ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

Bilingual Response Generation using Semi-Automatically-Induced Templates for a Mixed-Initiative Dialog System

Wing Lin Yip and Helen M. Meng

Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
{wlyip, hmmeng}@se.cuhk.edu.hk

ABSTRACT

We have previously developed a framework for natural language response generation for mixed-initiative dialogs in the CUHK Restaurants domain [1]. This paper investigates the use of semi-automatic technique for response templates generation. We adopt a semi-automatic approach for grammar induction [2] to capture the language structures of responses from un-annotated corpora. We wish to use this approach to induce a set of grammars from our response data. The induced grammars are coupled with a parser to produce response templates in a semi-automatic way. Our response data consists of 2349 waiter responses. It is used as the training corpus for grammar induction. Unsupervised grammar induction is first performed, followed by using the learned grammars as prior knowledge for seeding the clustering process. Results show that the semi-automatically-induced response templates cover more than 50% of the hand-designed templates in templates coverage and provide more realization options. Performance evaluation indicates that the task completion rate has at least 90%, and most of the Grice's maxims as well as the overall user satisfaction scored at 3.5 points or above.

1. INTRODUCTION

This paper extends our previous effort of using hand-designed text generation templates for response verbalization [1]. We have previously worked on response generation in the context of the CUHK Restaurants domain, where our prototype system simulates the interaction between a customer and a waiter. In our earlier framework, we used a set of hand-designed text generate templates to verbalize the response message with appropriate selection of semantic, syntactic and lexical structures, each of which is associated with a response dialog state. The templates specify sentential structures that can incorporate semantic categories parsed from the user requests to generate a coherent system response. In order to reduce the manual work involved in hand-designing the text templates, we adopt a semi-automatic approach for grammar induction [2] to capture the language structures of responses. This approach can reduce manual effort of handcrafting response grammar. This can achieve enhanced portability across domains and languages. The clustering algorithm was previously implemented for acquiring semantic structure and syntactic structures from un-annotated corpora. We tend to use this approach to induce a set of grammar from our response data. The induced grammar should be useful for producing response templates in a semi-automatic way. The grammar induction is a statistical approach that uses agglomerative clustering to group words spatially and temporally. We use our response data as the training corpus for grammar induction. We perform unsupervised grammar induction first and use the learned grammars as prior knowledge for seeding the clustering process. A set of semi-automatically-induced response templates can then derived by parsing our response data with the induced grammars.

2. THE RESPONSE DATA

Our response data contains 2349 waiter response utterances which are mainly extracted from the CUHK Restaurants corpus [1]. The corpus contains 260 dialogs (with 1785 customer request utterances and 2176 waiter response utterances) that capture interactions between a customer and a waiter in a restaurant. We use those waiter response utterances from the corpus and further expand our response data by collecting 173 waiter response utterances from books [3, 4, 5, 6]. Some examples of response utterances are illustrated in Table 1.

"Have you decided on anything else?"
"Do you have a reservation?"
"How can I help you?"
"I would recommend smoked salmon scallop."
"Thank you."

Table 1: Response examples extracted from the response data.

3. RESPONSE GRAMMAR INDUCTION

The clustering algorithm was previously implemented for acquiring semantic structure and syntactic structures from un-annotated corpora. Details have been described in [2]. Grammars are induced by agglomerative clustering which groups words *spatially* and *temporally*. Spatial clustering groups words or phrases with similar left and right linguistic contexts by minimizing the symmetric divergence (*Div*), which incorporates the Kullback-Liebler (*KL*) distance (See Equation 1)[1]. Temporal clustering captures key phrases which co-occur frequently by maximizing the mutual information (*MI*), which indicate the degree of co-occurrence of two consecutive entities (See Equation 2). Spatial clusters (*SC*s) and temporal clusters (*TC*s), which are semantic categories and phrasal structures respectively, are produced iteratively.

$$Dist_{KL}(e_1, e_2) = Div\left(p_1^{left}, p_2^{left}\right) + Div\left(p_1^{right}, p_2^{right}\right) \quad (1)$$

where

$$Div\left(p_1^{adj}, p_2^{adj}\right) = \sum_{i=1}^{V} p_1^{adj}(i)\log\frac{p_1^{adj}(i)}{p_2^{adj}(i)} + \sum_{i=1}^{V} p_1^{adj}(i)\log\frac{p_1^{adj}(i)}{p_2^{adj}(i)}$$

[1] *$p_1(i)$* is the probability of the entity *i* adjacent (*adj*) to entity *e_1*. *V* is the vocabulary size for adjacent context.

$$MI(e_1, e_2) = P(e_1, e_2)\log\frac{P(e_1 \mid e_2)}{P(e_2)} \qquad (2)$$

There are two free parameters required in the clustering process: the minimum count threshold (*M*) and the number of merges in clustering (*N*). In each iteration, only entities with frequencies of occurrence above count *M* are considered, the *N* entity pairs with lowest values for *Dist* are merged to form spatial clusters and the *N* entity pairs with highest values of *MI* are merged to form temporal clusters. We empirically set *M*=3 and *N*=5. If a larger M is used, those contributive entities (e.g., "*mushroom*", "prawns" which can be the grammar terminals of FOOD) will be filtered. We experimented with N for different values *N*=1, *N*=3, *N*=5 and *N*=10 and compare the grammar size and time consumption for different values of *N*. By using N=5, the grammar induction can produce equally good grammar using fewer iterations and less computation time when compared with that of N=1 and N=3. We will not choose N=10 else the clustering process becomes too aggressive and the induced grammar becomes over-generalized.

Clustering is allowed to proceed to 140 iterations. As shown in Figure 1, the growth of clusters number stopped beyond iteration 130. From the output grammar, we selected 30 categories that we regard as basic semantic categories and phrases for the CUHK Restaurants domain. Examples are FOOD, REST_NAME, NUM, UNIT, etc. We manually complete the terminals for these categories and use them as seed categories to catalyze the rerun of the agglomerative clustering process. From Figure 2, we found that the number of terminals saturates within 100 iterations, so the clustering process is terminated. The output grammar from both clustering processes is post-processed by hand-refinement Refinement involves (i) pruning irrelevant rules; (ii) consolidating similar rules; (iii) completing their set of terminals and (iv) giving meaningful labels to the grammar rules, e.g., FOOD, REST_NAME, etc. Post-processing took about three hours to produce a grammar with 109 non-terminals and 457 terminals. Some examples of grammar rules are depicted in Table 2.

Induced Grammar Rules
FOOD → egg | mushroom | rice | salad | steak | …
REST_NAME → abc | hilton …
NUM → 1 | 2 | … | one | two | …
REST → REST_NAME restaurant
TIME → NUM minutes | NUM o'clock | NUM pm …

Table 2: Examples of grammar rules induced from grammar induction process.

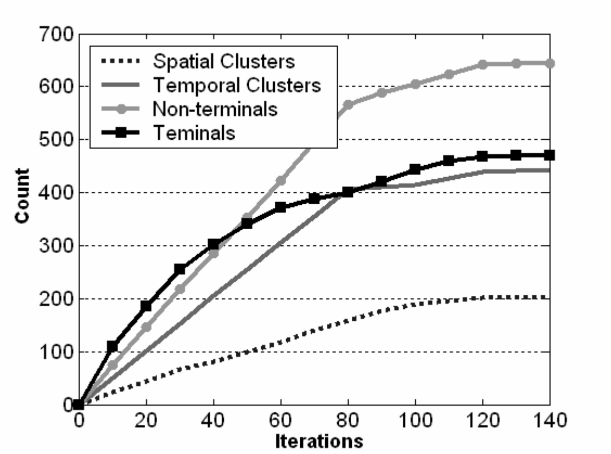

Figure 1: Growth of response grammar units in the grammar induction process.

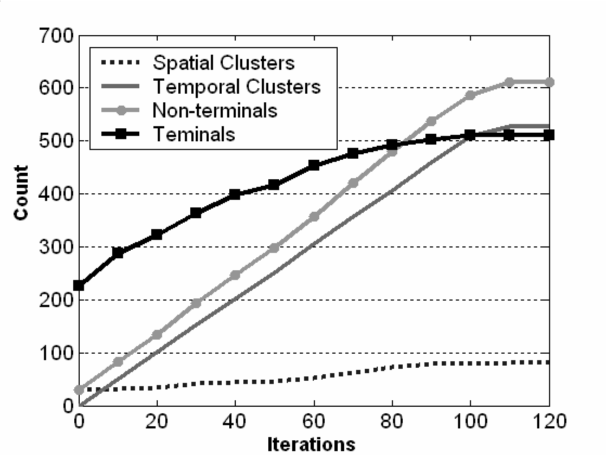

Figure 2: Growth of response grammar units for the grammar induction seeded 30 seed categories.

4. RESPONSE TEMPLATES GENERATION

The induced grammars are coupled with a parser and operated on our response data for retrieving key semantic concepts or structural phrases. The key semantic concepts and structural phrases are tagged automatically with their corresponding categories. All sentences in the response data are parsed with our induced grammars to produce tagged sentences. We obtain 642 distinct tagged sentences. Examples of tagged sentences obtained from our parser are shown in Table 3.

Before parsing:	*welcome to hilton restaurant*
After parsing:	WELCOME REST
Grammars:	WELCOME → welcome to REST_NAME → abc | hilton … REST → REST_NAME restaurant
Before parsing:	*you have reserved a table for four at 7 pm*
After parsing:	you have reserved TABLE_FOR at TIME
Grammars:	TABLE_FOR → a table for NUM NUM → 1 | 2 … TIME → NUM o'clock | NUM pm …

Table 3: Examples of tagged sentences parsed using induced grammars

Among those distinct tagged sentences, we have selected 278 tagged sentences with categories coverage greater than 30%. We define the categories coverage as the percentage of words that are covered by our induced grammars. Table 4 presents the computation of categories coverage. The first one will not be selected since its categories coverage is less than 30%. The selected tagged sentences are used as response template realization options.

Before parsing:	*i will look into the matter at once*
After parsing:	I_WILL look into the matter at once
Categories coverage:	2/8 = 25%
Before parsing:	*you have reserved a table for four at 7 pm*
After parsing:	you have reserved TABLE_FOR at TIME
Categories coverage:	6/10 = 60%

Table 4: Examples illustrate the computation of categories coverage.

We found that some tagged sentences actually belong to similar response structure. For example, the following tagged sentences are all referring to a response that offering further services. They are grouped into a single response template labeled ANYTHING_ELSE (See Table 5). Each template is associated with one or more tagged sentences that constitute a variety of realization options. Our approach generated 64 response templates in total. Some examples are shown in Table 6. The categories prefixed with '#' can be obtained either from grammar terminals or customer requests.

Template label: ANYTHING_ELSE	
Associated tagged sentences:	Realization options:
ANY_ELSE	*"Anything else?"*
WOULD_U_LIKE ANY_ELSE	*"Would you like anything else?"*
would there be ANY_ELSE	*"Would there be anything else?"*
do you need ANY_ELSE	*"Do you need anything else?"*
MODAL_I bring you ANY_ELSE	*"Can I bring you anything else?"*
MODAL_I serve you ANY_ELSE	*"May I serve you anything else?"*

Table 5: Template ANYTHING_ELSE offers multiple realization options by using its associated tagged sentences.

Template label: WELCOME_REST
Tagged sentence: WELCOME REST
Grammars: WELCOME → welcome to REST_NAME → abc | hilton … REST → REST_NAME restaurant
Content: welcome to #REST_NAME restaurant.
Template label: SUGGEST
Tagged sentence: HOW_ABT FOOD i RECOMMEND the FOOD WOULD_U_LIKE some FOOD
Grammars: HOW_ABT → how about i RECOMMEND → i would recommend WOULD_U_LIKE → would you like FOOD → seafood platter | steak | …
Content: How about #FOOD? I would recommend #FOOD. Would you like some #FOOD?

Table 6: Example templates WELCOME_REST (for welcoming customers) and SUGGEST (for suggesting food).

4.1. BILINGUAL RESPONSE TEMPLATES

We translated the 64 response templates from English to Chinese in order to achieve Chinese response generation as well. Table 7 depicts the translated response templates SUGGEST.

Template label: SUGGEST	
English responses:	Chinese responses:
How about #FOOD? I would recommend #FOOD. Would you like some #FOOD?	#FOOD 如何? 我推薦 #FOOD. 您喜歡 #FOOD 嗎?

Table 7: The English and Chinese responses for the template SUGGEST.

5. EVALUATION

We compared the resultant grammar-induced response templates with the hand-designed one, as described in [1]. Among the 64 semi-automatically-induced response templates, 57 of them carry similar semantic meaning and serve the same function as those hand-designed (101 templates). Our semi-automatically-induced response templates cover over 50% (57 out of 101) of the hand-designed templates. An extra 7 templates are discovered using the induced grammars and they do not appear in the hand-designed templates. One of the extra templates SHOW_LOC is shown in Table 8.

Template label: SHOW_LOC
Tagged sentence: SHOW you to the LOC
Grammars: SHOW → I will show LOC → bar | main restaurant …
Content: I will show you to the #LOC.

Table 8: The extra response template SHOW_LOC is used for showing location to customer.

Although the templates coverage of the semi-automatically-derived response templates is not as good as those hand-designed one, we observed that our semi-automatically-induced response templates can increase variability of response. It is because each response template offers more realization options than those hand-designed one. The number of realization options for each template has increased 50% in average. Take the template ANYTHING_ELSE as an example, the hand-designed template only gives 4 options for realizing a response, while the semi-automatically-induced response templates offers 6 response realizations (See Table 9).

Hand-designed Template	*"Anything else, sir?"*
	"Is that all?"
	"Is that anything else?"
	"Is there anything else, sir?"
Semi-automatically-induced Template	*"Anything else?"*
	"Would you like anything else?"
	"Would there be anything else?"
	"Do you need anything else?
	"Can I bring you anything else?"
	"May I serve you anything else?"

Table 9: Comparison on the realization options between hand-designed response template and semi-automatically-induced response template with template label ANYTHING_ELSE.

We have previously incorporated the cooperative response generation mechanism in an initial prototype of the interactive CUHK Restaurants system [1]. The system accepts typed natural language queries in English as input. We replaced the previous hand-designed templates with the semi-automatically-induced response templates for text generation. We asked ten subjects to interact with the system. Each subject is given three tasks: (i) reserve a table; (ii) order a meal; and (iii) ask for the bill. All interactions are automatically logged by the system. An example evaluation dialog is shown in Table 10. We can see that the food name "garden salad" and "orange juice" are parsed directly from the customer request into the category #FOOD, without English-to-Chinese translation, while those suggested food items are obtained from existing Chinese grammar terminals. The average number of dialog turns for each task is shown in Table 11.

C_1	*"Hello. I would like to order the Garden salad along with the orange juice."*	
W_1	*"Ok, garden salad, orange juice. Would you like anything else? I would recommend the Egg and mushrooms special."*	*"好的, garden salad, orange juice. 您還要其他東西嗎? 我推薦特色磨菇伴蛋."*
C_2	*"Er...Do you have any other main course to recommend?"*	
W_2	*"No problem. How about Roasted chicken with creamy cheese and mushroom sauce?"*	*"無問題. 芝士磨菇醬烤鷄如何?"*
C_3	*"Is there anything else?"*	
W_3	*"Would you like some Grilled fish fillet with tomato herb sauce."*	*"您喜歡番茄香草醬烤魚柳嗎?"*
C_4	*"I would like to have it."*	
W_4	*"OK! Would you like anything else?"*	*"好的! 您還要其他東西嗎?"*
C_5	*"No, thanks."*	
W_5	*"You are welcome. You have ordered garden salad, orange juice, grilled fish fillet with tomato herb sauce. They will be ready in 15 minutes."*	*"不用客氣. 您點選了garden salad, orange juice, 番茄香草醬烤魚柳. 您的食物將於十五分鐘內到."*

Table 10: An example evaluation dialog extracted from the system log (C_i: customer request, W_i: waiter response).

Tasks	Reservation	Order	Bill
Average #dialog turns	6.5	5.5	3.2

Table 11: Average number of dialog turns across the 10 evaluation dialogs for each of the three tasks.

We evaluate the dialogs in terms of the task completion rate, Grice's maxims [7] as well as overall user satisfaction.

5.1 Task Completion Rate

All the evaluation dialogs logged by the system have been checked for task completion. A task is considered complete if the appropriate confirmation message is present in the dialog. For the reservation task, we search for the system confirmation. A task is considered complete as long as the appropriate confirmation message exists, even if there are incoherent dialog turns involved. The simplicity of our evaluation tasks have led to high task completion rates across the evaluation dialogs (See Table 12).

Tasks	Reservation	Order	Bill
Task Completion Rate	90% (9/10)	100% (10/10)	100% (10/10)

Table 12: Task completion rates across the 10 evaluation dialogs for each of the tasks – reserving a table, ordering food and requesting the bill.

5.2 Grice's Maxims and Perceived User Satisfaction

We also evaluate response generation in terms of Grice's Maxims as well as overall user satisfaction. Each subject was asked to fill out a questionnaire that contains three sets of questions, one for each task (i.e. reservation, ordering food and requesting the bill). The set of questions is identical across the tasks and relate to Grice's Maxims as well as overall user satisfaction. The subjects were asked to respond to these questions on a five-point Likert scale and the results are shown in Table 13. A t-test shows that most of the maxims are significantly better than average (Likert score 3) at α=0.05, except the maxim of Quantity for reservation task and the maxim of Manner for order task.

	Reservation	Order	Bill
Maxim of Quality	4.1 (0.9)	3.9 (1.0)	3.7 (1.2)
Maxim of Quantity	**3.4 (1.0)**	4.0 (1.1)	4.0 (1.1)
Maxim of Relevance	4.2 (1.1)	3.8 (0.9)	4.3 (1.3)
Maxim of Manner	3.6 (0.8)	**3.6 (1.1)**	4.2 (1.3)
Satisfaction	3.5 (0.8)	3.8 (0.9)	4.0 (1.2)

Table 13: Average scores and standard deviations (in brackets) in a five-point Likert scale (1--very poor, 5—very good) obtained from evaluation in terms of Grice's Maxims and overall user satisfaction.

The score for Maxim of Quantity for reservation task is relatively low. This reflects that some of our semi-automatically-induced response templates do not give sufficient information to the customer. Table 14 presents an illustration – in the second and third dialog turns, the customer does not give a specific location, however, the system is expecting an answer with finite location such as "*near the window*". This situation does not happen when we use the hand-designed template. It is because the hand-designed one includes the location options for customer to choose (i.e., *"Where would you like to sit? By the window, in the main restaurant or in the bar?"*).

C_1	*"I want to book a table for two people tomorrow at 8 pm."*	
W_1	*"Where would you like to sit?"*	*"您想坐在那兒?"*
C_2	*"Anywhere."*	
W_2	*"Where would you like to sit?"*	*"您想坐在那兒?"*
C_3	*"You can assign the location to me."*	
W_3	*"Where would you like to sit?"*	*"您想坐在那兒?"*

Table 14: An example extracted from the evaluation dialogs to illustrate incoherent dialog turns.

6. CONCLUSIONS

This paper reports on our approach towards semi-automatic natural language response templates generation in the CUHK Restaurants domain. We have tried to develop a set of response templates semi-automatically from corpus. We have adopted a semi-automatic approach for grammar induction to capture the language structures of responses. Agglomerative clustering is used to group words spatially and temporally. Several experiments are conducted to determine the two free parameters M = 3 (minimum count threshold) and N = 5 (number of merges). The resultant grammar is post-processed by labeling tags with meaningful labels, completing the terminals for some categories, pruning irrelevant clusters and consolidating clusters that belonging to same categories. We have injected some prior knowledge that was learned from the unsupervised grammar induction. Seed categories obtained from unsupervised process are used to catalyze grammar induction so as to produce longer phrases using less iteration. A set of semi-automatically-induced response templates was derived by parsing our response data with the induced grammar. Those templates are compared with the hand-designed templates in terms of templates coverage and number of realization options. Although the semi-automatically-induced response templates cannot outperform the hand-designed one in templates coverage, they still have a competitive performance with coverage greater than 50% (57 out of 101). Our approach also increases the variability of response by providing more realization options. Performance evaluation based on the 30 interactive dialogs from 10 subjects showed at least 90% task completion rate. Most of the Grice's maxims as well as the overall user satisfaction scored at 3.5 points or above.

7. ACKNOWLEDGEMENTS

The work described in this paper was partially supported by grants from the Research Grants Council of the Hong Kong SAR Government, China (Project Nos. CUHK4326/02E).

8. REFERENCES

[1] H. Meng, W. L. Yip, M. Y. Mok, S. F. Chan, "Natural Language Response Generation in Mixed-Initiative Dialogs using Task Goals and Dialog Acts". In *Proceedings of European Conference on Speech Communication and Technology, 2003*

[2] H. Meng, K. C. Siu, "Semiautomatic Acquisition of Semantic Structures for Understanding Domain-Specific Natural language Queries". In *IEEE Transactions on Knowledge and Data Engineering, Vol.14, Issue 1, p.172-181, Jan/Feb 2002.*

[3] 翁顯雄, "人際互動英語". *萬里書店出版.*

[4] Carmen Tang, "醒目酒店英語". *世界出版社.*

[5] 陳惠編, "銷售員必備英語會話". *集英文化.*

[6] 胡潤生, "餐飲英語會話專集". *雄峰出版社.*

[7] R. Frederking, "Grice's maxims: do the right thing". In *Frederking, R. E., 1996.*