Evaluation of text generation system

Xinyu Deng and **Jun-ichi Nakamura**
Kyoto University
deng@pine.kuee.kyoto-u.ac.jp

1 Background

At present, the population of non-native speakers of English is twice that of native English speakers. In order to satisfy the needs of non-native users, we developed a text generation system – the SILK (Generation *S*ystem for *I*ntermediate *L*evel non-native spea*K*ers on discourse level) system (Deng and Nakamura, 2006).

In Natural Language Generation, evaluation is an important issue. Until now, researchers have been trying to find valid and widely accepted methods. Dale and Mellish (1998) mentioned three aspects of evaluation: evaluating properties of the theory; evaluating properties of the system; and evaluating the application potential for a system. Furthermore, researchers chose the evaluation methods which were appropriate for their needs. For example, Coch (1996) used a black-box methodology to assess three techniques for producing multisentential texts; Yeh and Mellish (1997) compared human-created results and computer-generated results.

In this paper, we introduce two methods of evaluating the SILK system. The rest of the paper is arranged as follows. Section 2 evaluates the system itself. Section 3 demonstrates how to assess the generated texts by human subjects. In Section 4, we draw a conclusion.

2 Evaluating the system

In this section, we evaluate the validity of the system by justifying the evaluation function of Genetic Algorithm (Cheng and Mellish, 2000).

2.1 Correlations of the values

The evaluation function of the SILK system was based on four features: position of nucleus, between-text-span punctuation, complex multiple cue phrases (CMCPs), and patterns of punctuation. Deng and Nakamura (2006) put forward four heuristics which show the preferences among the possible states of each feature. The main opinion of this study is that the preferences among the features rather than the features themselves decide the ease of a text on discourse level. That is, if a text is evaluated by using two or more values satisfying the preferences, the evaluation results would be consistent with each other. Based on this consideration, we examined five values and their correlations. Firstly, we generated five values satisfying the four heuristics by a constraint-based program. Table 1 shows three of them, whose ranges are: -10 $\sim$ 20, -30 $\sim$ 30, -20 $\sim$ 70.

Features	Values 1	2	3
Position of nucleus			
good position	15	19	69
normal positiom	1	3	2
bad position	-3	-27	-5
Between-text-span punctuation			
good punctuation	6	22	37
normal punctuation	2	10	11
bad punctuation	-9	-8	-6
CMCPs			
good CMCPs	7	12	65
normal CMCPs	5	9	33
bad CMCPs	-2	-3	-10
Pattern of punctuation			
good pattern	14	17	57
normal pattern	11	2	22
bad pattern	-7	-1	-19

Table 1: Three different values satisfying the same constraints

Using value 1 and value 3, we generated all possible combinations of a text. In order to describe whether and how strongly value 1 and value 3 are related, we drew the scatterplot of the scores (Figure 1). The X-axis and Y-axis represent the scores obtained from value 1 and value 3 respectively. We can see that the scatterplot has a linear pat-

tern which indicates a high correlation between the two values, i.e., high scores on the X-axis are associated with high scores on the Y-axis. In addition, we found that the frequency distribution of the two scores illustrated by the histogram follow a "Normal Distribution". Though the means of the two distributions are different, the shape of the two histograms are very similar. This shows that the behaviours of the two values are nearly the same while generating a text.

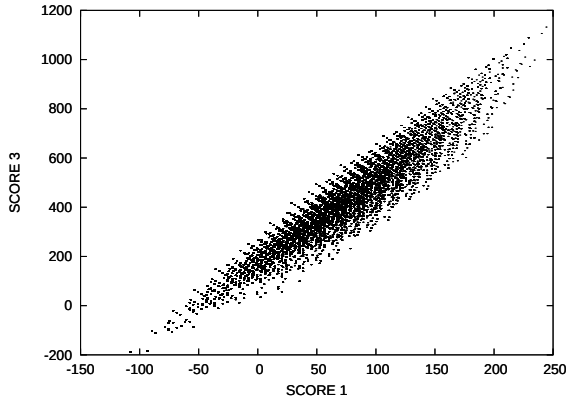

Figure 1: Scatterplot of Score 1 and Score 3

The most common measure of correlation is the Pearson's correlation. In order to illustrate that different values agree with each other on measuring the texts, we calculated the Pearson correlation coefficients between all pairs of the five values. The results (Table 2) show that all correlation coefficients are more than 0.9, which means that there is a nearly perfect positive linear relationship between the five values.

	Score 2	Score 3	Score 4	Score 5
Score 1	.914	.969	.902	.917
Score 2		.946	.912	.906
Score 3			.950	.921
Score 4				.937

Table 2: Correlations between five values

2.2 Comparison of original texts and generated texts

Generally, for two texts with the same rhetorical structure, the better one would have a higher score when evaluated by a value satisfying the preferences. In this section, we compare the scores of the generated texts and their original counterparts to examine the quality of the generated texts.

We chose three texts from corpus CNNSE, which contains texts appropriate for imtermediate level non-native speakers. Then we manually created the RST trees (Mann and Thompson, 1988) of these texts to represent the rhetorical structures of them.

We used value 1 shown in Table 1 and the evaluation function to score the three texts. We then ran the GA for 5000 iterations for 10 times on the RST tree of each texts using value 1. Table 3 shows the highest scores and the average scores of the generated texts and their original counterparts (CNNSE texts). In addition, Figure 2 illustrates these scores in detail. For TEXT 1, the highest score of the generated texts is the same as the original counterpart; for TEXT 2, the highest score of the generated texts is higher than the orignal one; for TEXT 3, though the highest score of generated texts is a little bit lower than the original one, the difference is not too much. The experiment results showed that the highest scores of the texts generated by GA were close to their original counterparts. This means that the evaluation function of GA is rational and the generated texts are appropriate for intermediate level non-native users.

	Score of CNNSE text	Generated text	
		Highest score	Average score
Text 1	372	372	329
Text 2	577	585	510
Text 3	240	236	214

Table 3: The scores of three generated texts and their original counterparts

3 Assessing text comprehensibility by human subjects

Although we have justified the evaluate function, we still don't know if the texts generated by the GA method are appropriate for intermediate level non-native speakers. In this section, we designed a questionnaire and asked the human subjects to rate the comprehensibility of the generated texts.

The questionnaire contained 9 generated texts with different lengths and structures. For each text, we ran five times with 5000 iterations, and chose the best result (the one with the highest score) to be used in the questionnaire. The comprehensibility of each text is to be rated using a

number between 0 and 10, where a higher number represents a text which is easier to understand. For example, "0" means "can not understand", "10" means "very easy to understand".

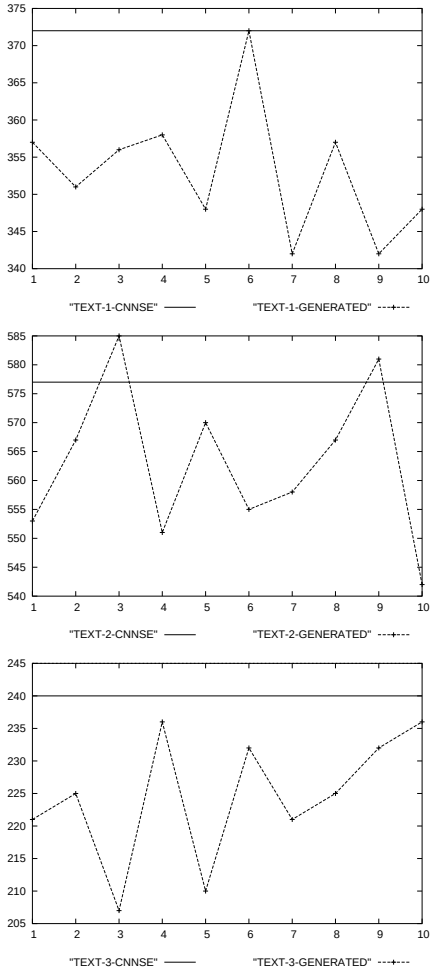

Figure 2: Scores for three texts

Group I			Group II		
No.	Age	Sex	No.	Age	Sex
Q1	10's	male	N1	10's	female
Q2	20's	female	N2	20's	male
Q3	20's	female	N3	20's	male
Q4	30's	male	N4	30's	male
Q5	40's	female	N5	40's	female

Table 4: Two groups of human subjects

In order to enhance the generalizability of the experimental results, we divided the subjects into two groups (Table 4): Group I did the experiment under a quiet environment; Group II did the experiment under a noisy environment. Both groups used the same questionnaire. 10 intermediate level non-native speakers took part in the experiment. Of the subjects, 5 persons were in Group I (2 were male, 3 were female); another 5 persons were in Group II (3 were male, 2 were female). The age of subjects ranged from 17 to 45; in Group I and Group II, the the mean was 29 and 30 years respectively.

We used two methods to analyse the questionnaire results. Firstly, we used the Kappa statistics (Carletta, 1996) to assess the inter-rater reliability. The average Kappa value of Group I was 0.4916 (Table 5); for Group II, was 0.4562 (Table 6). Moreover, the mean comprehensibility of the two groups were different. For the group under the quiet environment, the mean was 8.3; for the group under the noisy environment, was 7.8. In light of these results, an explanation is that noise affected the raters' reading ability, and their ability to concentrate on reading the texts. The results also showed that the generated texts could be understood quite well, because the mean comprehensibility was higher than 7.5 in each group.

In fact, the Kappa values of each group only represented "moderate" agreement. We think that there were two reasons which caused this results. One is that all the subjects were naive raters, so we could not guarantee that they used the same criteria to score the texts though we introduced the rating scale to them before the experiments. The second reason is that the length of the texts affect the reading ability of the raters. Generally, the longer the text is, the more complicated its structure is. In the tests, the measure of agreement got worse if a text was longer.

	Q2	Q3	Q4	Q5
Q1	.4405	.4489	.4726	.5739
Q2		.4405	.4726	.5804
Q3			.4874	.4319
Q4				.5672

Table 5: Kappa Statistics of Group I

	N2	N3	N4	N5
N1	.4405	.4650	.4489	.4489
N2		.5739	.4319	.4405
N3			.4571	.4229
N4				.4319

Table 6: Kappa Statistics of Group II

On the other hand, Pearson's correlation was used to measure the inter-rate agreement of the two groups as well. In the experiments, the raters showed a relatively high level of agreement, because most of the correlation coefficients ranged from 0.5 to 0.75. In the group under the quiet environment, 2 out of the 10 correlation coefficients were greater than 0.75, while none correlation efficient was lower than 0.5 (Table 7). The average Pearson's correlation coefficients was 0.668. In the group under noisy environment, one correlation coefficient was greater than 0.75, but none correlation coefficient was lower than 0.5 (Table 8). The average Pearson's correlation coefficients was 0.637. As expected, it was proved again that environmental noise affects the rater's reading ability. In addition, agreement appears to be independent from life environment, because Pearson's correlation coefficients showed a clear tendency for a positive correlation among the scores rated by subjects with different backgrouds.

The discussions above show that the evaluation function of GA are effective in generating texts for intermediate level non-native users. Though the inter-rater agreement was not very high, we think that the degree of agreement could be improved if the raters were trained. Moreover, the results of the questionnaire also showed that long texts are fairly difficult to understand for non-native speakers, especially under a noisy environment.

	Q2	Q3	Q4	Q5
Q1	.669	.636	.673	.796
Q2		.589	.585	.701
Q3			.791	.547
Q4				.696

Table 7: Pearson Correlations of Group II

	N2	N3	N4	N5
N1	.638	.624	.603	.612
N2		.771	.669	.698
N3			.589	.599
N4				.571

Table 8: Pearson Correlations of Group II

4 Conclusion

In this paper, we adopt two methods to evaluate the SILK system. The validity of using GA was proved by evaluating the generation system itself and assessing the comprehensibility of the generated texts by human subjects. The evaluation results show that the system is effective and the generated texts are appropriate for non-native users.

Reference

Xinyu Deng and Jun-ichi Nakamura (2006). Generating easier texts on discourse level. In *the 12th Annual Meeting of Natural Language Processing*, Japan.

Robert Dale and Chris Mellish (1998). Towards the evaluation of natural language generation. In *the First Conference on Language Resources and Evaluation*.

Hose Coch (1996). Evaluating and comparing three text-production techniques. In *COLING'96*.

Ching-Long Yeh and Chris Mellish (1997). An empirical study on the generation of anaphora in Chinese, *Computational Linguistics*, Vol.23, No.1.

Hua Cheng and Chris Mellish (2000). Capturing the interaction between aggregation and text planning in two generation systems. In *Proceedings of the First International Natural Language Generation Conference*.

William Mann and Sandra Thompson (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3).

Jean Carletta (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, Vol.22, No.2.