Evaluating an NLG System using Post-Editing

Somayajulu G. Sripada and Ehud Reiter and Lezan Hawizy
Department of Computing Science
University of Aberdeen
Aberdeen, AB24 3UE, UK
{ssripada,ereiter,lhawizy}@csd.abdn.ac.uk

Abstract

Computer-generated texts, whether from Natural Language Generation (NLG) or Machine Translation (MT) systems, are often post-edited by humans before being released to users. The frequency and type of post-edits is a measure of how well the system works, and can be used for evaluation. We describe how we have used post-edit data to evaluate **SUMTIME-MOUSAM**, an NLG system that produces weather forecasts.

1 Introduction

In this paper we describe an evaluation technique, which looks at how much humans need to post-edit texts generated by an NLG system before they are released to users. Post-edit evaluations are common in machine translation [Hutchins and Somers, 1992], but we believe that ours is the first large-scale post-edit evaluation of an NLG system. Mitkov and An Ha [2003] reported a small scale post-edit evaluation of their NLG system.

The system being evaluated is **SUMTIME-MOUSAM** [Sripada et al, 2003], an NLG system, which generates weather forecasts from Numerical Weather Prediction (NWP) data. The forecasts are marine forecasts for offshore oilrigs. **SUMTIME-MOUSAM** is operational and is used by Weathernews (UK) Ltd to generate 150 draft forecasts per day, which are post-edited by Weathernews forecasters before being released to clients.

Time	Wind Dir	Wind Spd 10m	Wind Spd 50m	Gust 10m	Gust 50m
06:00	W	10.0	12.0	12.0	16.0
09:00	W	11.0	14.0	14.0	17.0
12:00	WSW	10.0	12.0	12.0	16.0
15:00	SW	7.0	9.0	9.0	11.0
18:00	SSW	8.0	10.0	10.0	12.0
21:00	S	9.0	11.0	11.0	14.0
00:00	S	12.0	15.0	15.0	19.0

Table 1. Weather Data produced by an NWP model for 12-Jun 2002

Table 1 shows a small extract from the NWP data for 12-06-2002, and Table 2 shows part of the textual forecast that SumTime-Mousam generates from the NWP data.

Field	**Text**
WIND(KTS) 10M	W 8-13 backing SW by mid afternoon and S 10-15 by midnight.
WIND(KTS) 50M	W 10-15 backing SW by mid afternoon and S 13-18 by midnight.

Table 2. Extract from **SUMTIME-MOUSAM** Forecast Produced from NWP data in Table 1.

Weathernews uses SumTime-Mousam to generate draft forecasts; we call them 'Pre-edit Texts'. The forecasters then post-edit them to produce 'Post-edit texts'. When the forecaster is done, the complete forecast is sent to the customer.

2 Post-Edit Evaluation

The evaluation was carried out on 2728 forecasts, collected during period June to August 2003. Each forecast was roughly of 400 words, so there are about one million words in all in the corpus.

For each forecast, we have the following data:

- Data: The final edited NWP data
- Pre-edit text: The draft forecast produced by **SUMTIME-MOUSAM**.
- Post-edit text: The manually post-edited forecast, which was sent to the client.
- Background information: includes date, location, and forecaster

The following procedure is performed automatically by a software tool:

- First, we break sentences up into phrases, where each phrase describes the weather at one point in time.
- The second step is to align phrases from the previous step as a preparation for comparison in the next step. Our alignment procedure first generates an exhaustive

list of possible alignments and uses a scoring scheme to select aligned phrases.

- The third step is to compare aligned phrases and label each pre-edit/post-edit pair as match, replace, add, or delete.

For example, A and B of Figure 1 are analyzed as in Table 3 where phrases are shown separated by a shaded row.

A. Pre-edit Text: *SW 20-25 backing SSW 28-33 by midday, then gradually increasing 34-39 by midnight.*
B. Post-edit Text: *SW 22-27 gradually increasing SSW 34-39.*

Figure 1. Example pre-edit and post-edit texts from the post-edit corpus

POS	**A**	**B**	**label**
Direction	SW	SW	match
Speed	20-25	22-27	replace
Conjunction	then	<none>	delete
Adverb	gradually	gradually	match
Verb	increasing	increasing	match
Direction	<none>	SSW	add
Speed	34-39	34-39	match
Time	by midnight	<none>	delete

Table 3. Detailed Edit Analysis

We processed 2728 forecast pairs (pre-edited and post-edited). These were divided into 73041 phrases. Out of these, the alignment procedure failed to align 7608 (10%) phrases. Out of the successfully aligned phrases, 43914 (60%) are perfect matches, and the remaining 21519 (30%) are mismatches. Table 4 summarizes the mismatches and suggests that the major problem is ellipsis. Most (25235 out of 35874, 70%) of these errors are deletions, where the forecaster deletes words SumTime-Mousam's texts.

S. No.	Mismatch Type	Freq.	%
1.	Ellipses (word additions and deletions)	35874	65
2.	Data Related Replacements (range and direction replacements)	10781	20
3.	Lexical Replacements	8264	15
	Total	54919	

Table 4. Results of the evaluation showing summary categories and their frequencies

3 Discussion of Post-Edit Evaluations

We were attracted to post-edit evaluation because we believed that (A) people would only edit things that were clearly wrong; and (B) post-editing was an important usefulness metric from the perspective of our users.
Looking back, (B) was certainly true. The amount of post-editing that generated texts require is a crucial component of the cost of using **SUMTIME-MOUSAM**, and hence of the attractiveness of the system to users.

(A) however was perhaps less true than we had hoped. Wagner [1998] also described post-edited texts in MT as at times noisy. During the development of **SUMTIME-MOUSAM**, our analysis of manually written forecasts [Reiter and Sripada, 2002] had highlighted a number of "noise" elements that made it more difficult to extract information from such corpora. While collecting the post-edit data, we assumed that people would only post-edit mistakes, where the generated text was wrong or sub-optimal, and hence post-edit data would be better for evaluation purposes than corpus comparisons.

In fact, however, there were many justifications for post-edits. Some post-edits fixed problems in the generated texts (such as overuse of *then*); some post-edits refined/optimized the texts (such as using *for a time*); some post-edits reflected individual preferences (such as *easing* vs *decreasing*); and some post-edits were downstream consequences of earlier changes (such as introducing *SSW* before '*34-39*' in B, in the example of Section 2). We wanted to use our post-edit data to improve the system, not just to quantify its performance, and we discovered that we could not do this without attempting to analyze why post-edits were made. Probably the best way of doing this was to discuss post-edits with the forecasters.

4 Conclusion

We have used analysis of post-edits, a popular evaluation technique in machine translation, to evaluate SumTime-Mousam, an NLG system that generates weather forecasts. While we encountered some problems, such as the need to identify why post-edits were made, on the whole we found post-editing to be a useful evaluation technique which gave us valuable insights as to how to improve our system.

References

[Hutchins and Somers, 1992] John Hutchins and Harold L. Somers. *An Introduction to Machine Translation*, Academic Press, London, 1992.

[Mitkov and An Ha, 2003] Ruslan Mitkov and Le An Ha. Computer-Aided Generation of Multiple-Choice Tests, In *Proc. of the HLT-NAACL03 Workshop on Building Educational Applications Using NLP*, Edmonton, Canada, pages 17-22, 2003.

[Reiter and Sripada, 2002] Ehud Reiter and Somayajulu G. Sripada. Human Variation and Lexical Choice. *Computational Linguistics*, 28:545-553, 2002.

[Sripada et al., 2003] Somayajulu G. Sripada, Ehud Reiter, and Ian Davy. SUMTIME-MOUSAM: Configurable Marine Weather Forecast Generator. *Expert Update*, 6(3):4-10, 2003.

[Wagner, 1998] Simone Wagner. Small Scale Evaluation Methods In *Proc. of the Workshop on Evaluation of the Linguistic Performance of Machine Translation Systems at the KONVENS-98*, Bonn, pages 93-105, 1998.