A Flexible Dialogue System for Enhancing Web Usability

Marta Gatius Meritxell González

Technical University of Catalonia, Software Department
Campus Nord UPC, Jordi Girona, 1-3,
08034 Barcelona, Spain
34-93-4137797

{gatius,mgonzalez}@lsi.upc.edu

ABSTRACT

In this paper, we study how the performance and usability of web dialogue systems could be enhanced by using an appropriate representation of the different types of knowledge involved in communication: general dialogue mechanisms, specific domain-restricted linguistic and conceptual knowledge and information on how well the communication process is doing. We describe the experiments carried out to analyze how to improve this knowledge representation in the web dialogue system we developed.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *natural language, evaluation, user centered design.*

General Terms

Design, Experimentation, Human Factors.

Keywords

Web dialogue systems, mixed-initiative dialogues, adaptive dialogue systems, evaluation.

1. A FLEXIBLE MIXED-INITIATIVE WEB DIALOGUE SYSTEM

The improvement of NL and speech technologies has made possible dialogue systems (DSs) capable of dealing not only with dialogues in which the system drives the interaction but also with mixed-initiative dialogues, in which both the user and the system can take the initiative. The DS become more complex when dealing with user-initiative dialogues because the system has to consider several aspects of the interaction in order to understand what the user really intends and then decide what to respond to user's intervention and how to express it. For this reason, practical mixed-initiative DSs are mostly adapted to the functionality of a specific application. However, the cost of developing application-restricted DSs is high and they are not easily adaptable to new applications and, thus, those systems are not appropriate to guide the user accessing different types of the web contents. There are complex conversational systems having reusable components of discourse management, NL and speech, (such as [4]), but they cannot be easily adapted to new types of applications by non-experts. In order to face the problem of adapting the language modules, the mixed-initiative web DS we developed (described in [2]) uses a syntactic-semantic taxonomy to generate (semi)-automatically the system's messages from the conceptual representation of the application. To improve the adaptability of the system to the user's expertise and application complexity we have distinguished two types of messages: directed and open. Directed system's messages are explicit about the information the system needs from the user at each state of the communication. Open system's messages suggest the user to introduce the information needed, but not as strongly. Although different types of messages could be used, the initiative (or control) of the dialogue is always mixed because the user can decide to select a new task at any state of the communication and the system will always guide the user to introduce the needed information.

In order to improve the adaptability of the DS several systems ([1],[3]) dynamically adapt the dialogue strategy. In our DS we have incorporated an independent module that uses data on how well the communication is doing to determine the type of message and the confirmation policy. Following the methodology proposed in [1] we analyzed a corpus of dialogues to obtain the data that gives information about the most appropiate system stategy and the amount of evidence each data gives. The dialogue cues used by the adaptive module to determine the system's respond are related to the system's errors as well as to the content of the user's intervention (asking for help, giving new relevant data and giving not expected data).

2. THE EVALUATION

Although our DS was designed to support speech and text we only used the text mode in our experiments. The grammars used by the text analyzer can be easily converted into the VoiceXML grammars used by the speech recognizer. Our main goal in evaluating the system was to compare its performance when using different dialogue strategies and when accessing different service types. We also wanted to measure the improvement of the system when the linguistic resources generated by the system were extended by experts using corpus of domain dialogues.

We carried out two separated experiments. For the first experiment the system's messages were generated automatically adapting the general linguistic structures to the web service tasks and their parameters. In the second experiment the adaptive module was incorporated. The corpus of dialogues obtained in the first experiment were used to develop the adaptive dialogue model as well as to improve the linguistic resources. In both experiments the subjects were asked to access two web services: a transactional service that states a date for large objects collection and an informational service giving information about the cultural events in the city (Barcelona). The DS included descriptions of these services. The order in which each subject accessed each of the two services was random. All tests were done in Spanish, which it is supposed to be the first language of all users. In the second experiment the system used 253 grammatical rules and 2241 lexical entries plus dynamic entries obtained (semi)-

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

automatically from web resources (16173 street names, 1005 locations, 92 cultural events titles, 541 participants in the events and 266 lexical entries from a furniture taxonomy).

2.1 Metrics and Methodology

For the experiments the system was accessible through a web site. The dialogues collected were analyzed according to the four performance features proposed in the PARADISE ([5]) evaluation scheme: task success (user gives the system all information needed to perform the web service task), dialogue quality (system's errors and appropriateness of system messages), dialogue efficiency (time, number of turns), system usability (user's satisfaction). In order to gather the satisfaction of the participants they were asked to complete a questionnaire related to their immediate impression about several features: overall impression, appropriateness of system's interventions, performance, friendliness, usefulness and future use.

The 30 volunteers recruited for the first experiment were randomly divided into two groups: one group accessed the version using open messages and the other group the version using directed messages. The 31 volunteers recruited for the second experiment were divided in two groups: one group accessed the open-messages plus the adaptive version and the other group the directed messages and the adaptive version.

2.2 Trial Results

The results of the questionnaire on user satisfaction are shown in Table 1 (rating scales are from 0 to 5, 0 strongly disagree, 5 strongly agree). Those results show the usability of the DS is high, most of the users have a good impression of the system (3,58) and would use the system again in the future (3,42). From those results we could infer our DS could improve web usability, once the technical problems to facilitate its adaptation to different contents would be completely solved.

Table 1. Questions related to individual interaction

Average punctuation for the 10 Questions	(0..5)
1. Overall impression is good	3,58
2. The system provided the desired information	3,59
3. You feel understood by the system	3,32
4. You knew what the system expected from you	3,74
5. You perceived the dialogue as pleasant	3,68
6. You found the system help was appropriate	3,83
7. You were able to understand the system	4,09
8. You found the system is useful	3,64
9. In the future, you would use the system again	3,42
10. You would prefer to access this system by phone	2,91

Table 2. Test 1 versus Test 2

Performance Features	Test 1	Test 2	P-value
Task success(%)	82%	83,3%	0,783
System Errors	2,37	1,19	0,003
Time (seconds)	124	102	0,222
# user turns	6,42	6,10	0,469
User Satisfaction (1-5)	3,28	3,655	0,004
Ambiguous messages	0,297	0,154	0,085

As can be seen in Table 2, the results of the experiment using the linguistic resources improved by experts (Test 2) resulted in significatively better performance along dialogue quality and system usability (comparing them using independent sample t-test). Considering dialogue strategies, dialogues with directed system's messages had significant lower system's errors than those with open messages, mainly because the more directed system's messages result in shorter and predictable user's interventions, easier to process by the analyzer. We obtained similar results when comparing directed-messages and adaptive version (using pair sample t-test), as shown in Table 3 adaptive dialogues had more errors. When comparing open messages and adaptive version (using also pair t-test), dialogues involving the adapting version had a significant lower number of ambiguous messages, as can be seen in Table 4. Although this measure could contribute to the higher system usability of the adaptive version it does not contribute to higher rate in the task success dimension. Finally, we also used pair t-test to compare the performance of the DS when accessing the two web services. The only significant difference is related to the number of ambiguous messages which is higher for the informational service than for the transactional one, where the data that has to be obtained from the user is easier to ask.

Table 3. Directed messages versus adaptive messages

Paired-t Test – Directed vs. A	Directed	Adapt.	P-value
Task success(%)	82%	80%	0,766
System Errors	0,749	1,273	0,038
Time (seconds)	73,5	103,8	0,037
# user turns	6,192	6,198	0,990
User Satisfaction Mean (1..5)	3,720	3,604	0,383
Ambiguous messages	0,1434	0,1320	0,872

Table 4. Open messages versus adaptive messages

Paired-t Test – Open vs. A	Open	Adapt.	P-value
Task success(%)	100%	91,67%	0,339
System Errors	1,827	1,284	0,381
Time (seconds)	81,0	94,7	0,602
# user turns	6,217	5,486	0,409
User Satisfaction Mean (1..5)	3,352	3,500	0,426
Ambiguous messages	0,2675	0,0275	0,021

3. REFERENCES

[1] J. Chu-Carrol and H. S. Nickerson. Evaluating Automatic Dialogue Strategy Adaptation for a Spoken Dialogue System. In Proceedings of NAACL'02 , 202-209.

[2] M. Gatius, M. González and E. Comelles. An Information State-Based Dialogue Manager for Making Voice Web Smarter. In Proceedings of WWW'07. ACM Press, 1315-1316.

[3] D. Litman and S. Pan. Designing and Evaluating an Adaptive Spoken Dialogue System. User Modeling and User-Adapted Interaction 12, 2002, 111-137.

[4] J. Polifroni, G. Chung and S. Seneff. Towars the Automatic Generation of Mixed-Initiave Dialogue Systems from Web Contents. In Proceedings of EUROSPEECH'03, 193-195.

[5] M. A. Walker, D. J. Litman, C. A. Kamm and A. Abella. PARADISE: A general framework for evaluating spoken dialogue agents. In Proceedings of the ACL'97, 271-280.