Can we evaluate the quality of generated text?

David Hardcastle, Donia Scott

The Open University
Walton Hall, Milton Keynes, UK
d.w.hardcastle@open.ac.uk, d.scott@open.ac.uk

Abstract

Evaluating the output of NLG systems is notoriously difficult, and performing assessments of text quality even more so. A range of automated and subject-based approaches to the evaluation of text quality have been taken, including comparison with a putative gold standard text, analysis of specific linguistic features of the output, expert review and task-based evaluation. In this paper we present the results of a variety of such approaches in the context of a case study application. We discuss the problems encountered in the implementation of each approach in the context of the literature, and propose that a test based on the Turing test for machine intelligence offers a way forward in the evaluation of the subjective notion of text quality.

1. Introduction

Evaluations of NLG systems focus on a wide range of textual features, such as readability (Williams and Reiter, 2005), grammaticality (Habash, 2003), fluency (Mutton et al., 2007), and fidelity (Hartley and Scott, 2001), to cite just a small sample. While all of these measures are useful aids to development, none individually characterises text quality. As Dale and Mellish (1998, 3) point out "there is no agreed objective criterion for comparing the 'goodness' of texts". Indeed, we propose that text quality, in addition to being subjective, is a non-compositional, epiphenomenal property of the text – a property that emerges only through a holistic appraisal of the features of a specific text in some context, and one that is more than just the sum of its parts.

In this paper we relate the evaluation of a case study NLG application called ENIGMA (Hardcastle, 2007, 228-265) which automatically generates cryptic crossword clues. These are short texts, not dissimilar to newspaper headlines, which typically consist of a single clause and some ellipsis. Each clue appears to be a fragment of English prose, but in fact it also contains a series of instructions, encoded according to a set of conventions, that together form a word play puzzle (such as a palindrome, an anagram or some combination of word plays) which the reader has to solve.

Attempts to evaluate the quality of the output text using metrics, gold standard comparison, a task-based exercise and domain expert review all produced some results of interest, but they also threw up problems relating to the lack of a clear model of text quality. The most convincing test of the system was a Turing-style test in which the participants were presented with pairs of clues and told that one of each pair had been generated by a computer. Their goal was to correctly identify as many of the human-authored clues as possible, and they were also asked to provide comments on how they had reached their decision before being shown the answers. The Turing test (Turing, 1950) was designed to address a particular problem in AI, namely that there was no agreed definition of what constituted intelligence. Rather than try to define it, the researcher involves the test subjects in a game in which they try to spot the machine participant and in so doing make a measurable, decisive judgment, while also providing some insights into *their* notions of intelligence, given the domain. We propose that since text quality, like intelligence, is a subjective notion for which we have no working model, the Turing test is a useful way of assessing the quality of generated text.

In the following section, we report on the different evaluation approaches taken and discuss the shortcomings of each, with respect to the domain of the application and also in the wider context of NLG. In Section 3 we draw on user notions of text quality that emerged during the Turing-style test to build a picture of what it is about text quality that makes it so hard to evaluate. Finally, in Section 4, we reflect on the benefits of the Turing-style test in the context of the existing literature on the evaluation of text quality in NLG.

2. Case Study

In this section we report on tests performed as part of the evaluation of the ENIGMA system covering five approaches to evaluation and attempt to draw out some more general conclusions about the benefits of each approach in turn.

2.1. Metrics

Dale and Mellish (1998, 5f) propose a focus on evaluating the *component tasks* of an NLG system, such as lexicalization or aggregation, using agreed metrics to determine how each component of the system is performing. The merits of such an approach include the fact that it can be automated, that the results can be compared across systems, and that it takes us closer to being able to build component-based NLG architectures, which would promote reuse.

In the evalution of ENIGMA two experiments were performed which used metrics to evaluate specific components of the system. The first measured the ability of the system to represent the content of the cryptic puzzle accurately. For example, an anagram which forms part of the puzzle must be adjacent to a keyword indicating an anagram (such as *jumbled* or *messy*) in the final clue. If this is not the case then the clue will not be a valid puzzle. In the experiment several thousand generated clues were checked for accuracy and no errors were reported. This demonstrated that there were no bugs in the *implementation* of the application, but provided no effective evaluation of the *design*, because

the assumptions about what makes a cryptic puzzle valid – the system's *model* of the domain – was not itself under test.

While the first test measured fidelity, in a restricted sense, the second test measured the fluency of the clues by checking grammaticality – see also Mutton et al (2007). The test involved running ENIGMA with syntactic and semantic constraints turned off as a control and comparing the grammaticality of the output with clues generated by the system when running with full syntactic and semantic checking in place. As a point of comparison a sample of 150 human-authored clues taken from the Independent newspaper were also analysed. The parser used was a statistical parser – the Stanford parser (Klein and Manning, 2003) – and so a parse was returned for all clues in the experiment, including those with no syntactic constraints. The test involved checking whether the tags assigned by the parser matched the tags assigned by the generator[1]. Because of the elliptical nature of the clue texts the parser was not able to correctly parse many of the clues, indeed running the parser against a hand-tagged sample of 200 cryptic crossword clues revealed at least one mis-tagged lexical item in 66% of the sample, suggesting an upper bound for the test of 34%. Table 1 shows the results of the test; the percentage match is the percentage of clues in the sample where the pos tags assigned by the parser matched the pos tags assigned by the system. In the case of the human-authored clues the pos tags were assigned manually.

Test Set	Sample Size	Match
Generated Clues	3,000	34%
Control Set	3,000	3%
Human-Authored Clues	150	46%

Table 1: Results of the Grammaticality Test

The grammaticality test provides more information about the components tested than the correctness test because it makes use of an external program with a different (grammatical) model. However, while the metric delivers numerical results they remain open to interpretation. The generated clues performed better than the control set, indeed they matched the upper bound. However, they also performed worse than the human-authored clues, although the human-authored set out-performed the upper bound. The most pressing problem in interpreting the results was that, because the parser struggled with the unusual, elliptical syntax of the clues, the upper bound for performance was low, and so there were a lot of errors which may have been the fault of the generator, or may equally likely have been the responsibility of the parser. While the independence of the external program provides a solution to the problem of circularity, it opens new problems since the program is not well attuned to the specific demands of the application domain.

So component testing using metrics can be informative, but it is hard to make the metrics relevant to the domain of the application without introducing circularity and removing our assumptions about the domain model from the test. It is also hard to induce statements about quality from tests of fidelity and syntactic fluency; while a text which fails these tests might reasonably be judged to be of low quality, assuming that we know how to threshold failure, a pass does not guarantee quality as other factors are likely to be of influence.

2.2. Gold Standard Texts

The Machine Translation community commonly uses comparison metrics to gold standard reference texts – see, for example, Papineni et al (2002) – to evaluate and compare competing systems, and this has allowed the community to develop a collaborative framework to assess the progress of the state of the art. This approach has also been assessed as a possible mechanism through which to evaluate the quality of the output of NLG systems by Belz and Reiter (2006). The algorithm behind the metrics is controversial – see for example Callison-Burch et al (2006) – but in the context of NLG the very notion of a gold standard text is itself challenging. In the domain of cryptic crossword generation the problem is accentuated by the separation between the *content* (expressed by the puzzle) and the *form* (the surface reading of the clue), which means that the same input word can be clued in a myriad of very different ways. As a result, it was not possible to choose a useful reference text for any given puzzle, or even to select a small set of reference clues against which these metrics could be run. Although cryptic crossword clues are unusually variable, we believe that the notion of gold standard texts remains problematic even when the dependency between content and form in the text is unbroken, a point to which we return in Section 4.

2.3. Task-based Evaluation

Another approach is to measure the effectiveness of the system's output in the performance of a third party system or in a related task – see for example Williams and Reiter (2005). In the case of ENIGMA, we performed an automated task-based evaluation using a third party cryptic crossword solving application called *Crossword Maestro*[2]. In this test Crossword Maestro was given two sets of 30 clues to solve, the first set contained clues generated by the system and the second set clues for the same words written by professional crossword compilers for two national newspapers – the Independent and the Sun.

Test Set	Size	Top	Fail
Generated Clues	30	17	7
Authored Clues	30	17	4
The Independent Subset	15	5	4
The Sun Subset	15	12	0

Table 2: Results of the Task-based Test

Table 2 shows the results of the experiment. The *Top* column lists the number of times Crossword Maestro found

[1] Differences accounted for by a small number of common ambiguities in pos-tagging, such as between gerunds and adjectives, were not counted as mismatches.

[2] Crossword Maestro is a commercial cryptic crossword solving tool developed by William Tunstall-Pedoe. A detailed description of the system can be found at www.crosswordmaestro.com.

the correct solution to the clue and ranked it top amongst the candidate solutions, and the *Fail* column lists the number of times that none of the solutions suggested by Crossword Maestro was correct. The first two rows compare the performance of the system to the set of human-authored clues taken as a whole, and the results seem comparable although, as with the results of the metrics described above, it is not obvious what interpretation should be placed on the difference in the number of failed clues. The third and fourth row disaggregate the human-authored set according to the newspaper in which the clue appeared. The Independent is a broadsheet and the crossword is known to be challenging, in contrast the Sun is a tabloid and is often recommended as a good place for would-be crossword solvers to start out. It is clear that Crossword Maestro was much better equipped to solve the clues from the Sun than those from the Independent, but this does not mean that the clues in the Sun are higher quality than the clues in the Independent, just that the clues in the Independent are more difficult to solve. Returning to the figures for ENIGMA this might suggest that the number of failures is a positive result, perhaps indicating that the clues are difficult, like those in the Independent. Of course it could also be that the failures resulted from invalid clues. Conversely, the high number of top-ranking solutions shows that at least some of the clues formed valid puzzles, but perhaps also indicates that they were too easy. Moreover, whichever interpretation we place on the numbers, the results still do not entitle us to make claims about the quality of the text because Crossword Maestro takes no account of the fluency of the surface reading, it only tries to interpret the text as a puzzle[3]. This highlights a potential pitfall in turning to a task or application to evaluate text quality, and that is the assumption that suitability for the task in question entails quality in the text. In the specific case of cryptic crossword clues, the difficulty of the puzzle is just one feature that contributes to quality, and measuring it in isolation is not sufficient to evaluate the quality of the output.

A further problem which the experiment highlights – and this is a problem which only affects automated task-based evaluation – is the risk of dependencies between the system under test and the system used to test it. Every crossword clue must include a word or phrase which acts as a definition for the solution word, and while the definition need not be synonymous with the solution it must be plausibly substitutable. ENIGMA relied on Roget's thesaurus and WordNet for its definitions, with the result that human evaluators often found fault with the definitions that it supplied, most often because they were merely loosely connected co-hyponyms and not plausibly substitutable. Although Crossword Maestro has a much better definition set than ENIGMA, not least because many more man-hours have been invested in its development, some of the definition data is nonetheless sourced from Roget's thesaurus[4], and this indirect dependency could allow Crossword Maestro to solve clues with invalid definition words which human evaluators would dislike.

In more general terms we should expect indirect dependencies between the system under test and the testing system to result from their shared domain. This introduces something of a 'Catch-22' situation for task-based testing using a third party application. If the two systems operate in very different application domains then the models of quality may differ too starkly for the evaluation to be instructive, as we found when trying to parse crossword clues with a statistical parser; conversely, if they share the same application domain, as was the case with the crossword clue solving application, then there are likely to be shared assumptions in the design, or shared algorithms or data sets in the implementation, which could skew the results.

2.4. Domain Expert Review

A selection of clues generated by ENIGMA was sent to seven domain experts, including professional compilers, editors and commentators who were asked to rate the clues in comparison with cryptic clues written by professionals for publication. Between them the experts provided a range of criticisms of the sample clues ranging from high-level comments such as the propensity for the system to rely on anagram puzzles to fine-grained comments on such matters as the use of capitalization to engineer homograph puns. The expert commentary provided was hugely valuable in highlighting aspects of the design and specific components in the implementation which required reworking, but as a formal evaluation the exercise was less informative.

The commentary was valuable because the experts were assessing the quality of each clue as a whole and using their domain expertise to unpick which aspects of the text contributed to the quality, or lack of quality, of each clue. However, they frequently disagreed about the conventions of cryptic crossword clue compilation, with some highlighting as exemplars of good quality the very same clues that others dismissed as unacceptable. There were also many environmental factors that may well have contributed to their assessments, but which are very difficult to weight. For example, one expert was very busy and was only able to provide some very brief comments about the set of clues as a whole, while another was willing to talk through each clue individually and discuss its merits and demerits. Inevitably the latter made many more positive and negative assessments than the former, but as a result we cannot easily aggregate the reviews to draw out an agreed commentary. Another possible factor is the politics of the sector; computer programs are starting to make an appearance in crossword compilation, but many experts in the field feel that the use of machines cheapens the art of cryptic clue composition, and so it might be the case that some are predisposed to dislike clues generated by machines. For example, one expert commented "In much the same way that computer-generated jokes are rarely side-splittingly funny, these clues lack the humour I like to see in crossword clues".

So while the experts were able to provide a much fuller analysis of the quality of the texts than any automated process, it is hard to aggregate the comments and make claims

[3]Good quality cryptic crossword clues should have two quite different readings, a surface reading in which the clue appears to be a piece of English prose, and a puzzle reading in which the text is treated symbolically.

[4]Personal communication, William Tunstall-Pedoe, the author of Crossword Maestro.

about the system based on the evidence of their combined commentary. For this reason it seems that domain expert review makes the most sense with respect to a proof-of-concept or early milestone in a project, but is perhaps less appropriate as a tool for evaluating a completed application.

2.5. Turing-style Test

The final experiment in the evaluation of the ENIGMA system provides a solution of sorts to the problems raised by the other four approaches. The experiment consisted of a Turing-style test (Turing, 1950) in which subjects were presented with 30 pairs of clues, with each pair cluing the same answer word, and were told to choose the human-authored clue for each pair. Participants were also asked to detail the features or failings of the computer-generated clues that most often gave them away *before being presented with the 'answers'*. Figure 1 shows the first 6 questions as presented to the subjects in an online form[5].

Q1. (BROTHER)	☐ Double berth is awkward around rising gold (7) ☐ Sibling getting soup with hesitation (7)
Q2. (DISCARDED)	☐ Dad's cider mistakenly thrown out (9) ☐ Tossed wild cards in past (9)
Q3. (DRAGON)	☐ Note that groan is wild and berserk (6) ☐ Something boring about fire-breathing monster (6)
Q4. (EVENT)	☐ Issue proceeding (5) ☐ Incident involving head of English before opening (5)
Q5. (EXPRESS)	☐ Fast, say (7) ☐ Limited by mean (7)
Q6. (LAPSE)	☐ Slip plate above boiled peas (5) ☐ Slight error in peals, possibly (5)

Figure 1: Sample questions from the Turing-style test

Unlike the expert review process, the subjects in the Turing-style test were forced to make a choice about each of the thirty pairs of clues, and we can aggregate this data to make a single statement about the group's assessment of the quality of the texts; in the case of ENIGMA, the subjects correctly identifed the human-authored clues 72% of the time. This result needs to be contextualised, and we can provide some context for it by determining the upper and lower bound for performance – see Knight and Chander (1994). The lower bound is 100%, a scenario in which it is patently obvious which clues were generated by the computer program, and the upper bound is 50%, indicating that the generated clues were indistinguishable from the real thing. Even with these contextualising bounds the result remains open to interpretation and requires further analysis. For example, the quality of the human-authored clues would have played a role in the result, and the test would benefit from human-human and machine-machine control sets.

A key element of the design of the test is that the subjects were not asked to compare or rate the clues according to any specific set of criteria, other than the implicit criterion that a good quality clue should appear to have been composed by a human author. This ensures that the assumptions behind the design of the system do not leak into the evaluation and that we can, to some extent, have our cake and eat it by leaving judgments about text quality to be open and subjective while still ensuring that the subjects make decisions which we can measure, aggregate and compare.

The soft data post-coded from the comments provided by the test subjects was also of interest. Note that the subjects were not asked to choose which clue was *better*, but to decide which was computer-generated. Furthermore, they were offered no guidance on how to make this decision, but instead were asked to report the basis of their decisions to us. The resulting commentaries provided insights into the success or failure of many components of the system – for example, poor definitions were often highlighted as a give-away, whereas the plausibility of dependencies between items in the text was often highlighted as an authentic-seeming feature. In addition, they provided us with some data about notions of the nature of text quality in the context of the 60 cryptic crossword clues assessed, based on the views of a diverse group people with a range of experience and interest in the domain.

The final point about the Turing-style test is that it was clearly fun to do. Many of the subjects noted that they had enjoyed doing the test, and, since it was hosted on a public-facing web page, they forwarded the URL on to others. This meant that over 60 subjects participated in the experiment, while none of them had to be paid, and indeed many of them had received no direct communication from the evaluator.

3. Text Quality

The comments made by the subjects in the Turing-style test reinforce our intuitions about what makes text quality so subjective[6]. Over 60% of the test subjects who made specific comments on how they distinguished the generated clues from the human-authored ones made remarks about fluency, referring either to the flow of the text or the apparent meaningfulness of the surface reading. In referring to text flow the subjects used words such as "natural", "smooth", and "elegant" to describe the clues that they thought were human-authored and words such as "clunky", "inelegant" and "forced" to describe the lack of fluency that they felt characterised the generated clues. They described the meaningfulness of the text by saying that the human-authored clues had a "ring" to them, contained "connections", evoked "deep plays on words" or evoked coherent "images" or "logical" scenarios.

Although the subjects commented on many other aspects of the clues, there were no other features which were raised with such regularity; the dubiousness of some of the definitions was raised by 15% of those who made comments, and the lack of wit and/or comedy was raised by 12%. Although some of the clue pairs were very similar, the subjects drew no direct comparisons between the clues within

[5] The computer-generated clues are as follows: Q1 first clue, Q2 second clue, Q3 first clue, Q4 first clue, Q5 second clue, Q6 first clue.

[6] We assume that the fact that text quality is a subjective judgment is beyond argument, as anyone who has ever had an article peer-reviewed will surely agree.

the pairs, and none of them made any specific comments about grammaticality or about the validity of the puzzles[7], the two metrics-based tests against which the system was initially tested. Indeed, none of the participants made any specific comments about common computational linguistics metrics, such as length, syntax or lexis at all[8].

Of course, we should be wary of attaching too much salience to the comments made by the subjects, firstly because the notions of text quality that they present may be influenced by prejudice about computer-generated output, and secondly because they did not always correctly identify the human-authored clues and indeed in some instances used computer-generated clues to evidence their assertions about the characteristics of human-authored clues. Nonetheless, the comments made by the test subjects support the notion that text quality is not a just a composite of linguistic features, but a feature of the text when it is taken as a whole. Based on the evaluation exercise, and our own intuition, we suggest that the following three features of text quality contribute most to its subjectivity.

3.1. Non-compositionality

Text quality is a *non-compositional* feature of text, and even if we can measure a range of different features of the different linguistic levels of a text we cannot necessarily combine them in a way which predicts the quality of the text. As the saying goes, "it's all in the telling"; a commonly-held assumption that there is more to the quality of a narrative than the sum of its linguistic features.

3.2. Epiphenomenality

Text quality is *epiphenomenal* in that it only arises from the existence of a given text, and this means that we can't make formal statements about it without a text. This feature of text quality is a consequence of its non-compositionality; we don't have a model about how linguistic features translate into quality, although given a particular text we can make judgments about *its* quality. Of course, given a specific domain, we can make some general statements about the factors that we believe are indicators of text quality, but without a proper model we can't induce such judgments from a set of measurements based on those factors in a reliable or predictable way.

3.3. Context

Finally, text quality is also influenced by *context*, which means that judgments about the quality of a text are influenced by context, or at least by some presumed context. For example, in the Turing-style test described above the generated clue paired with each human-authored counterpart was selected independently based on the ranking assigned to it by the system in comparison with other generated clues for the same word. However, several of the participants noted that there were more clues reliant on anagram wordplay than they would expect to see in a single crossword (the standard context in which cryptic clues are encountered), and all of the domain experts made the same observation. So although the test was not set up with any *explicit* notion of context between the clues, several of the participants inferred a context and made judgements about text quality which took *that* context into account.

The results of the Turing-style test also showed a slight bias in favour of subjects who were regular solvers of broadsheet crosswords (they correctly identified the human-authored clue 73% rather than 66% of the time), highlighting the importance of *audience* in judgements about the quality of text. Taken together, these observations show that we should be wary of making assessments of text quality independently of context and audience.

4. Discussion

So, in the evaluation of ENIGMA we implemented some standard approaches to evaluating the quality of generated output and found that they did not provide us with what we wanted. This finding is not particularly suprising, and echoes the views of others – such as Dale and Mellish (1998). We also implemented an evaluation based on the Turing test, and found it to be much more effective, for a number of reasons. In this section we review what we learned about the different approaches that we took to evaluating ENIGMA and our view that the Turing-style test offers a viable alternative, in the context of some recent work on the evaluation of text quality in NLG systems.

4.1. Gold Standards

Belz and Reiter (2006) explore the prospects for using gold standard comparison metrics from the Machine Translation community – NIST, ROUGE and BLEU – in the evaluation of an NLG system. They note the lack of agreement between domain expert reviewers and report that "a NIST-based evaluation may produce more accurate results than an expert-based evaluation" (2006, 7), although they qualify this suggestion by proposing that new evaluation metrics designed specifically for NLG and high quality reference texts are likely to be required. There is some controversy over the use of such metrics to measure quality, see for example Callison-Burch et al (Callison-Burch et al., 2006), but, in our view, the most significant challenge in using such metrics for evaluating NLG is the task of identifying appropriate reference texts to serve as gold standards – see also Scott and Moore (2006). While in Machine Translation the source text can constrain the packaging of the information which is to be presented, in many NLG contexts we don't want to constrain the system's range of expression artificially.

In the evaluation of ENIGMA this issue meant that gold standard comparison with metrics was simply not viable. Figure 2 shows three cryptic clues for the word DAINTILY by way of example. Clue 1 is taken from a manual for cryptic crossword clue compilers (Macnutt, 2001) written by Ximenes, one of the best-known compilers in the UK – arguably a good candidate as a reference text – and it realizes a wordplay puzzle in which the word *daily*, defined as *char*, is written around the string *int*, which is itself an anagram

[7] Other than the point about the appropriateness of the definition words.

[8] Although two computational linguists who took the test made the (correct) assumption that the generated clues did not contain punctuation and commented that they had used the presence of punctuation as a meta-marker for human authorship.

of the word *tin*. Clues 2 and 3 express the same wordplay puzzle, which means that they express the same crossword puzzle *content* as Clue 1, although only one of them has a similar surface text. Clue 2 is the top-ranking clue generated by ENIGMA for this puzzle, Clue 3 was composed by one of the authors of this paper, a one-time compiler of crosswords for a student newspaper.

1. Char holds messy tin delicately (8)
2. Char holds battered tin delicately (8)
3. Carefully load molten tin into magazine (8)

Figure 2: Three clues for DAINTILY

It is clear that the clue generated by ENIGMA is much closer to the reference clue composed by Ximenes than the third clue, but this does not necessarily make it of better quality. Equally, had clue 3 been chosen as the reference clue, then clue 2 should not degrade in quality. If we were to show Clues 2 and 3 to a group of experts and they disagreed about which was best, this would illustrate the slipperiness of the concept of clue quality in the domain of crosswords, rather than a frailty inherent in expert judgments. Of course cryptic crossword clues are unusually unconstrained in terms of lexical choice and word order given the loose interaction between the puzzle content and the surface reading, but the constraint on expressive freedom that the example illustrates is likely to pose problems for many NLG systems. So, while gold standard comparison may be useful for domains in which language and expression are tightly controlled, in many contexts there are simply too many valid ways of expressing the same content for a gold standard to be helpful. The use of a reference corpus, rather than a single reference text (Bangalore et al., 2000), addresses this problem of excessive stricture, although, as a result, the evaluator has less control over the quality and authorship of the text, and this may mean that the reference text contains undesirable features (Reiter and Sripada, 2002). Furthermore, although a corpus-based comparison can tell us that the output text has the right 'feel' to it, because we are essentially able to show that we can re-classify (Sebastiani, 2002) our output into the correct sub-corpus, the success or failure of this test only tells us part of the story on text quality.

4.2. Other Metrics

Whether a reference text or some other metric, such as grammaticality (Mutton et al., 2007), is used to rate the *intelligibility* of the text, the text must also be analyzed to ensure *fidelity* – namely that the text "says what it [is] supposed to say" (Hartley and Scott, 2001, 1). The experiment conducted to check the fidelity of the clues generated by ENIGMA was rather circular, since the same model used to generate the plans from which the clues were assembled was, as a necessity, re-used to run the test, and so the model itself did not fall under scrutiny. Hartley and Scott (2001) describe in detail how the Generation String Accuracy metric (Bangalore et al., 2000) was used to test the model through which the AGILE system expressed the desired content, by porting the metric so that it measured the correspondence between the desired model and the model produced, rather than between a reference text and the text generated.

One might imagine, then, that a metric could be devised for NLG in which a suite of metrics could be employed in tandem to check all of the qualitative features of the output text: a fidelity test to ensure that it says what it is supposed to say, a grammaticality test to ensure that it reads fluently, a comparison to a reference corpus to ensure an appropriate balance of constructions and vocabulary, and so on. However, such a battery of tests would still be affected by the problems described in Section 2.1. – for example, there may be dependencies between the system and some of the software used to perform the test because of shared domain, the results of each test would require bounding and interpretation, and some of the tests may measure effects in the testing software more noisily than the components under test. Assuming that the tests could be thresholded, they would remain intrinsic, *system* tests which would indicate whether the system is performing as expected, but, given the non-compositional nature of text quality, they would not be sufficient for evaluation, but rather they would represent an essential pre-cursor to evaluating the application.

4.3. Task-Based Testing

We also showed, in Section 2.3., that it is possible to conduct an automated, task-based test of the output of an NLG system, a common practice in evaluating natural language processing systems – see for example, Budanitsky and Hirst (2001). This experiment highlighted some of the difficulties of working with third party software in evaluation, discussed above, and also illustrated the point that even a task that is central to the function of a text, in this case the solving of a crossword clue, is not necessarily fully indicative of quality. An important advantage of a task-based approach is that the evaluation can be designed around the task to induce measurable effects or decisions, and this is particularly helpful with human evaluators as it allows us to aggregate the results. However, not every text has a clear function associated with it, and even when it does it may be hard to disentangle the contribution of the quality of the text to the success or failure of the task. For example, the failure of the clinical trials of STOP (Reiter et al., 2001) might have resulted from a problem with the clinical assumptions behind them, rather than the quality of the texts generated by the system, and even if performance in the task is traced back to the quality of the text it is hard to break the result down and learn lessons about the design of the system. There are examples where a task has been shown to be a good measure of text quality: for instance, Williams and Reiter (2005) were able to evaluate the effectiveness of a system which generated texts for readers with low basic skills because they had a formal model mapping performance in a specific task (reading speed) to a key aspect of text quality for their system, underpinned by psycholinguistic theory, and Sripada et al (2005) used post-edit distance to evaluate generated weather reports motivated by an application requirement to generate texts that required minimal supervision. In many cases though, a specific and relevant task may not be available.

4.4. A Turing-Style Test

With this in mind we propose that the Turing-style test described in Section 2.5. may offer a way forward in the evaluation of text quality in NLG systems. The test has the benefit of an extrinsic test (Jones and Galliers, 1996, 19) in that the human subjects are forced to make decisions that can be measured and aggregated, but the evaluator does not need to find a task that fits the application, since the task is the same every time. The test also provides an intrinsic element, in that the subjects' comments about how they identified the machine-generated texts can be post-coded and aggregated so that they inform us about failings in the design or implementation of the system. Although the test uses reference texts as a point of comparison, the comparisons are made through subjective human judgements, and so there is no artificial restriction imposed on variation or expressivity by the choice of reference texts. Although each subject makes many judgements about quality, the evaluator does not have to set out the criteria against which these judgements should be made, and this ensures that all of the assumptions behind the system's model of quality are subject to testing. Finally, since the test is fun to do and requires only familiarity with the domain rather than suitability for a specific task, the process of finding subjects need not be problematic – compare, for example, Hallett et al (2007).

The down side of the Turing-style test is that the subjects may bring with them many prejudices and preconceptions about computers and machines, and these will inevitably influence what they look for in the text and what comments they make. The Turing-style test as conducted for ENIGMA is therefore incomplete, and would benefit from the addition of control sets of pairs which are entirely human- or entirely computer-generated.

5. Conclusion

Evaluating the quality of the texts generated by NLG systems is a thorny issue, not least because text quality is a subjective matter and we have no formal model through which can induce judgements about quality given data about a range of features of a given text. Researchers have taken a wide range of approaches to evaluating the quality of output text, but each of these approaches introduce problems such as circularity, lack of agreement between reviewers, lack of measurable results, constraints on variability, or difficulty finding subjects.

We propose the Turing-style test, as described in this paper, as a possible way forward which allows us to perform evaluations which force the subjects to make decisions which we can measure and aggregate while also providing independent verification of the model of quality expressed by the design of the system. The task behind the test is generic, and so it can be reused in any domain, and the test is easy to set up and fun to do. The test described in this paper related to cryptic crossword clues, which are short texts, typically containing only a single clause, with unusually high lexical variation. The next step is to explore the viability of the Turing-style test in the evaluation of larger texts comprising many sentences in more typical domains.

6. References

S. Bangalore, O. Rambow, and S. Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the 1st International Natural Language Generation Conference (INLG-2000)*, pages 1–8, Mitzpe Ramon, Israel.

A. Belz and E. Reiter. 2006. Comparing Automatic and Human Evaluation of NLG Systems. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL-2006)*, pages 313–320, Trento, Italy.

A. Budanitsky and G. Hirst. 2001. Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. In *Proceedings of the Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, pages 29–34, Pittsburgh, PA, USA.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL-2006)*, pages 249–256, Trento, Italy.

R. Dale and C. Mellish. 1998. Towards Evaluation in Natural Language Generation. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC-1998)*, pages 555–562, Granada, Spain.

N. Habash. 2003. Matador: A Large-Scale Spanish-English GHMT System. In *Proceedings of the 9th Machine Translation Summit (MT Summit IX)*, pages 149–156, New Orleans, USA.

C. Hallett, D. Scott, and R. Power. 2007. Composing queries through conceptual authoring. *Computational Linguistics*, 33(1):105–133.

D. Hardcastle. 2007. Riddle posed by computer (6): The Computer Generation of Cryptic Crossword Clues. PhD thesis, University of London.

A. Hartley and D. Scott. 2001. Evaluating text quality: judging output texts without a clear source. In *Proceedings of the 8th European Workshop on Natural Language Generation (EWNLG-01)*, pages 111–115, Toulouse, France.

K. Sparck Jones and J. R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer-Verlag New York Inc., New York, USA.

D. Klein and C. D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL-03)*, pages 423–430, Morristown, New Jersey, USA.

K. Knight and I. Chander. 1994. Automated postediting of documents. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pages 779–784, Seattle, USA.

D. S. Macnutt. 2001 [1966]. *Ximenes on the Art of the Crossword*. Swallowtail Books, Claverly, UK.

A. Mutton, M. Dras, S. Wan, and R. Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07)*, pages 344–351, Prague, Czech Republic.

K. Papineni, S.Roukos, T. Ward, and W-J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, USA.

E. Reiter and S. Sripada. 2002. Should Corpora Texts be Gold Standards for NLG? In *Proceedings of the 2nd International Natural Language Generation Conference (INLG-2002)*, pages 97–104, New York, USA.

E. Reiter, R. Robinson, A. S.Lennox, and L. Osman. 2001. Using a randomised controlled clinical trial to evaluate an NLG system. In *Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics (ACL-01)*, pages 442–449, Toulouse, France.

D. Scott and J. Moore. 2006. An NLG evaluation competition? Eight reasons to be cautious. Technical Report 2006/09, The Open University, Milton Keynes, UK.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

S. Sripada, E. Reiter, and L. Hawizy. 2005. Evaluation of an NLG System using Post-Edit Data: Lessons Learned. In *Proceedings of the 10th European Workshop on Natural Language Generation (EWNLG-05)*, pages 133–139, Helsinki, Finland.

A. Turing. 1950. Computing machinery and intelligence. *Mind*, (49):433–460.

S. Williams and E. Reiter. 2005. Generating readable texts for readers with low basic skills. In *Proceedings of the 10th European Workshop on Natural Language Generation (EWNLG-05)*, pages 140–147, Helsinki, Finland.