An evaluation of spoken and textual interaction in the RITEL interactive question answering system

Dave Toney[1], Sophie Rosset[1], Aurélien Max[1,2], Olivier Galibert[1], Eric Bilinski[1]

(1) LIMSI-CNRS, Orsay, France
(2) Université Paris-Sud 11, Orsay, France
{dave,rosset,amax,galibert,bilinski}@limsi.fr

Abstract

The RITEL project aims to integrate a spoken language dialogue system and an open-domain information retrieval system in order to enable human users to ask a general question and to refine their search for information interactively. This type of system is often referred to as an Interactive Question Answering (IQA) system. In this paper, we present an evaluation of how the performance of the RITEL system differs when users interact with it using spoken versus textual input and output. Our results indicate that while users do not perceive the two versions to perform significantly differently, many more questions are asked in a typical text-based dialogue.

1. Introduction

The RITEL project aims to integrate a spoken language dialogue system and an open-domain information retrieval system in order to enable human users to ask a general question and to refine their search for information interactively. This type of system is often referred to as an Interactive Question Answering (IQA) system. In the RITEL project, we have identified two important requirements: (i) the system's overall speed should be very fast; (ii) the user vocabulary should be preferably unlimited.

The RITEL system has been in development for over two years (Rosset et al., 2006). It was originally conceived as a speech-based IQA system. Recently, we have developed a text-based interface to the system. This allows users to interact with the RITEL system using a web browser. The main anticipated benefit of text-based interaction over speech-based interaction in a dialogue system is improved word recognition accuracy. Another potential benefit is that a user can visually review the dialogue history. On the other hand, speech is often perceived to be a more natural form of dialogue interaction. Issues like these raise some important questions. Does one mode perform more effectively than the other? Which mode of interaction do users prefer? Does the mode of interaction affect how users engage with the system?

In this paper, we present an evaluation of the latest version of the RITEL system. This evaluation has two main aims. Firstly, it is intended to provide a baseline system for comparison with future versions. We anticipate employing this evaluation framework on a periodic basis (approximately every six months). In each evaluation, we wish to identify how well the system performs, both in terms of: (i) objective measures, such as dialogue duration and recognition error rate; and (ii) subjective measures of users' satisfaction with the system performance. Secondly, we wish to assess whether the objective and subjective measures differ when comparing the original speech (telephone) interface with the new text (web) interface.

In section 2 we present related work on comparing speech and textual interaction in dialogue systems. Section 3 gives an overview of the RITEL system architecture. We describe our evaluation methodology in section 4 and present our evaluation results in section 5. We conclude in section 6 with a brief discussion of our findings and future work.

2. Related work

Relatively little work has been reported on the comparison of speech- and text-based interaction in spoken dialogue systems. A comparison of interaction modes has been undertaken in the development of a tutorial system (Litman et al., 2004; Litman et al., 2006). The main result of these studies was that there was little difference between the two modes of interaction with respect to the improvement of student learning. The reduced language understanding associated with a speech system did not reduce student learning gain. On the other hand, the facility for more natural, spontaneous input did not improve student learning. A similar result with speech-based tutoring is reported in Pon-Barry et al. (2004). In this evaluation of the RITEL system, we are interested in comparing the users' behaviour and assessment of the speech- and text-based interfaces. To our knowledge, this is the first evaluation of speech versus textual interaction in an IQA system.

3. System architecture

We provide a brief overview of the RITEL system architecture; a more detailed description is given in (van Schooten et al., 2007). The original system was based on speech only. The newly integrated web-based interface allows users to interact using text. The architecture (illustrated in figure 1) is highly distributed and based on servers and specialised modules that can exchange messages. The advantages of this type of architecture are twofold. First, computation-intensive processing can be performed on dedicated machines, assigning all available memory and CPU to a specific task, which will allow us to investigate smooth asynchronous operating in the future versions

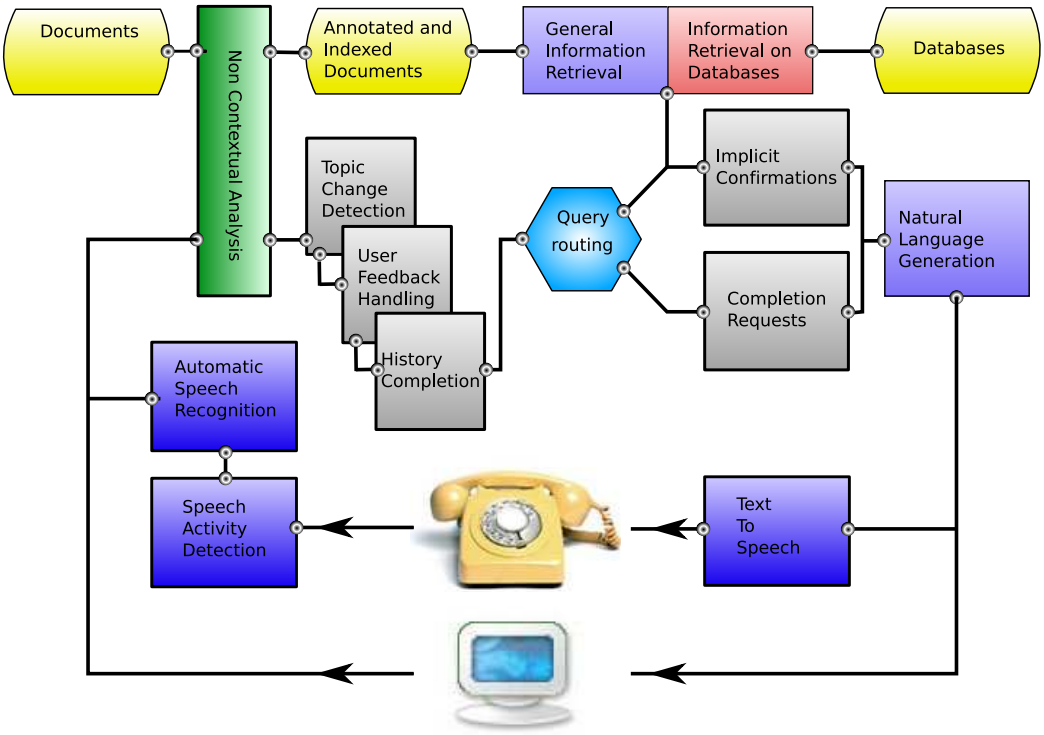

Figure 1: RITEL system architecture.

of the system. Second, the modular approach invites external contributions to be integrated into the system, thus offering a research framework for investigation in Human-Computer Interaction with competing strategies but also different application tasks.

A full account of the Speech Activity Detection and Automatic Speech Recognition components is given in (van Schooten et al., 2007). Naturally, these two components are not required by the text interface. Similarly, the Text-to-Speech module is not required for text output. Analysis of both indexed documents and user utterances are handled by the same module which is called *Non Contextual Analysis* (NCA) because no information from the dialogue or previous utterances is used. The general objective of NCA (see Rosset et al., 2006 for details) is to find the bits of information that can be of use for search and extraction, which we call *pertinent information chunks*. The Question Answering system is described in (Rosset et al., 2007). The indexing server's main role is to retrieve snippets, i.e. lines of documents corresponding to a given query. Queries take the form of a list of named entities and answer types. Candidate answers are ranked according to the scoring mechanism detailed in (Rosset et al., 2007).

The dialogue management function is decomposed into several steps. After an utterance has been annotated with NCA tags, it is processed in conjunction with the dialogue history. Consequently, an utterance may be considered a follow-up question, a change of topic, or some other user dialogue act (e.g. confirmation, rejection, goodbye etc.). At this point, information may be requested from the QA system. A system dialogue act is then generated and passed to the Natural Language Generation (NLG) component. The NLG module creates the full surface form of the system utterance and sends it either to the web interface or to the Text-to-Speech module, depending on the mode of interaction.

A crucial part of our dialogue management strategy is eliciting user feedback on the IR results that are being generated. Our system always invites the user to ask a question, while it communicates what it has understood through implicit confirmation. The system gives an answer whenever IR can be performed successfully. If the user does not react to the implicit confirmation, this strategy should still let the user judge whether the system has understood the user correctly, and repeat him/herself when appropriate. The user may choose to either give negative feedback and repeat (part of) the question (i.e. "No, I meant ..."), explicitly disconfirm anything mentioned by the system (i.e. "No, I didn't mean ..."), or implicitly or explicitly confirm anything mentioned by the system, and provide additional information as necessary (van Schooten et al., 2007).

4. Evaluation methodology

This evaluation employed a within-subjects design. Each participant was asked to conduct eight conversations with the RITEL system, four speech-based and four text-based. For the speech system, users dialled a free-phone number that connected to RITEL. For the text-based system, participants used a web-based interface. The order of the mode of interaction was counterbalanced to address the potential effect of order on the experimental results.

Statement	Metric
It was easy to get the information I wanted.	Task Ease
I found the system easy to understand.	Language Generation
I knew what I could say or do at each point in the dialogue.	User Expertise
The system worked the way I expected it to.	Expected Behaviour
The system's reaction time was appropriate.	System Response Time
Based on this experience, I would use this system regularly.	Future Use

Table 1: User satisfaction metrics.

Participants were randomly allocated to one of two order conditions (speech then text, text then speech). In each of the conversations, participants were given a topic to discuss with RITEL (e.g. "Superman", "Botswana"). Although RITEL is designed to be an open-domain system, the selection of topics was necessary to systematically compare speech- and text-based interaction.

For all speech-based conversations, the user utterances were recorded. The text of all dialogues (system and user utterances) was logged together with a variety of diagnostic information, including time-stamping information, and output from individual components: the speech recogniser, the natural language parser, the QA system and language generation system. This information was required to produce objective measures, such as dialogue duration, number of dialogue turns and Word Error Rate (WER).

After completing each conversation, participants were asked to complete a short questionnaire. Participants were asked to indicate their level of agreement with six statements using a Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). These statements were based on subjective measures of user satisfaction used in the Communicator evaluations (Walker et al., 2000). We included an additional measure of the system's response time. The measures were: Task Ease, Language Generation, User Expertise, Expected Behaviour, System Response Time and Future Use (Table 1).

5. Evaluation results

The participant sample comprised 11 men and 5 women. All participants were native French speakers. Of the 16 participants, 8 had no experience in the use of speech or language processing technology. It should be noted that there was little difference between participants in the two order conditions in terms of gender and previous experience with speech or language processing technology. Each of the 16 participants completed the 8 conversations. Consequently, the evaluation corpus comprised 128 dialogues (64 speech, 64 text), 2268 utterances and 25230 words.

With respect to objective measures, the mean duration of the speech-based dialogues was 138.45 seconds while the mean duration of the text-based dialogues was 215.25 seconds. A within-subjects t-test revealed that the difference between the two modes of interaction was highly significant ($p < .001$). However, measuring the duration for the text interface is not as meaningful as it is for speech, as an utterance is only considered finished when the submit button on the web interface is pressed by the user. The mean number of dialogue turns for the speech-based dialogues was 20.28 while the mean number of dialogue turns for the text-based dialogues was 15.15.

The mean duration per turn for speech dialogues was 6.82 seconds, while for text it was 14.20 seconds. The mean number of words per turn for speech was 10.32, while for text it was 12.19. The Word Error Rate (WER) for the speech-based dialogues ranged from 8.2% to 75.5% with a mean of 30.1%. For the text-based dialogues, we defined the Typing Error Rate (TER) to be the number of misspelled words. Unsurprisingly, the mean value for this measure was much less than WER – 5.5 %. These measures are summarised in Table 2.

Metric	Speech	Text
Mean duration	138.45	215.25
Mean #turns	20.28	15.15
Mean duration per turn	6.82	14.20
Mean words per turn	10.32	12.19
Word/Typing error rate	30.1	5.5

Table 2: Objective measure results.

With respect to the subjective measures, Table 3 summarises the mean user satisfaction scores for the speech- and text-based interaction modes. Mean scores tended to be low (below the mid point) for both modes, indicating a general lack of satisfaction with either mode. A t-test revealed no significant differences between the speech-based and the text-based mode on any of the six questions.

Metric	Speech	Text
Task Ease	1.81	2.04
Language Generation	3.65	3.75
User Expertise	3.17	3.31
Expected Behaviour	2.60	2.71
System Response Time	2.58	2.73
Future Use	1.73	1.85

Table 3: User satisfaction results.

With regard to user behaviour at the dialogue level, roughly similar proportions of self-contained questions, follow-up questions and yes/no questions were employed by users when comparing speech and text. However, there was a marked difference in how often questions as a whole were asked. In the spoken dialogues, 71% of utterances were questions; the remaining 29% of utterances were other dialogue acts, such as confirmations, rejections and repetitions. In the text-based dialogues, the ratio was 90:10.

In terms of the QA system, one unexpected but important observation was made: user utterances containing very few or very general search terms proved to be highly problematic. For example, one user asked the question "What is Botswana". A sensible (human) response might have been "Botswana is a country in southern Africa". But with only the search term 'Botswana', many thousands of responses were indexed in our QA system. Consequently, RITEL took several minutes to generate a QA response, in stark contrast to the mean response time of 0.1 seconds. This occurrence contravenes one of the system's key requirements: that its response time should be very fast. Naturally, the user perceived this time lag as an indication that the system had crashed and terminated the dialogue. This result is important because it demonstrates that using a QA system in a dialogue context requires modification even within the QA system itself; it is not simply a dialogue management issue. This situation did not occur during our participation in previous QA evaluation tasks (e.g. Rosset et al., 2007).

6. Conclusion and Future Work

While the spoken versus textual dialogues exhibited marked differences in objective terms, there was no significant difference in the perception by the users. The duration of spoken dialogues was much shorter (by 35%), but contained many more recognition errors. As a result, users spent considerably more effort on correcting and rejecting system responses in the speech-based system. In simple terms, it may be that the text-based system is superior in terms of fewer recognition errors, but that spoken language interaction is considered more natural. Consequently, neither mode of interaction is strongly preferred over the other. Alternative subjective measures may be needed in order to pinpoint the effect of the interaction mode on users' perceptions of system performance.

Ongoing analysis will also allow us to examine the correlation between the user satisfaction scores and objective metrics, perhaps using the PARADISE framework (Walker et al., 1998). We will also inspect the system logs to assess which components of the system are performing suboptimally and why. These analyses will facilitate the improvement of RITEL's performance before our next evaluation. We expect to continue comparing the performance of the speech- and text-based versions of the system. We are interested in adapting the dialogue and language generation strategies according to the mode of interaction. For example, it may be more helpful for the user if spoken system responses are shorter than textual responses since users are unable to process the content of the former in the same way.

7. Acknowledgements

This research was financed by The Leverhulme Trust and the Cap Digital business cluster project Infom@gic.

8. References

D. Litman, C. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman. 2004. Spoken versus typed human and computer dialogue tutoring. In *Proceedings of the Seventh International Conference on Intelligent Tutoring Systems*, Maceió, Brazil.

D. Litman, C. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman. 2006. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, 16:145–170.

H. Pon-Barry, B. Clark, E. Owen Bratt, K. Schultz, and S.Peters. 2004. Evaluating the effectiveness of SCoT: a spoken conversational tutor. In *Proceedings of the ITS 2004 Workshop on Dialog-based Intelligent Tutoring Systems*, Maceió, Brazil.

S. Rosset, O. Galibert, G. Illouz, and A. Max. 2006. Integrating spoken dialog and question answering: the Ritel project. In *Proceedings of the Ninth International Conference on Spoken Language Processing (INTERSPEECH)*, Pittsburgh, USA.

S. Rosset, O. Galibert, G. Adda, and E. Bilinski. 2007. The Limsi QAst systems: comparison between human and automatic rules generation for question-answering on speech transcriptions. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, Kyoto, Japan, December.

B. van Schooten, S. Rosset, O. Galibert, A. Max, R. op ken Akker, and G. Illouz. 2007. Handling speech input in the ritel qa dialogue system. In *Proceedings of the Tenth International Conference on Spoken Language Processing (INTERSPEECH)*, Antwerp, Belgium.

M. Walker, D. Litman, C. Kamm, and A. Abella. 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech and Language*, 12:317–347.

M. Walker, L. Hirschman, and J. Aberdeen. 2000. Evaluation for DARPA COMMUNICATOR Spoken Dialogue Systems. In *2nd International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, May–June.