Web resource selection for dialogue system generating natural responses

Masashi Inoue[1,2] , Takuya Matsuda[1] and Shoichi Yokoyama[1]

[1] Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan
[2] Collaborative Research Unit, National Institute of Informatics, Tokyo, Japan
mi@yz.yamagata-u.ac.jp

Abstract. Using Web information in example-based dialogue systems is considered to be a good way to increase the topical relevance of system responses. However, Web content is mostly written documents, not colloquial utterances. To alleviate the discrepancy in style between written and spoken language, we suggest that the corpus should be selected subsets of the Web, viz., online bulletin boards. The naturalness provided by casual utterances is especially important when the goal of a dialogue is chatting rather than formal question answering. By appropriately selecting the information source, our text-based dialogue system can generate friendly responses to users. These characteristics were evaluated with a questionnaire.

Keywords: Naturalness, Dialogue system, Web as corpus.

1   Introduction

In designing dialogue systems, two aspects must be considered. The first aspect is management of the dialogue's flow, and there have been statistical and rule-based studies on this [1]. The second aspect is refining the content of the textual responses. Chatting as entertainment is a functionality that we can expect from an advanced dialogue system that generates natural responses to user inputs. However, casual dialogue data are hard to collect. In this paper, we discuss our investigations into the utility of Web resources for this purpose. The style of Web text is an issue when using it as a corpus for a dialogue system. For example, although Wikipedia has plenty of well-organized entries, they are often too formal for the purpose of chatting. We overcame this problem in text style by utilizing an online bulletin forum as a corpus.

2   Colloquial Texts

We used the 2channel Internet bulletin board[1] as the source of natural colloquial utterances. This is considered to one of the biggest forums on the Web, and it was

---

[1] http://2ch.net

used in [2]. It has over a million posts per day, mostly in Japanese. Another smaller, but still large Internet forum, is Slashdot. For example, 17,948 articles and about 3.8 million comments on it were analyzed in cross-media retrieval [3]. However, Slashdot mostly covers technical topics. 2channel contains many boards for users to discuss diverse topics ranging from "hacking" to "today's supper". The breadth of the topics in the target corpus may be disadvantageous when the system suggests quite minor topics that the majority of users have no interest in. To minimize such mismatches in interaction, we limited the topics to "trending" ones. Such topics are the most current ones on Internet forums, and it is likely that people will have something to say about them. There is a site that hosts trending topics[2]. Around 500 topics are posted everyday on this site. For each topic, 267 comments are posted on average[3]. The number is about ten times larger than the movie review site used in a previous study [4]. Of these 500 topics, we randomly selected five as candidates and presented them to users. Table 1 shows an example of the selected topic and comments associated with it.

Table 1. Example topic. Each topic has its original article and comments referring to it.

Topic
Cadburys says today's youngsters prefer lighter tastes - are they immature?

Article
According to a study by the candy maker Cadbury, the younger generation prefers lighter tastes. Cadburys says "The young people stick to simple seasonings … they are less interested in eating and having meals while they are playing games or using their mobile phones.

Comments
1 If you can appreciate lighter tastes, it means you have a sensitive tongue.   **2** Traditional seasoning is light.   …   **7** This is typical bad-mouthing of young people by the media.   **8** Bogus!   …

3   System Overview

The process of chatting is depicted in Table 2. Since the goal of this research was to investigate the utility of selective online resources, the base dialogue system was kept simple. It incorporated only those processes that were needed to generate minimally reasonable responses. Also, since we focused on the naturalness in single exchanges, we did not consider dialogue management. In step 1, topics with less than 100 comments were removed because the number of comments defines the pool of example utterances; fewer comments mean there is a smaller base of examples. About

---

[2] http://www.2nn.jp/
[3] accessed on 23rd Nov. 2009

51% of the topics contained more than 100 comments in our dataset. We found that the precision of keyword detection in step 5 was about 80%. In step 7, the scores of comments were calculated as the sum of frequencies of initial keywords and related terms in them.

Online bulletin boards were used to extract input-response patterns to generate utterances in the chat bot system [4]. In contrast, our system utilizes the original sentences to preserve its colloquialism and does not limit utterances to logical responses. We consider this aspect contributed to the naturalness both positively or negatively. In the future, we are going to analyze what constitutes naturalness.

Table 2. Illustrative flow of dialogue between a user and the system.

	Internal Process	Visible Output and Input / [Internal Values]
1	The system shows a list of topics and the user selects a topic.	**System:** "Which topic do you want to talk about? *1: Cadburys says today's youngsters …*" **User:** "1"
2	The system shows the details on the selected topic.	**System:** "The details on the topic are as follows: *According to a study by the candy maker Cadbury…*"
3	The user inputs a response freely.	**User:** "This is typical bad-mouthing of the young people by the media."
4	The system extracts nouns and noun phrases from user input through morphological analysis[4].	[*bad-mouthing, younger generation*]
5	It obtains keywords from the list by using the word scoring API[5].	[*younger generation*]
6	It obtains related words by searching the Web[6] and counting co-occurrences.	[*older, people*]
7	It ranks all comments based on the frequencies of the keyword and related words in them.	
8	It outputs highest ranked comment as its response.	**System:** "The older generation always wants to say something bad about younger people."
9	The user quits the interaction.	**User:** "Bye."

---

[4] http://en.sourceforge.jp/projects/chasen-legacy/
[5] http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html
[6] http://developer.yahoo.co.jp/webapi/search/

4 Evaluation

4.1 Experimental Setting

Although openings and closings in dialogues are considered to be important [5] to make dialogue systems more human-like, we focused on the naturalness of responses in the middle of the interactions under the assumption that such content was more influenced by the stored examples. That is, the goodness of interaction was measured by comparing the user input and system response texts (step 3 and 8 in Table 2). The results may change over time, since the system uses a Web API. Therefore, we did not let the participants use the system and instead showed them sample inputs and outputs in the form of a questionnaire. The sample inputs were randomly selected from example pools after eliminating meaningless sentences. Sample responses were generated by the system. This procedure also eliminates the influence of system speed and user interfaces.

The evaluation was conducted by asking volunteers to rate the utterances of our dialogue system. The participants were 30 undergraduate students majoring in computer science who were informed that the examples were taken from the output of dialogue systems. They were asked to rate the 10 sample responses made for each of the samples of input for the five topics in terms of naturalness and topical relevance. The scores ranged from one (not natural or not relevant) to five (very natural or very relevant). Naturalness did not necessarily correspond to the use of colloquial expressions because formal expressions might be construed as being more natural for some topics.
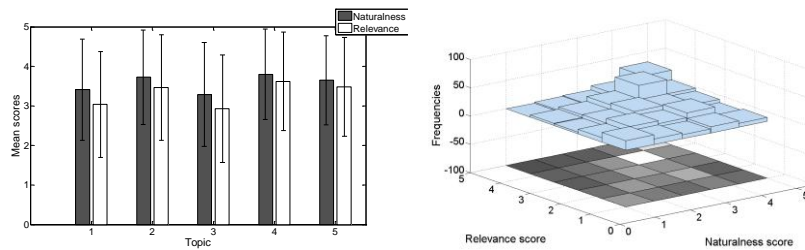

Fig. 1. Mean scores and standard deviations for five topics (left). Histograms for the first topics (right). The other four topics exhibit similar patterns.

4.2 Experimental Results

The average ratings awarded by the evaluators are plotted in the left graph of Figure 1. 300 scores were assigned to each topic. Although the scores for the same topic varied among evaluators and sample responses, as can be seen from the error bars, average naturalness scores exceed the borderline score 3.0 and the responses are categorized as natural. The right graph shows that there is a correlation between

relevance and naturalness since the diagonal elements are most frequent. Kendall's τ coefficients for the five topics are 0.39, 0.55, 0.47, 0.64, and 0.57.

5   Discussion

Our experiment showed a positive result in term of naturalness. However, it should be noted that the naturalness scores might have been higher because young students participated in the experiment. The source collection contained a great deal of online jargon and dialects seen only in online communication. Such expressions may have been regarded as being friendly. In addition to the evaluators' demography, if the entire dialogue rather than a single message exchange were to be evaluated, the score would be lower. Furthermore, we need to figure out what influence different sources have. We are currently looking at Twitter as an alternative source.

6   Conclusions

We analyzed the effect of corpus selection on generating friendly responses by a chatting system. Our experiment suggested that the use of original expressions in an online forum as the system's output is a promising way of achieving naturalness. Besides our reliance on the intrinsic naturalness in a corpus, there are sentence modification approaches. For example, one can add disfluencies to system utterances [6], make sentences shorter [6], or add modalities [7]. We need to investigate the utility of a web corpus when such approaches are used to increase naturalness.

References

1. Xu, Y and Seneff, S.: Dialogue management based on entities and constraints. In: 11th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pp. 87-90. (2010)
2. Ueno, M. Mori, N. and Matsumoto, K.: Novel Chatterbot System Utilizing Web Information, Advances in Soft Computing, Vol. 79, pp. 605-612 (2010)
3. Potthast, M.: Measuring the Descriptiveness of Web Comments. In: 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 724-725 (2009)
4. Huang, J., Zhou, M. and Yang D.: Extracting Chatbot Knowledge from Online Discussion Forums, In: 20th International Joint Conference on Artificial Intelligence, pp. 423-428 (2007)
5. Cheepen, C. and Monghan, J.: Designing for Naturalness in Automated Dialogues, chapter 11, pp. 127-142. Kluwer Academic Publishers, London (1999)
6. Marge, M., Miranda, M, Black, A.M. and Rudnicky, A.I.: Towards Improving the Naturalness of Social Conversations with Dialogue Systems. In: 11th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pp. 91-94 (2010)
7. Higuchi, S., Rzepka, R. and Araki, K.: A Casual Conversation System Using Modality and Word Associations Retrieved from the Web, Conference on Empirical Methods in Natural Language Processing, pp. 382-390 (2008)