Communication Papers of the Federated Conference on Computer Science and Information Systems pp. 37–42
DOI: 10.15439/2018F297
ISSN 2300-5963 ACSIS, Vol. 17

A Chatbot Based On AIML Rules Extracted From Twitter Dialogues

Hiroshi Yamaguchi, Maxim Mozgovoy
The University of Aizu
Tsuruga, Ikki-machi, Aizuwakamatsu, Fukushima,
965-8580 Japan
{m5201115, mozgovoy}@u-aizu.ac.jp

Anna Danielewicz-Betz
Ludwig-Maximilian University
EU Business School
Munich, Germany
anna.danielewicz_betz@euruni.edu

***Abstract*—A chat dialogue system, a chatbot, or a conversational agent is a computer program designed to hold a conversation using natural language. Many popular chat dialogue systems are based on handcrafted rules, written in Artificial Intelligence Markup Language (AIML). However, a manual design of rules requires significant efforts, as in practice most chatbots require hundreds if not thousands of rules. This paper presents the method of automated extraction of AIML rules from real Twitter conversation data. Our preliminary experimental results show the possibility of obtaining natural-language conversation between the user and a dialogue system without the necessity of handcrafting its knowledgebase.**

Introduction

A *CHAT dialogue system* or a *conversational agent* is a computer program designed to hold conversations in a human-like way and ideally "understand" the user's intent [1]. In general, chat dialogue systems can be categorised into two types: *task-oriented* systems that are used to assist the user in completing various tasks within a specified domain, and *open-domain* systems that aim at performing a natural conversation with the user [2]. Task-oriented chatbots are deployed in various business settings, such as social media marketing (personalised shopping, simplified buying procedures, customer help, statbots), typically to assist if not replace human customer service in live chats.

Conversational chatbots, on the other hand, may serve, e.g., as alternative interlocutors in healthcare or simply to keep lonely people company. Furthermore, open-domain conversational agents are good testbeds for the development and evaluation of "social" interfaces, deployable in a wide range of applications [3].

Many dialogue systems are rule-based, and one of the most popular mechanisms of representing rules is AIML (Artificial Intelligence Markup Language). AIML is a simple XML-based markup language that gained popularity after being used in the successful dialogue system A.L.I.C.E. [4] that won the Loebner Prize three times. The main drawback of AIML-based systems lies in the large number of rules required to imitate a natural conversation, especially in the case of open-domain systems. Therefore, an AIML-based dialogue system requires considerable manual effort to describe its knowledgebase, leading to an expensive and error-prone development process.

The goals of the present research are to:

- develop a method for AIML rules generation on the basis of existing conversational corpus;
- address approximate matching and context analysis;
- test the chosen approach using Japanese Twitter as a corpus of natural dialogues;
- evaluate the performance of the resulting chatbot.

Perhaps, the most attractive feature of rule-based systems is their simplicity. However, the need to design numerous rules can become a major obstacle in practice. We intend to show that it is possible to reduce the amount of human effort by automating the rule generation process, using a large dataset of authentic human conversations, such as Twitter dialogues. Certain steps of this process were demonstrated in our earlier system [5], but it notably lacked the ability to track the context of conversation and approximate matching capabilities.

Twitter as a Corpus of Conversations

The Internet serves as a vast corpus of conversational data. One can assume that Twitter dialogues come relatively close to informal daily conversations. There is, moreover, an API available to retrieve individual tweets and tweet streams [6]. This observation motivated us to use Twitter as a source of dialogues that can be converted into AIML rules.

For the previous version of the system [5], we retrieved a dataset of tweets posted between October 2016 and April 2017, using Streaming API [7]. Individual tweets were tagged with several attributes, including unique tweet ID, tweet language, timestamp, and in-reply-to field. For the present work, we used Rest API to extract replies to the tweets already present in our collection. Our goal was to obtain chains of three consecutive dialogue lines: the original tweet, a reply to the tweet, and a reply to the reply. As a result, we gathered a corpus of 49,971 dialogues of three lines or longer, and extracted 614,271 triples from the corpus.

CONTEXTUALISED AIML RULES

AIML is based on XML, and thus consists of hierarchically organised elements. Individual "units of knowledge" are known as *categories* in AIML. Each category should define at least two compulsory elements: a *pattern* that contains a sample input, and a *template* that contains the corresponding response of the chatbot [8].

In the following example, if the user inputs おはよう! ("Morning!"), the bot should reply おはようございます ("Good morning"):

```
<category>
 <pattern>おはよう!</pattern>
 <template>おはようございます。</template>
</category>
```

AIML syntax supports approximate matching with wildcard symbols and a mechanism of redirection, used to handle different situations with the same rules. AIML also allows specifying the context where the given rule is applicable, and thus keeping dialogues coherent. We rely on this capability when converting Twitter dialogues into AIML rules. The resulting system uses the rules including all three elements:

```
<category>
 <context>おはよう!</context>
 <pattern>おはようございます。</pattern>
 <template>今日はいい天気ですね。</template>
</category>
```

Here the bot will reply 今日はいい天気ですね。("It's a good weather today") only if the two preceding dialogue lines were おはよう! ("Morning!") and おはようございます ("Good morning").

The set of AIML rules forming a chatbot's knowledgebase is processed with an *AIML interpreter*, responsible for actual dialogues with the user. We use an in-house developed interpreter that implements approximate matching and tokenization of Japanese texts.

CONVERTING DIALOGUES INTO AIML RULES

The process of converting the raw tweet dataset into a set of AIML rules consists of the following steps.

Preprocessing: Raw tweet data contain numerous messages without any conversational meaning that have to be considered irrelevant for our purposes. These tweets typically consist of hyperlinks, hashtags and/or user names, or contain no Japanese characters. We remove such tweets from the source collection.

Normalisation: Each element in our collection contains three consecutive dialogue lines that are to be mapped to the AIML tags `<context>`, `<pattern>` and `<template>`. Our system attempts to identify the best matching context and pattern for the current situation using TF-IDF approach [9]. This process requires tokenization into individual morphemes, and stop-words removal. This is done with the help of Japanese morphological analyser MeCab [10] which splits the text into individual part-of-speech tagged morphemes. We use part-of-speech tags to eliminate non-significant morphemes, such as auxiliary verbs, postpositional particles, conjunctions, and pre-noun adjectivals.

Rule Generation. Each triple is transformed into an individual AIML rule. Triple elements are mapped to the AIML tags `<context>`, `<pattern>` and `<template>`.

INTERPRETING AIML

The semantics of individual AIML elements is well documented, which makes it possible to develop a universal AIML interpreter, able to serve as a chatbot powered with any given set of AIML rules. Indeed, some interpreters, such as Program AB [11] or pyAIML [12] are freely available.

Since we wanted to make use of TF-IDF based approximate matching, we had to implement our own AIML interpreter. The present AIML specification supports approximate matching, but this capability is based on wildcard characters rather than a text similarity analysis. On the other hand, we had to support the most basic AIML syntax that relies on `<context>`, `<pattern>`, and `<template>` tags only, so the resulting interpretation algorithm is relatively straightforward.

The current version of the system operates as follows. The chatbot starts a dialogue with a line こんにちは ("Hello"). The user's reply provides a context/pattern pair that is used to retrieve the next dialogue line of the chatbot (the highest scoring match according to TF-IDF is chosen). The last two replies become the new context and pattern, and the whole process is repeated.

EVALUATION CRITERIA

Different evaluation criteria can be applied, depending on the goals a given chatbot has been created to fulfil or tasks to perform. In other words, the evaluation criteria depend on the metrics applied at conceptual, operational, and qualitative levels. As for quality, here cohesion, cooperation, likeability, engagement, trust, reduction of frustration or ability to comment and provide feedback play a role. Cognitive linguistic quality criteria include, broadly speaking, conversational flow, understanding, and accuracy. Liu et al. [13] refer to task (completion)-focused responses and user satisfaction scores based on "model responses" and "appropriateness" of the proposed response to the conversation at hand, whereby a semantic match — co-occurrence in a given context — has to take place, especially for "informative words", as opposed to the "common" ones. They introduce a metric that correlates more strongly with human judgement, with the goal to

automatically evaluate how "appropriate" the proposed response is to the conversation, resorting to two approaches: word-based similarity metrics and word-embedding-based similarity metrics. Chakrabarti and Luger [14] refer to a goal-fulfilment map fostering evaluation that is to perform adequately not only in an isolated question-answer exchange, but also in a longer, sustained conversation, with the dialogue agent's enhanced ability to adhere to context in conversations, to hold a longer conversation, and more closely emulate a human-like conversation. This entails knowing what to say (content), knowing how to express it through a conversation (semantics), and having a standard benchmark to assess conversations (pragmatics-based evaluation, including input that is relevant to the context and within a given domain).

According to [15], adaptation to new information/request (that is, e.g., matching a given speech act while observing conversational rules) is an important quality factor, along with usability connected to effectiveness, efficiency, and satisfaction with contextually-bound goal fulfilment. On the other hand, a conversational agent should not aim at acting human. The evaluation categories encompass performance, manifested in such quality attributes as "robustness to unexpected input", "appropriateness and ability to perform damage control", "effective function allocation — provision of escalation channels"; functionality — manifested in "accurate speech synthesis", "accurate interpretation of commands", "appropriate degree of formality and accuracy of outputs", and "execution of requested tasks". Moreover, in terms of humanity criterion, the interaction should be "convincing, satisfying and natural", with the chatbot's ability to "respond to specific questions and to maintain themed discussion".

Overall, the quality attributes proposed include, among other things, robustness to unexpected input, provision of appropriate escalation channels, ability to maintain "themed discussion", i.e. within a given domain; being entertaining and engaging, ability to detect meaning and intent, and ability to respond to social cues.

Saygin and Cicekli [16] point out that the principles guiding human-computer conversation may be slightly different from those guiding inter-human communication. They propose that Grice's "cooperative principle" [17], consisting of conversational maxims, be taken into account when evaluating human-machine communication, albeit in a modified form. In particular, relevance maxim should not be violated by a dialogue system since, contrary to the human intention to change the subject, joke or use a metaphor, this is interpreted as inability to understand input utterances. By contrast, violations of manner have a positive effect on imitating human-like behaviour in that overreactive displays of emotions and impoliteness are normally associated with humans. So violations of relevance tend to create a machine-like effect and those of manner tend to create a human-like effect. Furthermore, violation of quantity maxim creates a machine-like effect since there is a strong correlation between this maxim and "artificial language use". As for quality maxim, no strong conclusions were reached since its violations tended to occur together with those of quantity, manner, and especially relevance. Since the deployment of a conversational agent cannot involve any cooperation *per se*, but rather imitation or simulation of thereof, the authors propose that the conversational principle be modified to accommodate human-computer interaction. In human-human conversations, the maxims are regularly flouted on purpose or violated unintentionally, yet this will not necessarily result in a communication breakdown, unlike in human-machine interaction.

Since the business goals of intelligent agents differ from purely conversational purposes (e.g. increased customer satisfaction, personalised solutions), so do the evaluation criteria as users expect intuitive, fast and "valuable" conversations. Such chatbot applications fall, however, outside the scope of the present paper.

EVALUATION AND FINDINGS

We conducted a pilot evaluation test of our chatbot, involving 10 respondents (5 female and 5 male), all speakers of Japanese (6 undergraduate students and 4 older adults aged 29-58). Each person made 3 attempts at chatting with the bot, resulting in total 30 chats (23 of which were 10 lines long on average; and 7 ranging from 17 to 41 lines). The evaluation questions on a 3-point Likert scale were adapted from [16] and answered by each respondent, following the three chat attempts. Due to the convenience sample of 10 respondents with rather limited exposure to the chatbot, as well as varying conversational skills and the negativity bias toward a machine interlocutor, the evaluation findings should be treated as preliminary. For the sake of reading convenience and comprehension, we only provide conversations translated into English in the subsequent sections.

A. Pragmatic Analysis

In pragmatics, the term "adjacency pairs", in connection to speech acts [18, 19], refers to those turns in conversations that have specific follow-ups, e.g. greeting-greeting, question-answer, invitation-acceptance or apology-acknowledgement. Opening sequences serve to initiate a conversation by means of greetings and small talk (general questions or comments about the weather, sports, etc.); whereas closing sequences signal an ending of a conversation (e.g. okay, all right then, well), followed by repetitions of farewells (okay, goodbye then; okay bye). Openings and closings are more conventionalised than are other parts of the conversation. The term "repair" refers to the clarification of previous intentions or the need of editing a preceding statement, i.e. "fixing" the utterance in some way. Politeness refers to conventionalised ways of conversing in an appropriate way that may involve titles and address forms, being indirect, that is generally avoiding any face threating acts (see Politeness Principle [20]).

Based on the pragmatic analysis of the 30 chats in our sample dataset, certain tendencies regarding "conversational behaviour" of both the chatbot (B) and that of the users (U) were observed. It can be noted that the chatbot tends to successfully complete adherent speech acts in opening sequences, such as greetings ("Hello" — "Hello"); small talk questions about general well-being (U: "How are you?" — B: "I'm well", reinforced at times by emoticons, or questions or remarks about the weather ("The weather is hot").

Generally, it tackles question-answer sequences in a satisfactory way by providing a generic answer (U: "Will you go shopping?"/"Where will you go?" — B: "I don't know"); by answering a yes/no question (U: "Are you hungry?" – B: "Yes, I am"; U: "Do you like music?" — B: "Yes, I like it [*note emoji*]") or by giving a more detail answer (U: "Where do you work?" — B: "I work in a factory [*thumbs up emoji*]"). Depending on the domain and question complexity, an attempt at a more elaborate answer may stretch over a number of lines, if not "interrupted" by the user's impatience and an abrupt change of subject (U: "Do you enjoy painting?" — B: (…) "I'm not good at sketching in five minutes [*emoticons*]. But I'll get experience [*emoticons*]. I want to improve my skills. I will try"). In most cases, however, the users, not aiming at exploration of a given topic, fail to ask further questions, (e.g., "What kind of music do you like? Do you like jazz?"), and change the topic abruptly.

In addition, the conversational agent simulates emotive reactions (with reference to senses) on a number of occasions, reinforced by punctuation marks, emoticons and/ or emoji ("Lean on me" — "I want to pat you on the head" — "Oh… you're fluffy" — all in one conversational sequence). Unfortunately, the users tend not to follow up on such "emotional vibes", resorting to dispreferred options (causing a mismatch in speech acts) in their responses, as illustrated by this example:

> B: I love you
> U: No, thank you.

Here the user exhibits outright violation of maxim of relevance, as well as that of politeness.

The chatbot makes rather successful attempts at simulating emotions — frequently reinforced by emoticons/ emoji ("Oh, I'm embarrassing", "Envy"), It also "expresses" concern, but, again, such attempts are not pursued further by the users, reluctant to explore the topic of emotions:

> B: My heart is frozen by your reply
> U: My muscles are aching!

Emotional content is thus commonly met with inappropriate answers, repetition or change of subject.

B. User-Induced Failure in Communication

The following example represents a dispreferred pair, hence the subsequent problems in matching an appropriate response by the bot:

> B: Hello
> U: What's your name?

A failure to reciprocate a greeting results in communication difficulties as the bot repeats its attempts at greeting (B: "Hello" — U: "What kind of people do you like?" — B: "Hello"). The same applies to conversations initiated by the user: Since the greeting speech act is not performed on numerous occasions, this causes a mismatch in adjacency pairs (a dispreferred second part).

Moreover, the user's mono-word utterances, such as "OK", "Who?", "What?", "Yes", "Really" — violating the conversational maxim of quantity [17] — pose a great challenge to the bot searching for appropriate follow-up candidate responses to match its response with, as such common words devoid of context cannot be assigned to any specific thematic domain or speech act adjacency pair (U: "Are you OK?" — B: "Thanks" — U: "What?" — B: "Kittyoppa-"). They tend to lead to a breakdown in communication, especially if used in a context requiring a more elaborate reaction (B: "You're a liar" — U: "What?"). By simply confirming with "Yes", for instance, and not exploiting the topic further, the user signals that she/he wishes to finish the interaction (B: "Let's hang out!" — U: "What are we doing?" — B: "Today?" — U: "Yes").

As mentioned before, the users also tend to change the subject abruptly (or ask questions that do not make much sense), thus violating the maxim of relevance. This poses immense difficulties for the bot. Simultaneously, the users appear to exhibit negative bias toward the computer system [16], demonstrated in questioning the chatbot's overall ability to communicate ("You cannot communicate"; "You really don't understand"; "What are you talking about?"). Consequently, instead of focusing on a conversation at hand, the users switch to the "evaluation mode", which hinders conversational flow and "confuses" the chatbot entirely. Generally, in the face of any slight difficulty, rather than rephrasing or clarifying their utterance (repairs are not common) or resorting to another conversational domain (since only limited domains of conversation such as greeting, weather, and food are explored) where a conversation might be more successful, the users accuse or reproach the bot ("I don't understand"/"What?"), thus violating the principles of cooperation and politeness themselves.

The challenging speech act of reprimand-apology is, however, tackled rather well (U: "You're a fool" — B: "It is regrettable" — U: "I'm sorry"), whereby the ensuing bot's utterance attempts to pick on the unjust reprimand ("It's so dangerous that you are strange"), but this rather sophisticated attempt at a further exchange on the topic at hand is not recognised as a follow-up sequence by the user.

In longer conversations (20 lines onward), there is no flow or cohesion that is normally demonstrated in human-human turn taking with smooth transitions from one topic to the other, but rather incohesive utterance chunks divide the whole "conversation" into unrelated parts. This is, again, at least partly due to the users' inability to lead and steer a

conversation skillfully to a variety of topics, by providing longer utterances, if not some contextual information; or by not recognising the attempts to continue a topic broached upon a couple of lines earlier. Such behaviour does not facilitate further conversation.

Overall, the users appear to violate the quantity, relevance, quality and manner maxims. They answer very briefly, even in single words and discuss a very limited range of topics — mostly food in a repetitive manner, then weather or music, and rarely work ("I want meat", "I'm thinking what to eat"). Moreover, as mentioned before, the users frequently change the subject in an unexpected way, and then accuse the chatbot of being unable to communicate ("We don't communicate well"):

> B: I hope it'll be a nice day today
> U: I want to eat grilled meat

The users also tend to use imprecise or awkward expressions, with an unspecified reference ("I said it to you"); give incorrect answers (B: "What is the difference between virginity and non-virginity?" — U: There's no difference"), and ask either irrelevant questions ("Do you hear me?") or too general ones ("What do you know?"), as well as echo questions ("Job?"). This all contributes to the overall lack in coherence and cohesion.

Additionally, the users prefer repetition to new attempts at evolving conversation:

> U: It's good.
> B: A little.
> U: You are wonderful.
> B: Exactly. That's being said…
> U: You are really wonderful.
> B: :) It's hard.
> U: I am thinking what to eat.

The above is a representative example of a conversation with the chatbot, illustrating that there is not much of a difference in the conversational styles of the chatbot and the user, whereby the user violates the cooperation principle on all fronts. Non-observance of manner and relevance is also demonstrated in scolding, ignoring attempts to show emotions, as well as nonsensical questions and responses ("Will you fire something?", "That's my line!").

Dissatisfied with the course of a given chat (yet not blaming themselves), the users also signal prematurely that they wish to abort it (e.g. by typing "Bye" in the middle of a conversation), hence triggering a communication breakdown due to inability to rephrase, choose a different topic, make their contribution longer or, in general, avoid nonsensical turns in conversation.

This all demonstrates that chatbot developers cannot expect that the users will adhere to conversational maxims when dealing with a computer system. Thus, training on real human conversations can pose limitations when dealing with adjacency pairs appearing in actual dialogues with chatbots.

CONCLUSION

In this paper, we outlined the process of creating a rule-based chatbot system with a set of rules derived automatically from Twitter conversations. Our experience shows that Twitter can serve as a source of relatively long (10 or more lines) casual conversations between people, rich in informal language constructions.

The resulting system has a simple architecture, somewhat compensated with a large number of AIML rules (over 600,000 in the current implementation). Our experiments show that the system is able to engage in conversations with people, and keep track of the dialogue context to some extent. However, it appears that TF-IDF is not adequate enough to serve as a reliable relevance measurement in this task. It is clear that the chatbot's responses are often irrelevant to the conversation at hand. It is also plausible that a single-line context, used by the bot, is not sufficient to keep track of the ongoing conversation. Furthermore, the selection of the most relevant dialogue lines according to TF-IDF measure produces predictable results. We are therefore planning to introduce random factors into the system.

Interestingly, since the users do not observe conversational rules on numerous occasions, those rules that definitely must not be violated should be specified both for the bot and for the user. It seems that the users tend to switch into the "evaluation mode" and "play" with the system to find out the chatbot's response to their particular remarks, rather than take turns in a genuine dialogue. As our system is trained entirely on real conversations, it typically fails to find an adequate answer when faced with such challenges. Hence, the improved version of our chatbot should be able to recognise irrelevant input and attempt to steer the conversation back on track. Moreover, the users should either undergo training to learn to converse with a chatbot more successfully or not be informed about the fact that their conversation partner is non-human to avoid bias.

REFERENCES

[1] B. AbuShawar and E. Atwell, "ALICE chatbot: trials and outputs," *Computación y Sistemas*, vol. 19, no. 4, pp. 625–632, 2015.

[2] R. Higashinaka *et al.,* "Towards an open-domain conversational system fully based on natural language processing," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 928–939.

[3] T. Bickmore and J. Cassell, "Relational agents: a model and implementation of building user trust," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2001, pp. 396–403.

[4] B. A. Shawar and E. Atwell, *A comparison between Alice and Elizabeth chatbot systems*: University of Leeds, School of Computing research report 2002. 19, 2002.

[5] H. Yamaguchi and M. Mozgovoy, “Generating AIML Rules from Twitter Conversations,” vol. Communication Papers of the 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 59–61, 2017.

[6] F. Bessho, T. Harada, and Y. Kuniyoshi, “Dialog system using real-time crowdsourcing and twitter large-scale corpus,” in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2012, pp. 227–231.

[7] Twitter Inc, *Twitter Streaming API.* Available: https://dev.twitter.com/streaming/overview.

[8] R. Wallace, “The elements of AIML style,” *Alice AI Foundation*, 2003.

[9] J. Ramos and others, “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, 2003, pp. 133–142.

[10] T. Kudo, *MeCab: Yet Another Part-of-Speech and Morphological Analyzer.* Available: https://sourceforge.net/projects/mecab.

[11] ALICE A.I. Foundation, *Program AB.* Available: https://code.google.com/archive/p/program-ab.

[12] C. Stratton, *PyAIML -- The Python AIML Interpreter.* Available: https://github.com/creatorrr/pyAIML.

[13] C-W. Liu *et al.,* “How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation” *arXiv preprint arXiv:1603.08023*, 2016.

[14] C. Chakrabarti and G. F. Luger, “A Framework for Simulating and Evaluating Artificial Chatter Bot Conversations,” in *FLAIRS Conference*, 2013.

[15] N. M. Radziwill and M. C. Benton, “Evaluating Quality of Chatbots and Intelligent Conversational Agents,” *arXiv preprint arXiv:1704.04579*, 2017.

[16] A. P. Saygin and I. Cicekli, “Pragmatics in human-computer conversations,” *Journal of Pragmatics*, vol. 34, no. 3, pp. 227–258, 2002.

[17] H. P. Grice, “Logic and conversation”, in *Syntax and Semantics*, Vol. 3, *Speech Acts*, P. Cole, & J. L. Morgan, Eds. New York: Academic Press, 1975, pp. 41–58.

[18] J. L. Austin, *How to Do Things with Words* Cambridge, MA: Harvard University Press, vol. 13, 1962.

[19] J. R. Searle, *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press, 1969.

[20] G. N. Leech, *Principles of pragmatics*. London: Longman, 1983.