Be More Eloquent, Professor ELIZA – Comparison of Utterance Generation Methods for Artificial Second Language Tutor

Taku Nakamura[1], Rafal Rzepka[2], Kenji Araki[2], and Kentaro Inui[1]

[1]Graduate School of Information Science, Tohoku University
[2]Graduate School of Information and Technology, Hokkaido University
[1]{tnakamura, inui}@ecei.tohoku.ac.jp
[2]{rzepka, araki}@ist.hokudai.ac.jp

Abstract

This paper presents utterance generation methods for artificial foreign language tutors and discusses some problems of more autonomous educational tools. To tackle problem of keeping learners interested, we propose a hybrid, half automatic (for semantics), half rule-based (for syntax) approach that utilizes topic expansion by retrieving the conversational subjects related to users' utterances. We compared the utterances generated by our methods with those of other dialogue systems. The evaluation results show that the topic expansion enriches vocabulary of the utterances. On the other hand, ELIZA-like confirmations and follow-ups were preferred by Japanese subjects when practicing conversational English was considered. Although our project is in its initial stage, we have decided to share our findings and thoughts on autonomy resulting from various trials, and thereby spark a discussion on the pros and cons of next generation of teaching applications.

1 Introduction

Applications supporting second language acquisition have evolved from simple flashcards for memorizing words to more sophisticated tools using gamification, voice analysis, etc. Computer applications and socializing online help to improve stickiness [Chen, 2014], which is often one of the biggest obstacles on the way of mastering a given topic. This problem is visible in software solutions not demanding any involvement from tutors and other peers (which is the majority of self-study mobile applications) but software-led teaching is preferable in scenarios where learners wish to improve skills without feeling ashamed. Our task, helping Japanese practice their communicational skills in English, is an example of such scenarios. Japanese students are not eager to use the language in everyday life for social and cultural reasons [Doyon, 2000], although they are often interested in foreign languages and possess wide knowledge about grammar and vocabulary. Artificial tutor is one possible solution and we decided to start a project aiming at creating a chat system that could be not only conversational partner, but also a second language acquisition supporter that learns user preferences (from language level to hobbies). However, there are difficult problems to be solved. First, the system must be linguistically correct. Second, the autonomy level of the software is important when using external corpora as its world knowledge. Presumably, controlled conversation by artificial templates must be balanced with the learning of a user's preferences and topic retrieval from the big textual data, which is more interesting but can be dangerous when left completely uncontrolled (as in case of Tay bot from Microsoft [Lee, 2016]).

The present paper introduces our prototype methods, which focus on providing the responses affected by learners' utterances based on rules and comparatively reliable knowledge resources. It does not necessarily extend the state-of-the-art techniques in the language generation domain per se, but we believe it is more efficient for this specific educational purpose. We compare various utterance generation methods, present experimental results and discuss other findings including user preferences for error corrections.

This paper concludes with ideas of measures that could be implemented to maintain balance between interesting and potentially dangerous Web-based tutors.

1.1 Traditional vs. Web-based Dialogue Systems

Well-known chatbots are ELIZA [Weizenbaum, 1966] and ALICEBOT[1]. ELIZA can respond to any input, but never provides new topics related to the user's utterances. ALICEBOT responds based on manually created databases that already exist. Although creating or extending databases will expand conversational topics, it is costly and nearly impossible to build a database that covers many fields and a broad range of users' interests.

Modalin [Higuchi *et al.*, 2008] is a Japanese text-based dialogue system that uses word associations retrieved from the Web and randomly adds modality to generated utterances. To sustain motivated conversation with users, the Modalin system generates input-related utterances using word associations. Presuming that a similar approach could enhance the conversation opportunities for English learners, we adopt the idea of word associations in our proposed system.

1.2 System for Language Learning

Jia [Jia, 2009] developed CSIEC (Computer Simulation in Educational Communication) system with multiple functions

[1] http://www.alicebot.org

for English learning, including a chatbot as a conversational partner. CSIEC system has a free conversation function based on textual knowledge and reasoning, aiming to overcome the problem in ELIZA-like systems, which require numerous predefined patterns fitted to the various utterances of users. The author suggested that databases for the system responses can be enriched by users' inputs, which need to be created beforehand. The CSIEC system still had insurmountable content shortcomings, and the project has been discontinued.

2 System Overview

2.1 CoAPM

Figure 1 outlines our first proposed method, the Co-occurring Action Phrases-based Method (CoAPM). The CoAPM method adopts the word associations utilized in Modalin [Higuchi *et al.*, 2008] on the hypothesis that input-related utterances could maintain users' interest in the conversation.

The present research applies this idea to English by replacing the Web with the British National Corpus (BNC)[2]. The BNC was chosen primarily because Web search engines restrict the number of searches, and because the BNC (being taken from trustful sources like newspapers or books) is expected to contain more correct English than other Web-based corpora. Therefore the English in the BNC was deemed suitable for educational purposes. Learners of English as a second language, who will mainly use common English, need not necessarily be familiar with native standard English, especially with natural expressions that rarely appear in textbooks. Nevertheless, resources with more input from non-native contributors might contain dialects proper to specific regions, which could baffle some leaners, whereas the BNC seems to maintain a more unified style with less potential for confusion. Thus, we assume that a standard English corpus such as the BNC is still useful for realizing a system as a widely acceptable English teacher.

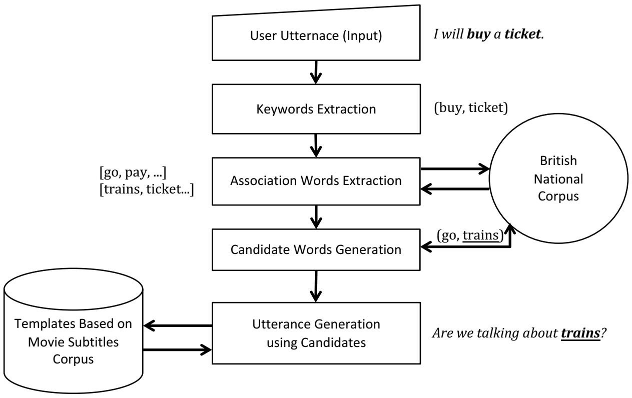
User Utternace (Input)
*I will **buy** a **ticket**.*
Keywords Extraction
(buy, ticket)
Association Words Extraction
[go, pay, ...]
[trains, ticket...]
British National Corpus
Candidate Words Generation
(go, trains)
Templates Based on Movie Subtitles Corpus
Utterance Generation using Candidates
*Are we talking about **trains**?*

Figure 1: Overview of the proposed Co-occurring Action Phrases-based Method (CoAPM) and utterance generation examples.

Extracting Keywords and Word Associations

In the first step, the method analyzes users' utterances using the Stanford Log-linear Part-Of-Speech Tagger (POS Tagger) [Toutanova and Manning, 2000; Toutanova *et al.*, 2003] to spot query keywords for extracting word associations lists. As the query keywords, we selected nouns and verbs (excluding some stop-words) because they constitute the core semantic elements of English sentence structures, and to some extent, describe the context of the utterances. This concentration also helps to reduce the exact co-occurrence matching costs when searching words of interest. Nouns identified as proper nouns by the POS Tagger are further analyzed using the Stanford Named Entity Recognizer (NER) [Finkel *et al.*, 2005] and are assigned to labels such as "PERSON", "LOCATION", "ORGANIZATION". In the next step, the method searches the BNC using these keywords (nouns or named entities, verbs) as queries and extracts sentences containing these words from the corpus. Nouns and verbs in the extracted sentences are listed and sorted in frequency order as word associations. This process is exemplified in Table 1.

Table 1: Keywords and association words extracted from the user utterance "I drink a glass of water in the morning."

keywords	'drink', 'glass', 'water', 'morning'
association verbs	'rising', 'braised', 'cooked', 'fried', 'chopped'
association nouns	'fruit', 'juice', 'glasses', 'piece', 'salad'

Generation of Words Candidates for Utterances

Using the sorted lists of extracted nouns and verbs related to the input keywords, the method generates a single verb and a single noun pair from the most frequent words in the lists. This verb-noun pair can be a candidate for utterance generation. To verify the existence of the verb-noun combination, the method then checks for co-occurrences of the given pair in the BNC. That is, the method first selects the top noun and top verb word associations, and then searches for the co-occurrence in each sentence in the BNC using exact-matching. Even if only one pair is found in the BNC, the verb-noun combination is regarded as possible in English. If the noun and verb are not found in the same sentence of the corpus, the method tests another verb-noun pair (the second most frequent verb and top noun in the list). The method repeats this process up to the three most frequent verbs and nouns, advancing to the next verb in stepwise fashion until a proper combination is found. We prioritize nouns because of the assumption that nouns describe the context of an utterance more specifically than verbs, which influence a topic shift more often. However, this assumption must be confirmed empirically in the future.

Utterance Generation

A CoAPM response is generated by applying the proposed verb-noun or one of the pair to a template. We prepared the templates for utterances half-manually, based on the most frequent sentences in English movie subtitles retrieved from

[2] The British National Corpus, version 3 (BNC XML Edition), 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. `http://www.natcorp.ox.ac.uk/`

OPUS corpus[3] [Tiedemann, 2012; Lison and Tiedemann, 2016]. The sentences were automatically abstracted using POS tagging and NER, then ranked by frequency. Movie subtitles were selected for their adequately large corpus size and their potential suitability for conversational templates. Examples of templates are shown in Figure 2. Using POS tag analysis, the method selects the templates that fit the proposed words or words in users' input. It then randomly selects a template and applies the previously chosen candidate words or input words. To confirm the correctness of the expression in an applied template, the method searches the core phrase of the given template (such as "visit* Tokyo" for "Would you like to visit Tokyo?", where * is a wildcard for matching various forms of a verb; in this case, visited, visits or visiting) in the corpus by exact matching. If more than five matches occur in the BNC, the method outputs that template inserted with the retrieved words or input words. The number of matches is set experimentally, accounting for the processing time and validity of the output. If no template satisfies the condition, CoAPM tries another combination of candidate words.

Figure 2: Examples of CoAPM templates

Speaking of (noun from user utterance), do you (retrieved verb)?
Would you like to visit LOCATION?
What do you think about (retrieved noun)s?
Everybody (retrieved verb), right?
Does (noun from user utterance) belong to ORGANIZATION?

2.2 CiAPM and RAPM

The BNC used in CoAPM contains formal and reliable English, which could be suitable for learners of English. However, the corpus covers few expressions of latest events or trends. In our next models, we relied on a more up-to-date ontology, ConceptNet[4], enabling response to ongoing topics. Based on the evaluation outcome and analysis of the first method evaluation, which we describes in Section 3.3, we developed two variations of our second method, named "Cited Action Phrases-based Method (CiAPM)" and "Related Action Phrases-based Method (RAPM)". CiAPM uses the cited phrases from user utterances without replacing the relevant text. RAPM retrieves the input-related concepts using the semantic network, ConceptNet, which contains natural language phrases. The method is outlined in Figure 3.

ConceptNet

ConceptNet is a large-scale semantic network providing general human knowledge [Speer and Havasi, 2012] expressed in natural language. It includes words, common phrases and the relations between them.

In the course of our study for better system utterances, we considered to employ sequence to sequence model introduced in [Cho *et al.*, 2014]. Inspired by [Vinyals and Le, 2015], we tried to apply this model to build a conversational system. However, it was difficult to find a training corpus with sufficient size and compatibility with our objective: conversational practice for language learning. Mainly because of this difficulty, we abandoned this attempt after a few trials.

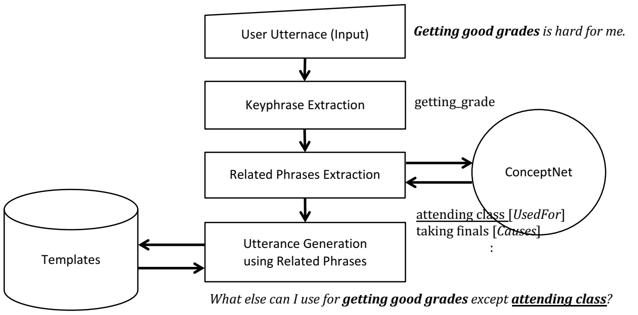
User Utternace (Input)
Getting good grades *is hard for me.*
Keyphrase Extraction
getting_grade
Related Phrases Extraction
ConceptNet
attending class [*UsedFor*]
taking finals [*Causes*]
:
Templates
Utterance Generation using Related Phrases
What else can I use for ***getting good grades*** *except* ***attending class****?*

Figure 3: Overview of the proposed Related Action Phrases-based Method (RAPM) and utterance generation examples.

At that time, the latest iteration of ConceptNet was announced, which can be regarded as reliable, up-to-date and one of the biggest freely available common sense knowledge resources. Commonsensical utterances are known to be a factor for enriching the naturalness of system responses: consequently, they enhance users' will to continue conversations [Rzepka *et al.*, 2005]. Therefore, we adopted ConceptNet, which includes knowledge from ConceptNet 5[5] and many different sources, in our methods.

Extracting the Key Phrase and Related Phrases

CoAPM identifies single words, so cannot handle idiomatic phrasal expressions. CiAPM and RAPM, which detect phrases including a gerund and a noun, can handle multiword expressions in a limited syntactic form, but they do not cover inflections in the phrase or other syntactic forms. For example, CiAPM and RAPM will detect "making a mistake", but ignore variations such as "made a mistake" or the phrasal verb "break down".

In the first step, the method parses the input utterances using the Stanford POS tagger to detect action phrases consisting of the -ing (gerund) form of a verb and a noun. Articles and adjectives between the verb and the noun are also captured. As key phrases, this form of action phrases is selected because they play various grammatical roles in English sentences without inflection, and to a certain degree, represent the semantic essence of utterances. In this stage, we partially detect the action phrases using the gerund without lemmatization, which facilitates the maintenance of grammatical validity. However, a fully developed system should respond to any utterances, requiring a more flexible method. If there are more than two action phrases in the input, the method selects the first phrase, based on an assumption that the first phrase has priority over other action phrases in the utterance context in English. The extracted phrase is transformed into a form of query phrase for ConceptNet API. Next, RAPM searches ConceptNet using this key phrase as a query. Finally, the method extracts the related action phrases from the results in natural language form. The phrase-extraction process is

[3] http://opus.lingfil.uu.se/OpenSubtitles.php

[4] http://conceptnet.io/

[5] http://conceptnet5.media.mit.edu

Table 2: Example of key phrase and related phrases extraction.

User utterance	"I was reading a newspaper, listening to music."
Key phrase	"reading_newspaper"
Related phrases and relations	(*HasSubevent*: "learning about current events") (*HasPrerequisite*: "getting a newspaper")

demonstrated in Table 2.

Utterance Generation (RAPM / CiAPM)

To generate responses from the proposed methods (RAPM, CiAPM), a related phrase or a cited phrase from an input is applied to a template. The related phrase and template are selected randomly. The templates were manually prepared based on the analysis of the first method (Section 3.3).

They were divided into two types: templates for any relation and templates for specific relations. Referring to the statistics of common relations [Ferschke *et al.*, 2013] in ConceptNet 5, we selected 11 relations in ConceptNet, namely, *IsA, PartOf, RelatedTo, HasProperty, UsedFor, DerivedForm, Cause, CapableOf, MotivatedbyGoal, HasSubevent, HasPrerequiste*. CiAPM applies phrases to the former type of templates, without using relations. In the template examples of Figure 4, 'V-ing N' denotes an action phrase which comprising a verb in gerund form and a noun.

Templates for any relation
Talking about [V-ing N (related phrase)]... What is your opinion on that topic?
Speaking of that, what do you think about [V-ing N (related phrase)]?

Templates for specific relations
relation: *RelatedTo*
Often [V'-ing N' (action phrase from input)] and [V-ing N (related phrase)] are a good combination.
What do you think?
relation: *HasProperty*
What about [V-ing N (related phrase)] while [V'-ing N' (action phrase from input)]?

Figure 4: Examples of templates for RAPM / CiAPM.

Error Correction

To improve the tutoring ability of our method, we aim at detecting the spelling or grammatical mistakes in users' utterances. We integrates LanguageTool, an open source writing style (including spelling) and grammar checker, calling it as a service via the HTTP API. Our method indicates errors in English usage by presenting a candidate correction with the error description message returned by LanguageTool. The correction candidate is taken from the top of the suggestions list generated by LanguageTool, in "Recast" form, which was preferred in the preliminary survey described in Section 3.1, and is displayed before the method utterance.

3 Experiments and Results

3.1 Survey on Error Correction Methods

Since we plan to equip our system with the function that detects the mistakes in users' utterances and convey these mistakes to the users in the dialogue, we conducted a questionnaire about how people prefer to be corrected. Five evaluators (four male students in their early 20s, one male in his early 30s), selected from among the potential users of an automated tutor, chose their preference as learners from three error correction methods, "Explicit-correction", "Recast", and "Prompt" (or "Elicitation") (see Table 3). These options were based previous studies of error correction in a second language classroom [Loewen, 2007; Tedick, 1986]. "Explicit-correction" refers to the direct indication and correction of mistakes. "Recast" is implicit reformulation of errors to the correct form. "Prompt" induces self-correction instead of providing the corrected form. Among many types of error correction, these three methods were selected for their efficiency and applicability to automatic dialogue generation methods.

This survey and the evaluation experiment of CoAPM in Section 3.2 were conducted online in a bundle. The survey presents participants with an erroneous utterance and its corrections by each method.

Majority of evaluators answered that "Explicit-correction" (40%) or "Recast" (40%) is preferable for learners, while the remaining 20% supported "Prompt" (Table 3). According to the result, "Explicit-correction" and "Recast" were considered to be more suitable than "Prompt" for error correction in utterances, although a broader survey is needed to reach a more definite conclusion.

The lower score for prompting might be related to the fact that we are not willing to keep people waiting and feel embarrassed when we are not sure what is the correct form. However, replacing a human teacher by a patient machine might significantly alter these results. This possibility requires evaluation in future study.

3.2 CoAPM Evaluation

To see how learners react to generated utterances, we compared CoAPM with ELIZA [Weizenbaum, 1966]. A possible benchmark, CSIEC [Jia, 2009], mentioned in Section 1.2, utilizes the conversational history. Because we evaluated only one-turn utterance exchanges this time, we instead used ELIZA as a baseline, which is independent of the preceding conversation and whose utterance rules are freely available. We employed python implementation of ELIZA by Jez Higgins[6].

As the user inputs, we used the utterances of English learners' in The NICT JLE (Japanese Learner English) Corpus[7]. This corpus comprises transcriptions of English oral proficiency interview tests for native Japanese speakers. The utterances include errors in English, some of which are tagged. Among the error-tagged data, we chose test takers' utterances

[6] `http://www.jezuk.co.uk/cgi-bin/view/software/eliza`

[7] `https://alaginrc.nict.go.jp/nict_jle/index_E.html`

Table 3: Examples of error correction methods for the user utterance: “I spend time listening music” and the survey results.

Methods	Examples	Respondents[9]
Explicit correction	“No, **listening to**”	40% (2)
Recast	“listening to”	40% (2)
Prompt	“listening...”	20% (1)

Table 4: Average scores in the three evaluation criteria. (Standard deviations are shown in parentheses.)

	CoAPM	ELIZA
Grammatical naturalness	3.50 (1.25)	3.74 (1.45)
Semantic naturalness	2.20 (1.43)	2.25 (1.49)
Motivation to keep studying	2.17 (1.37)	2.39 (1.46)

including at least a verb and a noun that appear more than five times in the BNC, and applied them as the input data (to ensure that the utterances convey a rich meaning, the 10 most frequent verbs in the BNC, expecting to include auxiliary and delexical verbs, were excluded from the condition). Under these restrictions, 19.6% of the examinees’ utterances were used as potential inputs. We used error-tagged utterances[8] for the convenience of evaluation when introducing the error suggestion function into our system. As mentioned above, the 10 most frequent verbs were excluded because they include verbs with low semantic meaning such as auxiliary and delexical verbs, although a more principled approach could be taken.

We asked the five evaluators (described in Section 3.1) to assess each of 20 utterance pairs (identical for all evaluators). Evaluators were asked to rate the input and response utterances generated by two methods in three categories: “grammatical naturalness”, “semantic naturalness” and “motivation to keep studying as a learner” on a 5-point scale (where 1 indicates unnatural language or lowest motivator of continued study, and 5 denotes natural language or highest motivator of continued study).

3.3 Results and Analysis (CoAPM)

Table 4 shows the average scores of all evaluators in each criteria for both systems, rated on a 1-5 scale. The inter-rater agreement of the five evaluators was 0.48 (Kendall’s coefficient of concordance).

On average, the preliminary version of our proposed method (CoAPM) scored slightly lower than ELIZA, although there were no statistically significant difference ($p > 0.05$) between CoAPM and ELIZA in all three evaluation criteria (Mann-Whitney U-test, $p = 0.42$ for grammar, $p = 0.29$ for semantics, $p = 0.21$ for motivation).

Figure 5 shows how CoAPM and ELIZA responded to several input utterances. The lower average scores of CoAPM than ELIZA (especially for grammatical and semantic naturalness) were mainly caused by insufficient utterance templates and incorrect POS analysis.

Figure 5: Examples of CoAPM and ELIZA outputs

Input	“In free time, I like to read books.”
CoAPM	“Does chapter read?”
ELIZA	“Very interesting.”
Input	“What did you watch?”
CoAPM	“Are we talking about watch?”
ELIZA	“Please consider whether you can answer your own question.”.

Among more than 100 types of templates, the POS restrictions admitted only six templates for 20 utterances of CoAPM.

In addition, we presumed that in second-language acquisition, the questioning or confirming style of ELIZA frequently surpassed the association-based strategy of CoAPM, although people preferred Modalin [Higuchi *et al.*, 2008] over ELIZA during normal chatting with no educational inclinations. This implies that follow-up questions are often more important than input-related statements in language tutoring tasks. Considering to evaluate each turn (each utterance pair) separately, we here set all templates as interrogatives. However, a deployed system should acknowledge as well as question a user’s utterance.

3.4 CiAPM and RAPM Evaluation

The following five systems were experimentally evaluated:

- Baselines
  (I) ELIZA
  (II) ALICEBOT
- Proposed methods
  (IV) CiAPM
  (V) RAPM-NOREL (not using relations)
  (VI) RAPM-REL (using relations)

We used the same implementation of ELIZA as described in a previous experiment (Section 3.2). In conversation, ALICEBOT needs the AIML (Artificial Intelligence Markup Language) set, which contains the contents of the ALICE brain written in AIML. Therefore, we adopted the standard free AIML set, “AIML-en-us-foundation-ALICE”[10]. By comparing with ELIZA and ALICE, we expected to observe whether chatting with simple dialogue systems is intrinsically efficient or not for educational purposes. Although it might be discussable, we believe that among the conversational systems, ELIZA and ALICE have been well-known and cited because other rule-based dialogue systems adopt similar processing of scarce context, or are unavailable for commercial or disclosed specification.

We created two versions of RAPM, which generate utterances from different templates. Specifically, RAPM-NOREL

[8] In the evaluations, we used the original utterances without error corrections as inputs, so the examples may contain erroneous expressions.

[9] The number of respondents is shown in brackets.

[10] `https://code.google.com/archive/p/aiml-en-us-foundation-alice/`

employs templates for any relations, while RAPM-REL utilizes templates for specific relations.

The user inputs were sentences in English learners' utterances extracted from The NICT JLE Corpus described in Section 3.2. Considering there were long utterances with many sentences, we used sentences here. From test takers' utterances, we selected sentences including at least one action phrase comprising a verb in gerund form and a noun. This condition is set on the assumption that action phrases have richer context in sentences, and facilitate the generation of grammatically correct utterances. Under this condition, 6.12% of the examinees' original sentences were retained as potential user inputs.

The evaluators were six male Japanese university students majoring in science (three undergraduates and three graduates in their 20s), who were potential targets of a full-fledged tutoring system. The subjects were intermediate English learners with basic knowledge of English grammar and vocabulary, but with low proficiency especially in speaking English.

The six evaluators assessed the utterances generated by all five systems, in response to each of 10 inputs chosen randomly from the utterances of test takers. The examinees' utterances were originally separated from the interviewer's utterances in the corpus. That is, each evaluator was given the same 50 utterances from the systems. The participants received pairs of utterances in a specific order. In contrast, in the former CoAPM evaluation (Section 3.2), the utterance pairs were presented in mixed order.

The system utterances were rated on a 5-point scale (where 1 means 'poor' and 5 represents 'excellent') in the following six categories.

(A) "Will to continue the conversation"

(B) "Semantical naturalness of dialogue"

(C) "Appropriateness in English conversation practice"

(D) "Vocabulary richness"

(E) "Knowledge richness"

(F) "Human-likeness of the system"

These evaluation criteria were based on the benchmark used in a related work [Higuchi *et al.*, 2008]. However, by focusing on the action phrases, the proposed methods are supposed to ensure a degree of grammatical naturalness in the utterances. Therefore, the original criterion "grammatical naturalness of dialogues" was changed to "appropriateness in English conversation practice", which is considered to be more important for evaluating English-teaching dialogue systems.

In the "vocabulary richness" evaluation, we expected subjects to rate utterances on a scale from "laconic" to "wordy". Some of these criteria could be evaluated by specialists familiar with English education, or at least by native English speakers. However, at this stage of our project, we focus on the user experience of learners who are easily bored with learning. Therefore we set the criteria in terms of the user experience, expecting evaluation from the learners' standpoint. In the questionnaire, the criteria (without specific descriptions) were presented to the evaluators in the Japanese language.

Table 5: Average scores in six evaluation criteria (A - F) and standard deviations (in parentheses). The highest scores for each criterion are highlighted in bold font.

	(A)	(B)	(C)	(D)	(E)	(F)
ELIZA	2.35	2.80	2.57	2.32	2.50	2.72
	(1.19)	(1.28)	(1.21)	(1.16)	(1.16)	(1.21)
ALICEBOT	2.78	2.67	2.90	2.65	2.80	2.88
	(1.21)	(1.29)	(1.31)	(1.14)	(1.09)	(1.32)
CiAPM	**3.13**	**3.15**	**3.37**	3.02^{*}	**3.20**	**3.17**
	(1.09)	(1.11)	(1.09)	(0.94)	(0.94)	(0.95)
RAPM	3.00	2.82	3.10	**3.23***	3.03	2.92
-NONREL	(1.12)	(1.09)	(1.14)	(0.91)	(0.95)	(1.08)
RAPM	2.97	2.62	3.03	3.18^{*}	2.88	2.90
-REL	(1.21)	(1.17)	(1.20)	(0.89)	(1.05)	(1.23)

* statistically significant difference between the model and ELIZA scores

Figure 6: Examples of outputs in CiAPM / RAPM evaluation

Input	"And a woman is playing piano."
ELIZA	"Let's change focus a bit... Tell me about your family."
ALICEBOT	"Aren't all woman playing piano ?"
CiAPM	"Let's talk about playing piano. What do you think about it?"
RAPM-NONREL	"Learning something about music... why some people love it and some don't?"
RAPM-REL	"I guess you were learning something about music - can you tell me how you did that?"
Input	"And they enjoyed eating delicious food and alcohol."
ELIZA	"How does that make you feel?"
ALICEBOT	"Who, specifically?"
CiAPM	"Let's talk about eating delicious food. What do you think about it?"
RAPM-NONREL	"Talking about diminishing your own hunger... What is your opinion on that topic?"
RAPM-REL	"I guess you were diminishing your own hunger - can you tell me how you did that?"

3.5 Results and Analysis (CiAPM / RAPM)

Table 5 shows the average scores and standard deviations of all evaluators in each criterion for the five systems (rated from 1 to 5). The Kendall's coefficient of concordance among the six raters was 0.369. One of the proposed methods, RAPM-NONREL with templates not using relations, scored highest in "vocabulary richness (criterion D)", and scored second-highest in other criteria. In all criteria except vocabulary richness, CiAPM achieved the highest score. The other proposed methods, RAPM-REL, also achieved a high average score in vocabulary richness. According to the Steel-Dwass test (evaluated by the asymptotic method), the "vocabulary richness (D)" score of our three methods significantly differed from the ELIZA score ($p < 0.05$), but no statistically significant differences were observed in the other criteria. Figure 6 shows some responses of each method to different input utterances.

The result suggests that the input-related phrases from ConceptNet are useful to expand the vocabulary of the system, and hopefully that of interacting users. For instance, the input “And a woman is playing piano.” elicited the responses “Learning something about music... why some people love it and some don’t?” (RAPM-NONREL) and “Let’s talk about playing piano. What do you think about it?” (CiAPM). The retrieved phrase ‘learning something about music’, which had a relation to the input phrase ‘playing piano’, and appears to enrich the vocabulary over merely repeating the input phrase. In RAPM-NONREL and CiAPM, the criterion “vocabulary richness” was rated 4 by 6/6 and 2/6 evaluators, respectively. This example indicates the potential usefulness of expanding the variety of expressions with phrases including hypernyms or hyponyms, based on the relations in ConceptNet.

However, when the action phrases from a user input are inserted into the system output, the utterances may sound more natural, as demonstrated in the following example. The input “And they enjoyed eating delicious food and alcohol.”, brought the outputs “Let’s talk about eating delicious food. What do you think about it?” (CiAPM) and “Talking about diminishing your own hunger... What is your opinion on that topic?” (RAPM-NONREL). In this case, a discussion about human needs, suggesting the related subject of ‘diminishing your own hunger’ to ‘eating delicious food’, would be a good topic for a deeper conversation. However, the preference of the conversational topic depends on the user, his or her interests and their English levels. For this reason, the repeating method (CiAPM) is considered to score above the other methods on average in all criteria except vocabulary richness.

We presume that the related concepts in ConceptNet are not always compatible with the dialogue context. In such cases, the responses are unsuited to the user’s need. This could be partly attributable to random selection of the related concepts. To avoid wandering away from the subject of the conversation, the related phrases must be carefully chosen to suit the context and the individual user, especially when applying phrases with their relations. In future work, the random selection must be replaced by a context processing module, a user profiler, and a language level estimator. A context processing module could select proper phrases by semantic analysis. Considering the ambiguity of multi-word expressions, detecting phrases after applying a topic modeling such as latent Dirichlet allocation might be useful for this purpose. In addition, complete reliance on ConceptNet, which lacks knowledge of some items and includes dubious entries, is also problematic.

As wrong inputs were not corrected, the open source checker found no mistakes. We might require a more powerful error correction approach. For error detection and correction suggestions, a promising solution is the Grammatical Error Correction (GEC) system based on the Neural Machine Translation (NMT) approach [Yuan, 2017]. In the experiments [Yuan, 2017], the NMT-based GEC system outperformed the SMT (Statistical Machine Translation)-based system even in a difficult subject-verb agreement problem.

From these results we can assume that repetition for confirmation plays an important part in conversation practice by Japanese learners of English. However, to assess whether this observation extends to other cultural backgrounds and individuals, broader experiments with more evaluators are needed. Furthermore, the evaluated conversations were very short, limited to one-turn dialogue (a user’s utterance and the corresponding system utterances). Whether the proposed methods maintain users’ interest in an actual conversation cannot be known at this stage. For this purpose, we must evaluate a fully developed system on multiple turns of free conversation. In long conversations for second language acquisition, a system that generates only repetitive utterances would bore users. The wide vocabulary of RAPM, providing related topics to user utterances, could potentially mitigate conversational deadlocks. Thus, combining the two methods (one that with repetitive utterances, the other using related topics) might be more efficient for language tutoring tasks.

4 Conclusion and Future Works

We proposed methods that automatically generate utterances for an English language tutor, and compared their performances with those of classic chatbots. Specifically, we evaluated how the generated expressions were received by Japanese subjects. Although our small-scale experiment does not allow drawing any conclusions about the stickiness level of these approaches yet, we found that ELIZA-like outputs offer more encouragement to users than Web- or common sense-based approaches. These inferences oppose the findings of [Rzepka *et al.*, 2005], who evaluated non-learning dialogues. In enriching the vocabulary of the system utterances, the proposed methods had shown their superiority, which could be potentially useful to improve users’ command of a foreign language.

However, using external corpora or crowd-sourced knowledge sources might incur serious drawbacks. Allowing the tutor excessive freedom, especially in learning material beyond the preferences of the user, risks misuse, as has occurred in Microsoft’s Tay and other chatbots [Michael, 2016]. In our approach, adaption of hand-crafted syntactic rules seem to be the only restriction, but because of majority voting in both British National Corpus- and ConceptNet-based methods we indirectly try to avoid semantic strangeness. This does not mean that corpora guarantee safe communication, and some topic restrictions might be needed from the outset. However, blocking slang and offensive words completely can be problematic, especially when considering more sophisticated personality modeling, which is required in longer-term conversational sessions.

As the next step, we plan to combine our method with estimating language level and supporting vocabulary acquisition algorithms [Mazur, 2016]. Error corrections could be improved by the annotated data[11], taking into account that Japanese students often make non-word spelling errors (making not existing spellings) [Nagata and Neubig, 2017]. Although our dialogue system is not yet ready for long-run conversational sessions, we should experiment on the tutor’s autonomy level in choosing topics related to user’s input, prior to larger scale testing. We plan to analyze which outputs

[11] http://www.gsk.or.jp/catalog/gsk2016-b/

are potentially harmful, and to determine appropriate countermeasures against these expressions.

Acknowledgements

The authors would like to thank DENSO CORPORATION for funding this work. We are grateful to anonymous reviewers for their detailed comments and helpful suggestions.

References

[Chen, 2014] Yi-Cheng Chen. An empirical examination of factors affecting college students' proactive stickiness with a web-based english learning environment. *Computers in Human Behavior*, 31:159 – 171, 2014.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

[Doyon, 2000] Paul Doyon. Shyness in the Japanese EFL class: Why it is a problem, what it is, what causes it, and what to do about it. *The Language Teacher*, 24(1):11–16, 2000.

[Ferschke *et al.*, 2013] Oliver Ferschke, Johannes Daxenberger, and Iryna Gurevych. The People's Web Meets NLP. In *Theory and Applications of Natural Language Processing*, pages 121–160. Springer, 2013.

[Finkel *et al.*, 2005] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[Higuchi *et al.*, 2008] Shinsuke Higuchi, Rafal Rzepka, and Kenji Araki. A casual conversation system using modality and word associations retrieved from the Web. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 382–390, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[Jia, 2009] Jiyou Jia. CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning. *Knowledge-Based Systems*, 22(4):249–255, 2009.

[Lee, 2016] Peter Lee. Learning from Tay's introduction. `https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/`, 2016. (accessed May 8 2017).

[Lison and Tiedemann, 2016] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[Loewen, 2007] Shawn Loewen. Error correction in the second language classroom. *Clear News*, 11(12):1–7, 2007.

[Mazur, 2016] Michal Mazur. *A Study on English Language Tutoring System Using Code-Switching Based Second Language Vocabulary Acquisition Method*. PhD thesis, Hokkaido University, 2016. `https://eprints.lib.hokudai.ac.jp/dspace/handle/2115/61833`.

[Michael, 2016] Katina Michael. Science fiction is full of bots that hurt people:... but these bots are here now. *IEEE Consumer Electronics Magazine*, 5(4):112–117, 2016.

[Nagata and Neubig, 2017] Ryo Nagata and Graham Neubig. Construction of japanese efl learner corpus for a study of spelling mistakes (in Japanese). In *Proceedings of the Twenty-third Annual Meeting of the Association for Natural Language Processing*, pages 1030–1033, 2017.

[Rzepka *et al.*, 2005] Rafal Rzepka, Yali Ge, and Kenji Araki. Naturalness of an utterance based on the automatically retrieved commonsense. In *Proceedings of IJCAI 2005 - Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland*, pages 996–998, August 2005.

[Speer and Havasi, 2012] Robert Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012.

[Tedick, 1986] Diane J Tedick. Research on error correction and implications for classroom. *ACIE Newsletter*, 1986.

[Tiedemann, 2012] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, pages 2214–2218, 2012.

[Toutanova and Manning, 2000] Kristina Toutanova and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical Methods in Natural Language Processing and very large corpora held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, volume 13, pages 63–70, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[Toutanova *et al.*, 2003] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, volume 1, pages 173–180, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[Vinyals and Le, 2015] Oriol Vinyals and Quoc Le. A Neural Conversational Model. *arXiv preprint arXiv:1506.05869*, 2015.

[Weizenbaum, 1966] Joseph Weizenbaum. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[Yuan, 2017] Zheng Yuan. Grammatical error correction in non-native english. Technical report, University of Cambridge, Computer Laboratory, 2017.