WEBTALK: TOWARDS AUTOMATICALLY BUILDING SPOKEN DIALOG SYSTEMS THROUGH MINING WEBSITES

Junlan Feng[1], Dilek Hakkani-Tür[2], Giuseppe Di Fabbrizio[1], Mazin Gilbert[1], Mark Beutnagel[1]

[1] AT&T Labs – Research
180 Park Ave. – Florham Park, NJ 07932 – USA
{junlan, pino, mazin, mcb}@research.att.com

[2] International Computer Science Institute
1947 Center Street. Suite 600 Berkeley, CA 94704 – USA
dilek@icsi.berkeley.edu

ABSTRACT

WebTalk is a system for analyzing unstructured information from company websites to support automatic creation of spoken dialog applications. The goal is to completely automate the process of building, maintaining and deploying dialog applications by leveraging the wealth of information on the World Wide Web. WebTalk employs technologies in web mining, document understanding, question/answering, and speech and language processing. In this paper, we review extensions to these technologies to make them suitable for creating a WebTalk application. We present an evaluation study of a WebTalk spoken dialog system that has been instantiated on a telecom company website. Experiments with 30 different scenarios indicate promising results and provide evidence that such systems can potentially revolutionize the paradigm for creating and scaling spoken dialog services.

1. INTRODUCTION

Spoken dialog systems provide individuals and companies with a cost effective means of communicating with customers. Although there are a number of successfully deployed dialog applications [1], there remain several barriers that hinder the rapid portability of such systems to new services. The most significant challenge is minimizing the human effort and the knowledge required in building and maintaining dialog applications. These applications are expensive to create and require extensive efforts on data collection and user interface design. As a result, only few of the large companies today take advantage in deploying spoken dialog systems for their customer care. On the other hand, the majority of companies worldwide invest a significant effort to develop and maintain their websites. By the end of 2004, the total number of live *.com* domains was at a record high of 20 million [2], whereas the number of deployed speech applications is in the order of few thousands and many are very limited in their functionalities.

WebTalk attempts to completely automate the process of creating spoken dialog applications by leveraging the wealth of information on company websites. The goal is to be able

to mine a website and instantly create an interactive dialog system that can answer questions and perform transactional requests.

Although, we are not aware of any literature towards building conversational systems automatically based on the content of websites, there is a vast literature in the World Wide Web community on extraction of information from websites [3] and automated question-answering based on a collection of documents [4][5]. WebTalk incorporates some of the techniques present in this literature and extends them to provide a conversational interface to the underlying web content [6].

In this paper, we describe the major technologies behind the WebTalk system including website understanding, automatic speech recognition, speech synthesis, question/answering, and dialog management. We present a usability study for a WebTalk application that has been automatically generated by mining a telecom website. This website provides general information and features about communication services, frequently asked question/answer pairs, promotions, and general marketing information. Although complete automation in creating spoken dialog applications remains an extremely difficult problem, this paper shows that the performance of such a system is rather reasonable and in some cases acceptable based on a user study of 30 scenarios. Based on these results, we expect that over the next few years, the technology will become sufficiently mature to be able to deploy such services with minimal or zero human intervention.

The organization of this paper is as follows. In the next section, we describe the technology components of the WebTalk system and address the research challenges. In Section 3, we describe user scenarios and present evaluation results. A summary of this paper is presented in Section 4.

2. ARCHTECTURE OF WEBTALK

The five major technology components in WebTalk include Website Understanding, Automatic Speech Recognition

[2]The work was done while the second author worked at AT&T.

(ASR), Question Answering (QA), Dialog Management (DM), and Text-to-Speech synthesis (TTS).

2.1. Website Understanding

Dialog technologies are not currently capable of leveraging the information in free-form documents such as websites. Website Understanding is a component of WebTalk to automatically convert website contents into more structured formats that enable the DM to cope with the user's spoken requests. It includes a Web Page Parser, Website Data Mining, and Website Structure Understanding.

Web Page Parser. Information conveyed on web pages is carried out not only by their stream of texts, but also by the semantic structure of these pages, which are implicitly encoded in web documents. A Web Page Parser segments the web page content into smaller semantic units and identifies their semantic categories. A semantic unit is defined as a coherent topic area according to its content or a coherent functional area according to its associated behavior. It is classified into 12 semantic categories including *Page-Title, Form, Table-Data, FAQ-Answer, Menu, Bulletined-List, Heading, Heading-List, Normal-Content, Heading-Content, Picture-Label,* and *Other*. Classification results using two machine learning algorithms, Adaboost and Support Vector Machines, have been reported in [7].

Website Data Mining. The second task for Website Understanding is to extract structured task knowledge, such as names and properties of products and services, corporate contact information, as well as acronyms defined on the website. Structured task knowledge would facilitate the QA and DM components to more precisely respond to user requests. We developed a boosting algorithm to extract products and services and implemented a set of rules for extracting other entities. Results will be reported in a future publication.

Website Structure Understanding. Web pages on a company website are often systematically organized into subdirectories and are linked to each other through meaningful hyperlinks. Most web pages have meaningful page titles. Website Structure Understanding takes advantage of web page titles and hyperlinks to create a summary for each website subdirectory.

The output of the Website Understanding component includes five types of data: web sentences, semantic text data units, transaction forms, structured task knowledge, as well as website directory summaries. These data are used by the various components of WebTalk as will be shown next.

2.2. ASR

One of the biggest challenges in creating a WebTalk application is being able to use website data to train a statistical language model. The web language is significantly different than conversational utterances that are typically observed in a spoken dialog system. For example, disfluencies such as filled pauses or first/third person pronouns which are very common in spoken language are rarely observed in the web data. Instead, there are frequent word sequences, related to the web, such as *"… click on the link …"*, etc. In order to take advantage of the website content, we translate the web sentences provided by the Web Page Parser into conversational style utterances using the following three steps: filtering, predicate/argument extraction, and stitching predicate and arguments to conversational templates to generate utterances.

The first step, filtering, removes the common task-independent sentences from the web text. The common task-independent sentences are obtained by taking the frequently occurring subset of sentences from multiple websites. In the second step, we semantically parse the web sentences, using the ASSERT tool [15] from the University of Colorado, and extract the predicate/argument pairs. The final step inserts the predicate and arguments to the corresponding slots in the conversational templates, which are sequences like:

I would like to <PRED> <ARG>.

The conversational templates are manually written, or learned from previously collected utterances of other spoken dialog applications. These templates are used to generate new utterances which are then merged with data collected from other applications to create an n-gram language model.

The acoustic model was trained using utterances collected from other deployed spoken dialog services. Both the acoustic and language models are used in speech recognition with the AT&T Watson speech recognizer [8].

2.3. Website-Based Question Answering

We incorporate a QA component into WebTalk, which takes a natural language question and dialog context as input and finds a number of responsive answers from the task data. The DM prepares appropriate dialog context and determines the way to negotiate with the user based on the returned answers from the QA component.

The QA process consists of five stages, namely, question parsing, question classification, query formulation, answer retrieval, and answer extraction. Question parsing labels the recognized speech with part-of-speech tags, general named-entity tags and company-specific product and service entities that are extracted by the Website Understanding component. The second module is a question classifier, which categorizes the question into one of the following five categories - *Generic Information Request, Problem Reporting, Factoid Questions, Transaction Request,* and *Information Search.* The third module called query formulation which transforms a natural language question and the dialog context into a set of query terms. The fourth module is an answer retrieval engine which takes a query as input and returns a list of answer candidates deemed to be relevant to the query in a ranked manner. The answer retrieving algorithm has been described in [9]. The fifth module, answer extraction, checks the ranked list of answer

candidates and outputs those that contain product or service entities that are mentioned in the question, match the question type and have confidence scores above a predefined threshold. If none of the answer candidates meet these conditions, the system returns the string "NIL". The evaluation of the QA component of WebTalk was presented in [9]. It showed comparable performance against a handcrafted company specific question answering system.

2.4. DM

Conversational interfaces to Question Answering systems allow the user to retrieve information in a natural unconstrained manner. However, natural language is, by its nature, ambiguous. Requests from a user could be vague, incomplete, referring to previous context, or have multiple answers. To cope with this complexity, an automated system should be able to ask clarification questions, resolve anaphora and ellipsis references, render multiple responses, properly summarize relevant information, and generate surface realizations in a conversational style. Existing literature shows promising results using text input/output [5], but real-time speech-to-speech systems are still in their infancy [10][11]. Our approach is an initial attempt to address the general problem of Question Answering interaction with natural spoken language input.

To properly capture the user's intention, we used a generic goal-oriented call classifier which is able to classify the intent of the user into one of a predefined set of call-types [12]. This classifier was specifically trained on application independent data in order to capture generic discourse illocutionary acts like vague questions, greetings, thanks and agreements that have little or no relevance to the actual service task. Other requests are classified as relevant questions and directed to the QA module for further processing. The QA returns a list of possible answers with associated confidence scores. The DM keeps track of the specific discourse context and provides clarification strategies when the call-types are ambiguous (e.g., with similar confidence score) or have associated low confidence scores. It then retrieves the best answer to the user based on the QA confidence scores and provides a navigation mechanism when the answers are summarized in multiple segments. This pragmatic dialog strategy does not keep in consideration contextual questions and anaphora resolution in the case of follow-up questions. We are currently investigating other promising approaches to the problems which will be presented in future publications.

2.5. TTS

Awkward or unintelligible responses can dramatically reduce the perceived quality of a service. Crafting responses with TTS in mind is a necessary part of the system design to insure that useful information about structure and content is fully applied. This is clearly essential in WebTalk as the system generates large chunks of text blocks that are highly unsuitable for TTS. Acronyms, abbreviations and other web-specific language generate highly undesirable synthesized speech. In this study, we applied several general-purpose techniques to improve the quality of the synthesized speech:

- Employ commas, periods and TTS tags for audible cues, replacing the visual structure of HTML tables, navigation bars, and other web-specific artifacts.
- Implement changes using application-specific dictionaries and rewrite scripts, because all manual editing will be lost when data is regenerated.

Besides the above procedures, we did a fast visual inspection and random listening of a small subset of the 1000 prompts to find unexpected issues that may have been overlooked by systematic searches.

3. EVALUATIONS

In this section, we elaborate on an evaluation of a WebTalk spoken dialog system when instantiated on a telecom company website. Although, there have been many proposals for how to evaluate spoken dialog systems such as monitoring the number of turns, or the duration of the dialog [13], dialog evaluation remains a challenging task. We conducted our evaluation using a similar approach to the evaluation of the W99 spoken dialog system [14]. We manually crafted 30 scenarios, of which 24 scenarios are in-domain requests and 6 scenarios are out-of-domain. Table 1 provides two scenario examples. When designing these scenarios, we tried to phrase them as broad as possible so that the evaluators can express the requests in their own words. Scenarios are randomly assigned to the evaluators.

Table 1: Scenario examples

In-Domain:	You would like to know what type of hardware or equipment you would need in order to access the phone service.
Out-Domain:	You are taking a trip to Florida this Thursday, and you want to check out the weather there.

Table 2: Survey questions for the evaluation

Q1: Did you get the information you requested successfully?
Q2: When the system was unable to give you the information you wanted, were its responses sensible?
Q3: In this conversation, did the system understand what you said?
Q4: In this conversation, did you understand what the system said?
Q5: In this conversation, was it easy to find the information you wanted?
Q6: In this conversation, how would you rate your overall impression and interaction with the system?

In our evaluation, we received 100 calls from 16 volunteered callers. Table 3 provides a summary of the results of our experiments.

We designed a web interface to present scenarios, call instructions and survey questions. After each call, we ask the user to fill in a survey of 6 questions related to the success of

the dialog. Table 2 provides the list of our survey questions. The first question Q1 is a Yes/No question. The other questions are expected to be rated on a scale from 1 to 5 with 1 being very difficult, or almost never and 5 being very easy or almost always.

Table 3: Evaluation Results

	In-domain	Out-of-domain	Total
# dialogs	79	21	100
Q1 (% of yes)	49%	0%	37%
Q2	2.9	2.3	2.8
Q3	2.7	1.5	2.4
Q4	3.9	3.8	3.9
Q5	2.4	1.2	2.1
Q6	2.7	1.7	2.5

Our results show that users were able to successfully obtain the information they requested in 49% of the dialogs for in-domain requests (see Q1). As a sanity check, this number was 0 for out-of-domain requests indicating that all users were unable to find information that the system never had in the first place. Q2 scored an average of 2.8 which indicates the system's ability to converse with users in a sensible manner when it failed to respond with the exact answer. Subjects were generally not satisfied that the system "understood" them, giving Q3 an average score of 2.4 but that number is 2.7 for in-domain scenarios. This may be attributed to lower recognition and understanding accuracy. Q4 receives the highest average score of 3.9 which indicates that users considered the quality of the language generation and synthesized speech to be good independent of the domain of the scenarios. Q5 is related to the ease-of-use of the system and receives the lowest average of 2.1 which is certainly related to the low performance of Q3 and may also be attributed to the fact that users engaged in multiple turns before they were able to retrieve the right answer. Finally, in terms of overall rating (Q6), this system scored an average of 2.7 for in-domain scenarios which incidentally compares favorably with the 3.2 that was obtained for the W99 spoken dialog system [14]. One should note, however, the W99 system was designed manually, and models were built from a corpus of collected data, versus WebTalk, which was constructed automatically with minimal human intervention.

4. SUMMARY

This paper describes WebTalk, a framework towards automatically building spoken dialog applications from given websites. The goal is to enable companies with websites to extend their customer service with a spoken dialog interface over the phone.

In this paper, we addressed the technical challenges in WebTalk, described the technology components including website data mining, website–oriented language modeling, website-based question answering, dialog management, text-to-speech synthesis, automatic speech recognition, and text normalization. We presented an evaluation of a WebTalk spoken dialog system instantiated on a telecom company website. Our usability study shows that the overall system scored quite favorably compared to a dialog application that we manually created and evaluated few years back. These results are encouraging and suggest that such systems can potentially revolutionize the paradigm for creating and scaling spoken dialog services.

5. REFERENCES

[1] A.L. Gorin, B.A. Parker, R.M. Sachs and J.G. Wilpon , *"How May I Help You"*, Proc. of IVTTA 1996

[2] http://www.verisign.com/

[3] G. Salton and M. McGill, *"Introduction to Modern Information Retrieval"*, McGraw-Hill, 1983.

[4] B. Katz, J. Lin, D. Loreto, W. Hildebrandt, M. Bilotti, S. Felshin, A. Fernandes, G. Marton, and F. Mora, *"Integrating Web-based and Corpus-based Techniques for Question Answering"*, Proceedings of the Twelfth Text REtrieval Conference (TREC 2003), November 2003, Gaithersburg, Maryland

[5] S. Small, T. Liu, N. Shimizu, T. Strzalkowski, *"HITIQA: An Interactive Question Answering System: A Preliminary Report"*, Proc. of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, 2003.

[6] J. Feng, S. Bangalore, M. Rahim, *"WebTalk: Mining websites for Automatically Building Dialog systems"*, Proc. of IEEE ASRU 2003.

[7] J. Feng, P. Haffner, and M. Gilbert, *"A Learning approach to Discovering Web Page Semantic Structures"*, ICDAR2005, Korea

[8] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tür, A. Ljolje, S. Parthasarathy, M. Rahim, G. Riccardi, and M. Saraclar, *"The AT&T Watson Speech Recognizer*". Proc. of ICASSP, 2005, Philadelphia, PA.

[9] J. Feng, B. Srinivas, and M. Rahim, *"An Evaluation Study of WebTalk Question/Answering"*, International Conference on Speech and Language Processing, Korea, October 2004.

[10] Y. Kiyota, S. Kurohashi, T. Misu, K. Komatani, T. Kawahara, F. Kido, *"Dialog Navigator : A Spoken Dialog Q-A System based on Large Text Knowledge Base"*, Proceedings of 41st Annual Meeting of the ACL , July 2003, pp. 149-152

[11] C. Hori, T. Hori, H. Tsukada, H. Isozaki, Y. Sasaki, E. Maeda, *"Spoken Interactive ODQA System: SPIQA"*, Proceedings of 41st Annual Meeting of the ACL , July 2003, pp. 153-156.

[12] N. Gupta, G. Tur, D. Hakkani-Tür, S. Bangalore, G. Riccardi, M. Rahim, *"The AT&T Spoken Language Understanding System"*, IEEE Transactions on Speech and Audio Processing, January, 2006.

[13] M. Walker, D. Litman, C. Kamm, and A. Abella, *"PARADISE: A General Framework for Evaluating Spoken Dialogue Agents",* Proc. ACL/EACL, 271–280, 1997.

[14] M. Rahim, R. Pieraccini, W. Eckert, E. Levin, G. Di Fabbrizio, G. Riccardi, C. Lin, C. Kamm, *"W99 - A Spoken Dialogue System For The ASRU'99 Workshop",* IEEE Automatic Speech Recognition and Understanding Workshop, Keystone, Colorado, USA, December 12 -15, 1999.

[15] http://oak.colorado.edu/assert/