[1]Evaluating an Authoring Tool for Mini-Dialogs

Rohit KUMAR[a,b], Alicia SAGAE[a,b] and W. Lewis JOHNSON[a]

[a] *Alelo, Inc., 11965 Venice Blvd., Los Angeles, CA 90066 USA*
[b] *Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA*

Abstract. In order to facilitate efficient creation of instructional content for high proficiency language learning systems, we investigate the use of utterance templates through an intuitive authoring tool. Evaluation in the context of authoring mini-dialogs shows that the tool can help authors can achieve higher coverage of the target language.

Keywords. Authoring Tools, Language Learning, Mini-Dialogs, Utterance Templates

Introduction

Alelo's Tactical Language and Culture Training Systems (TLCTS) [1] employ a task-based approach, where the learner acquires the skills needed to accomplish particular communicative tasks [2]. Heavy emphasis is placed on spoken communication.

Our TLCTS courses, including Tactical Iraqi[TM] and Tactical French[TM] are in widespread use. As we continue to develop TLCTS courses for new languages, we are also developing instructional content which will allow the learners to practice the use of many more utterances in the game's communicative tasks, to help them achieve higher proficiency [3]. In this work, we describe development and evaluation of tools for authoring mini-dialogs which is one of the many types of instructional content used in TLCTS courses. There are about 800 mini-dialogs in Tactical Iraqi[TM] and over 300 in Tactical French[TM]. Any improvement in the mini-dialog authoring tools is likely to have a measurable and meaningful impact on our systems.

1. Designing a new Mini-Dialog Editor

Figure 1. Authored set of responses for a Mini-Dialog

The process of authoring mini-dialogs involves specification of an audio prompt that the game character would say and a text question that the tutoring agent would display. The learner is expected to respond to the prompt as guided by the question. In order to

give feedback on the learner response, we list a number of responses that the learner may say. Each response is annotated with a correctness label and a feedback. Figure 1 shows the responses underlying a typical mini-dialog .

From Figure 1 we observe that although the authored set covers some of the most common learner responses, the set is not nearly exhaustive in its coverage of possible responses. We also observe that a large number of relevant learner responses for the mini-dialog are simple variations of a small number of responses.

These observations motivated the design of new tools for creating variations of responses using Utterance Templates (UTs) [4].

$$\$chunk1 = (va\ commencer\ |commencera\ );$$
$$\$chunk2 = (par\ |avec\ );\ \$chunk3 = (les\ |des\ );$$
$$\$answer = On\ \$chunk1\ \$chunk2\ \$chunk3\ armes\ individuelles.; \qquad (1)$$

The generative power of UTs can be used to create variations from authored responses through a process we called *Templatization*. For example: Templatizing the response *On va commencer avec les armes individuelles* into the utterance template shown in (1), creates seven additional variations of the response. It must be noted that utterance templates can often over-generate responses which do not share the same correctness label or feedback as the response that was templatized.
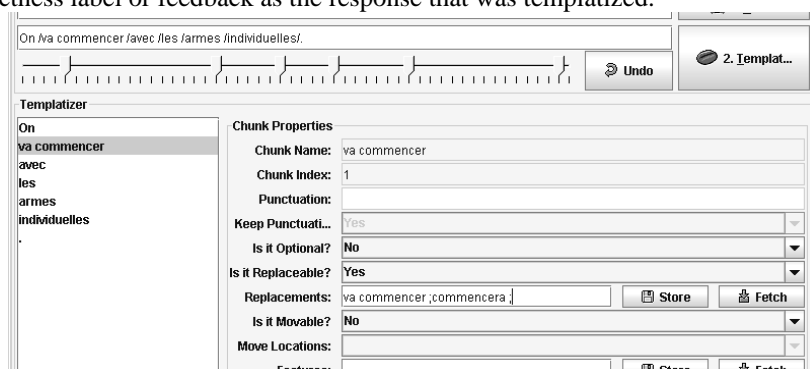

Figure 2. The Templatizer

Developing an authoring tool that uses utterance templates is a challenge, as a consequence of the fact that the target users are not necessarily specialists in AI and NLP. Figure 2 shows a tool called the *Templatizer* that we have developed to help the content authors use utterance templates. Authors can use the *templatizer* to create utterance templates without requiring them to write utterance templates. This is accomplished through the use of *power operations*. Currently the *templatizer* has three power operations.

- Power Operation 1: Is the chunk *Optional*?
- Power Operation 2: Is the chunk *Replaceable*? If so, specify the replacements.
- Power Operation 3: Is the chunk *Movable*? If so, specify the move locations..

2. Evaluation

We conducted an experiment to evaluate the new mini-dialog editor. Four members of Alelo's authoring team participated as subjects. The experiment was conducted over three one hour long sessions. During each session the subjects authored a different mini-dialog using the new tools. In an attempt to compensate for the relatively small

number of subjects, we divided each of the sessions into four sub-sessions referred to as *edits* here-on. During the second, third and fourth edits, the subjects circulated the mini-dialogs authored in the first edit among themselves. All subjects made improvements to each other's mini-dialog atleast once.

We extracted the existing content for the mini-dialogs being authored in this experiment from our Tactical French^{TM} system to measure the relative benefit of using the new editor. An ideal mini-dialog is one which captures all the possible learner responses, i.e. high recall as well as one which provides accurate feedback for each response, i.e. high precision. Hence we chose precision and recall as our outcome metrics. We used the set of all responses authored by any author in any of the four edits of each task as an approximation for the set of all possible learner responses for that mini-dialog. This set comprises of both relevant as well as irrelevant responses. We asked a French language instruction expert to rate each unique responses authored for each task on a six-point Likert scale. The rating represented the pedagogical usefulness of a response, 0 being not useful at all and 5 begin absolutely useful. The expert also specified if the correctness labels of each authored response were appropriate. For the purposes of computing our outcome metrics, we consider responses which have an appropriate correctness label and a usefulness rating above three to be relevant.

3. Results & Conclusions

We observe that using the new tool, authors can create a large number of responses in a short amount of time. There were about 5 to 10 times more relevant responses in the new mini-dialogs. However the precision of newly authored content was lower than the existing content. These observations suggest that the new tools improve the coverage of responses at the cost of introducing some irrelevant ones into the content. An ANOVA on the F-measure of each mini-dialog authored by each subject during each edit using task, subject and edit as factors revealed a significant effect of task ($F(2,47)=20.7$, $p<0.001$) and edit ($F(3,47)=1.8$, $p<0.001$) on the metric. We notice that as the mini-dialog undergoes multiple edits, its recall improves at the cost of precision.

On a survey, all four subjects indicated that the new mini-dialog editor was helpful. However, three of the four participants suggested that they would prefer to use the existing mini-dialog editor while authoring very simple mini-dialogs.

To summarize, it is evident from the evaluation presented here that the use of techniques like utterance templates can help instructional designers in creating better content for TLCTS courses. However, in order to further validate the benefit of using the new tools, we need to follow this work up with user evaluations in which content authored using the new authoring tools will be used by real language learners.

References

[1] Alelo Inc., Tactical Language & Culture Training Systems (TLCTS), *www.alelo.com,* 2008

[2] C. J. Doughty & M. H. Long, M.H. Optimal psycholinguistic environments for distance foreign language learning, *Language Learning & Technology 7(3)*, 50-80. 2003

[3] W.L. Johnson & A. Valente, Collaborative Authoring of Serious Games for Language and Culture, *SimTecT 2008*, March 2008

[4] J. Meron, A. Valente, W. L. Johnson, Improving the Authoring of Foreign Language Interactive Lessons in the Tactical Language Training System, *Workshop on Speech and Language Technology in Education (SLaTE)*, 2007