Evaluating and comparing three text-production techniques

José Coch
GSI-Erli
1, pl. des Marseillais
F-94227 Charenton-le-Pont Cedex
France
jose.coch@erli.fr

Abstract

What are the benefits of using Natural Language Generation in an industrial application? We have attempt to answer part of this question with a description of an assessment of three techniques for producing multisentential text: semi-automatic fill-in-the-blank interfacing, automatic linguistic-and-templates hybrid generation, and human writing. This assessment used a black-box methodology, with an independent blind-tested jury that gave different quality levels in relation to a set of criteria. The texts used for the assessment were business reply letters.

1 Introduction

There are many more industrial projects in Analysis than in Natural Language Generation. Therefore the benefits of using applied NLG would appear a crucial issue. We have provided a partial response to this issue by analysing the assessment of three different techniques for producing multisentential text (in this case, business reply letters).

In the following section, we have described the three techniques under assessment: semi-automatic non-linguistic fill-in-the-blank interfacing, automatic linguistic-and-template hybrid generation, and human writing.

The third section deals with the black-box methodology and quality criteria used for the assessment.

The fourth section describes the results of the assessment.

The fifth section gives examples of letters produced by both the semi-automatic system, and the linguistic-and-template hybrid system.

The last section analyses the results of the assessment.

2 Three techniques for producing multisentential text

This section describes the three text-production techniques under assessment.

2.1 Fill-in-the-blank semi-automatic technique

Since 1975, the mail department of La Redoute (a European mail-order company) has been using a semi-automatic reply system, referred to below as "SA", consisting of a number of predefined and fill-in-the-blank sentences or paragraphs which are identified by codes that the writers memorise. Writing a letter therefore involves typing the code that corresponds to the desired paragraph and inserting the relevant elements. The sentences or paragraphs thus produced are therefore concatenations of predefined and inserted texts.

1. A relatively high number of predefined sentences and paragraphs have to be provided, to cover the writers' needs, but:
2. In fact, writers use only a reduced set of predefined paragraphs, the number of which depending on the writer.
3. The quality of the resulting style of reply varies widely.

2.2 Automatic Hybrid Generation (Linguistic + Template approach)

La Redoute and GSI-Erli have developed a real-situation pilot system (for details on this project, see (Coch, David, and Magnoler, 1995)) which builds up a text (i.e. a letter) from data entered by the human operator who processes the request; a customer database; and knowledge bases. It uses GSI-Erli's AlethGen text generation toolbox (see (Coch, 1996)). The overall system is composed of two main modules: the Decision module and the Generation module.

The Decision module has the following functions:

- it allows the writer (who reads the request letter) to identify the author and subject of the request letter;
- it asks the writer for relevant information;
- it suggests a decision (for example, order cancellation, renewal, etc.), after consulting the customer database and the domain knowledge;
- it asks the writer to validate the decision (or make a different choice);
- it communicates the relevant information to the Generation module.

The Generation module automatically produces the reply letter in a standard format (SGML). This module consists of several submodules (for more details, see (Coch, David and Magnoler 1995) and (Coch and David, 1994)): the direct generator; the text deep-structure planner (or conceptual planner); the text surface-structure planner (or rhetorical planner); and linguistic realisation, inspired by the Meaning-Text Theory.

The direct generator has two functions:

1. planning the text in direct mode (top-down), and
2. generating more or less fixed expressions or non-linguistic texts (i.e. tables, addresses, lists, etc.).

The direct generator could be used without the other submodules to generate texts in an automatic but non-linguistic way (manipulation of character strings). Reiter (Reiter, 1995) calls this technique "the template approach".

The output of the conceptual planner is the text's deep structure, in which the events to be carried out are not yet in a definitive order. The conceptual planner uses logical, causality, and time rules (see (Coch and David, 1994)).

The rhetorical module chooses concrete operators, modalities and surface order, according to rhetorical rules. The choices made depend on certain attributes, e.g. whether the addressee is aware of an event, whether an event is in the addressee's favour, and so on.

Lastly, the linguistic generation submodule realises each event from the text surface structure. It uses anaphora (see (Coch, David and Wonsever, 1994)), semantic, deep-syntactic, surface-syntactic, and morphological rules. This sub-module is inspired mainly by the Meaning-Text Theory (as developed for example in (Mel'čuk, 1988) and (Mel'čuk and Polguère 1988)).

In accordance with Reiter (Reiter, 1995), La Redoute and GSI-Erli's system can be defined as "hybrid", because it uses both linguistic and template techniques.

2.3 Human writing

The third technique used was human writing in "ideal" conditions: one of La Redoute's best writers wrote the letters with no time constraints.

2.4 Functional differences

It is to be noted that the three techniques described differ from an external functional point of view:

- in the semi-automatic approach, the writer compose the letter themselves, even if assisted by a set of predefined-paragraph codes;
- in the automatic hybrid approach, the operator enters data on the addressee and letter, but does not have to compose the reply letter;
- in the third case, the writer has to write the letter.

Reiter (Reiter, 1995) studied the difference between the linguistic generation and template approaches. The two techniques do not differ from an external functional point of view.

3. Methodology

3.1 Evaluation Tests

Black-box methodology was used for the assessment, which was carried out by an independent jury of 14 people, who were representative of end users, in a blind-test context. The jury was not informed of the automatic generation project.

Each member of the jury examined the quality of a set of 60 letters (20 produced by the SA system, 20 by the automatic hybrid system, and 20 human-written, for identical cases). No member of the jury knew which technique had been used for producing each of the letters.

Each member of the jury wrote a report on each letter, with assessment values according to quality criteria. Examples of these criteria are:

- correct spelling,
- good grammar,
- comprehensiveness,
- rhythm and flow,
- appropriateness of the tone,
- proximity, personalisation,
- absence of repetition,
- correct choice and precision of the terminology used.

The first three criteria were considered as eliminatory, and were marked 0 or 1. The other criteria were marked out of 20.

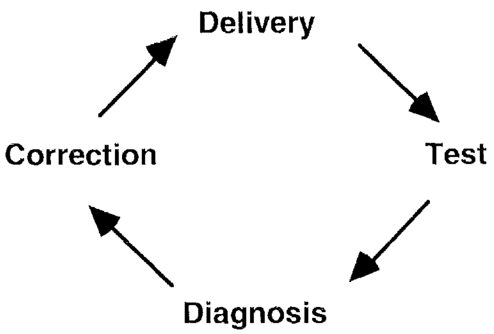
There were also other criteria, but they were too application-oriented and confidential.

3.2 Representativity of the results

Given that the tests used only 20 letters of each type, one might question their representativity.

In fact, representativity is ensured by the projection of the results of the previous phase (system tests) which used the same quality criteria, involved a reduced Jury (2 to 6 members), and was based on 200 test cases (200 letters of each type).

The test cycle was performed six times:



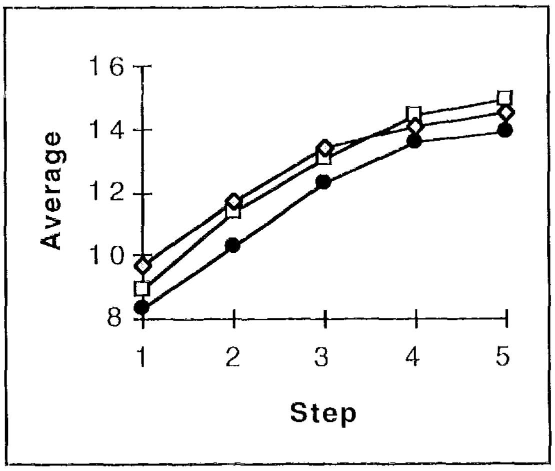
After the sixth cycle, the average quality scores showed that the results would be sufficiently representative.

For example, for the following criteria:

● rhythm and flow

❑ precision of terminology

◊ absence of repetitions



We can thus conclude that, for the automatic letters, the results are representative.

The semi-automatic letters were produced by human "writers" in a real situation. There is no proof of this, but several people who know the semi-automatic system were of the opinion that the semi-automatic letters used in the test were better than the average semi-automatic letter.

4. Assessment results

4.1 Eliminatory criteria and overall average

All the automatic and human letters met the eliminatory criteria standards. However, this was not the case for the semi-automatic system, in particular due to problems of comprehension, but also due to grammatical mistakes in the fill-in-the-blank system.

The overall averages of the entire jury, for all the quality criteria (including application-oriented criteria), and for all the letters were as follows.

• semi-automatic system:	11 out of 20
• automatic hybrid system:	14.5 out of 20
• human-written letters:	15.5 out of 20.

It can be seen that the quality of the letters generated by the pilot system using AlethGen was far superior to that of the semi-automatic system using predefined paragraphs.

These tests show that the "Ideal" human-written letters are, obviously, the best. However, the differences between the human-written letters and those produced by the automatic hybrid system are relatively slight.

4.2 Detailed results

Below are the averages for the whole jury and all the letters, as regards the non-eliminatory criteria:

4.2.1 Rhythm and flow

• semi-automatic system:	12.8 out of 20
• automatic hybrid system:	14 out of 20
• human-written letters:	16.8 out of 20

Differences:

• ideal human letters	vs.	automatic letters:	2.8
• automatic letters	vs.	SA letters:	1.2
• ideal human letters	vs.	SA letters:	4

The difference between the ideal human letters and those obtained with the automatic hybrid system is considerable: 2.8 out of 20.

4.2.2 Right tone

• semi-automatic system:	11.6 out of 20
• automatic hybrid system:	13.6 out of 20
• human-written letters:	14.4 out of 20

Differences:

• ideal human letters	vs.	automatic letters:	0.8
• automatic letters	vs.	SA letters:	2

• ideal human letters	vs.	SA letters:	2.8

The results obtained by the ideal human letters and those generated automatically are close. However, the difference between automatic and semi-automatic letters is considerable: 2 out of 20.

4.2.3 Proximity, personalisation

• semi-automatic system	12 out of 20
• automatic hybrid system	15.2 out of 20
• human-written letters	17.6 out of 20

Differences:

• ideal human letters	vs.	automatic letters:	2.4
• automatic letters	vs.	SA letters:	3.2
• ideal human letters	vs.	SA letters:	5.6

Here, all the differences are considerable. The human letters are obviously the best, but the difference between the automatic and semi-automatic letters is very great: 3.2 out of 20.

4.2.4 Absence of repetition

• semi-automatic system	11.2 out of 20
• automatic hybrid system	14.8 out of 20
• human written-letters	17.6 out of 20

Differences:

• ideal human letters	vs.	automatic letters:	2.8
• automatic letters	vs.	SA letters:	3.6
• ideal human letters	vs.	SA letters:	6.4

For this last point, all the differences are considerable, but that between the automatic and semi-automatic letters is very great: 3.6 out of 20.

4.2.5 Correct choice of terminology

• semi-automatic system	11.6 out of 20
• automatic hybrid system	14 out of 20
• human written-letters	16 out of 20

Differences:

• ideal human letters	vs.	automatic letters:	2
• automatic letters	vs.	SA letters:	2.4
• ideal human letters	vs.	SA letters:	4.4

Here, all differences are relatively great. That between the automatic and semi-automatic letters is considerable: 2.4 out of 20.

5. Examples

Below are several examples of letters produced using the semi-automatic fill-in-the-blanks system and the automatic linguistic-and-template hybrid system.

5.1 Semi-automatic letter

Chère Madame,

J'ai bien reçu votre courrier du 3 Otobre [sic] et je comprends tout à fait votre mécontentement.

Nous faisons le maximum pour contenter nos clients, mais nous sommes dépendants des délais de livraison que nous imposent certains fournisseurs.

Je suis désolée de ne pouvoir vous donner une date précise de livraison, croyez bien que je regrette vivement ce retard.

Restans à votre entière disposition, je vous prie de croire, Chère Madame, à l'expression de mes sentiments dévoués.

[Dear Madam,

In reply to your letter of 3rd Otober [sic], I can completely understand your dissatisfaction.

We do our utmost to satisfy our customers, but are dependent on the delivery times imposed on us by certain suppliers.

I am afraid that I cannot give you an exact delivery date, and sincerely apologise for this delay.

I remain at your entire disposal should you require any further assistance.

Yours sincerely,]

5.2 Linguistic and template example

Chère Madame,

Je suis désolée que vous n'ayez pas reçu les chaussures de sport blanches.

Comme vous en avez été informée lors de l'enregistrement de votre commande, elles n'étaient pas disponibles. La livraison était différée de deux semaines.

Ce délai sera un peu plus long que prévu.

Dès la rentrée en stock de ces chaussures de sport, je vous les enverrai immédiatement, en priorité.

J'espère que vous nous pardonnerez cette attente et que vous voudrez bien patienter.

Je vous prie d'agréer, Chère Madame, l'expression de mon entier dévouement.

[Dear Madam,

> *I am very sorry that you have not received the white sports shoes.*
>
> *These items were not available when your order was recorded, as you were informed at the time. The delivery was postponed by two weeks.*
>
> *The delivery will in fact take a little longer than planned.*
>
> *As soon as these sports shoes are in stock I will send them to you in priority.*
>
> *I hope that you will forgive us for this delay, and are prepared to wait for your delivery.*
>
> *Yours sincerely,]*

5.3 Comments

a) Spelling error in the semi-automatic letter due to the date written by the operator in a blank of a predefined sentence

b) Personalisation: the article and its colour are mentioned only in the automatic letter

c) Precision of terminology (precision of the explanation): clearly, the automatic letter is much more precise

5.4 Semi-automatic example

The following example shows the typical problem of <u>repetition</u> in the semi-automatic letters.

Cher Monsieur,

> J'ai bien reçu votre lettre qui a retenu toute mon attention.
>
> Je réponds à votre demande concernant la <u>marchandise</u> différée suivante : cardigan 4566654 taille 114.
>
> La <u>marchandise</u> a été enregistrée sous le no 176 788956.
>
> Un envoi a été fait le 23 juin.
>
> Normalement, vous devriez déjà avoir reçu la livraison de ce paquet, veuillez m'adresser de préférence un chèque pour régler la <u>marchandise</u> que nous vous avons envoyée.
>
> Restant à votre entière disposition, je vous prie de croire, Cher Monsieur, en mes sentiments dévoués.

[Dear Sir,

> *I have received your letter, which I have read with great attention.*
>
> *I am writing in reply to your request concerning the following postponed <u>merchandise</u>: cardigan 4566654 size 114.*
>
> *The <u>merchandise</u> was recorded with the number 176 788956.*
>
> *«Sending occured» on June 23rd.*
>
> *You should already have received this parcel, therefore would you please send me a cheque in payment of the <u>merchandise</u> that we have sent to you.*
>
> *I remain at your entire disposal.*
>
> *Yours sincerely,]*

6. Analysis of results and Conclusion

6.1 Analysis of results

The order of results for the different techniques is always the same for all the criteria: first, human writing; second, the automatic hybrid approach; and third, the semi-automatic system. Let us now examine the salient points of each type of technique.

Semi-automatic system

The principal weak points of the semi-automatic system are as follows, in decreasing order of variation in relation to the human averages.

- Eliminatory criteria not always met due to problems of comprehension and grammar.
- Excessive repetition (a difference of 6.4 out of 20 in relation to human writing, and of 3.6 in relation to the automatic system).
- Lack of personalisation (5.6 and 3.2).
- Lack of precision in the choice of vocabulary (4.4 and 2.4).

Automatic hybrid system

The principal strong points of the automatic linguistic-and-templates system based on AlethGen are as follows, in decreasing order of variation in relation to the semi-automatic averages.

- Eliminatory criteria always met.
- Absence of repetition (3.6 out of 20 better than the semi-automatic system).
- Proximity, personalisation (3.2 better than the semi-automatic system).
- Precision in the choice of vocabulary (2.4 better).

The main points for improvement for the automatic system are as follows, in decreasing order of variation in relation to the human averages.

- Absence of repetition (human letters 2.8 out of 20 better).
- Rhythm and flow (human letters 2.8 better).
- Proximity, personalisation (human letters 2.4 better).

Human writing

The best characteristics of the human letters were **absence of repetition**, and **proximity / personalisation**, which were both given scores of 17.6 out of 20.

It can be seen that the jury considers the tone of the human letters as being not very good: only 14.4 out of 20. This would appear to be mainly for reasons related to commercial communication rather than computational linguistics.

6.2 Conclusion

The first conclusion is that semi-automatic systems (just as real-situation human writing) are subject to human mistakes, and that the texts they produce may be difficult to understand.

The second conclusion is that the weak points of the semi-automatic systems are the strong points of the automatic hybrid systems, **in the same order**.

We can conclude that, even if current automatic generation systems could do better (and we believe that this will soon be the case), one of the two main reasons for using linguistic-and-template hybrid systems such as that developed by La Redoute and GSI-Erli, rather than using semi-automatic systems, is the improvement in quality (the other being, of course, productivity).

Although there are more research and industrial projects in Analysis than in Natural Language Generation, Generation has great potential, since the gains in terms of **quality and productivity** largely justify the investment.

References

José Coch and Raphaël David. 1994. Representing knowledge for planning multisentential text. *Proceedings of the 4th Conference on Applied Natural Language Processing*, Stuttgart, Germany.

José Coch, Raphaël David, and Dina Wonsever. 1994. Plans, rhetoric and anaphora in a text generation tool. *Working papers of the IBM Institute for Logic and Linguistics. Special Issue on Focus and Natural Language Processing*, IBM Deutschland Informationssysteme GmbH, Scientific Centre, Heidelberg, Germany.

José Coch, Raphaël David, and Jeannine Magnoler. 1995. Quality test for a mail generation system. *Proceedings of Linguistic Engineering 95*, Montpellier, France.

José Coch. 1996. Overview of AlethGen. *Proceedings of the International Workshop on Natural Language Generation (INLG-96).* Herstmonceux, England, 1996.

Igor Mel'čuk. 1988. Dependency Syntax: Theory and Practice. *State University of New York Press*, Albany, NY, USA.

Igor Mel'čuk and Alain Polguère. 1987. A Formal Lexicon in the Meaning-Text Theory (or How to Do Lexica with Words). *Computational Linguistics*, 13(3-4):276–289.

Ehud Reiter. 1994. Has a consensus NL Generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 163–170.

Ehud Reiter. 1995. NLG vs. Templates. In *Proceedings of the 1995 European NL Generation Workshop*, Holland.