Automated Metrics That Agree With Human Judgements On Generated Output for an Embodied Conversational Agent

Mary Ellen Foster
Informatik VI: Robotics and Embedded Systems
Technische Universität München
Boltzmannstraße 3, D-85748 Garching bei München, Germany
foster@in.tum.de

Abstract

When evaluating a generation system, if a corpus of target outputs is available, a common and simple strategy is to compare the system output against the corpus contents. However, cross-validation metrics that test whether the system makes exactly the same choices as the corpus on each item have recently been shown not to correlate well with human judgements of quality. An alternative evaluation strategy is to compute intrinsic, task-specific properties of the generated output; this requires more domain-specific metrics, but can often produce a better assessment of the output. In this paper, a range of metrics using both of these techniques are used to evaluate three methods for selecting the facial displays of an embodied conversational agent, and the predictions of the metrics are compared with human judgements of the same generated output. The corpus-reproduction metrics show no relationship with the human judgements, while the intrinsic metrics that capture the number and variety of facial displays show a significant correlation with the preferences of the human users.

1 Introduction

Evaluating the output of a generation system is known to be difficult: since generation is an open-ended task, the criteria for success can be difficult to define (cf. Mellish and Dale, 1998). In the current state of the art, there are two main strategies for evaluating the output of a generation system: the behaviour or preferences of humans in response to the output may be measured, or automated measures may be computed on the output itself. A study involving human judges is the most complete and convincing evaluation of generated output. However, such a study is not always practical, as recruiting sufficient subjects can be time-consuming and expensive. So automated metrics are also used in addition to—or instead of—human studies.

When automatically evaluating generated output, the goal is to find metrics that can easily be computed and that can also be shown to correlate with human judgements of quality. Such metrics have been introduced in other fields, including PARADISE (Walker et al., 1997) for spoken dialogue systems, BLEU (Papineni et al., 2002) for machine translation,[1] and ROUGE (Lin, 2004) for summarisation. Many automated generation evaluations measure the similarity between the generated output and a corpus of "gold-standard" target outputs, often using measures such as precision and recall. Such measures of corpus similarity are straightforward to compute and easy to interpret; however, they are not always appropriate for generation systems. One of the main advantages of choosing dynamic generation over canned output is its flexibility and its ability to produce a range of different outputs; as pointed out by Paris et al. (2007), "[e]valuation studies that ignore the potential of the system to generate a range of appropriate outputs will be necessarily limited." Indeed, several recent studies (Stent et al., 2005; Belz and Reiter, 2006; Foster and White, 2007) have shown that strict corpus-similarity measures tend to favour repetitive generation strategies that do not diverge much, on average, from the corpus data, while human judges often prefer output with more variety.

[1] Although Callison-Burch et al. (2006) have recently called into question the utility of BLEU.

Automated metrics that take into account other properties than strict corpus similarity have also been used to evaluate the output of generation systems. Walker (2005) describes several evaluations that used corpus data in a different way: each of the corpus examples was associated with some reward function (e.g., subjective user evaluation or task success), and machine-learning techniques such as reinforcement learning or boosting were then used to train the output planner. Foster and White (2007) found that automated metrics based on factors other than corpus similarity (e.g., the amount of variation in the output) agreed better with user preferences than did the corpus-similarity scores. Belz and Gatt (2008) compare the predictions of a range of measures, both intrinsic and extrinsic, that were used to evaluate the systems in a shared-task referring-expression generation challenge. One main finding from this comparison was that there was no significant correlation between the intrinsic and extrinsic (task success) measures for this task.

All of the above studies considered only systems that generate text, but many of the same factors also apply to the generation of non-verbal behaviours for an embodied conversational agent (ECA) (Cassell et al., 2000). The behaviour of such an agent is normally based on recorded human behaviour, which can provide targets similar to those used in corpus-based evaluations of text-generation systems. However, just as in text generation, a multimodal system that scores well on corpus similarity tends to produce highly repetitive non-verbal behaviours, so it is equally important to gather human judgements to accompany any automated evaluation.

This paper presents three corpus-driven methods of selecting facial displays for an embodied conversational agent and describes two studies comparing the output of the different methods. All methods are based on annotated data drawn from a corpus of human facial displays, and each uses the corpus data in a different way. The first evaluation study uses human judges to compare the output of the selection methods against one another, while the second study uses a range of automated metrics: several corpus-reproduction measures, along with metrics based on intrinsic properties of the outputs. The results of the two studies are compared using multiple regression, and the implications are discussed.

2 Corpus-based generation of facial displays for an ECA

The experiments in this paper make use of the output components of the COMIC multimodal dialogue system (Foster et al., 2005), which adds a multimodal talking-head interface to a CAD-style system for redesigning bathrooms. The studies focus on the task of selecting appropriate ECA head and eyebrow motions to accompany the turns in which the system describes and compares the options for tiling the room, as those are the parts of the output with the most interesting and varied content.

The implementations were based on a corpus of conversational facial displays derived from the behaviour of a single speaker reading approximately 450 scripted sentences generated by the COMIC output-generation system. The OpenCCG syntactic derivation trees (White, 2006) for the sentences form the basis of the corpus. The leaf nodes in these trees correspond to the individual words, while the internal nodes correspond to multi-word constituents. Every node in each tree was initially labelled with all of the applicable contextual features produced by the output planner: the user-preference evaluation of the tile design being described (positive/negative/neutral), the information status (given/new) of each piece of information, and the predicted speech-synthesiser prosody. The annotators then linked each facial display produced by the speaker to the node or span of nodes in the derivation tree covering the words temporally associated with the display. Full details of this corpus are given in Foster (2007a).

The most common display used by the speaker was a downward nod, while the user-preference evaluation had the single largest differential effect on the displays used. When the speaker described features of the design that the user was expected to like, he was relatively more likely to turn to the right and to raise his eyebrows (Figure 1(a)); on features that the user was expected to dislike, on the other hand, there was a higher probability of left leaning, lowered eyebrows, and narrowed eyes (Figure 1(b)). In a previous study, users were generally able to recognise these "positive" and "negative" displays when they were resynthesised on an embodied conversational agent (Foster, 2007b).

(a) Positive (b) Negative

Figure 1: Characteristic facial displays from the corpus

Based on this corpus, three different strategies were implemented for selecting facial displays to accompany the synthesised speech: one strategy using only the three characteristic displays described above, along with two data-driven strategies drawing on the full corpus data. All of the strategies use the same basic process to select the displays to accompany a sentence. Beginning with the contextually-annotated syntactic tree for the sentence, the system proceeds depth-first, selecting a face-display combination to accompany each node in turn. The main difference among the strategies is the way that each selects the displays for a node as it is encountered.

The **rule-based** strategy includes displays only on derivation-tree nodes corresponding to specific tile-design properties: that is, manufacturer and series names, colours, and decorative motifs. The displays for such a node are entirely determined by the user-preference evaluation of the property being described, and are based on the corpus patterns described above: for every node associated with a positive evaluation, this strategy selects a right turn and brow raise; for a negative node, it selects a left turn, brow lower, and eye squint; while for all other design-property nodes, it chooses a downward nod.

While the rule-based strategy selects displays only on nodes describing tile-design features, the two data-driven strategies consider all nodes in the syntactic tree for a sentence as possible sites for a facial display. To choose the displays for a given node, the system considers the set of displays that occurred on all nodes in the corpus with the same syntactic, semantic, and pragmatic context, and then chooses a display from this set in one of two ways. The **majority** strategy selects the most common option in all cases, while the **weighted** strategy makes a stochastic choice among all of the options based on the relative frequency. As a concrete example, consider a hypothetical context in which the speaker made no motion 80% of the time, a downward nod 15% of the time, and a downward nod with a brow raise the other 5% of the time. For nodes with this context, the majority strategy would always choose no motion, while the weighted strategy would choose no motion with probability 0.8, a downward nod with probability 0.15, and a nod with a brow raise with probability 0.05.

Table 1 shows a sample sentence from the corpus, the original facial displays used by the speaker, and the displays selected by each of the strategies. In the figure, *nd=d* indicates a downward nod, *bw=u* and *bw=d* a brow raise and lower, respectively, *sq* an eye squint, *ln=l* a left lean, and *tn=r* a right turn. Most of the displays in these schedules are associated with leaf nodes in the derivation tree, and therefore with single words in the output. However, both the left lean in the original schedule and the right turn in the weighted schedule are associated with internal nodes in the tree, and therefore cover more than one word in the surface string.

3 User-preference studies

As a first comparison of the evaluation strategies, human judges were asked to compare videos based on the output of each of the generation strategies to one another and to resynthesised versions of the original displays from the corpus. This section gives the details of a study in which the judges chose

	Although	*it's*	*in*	*the*	*family*	*style,*	*the*	*tiles*	*are*	*by*	*Alessi.*
Original	nd=d	nd=d	nd=d		nd=d				nd=d,bw=u		
	. . . . . ln=l										
Rule-based					ln=l,bw=d						tn=r,bw=u
Majority					nd=d						nd=d
Weighted	nd=d				nd=d		. . tn=r . .				

Table 1: Face-display schedules for a sample sentence

Figure 2: RUTH talking head

among the original displays and the output of the weighted and rule-based strategies. At the end of the section, the results of this study are discussed together with the results of a similar previous study comparing the two data-driven strategies to each other; the full details of the earlier study are given in Foster and Oberlander (2007).

3.1 Subjects

Subjects were recruited for this experiment through the Language Experiments Portal,[2] a website dedicated to online psycholinguistic experiments. There were 36 subjects (20 female), 50% of whom identified themselves as native English speakers; most of the subjects were between 20 and 29 years old.

3.2 Materials

The materials for this experiment were based on 18 randomly-selected sentences from the corpus. For each sentence, face-display schedules were generated using both the rule-based and the weighted strategies. The Festival speech synthesiser (Clark et al., 2004) and the RUTH animated talking head (DeCarlo et al., 2004) (Figure 2) were used to create video clips of the two generated schedules for each sentence, along with a video clip showing the original facial displays annotated in the corpus.

[2] http://www.language-experiments.org/

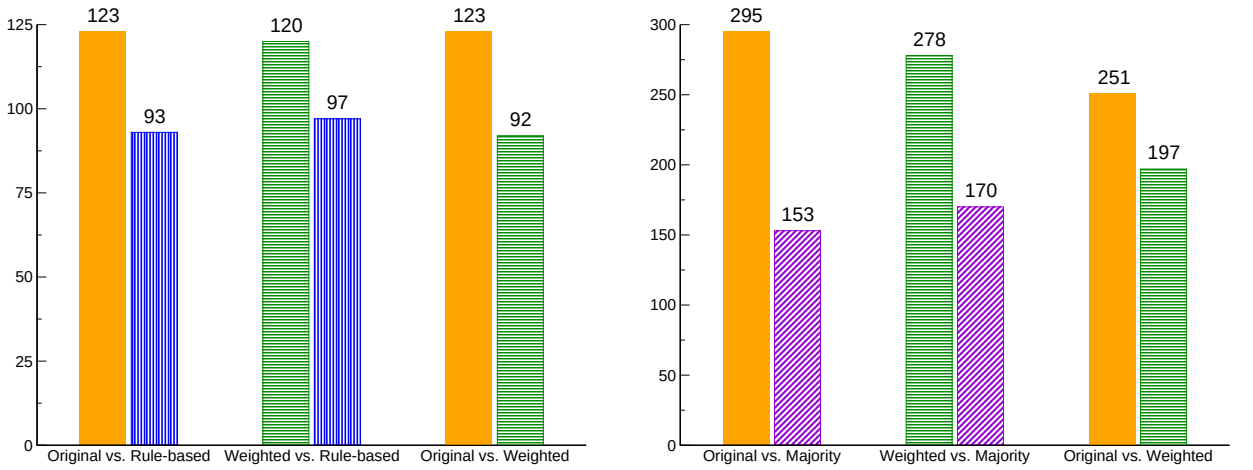
3.3 Method

Each subject saw a series of pairs of videos. Both videos in a pair had identical spoken content, but the face-display schedules differed: each trial included two of rule-based, weighted, and original. For each pair of videos, the subject was asked to select which of the two versions they preferred. Subjects made each pairwise comparison between schedule types six times—three times in each order—for a total of 18 judgements. All subjects saw the same set of sentences, in an individually randomly-selected order: the pairwise choices between schedule types were also allocated to items at random.

3.4 Results and analysis

The overall pairwise preferences of the subjects in this study are shown in Figure 3(a). A χ^2 goodness-of-fit test can be used to evaluate the significance of the choices made on each individual comparison. For the comparison between original and rule-based schedules, the preference is significant: $\chi^2(1, N = 216) = 4.17$, $p < 0.05$. The results are similar for the original vs. weighted comparison: $\chi^2(1, N = 215) = 4.47$, $p < 0.05$. However, the preferences for the weighted vs. rule-based comparison are not significant: $\chi^2(1, N = 217) = 2.44$, $p \approx 0.12$.

Figure 3(b) shows the results from a similar previous study (Foster and Oberlander, 2007) in which the subjects compared the two data-driven strategies to the original displays, using a design identical to that used in the current study with 54 subjects and 24 sentences. The responses given by the subjects in this study also showed a signifi-



(a) Original, weighted, rule-based

(b) Original, weighted, majority (Foster and Oberlander, 2007)

Figure 3: Pairwise preferences from the user evaluations

cant preference for the original schedules over the weighted ones ($\chi^2(1, N = 448) = 6.51$, $p < 0.05$). Both the weighted and the original schedules were very strongly preferred over the majority schedules ($\chi^2(1, N = 448) = 45$ and 26, respectively; $p \ll 0.0001$). The original vs. weighted comparison was included in both studies (the rightmost pair of bars on the two graphs in Figure 3), and the response patterns across the two studies for this comparison did not differ significantly from each other: $\chi^2(1, N = 664) = 0.02$, $p \approx 0.89$.

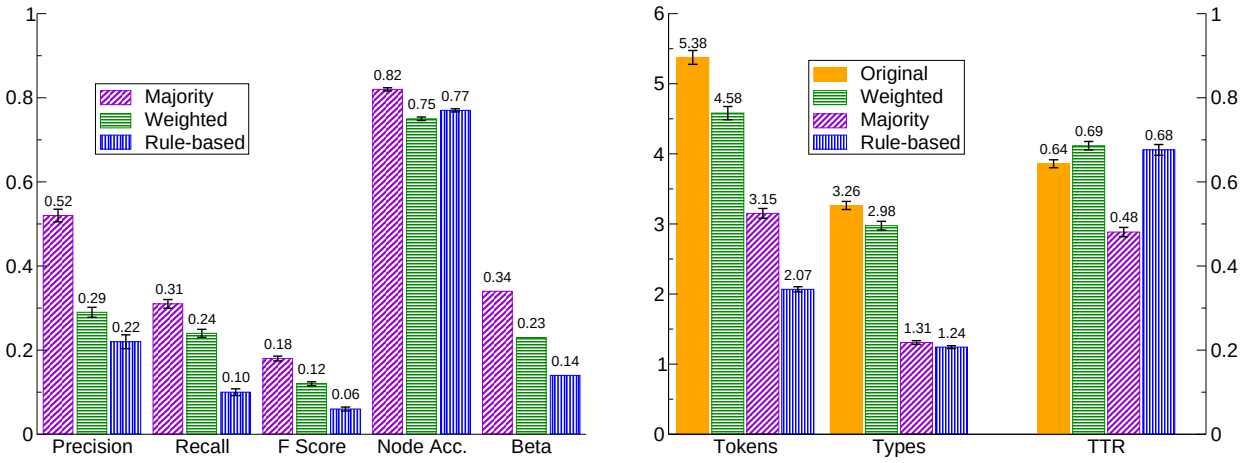
3.5 Discussion

Taken together, the results of these two studies suggest a rough preference ordering among the different strategies for generating facial displays. In both studies, the judges significantly preferred the original displays from the corpus over any of the automatically-generated alternatives. This suggests that, for this generation task, the data in the corpus can indeed be treated as a "gold standard"—unlike, for example, the corpus used by Belz and Reiter (2006), where the human judges sometimes preferred generated output to the corpus data. The schedules generated by the majority strategy, on the other hand, were very obviously disliked by the judges in the Foster and Oberlander (2007) study. The ranking between the rule-based and weighted schedules from the current study is less clear, although there was a tendency to prefer the latter.

4 Automated evaluation

Since the subjects in the user-preference studies generally selected the corpus schedules over any of the alternatives, any automated metric for this task should favour output that resembles the examples in the corpus. The most obvious form of corpus similarity is exact reproduction of the displays in the corpus, which suggests using metrics such as precision and recall that favour generation strategies whose output on every item is as close as possible to what was annotated in the corpus for that sentence. In Section 4.1, several such **corpus-reproduction** metrics are described and their results presented.

For this type of open-ended generation task, though, it can be overly restrictive to allow *only* the displays that were annotated in the corpus for a sentence and to penalise any deviation. Indeed, as mentioned in the introduction, a number of previous studies have found that the output of generation systems that score well on this type of metric is often disliked in practice by users. Section 4.2 therefore presents several **intrinsic** metrics that aim to capture corpus similarity of a different type: rather than requiring the system to exactly reproduce the corpus on each sentence, these metrics instead favour strategies resulting in global behaviour that exhibits similar patterns to those found in the corpus, without necessarily agreeing exactly with the corpus on any specific sentence.



(a) Corpus-reproduction metrics

(b) Intrinsic metrics

Figure 4: Results of the automated evaluations

4.1 Corpus-reproduction metrics

This first set of metrics compared the generated schedules against the original schedules annotated in the corpus, using 10-fold cross-validation. The first three metrics that were tested are standard for this sort of corpus-comparison task: recall, precision, and F score. Recall was computed as the proportion of the corpus displays for a sentence that were reproduced exactly in the generated output, while precision was the proportion of generated displays that had exact matches in the corpus; the F score for a sentence is then the harmonic mean of these two values, as usual. The leftmost three columns in Table 2 show the precision, recall, and F score for the sample schedules in Table 1.

In addition to the above commonly-used metrics, two other corpus-reproduction metrics were also computed. The first, *node accuracy*, represents the proportion of nodes in the derivation tree for a sentence where the proposed displays were correct, including those nodes where the system correctly selected no motion—a baseline system that never proposes any motion scores 0.79 on this measure. The fourth column of Table 2 shows the node-accuracy score for the sample sentences. The final corpus-reproduction metric compared the proposed displays to the annotated corpus displays using the β agreement measure (Artstein and Poesio, 2005). β is a weighted measure that permits different levels of agreement when annotations overlap, and that can therefore capture a more fine-grained form of agreement than other measures such as κ.

	P	R	F	NAcc	Tok	Typ	TTR
Original	–	–	–	–	6	3	0.5
Rule-based	0	0	0	0.65	2	2	1
Majority	0.50	0.14	0.11	0.70	2	1	0.5
Weighted	0.67	0.29	0.20	0.74	3	2	0.67

Table 2: Automated evaluation of the sample schedules

Figure 4(a) shows the results for all of these corpus-reproduction measures, averaged across the sentences in the corpus; the results for the weighted and majority strategies are from Foster and Oberlander (2007). The majority strategy scored uniformly higher than the weighted strategy on all of these measures—particularly on precision—while the weighted strategy in turn scored higher than the rule-based strategy on all measures except for node accuracy. Using a Wilcoxon rank sum test with a Bonferroni correction for multiple comparisons, all of the differences among the strategies on precision, recall, F score, and node accuracy are significant at $p < 0.001$. Significance cannot be assessed for the differences in β scores, as noted by Artstein and Poesio (2005), but the results are similar. Also, the node accuracy score for the majority strategy is significantly better than the no-motion baseline of 0.79, while those for the weighted and rule-based strategies are worse (also all $p < 0.001$).

As expected—and as noted by Foster and Oberlander (2007)—all of the corpus-reproduction metrics strongly favoured the weighted strategy over the weighted strategy and generally penalised the rule-based strategy. Since the majority strategy always chooses the most probable option, it is not surprising that it agrees more often with the corpus than do the other strategies, which deliberately select less frequent options; this led to its relatively high scores on the corpus-reproduction metrics. It is also not surprising that the weighted strategy beat the rule-based strategy on most of these metrics, as the former selects from the most frequent options, while the latter uses the most *marked* options, which are not generally the most frequent.

4.2 Intrinsic metrics

The metrics in the preceding section compared the displays selected for a sentence against the displays found in the corpus for that sentence. This section describes other measures that are computed directly on the generated schedules, without any reference to the corpus data. For each sentence, the following values were counted: the total number of face-display combinations (i.e., the number of display *tokens*), and the number of different combinations (*types*). In addition to being used as metrics themselves, these two counts were also used to compute a third value: the *type/token ratio* (TTR) (i.e., $\frac{\#\ types}{\#\ tokens}$), which captures the diversity of the displays selected for each sentence.

These intrinsic metrics were computed on each sentence produced in the cross-validation study from the preceding section and then averaged to produce the final results. Since these metrics do not require the original corpus data for comparison, they were also computed on the original corpus schedules. The rightmost columns in Table 2 show the intrinsic results for the sample schedules in Table 1.

The overall results for these metrics across the entire corpus are shown in Figure 4(b). The original corpus had both the most displays types and the most tokens; the values for weighted choice were a fairly close second, those for majority choice third, while the rule-based strategy scored lowest on both of these metrics. Except for the difference between majority and rule-based on the facial-display types—which is not significant—all of the differences between schedule types on these two measures are significant at $p < 0.001$ on a Wilcoxon rank sum test with Bonferroni correction. When it comes to the type/token ratio, the value for the majority-choice schedule is significantly lower than that for the other three schedule types (all $p < 0.0001$), while the value for weighted choice is somewhat higher than that for the original schedules ($p < 0.01$); no other differences are significant.

Since the original corpus schedules scored the highest on the user study, these metrics should be considered in the context of of how close the results are to those of the corpus. Figure 4(b) shows that the weighted strategy is most similar to the corpus in both the number and the diversity of displays it selects, while the other two strategies have much lower diversity. However, even though the rule-based strategy selects fewer displays all of the other strategies, its TTR is more similar to that of the corpus and the weighted strategy, while the majority strategy has a much lower TTR. In fact, in the schedules generated by the majority-choice strategy, nearly 90% of the displays that were selected were downward nods.

5 Comparing the automated metrics with human preferences

Qualitatively, the results of the corpus-reproduction metrics differ greatly from the preferences of the human judges. The users generally liked the majority schedules the least, while all of these metrics scored this strategy the highest. Among the intrinsic metrics, the type and token counts placed the weighted schedules closest to the corpus, while the majority and rule-based strategies were further away; this agrees with the human results for the two data-driven strategies, but not for the rule-based strategy. On the other hand, the TTR indicated that the output of the rule-based and weighted schedules was similar to the schedules found in the corpus, while the majority-choice strategy produced sentences with TTRs more different from the corpus, generally agreeing with the human results.

To permit a more quantitative comparison between the predictions of the automated metrics and the judges' preferences, the pairwise preferences from the user study were converted into a numeric value called the *selection ratio*. The selection ratio

for an item (i.e., a sentence with a particular set of facial displays) was computed as the number of trials on which that item was selected, divided by the total number of trials on which that item was an option. For example, an item that was always preferred over any of the alternatives on all trials would score 1.0 on this measure, while an item that was selected a quarter of the time would score 0.25. The selection ratios of the items used in the human-preference studies ranged from 0.13 to 0.85. As a concrete example, when the sentence in Table 1 was used in the Foster and Oberlander (2007) study, the selection ratios were 0.43 for the original version, 0.33 for the majority version, and 0.24 for the weighted version.

The relationship between the selection ratio and the full set of automated metrics from the preceding section was assessed through multiple linear regression. An initial model including all of the automated metrics as predictor variables had an adjusted R^2 value of 0.413. Performing stepwise selection on this initial model resulted in a final model with two significant predictor variables—display tokens and TTR—and an adjusted R^2 of 0.422. The regression coefficients for both of these predictor variables are positive, with high significance ($p < 0.001$). While the R^2 values indicate that neither the initial nor the final model fully explains the selection ratios from the user study, the details of the models themselves are relevant to the overall goal of finding automated metrics that agree with human preferences.

The results of the stepwise selection have backed up the qualitative intuition that none of the corpus-reproduction metrics had any relationship to the users' preferences, while the number and diversity of displays per sentence appear to have contributed much more strongly to the choices made by the human judges. This adds to the growing body of evidence that intrinsic measures are the preferred option for evaluating the output of generation systems, particularly those that are designed to incorporate variation into their output, while measures based on strict corpus similarity are less likely to be useful.

6 Conclusions

This paper has presented three methods for using corpus data to select facial displays for an embodied agent and shown the results from two studies comparing the output generated by these methods. When human judges rated the output, they preferred the original displays from the corpus and strongly disliked the displays selected by a majority-choice strategy, with the weighted and rule-based strategies in between. In the automated evaluation, the metrics that directly compared the generated output against the corpus data favoured the majority strategy and did not show any relationship with the user preferences. On the other hand, the number of displays accompanying a sentence and the diversity of those displays both had a positive relationship with the rate at which users selected that display schedule.

These results confirm those of previous text-generation evaluations and extend these results to the multimodal-generation case. This adds to the body of evidence that, even though direct corpus reproduction is often the easiest factor to analyse automatically, it is rarely an accurate reflection of user reactions to generated output. If a system performs well on this type of metric, its output tends to be constrained to a small space; for example, the majority-choice strategy used in these studies nearly always selected a nodding expression. For most generation tasks, output options beyond those in the corpus are often equally valid, and users seem to prefer a system that makes use of this wider space of variations.

This suggests that corpus-based generation systems should use strategies that retain the full range of variation, and—perhaps most importantly—that metrics based on factors other than strict similarity are more likely to capture human preferences when evaluating generated output.

The user study described here was based only on the preferences of human judges. In future, it would be informative to include more task-based measures such as task success and time taken, as user preferences do not always correlate with performance (Nielsen and Levy, 1994), to see if a different style of automated measure agrees better with the results of this sort of user study.

Acknowledgements

This work was partly supported by the EU projects COMIC (IST-2001-32311) and JAST (FP6-003747-IP). Thanks to Jon Oberlander, Jean Carletta, and the INLG reviewers for helpful feedback.

References

R. Artstein and M. Poesio. 2005. Kappa^3 = alpha (or beta). Technical Report CSM-437, University of Essex Department of Computer Science. http://cswww.essex.ac.uk/technical-reports/2005/csm-437.pdf.

A. Belz and A. Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL/HLT 2008*.

A. Belz and E. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of EACL 2006*. acl:E06-1040.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL 2006*. acl:E06-1032.

J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. 2000. *Embodied Conversational Agents*. MIT Press.

R. A. J. Clark, K. Richmond, and S. King. 2004. Festival 2 – build your own general purpose unit selection speech synthesiser. In *Proceedings of the 5th ISCA Workshop on Speech Synthesis*. http://www.ssw5.org/papers/1047.pdf.

R. Dale and M. White, editors. 2007. *Report from the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*. http://ling.ohio-state.edu/nlgeval07/NLGEval07-Report.pdf.

D. DeCarlo, M. Stone, C. Revilla, and J. Venditti. 2004. Specifying and animating facial signals for discourse in embodied conversational agents. *Computer Animation and Virtual Worlds*, 15(1):27–38. doi:10.1002/cav.5.

M. E. Foster. 2007a. Associating facial displays with syntactic constituents for generation. In *Proceedings of the ACL 2007 Linguistic Annotation Workshop*. acl:W07-1504.

M. E. Foster. 2007b. Generating embodied descriptions tailored to user preferences. In *Proceedings of Intelligent Virtual Agents 2007*. doi:10.1007/978-3-540-74997-4_24.

M. E. Foster and J. Oberlander. 2007. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, 41(3–4):305–323. doi:10.1007/s10579-007-9055-3.

M. E. Foster and M. White. 2007. Avoiding repetition in generated text. In *Proceedings of ENLG 2007*. acl:W07-2305.

M. E. Foster, M. White, A. Setzer, and R. Catizone. 2005. Multimodal generation in the COMIC dialogue system. In *Proceedings of the ACL 2005 Demo Session*. acl:P05-3012.

C. Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL 2004 Workshop on Text Summarization*. acl:W04-1013.

C. Mellish and R. Dale. 1998. Evaluation in the context of natural language generation. *Computer Speech and Language*, 12(4):349–373. doi:10.1006/csla.1998.0106.

J. Nielsen and J. Levy. 1994. Measuring usability: preference vs. performance. *Communications of the ACM*, 37(4):66–75. doi:10.1145/175276.175282.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*. acl:P02-1040.

C. Paris, D. Scott, N. Green, K. McCoy, , and D. McDonald. 2007. Desiderata for evaluation of natural language generation. In Dale and White (2007), chapter 2.

A. Stent, M. Marge, and M. Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Computational Linguistics and Intelligent Text Processing*, pages 341–351. Springer. doi:10.1007/b105772.

M. A. Walker. 2005. Can we talk? Methods for evaluation and training of spoken dialogue systems. *Language Resources and Evaluation*, 39(1):65–75. doi:10.1007/s10579-005-2696-1.

M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. 1997. PARADISE: A general framework for evaluating spoken dialogue agents. In *Proceedings of ACL/EACL 1997*. acl:P97-1035.

M. White. 2006. Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75. doi:10.1007/s11168-006-9010-2.