# On the Difficulty of Improving Hand-crafted Rules in Chat-oriented Dialogue Systems

Ryuichiro Higashinaka* and Toyomi Meguro* and Hiroaki Sugiyama*
Toshiro Makino* and Yoshihiro Matsuo*
* NTT Corporation, Japan
E-mail: {higashinaka.ryuichiro,meguro.toyomi,sugiyama.hiroaki,makino.toshiro,matsuo.yoshihiro}@lab.ntt.co.jp
Tel: +81-46-859-2027 Fax: +81-46-855-1054

***Abstract*—In this work, we explore the difficulty of improving hand-crafted rules in chat-oriented dialogue systems. We first created an initial rule set in artificial intelligence markup language and revised it through an iterative cycle. Then, we tested the initial and revised rule sets by using human participants. The dialogue experiment showed that, despite the intensive revision process, the overall performance had little improvement. We investigated the errors made by the systems with each of the two rule sets, leading to the conclusion that it is the tracking of the context and the coverage of questions that are hindering the improvement of hand-crafted rules.**

## I. Introduction

After decades of research and development in task-oriented dialogue systems, chat-oriented dialogue systems (or chatbots) have been attracting attention in the social and entertainment spheres [1], [2]. Chat-oriented dialogue systems need to handle open-domain utterances from users, but current natural language processing (NLP) techniques are not mature enough to handle them correctly. Therefore, recent systems in chatbot contests or in the market typically depend on hand-crafted rules. Figure 1 shows some example rules in artificial intelligence markup language (AIML) [3], which is widely used in the chatbot community.

```
<category>
 <pattern>osake (alcohol) * nome (can drink) * ?</pattern>
 <template>nomemasu yo. osake wa sukidesuka (Yes, I can. Do you like alcohol?)</template>
</category>
<category>
 <pattern>tabako (cigarette) * sui (smoking) * ?</pattern>
 <template>tabako wa zenzen suimasen. donnna ajiga surundeshou (I don't smoke. How does it taste?)</template>
</category>
```

Fig. 1. Dialogue rules in AIML in our rule set. The italicized words denote Japanese and our translations are in parentheses.

Rule-based systems have various pros and cons. In terms of pros, the system can work reasonably well if we write many rules. The rules work because they can incorporate human insight and common sense to generate human-like utterances. The most significant con is the high cost for creating the rules: tens of thousands of rules need to be created for a workable system [3]. In addition, improving the rules is difficult because the rules are created mostly by human intuition and it is not known which types of rules are necessary for making improvements.

In this paper, we discuss the difficulty of improving hand-crafted rules. We first create an initial set of rules and revise them with an intensive human effort and then compare the performance of systems based on the initial and revised rule sets in order to identify the difficult points in improving the rules. By recognizing this difficulty, it will be possible to consider more strategically the ways to improve systems.

The following section describes our two rule sets: initial and revised. Section III describes the dialogue experiment using the two rule sets. In Section IV, we describe our analysis, where we perform dialogue breakdown annotation on the collected dialogues to identify which parts of the dialogue each of the rule sets cannot handle user utterances correctly, enabling us to analyze the difficulty of revising the rules. Section V summarizes the paper and mentions future work.

## II. Rule sets

We first created an initial rule set and then revised it into a revised rule set. To ascertain the quality of the rules, for the initial rule set, we made it a requirement that over 90% of utterances must be responded correctly in a *one-shot* interaction. In the revised rule set, to better handle the context, the criterion with a *two-shot* interaction was used. Below, we show how the initial rule set was created and revised.

### A. Initial rule set

The creation of our initial rule set is described in [4]. To overview it briefly here, one text analyst created 149,300 rules by encoding question-answer (QA) pairs for personas [5], utterance pairs in our chat dialogue corpus [4], and pairs of a stimulus word and an utterance into AIML rules. He also came up with topic-dependent rules; in AIML, once a topic is set by a triggering utterance using the `topic` tag, the rules under that topic are prioritized for pattern matching. Topics covered are movies, music, drama, animals, travel, and fortune-telling. He made some additional rules to output at least some response (such as back-channels) using wild-card patterns.

For evaluating the rules, 100 utterances from the chat dialogue corpus (a different portion from that used for creating the rules) were randomly sampled and used as input to the system loaded with the rules. Here, ProgramD, an AIML interpreter (http://aitools.org/Program_D), was used for implementing the system. Since Japanese does not have word boundaries, we used JTAG, NTT's morphological analyzer, to separate an utterance into word tokens. Note that the matching part of the rules (within `pattern` tags) is also composed of tokens, each of which corresponds to a morpheme.

An external judge subjectively evaluated the quality of the responses, and only when more than 90% of the responses were above average (over 6 points out of 10) was the rule-creation terminated. After six iterations of this evaluation process, the criterion was satisfied. Table I shows the statistics of the tags in the initial rule set. Refer to [3] for the meaning of the AIML tags. As far as we know, this is one of the largest AIML rule sets in Japanese except for our revised rule set described below.

### B. Revised rule set

To revise our rule set, we followed a careful procedure. We first performed a dialogue data collection using the initial rule set. For this, we recruited 60 participants to chat with our system. Each participant chatted four times, resulting in 240 dialogues (2,094 user turns). Each user utterance in the collected data was input to the system loaded with the initial rule set and the system's output was subjectively evaluated by two external judges. The rule update process was continued until both judges rated all system utterances as above average (over 6 points out of 10). Then, to further improve the rule set and to incorporate some contextual issues, we performed an online evaluation, where one external judge chatted for two turns with the system and evaluated the interactions subjectively. To avoid non-content-bearing interactions (e.g., greetings or backchannels), we instructed the judge to include at least one content word (noun, verb, or adjective) in all input utterances. The rule-revision process terminated only when the judge was satisfied (same criterion as above) 90% of the times within 100 interactions. We ran eight iterations of this procedure to finalize the revised rule set. The entire revision process took approximately three months.

### C. Statistics of the rule sets

As we have described, the initial rule set was meticulously revised. Table I shows that the rules were significantly augmented. One qualitative difference from the initial rule set was the removal of `topic` tags, since the persistence to the current topic was observed in the revision. Table II shows the statistics of the utterances that can be generated by each rule set. The revised rule set has more unique sentences and words. Here, the number of sentences for the initial rule set is larger because it does not use `srai` tags that can gather patterns with the same output utterance as frequently as the revised rule set.

## III. Experiment

Having created the two rule sets (initial and revised), we compared them by a dialogue experiment using human participants. We recruited 30 participants to use the systems loaded with the initial and revised rule sets. Each participant chatted (in text) with each system four times in a randomized order. The duration of a dialogue was restricted to four minutes.

TABLE I
Statistics of AIML tags for the initial and revised rule sets.

| AIML tag | Initial | Revised |
|---|---|---|
| aiml | 18 | 22 |
| category | 149,300 | 333,295 |
| template | 149,300 | 333,295 |
| pattern | 149,300 | 333,295 |
| topic | 6 | 0 |
| srai | 38,586 | 291,928 |
| li | 107,429 | 53,571 |
| that | 1,962 | 9,416 |
| random | 2,495 | 9,138 |
| star | 14 | 709 |
| sr | 0 | 12 |

TABLE II
Statistics of generatable utterances by each rule set.

| | Initial | Revised |
|---|---|---|
| No. of sentences | 215,649 | 85,800 |
| No. of unique sentences | 51,514 | 59,689 |
| No. of words | 3,107,976 | 1,231,801 |
| No. of unique words | 22,202 | 24,124 |

After each dialogue, each participant filled out a questionnaire asking for the subjective evaluation of the dialogue. The questionnaire was comprised of eight items that include the naturalness of the conversation, the diversity of system utterances, and the overall user satisfaction with the dialogue. The ratings were on a 7-point Likert scale.

Table III shows the subjective evaluation results. As shown in the table, there was little improvement made by the revised rule set. We tested the difference by Welch's t-testing and found that the ratings were not significantly different. Figures 2 and 3 show the dialogues by the system with the revised rule set with low and high user satisfaction ratings. The last column shows the matched patterns used for responding. It is surprising that the subjective evaluation did not improve at all. In the next section, we investigate the reason for this lackluster performance by comparing the errors made by the two rule sets.

## IV. Analysis

### A. Dialogue breakdown annotation

To identify the areas of dialogue that went wrong, we performed dialogue breakdown annotation on the system utterances collected in the experiment using three dialogue breakdown labels. This labeling scheme has been utilized in the error analysis of a conventional NLP-based chat-oriented dialogue systems [6]. Here, dialogue breakdown means a point in dialogue where users cannot proceed with the dialogue [7]. The three labels are as follows:

- **NB** Not a breakdown: It is easy to continue the conversation after the system utterance in question.
- **PB** Possible breakdown: It is difficult to continue the conversation smoothly after the system utterance in question.
- **B** Breakdown: It is difficult to continue the conversation at all after the system utterance in question.

Two annotators annotated each system utterance with one of the above breakdown labels. One of the annotators (Annotator-1) also provided written comments describing the error that

| Turn | User utterance | System utterance | Matched pattern |
|---|---|---|---|
| 1 | INIT | Hello | INIT |
| 2 | Hello! | Hello. | *konnichiwa* (hello) |
| 3 | How are you recently? | I'm doing great. How about you? | *saikin* (recently) *dou* (how) |
| 4 | I'm doing great. | I see. It's good that we live peacefully. | `that` |
| 5 | You're right. Peace is the most important. | Oh, is it? | *desu* (it is ...) |
| 6 | What is your recent pleasant experience? | I can't say. Do you have something that you can't say? | *nanika* (what) * *ka* (question marker) ? |
| 7 | I have a lot. | Oh, is it? | *rimasu* (have) *ne* (tag-question) |
| 8 | Do you have a lot of secrets? | I think I have many. | *takusan* (many) *motsu* (have) * ? |
| 9 | I see. You are mysterious. | Yes, I am. | *naruhodo* (I see) |
| 10 | It's fun to share secrets, isn't it. | Yes, indeed. | *idesu* (it is...) *ne* (tag-question) |
| 11 | I want to be someone with a lot of secrets. | Oh, please. | *taidesu* (want to be) |
| 12 | How can I be? | Well, I don't know how. | *dou* (how) * ? |
| 13 | It is difficult, isn't it. | Yes, indeed. | *idesu* (it is...) *ne* (tag-question) |
| 14 | Do you have hobbies? | Recently, I'm into dance. Can you dance? | *shumi* (hobby) * *motsu* (have) * ? |
| 15 | That's great. I'm not really good at dancing. | There are many songs with dancing performances. | `that` |
| 16 | Yes. When I go to see them, it's amazing. | You like it, don't you? How do you like it? | * *mini iccha* (go to see) * |
| 17 | When several people dance in coordination, it's beautiful. | It's beautiful, isn't it. | *kirei* (beautiful) |

Fig. 2. Example dialogue by the revised rule set. The user satisfaction rating for this dialogue was 7 (best). INIT indicates the initializing symbol for the system. The utterances were originally in Japanese and have been translated by the authors. The italicized words are Japanese words/expressions followed by English translations in parentheses. '`that`' indicates a triple match (previous system utterance, user utterance, system utterance).

| Turn | User utterance | System utterance | Matched pattern |
|---|---|---|---|
| 1 | INIT | Hello | INIT |
| 2 | Evening. | Yes, it is. | * |
| 3 | What is "it"? | Well, I really don't know. | *nani* (what) * *desu* (it is ...) ? |
| 4 | Tell me something interesting. | I'm becoming interested. | *omoshiroi* (interesting) |
| 5 | So, go on with your interesting story. | I'm becoming interested. | *omoshiroi* (interesting) |
| 6 | You are persistent. So, interesting story. | I'm sorry. Let's change the topic. | * *shitsuko* (persistent) * |
| 7 | It is the topic we are talking about. What is the topic? | Yes, it is. | *desu* (it is ...)? |
| 8 | Are you out of your mind? | Yes, I am. Are you surprised? | * *desu* (it is ...) ? |
| 9 | Yes, I'm surprised. | Oh, is it so? | *desu* (it is ...) |
| 10 | So, please tell me a story. | Okay. | *kudasai* (please do) |
| 11 | Please. | Thank you. | *douzo* (please) |
| 12 | I think you are really out of your mind. Please, go on with your story. | Thank you very much. | *douzo* (please) |
| 13 | I don't want to talk with you. | Me neither. | *naidesu* (I don't) |

Fig. 3. Example dialogue by the revised rule set. The user satisfaction rating for this dialogue was 1 (worst). See Figure 2 for the notations in the table.

TABLE III
SUBJECTIVE EVALUATION RESULTS: RATINGS AVERAGED OVER ALL DIALOGUES FOR EACH RULE SET.

| | Initial | Revised |
|---|---|---|
| Q1 Naturalness | 3.48 | 3.58 |
| Q2 Generation | 3.90 | 3.90 |
| Q3 Understanding | 3.48 | 3.42 |
| Q4 Informativeness | 2.76 | 2.72 |
| Q5 Diversity | 3.30 | 3.17 |
| Q6 Continuity | 3.39 | 3.25 |
| Q7 Willingness | 3.17 | 3.13 |
| Q8 Satisfaction | 3.01 | 3.05 |

TABLE IV
COUNTS OF DIALOGUE BREAKDOWN ANNOTATIONS FOR EACH RULE SET (INITIAL AND REVISED). THE RATIOS ARE SHOWN IN PARENTHESES.

| | Initial | Revised |
|---|---|---|
| Annotator-1 NB | 649 (0.45) | 755 (0.47) |
| PB | 347 (0.24) | 315 (0.20) |
| B | 458 (0.31) | 523 (0.33) |
| Annotator-2 NB | 926 (0.64) | 856 (0.54) |
| PB | 392 (0.27) | 451 (0.28) |
| B | 136 (0.09) | 286 (0.18) |

caused the dialogue breakdown (i.e., for PB and B labels). Table IV shows the distribution of the labels for the initial and revised rule sets. As can be seen, the difference in the ratios is marginal between the rule sets. In fact, for Annotator-2, there were more breakdowns in the revised rule set.

The inter-annotator agreement in Fleiss' $\kappa$ was 0.222 and 0.323 for the initial and revised rule set, respectively. When we merge PB and B and make it a two-class annotation, $\kappa$ values were 0.420 and 0.491, respectively. Since the agreement for the two-class annotation is moderate, we consider it meaningful to further investigate the utterances annotated as causing dialogue breakdowns.

### B. Clustering of the comments

To investigate the errors made by each of the rule sets, we used a text mining-based approach. Specifically, we mined the comments given to breakdowns by Annotator-1 using an automatic clustering method in order to obtain clusters of comments, each of which is likely to represent a particular error type by the rule set. Since we do not know the number of clusters in advance, we used a non-parametric Bayesian method called the Chinese restaurant process (CRP) for clustering. CRP can infer the number of clusters from data. We used word vector representation for comments. For hyper-parameters, we used 0.1 for $\alpha$ and $\beta$, and 10,000 iterations were performed for Gibbs sampling. See [8] for the details of CRP and these parameters.

Tables V and VI show the clusters obtained for the comments for the initial and revised rule set, respectively. For both rule sets, we obtained 18 clusters. Here, we only show the top-10 clusters ranked by cluster size. The tables show representative words as well as notable comments picked up by the authors. The representative words are the ones we found to be

TABLE V
CLUSTERS OBTAINED BY CRP FOR 803 COMMENTS GIVEN TO BREAKDOWNS (PB+B) IN THE DIALOGUES WITH THE INITIAL RULE SET.

| ID | Size | Interpretation | Representative words | Comments |
|---|---|---|---|---|
| 8 | 86 | Repetition | already, question, ignore, greeting, through | the system repeats the same utterance |
| 7 | 74 | Social error | feel, bad, give, change, impolite | the utterance is impolite |
| 2 | 72 | Not understandable | what, understand, say | the utterance is not understandable |
| 9 | 71 | Unable to answer question | answer, question, response | the system does not answer the question |
| 17 | 70 | Question detection error | question, place, strange, understand, feel | the system does not recognize that it is asked a question |
| 13 | 66 | Unable to answer about self | self, begin to say, strange, thing, start | the system cannot answer about itself or the topic it introduced |
| 4 | 62 | Contextual understanding error | see, say, have, happen | the system's utterance does not reflect the context |
| 16 | 55 | Repeated content | greeting | the system greets for the second time |
| 10 | 51 | Expression error | self, strange, uneasy, work | the usage of a particle is strange |
| 0 | 33 | Mismatched response | strange | the system's utterance is a little strange |

TABLE VI
CLUSTERS OBTAINED BY CRP FOR 837 COMMENTS GIVEN TO BREAKDOWNS (PB+B) IN THE DIALOGUES WITH THE REVISED RULE SET.

| ID | Size | Interpretation | Representative words | Comments |
|---|---|---|---|---|
| 10 | 132 | Repetition | thing, say | the system repeats the same utterance |
| 5 | 107 | Contextual understanding error | thing, strange, say, response | the response does not reflect what the user/system said |
| 3 | 92 | Unable to answer question | answer, question, introduce, say, beginning | the system does not answer the question |
| 16 | 85 | Unable to answer about self | strange, persistent, self, thing | the system cannot answer about itself |
| 9 | 77 | Not understandable | understand, what, happen, recognize | the utterance is not understandable |
| 1 | 51 | Social error | think, pollen, thing | the system doesn't care about the user |
| 15 | 49 | Mismatched response | mismatch, complement | the system misunderstands user and creates mismatched response |
| 4 | 41 | Repeated content | already, repeat, greeting | the system repeats the same content |
| 6 | 39 | Unclear intention | what, understand | it is difficult to understand what the system means |
| 8 | 35 | Question detection error | ask, thing, self | the system does not recognize that it is asked a question |

TABLE VII
CHANGE OF THE RANKING OF ERROR TYPES FROM THE INITIAL TO REVISED RULE SET.

| Interpretation | Initial | Revised | Change in rank |
|---|---|---|---|
| Repetition | 1 | 1 | → |
| Contextual understanding error | 7 | 2 | ↗ |
| Unable to answer question | 4 | 3 | ↗ |
| Unable to answer about self | 6 | 4 | ↗ |
| Not understandable | 3 | 5 | ↘ |
| Social error | 2 | 6 | ↘ |
| Mismatched response | 10 | 7 | ↗ |
| Repeated content | 8 | 8 | → |
| Unclear intention | – | 9 | ↗ |
| Question detection error | 5 | 10 | ↘ |

significantly dependent on the clusters by log-likelihood ratio testing. The interpretation column indicates our interpretations of the clusters on the basis of the representative words and the raw comments. We used the same interpretation labels for similar clusters across the rule sets; here, the similarity of the clusters was manually evaluated by comparing their representative words and comments.

Since we are interested in knowing what kind of error can be successfully revised and what kind of error persists, we compared the rankings of discovered errors types (interpretations). Here, we assume that the ranks of the error types are determined by cluster size.

Table VII shows how the ranking of the error types changed from the initial to the revised rule set. Note that only "Unclear intention" was not found within the top-10 error types for the initial rule set. It can be seen that the revision succeeded in reducing the errors related to "Not understandable", "Social error", and "Question detection error". It is reasonable that these errors lowered their rankings because they are errors concerning one-shot interaction. On the other hand, we could not remove the repetition of the same utterance. It remained the most salient error; the two-shot evaluation criterion could not suppress the repetition. This indicates that we need to consider longer context. The same can be said for "Contextual understanding error". The errors "Unable to answer question" and "Unable to answer about self" had higher rankings in the revised rule set. These errors mainly concern one-shot interaction but still cause dialogue breakdowns. This is presumably because of the wide variety of questions.

At the bottom line, it is the difficulty of writing rules depending on the context and of covering the possible questions that was hindering the improvement. Since it would be prohibitive to write all possible rules by hand because of the complexity, we need to use some means for tracking the context (e.g., information state [9]) and to adopt open-domain question answering technologies [10]; otherwise, the performance of the system will not improve, exactly as we have just experienced. Although some one-shot interactions can be covered by rules, they all simply cannot be covered by hand.

## V. SUMMARY AND FUTURE WORK

This paper discussed the difficulty of improving hand-crafted rules in chat-oriented dialogue systems. We compared two rule sets, an initial rule set and its revised version. By comparing the performance of the two rule sets, we found that the tracking of the context and the coverage of questions are the two main issues hindering the improvement of a rule-based system. Currently, there is emerging work to integrate rule-based and statistical-based systems [11], [12]. Acknowledging the usefulness of hand-crafted rules, we would like to pursue ways to combine NLP-based techniques with hand-crafted rules to improve the overall quality of chat-oriented dialogue systems.

## REFERENCES


[1] T. W. Bickmore and J. Cassell, "Relational agents: a model and implementation of building user trust," in *Proc. CHI*, 2001, pp. 396–403.

[2] R. E. Banchs and H. Li, "IRIS: a chat-oriented dialogue system based on the vector space model," in *Proc. the ACL 2012 System Demonstrations*, 2012, pp. 37–42.

[3] R. S. Wallace, *The Anatomy of A.L.I.C.E.* A.L.I.C.E. Artificial Intelligence Foundation, Inc., 2004.

[4] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing," in *Proc. COLING*, 2014, pp. 928–939.

[5] H. Sugiyama, T. Meguro, R. Higashinaka, and Y. Minami, "Large-scale collection and analysis of personal question-answer pairs for conversational agents," in *Proc. IVA*, 2014, pp. 420–433.

[6] R. Higashinaka, K. Funakoshi, M. Araki, H. Tsukahara, Y. Kobayashi, and M. Mizukami, "Towards taxonomy of errors in chat-oriented dialogue systems," in *Proc. SIGDIAL*, 2015, pp. 87–95.

[7] B. Martinovsky and D. Traum, "The error is the clue: Breakdown in human-machine interaction," in *Proc. ISCA Workshop on Error Handling in Spoken Dialogue Systems*, 2003, pp. 11–16.

[8] R. Higashinaka, N. Kawamae, K. Sadamitsu, Y. Minami, T. Meguro, K. Dohsaka, and H. Inagaki, "Unsupervised clustering of utterances using non-parametric Bayesian methods," in *Proc. INTERSPEECH*, 2011, pp. 2081–2084.

[9] S. Larsson and D. R. Traum, "Information state and dialogue management in the trindi dialogue move engine toolkit," *Natural language engineering*, vol. 6, no. 3&4, pp. 323–340, 2000.

[10] E. M. Voorhees, "The TREC question answering track," *Natural Language Engineering*, vol. 7, no. 4, pp. 361–378, 2001.

[11] S. Watanabe, J. R. Hershey, T. K. Marks, Y. Fujii, and Y. Koji, "Cost-level integration of statistical and rule-based dialog managers," in *Proc. INTERSPEECH*, 2014, pp. 323–327.

[12] T. Meguro, H. Sugiyama, R. Higashinaka, and Y. Minami, "Building a conversational system based on the fusion of rule-based and stochastic utterance generation," in *Proc. The 28th Annual Conference of the Japanese Society for Articial Intelligence*, 2014, (In Japanese).