Strategies for Mixed-Initiative Conversation Management using Question-Answer Pairs

Wilson Wong, Lawrence Cavedon, John Thangarajah, Lin Padgham
RMIT University, Melbourne, Australia
{wilson.wong,lawrence.cavedon,john.thangarajah,lin.padgham}@rmit.edu.au

ABSTRACT
One of the biggest bottlenecks for conversational systems is large-scale provision of suitable content. In this paper, we present the use of content mined from online question-and-answer forums to automatically construct system utterances. Although this content is mined in the form of question-answer pairs, our system is able to use it to formulate utterances that drive a conversation, not just for answering user questions as has been done in previous work. We use a collection of strategies that specify how and when the question-answer pairs can be used and augmented with a small number of generic hand-crafted text snippets to generate natural and coherent system utterances. Our experiments involving 11 human participants demonstrated that this approach can indeed produce relatively natural and coherent interaction.

KEYWORDS: conversational system; question-answer pairs; conversational strategies.

1 Introduction

A major bottleneck for any kind of *conversational system* is provision of sufficient content, and if one wishes to avoid repetition and to cover the many possible user inputs, there must be a substantial amount of content. In the case of chatbots such as ALICE (Wallace, 2009), the content is custom-crafted to be as generic and deflective as possible to cover the needs of open-ended conversations. This approach, which reduces the amount of content that the system needs despite the relatively broad range of user inputs, is essentially only capable of content-free, small talk. This kind of conversational ability is, however, inadequate for virtual characters that need to contribute relevant content in their conversations. In systems that require domain-specific content, for example about health in a virtual nurse (Bickmore et al., 2009) or about children-related topics in an intelligent toy (Chen et al., 2011), the custom-crafting of content to cover the depth and breadth of the domain as well as to keep it current becomes an impractical task. This work reports on our use of an abundant type of data, sourced from the Web, to equip a conversational agent with the content it needs to engage users in a coherent and natural way over a wide range of domains.

To overcome the bottleneck of custom- or hand-crafting conversational content, researchers have recently investigated mining online resources such as Web pages returned by search engines (e.g., `Yahoo!` (Shibata et al., 2009)). The problem with this approach, however, is that most content on the Web is not suitable for conversational agents, for various reasons. The verbose, non-colloquial and monologue nature of much of Web text is not a good match for the characteristics of human-human dialogue. However, there are some other sources on the Web that have more potential for this purpose, such as human-written comments on news websites (e.g., `NYTimes.com` (Marge et al., 2010)) and online forums (e.g., `RottenTomatoes.com`

Proceedings of COLING 2012: Technical Papers, pages 2821–2834,
COLING 2012, Mumbai, December 2012.

(Huang et al., 2007), `2ch.net` (Inoue et al., 2011)). We use text of this nature in the current work for obtaining content with which our conversational agent constructs its utterances.

In this paper, we propose the use of question-answer (QA) pairs from community-driven question-and-answer websites such as `AskKids.com` and `Answers.com` as content for our conversational agent. QA pairs from the Web are essentially individual pairings of questions and the corresponding answers contributed by online communities. By nature, the QA pairs extracted from a particular source on a certain topic are disjointed, in that they do not have any temporal or structural information that could immediately lend themselves to straightforwardly building conversations. We see the potential of QA pairs for use in a conversational agent as they are a reasonable embodiment of human-human communication, unlike text extracted from other sources such as Wikipedia or online news articles. The abundance and self-contained nature of QA pairs means that they can be extracted easily over a wide range of topics covered by the respective online communities. QA pairs have also been used in the past with some success for interactive question answering (IQA), driven solely by the user's information needs. We developed an IQA system (Wong et al., 2011) in the past that supports interactivity and does not require inputs to be formulated as questions. This system, however, still interprets inputs as questions and service them with answers. The approach in this paper, on the other hand, is able to share the initiative with the human user for determining conversation content and direction, and is able to engage with the user using a range of different *conversational strategies* such as 'question asking', 'fact telling' and 'question answering'. Context is maintained, similar to our previous work (Wong et al., 2011), but instead of using this purely to identify a question, which is then answered, it is now used to identify a QA pair of relevance for crafting system utterances to further a conversation.

The key contributions of this paper lie in the conversational strategies that we describe that specify how the different parts of the QA pairs can be used and combined with a small number of generic hand-crafted fragments to generate natural system utterances. These system contributions, when interplayed with cooperative user inputs, will produce seemingly coherent conversations. Unlike existing dialogue systems that rely on deep language processing to achieve 'understanding' from input (e.g., (Schulman et al., 2011)), our conversational agent relies only on shallow analysis for efficiency to extract and assign weights to terms and to identify domain information for building *conversational context*. To examine the level of coherence and naturalness that our system can achieve in the absence of standard dialogue management, custom language resources, and deep semantic analysis of user inputs, we perform a pilot study involving 11 human assessors. The results, which show that our approach does indeed produce natural and coherent interaction relative to the human assessment, are discussed in Section 5. Prior to that, we look at the state of the art in conversational agents in terms of their dialogue management approaches and language resource requirements in Section 2. In Section 3, we provide some background on the system architecture and the overall process, including the structure of our QA pairs, the processing of input and the building of context. We then describe in Section 4 the way in which we use the QA pairs according to several strategies, to form a variety of system utterances, using these to build natural, reasonably coherent conversations. After the system evaluation in Section 5, we conclude in Section 6 with discussion of future work to incorporate this into a more complete virtual companion, integrating these conversational agent abilities into a fully fledged companion with a range of capabilities.

2 Related Work

Mining content from the Web to support human-computer interaction applications (e.g., question answering (Radev et al., 2001; Clarke et al., 2001; Yao et al., 2012), interactive question answering (Wong et al., 2011), conversational systems (Huang et al., 2007; Wu et al., 2008; Shibata et al., 2009)) has been investigated since the advent of the read-write Web. In particular, the practice of mining the Web to alleviate the bottleneck of conversational content acquisition is increasing in popularity. Huang et al. (Huang et al., 2007), for example, proposed a supervised technique for extracting `<thread-title><reply>` pairs from online forums as 'knowledge' for chatbots. The authors used `RottenTomatoes.com` as their source of title-reply pairs. The extraction is done in a cascaded manner, given the semi-structured and relatively noisy nature of text in online forums. All replies relevant to the thread title are first extracted using an SVM classifier based on features related to the structure and content. The candidate pairs are then ranked according to the quality of the content, and the top *n* pairs are selected as chatbot 'knowledge'. However, no details were provided on how the pairs could be used in a chatbot. Yao et al. (Yao et al., 2012), on the other hand, describe a technique for extracting factoid QA pairs from standard text, such as Wikipedia[1]. In their approach, natural language processing techniques are used to create a QA pair from a declarative sentence. These QA pairs are then used by virtual characters to answer queries from a user. As such, they are only using the QA pairs reactively, whereas our novel proposal is to also use them proactively. Moreover, by mining our content from social media sites, we are more likely to obtain more conversational fragments, as opposed to the strictly factoid style QA pairs that Yao et al. are interested in. However, they do demonstrate encouraging performance for their techniques and it may be possible to use such an approach to further increase our content with such (factoid) QA pairs.

Overall, despite the rise in interest in deriving content from the Web for conversational systems, there is little research on how this content can be used more effectively to maximise the traits that should ideally be present in human-computer dialogue. These traits include the coherence of system-generated outputs with respect to preceding exchanges, the naturalness of system utterances (e.g., is the utterance too lengthy or formal for supposedly casual conversations), and the ability to support mixed-initiative interactions (i.e., no strict adherence to whether the system or the user leads the conversations). In the next section, we present our approach to a system which mines the Web for QA pairs, and specifies how and when this content can be used to generate coherent, natural system contributions to mixed-initiative conversation.

3 System Overview

An overview of the main components of our conversational system is shown in Figure 1. We briefly describe the collection of QA pairs, and the *Input Analysis*, *Context Management* and *QA pair Retrieval and Ranking* components in this section. We focus in more detail in Section 4 on the *Strategiser*, which is the major novel contribution of this paper.

3.1 QA Pair Collection

The content for generating system utterances is primarily a set of QA pairs, mined from community-driven question-and-answer websites. During output generation, the different parts of the selected QA pairs are augmented by a small number of strategy-specific fragments, which we describe together with the strategies, to create natural system utterances. As our focus

[1]They actually use the Simple English version of Wikipedia.

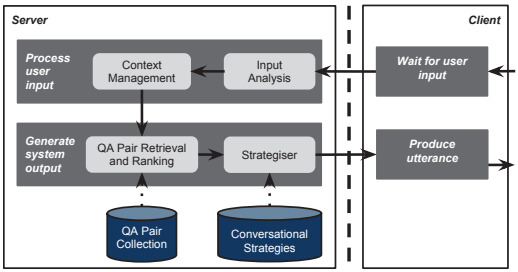


Figure 1: Conversational system architecture

is a companion for children, in this work we have used websites such as `AskKids.com` and `Answers.com` for collecting our QA pairs. This collection is constructed offline, using an interface similar to that described in (Wong et al., 2011) which allows the system administrator to specify the source (e.g., *"askkids*"), domain (e.g., *"animal*") and a seed word (e.g., *"lion*") to represent a topic in the domain for localising and indexing QA pairs. Each QA pair consists of the question and the answer to that question, and some metadata, which are the domain tag, the topic tag and the source tag. These QA pairs are then organised into a database according to the metadata. The tags are used to move between topics of the same domain during a conversation, and to define the scope (i.e., subset of fragments) that the conversational agent has to deal with at any one time. Currently, the collection contains a total of 23,295 QA pairs across 150 topics in the animal domain.

3.2 Input Analysis

Analysis of input utterances is relatively shallow, deliberately avoiding any attempt at sophisticated natural language understanding. The input is first tagged with parts of speech using FastTag[2] for its speed. Noun phrases are then identified using simple regular expression patterns. Pronouns are resolved to the most prominent entities from previous inputs, using a method loosely similar to the backward looking centering approach (Mitkov, 2001). The phrases and words are then assigned weights to reflect their content bearing property. These weighted phrases and words are referred to as terms. Our system has two different weighting schemes implemented, the deviation from Poisson approach (Church and Gale, 1995) and *tf-idf*, with the former being the default. The higher the weight of a term, the more content-bearing it is. Content-bearing terms play a more discriminatory role in selecting relevant QA pairs for utterance generation during the retrieval and ranking process. Stopwords, on the other hand, are removed by virtue of them being non-content bearing. In addition, a dictionary-based approach, operating over the input string, is used to detect if an input is an explicit question or a request for session termination.

3.3 Context Management

Context, which is essentially a collection of weighted terms decayed over time, is used to select QA pairs which are sufficiently relevant to the user inputs to generate system utterances. By

[2] markwatson.com/opensource

collecting the weighted terms from the user inputs (and potentially system output), and then decaying them over time, we maintain sufficient contextual information to make good choices of QA pairs for building system utterances. Each time a new input is processed, the corresponding weighted terms are combined into the existing conversational context. For example, if the current input is the $5th$ utterance by the user, then the context at that turn would contain all the terms and their weights extracted from the previous four inputs. The context management process at turn 5 would then assimilate the terms from the recent input into the current context to create a revised conversational context for turn 6.

The rule for combining new terms to the context is as follows. If a term is already present in the context, the weight associated with the recent occurrence is added to the term's existing decayed weight in the context to reinforce what we perceive as an important term. If the term is absent from the context, the new term and its weight are added to the context. In other words, recurrence is an important factor for a term to maintain its prominence in the context, where the more times a term occurs in the conversation, the more significant it will become through the compounding of its weights. During the process of combining, the weights of all other terms in the context that are not encountered in the most recent input are decayed based on the turn in the conversation, the decay factor and the part of speech. Those term used less recently thus have a greatly reduced weight. Different parts of speech are decayed at different rates, with nouns decaying less quickly than verbs, which decay slower than adjectives, adverbs and other parts of speech. Terms are decayed according to $v' = v \exp(-t\lambda\alpha)$ where v' is the decayed weight, v is the original weight, λ is the decay factor set to 0.8 during our experiments. The α value has been set to the following empirically-selected values to reflect the bias that we introduced according to a term's part of speech: 0.25 for nouns, 0.75 for adjectives and adverbs, 1.25 for verbs, and 2.50 for others.

3.4 QA Pair Retrieval and Ranking

QA pair retrieval and ranking are performed to determine the QA pair that has the highest relevance to the conversational context. Initially, a set of all candidate QA pairs containing at least one term from the context is retrieved. The relevance of the candidates with respect to the conversational context is then determined in terms of (1) the extent of the overlap of words in the question and answer parts of QA pairs with those in the conversational context, and (2) the string similarity between the question component and the user input. Edit distance (Levenshtein, 1966) is used to determine the similarity between the user input and the question part of the candidate QA pairs. As for overlap, the more highly weighted terms from the conversational context that appear in a QA pair, the higher the pair's score will be. The sum of the weights of terms from the context that appear in the question as well as the answer of a QA pair is determined. This sum is augmented by the location of appearance, where term matches in the questions are scored twice as high as matches in the answers. At the end of retrieval and ranking, the top scoring QA pair is selected for utterance generation.

4 Conversational Strategies

The core of our system's abilities are two *reactive* and four *proactive* strategies. These strategies attempt to recreate the types of human-to-human communications that are commonly encountered. The strategies is also divided into *progression* or *conclusion* depending on their roles, to either progress or conclude a conversation. The content used by these strategies to produce utterances are the QA pairs and a small number of specialised *hand crafted fragments*

which include speech disfluencies to produce more natural outputs (Marge et al., 2010). These fragments are assigned tags and are contained in the strategy library. This library can be extended to increase the variety of specialised fragments, independently of the actual strategies.

4.1 Reactive Strategies

This group of strategies is characterised by the agent following the user's lead. There are two strategies in this group, the first being a *reactive progression* strategy called UAQ, and a *reactive conclusion* strategy called UEC. The details are discussed below.

UAQ (User Asks Questions) Strategy: This strategy is initiated when the input parsing detects that the user has asked a question. In this case the keywords from the parsed input, weighted as described in *Context Management*, are used to select a QA pair where the context has an optimal match with the question portion of the QA pair. The answer portion of the pair (or if it is too long, the initial part) is then used as output. If it is desired to continue this strategy beyond a single interaction in the conversation, then subsequent inputs are essentially treated as refinements of the initial question. This is essentially the approach used in our IQA system (Wong et al., 2011). We do not use any specialised fragments to augment the selected QA pair in this strategy.

UEC (User Ends Conversation) Strategy: This strategy is used when input processing detects signs that the user wants to conclude the conversation. When a termination is detected, the Strategiser removes all pending outputs in the queue and adopts this strategy to access the associated fragments shown in Table 1, which are used to generate farewell messages.

1	OK. Nice chatting with you. Cya.	[UEC]
2	Bye.	[UEC]

Table 1: Some fragments for the UEC strategy

4.2 Proactive Strategies

We have five proactive strategies, two for progressing and three for concluding a conversation. The *proactive progression* strategies are the ones that we make most use of in our conversational agent. In these strategies the system takes the initiative by either asking a question or sharing some information. The three *proactive conclusion* strategies deal with certain circumstances that necessitate the conclusion of conversation sessions, namely, (1) lack of user participation, and (2) system's inability to carry on the conversation. These conclusion strategies are the result of the system interpreting various signs and initiating the ending. The conclusion strategies do not use any components of QA pairs for generating system utterances.

SSK (System Shares Knowledge) Strategy: This strategy attempts to simulate conversations involving the sharing of knowledge between two participants in a chat. In the SSK strategy, the system assumes the role of imparting knowledge, with the user on the receiving end. This strategy uses the answer component of QA pairs to formulate system utterances to create this effect. To illustrate, consider the following QA pair:

```
Q: What is a panda?
```

```
A: The Panda is the bear-like black and white mammals which lives in
China. These cute and cuddly like bears eats mainly bamboo shoots
and are becoming extinct. Some experts refer to them as bears while
others believe they are more like a raccoon.
```

The SSK strategy uses the first sentence in the answer and augments it with a selection of fragments such as *"Did you know that $X*" and *"I just learnt that $X*". In this example, the first system utterance generated using this strategy, which will be placed in a queue, is *"Did you know that the panda is the bear-like black and white mammals which lives in China.*". The subsequent sentences in the answer are stored in the queue for generation but their order of use will depend on the evolving conversation context in the next n iterations. For example, assume that the system sends the first utterance in the queue to the user, and the user provides *"They look like a racoon*" as the next input. Since this new input is not an explicit question and the queue of the next things to utter is not empty, the strategy remains the same, i.e., SSK, and the system utterance generation process does not invoke the QA pair retrieval and ranking process. The Strategiser is used to reorder the utterances in the queue based on their overlap with the current conversational context that contains the term *"racoon*". Since the overlap of the words in the second sentence in the queue (i.e., the third sentence in the answer of the selected QA pair) with the context is likely to be higher (due to the word *"racoon*"), this sentence will be selected, augmented with the appropriate fragments and moved to the front of the queue as the next system utterance. This next utterance will appear as *"And some experts refer to them as bears while others believe they are more like a raccoon*". In the third iteration, only one sentence from the 3-sentence answer remains. Unless the next user input is a question, the system will generate its next utterance as *"What's more is that these cute and cuddly like bears eats mainly bamboo shoots and are becoming extinct*". This process of reordering the sentences that the system utters takes place for all sentences in the answer of the selected QA pair.

1	Did you know that $\$X$?	[SSK-1]
2	I just learnt that $\$X$.	[SSK-1]
3	What's more, $\$X$.	[SSK-2]
4	What's more is that $\$X$.	[SSK-2]
5	And $\$X_3$... Well... $\$X$.	[SSK-2]
6	And $\$X_3$... Umm... $\$X$.	[SSK-2]

Table 2: Some fragments for the SSK strategy

Table 2 shows the list of fragments associated with the SSK strategy. Fragments labelled [SSK-1] are used by the strategy to prepend to the first sentence of the selected answer, while those labelled [SSK-2] are used for augmenting the subsequent sentences. Fragments 5 and 6 are three examples containing false starts, where $\$X_3$ and $\$X_4$ represents the first three and four words in $\$X$, respectively.

SAQ (System Asks Questions) Strategy: This strategy attempts to recreate a conversational scenario where a conversation participant explicitly asks the other participant a question. In this strategy, the system poses as the participant asking the question. Both the question as well as the answer components of QA pairs are used to generate the system utterances in this strategy.

1	Can you tell me, $\$X$?	[SAQ-1]
2	Tell me, $\$X$?	[SAQ-1]
3	Erm... $\$X$.	[SAQ-2]
4	$\$X_3$... Well... $\$X$.	[SAQ-2]

Table 3: Some fragments for the SAQ strategy

When this strategy is selected, the question component of the selected QA pair is used to construct the first system utterance. Fragments shown in Table 3 which are labelled [SAQ-1] are used to augment the question. The second utterance is then constructed using the first three sentences of the corresponding answer and a fragment tagged with [SAQ-2]. With regard to [SAQ-2] fragments, 3 contains fillers while 4 contains a false start to add to the 'natural' effect of the resulting utterances. To illustrate, consider the QA pair above. The first and second utterances in the queue would appear as *"Can you tell me, what is a panda?*" and *"The Panda is... Well... The Panda is the bear-like black and white mammals which lives in China. These cute and cuddly like bears eats mainly bamboo shoots and are becoming extinct. Some experts refer to them as bears while others believe they are more like a raccoon.*". The use of only the first three sentences in the answer to generate the second system utterance is motivated by the average sentences in answers of QA pairs from AskKids.com being around three.

UNI (User Not Interested) Strategy: This strategy is chosen by the Strategiser when input processing detects three empty input sentences in a row, likely to be the result of the expiration of the 60-second timer imposed by the input waiting process at the client side. The system assumes the three successive timeouts to be a sign of the user losing focus to other tasks, or lacking interest in pursuing the conversation further. Similar to the UEC strategy, the Strategiser cancels all pending utterances when this strategy is chosen, and uses one of the fragments lablled [UNI] shown in Table 4 to generate the system's final utterance.

1	You don't seem very interested in continuing the chat. Let's do this some other time. Bye!	[UNI]
2	It seems that you're in the middle of something. Let's talk some other time. Cya.	[UNI]

Table 4: Some fragments for the UNI strategy

LTT (Lost Train of Thought) and SEC (System Ends Conversation) Strategies: The second circumstance that necessitates the conclusion of a conversation, from the system's point of view, is the repeated inability of the retrieval and ranking modules to select a QA pair that the Strategiser can work on using the conversational context maintained by the system. There are two main causes of this: (1) the user's successive use of non-content bearing words in his or her inputs, and (2) the exhaustion of QA pairs. Two strategies are involved in this circumstance. Whenever the Strategiser does not receive a selected QA pair from the retrieval and ranking modules, the LTT strategy is used to generate utterances with fragments, shown in Table 5, designed to prompt the user to repeat the recent input(s). When this strategy is used, the conversational context is emptied and rebuilt using the latest input with the hope that the new context can be used to select some useful QA pairs. In the event that the LTT strategy is being selected for the third time in a row, the system will consider this as a sign that it is unable to further pursue the conversation. In such cases, the Strategiser uses the SEC strategy

to conclude the current session. Table 6 shows the fragments associated with the SEC strategy for ending a conversation.

1	Oops. I got distracted. You were saying?	[LTT]
2	Sorry, can you please repeat what you just said?	[LTT]
3	Where were we? I was talking to someone else.	[LTT]

Table 5: Some fragments for the LTT strategy

1	I've got something that I have to attend to now. Talk to you later.	[SEC]
2	Sorry. I'm tired now. Let's continue this some other time. Bye.	[SEC]
3	We'll need to stop here now. I've got to run. Bye.	[SEC]

Table 6: Some fragments for the SEC strategy

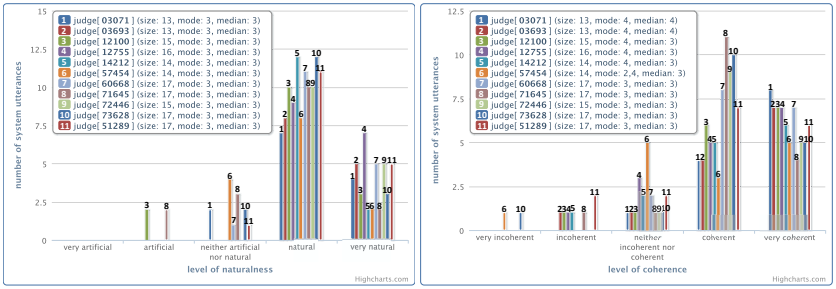
5 Evaluation

We conducted a pilot study to assess the naturalness and coherence of utterances generated by our conversational agent using only QA pairs and 41 generic fragments. We first discuss the experimental setup, before moving on to present the results of humans' judgements of our system outputs. Error analysis is also performed, where we analyse the causes of those system utterances that were perceived as less natural and coherent by the judges. An actual interaction between the system and one of the judges is presented at the end to illustrate a typical conversation with the conversational agent.

5.1 Experimental Setup

The domain that we have selected to test our conversational agent is that of *"animal*". The question-answer pairs are downloaded from two sources, namely, `AskKids.com` and `Answers.com`, solely for this experiment. 150 topics in the *"animal*" domain were manually identified and used as seed words, where 17,790 QA pairs were retrieved from `Answers.com` and 5,505 from `AskKids.com`. We were only able to download QA pairs about 96 out of the total 150 topics from `AskKids.com`. To ensure that the Strategiser has the option of using QA pairs from both sources during utterance generation, a topic is randomly selected by the system from amongst the 96 seed words to initiate every conversation with the participants.

For the experiment, we obtained the assistance of 11 human participants. For privacy reasons, the participants' names, genders and other personal information were not collected, and their names were never used by the conversational agent during conversations. The experiments were conducted over the Internet where the participants were provided with a text interface accessible via a Web browser. The interface is made up of two main parts, a user input field at the bottom, and a panel to display the exchanges between the user and the system. The instructions that we provide to the participants were as follows: (1) We requested the participants to limit their conversations to a 10-minute duration; and (2) At the end of every conversation, the participants were requested to rate the system utterances for naturalness and coherence using two 5-point Likert scales (the naturalness scale has scores that range from 0-`very artificial` to 4-`very natural`, while the coherence scale ranges from 0-`very incoherent` to 4-`very coherent`).



(a) Distribution of "naturalness"

(b) Distribution of "coherence"

Figure 2: Distribution of human judgements

5.2 Results

A total of 168 system utterances were recorded for all 11 conversations involving the 11 participants, with 16 − 17 system utterances on average per conversation.

Naturalness of system utterances: Figure 2(a) shows the distribution of the naturalness rating of system-generated outputs. Out of the 168 system utterances, only about 10% (17 utterances) are rated at or below level 2, with 13 system utterances being judged as `neither artificial nor natural` and 4 as `artificial`. The other 26% (43 utterances) and 64% (108 utterances) are judged as `very natural` and `natural`, respectively. In other words, the chart is left-skewed, with the majority (90%) of the judgements tending towards the `very natural` end of the scale, with both the mode and median for all 11 conversations being at level 3, which is `natural`.

Coherence of system utterances: Figure 2(b) summarises the human judgements of coherence of system-generated outputs. The trend appears to be quite similar to the judgements of naturalness in terms of skewness. However, instead of being concentrated at level 3, the judgements of coherence are slightly more spread over the neighbouring levels 2 and 4. The modes and medians shown in the legend in Figure 2(b) also demonstrate this spread. About 17% (29 utterances) of all system outputs were considered as `neither coherent nor incoherent` (20 utterances), `incoherent` (7 utterances) or `very incoherent` (2 utterances). At the same time, the numbers of utterances rated as `very coherent` (67 utterances) and as `coherent` (72 utterances) are approximately equal and when tallied up constitute 83% of all system-generated outputs.

Ability to maintain coherence: While the majority of the system-generated outputs are coherent (i.e., 83% of utterances are `coherent` or better), we are also interested in whether coherence of the conversations is sustained throughout, or whether coherence deteriorates with conversation length. Figure 3 shows coherence of individual system utterances over time for all 11 conversations. The figure shows that the coherence of the system contributions was consistently above level 2 (i.e., `neither incoherent nor coherent`), with 9 sporadic valleys reaching level 1 and 0.

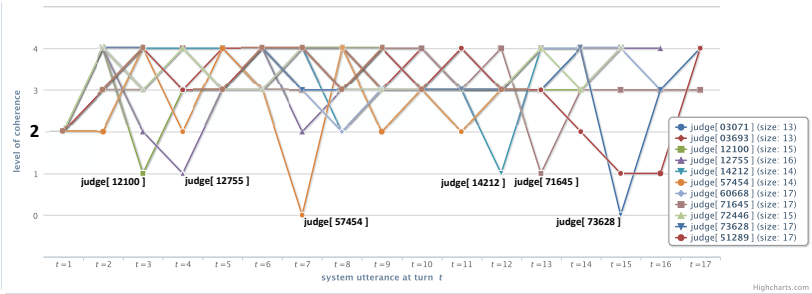


Figure 3: Level of coherence over the course of a conversation

causes	1	2	total
repetition of utterances	3	2	5
verbose utterances	0	4	4
perceived non-responsiveness to user questions	0	4	4
quality of QA pairs and wrongly judged utterances	1	3	4

Table 7: The causes behind the system utterances that were rated as `artificial` (score 1) or `neither artificial nor natural` (score 2) on the naturalness scale by the judges.

5.3 Error Analysis

After collecting judgements from the human participants, we analysed chat logs for naturalness and coherence ratings of the individual system utterances. We identified a number of causes behind the average and poorly-rated outputs: we discuss these here. Possible ways to remove these problems are discussed in Section 6.

Causes of artificial utterances: We identified four main causes behind the 17 system utterances that were judged as `artificial` or `neither artificial nor natural`, as summarised in Table 7. The first cause, i.e., repetition occurred when QA pairs which are different on the surface but semantically similar were used in succession to generate system utterances. For example, there were two distinct QA pairs with the questions *"What is cougar*" and *"What is the cougar?*" in our collection. During the system's conversation with judge 71645, these two pairs were used to generate the two different utterances *"Can you tell me, what is cougar?*" and *"Tell me, what is the cougar?*", four turns apart. From the judge's point of view, these utterances appeared as unnatural in that such blatant repetitions would not normally be encountered in human-human dialogue. The second cause, not surprisingly, is the artificiality of verbose utterances. The number of words in a typical system output ranges between 8 to 20. There are, however, a small number of utterances that go beyond this range, with some approaching or exceeding length 60. In this experiment, 8 utterances were considered to be lengthy, and of these, 4 were rated down by the participants (were rated as `neither artificial nor natural`) that encountered them during their interactions with the system. The third cause is the user perceiving the system as not responding to their questions. This problem occurred due to the system's more restrictive interpretation of what

causes	0	1	2	total
alternation between topics	0	4	2	6
utterances which are out-of-place	1	1	1	3
perceived non-responsiveness to user questions	1	0	3	4
issues with context management	0	2	3	5
initial system utterance	0	0	11	11

Table 8: The causes behind the system utterances that were rated as `very incoherent` (score 0), `incoherent` (score 1) or `neither incoherent nor coherent` (score 2) on the coherence scale by the judges.

constitutes a question. Out of the 4 utterances in the `quality of QA pairs and wrongly judged utterances` row of Table 7, 2 were the result of poorly-written QA pairs by the online human contributors. As an example, one of the sentences in the answer of a selected QA pair was incomplete, producing the system utterance *"And the interesting bit is that the giant panda lives in Asia and is most commonly known [sic]*", which was rated as 1 for naturalness by judge 12100. As for the other two utterances rated as `neither artificial nor natural` in the same row, the causes were unclear.

Causes of incoherent utterances: Next, we discuss the causes that contributed to some of the system utterances being judged poorly on the coherence scale. Table 8 summarises these causes. The first cause is the generation of utterances that give the users the impression of abrupt transitions. This may happen when the system moves to a new topic that is semantically less related (hence the abruptness), or alternates between topics due to the presence of multiple influential terms within the same domain in the context. The second cause is the generation of utterances that appear out of context. This may be due to the ways the QA pairs were written (poorly) or structured (i.e., the sentences in the answers may not lend well to the generation of individual, successive system utterances). Consider the following QA pair as an example:

```
Q: What do american cockroaches eat?
A: Just about everything! American cockroaches eat plants, garbage,
   people food, glue found in book bindings and other glues and
   pet foods are just a few.
```

This particular QA pair was used as part of the SSK strategy to produce the utterance *"Do you know that just about everything?*" when interacting with judge 03693. The utterance, which appeared in the following context, was in no way related to the previous exchanges and was ill-formed. For this reason, the judge rated its coherence as 1.

```
SYSM: erm... some types of cockroaches can fly but not all of them.
      the american cockroach can fly but the german one cannot.
      these are the most common cockroaches found.
USER: american cockroach?
SYSM: do you know that just about everything?
...
```

The third cause is the system's inability to detect and handle questions that were not formulated using typical *wh*-words. This cause also negatively affects the naturalness of system utterances,

as previously discussed. The fourth cause is due to problems with context management. As we have discussed before, the LTT strategy is used to recover from inadequate context for QA pair selection. This can happen when users repeatedly provide inputs that do not contain words that are useful for selecting QA pairs. Unfortunately, some of the judges do perceive this system-initiated solicitation for more useful inputs as unnatural. As for the 11 utterances judged as `neither incoherent nor coherent` on the fifth row of Table 8, these outputs were simply the first utterances generated by the system in their respective conversations. Since these utterances do not have any preceding inputs or outputs to be benchmarked against for coherence, these level 2 ratings were expected.

6 Discussion

Conversational systems that are required to engage users in conversations that cover the breadth and depth of certain topics face the bottleneck of custom- or hand-crafting the necessary content. Mining conversational content from the Web is increasingly being seen as a promising solution to this problem. However, the verbose, non-colloquial and monologue nature of typical Web text means that such content is not straightforwardly usable for responding to users by conversational systems for various reasons. In this work, we propose the use of question-answer (QA) pairs mined from community-driven question-and-answer websites as content for a conversational system. The main contribution of this paper lies in the conversational strategies we have defined that specify how and when the different parts of QA pairs can be used and augmented with a small number of generic fragments to generate natural system utterances. The coherence of system contributions in a conversation is managed using context. To assess the naturalness and coherence of system-generated utterances, we conducted a small pilot study involving 11 human participants. Out of the 168 system outputs, over 80% were judged as natural and coherent by the participants. The coherence of system contributions is generally maintained throughout the course of all 11 conversations, with sporadic incoherences. We also analysed the causes behind the $10-20$% of artificial and incoherent system outputs.

Limitations and potential improvements on the naturalness of system utterances: The two main causes that contributed to a number of outputs being judged as artificial were the appearance of repetitive and lengthy utterances. The problem of repetition arises when QA pairs with different surface patterns that are semantically similar are selected within the same conversation to generate system utterances. One possible way to overcome this is to use a string similarity measure to compare the questions of potential QA pairs against the QA pairs that have already been used. Candidate pairs that are similar to the previously used ones on the surface can then be excluded from future use. As for coping with lengthy utterances, a range of approaches are available, from detecting sentence boundaries with some heuristics for recombining them, to more elaborate summarisation techniques.

Limitations and potential improvements on the coherence of system utterances: Incoherent utterances are caused mainly by the lack of management of the implicit transitions between topics in the same domain during a conversation. The problem of topic transitioning can be managed, for example, using local *wellformedness* constraints, e.g., semantic relatedness measures. The other causes of poor coherence and artificial outputs, i.e., the non-responsiveness to certain questions (that have surface patterns not recognised by the system) and the quality of QA pairs, are more difficult to address. Adding new patterns to support more types of questions or using punctuation to detect questions may seem to be straightforward solutions. However,

our intention to potentially incorporate speech input will render the latter option useless, while the former solution will increase false positives during question detection. In regard to the quality of QA pairs, manual intervention is necessary to proof-read and edit the pairs. Currently, no manual effort is required to make our system operational, except providing seed terms to populate the QA pair collection.

Acknowledgment

This work is partially supported by the Australian Research Council and Real Thing Entertainment Pty. Ltd. under Linkage grant number LP110100050.

References

Bickmore, T., Pfeifer, L., and Jack, B. (2009). Taking the time to care: Empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the 27th CHI*, Boston, USA.

Chen, Z., Liao, C., Chien, T., and Chan, T. (2011). Animal companions: Fostering children's effort-making by nurturing virtual pets. *British Journal of Educational Technology*, 42(1):166–180.

Church, K. and Gale, W. (1995). Inverse document frequency (idf): A measure of deviations from poisson. In *Proceedings of the ACL 3rd Workshop on Very Large Corpora*, pages 121–130.

Clarke, C., Cormack, G., and Lynam, T. (2001). Exploiting redundancy in question answering. In *Proceedings of the 24th International Conference on Research and Development in Information Retrieval*.

Huang, J., Zhou, M., and Yang, D. (2007). Extracting chatbot knowledge from online discussion forums. In *Proceedings of the 20th IJCAI*, Hyderabad.

Inoue, M., Matsuda, T., and Yokoyama, S. (2011). Web resource selection for dialogue system generating natural responses. In *Proceedings of the 14th HCI*, pages 571–575, Orlando, Florida, USA.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Marge, M., Miranda, J., Black, A., and Rudnicky, A. (2010). Towards improving the naturalness of social conversations with dialogue systems. In *Proceedings of the 11th SIGDIAL*, University of Tokyo, Japan.

Mitkov, R. (2001). Outstanding issues in anaphora resolution. In *Computational Linguistics and Intelligent Text Processing*, pages 110–125. Springer.

Radev, D., Qi, H., Zheng, Z., Zhang, Z., Fan, W., Blair-Goldensohn, S., and Prager, J. (2001). Mining the web for answers to natural language questions. In *Proceedings of the 10th CIKM*.

Schulman, D., Bickmore, T., and Sidner, C. (2011). An intelligent conversational agent for promoting long-term health behavior change using motivational interviewing. In *Proceedings of the AAAI Spring Symposium*, Stanford University.

Shibata, M., Nishiguchi, T., and Tomiura, Y. (2009). Dialog system for open-ended conversation using web documents. *Informatica*, 33(3):277–284.

Wallace, R. (2009). The anatomy of a.l.i.c.e. In Epstein, R., Roberts, G., and Beber, G., editors, *Parsing the Turing Test*. Springer.

Wong, W., Thangarajah, J., and Padgham, L. (2011). Health conversational system based on contextual matching of community-driven question-answer pairs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2577–2580, Glasgow, UK.

Wu, Y., Wang, G., Li, W., and Li, Z. (2008). Automatic chatbot knowledge acquisition from online forum via rough set and ensemble learning. In *Proceedings of the IFIP International Conference on Network and Parallel Computing*, Shanghai, China.

Yao, X., Tosch, E., Chen, G., Nouri, E., Artstein, R., Leuski, A., Sagae, K., and Traum, D. (2012). Creating conversational characters using question generation tools. *Dialogue and Discourse*, 3(2):125–146.