Multimedia Blog Creation System using Dialogue with Intelligent Robot

Akitoshi Okumura, Takahiro Ikeda, Toshihiro Nishizawa, Shin-ichi Ando, and Fumihiro Adachi
Common Platform Software Research Laboratries,
NEC Corporation
1753 Shimonumabe Nakahara-ku, Kawasaki-city, Kanagawa 211-8666 JAPAN
{a-okumura@bx,nishizawa@bk,s-ando@cw,f-adachi@aj}.jp.nec.com

Abstract

A multimedia blog creation system is described that uses Japanese dialogue with an intelligent robot. Although multimedia blogs are increasing in popularity, creating blogs is not easy for users who lack high-level information literacy skills. Even skilled users have to waste time creating and assigning text descriptions to their blogs and searching related multimedia such as images, music, and illustrations. To enable effortless and enjoyable creation of multimedia blogs, we developed the system on a prototype robot called PaPeRo. Video messages are recorded and converted into text descriptions by PaPeRo using continuous speech recognition. PaPeRo then searches for suitable multimedia contents on the internet and databases, and then, based on the search results, chooses appropriate sympathetic comments by using natural language text retrieval. The retrieved contents, PaPeRo's comments, and the video recording on the user's blog is automatically uploaded and edited. The system was evaluated by 10 users for creating travel blogs and proved to be helpful for both inexperienced and experienced users. The system enabled easy multimedia-rich blog creation and even provided users the pleasure of chatting with PaPeRo.

1 Introduction

Blogs have become popular and are used in a variety of settings not only for personal use, but are also used in the internal communications of organizations. A multimedia blog, which contains videos, music, and illustrations, is increasing in popularity because it enables users to express their thoughts creatively. However, users are unsatisfied with the current multimedia blog creation methods. Users have three requirements. First, they need easier methods to create blogs. Most multimedia blogs are created in one of two ways: 1) A user creates audio-visual contents by cameras and or some other recording devices, and then assigns a text description to the contents as indexes. 2) A user creates a text blog, and then searches for multimedia contents on the internet and databases to attach them to his blog. Both methods require high-level information literacy skills. Second, they would like to reduce their blog-creation time. Even skilled users have to waste time assigning text description and searching related multimedia contents. Third, they like to be encouraged by other peoples' comments on their blogs. Although some users utilize pet-type agents making automatic comments to their blogs, the agents do not always satisfy them because the comments do not consider users' moods. To meet the three requirements, we developed a multimedia blog creation system using Japanese dialogue with an intelligent robot. The system was developed on a prototype robot called PaPeRo (Fujita, 2002), which has the same CPU and memory as a mobile PC. In this paper, we describe the multimedia blog creation method and the evaluation results in a practical setting.

2 Multimedia Blog Creation

2.1 Outline of system processes

The system has four sequential processes: video message recording, continuous speech recognition, natural language text retrieval, and blog coordina-

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 9–12,
Prague, June 2007.

tion. The first process is activated when a user begins a conversation with PaPeRo. The process stores a video message recorded on PaPeRo's microphones and CCD cameras, and the second process converts the speech contents of the video message into a text description to extract important keywords. Then, the third process searches for suitable multimedia contents on pre-specified web sites and databases based on the text description. The first three processes can simplify multimedia blog creation and reduce creation costs. The last process detects a user's mood, such as delight, anger, sorrow, and pleasure, by extracting typical expressions from the text description, and then chooses appropriate sympathetic comments to encourage the user. Finally, the last process coordinates uploading the recorded video message, the text description, the extracted keywords, the searched contents, and the sympathetic comments on the user's blog.

2.2 Continuous Speech Recognition

The system converts the speech content of the video message into text descriptions and extracts important keywords based on their lexical information. The system should, therefore, be equipped with a large-vocabulary continuous speech recognition engine capable of dealing with spontaneous speech. This is because blog messages usually contain various kinds of words and expressions. As this kind of engine needs a large amount of memory and computational resources, it is generally difficult to install the engine on small intelligent robots because the robot requires its own computational resources for their own intelligent operations, such as image recognition and movement control. To solve this problem, we used a compact and scalable large-vocabulary continuous speech recognition framework, which has been shown to work on low-power devices, such as PDAs (Isotani et al.,2005). The framework achieves compact and high-speed processing by the following techniques:

- Efficient reduction of Gaussian components using MDL criterion (Shinoda, et al., 2002)
- High-speed likelihood calculation using tree-structured probability density functions (Watanabe, et al., 1995)
- Compaction of search algorithm using lexical prefix tree and shared usage of calculated language model scores (Isotani et al., 2005)

The framework we developed contained a Japanese lexicon of 50,000 words typically used in travel conversations based on a speech translation system (Isotani, et al., 2002). We were able to evaluate the developed system by making a travel blog using Japanese dialogue with PaPeRo.

2.3 Natural Language Text Retrieval

The system generates a query sentence from a text description converted using the above-mentioned framework. As few multimedia contents contain retrieval keywords, the system matches the query to text in web pages and documents containing multimedia contents. The system then chooses multimedia contents located adjacent to the highly-matched text as retrieval results. To achieve high precision for avoiding user interaction with the retrieved results, the system is enhanced using the Okapi BM25 model (Robertson, et al., 1995) by the following techniques (Ikeda, et al., 2005):

(1) Utilization of syntactic relationships

The system needs to distinguish illustrations based on the context. For example, an illustration of fish to be eaten in a restaurant should be different from that of fish to be seen in an aquarium. To achieve this, the system utilizes the syntactic relationships between a pair of words. The system produces a higher score for text containing the same syntactic relationship as that of a pair of words in a query sentence when calculating the matching score.

(2) Distinction of negation and affirmation

The system needs to distinguish negative and affirmative expressions because their meanings are clearly opposite. To achieve this, the system checks adjuncts attached to the expressions when matching a query sentence and text.

(3) Identification of synonyms

As different expressions have the same meaning, the system normalizes expressions by using a synonym dictionary containing 500 words before matching a query sentence and text.

2.4 Blog Coordination

The system detects users' mood to choose encouraging comments. Users' moods are sometimes detected by the expressions used and the manner in which the utterances are spoken. Although speaking manner can clearly detect emotions, such as laughing or crying, some emotions are not always indicated. Expressions that clearly identify a person's mood can be indicated (Nakamura, 1993). By studying moods that are easily detectable from expressions, including modality, we developed a database of 10 moods (delight, anger, sorrow, pleasure, desire, fear, shame, relief, surprise, and normal) individually linked with 100 kinds of specific expressions. The database is searched based on the above-mentioned natural language text retrieval, which considers syntactic relationships, negative and affirmative responses, and synonyms. The database is also linked to PaPeRo's response to convey the most appropriate sympathy for each mood. The response includes verbal comments, such as "I'm happy for you" and "It's really too bad", and gestures, such as dancing and crying depicted using GIF animation files. Responses are chosen based on the mood detected. Finally, the system coordinates uploading a recorded video message, the text description, the extracted important keywords, the searched multimedia contents, and PaPeRo's responses on the user's blog.

3 Example Use in Practical Setting

We developed a prototype system for creating a travel blog on PaPeRo, which can retrieve 2000 web pages containing 1500 illustrations and 550 songs. PaPeRo is activated by hearing the phrase, "can you help me make my blog please?", as listed in Table. 1, and creates a blog, as shown in Figure 1. Figure 1 shows a screen shot of a video message attached to the blog, a text description converted by the speech recognition and a button for playing the video message (A). Keywords, in this case Yosemite, Las Vegas, and Roulette, extracted from the text description are displayed (B). Three illustrations searched based on a query using the text description are displayed (C). A button for playing a searched song is available (D). PaPeRo's comments, such as "I hope that happens", are displayed (E). The user's mood is detected as desire from her saying "I ***would like to*** go there again." The comment is displayed together with the appropriate PaPeRo's response.

Table 1. Dialogue example (Excerpt)

A user : Can you help me make my blog please?
PaPeRo: Yes, please push the button on my head.
A user : I went to Yosemite for my winter vacation. I played the roulette for the first time in Las Vegas. I would like to go there again.
PaPeRo: Ok, now your blog is ready for viewing.

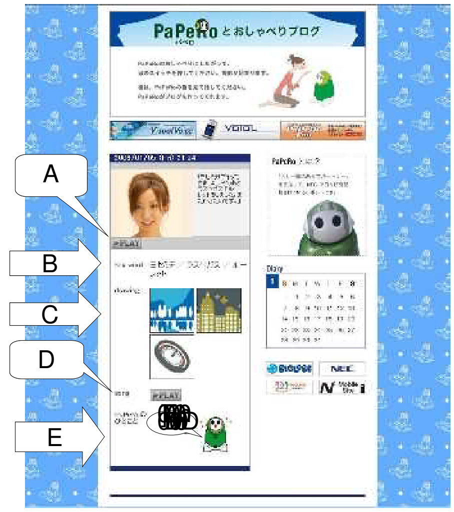

Figure 1. Example of Created Blog

4 Evaluation and Discussion

4.1 Evaluation

The system needs to be evaluated from two perspectives. The first is to individually evaluate the performance of each process mentioned in section 2. The second is to evaluate total performance, including the users' subjective opinions. As performance has been evaluated using different application systems, such as an automatic speech translation system (Isotani, et al., 2002) and a speech-activated text retrieval system (Ikeda, et al., 2005), we concentrated on evaluating the total performance based on surveying users' opinions about the blogs they created using the developed system. The survey results were analyzed in terms of speech recognition accuracy and users' blog making experience to improve the system.

4.2 Results and Discussion

The system was evaluated by 10 users. Half had blog making experiences, and the other half had no experience at all. All users input 20 sentences, and half of the sentences input were on travel issues, but the other half were unrelated because we needed opinions based on the results from low speech recognition accuracy. Users were interviewed on their automatically created blogs. Their opinions are listed in Table 2. The first row contains opinions about blogs created based on speech recognition results that had high word accuracy (85-95%). The second row contains opinions that had low accuracy (50-84%). The third row shows opinions regardless of the accuracy.

The left column contains opinions of users with blog-making experience. The middle column contains opinions of inexperienced users. The right column shows opinions regardless of the experience. The table leads to the following discussion:
(1) Expectations for multimedia blog creation

Users were satisfied with the system when high speech recognition accuracy was used regardless of their blog-making experience. Some users expected that the system could promote spread of multimedia contents with index keywords, even though few multimedia contents currently have indexes for retrieval.

(2) Robustness and tolerability for low accuracy

Users understood the results when low speech recognition accuracy was used because the multimedia content search is still fairly successful when keywords are accurately recognized, even though the total accuracy is not high. Users can appreciate the funny side of speech recognition errors and unexpected multimedia contents from PaPeRo's mistakes. However, as the errors do not always lead to amusing results, an edit interface should be equipped to improve keywords, illustrations and the total blog page layout.

(3) More expectations of dialogue with PaPeRo

Users would like to more enjoy themselves with PaPeRo, regardless of the speech recognition accuracy. They expect PaPeRo to give them more information, such as track-back and comments, based on dialogue history. As PaPeRo stores all the messages in himself, he has the ability to generate more sophisticated comments and track-back messages with users. Also, when the dialogue scenario is improved, he can ask the users some encouraging questions to make their blog more interesting and attractive while recording their video messages.

5 Conclusion

We developed a multimedia blog creation system using Japanese dialogue with an intelligent robot. The system was developed on PaPeRo for creating travel blogs and was evaluated by 10 users. Results showed that the system was effective for inexperienced and experienced users. The system enabled easy and simple creation of multimedia-rich blogs, while enabling users the pleasure of chatting with PaPeRo. We plan to improve the system by supporting the edit interface and enhancing the dialogue scenario so that users can enjoy themselves with more sophisticated and complex interaction with PaPeRo.

Table 2. Survey of Users' Opinions

		Blog-making experience		
		Experienced	Inexperienced	Either
Speech recognition accuracy	High	-This system makes multimedia contents more searchable on the internet.	-I would like to create blogs with PaPeRo.	-Easy to create blog only by chatting. -PaPeRo's comments are nice.
	Low	-Keywords, searched contents, and the total lay-out of blogs should be edited.	-Searched contents are good. -Even unexpectedly searched contents because of recognition errors are funny.	-PaPeRo could be allowed for his mistake. - Unexpected texts tempt users to play the video.
	Either	-PaPeRo's track-back is wanted as well as more dialogue variation.	-PaPeRo should talk on reasons of his choosing a song.	-PaPeRo should consider a history of recorded messages and his comments.

References

Yoshihiro Fujita. 2002. Personal Robot PaPeRo. *Journal of Robotics and Mechatronics,* 14(1): 60–63.

Takahiro Ikeda, at al. 2005. Speech-Activated Text Retrieval System for Cellular Phones with Web Browsing Capability. In *Proceedings of PACLIC19*, 265–272.

Ryosuke Isotani, et al. 2002. An Automatic Speech Translation System on PDAs for Travel Conversation. In *Proceedings of ICMI2002*, 211–216.

Ryosuke Isotani, et al. 2005. Basic Technologies for Spontaneous Speech Recognition and Its Applications. In *IPSJ-SIGNL*, 2005-NL-169, 209–116 (in Japanese).

Akira Nakamura (ed.). 1993. *Kanjo hyogen jiten (Emotional Expressions Dictionary).* Tokyodo Shuppan, Tokyo (in Japanese).

Stephen E. Robertson, et al. 1995. Okapi at TREC-3. In *Proceedings of TREC-3*, 109–126.

Koichi Shinoda and et al. 2002. Efficient Reduction of Gaussian Components Using MDL Criterion for HMM-based Speech Recognition, In *Proceedings of ICASSP-2002,* 869–872.

Takao Watanabe, et al. 1995. High Speed Speech Recognition Using Tree-Structured Probability Density Function. In *Proceedings of ICASSP-1995,* 556–559.