, we found its answer on other websites such as Wikipedia or personal websites. The other 138 queries are hardly understandable even for humans or are not related to the domain.

Evaluation Data using Real Service Queries. We also evaluate our retrieval method on real inquiries that Tamba QA chatbot service received. Figure 7 shows some examples. Compared to queries collected via crowdsourcing, real queries are informal and contain more real-time topics.

We sampled 1,075 inquiries that the Tamba QA chatbot received from the citizens and examined whether the Tamba QA chatbot could reply with the appropriate information. For the inappropriate cases, we further categorized them by the cause of the errors. In the end, we obtained 73 queries that the Tamba QA chatbot could not answer correctly due to the lack of relevant QA pairs.

We took these 73 queries and used them as the input of our retrieval component. We manually evaluated the results in order to see the effect of data we constructed from web pages. We used

[5] In the following, examples taken from Japanese websites are translated into English by the authors

Table 2: Randomly sampled nodes from Harvard website. (The URL at the end of the text part is ommitted.)

Title	Text
Data Science FAQs > When and how can I apply? Does the program have rolling admissions?	Higher Degrees (BIO) The SM in Data Science uses the online application for the Harvard Graduate School of Arts and Sciences. You will find it here: https://gsas.harvard.edu/admissions/apply. We do not have rolling admissions. There is a single application deadline each year - December 15th.
Graduate Program in Computer Science	Choose one of the following: 1. Degrees Offered 2. Engineering Sciences (A.B.) - Mechanical and Materi als Science and Engineering Track 3. Mechanical Engineering (S.B.)
Ferran Adrià visits Harvard > Sponsors	Sponsored by the Materials Research Science and Engineering Center at Harvard University. -Hosted by the Harvard Department of Physics and the Harvard School of Engineering and Applied Sciences.
Video Posting > Converting videos from one format to another	If you need to convert your video from one format to another, e.g. from flash to .mp4, one useful and easy tool is HandBrake.Handbrake is available at: http://handbrake.fr/

Table 3: Randomly sampled nodes from Tamba city website. (The URL at the end of the text part is ommitted.)

Title	Text
Disposal of motorcycles and motorbikes	Motorcycles and motorbikes are recycled using a "two-wheeled vehicle recycling system" by voluntary efforts by domestic manufacturers and importers. Please use the motorcycle recycling system by yourself or ask the dealer to dispose of your item. A scrapped car confirmation document is required at the time of disposal. Please dispose of it at the transportation branch office or city counter, depending on the type of the vehicle. For more detail, please contact the Car Recycling Call Center (Phone: 050-3000-0727).
Recruitment of students from Kashiwabara Kinone University > Kashiwabara Kinone University circle committee	The Kashiwabara Kinone University circle committee, which carries out hobby club activities for the elderly, holds various classes for people over the age of 65. If you would like to attend, please contact the Tamba City Hall Kashiwabara Branch or Kashiwabara Resident Center. The membership fee varies depending on the class.
Vaccination > DPT-IPV	About DPT-IPV vaccine. Choose one of the following: (1) Target diseases (2) Target person (3) Standard inoculation period and number of inoculations (4) Important notice (5) Things to bring to a medical institution on the day of vaccination (6) City contract medical institution

Table 4: The node quality evaluation.

Name	Austin	Harvard	Tamba	Kyoto Univ.
Good	89	95	92	97
Inappropriate Title	4	4	5	1
Valueless Information	4	0	3	0
Error	3	1	0	2

the same label as the evaluation data by crowdsourcing. See Table 6 for the statistics of the final resultant evaluation set.

- I'd like you to issue a copy of family register, but how much does it cost?
- I'd like you to publish a maternal and child health handbook, but what is required for the procedure?
- I'm thinking of purchasing a new housing, so I want to know about the reduction measure.
- From which station does the pick-up bus of the Center Pool come out?

Figure 6: Examples of queries collected via crowdsourcing.

Settings. SR@k (Success Rate)[6] and nDCG (normalized Discounted Cumulative Gain) were used as our evaluation measures. The evaluation level of categories A, B, and C were regarded as merely correct for SR@k. For nDCG, A, B, and C were regarded as 3, 2, and 1, respectively.

We adopted TSUBAKI [29] as BM25-based retrieval system. TSUBAKI accounts for a dependency structure of a sentence, not just its words, to provide accurate retrieval. In order for flexible matching,

[6] Success Rate is the fraction of questions for which at least one related question is ranked among the top k.

Table 5: Error analysis of nodes. (The URL at the end of the text part is ommitted.)

Title	Text	Label	Website
Committee Info > Members	We also have a safety-committee email list serve. Safety officers are automatically added; if you are not a lab safety officer and would like to be added, please email Lin.	Inappropriate Title	Harvard
Eastlink Trail Project > Community Engagement > Technical Advisory Group Meetings	TAG Meeting #1 May 29, 2019, 9 to 11 a.m. AISD Performing Arts Center...	Valueless Information	Austin
Open lecture > documents	There are currently no items in this folder.	Error	Kyoto Univ.

- The 5th Kashiwabara Hospital Festival
- School tomorrow
- Is there any weekend event?
- Please tell me if we need to apply again for infant care if the insurance card changes

Figure 7: Examples of real inquiries in the Tamba city.

Table 6: Evaluation set statistics: Avg q length means the average length of queries, and #A means the number of pairs of a query and a node with label A. The same applies to #B and #C.

type	#queries	avg q length	#A	#B	#C
Crowdsourcing	795	13.7	1491	1040	881
Real queries	73	4.7	56	49	56

it also uses synonyms that automatically extracted from dictionaries and Web corpus. The pre-training of BERT was performed using Japanese Wikipedia, which consists of approximately 18M sentences. The fine-tuning was performed using title-text pairs in all nodes of constructed dialogue flowcharts. For both datasets, TSUBAKI was applied only in node titles. We omit the TSUBAKI's results using the text part as we got the worse scores than the only title setting.

Results. Table 7 shows an experimental result on the crowdsourced queries. We can see that the system can output the related information in top-5 candidates to more than 80% of the queries. The hybrid method overwhelmed TSUBAKI (BM25-based model). The weak-supervised BERT ranking model is effective even on our proposed flowchart. T + B (Manual) shows the score when performing a retrieval using only manually-created QA pairs as an information source. We can see that the number of questions our system could answer has increased significantly by using the information on the website.

Table 8 shows the output examples of our retrieval model (translated from Japanese). ✓and × in the table mean correct and incorrect, respectively, where the evaluation categories A, B, and C are regarded as correct. In the first example, the system could answer correctly through symbol matching by TSUBAKI. In the second example, the system answer is correct thanks to BERT, although words of the query and the correct node do not highly overlap with each other. In the third example, the system could not find the correct node. There is the correct node titled "Take measures against harmful birds and beasts," but the system could not find this because it does not know that a "stray dog" is a poisonous beast.

Table 7: Retrieval scores. T+B means the hybrid model of TSUBAKI and BERT.

Model	SR@1	SR@5	NDCG
TSUBAKI	0.469	0.764	0.535
BERT	0.390	0.777	0.492
T + B (Manual)	0.491	0.613	0.372
T + B	**0.586**	**0.805**	**0.635**

Table 9 shows the result on the real service queries. In the table, "Correct" corresponds to the label A and "Related" corresponds to the label B and C. The system founds the correct or related information in the top-5 candidates to nearly 70% of queries. We can see our automatically-constructed QA chatbot can answer the real inquiries that the original Tamba QA chatbot could not answer.

Table 10 shows some examples of queries and the system outputs (translated from Japanese). The system answers appropriately in the first two examples. In both cases, the information that the system provided matches the user's intention. In the last example, the system did not respond correctly because the official website did not contain relevant information. In Tamba City, event information is generally posted on the website of the tourist association, not on the official website. In this case, we can expect significant improvements by using information from multiple websites. Another thing to be considered is to equipped with a function that does not output any response when the confidence is low.

5 RELATED STUDY

Dialogue systems in early days performed conversation using pattern matching based on prepared keywords set (e.g., *ELIZA* [34], *A.L.I.C.E* [33]), which is mainly developed for passing Turing test [32] or for fun.

In recent years, various chatbot services have been launched [9] thanks to improvements in voice recognition and generation functions and the development of data-driven methods [22]. Some chatbots are designed for individual tasks such as booking (e.g., restaurant reservations), selling (e.g., receiving orders), and entertainment (e.g., chit-chat and games). The others are supposed to be used for

Table 8: Examples of system outputs and their manual evaluations (crowdsourced-queries).

Query		Output
Do you have enough support for single mothers?	✓	Here is an information about "The support for the single parent": Please check the page about a support for a single parent. URL : ...
I have just moved here, so I would like to participate if there is an event to know more about this area. Where is the information?	✓	Here is an information about "We are looking for participants of early morning trekking at Kuroi Castle Ruins on Tamba City Sports Day": Residents, students and employees in Tamba City can participate. When children under elementary school age participate, please join with parents. For more details, visit http:// ...
Recently, stray dogs have settled in the park	×	Here is an information about "I became difficult to keep a dog. What should I do?": Responsible to the end, try to find a new owner... If it is unavoidable, ...

Table 9: The experimental results on the real service queries. Showing what percentage of the queries our system can output correct/related information.

	Correct	Related
TOP1	25.7%	34.3%
TOP5	36.9%	31.6%

Table 10: Examples of system outputs and their manual evaluations (real queries).

Query		Output
Is there a lecture on human rights?	✓	Here is an information about "Information for the 1st Human Rights Seminar": As part of the "Citizens' Movement to Promote Human Rights Culture", we hold the "Human Rights Seminar" to promote town development where the human rights of each and every citizen are respected.
Is there a women-only consultation service?	✓	Here is an information about "Health of mother and child > The information of child care generation support center": In Hikami Health Center, you can have a healthy and safe childbirth and raise children ...
Do you have weekend events?	×	Here is an information about "We raise *Person with a disability sports promotion committees in Tamba city*! > Committee schedule": It is held about three times a year. (A meeting is held during the daytime on weekdays.) "Sports Gathering" held in October...

various purposes, called "smart speakers". They are mainly developed by big technology companies (e.g., *Alexa*, *Google assistant*, and *XioIce* [41]).

In the field of NLP/AI research, there are many attempts to realize more powerful conversations using machine learning approaches. For example, the generation based dialogue system is developing fast, which is based on the SEQ2SEQ model using utterance response pairs as training data [27, 28]. This method enables dialogue systems to speak sentences that do not exist in the data and to respond to various questions more flexibly [39].

The machine reading comprehension system, which reads a document and gives appropriate answers to user questions, is also actively studied [8, 23, 40]. Some studies aim to respond appropriately from large-scale documents [1, 2, 15], and others aim to conduct question answering interactively about the content of passages such as children's stories or news articles [3, 24].

However, it is pointed out that these systems are black box-like and sometimes make irresponsible or problematic statements, and there is a growing need for a transparent and honest system [12, 16].

There are a lot of studies on retrieval-based dialogue systems, including single-turn models [7, 36, 38] and multi-turn models [17, 18, 30, 37]. Such systems do not output sentences that do not appear in the search source. As this can reduce the black box-like behavior, our research uses the retrieval-based method. In many existing studies on retrieval-based dialogue systems, training and evaluation are performed by using manually created data sets. However, when we apply the retrieval-based method to new domains, the knowledge source construction and manual annotation cost a lot. The method proposed in this study does not require any additional training data, even in low-resource languages and domains.

One of the biggest motivations of this study is also to establish a framework to construct transparent QA chatbot with the minimum cost.

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed a framework to construct a transparent QA chatbot directly from an organization's official website. Since the proposed method operates on only the existing official website as the information source, we can minimize the costs of the introduction and maintenance of QA chatbots. To understand user's intention deeply, we adopted the weak supervision-based neural ranking model, which works well without additionally-constructed training data. In the experiment, we demonstrated that (1) our method can construct the dialogue flowcharts of good quality from website and (2) the constructed chatbot replies with the appropriate information considering the user's intentions. We are planning to evaluate and sophisticate multi-turn dialogue as future work.

REFERENCES

[1] Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2019. Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering. In *ICLR*. Addis Ababa, (to appear). https://openreview.net/forum?id=SJgVHkrYDH

[2] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *ACL*. Association for Computational Linguistics, Vancouver, Canada, 1870–1879.

[3] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *EMNLP*. Association for Computational Linguistics, Brussels, Belgium, 2174–2184. https://doi.org/10.18653/v1/D18-1241

[4] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *SIGIR (SIGIR'19)*. ACM, New York, NY, USA, 985–988. https://doi.org/10.1145/3331184.3331303

[5] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *SIGIR*. ACM, Tokyo, Japan, 65–74.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Jiazhan Feng, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2019. Learning a Matching Model with Co-teaching for Multi-turn Response Selection in Retrieval-based Dialogue Systems. In *ACL*. Association for Computational Linguistics, Florence, Italy, 3805–3815. https://doi.org/10.18653/v1/P19-1370

[8] David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31, 3 (2010), 59–79.

[9] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the New World of HCI. *Interactions* 24, 4 (June 2017), 38–42. https://doi.org/10.1145/3085558

[10] Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for Customer Service: User Experience and Motivation. In *CUI (CUI '19)*. ACM, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3342775.3342784

[11] Mor Geva and Jonathan Berant. 2018. Learning to Search in Long Documents Using Document Structure. In *ACL*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 161–176.

[12] Jonathan Grudin and Richard Jacques. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *CHI (CHI '19)*. ACM, New York, NY, USA, Article 209, 11 pages. https://doi.org/10.1145/3290605.3300439

[13] Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Mike Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic Text Matching for Long-Form Documents. In *WWW*. ACM, San Francisco, CA, USA, 795–806.

[14] Kit Kuksenok and Nina Praß. 2019. Transparency in Maintenance of Recruitment Chatbots. , 4 pages. arXiv:cs.AI/1905.03640

[15] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *ACL*. Association for Computational Linguistics, Florence, Italy, 6086–6096. https://doi.org/10.18653/v1/P19-1612

[16] Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019. Exploring Social Bias in Chatbots using Stereotype Knowledge. In *Proceedings of the 2019 Workshop on Widening NLP*. Association for Computational Linguistics, Florence, Italy, 177–180.

[17] Wentao Ma, Yiming Cui, Ting Liu, Dong Wang, Shijin Wang, and Guoping Hu. 2020. Conversational Word Embedding for Retrieval-Based Dialog System. In *ACL*. Association for Computational Linguistics, To Appear. ToAppear

[18] Wentao Ma, Yiming Cui, Nan Shao, Su He, Wei-Nan Zhang, Ting Liu, Shijin Wang, and Guoping Hu. 2019. TripleNet: Triple Attention Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *CoNLL*. Association for Computational Linguistics, Hong Kong, China, 737–746. https://doi.org/10.18653/v1/K19-1069

[19] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized Embeddings for Document Ranking. In *SIGIR*. ACM, Paris, France, 1101–1104. https://doi.org/10.1145/3331184.3331317

[20] Sean MacAvaney, Andrew Yates, Kai Hui, and Ophir Frieder. 2019. Content-Based Weak Supervision for Ad-Hoc Re-Ranking. In *SIGIR*. ACM, Paris, France, 993–996. https://doi.org/10.1145/3331184.3331316

[21] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL*. Association for Computational Linguistics, Baltimore, Maryland, 55–60. https://doi.org/10.3115/v1/P14-5010

[22] Zhenhui Peng and Xiaojuan Ma. 2019. A survey on construction and enhancement methods in service chatbots design. *CCF Transactions on Pervasive Computing and Interaction* 1 (24 Sep 2019), 204—-223. https://doi.org/10.1007/s42486-019-00012-3

[23] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*. Association for Computational Linguistics, 2383–2392.

[24] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266. https://doi.org/10.1162/tacl_a_00266

[25] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. 1992. Okapi at TREC. In *TREC*. 21–30.

[26] Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. FAQ Retrieval Using Query-Question Similarity and BERT-Based Query-Answer Relevance. In *SIGIR (SIGIR'19)*. ACM, New York, NY, USA, 1113–1116. https://doi.org/10.1145/3331184.3331326

[27] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building End-to-end Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI (AAAI'16)*. AAAI Press, 3776–3783. http://dl.acm.org/citation.cfm?id=3016387.3016435

[28] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *ACL*. Association for Computational Linguistics, Beijing, China, 1577–1586. https://doi.org/10.3115/v1/P15-1152

[29] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. 2008. TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology. In *IJCNLP*. ACL, Hyderabad, India, 189–196.

[30] Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One Time of Interaction May Not Be Enough: Go Deep with an Interaction-over-Interaction Network for Response Selection in Dialogues. In *ACL*. Association for Computational Linguistics, Florence, Italy, 1–11. https://doi.org/10.18653/v1/P19-1001

[31] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A Morphological Analysis Toolkit for Scriptio Continua. In *EMNLP*. Association for Computational Linguistics, Brussels, Belgium, 54–59. https://doi.org/10.18653/v1/D18-2010

[32] A. M. Turing. 1950. Computing Machinery and Intelligence. *Mind* 59, 236 (1950), 433–460. http://www.jstor.org/stable/2251299

[33] Richard Wallace. 2009. *The anatomy of A.L.I.C.E.* 181–210. https://doi.org/10.1007/978-1-4020-6710-5_13

[34] Joseph Weizenbaum. 1966. ELIZA&Mdash;a Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun. ACM* 9, 1 (Jan. 1966), 36–45. https://doi.org/10.1145/365153.365168

[35] Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2018. Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots. In *ACL*. Association for Computational Linguistics, Melbourne, Australia, 420–425. https://doi.org/10.18653/v1/P18-2067

[36] Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2018. Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots. In *ACL*. Association for Computational Linguistics, 420–425.

[37] Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2019. A Sequential Matching Framework for Multi-Turn Response Selection in Retrieval-Based Chatbots. *Computational Linguistics* 45, 1 (March 2019), 163–197. https://doi.org/10.1162/coli_a_00345

[38] Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, Zhoujun Li, and Jianshe Zhou. 2016. DocChat: An Information Retrieval Approach for Chatbot Engines Using Unstructured Documents. In *ACL*. Association for Computational Linguistics, Berlin, Germany, 516–525. https://doi.org/10.18653/v1/P16-1049

[39] Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W. Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A Hybrid Retrieval-Generation Neural Conversation Model. In *CIKM*. ACM, Beijing, China, 1341—-1350.

[40] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *EMNLP*. Association for Computational Linguistics, Brussels, Belgium, 2369–2380. https://doi.org/10.18653/v1/D18-1259

[41] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *CoRR* abs/1812.08989 (2018), 26.