Applied NLG system evaluation: FlexyCAT

Nestor Miliaev, Alison Cawsey and Greg Michaelson
Department of Computer Science
Heriot-Watt University
{ceenym, alison, greg}@macs.hw.ac.uk

Abstract

Evaluation is an important part of NLG projects, however NLG system evaluation often consists of usability or static text quality assessment. This paper presents an NLG system, FlexyCAT, and experiments that enabled us to evaluate the degree of knowledge re-use and the task-specific value of generated texts.

1 Introduction

Many applied natural language generation (NLG) systems have been created recently producing various kinds of texts for different applications. However, only a few of these systems have been formally evaluated, and the evaluation performed have focused on the grammaticality and fluency of the output text, rather that on its effectiveness (Colineau et al., 2002). Issues such as knowledge re-use were largely ignored in these projects.

This paper describes the evaluation of an NLG system Flexible Computer-Aided Technical Writer (FlexyCAT) focusing on the assessment of the task-specific quality of generated texts and knowledge re-use when the system is used for the description of different devices. This work has shown, firstly, that the task-oriented quality of generated texts can be comparable to that of human-crafted texts; and secondly, how knowledge re-use allows us to extend the applicability of an NLG system for the description of different technical systems and to reduce time taken to create a document.

2 Existing practise of NLG System Evaluation

Recently, it had become widely accepted that work in NLG should pay closer attention to the evaluation of results. The aspects of an NLG system to evaluate and the metrics to use for this are defined by the goals of the NLG system. The most common advantages of using NLG comparable with, e.g., Machine Translation (MT) are considered the following (Reiter and Dale, 2000):

- High quality output text that is generated based on machine data and does not require post-processing
- Simultaneous production of text versions in different languages
- Consistency of text between the versions and with a domain model
- Lower cost and time of text revision (e.g., when domain changes)
- Dynamic text generation upon user query
- Potentially good knowledge re-use

Unfortunately, most NLG systems evaluation carried out to date did not cover all these aspects. Usually NLG evaluation deals with 1) System quality; 2) Text quality.

System quality evaluation consists of measuring characteristics like usability, generation speed

and system robustness. Such evaluation follows common patterns of software systems evaluation, see e.g., (Newman and Lamming, 1995). NLG-specific principles and metrics of system evaluation are described in (Mellish and Dale, 1998) and (EAGLES, 1995).

One of the most thorough usability evaluation of an NLG system has been carried out in the AGILE project; the course of experiments and their results are presented in (Hartley et al., 2000).

Assessing the quality of a generated text (based on (Mellish and Dale, 1998) and (Hartley et al., 2000)) includes rating of the text against such criteria as *Accuracy*, *Fluency* (including *Acceptability* and *Grammaticality*) and *Lexico-grammar coverage*.

Methods of evaluation appropriate to these criteria can be grouped into three classes, as suggested in (Bangalore et al., 1998):

Intrinsic that typically consists in asking human judges to rate the quality of generated texts. Intrinsic evaluation is most common, and was carried out e.g., in AGILE (Hartley et al., 2000) and MIRADOR (Eddy and Cawsey, 2002); it is fairly simple and straightforward, however depends a great deal on the evaluators' expertise and personal preferences.

Extrinsic or *task evaluation*, where the user's ability to perform some task using a generated text or the impact of the text on the user behaviour is assessed. Extrinsic evaluation is less common because experiments are more difficult to carry out. However, it allows us to estimate the task-oriented worth of a document. This kind of evaluation was used in Isolde (Colineau et al., 2002), STOP (Reiter et al., 2001), and other NLG projects.

Comparative where the objective is to directly compare the performance of different generation systems and formalisms. Comparative evaluation is not often used because of its complexity. An example of this kind of trial for the XTAG project is described in (Bangalore et al., 1998).

Other aspects of NLG, such as knowledge reuse, time of document creation or system flexibility have not received enough attention so far. Although there were some reports on task-oriented evaluation, these experiments are not common and many of those did not include the comparison of the user performance on human-crafted and generated manuals, like in IDAS (Levine and Mellish, 1995).

During FlexyCAT evaluation, we paid extra attention to task-oriented text quality, knowledge reuse and the benefits the author obtains by using a flexible planning technique.

3 FlexyCAT: System Description

FlexyCAT is an NLG system for producing manuals for technical devices, mainly home appliances, e.g., TVs, VCRs, cameras, etc. Manuals consist of texts divided up into sections, each of which describes an individual procedure, consisting of a sequence of steps that the user is to perform to attain his/her goal. No generation of explanatory or warning information is included; generated texts are purely instructional and are thus like the 'minimalist instructions' described in (Colineau et al., 2002).

Texts are generated in two languages (English and Russian) given a domain model. The domain model represents an object-oriented description of a device being documented. Such a representation contains a description of all device elementary constituents and their functioning. The description of the functioning of elementary objects includes events (user actions) and object functions, that have preconditions and actions. Both preconditions and actions are described in terms of elementary object properties. Objects, events and actions are represented using linguistic classes. The generation system uses both linguistic classes and domain knowledge for the production of a grammatical manual text in two languages. The output of FlexyCAT is a ready to use manual for the device. FlexyCAT's GUI provides the user with facilities for creating and editing a domain model and dictionaries of linguistic classes. It also allows the user to easily assign linguistic values to the elements of the domain model. For more information see (Miliaev et al., 2002).

Two main advantages of FlexyCAT compared with other NLG systems are:

- FlexyCAT is an integrated tool allowing extensive editing of both linguistic and domain knowledge. This, and an object-oriented design of the knowledge, enable firstly, good knowledge re-use, and, secondly, the production of manuals for a variety of different technical devices using a single NLG tool.
- FlexyCAT offers a flexible approach to text planning. The planner produces a candidate text plan automatically, based on the domain model. When there is a necessity to obtain a better quality text, the user edits this plan using an interactive planning utility. A text is generated based on the text plan. For automatically generated plan it is of a draft quality, while a better quality text is obtained from an edited plan. The automatic feature facilitates the work of the technical author by producing a text draft quickly and at fairly little cost.

An example original text from the corpus is given below; it contains a considerable proportion of explanatory and causative information:

```
How to play a CD
1. Press the PLAY/PAUSE button
The CD will begin to play and
the track number will be shown
in the display
2. To pause the CD press the
PLAY/PAUSE button
The CD will stop and the PAUSE
INDICATOR will flash in the
display
```

An automatically generated corresponding text looks like this:

```
Playback CD
1. Press PLAY/PAUSE button to
start CD playback
2. Press PLAY/PAUSE button to
pause CD
```

After the plan refinement, we get the following text, with richer rhetorical structure (the rhetorical relation in the first sentence is changed from 'motivation' to 'means'; in the second sentence the sequence of the nucleus and satellite clauses changed) and, we believe, a better quality:

```
Playback CD
1. Start CD playback by pressing
PLAY/PAUSE button
2. To pause CD, press PLAY/PAUSE
button
```

Below we will describe our experience and results of the evaluation of the FlexyCAT NLG system.

4 FlexyCAT Evaluation

The evaluation of FlexyCAT is ongoing. This section describes evaluation experiments that have been carried out to date and their results.

So far, three stages of the evaluation have been performed. These are:

- Experiments on knowledge re-use and text production for different devices
- Subjective assessment of text quality
- Task-oriented text quality evaluation

4.1 Knowledge Re-use and Resource Management

This experiment was targeted, firstly, to assess the appropriateness and effectiveness of the chosen domain description structure and the domain engineering tool for documenting different technical devices. The second goal was to assess the degree of knowledge (both linguistic and domain) re-use.

The experiment was set up as follows. Manual texts, each about 3 pages long, were selected for three different pieces of hardware. All texts contained sections, subsections and paragraphs. FlexyCAT was used to build a domain model for the pertinent subset of each of these devices and manual texts were generated in both English and Russian. The interactive planning utility was used to obtain text as close to the original manual as possible, in terms of wording and word order, see the description of the metrics for generation accuracy and their importance in (Hartley et al., 2000). In many cases the generated texts were identical to the original ones. The time taken to complete the

entire process of document creation was measured and the number of knowledge base (KB) elements re-used was estimated for each task.

The course of the experiment was as follows. A domain model and a manual for the first VCR (VCR1) was created. Then the manual for a similar device, VCR2 was created. Both manuals contained similar set of functions (however the controls and operation were different in some cases). During the creation of the manual for VCR2 both linguistic and domain resources created during the production of the manual for VCR1 were available and were re-used.

The second phase of the experiment consisted in the creation of a manual for a fairly distant device, a combined CD-Radio-Cassette (Combine). That device contained parts similar to that of a VCR, namely a tape recorder and a CD player; and a part different from any known in a VCR, a radio tuner. The manual for Combine was created twice – once from scratch without using any of the existing resources, and a second time making use of the resources created for VCR1. The role of a technical author was played by the system author, who knows the system very well and had created manuals for these and similar devices number of times in the course of the system development and evaluation. Thus, the learning effect that may have affected time of subsequent document creation may be largely ignored.

FlexyCAT uses three classes of linguistic primitives: nominal expressions (consisting of up to two nouns), verbs and adverbials. Adverbials were not considered in the evaluation experiments. There exist two versions of each nominal expression and verb, one for English and one for Russian. Complex classes, that encapsulate the references to both versions, are called *Nominals* and *Verbals* respectively. *Nominals* denote nominal expressions that could be used to name a domain entry. *Verbals* are used to specify events and actions in a domain model. A *Domain object* represents a single elementary constituent of a device; an object's description may include a number of references to nominals and verbals.

Figures 1, 2 and 3 show the results of knowledge re-use which occurred in the course of the experiment.

KB element	Reused	Total	Percentage
Nominals	30	84	36%
Nominals*	12	84	14%
English nouns	41	83	49%
Russian nouns	52	87	60%
Verbals	32	46	70%
English verbs	31	38	82%
Russian verbs	33	35	94%
Domain objects	23	38	61%

Figure 1: Knowledge re-use for VCR2

In Figures 1, 2 and 3 Nominals* denote nominals that were amended slightly for the description of new objects in a new device.

Domain objects almost always underwent change to conform the description of a new device. 'Re-used' objects, hence, are those that were amended only slightly.

KB element	Reused	Total	Percentage
Nominals	12	24	50%
Nominals*	5	24	21%
English nouns	23	30	77%
Russian nouns	23	27	85%
Verbals	19	24	79%
English verbs	18	18	100%
Russian verbs	20	21	95%
Domain objects	10	15	67%

Figure 2: Knowledge re-use for CD: Tape recorder sub-set

KB element	Reused	Total	Percentage
Nominals	13	51	25%
Nominals*	8	51	16%
English nouns	28	58	48%
Russian nouns	30	59	51%
Verbals	22	31	71%
English verbs	23	31	74%
Russian verbs	23	34	68%
Domain objects	15	28	54%

Figure 3: Knowledge re-use for CD: whole device

As can be seen from Figures 1 - 3, the percentage of knowledge re-use is fairly high. By that, the percentage is higher between similar devices and increases for simpler elements; nouns and

verbs, show the best degree of re-use, with verbs in some cases being re-used by 100%. This could be explained by the fact that the way controls operate are fairly persistent across many devices. Nominals are re-used a little less. That is caused mainly by the difference in control names in different devices.

For the comparison of time taken to produce documents using FlexyCAT it can be assumed that manuals for VCR1 and VCR2 are pretty much identical in size and labour-intensity to produce. The manual for Combine is slightly shorter. That explains the necessity of producing it from scratch first – we needed it to assess the initial time required to compare it with that when re-using existing resources. Figure 4 shows the time measures for each experiment.

Experiment	Time
VCR1, no resource exists	580 mins
VCR2, VCR1 resource used	340 mins
Combine, no resource used	430 mins
Combine, VCR1 resource used	270 mins

Figure 4: Time of manual creation in FlexyCAT

As can be seen, the time to create a manual is being reduced significantly when re-using existing resources. This suggests good knowledge re-use and effort reduction when making use of existing resources, even for a fairly different device.

4.2 Subjective Assessment of Text Quality

An experiment on subjective assessment of text quality was conceived as a pilot study for the task-oriented text quality evaluation. However, it has yielded some interesting results, especially in comparison with the latter, so we will refer to both.

The experiment was set up as follows. We used a within subject design. A group of nine native speakers of English were offered a set of nine text excerpts each. Text excerpts were short instructions up to half a page long. Texts were given in three different versions: original manual text; automatically generated draft; and a text generated after plan editing (further regarded as 'original', 'generated' and 'edited' respectively). See text examples in section 3. Each set contained only one version of each text (either original, automatically generated or generated after plan editing). Text versions were evenly distributed across different sets and the evaluators did not know which version of each text excerpt s/he got. Illustrations from original manuals were included in all versions, where applicable. Text layout was the same in all versions.

The evaluators were asked to assess quality and understandability of each text on the scale from 1 to 5 (1 is bad, 5 is excellent).

The average scores and standard deviation across different document versions were as follows, see Figure 5.

Text quality	
Original	4.07(0.92)
Generated	4.19(0.79)
Generated+edited	4.00(0.68)
Text understandability	
Original	3.85(1.03)
Generated	3.93(1.04)
Generated+edited	3.81(1.08)

Figure 5: Mean(standard deviation) of subjective scores of text quality

This figure shows that different text versions were ranked quite closely, which lets us conclude that all text versions are similar from the reader's point of view. This is similar to the results obtained in the AGILE evaluation.

Surprisingly, automatically generated text drafts (having no diverse sentence structures and lacking rhetorical markers) have been preferred over other text versions. However, a t-test has shown that the difference in user preference is not significant. The results of the t-test at df=57 are: t=-0.408, p=0.685 between original and generated; t=0.272, p=0.7866 between original and edited; and t=-0.680, p=0.4993 between generated and edited texts.

After this pilot study, a larger scale task-oriented experiment was run.

4.3 Task-oriented text quality evaluation

This time the user had to carry out real tasks with actual pieces of hardware, instead of subjectively assessing text qualities. There were three sets of hardware: VCR1+TV, VCR2 and a CD player.

The tasks included: 1) Assemble all components of VCR1+TV; start, stop and eject the videotape; 2) Program VCR1 to automatically record a certain program; 3) Set clock of VCR2 to a certain time; 4) Power a CD player, load a CD, start playback and program the CD player to play certain tracks.

As in the pilot study, sets of manuals were prepared, produced according the same principles; the only difference was that this time the manuals were longer– up to 3 pages. There were 21 evaluators – seven for each set of manuals.

The users were encouraged to use the manual as much as possible, but they were free to use their prior knowledge as well. However, because of the complexity of the tasks the users had to use manuals for almost all tasks.

The user performance of each task was timed. At the end of each task, the users were asked to assess the following: "Task difficulty", "How useful was manual to cope with the task" and "Quality of the manual text". The scale was 1 to 5 (1 is an easy task or useless manual or bad text quality; 5 is difficult task, helpful manual or good text quality respectively). Also the users were encouraged to give their informal comments about tasks and used manuals at the end of the experimental session.

Figure 6 shows average time taken to perform tasks given different sets of manuals and standard deviation of the task accomplishment time. As can be seen, the time of task completion varied greatly between different tasks and users.

Task.No	Orig	Gen	Edit
1	3:37(0:58)	3:19(0:47)	2:58(0:31)
2	6:02(1:26)	5:53(1:21)	4:38(1:11)
3	3:59(1:17)	2:43(0:58)	4:20(2:26)
4	4:29(1:44)	3:48(0:50)	2:57(0:37)

Figure 6: Mean(standard deviation) of task accomplishment time

While performing the first task the users often did not resort to the manual, so we discard the results of this test from the further discussion as not being reliable. After normalising experiment time (mean time(deviation) is 1.10(0.22), 0.96(0.29) and 0.98(0.48) for original, generated and edited texts respectively), a t-test has been done to estimate the difference in user performance depending on the text version used.

The results of the t-test indicate that users performed faster on generated versions of the manuals; however the difference in the user performance on different text versions is not significant, only approaching the level of significance between original and generated texts. The results at df=40 are: t=1.731, p=0.0912 between original and generated; t=1.032, p=0.3083; and t=0.152, p=0.880 between edited and generated text versions.

The small dependency of the task accomplishment time on the text version may suggest that users seldom pay much attention to the text detail, preferring to quickly skim through the text looking for the pertinent information or use the diagrams. Any attempts to improve text quality (adding rhetorical markers, etc.) in our case made little difference to the user performance.

The results of user assessment of the text quality/usefulness after performing their tasks are presented in Figure 7. This figure shows average score of text quality and usefulness for completing the task and standard deviation of these scores.

Text quality	
Original	3.45(0.80)
Generated	2.75(1.36)
Generated+edited	3.25(1.62)
Text usefulness	
Original	3.90(0.92)
Generated	3.25(1.18)
Generated+edited	3.90(1.64)

Figure 7: Mean(standard deviation) of text quality/usefulness results

As is the case with the time of task accomplishment, scores varied a lot between different users.

The results of a t-test of text quality at df=40 are: t=1.936, p=0.06 between original and gen-

erated; t=0.481, p=0.633 between original and edited; and t=1.030, p=0.3092 between edited and generated texts.

The results of a t-test of text usefulness at df=40 are: t=1.898, p=0.065 between original and generated; t=0, p=1 between original and edited; and t=1.403, p=0.168 between edited and generated texts.

These results show that on average the users regarded the quality and usefulness of the original texts being equal or slightly better than these of the generated or edited texts. However the difference is not significant.

The comparison of data in Figures 6 and 7 gives rise to an interesting paradox. The users have assessed the quality and usefulness of original manuals slightly higher than of their generated counterparts. Nonetheless, they performed slightly better on the generated versions. This shows that the subjective judgement of text quality may not reliably represent its task-oriented worth. A similar opinion was expressed by the participants of the AGILE evaluation that 'it is difficult to evaluate the acceptability of a technical instruction text *per se*, without real knowledge of the ...system it describes'.

Another discrepancy is that in the pilot study the users preferred generated texts over other versions, whereas in the main experiment original texts were favourites. We do not have other explanation than either users' subjectivity or a low number of participants that played role. The first premise indicates the insufficiency of subjective methods of text quality evaluation; the second one entails us to re-do the experiments with greater number of participants.

However, any differences in user performance and text quality scores described above were insignificant, which means that both subjective and task-specific quality of all manual versions was pretty much equal.

5 Evaluation Results

The results obtained in the course of evaluation of FlexyCAT have confirmed again the advantages of NLG systems and emphasised the aspects which had previously received less attention.

Firstly, FlexyCAT has proved a versatile tool allowing us to create manuals for different devices. The extent of knowledge re-use (both linguistic and domain) is very promising not only across similar devices, but even between fairly distant ones.

Re-using existing knowledge also allows the user to save valuable time in the production of subsequent documents. We have found that the time of manual creation in two languages, when re-using existing linguistic and domain resources, is comparable to that of manual document creation. These results suggest that NLG could become a promising competitor to other ways of multilingual document creation in terms of time and effort saving.

Secondly, the users did not show any significant preference of any text version over others, not did they perform significantly better on a certain version of text. This is a very promising result suggesting that the quality of generated texts was as good as that of manually-crafted ones.

Thirdly, time of task completion varies a great deal between different users, but generally depends little on the manual text quality. An attempt to improve text quality by enriching its rhetorical structure made little difference to the user performance. This also indicates that text *fluency* matters little for accomplishing user task. Sometimes users subjectively assessed a manual as having poor quality, nonetheless finding it useful for completing their task – 'Bad manual is better than no manual'.

Fourthly, a subjective evaluation of text quality may not be a true indication of manual worth with regard to performing a task. Task-oriented evaluation is a more objective way of assessing how helpful the manual is.

The collected data and informal comments have given rise to the following suggestions to improve manual quality, that support those found in (Haydon, 1995):

Firstly, the consistency and structure of a text are very important. All pertinent information should be presented in an overt form and at the place where it is vital.

Secondly, small pieces of text are better than big dense chunks. Each piece should describe a

single function.

6 Conclusions and Further Work

In the course of FlexyCAT evaluation some novel experiments were carried out that have shown that the NLG system may be used to produce texts for different devices and that it is possible to achieve good knowledge re-use, that allows significant saving of time and effort for the production of subsequent manuals.

The text quality experiments have indicated that an NLG system is capable of producing good quality texts. Our task-oriented experiment have shown the advantages of using generated manuals for performing a task and that that subjective methods of assessing text quality are not always adequate for the estimation of text usefulness.

Although FlexyCAT is a bilingual text generation tool, so far it has not been possible to assess the Russian part of it because of the lack of both original manuals in Russian and of Russian-speaking evaluators. All experiments described in this paper are pertinent to the English part of the generation; it was assumed that Russian texts have comparable quality and properties. Further experiments to assess the quality of Russian texts and their consistency with English ones are required.

Also, larger-scale experiments are desirable to evaluate the our conclusions.

This work has indicated the importance of evaluation in NLG, especially task-oriented evaluation and the necessity of broadening of scopes of NLG evaluation.

Acknowledgements

We thank Dr. Diana Bental for her help with statistical processing of experimental data.

References

Srinivas Bangalore, Anoop Sarkar, Christine Doran, and Beth Ann Hockey. 1998. Grammar and Parser Evaluation in the XTAG Project. In *Workshop on The Evaluation of Parsing Systems*, Granada, Spain.

Nathalie Colineau, Ccile Paris, Keith Vander Linden. 2002. An Evaluation of Procedural Instructional Text. In *Proceedings of INLG 2002. pp 128-135.*

1995 EAGLES, Evaluation of Natural Language Processing Systems. FINAL REPORT, EAGLES DOCUMENT EWG-PR. Version of September 1995.

Bruce Eddy and Alison Cawsey. 2002. Balancing Conciseness, Readability and Salience in Generated Text. In *Proceedings of the 3rd International Workshop on Natural Language and Information Systems*, Aix-en-Provence, France.

Antony Hartley, Donia Scott, I. Kruijff-Korbayouva, Serge Sharoff, E. Teich, Lena Sokolova, Kamenka Staykova, Danail Dochev, Martin Cmajrek, and Jiri Hana. 2000. Evaluation of the final prototype. Technical report, Brighton University, October, 12.

Leslie M. Haydon July 1995. The Complete Guide to Writing and Producing Technical Manuals. John Wiley and Sons, Ltd.

John Levine and Chris Mellish. 1995. The IDAS User Trials: Quantitative Evaluation. In *Proceedings of the 5th European Workshop on NLG*, pages 75–93. Rijks Universiteit Leiden.

Chris Mellish and Robert Dale. 1998. Evaluation in the context of natural language generation.

Nestor Miliaev, Alison Cawsey, and Greg Michaelson. 2002. Technical Documentation: An Integrated Architecture for Supporting the Author in Generation and Resource Editing. In *The Tenth International Conference on Artificial Intelligence: Methodology, Systems, Applications AIMSA 2002*, Varna, Bulgaria, September 4-6. Springer-Verlag.

Ehud Reiter and Robert Dale. 2000. *Building Applied Natural Language Generation System.* Natural Language Engineering. Cambridge University Pres.

Ehud Reiter, Roma Robertson, A. Scott Lennox, and Liesl Osman. 2001. Using a Randomised Controlled Clinical Trial to Evaluate an NLG System. In *In Proceedings of ACL-2001*, pages pp. 434–441.

William M. Newman, Michael G. Lamming 1995. Interactive System Design. Addison-Wesley Pub Co. 1st edition.