Comparing Automatic and Human Evaluation of NLG Systems

Anja Belz
Natural Language Technology Group
CMIS, University of Brighton
UK
A.S.Belz@brighton.ac.uk

Ehud Reiter
Dept of Computing Science
University of Aberdeen
UK
ereiter@csd.abdn.ac.uk

Abstract

We consider the evaluation problem in Natural Language Generation (NLG) and present results for evaluating several NLG systems with similar functionality, including a knowledge-based generator and several statistical systems. We compare evaluation results for these systems by human domain experts, human non-experts, and several automatic evaluation metrics, including NIST, BLEU, and ROUGE. We find that NIST scores correlate best ($>$ 0.8) with human judgments, but that all automatic metrics we examined are biased in favour of generators that select on the basis of frequency alone. We conclude that automatic evaluation of NLG systems has considerable potential, in particular where high-quality reference texts and only a small number of human evaluators are available. However, in general it is probably best for automatic evaluations to be supported by human-based evaluations, or at least by studies that demonstrate that a particular metric correlates well with human judgments in a given domain.

1 Introduction

Evaluation is becoming an increasingly important topic in Natural Language Generation (NLG), as in other fields of computational linguistics. Some NLG researchers are impressed by the success of the BLEU evaluation metric (Papineni et al., 2002) in Machine Translation (MT), which has transformed the MT field by allowing researchers to quickly and cheaply evaluate the impact of new ideas, algorithms, and data sets. BLEU and related metrics work by comparing the output of an MT system to a set of reference ('gold standard') translations, and in principle this kind of evaluation could be done with NLG systems as well. Indeed NLG researchers are already starting to use BLEU (Habash, 2004; Belz, 2005) in their evaluations, as this is much cheaper and easier to organise than the human evaluations that have traditionally been used to evaluate NLG systems.

However, the use of such corpus-based evaluation metrics is only sensible if they are known to be correlated with the results of human-based evaluations. While studies have shown that ratings of MT systems by BLEU and similar metrics correlate well with human judgments (Papineni et al., 2002; Doddington, 2002), we are not aware of any studies that have shown that corpus-based evaluation metrics of NLG systems are correlated with human judgments; correlation studies have been made of individual components (Bangalore et al., 2000), but not of systems.

In this paper we present an empirical study of how well various corpus-based metrics agree with human judgments, when evaluating several NLG systems that generate sentences which describe changes in the wind (for weather forecasts). These systems do not perform content determination (they are limited to microplanning and realisation), so our study does not address corpus-based evaluation of content determination.

2 Background

2.1 Evaluation of NLG systems

NLG systems have traditionally been evaluated using human subjects (Mellish and Dale, 1998). NLG evaluations have tended to be of the *intrinsic* type (Sparck Jones and Galliers, 1996), involving subjects reading and rating texts; usually subjects

are shown both NLG and human-written texts, and the NLG system is evaluated by comparing the ratings of its texts and human texts. In some cases, subjects are shown texts generated by several NLG systems, including a baseline system which serves as another point of comparison. This methodology was first used in NLG in the mid-1990s by Coch (1996) and Lester and Porter (1997), and continues to be popular today.

Other, *extrinsic*, types of human evaluations of NLG systems include measuring the impact of different generated texts on task performance (Young, 1999), measuring how much experts post-edit generated texts (Sripada et al., 2005), and measuring how quickly people read generated texts (Williams and Reiter, 2005).

In recent years there has been growing interest in evaluating NLG texts by comparing them to a corpus of human-written texts. As in other areas of NLP, the advantages of automatic corpus-based evaluation are that it is potentially much cheaper and quicker than human-based evaluation, and also that it is repeatable. Corpus-based evaluation was first used in NLG by Langkilde (1998), who parsed texts from a corpus, fed the output of her parser to her NLG system, and then compared the generated texts to the original corpus texts. Similar evaluations have been used e.g. by Bangalore et al. (2000) and Marciniak and Strube (2004).

Such corpus-based evaluations have sometimes been criticised in the NLG community, for example by Reiter and Sripada (2002). Grounds for criticism include the fact that regenerating a parsed text is not a realistic NLG task; that texts can be very different from a corpus text but still effectively meet the system's communicative goal; and that corpus texts are often not of high enough quality to form a realistic test.

2.2 Automatic evaluation of generated texts in MT and Summarisation

The MT and document summarisation communities have developed evaluation metrics based on comparing output texts to a corpus of human texts, and have shown that some of these metrics are highly correlated with human judgments.

The BLEU metric (Papineni et al., 2002) in MT has been particularly successful; for example MT-05, the 2005 NIST MT evaluation exercise, used BLEU-4 as the only method of evaluation. BLEU is a precision metric that assesses the quality of a translation in terms of the proportion of its word n-grams ($n = 4$ has become standard) that it shares with one or more high-quality reference translations. BLEU scores range from 0 to 1, 1 being the highest which can only be achieved by a translation if all its substrings can be found in one of the reference texts (hence a reference text will always score 1). BLEU should be calculated on a large test set with several reference translations (four appears to be standard in MT). Properly calculated BLEU scores have been shown to correlate reliably with human judgments (Papineni et al., 2002).

The NIST MT evaluation metric (Doddington, 2002) is an adaptation of BLEU, but where BLEU gives equal weight to all n-grams, NIST gives more importance to less frequent (hence more informative) n-grams. BLEU's ability to detect subtle but important differences in translation quality has been questioned, some research showing NIST to be more sensitive (Doddington, 2002; Riezler and Maxwell III, 2005).

The ROUGE metric (Lin and Hovy, 2003) was conceived as document summarisation's answer to BLEU, but it does not appear to have met with the same degree of enthusiasm. There are several different ROUGE metrics. The simplest is ROUGE-N, which computes the highest proportion in any reference summary of n-grams that are matched by the system-generated summary. A procedure is applied that averages the score across leave-one-out subsets of the set of reference texts. ROUGE-N is an almost straightforward n-gram recall metric between two texts, and has several counter-intuitive properties, including that even a text composed entirely of sentences from reference texts cannot score 1 (unless there is only one reference text). There are several other variants of the ROUGE metric, and ROUGE-2, along with ROUGE-SU (based on skip bigrams and unigrams), were among the official scores for the DUC 2005 summarisation task.

2.3 SUMTIME

The SUMTIME project (Reiter et al., 2005) developed an NLG system which generated textual weather forecasts from numerical forecast data. The SUMTIME system generates specialist forecasts for offshore oil rigs. It has two modules: a content-determination module that determines the content of the weather forecast by analysing the numerical data using linear segmentation and

other data analysis techniques; and a microplanning and realisation module which generates texts based on this content by choosing appropriate words, deciding on aggregation, enforcing the sublanguage grammar, and so forth. SUMTIME generates very high-quality texts, in some cases forecast users believe SUMTIME texts are better than human-written texts (Reiter et al., 2005).

SUMTIME is a knowledge-based NLG system. While its design was informed by corpus analysis (Reiter et al., 2003), the system is based on manually authored rules and code.

As part of the project, the SUMTIME team created a corpus of 1045 forecasts from the commercial output of five different forecasters and the input data (numerical predictions of wind, temperature, etc) that the forecasters examined when they wrote the forecasts (Sripada et al., 2003). In other words, the SUMTIME corpus contains both the inputs (numerical weather predictions) and the outputs (forecast texts) of the forecast-generation process. The SUMTIME team also derived a content representation (called 'tuples') from the corpus texts similar to that produced by SUMTIME's content-determination module. The SUMTIME microplanner/realiser can be driven by these tuples; this mode (combining human content determination with SUMTIME microplanning and realisation) is called SUMTIME-Hybrid. Table 1 includes an example of the tuples extracted from the corpus text (row 1), and a SUMTIME-Hybrid text produced from the tuples (row 5).

2.4 *p*CRU language generation

Statistical NLG has focused on generate-and-select models: a set of alternatives is generated and one is selected with a language model. This technique is computationally very expensive. Moreover, the only type of language model used in NLG are n-gram models which have the additional disadvantage of a general preference for shorter realisations, which can be harmful in NLG (Belz, 2005).

*p*CRU[1] language generation (Belz, 2006) is a language generation framework that was designed to facilitate statistical generation techniques that are more efficient and less biased. In *p*CRU generation, a base generator is encoded as a set of generation rules made up of relations with zero or more atomic arguments. The base generator is then trained on raw text corpora to provide a probability distribution over generation rules. The resulting PCRU generator can be run in several modes, including the following:

Random: ignoring *p*CRU probabilities, randomly select generation rules.

N-gram: ignoring *p*CRU probabilities, generate set of alternatives and select the most likely according to a given n-gram language model.

Greedy: select the most likely among each set of candidate generation rules.

Greedy roulette: select rules with likelihood proportional to their *p*CRU probability.

The greedy modes are deterministic and therefore considerably cheaper in computational terms than the equivalent n-gram method (Belz, 2005).

3 Experimental Procedure

The main goal of our experiments was to determine how well a variety of automatic evaluation metrics correlated with human judgments of text quality in NLG. A secondary goal was to determine if there were types of NLG systems for which the correlation of automatic and human evaluation was particularly good or bad.

Data: We extracted from each forecast in the SUMTIME corpus the first description of wind (at 10m height) from every morning forecast (the text shown in Table 1 is a typical example), which resulted in a set of about 500 wind forecasts. We excluded several forecasts for which we had no input data (numerical weather predictions) or an incomplete set of system outputs; this left 465 texts, which we used in our evaluation.

The inputs to the generators were tuples composed of an index, timestamp, wind direction, wind speed range, and gust speed range (see examples at top of Table 1).

We randomly selected a subset of 21 forecast dates for use in human evaluations. For these 21 forecast dates, we also asked two meteorologists who had not contributed to the original SUMTIME corpus to write new forecasts texts; we used these as reference texts for the automatic metrics. The forecasters created these texts by rewriting the corpus texts, as this was a more natural task for them than writing texts based on tuples.

500 wind descriptions may seem like a small corpus, but in fact provides very good coverage as

[1] Probabilistic Context-free Representational Underspecification.

Input	[[0,0600,SSW,16,20,-,-],[1,NOTIME,SSE,-,-,-,-],[2,0000,VAR,04,08,-,-]]
Corpus	SSW 16-20 GRADUALLY BACKING SSE THEN FALLING VARIABLE 4-8 BY LATE EVENING
Human1	SSW'LY 16-20 GRADUALLY BACKING SSE'LY THEN DECREASING VARIABLE 4-8 BY LATE EVENING
Human2	SSW 16-20 GRADUALLY BACKING SSE BY 1800 THEN FALLING VARIABLE 4-8 BY LATE EVENING
SumTime	SSW 16-20 GRADUALLY BACKING SSE THEN BECOMING VARIABLE 10 OR LESS BY MIDNIGHT
*p*CRU	
-greedy	SSW 16-20 BACKING SSE FOR A TIME THEN FALLING VARIABLE 4-8 BY LATE EVENING
-roulette	SSW 16-20 GRADUALLY BACKING SSE AND VARIABLE 4-8
-2gram	SSW 16-20 BACKING SSE VARIABLE 4-8 LATER
-random	SSW 16-20 AT FIRST FROM MIDDAY BECOMING SSE DURING THE AFTERNOON THEN VARIABLE 4-8

Table 1: Input tuples with corresponding forecasts in corpus, written by two experts and generated by all systems (for 5 Oct 2000).

the domain language is extremely simple, involving only about 90 word forms (not counting numbers and wind directions) and a small handful of different syntactic structures.

Systems and texts evaluated: We evaluated four *p*CRU generators and the SUMTIME system, operating in Hybrid mode (Section 2.3) for better comparability because the *p*CRU generators do not perform content determination.

A base *p*CRU generator was created semi-automatically by running a chunker over the corpus, extracting generation rules and adding some higher-level rules taking care of aggregation, elision etc. This base generator was then trained on 9/10 of the corpus (the training data). 5 different random divisions of the corpus into training and testing data were used (i.e. all results were validated by 5-fold hold-out cross-validation). Additionally, a back-off 2-gram model with Good-Turing discounting and no lexical classes was built from the same training data, using the SRILM toolkit (Stolcke, 2002). Forecasts were then generated for all corpus inputs, in all four generation modes (Section 2.4).

Table 1 shows an example of an input to the systems, along with the three human texts (Corpus, Human1, Human2) and the texts produced by all five NLG systems from this data.

Automatic evaluations: We used NIST[2], BLEU[3], and ROUGE[4] to automatically evaluate the above systems and texts. We computed BLEU-N for $N = 1..4$ (using BLEU-4 as our main BLEU score). We also computed NIST-5 and ROUGE-4. As a baseline we used string-edit (SE) distance with substitution at cost 2, and deletion and insertion at cost 1, and normalised to range 0 to 1 (perfect match). When multiple reference texts are used, the SE score for a generator forecast is the average of its scores against the reference texts; the SE score for a set of generator forecasts is the average of scores for individual forecasts.

Human evaluations: We recruited 9 experts (people with experience reading forecasts for offshore oil rigs) and 21 non-experts (people with no such experience). Subjects did not have a background in NLP, and were native speakers of English. They were shown forecast texts from all the generators and from the corpus, and asked to score them on a scale of 0 to 5, for readability, clarity and general appropriateness. Experts were additionally shown the numerical weather data that the forecast text was based on. At the start, subjects were shown two practice examples. The experiments were carried out over the web. Subjects completed the experiment unsupervised, at a time and place of their choosing.

Expert subjects were shown a randomly selected forecast for 18 of the dates. The non-experts were shown 21 forecast texts, in a repeated Latin squares (non-repeating column and row entries) experimental design where each combination of date and system is assigned one evaluation.

4 Results

Table 2 shows evaluation scores for the five NLG systems and the corpus texts as assessed by experts, non-experts, NIST-5, BLEU-4, ROUGE-4 and SE. Scores are averaged over the 18 forecasts that were used in the expert experiments (for which we had scores by all metrics and humans) in order to make results as directly comparable as possi-

[2] http://cio.nist.gov/esd/emaildir/lists/mt_list/bin00000.bin
[3] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl
[4] http://www.isi.edu/~cyl/ROUGE/latest.html

System	Experts	Non-experts	NIST-5	BLEU-4	ROUGE-4	SE
SUMTIME-Hybrid	0.762 (1)	0.77 (1)	5.985 (2)	0.552 (2)	0.192 (3)	0.582 (3)
*p*CRU-greedy	0.716 (2)	0.68 (3)	6.549 (1)	0.613 (1)	0.315 (1)	0.673 (1)
SUMTIME-Corpus	*0.644 (-)*	*0.736 (-)*	*8.262 (-)*	*0.877 (-)*	*0.569 (-)*	*0.835 (-)*
*p*CRU-roulette	0.622 (3)	0.714 (2)	5.833 (3)	0.478 (4)	0.156 (4)	0.571 (4)
*p*CRU-2gram	0.536 (4)	0.65 (4)	5.592 (4)	0.519 (3)	0.223 (2)	0.626 (2)
*p*CRU-random	0.484 (5)	0.496 (5)	4.287 (5)	0.296 (5)	0.075 (5)	0.464 (5)

Table 2: Evaluation scores against 2 reference texts, for set of 18 forecasts used in expert evaluation.

	Experts	Non-experts	NIST-5	BLEU-4	ROUGE-4	SE
Experts	1 (0.799)	0.845 (0.510)	0.825	0.791	0.606	0.576
Non-experts	0.845 (0.496)	1 (0.609)	0.836	0.812	0.534	0.627
NIST-5	0.825 (0.822)	0.836 (0.83)	1 (0.991)	0.973	0.884	0.911
BLEU-4	0.791 (0.790)	0.812 (0.808)	0.973	1 (0.995)	0.925	0.949
ROUGE-4	0.606 (0.604)	0.534 (0.534)	0.884	0.925	1 (0.995)	0.974
SE	0.576 (0.568)	0.627 (0.614)	0.911	0.949	0.974	1 (0.984)

Table 3: Pearson correlation coefficients between all scores for systems in Table 2.

ble. Human scores are normalised to range 0 to 1. Systems are ranked in order of the scores given to them by experts. All ranks are shown in brackets behind the absolute scores.

Both experts and non-experts score SUMTIME-Hybrid the highest, and *p*CRU-2gram and *p*CRU-random the lowest. The experts have *p*CRU-greedy in second place, where the non-experts have *p*CRU-roulette. The experts rank the corpus forecasts fourth, the non-experts second.

We used approximate randomisation (AR) as our significance test, as recommended by Riezler and Maxwell III (2005). Pair-wise tests between results in Table 2 showed all but three differences to be significant with the likelihood of incorrectly rejecting the null hypothesis $p < 0.05$ (the standard threshold in NLP). The exceptions were the differences in NIST and SE scores for SUMTIME-Hybrid/*p*CRU-roulette, and the difference in BLEU scores for SUMTIME-Hybrid/*p*CRU-2gram.

Table 3 shows Pearson correlation coefficients (PCC) for the metrics and humans in Table 2. The strongest correlation with experts and non-experts is achieved by NIST-5 (0.82 and 0.83), with ROUGE-4 and SE showing especially poor correlation. BLEU-4 correlates fairly well with the non-experts but less with the experts.

We computed another correlation statistic (shown in brackets in Table 3) which measures how well scores by an arbitrary single human or run of a metric correlate with the average scores by a set of humans or runs of a metric. This is computed as the average PCC between the scores assigned by individual humans/runs of a metric (indexing the rows in Table 3) and the average scores assigned by a set of humans/runs of a metric (indexing the columns in Table 3). For example, the PCC for non-experts and experts is 0.845, but the average PCC between individual non-experts and average expert judgment is only 0.496, implying that an arbitrary non-expert is not very likely to correlate well with average expert judgments. Experts are better predictors for each other's judgments (0.799) than non-experts (0.609). Interestingly, it turns out that an arbitrary NIST-5 run is a better predictor (0.822) of average expert opinion than an arbitrary single expert (0.799).

The number of forecasts we were able to use in our human experiments was small, and to back up the results presented in Table 2 we report NIST-5, BLEU-4, ROUGE-4 and SE scores averaged across the five test sets from the *p*CRU validation runs, in Table 4. The picture is similar to results for the smaller data set: the rankings assigned by all metrics are the same, except that NIST-5 and SE have swapped the ranks of SUMTIME-Hybrid and *p*CRU-roulette. Pair-wise AR tests showed all differences to be significant with $p < 0.05$, except for the differences in BLEU, NIST and ROUGE scores for SUMTIME-Hybrid/*p*CRU-roulette, and the difference in BLEU scores for SUMTIME-Hybrid/*p*CRU-2gram.

In both Tables 2 and 4, there are two major differences between the rankings assigned by hu-

System	Experts	NIST-5	BLEU-4	ROUGE-4	SE
SUMTIME-Hybrid	1	6.076 (3)	0.527 (2)	0.278 (3)	0.607 (4)
*p*CRU-greedy	2	6.925 (1)	0.641 (1)	0.425 (1)	0.758 (1)
SUMTIME-Corpus	-	*9.317 (-)*	*1 (-)*	*1 (-)*	*1 (-)*
*p*CRU-roulette	3	6.175 (2)	0.497 (4)	0.242 (4)	0.679 (3)
*p*CRU-2gram	4	5.685 (4)	0.519 (3)	0.315 (2)	0.712 (2)
*p*CRU-random	5	4.515 (5)	0.313 (5)	0.098 (5)	0.551 (5)

Table 4: Evaluation scores against the SUMTIME corpus, on 5 test sets from *p*CRU validation.

man and automatic evaluation: (i) Human evaluators prefer SUMTIME-Hybrid over *p*CRU-greedy, whereas all the automatic metrics have it the other way around; and (ii) human evaluators score *p*CRU-roulette highly (second and third respectively), whereas the automatic metrics score it very low, second worst to random generation (except for NIST which puts it second).

There are two clear tendencies in scores going from left (humans) to right (SE) across Tables 2 and 4: SUMTIME-Hybrid goes down in rank, and *p*CRU-2gram comes up.

In addition to the BLEU-4 scores shown in the tables, we also calculated BLEU-1, BLEU-2, BLEU-3 scores. These give similar results, except that BLEU-1 and BLEU-2 rank *p*CRU-roulette as highly as the human judges.

It is striking how low the experts rank the corpus texts, and to what extent they disagree on their quality. This appears to indicate that corpus quality is not ideal. If an imperfect corpus is used as the gold standard for the automatic metrics, then high correlation with human judgments is less likely, and this may explain the difference in human and automatic scores for SUMTIME-Hybrid.

5 Discussion

If we assume that the human evaluation scores are the most valid, then the automatic metrics do not do a good job of comparing the knowledge-based SUMTIME system to the statistical systems.

One reason for this could be that there are cases where SUMTIME deliberately does not choose the most common option in the corpus, because its developers believed that it was not the best for readers. For example, in Table 1, the human forecasters and *p*CRU-greedy use the phrase *by late evening* to refer to 0000, *p*CRU-2gram uses the phrase *later*, while SUMTIME-Hybrid uses the phrase *by midnight*. The *p*CRU choices reflect frequency in the SUMTIME corpus: *later* (837 instances) and *by late evening* (327 instances) are more common than *by midnight* (184 instances). However, forecast readers dislike this use of *later* (because *later* is used to mean something else in a different type of forecast), and also dislike variants of *by evening*, because they are unsure how to interpret them (Reiter et al., 2005); this is why SUMTIME uses *by midnight*.

The SUMTIME system builders believe deviating from corpus frequency in such cases makes SUMTIME texts better from the reader's perspective, and it does appear to increase human ratings of the system; but deviating from the corpus in such a way *decreases* the system's score under corpus-similarity metrics. In other words, judging the output of an NLG system by comparing it to corpus texts by a method that rewards corpus similarity will penalise systems which do not base choice on highest frequency of occurrence in the corpus, even if this is motivated by careful studies of what is best for text readers.

The MT community recognises that BLEU is not effective at evaluating texts which are as good as (or better than) the reference texts. This is not a problem for MT, because the output of current (wide-coverage) MT systems is generally worse than human translations. But it is an issue for NLG, where systems are domain-specific and can generate texts that are judged better by humans than human-written texts (as seen in Tables 4 and 2).

Although the automatic evaluation metrics generally replicated human judgments fairly well when comparing different statistical NLG systems, there was a discrepancy in the ranking of *p*CRU-roulette (ranked high by humans, low by several of the automatic metrics). *p*CRU-roulette differs from the other statistical generators because it does not always try to make the most common choice (maximise the likelihood of the corpus), instead it tries to vary choices. In particular, if there are several competing words and phrases with similar prob-

abilities, *p*CRU-roulette will tend to use different words and phrases in different texts, whereas the other statistical generators will stick to those with the highest frequency. This behaviour is penalised by the automatic evaluation metrics, but the human evaluators do not seem to mind it.

One of the classic rules of writing is to vary lexical and syntactic choices, in order to keep text interesting. However, this behaviour (variation for variation's sake) will always reduce a system's score under corpus-similarity metrics, even if it enhances text quality from the perspective of readers. Foster and Oberlander (2006), in their study of facial gestures, have also noted that humans do not mind and indeed in some cases prefer variation, whereas corpus-based evaluations give higher ratings to systems which follow corpus frequency.

Using more reference texts does counteract this tendency, but only up to a point: no matter how many reference texts are used, there will still be one, or a small number of, most frequent variants, and using anything else will still worsen corpus-similarity scores.

Canvassing expert opinion of text quality and averaging the results is also in a sense frequency-based, as results reflect what the majority of experts consider good variants. Expert opinions can vary considerably, as shown by the low correlation among experts in our study (and as seen in corpus studies, e.g. Reiter et al., 2005), and evaluations by a small number of experts may also be problematic, unless we have good reason to believe that expert opinions are highly correlated in the domain (which was certainly not the case in our weather forecast domain). Ultimately, such disagreement between experts suggests that (intrinsic) judgments of the text quality — whether by human or metric — really should be be backed up by (extrinsic) judgments of the effectiveness of a text in helping real users perform tasks or otherwise achieving its communicative goal.

6 Future Work

We plan to further investigate the performance of automatic evaluation measures in NLG in the future: (i) performing similar experiments to the one described here in other domains, and with more subjects and larger test sets; (ii) investigating whether automatic corpus-based techniques can evaluate content determination; (iii) investigating how well both human ratings and corpus-based measures correlate with extrinsic evaluations of the effectiveness of generated texts. Ultimately, we would like to move beyond critiques of existing corpus-based metrics to proposing (and validating) new metrics which work well for NLG.

7 Conclusions

Corpus quality plays a significant role in automatic evaluation of NLG texts. Automatic metrics can be expected to correlate very highly with human judgments only if the reference texts used are of high quality, or rather, can be expected to be judged high quality by the human evaluators. This is especially important when the generated texts are of similar quality to human-written texts.

In MT, high-quality texts vary less than generally in NLG, so BLEU scores against 4 reference translations from reputable sources (as in MT '05) are a feasible evaluation regime. It seems likely that for automatic evaluation in NLG, a larger number of reference texts than four are needed.

In our experiments, we have found NIST a more reliable evaluation metric than BLEU and in particular ROUGE which did not seem to offer any advantage over simple string-edit distance. We also found individual experts' judgments are not likely to correlate highly with average expert opinion, in fact less likely than NIST scores. This seems to imply that if expert evaluation can only be done with one or two experts, but a high-quality reference corpus is available, then a NIST-based evaluation may produce more accurate results than an expert-based evaluation.

It seems clear that for automatic corpus-based evaluation to work well, we need high-quality reference texts written by many different authors and large enough to give reasonable coverage of phenomena such as variation for variation's sake. Metrics that do not exclusively reward similarity with reference texts (such as NIST) are more likely to correlate well with human judges, but all of the existing metrics that we looked at still penalised generators that do not always choose the most frequent variant.

The results we have reported here are for a relatively simple sublanguage and domain, and more empirical research needs to be done on how well different evaluation metrics and methodologies (including different types of human evaluations) correlate with each other. In order to establish reliable and trusted automatic cross-system

evaluation methodologies, it seems likely that the NLG community will need to establish how to collect large amounts of high-quality reference texts and develop new evaluation metrics specifically for NLG that correlate more reliably with human judgments of text quality and appropriateness. Ultimately, research should also look at developing new evaluation techniques that correlate reliably with the real world usefulness of generated texts. In the shorter term, we recommend that automatic evaluations of NLG systems be supported by conventional large-scale human-based evaluations.

Acknowledgments

Anja Belz's part of the research reported in this paper was supported under UK EPSRC Grant GR/S24480/01. Many thanks to John Carroll, Roger Evans and the anonymous reviewers for very helpful comments.

References

S. Bangalore, O. Rambow, and S. Whittaker. 2000. Evaluation metrics for generation. In *Proc. 1st International Conference on Natural Language Generation*, pages 1–8.

A. Belz. 2005. Statistical generation: Three methods compared and evaluated. In *Proc. 10th European Workshop on Natural Language Generation (ENLG'05)*, pages 15–23.

A. Belz. 2006. pCRU: Probabilistic generation using representational underspecification. Technical Report ITRI-06-01, ITRI, University of Brighton.

J. Coch. 1996. Evaluating and comparing three text production techniques. In *Proc. 16th International Conference on Computational Linguistics (COLING-1996)*.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology*.

M. E. Foster and J. Oberlander. 2006. Data-driven generation of emphatic facial displays. In *Proceedings of EACL-2006*.

N. Habash. 2004. The use of a structural n-gram language model in generation-heavy hybrid machine translation. In *Proc. 3rd International Conference on Natural Language Generation (INLG '04)*, volume 3123 of *LNAI*, pages 61–69. Springer.

I. Langkilde. 1998. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proc. 2nd International Natural Language Generation Conference (INLG '02)*.

J. Lester and B. Porter. 1997. Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics*, 23(1):65–101.

C.-Y. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. HLT-NAACL 2003*, pages 71–78.

T. Marciniak and M. Strube. 2004. Classification-based generation using TAG. In *Natural Language Generation: Proceedings of INLG-2994*, pages 100–109. Springer.

C. Mellish and R. Dale. 1998. Evaluation in the context of natural language generation. *Computer Speech and Language*, 12:349–373.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. ACL-2002*, pages 311–318.

E. Reiter and S. Sripada. 2002. Should corpora texts be gold standards for NLG? In *Proc. 2nd International Conference on Natural Language Generation*, pages 97–104.

E. Reiter, S. Sripada, and R. Robertson. 2003. Acquiring correct knowledge for natural language generation. *Journal of Artificial Intelligence Research*, 18:491–516.

E. Reiter, S. Sripada, J. Hunter, and J. Yu. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

S. Riezler and J. T. Maxwell III. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 57–64.

K. Sparck Jones and J. R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer Verlag.

S. Sripada, E. Reiter, J. Hunter, and J. Yu. 2003. Exploiting a parallel TEXT-DATA corpus. In *Proc. Corpus Linguistics 2003*, pages 734–743.

S. Sripada, E. Reiter, and L. Hawizy. 2005. Evaluation of an NLG system used post-edit data: Lessons learned. In *Proc. ENLG-2005*, pages 133–139.

A. Stolcke. 2002. SRILM: An extensible language modeling toolkit. In *Proc. 7th International Conference on Spoken Language Processing (ICSLP '02)*, pages 901–904,.

S. Williams and E. Reiter. 2005. Generating readable texts for readers with low basic skills. In *Proc. ENLG-2005*, pages 140–147.

M. Young. 1999. Using Grice's maxim of quantity to select the content of plan descriptions. *Artificial Intelligence*, 115:215–256.