Computer Speech and Language (1998) **12,** 349–373
Article No. la980106

Evaluation in the context of natural language generation

C. Mellish[†] and R. Dale[‡]

[†] *Department of Artificial Intelligence, University of Edinburgh, Scotland,*
[‡] *Microsoft Research Institute, Macquarie University, Australia*

Abstract

What role should evaluation play in the development of natural language generation (NLG) techniques and systems? In this paper we describe what is involved in natural language generation, and survey how evaluation has figured in work in this area to date. We comment on the issues raised by this existing work and on how the problems of NLG evaluation are different from the problems of evaluating work in natural language *understanding*. The paper is concluded by suggesting a way forward by looking more closely at the component problems that are addressed in natural language generation research; a particular text generation application is examined and the issues that are raised in assessing its performance on a variety of dimensions are looked at.

1. Introduction

Over the last 10 years, the level of interest and concern expressed by the natural language processing community with regard to evaluation has increased substantially. Although this has been driven most notably by the evaluative focus of the DARPA-sponsored MUC and SLS conferences, the effects have been much more widespread than this, and not just restricted to those working in those particular areas. It has now become widely accepted that work in NLP more generally should pay close attention to the evaluation of results. Sparck, Jones and Galliers (1996) provide an important step towards consolidating what we have learned so far and its consequences for the field of NLP as a whole.

Natural language *generation* (NLG) is, in some sense at least, exactly half of the problem of natural language processing; but the evaluation of NLG systems is barely mentioned by Sparck Jones and Galliers. This is not, of course, surprising, and should not be taken as a criticism of that work: just as there is less work in NLG as compared to NLU as a whole, there is similarly much less that has been written on the evaluation of NLG systems than has been written on the evaluation of NLU systems or their components. But neither have researchers in NLG been silent on the matter: in 1990, an AAAI workshop was held on the theme of the *Evaluation of Natural Language*

0885–2308/98/040349+25 $30.00/0

Generation Systems, and papers by Meteer and McDonald (1991) and Moore (1991) provide a fair expression of the NLG community's views on evaluation at that time. In the intervening years, there has been a noticeable increase in empirical work. A common request from reviewers of NLG submissions to journals and conferences is for some material on evaluation to be included, and researchers in the field have tried to respond to such requests. In the last 5 or so years several studies have appeared, which address questions of evaluation in NLG more directly than was previously the case.

The aim of this paper is to survey some of this work to see how much further forward we have come since the beginning of the decade. Do we now have a clearer understanding of how evaluation can play a role in the development of NLG systems and techniques? On the basis of what has been done in recent years, we try here to identify the problematic issues evaluation raises, and suggest a way in which the evaluation of NLG systems and subtasks might be considered in the future. We explore some of these ideas by taking a simple case study of NLG system development and exploring how evaluation might play a role in this work.

Natural language generation is a research area whose content is often unclear to those working outside of the area. We begin, therefore, by providing in Section 2 an overview of what is involved in natural language generation, and elaborate upon the relationship between NLG and the process of natural language understanding.

In Section 3 the question of what it might mean to carry out evaluation in the context of NLG is addressed. We distinguish the evaluation of systems from the evaluation of their underlying theories, and distinguish both of these from task evaluation; each of these aspects is considered by looking at how evaluation has been carried out in the field so far.

On the basis of the preceding material, in Section 4 a number of questions that anyone attempting evaluation in the context of NLG must face are identified. In Section 5, we suggest that rather than attempting to mirror the methods of evaluation used for some NLU tasks, at this stage in the development of the field it is best to reconsider the problem of NLG evaluation from the inside, looking at the component problems that make up NLG.

2. What is Natural Language Generation?

Natural Language Generation is the name we give to a body of research that is concerned with the process of mapping from some underlying representation of information to a presentation of that information in linguistic form, whether textual or spoken. The underlying representation may be symbolic (for example, an expert system knowledge base) or numeric (for example, a database containing stock market prices) but it is generally non-linguistic.

From a theoretical perspective, the task of NLG is to go from some communicative goal (such as a speaker's desire to inform the hearer of something or to persuade them of something) to a text, written or spoken, that satisfies this goal. The key issues for generation often revolve around the notion of *choice*. Of all the possible texts that could be generated, which is the most appropriate to the current context?

A complete NLG system must take a great many decisions to produce such a result. Given the complexity of the task, individual pieces of work in NLG often focus on what are viewed as component parts of this overall problem; this is no different, of course, to the situation in natural language understanding. The most common decomposition of the NLG task, going back at least as far as Thompson (1977), separates decisions about what to say from decisions about how to say it, sometimes referred to as a distinction between STRATEGIC and TACTICAL decisions. During much of the 1980s, this distinction manifested itself within NLG systems as an architectural decomposition into a TEXT PLANNING component and a LINGUISTIC REALIZATION component. However, these terms can mean very different things to different researchers. When we look at the details, we see that there is, as yet, no universally accepted architectural model for natural language generation. This has perhaps been most evident with respect to the question of LEXICAL CHOICE, where the task is to determine what word or words should be used to express some underlying concept: for some researchers this is part of text planning, while for others it belongs firmly within linguistic realization.

Ultimately, when we look at research in the field, it becomes apparent that there are a number of somewhat more specific problems to be dealt with in generating language, with the differences of view being with regard to how these are best combined into components in a system. At the time of writing, we can identify six main categories of problems that are discussed in the literature:

Content Determination: deciding what information should be included in the text, and what should be omitted. Many NLG systems operate in contexts where the information that is to be communicated is selected from a larger body of information, with that selection depending upon a variety of contextual factors, including the intended purpose of the text to be generated and the particular audience at whom it is to be directed.

Document Structuring: deciding how the text should be organized and structured. As soon as we look at real multi-sentential texts, it becomes obvious that there is considerable structure above the level of the sentence. This can be easily demonstrated by taking, for example, a newspaper story and randomly re-ordering the paragraphs and their constituent sentences: the results are invariably incoherent. This means that, given some body of information to convey, an NLG system has to choose an appropriate organization for that information.

Lexicalization: choosing the particular words or phrases that are required in order to communicate the specified information. In some cases the underlying elements of the domain in which the NLG system is operating, will map straightforwardly onto specific words and phrases. However, we cannot assume that the information the system has to work with is linguistic in nature, and so the system may have to choose between different ways of expressing underlying concepts in words. This is most obviously true of systems which are intended to generate texts in more than one language from a common underlying source, although monolingual systems may also need to do work here: for example, it may be appropriate to vary the words used for stylistic effect.

Aggregation: deciding how information should be composed into sentence-sized chunks. Again, we cannot assume that the underlying information is in the form of elements that can be straightforwardly expressed as sentences: often, in the interests of fluency, it will be appropriate to fold several elements of information into one sentence.

Referring Expression Generation: determining what properties of an entity should be used in referring to that entity. In real natural language texts, anaphoric resources such as pronouns and reduced definite noun phrases are used to refer to entities once they have been introduced. This means that an NLG system must decide how to refer to a given entity in the most appropriate fashion; otherwise the risk is redundant and stilted text.

Surface Realization: determining how the underlying content of a text should be mapped into a sequence of grammatically correct sentences. A generally held view is that the same propositional content can often be realized using different sentential forms (for example, a proposition may be expressed via either an active or a passive sentence). An NLG system has to decide which syntactic form to use, and it has to ensure that the resulting text is syntactically and morphologically correct.

This set of categories does not necessarily exhaust the range of problems to be dealt with in natural language generation; and it is not the case that each necessarily corresponds to a specific task or module in an NLG system. However, for many researchers, each of the above categories constitutes a research area in its own right, and serves as the focus for the development of theories and techniques.

As the authors hope to have demonstrated by cataloguing this inventory of concerns, the process of natural language generation can be seen as the inverse of the process of natural language understanding (NLU) as a whole. Whereas in NLU the concern is to map from text to some representation of meaning, NLG is concerned with mapping from some representation of meaning to text. It is important to appreciate that NLG is not simply the inverse of the parsing task in NLU: if any component part of the overall process of NLG is the inverse of parsing (and this itself is questionable), then it is linguistic realization. The overall problem of NLG is best seen as the inverse of the broader NLU problem of determining the speaker's intention that underlies a text.

As noted earlier, while it makes sense to see natural language generation and natural language understanding as the two halves of the puzzle of natural language processing, this symmetry has not so far been reflected in the balance of effort. Although interest in NLG has increased considerably in the last 10 years, it is still the case that the bulk of research in natural language processing is carried out in the context of NLU. There are a number of reasons for this, perhaps the most important of which is that from a practical perspective we are faced with a world where there is a great deal of textual material whose value might be leveraged by the successful achievement of partial achievement of the goals of NLU. Put simply, there is plenty of raw material for researchers in NLU to work with. It is less clear what the appropriate raw material for NLG is, and as a consequence less clear what the real benefits of such research might be. This point will be returned to below; for the moment, it must be stressed that just because there is less work in NLG does not mean that the problems to be dealt with are any less significant than those in NLU; indeed, there are some who would argue that the effort that has been expended in NLP as a whole to date might have been better spread across the field more equally. Whereas a considerable body of work in NLU is still concerned with the parsing of single-sentence utterances to obtain representations of literal semantic content, it is noteworthy that work in NLG is generally more concerned with pragmatic aspects of meaning (for example, the purposes that underlie an utterance, and questions of how information can be structured for presentation in a text) and with larger multi-sentential discourses. An important question here is the extent to which this difference in emphasis may impact on methods of evaluation.

3. What does it mean to evaluate NLG?

There are clear reasons why we should consider evaluation to be important for NLG, as it is for NLU. We need to be able to demonstrate progress, both to fellow researchers and to potential funders, and the field needs to be able to support possible users of the technology by helping them make decisions about whether it is good enough for their purposes. There is, however, no extant wisdom as to how work in NLG should be evaluated.

As Sparck Jones and Galliers (1996) and other commentators on evaluation have pointed out, evaluation can have many objectives and can consider many different dimensions of a system or theory. We will suggest here that it can be fruitful to consider the evaluation of NLG techniques to break down into the following three main categories; the order of presentation here reflects roughly the order in which they would be relevant in the development of an NLG system.

> Evaluating Properties of the Theory: assessing the characteristics (e.g. coverage, domain-independence) of some *theory* underlying an NLG system or one of its parts. We might want to determine whether full implementation of the theory in a domain, or extenstion of an implementation to a new domain, will be productive. So, for example, we might ask whether Rhetorical Structure Theory is an appropriate theory for characterizing the structures of texts to be generated in some domain.
>
> Evaluating Properties of the System: assessing certain characteristics (e.g. coverage, speed, correctness) of some NLG *system* or one of its parts. We might want to compare two NLG systems or algorithms to determine which provides the better results; or we might want to develop some metric by means of which we can determine when the performance of a given system has improved in some way.
>
> Applications Potential: evaluating the potential utility of an NLG system in some environment. We might want to determine whether the use of NLG provides a better solution than some other approach (for example, the use of a sophisticated mail-merge tool, or the presentation of information via graphics, or the hiring of a human author).

These categories can be thought of as to a large extent cumulative, in that an evaluation of a system in general gives some information indirectly about the theories behind its construction and an evaluation of applications potential in general gives some information about both the properties of the system and the underlying theories. To evaluate on applications potential, pragmatic issues that are not necessarily of primary interest to the NLG researcher and which arise when considering the application of any knowledge-based system need to be considered. For instance, how cost effective is the system? What is its impact on the existing (e.g. social) environment? Is the system maintainable? Is it fast enough, robust enough? How does it compare to its rivals (not necessarily all other NLG systems)? These issues are largely ignored here, as they are concerned with what Sparck Jones and Galliers refer to as SETUP rather than SYSTEM; see Reiter & Dale (1997) for some discussion of the issues and alternatives that arise in considering NLG as a solution. Our intended focus in the current paper is on the second category of evaluation, although it can often be difficult to distinguish aspects of the theory from aspects of implementation.

In the next two sections, we summarize the main approaches that have been used up to the present to evaluate NLG theories as such (rather briefly) and systems (more fully). We make use of examples from neighbouring fields (for instance, Machine Translation and Hypertext) when this seems relevant. In an area of this kind, it is impossible to be exhaustive, but it is hoped that examples of the most significant evaluation methods that have been used have been covered.

3.1. Previous approaches to evaluation of NLG theories as such

Good NLG systems are based on theories of various kinds, and there are many reasons why an implementation may not truly represent the theory on which it is built. In particular, constructing a complete NLG system involves solving many practical problems that may be unconnected to the original motivating theory, and it would be foolish to judge a theory by the quality of the generated text if these other problems are influencing it significantly (Sparck Jones & Galliers, 1996).

Properties of a theory (e.g. completeness, computational complexity) can sometimes be demonstrated purely analytically. Such results have been rare in NLG. On the other hand, there is a tradition of evaluating theories by hand simulation in disciplines such as Linguistics, though such evaluations are often not carried out on a large enough scale to yield general conclusions. One might evaluate a grammar, for example, by simulating by hand its ability to characterize a set of test sentences, extending the grammar appropriately whenever a deficiency in coverage is identified. In matters related to NLG, the development of Rhetorical Structure Theory (Mann & Thompson, 1987) was supported by the fact that hand analysis of many types of text was possible using it. The following two examples show other places where it was useful to assess by hand how much of (relevant aspects of) a corpus of natural text could be produced by different algorithms.

Yeh and Mellish (1997) used hand simulation during the development stage of their algorithms to generate referring expressions in Chinese. A human-generated corpus was analysed and different algorithms for referring expression generation were simulated, with intuition being used to guess what knowledge would be available to an actual text generator. For each algorithm, the similarity between the referring expressions in the original corpus and those generated in the simulation gave a basis for evaluating the algorithm. A sequence of increasingly sophisticated algorithms was motivated by this incremental evaluation method; this is not unlike the incremental extension of the coverage of a grammar as described above.

Robin (1994) used a human-generated corpus to evaluate his theory of revision-based generation, which was implemented for the domain of baseball summaries. Whereas Yeh and Mellish were concerned with fluency, in this case it was coverage, extensibility and protability between domains that were under investigation. The first set of evaluations estimated how much of the sub-language for the domain was captured by the results of an analysis of one year's and two years' texts (by examining texts for the succeeding year). This gave a basis for estimating how large a corpus would need to be analysed in order for the whole of the sub-language to be covered. It was also possible to estimate what difference it made using Robin's revision-based model of generation, compared to more traditional models. The second set of evaluations concerned portability to a new financial domain. It estimated to what extent the abstract rules of the system would do useful work in the new domain, thus giving a measure of the domain-independence of the model.

Evaluating a theory by hand simulation can, of course, be dangerous, because the simulator can inadvertently introduce errors or bias. This must be balanced against the incidental problems that arise from actually trying to implement a theory. Overall, hand simulation seems to be a useful tool to support the early stages of the development of theories and to measure parameters that would otherwise be inaccessible to any practical quantitative analysis. However most reported work on NLG evaluates its theories indirectly through the systems that implement them.

3.2. Previous approaches to evaluation of NLG systems

We now move on to consider the evaluation of NLG *systems*, concentrating on approaches which are formal and objective. All computer systems have to be evaluated throughout their development, and inevitably much of this involves informal analysis by the system developers, the development of examples to demonstrate theories and techniques, the tuning of systems to deal with these examples, and so on. Informal evaluation during system development is necessary and cannot be replaced. Here evaluation is looked at more as a way of answering the question (which not everybody necessarily wants or needs to ask) *how good is your NLG system?*

In discussing the evaluation of NLG systems, Meteer and McDonald (1991) echo a distinction made elsewhere in the evaluation literature between *black box* and *glass box* evaluation. Black box evaluation seeks to assess a system as a whole (in terms of its inputs and outputs), whereas glass box evaluation seeks additionally to reflect on the contributions of the parts of a system to its overall performance. The authors begin below by looking at the black box evaluation techniques that have been used in NLG; Section 3.2.4 then considers how these can be extended to glass box techniques.

There are a number of different kinds of evaluation that can be pursued; these are referred to here as *accuracy* evaluation, *fluency/intelligibility* evaluation and *task* evaluation. Each is described in more detail below.

3.2.1. Accuracy evaluation

By *accuracy* we mean the extent to which the generated text conveys the desired meaning to the reader. This is not necessarily correlated with *fluency*, which relates to the extent to which the text flows and is readable, although we might expect there to be some relationship between the two: it may be very difficult to determine whether a severely disfluent text conveys a desired meaning.

As a place to start, we can look at work that has been carried out in Machine Translation. Although the input to the MT task is quite different to that in NLG, both tasks have in common the creation of natural language text as output.

Several researchers (Jordan, Door & Benoit, 1993; Shiwen, 1993) have attempted to assess the accuracy of the texts produced by MT systems. This involved the following three stages:

(1) Selecting a test suite of source texts.
(2) Running the MT system on this suite.
(3) Evaluating the output by comparing with expert human output for the same task, or simply asking an expert to rate the accuracy of the generated texts.

```
R    ——  ——►  (492:1) *  $A [possessive] $B $C *
R    ——  ——►  (492:0) *
$A   ——  ——►  zaochen / shengwu
$B   ——  ——►  liu / 6
$C   ——  ——►  dianzhong / dian / shi
```

Figure 1. Test Description Language example.

In the case of Jordan *et al.* (1993) the test suite was selected in a way intended to reflect the nature of "real" data. Shiwen (1993), on the other hand, constructed a test suite by hand, in such a way as to test known problem areas in Chinese–English translation. This second approach is, of course, only useful if the suite construction is undertaken by somebody not involved in any of the systems to be evaluated.

The comparison between the output of the system and that of expert humans can itself be carried out by humans. Jordon *et al.* (1992) asked subjects to paraphrase the target sentences (hence removing uncertainty about the meaning) and then to assess these paraphrases (compared to the source sentences) as "right", "nearly right" and "wrong". They found a strong correlation between the correctness of the paraphrases and subjects' assessment of clarity of meaning and well-formedness (which are concepts related to intelligibility; see Section 3.2.2. below). However, 8% of the time the evaluators thought they understood the meaning of a sentence (i.e. it was intelligible) when they actually did not (i.e. it was not accurate). This demonstration that accuracy and fluency/intelligibility are not always correlated provides some justification for considering the two factors separately.

Shiwen's approach was to have automatic assessment of the accuracy of the texts, with no involvement of humans. This was made possible by the fact that the test suite was designed by hand to test particular features. For each sentence in the test suite, a set of expressions in a special Test Description Language (TDL) indicated how the translation was to be marked. The example in Figure 1 is for Shiwen's test 492, with possible results 1 or 0. * matches arbitrary characters, [and] indicate optionality and / separates alternatives. If the expression on the right of the first R rule matches the input, then the translation will be given the mark 1; if the second one matches, the mark will be 0.

Accuracy evaluation requires an assessment of the relationship between input and output. With MT, one has the advantages of a source text whose content can be compared (by an expert) with that of the target. Similarly, in automatic summarization, where the task is to select important sentences (Brandow, Mitze & Rau, 1995) or paragraphs (Salton, Singhal, Mitra & Buckley, 1997) from a document, an expert can be presented with exactly the same task as the system. In order to apply these accuracy evaluation techniques to NLG, one would have to find a way of presenting the input to an impartial human evaluator in a way that they would understand. If an expert judges the output of NLG without understanding its input (as was done, for instance, in the KNIGHT experiments discussed below) then accuracy assessments will not necessarily reflect directly on the NLG system. In this case, one is really assessing the performance of a larger system including not just the NLG system but also whatever provides its input. This could be thought of as a kind of task evaluation (see below).

Of course, although with MT one does have an input that an expert can understand, nevertheless in the case of an MT system a failure in accuracy could be due to either the process of interpreting the source text to produce the internal representation from which the target text is generated; or it could be due to a failure in the mapping from this internal representation to the target text. There is no obvious way to determine which aspect of the overall process is the source of error (although intuition suggests that the error is more likely to be due to mistakes in analysis rather than in generation).

Whereas getting an absolute characterization of a system's accuracy may be hard, to get a useful *comparative* evaluation of the accuracy of two systems it might not be necessary for the human evaluator to have perfect understanding of the input (e.g. the input might be paraphrased for them by a knowledge engineer).

3.2.2. Fluency/intelligibility evaluation

Fluency concerns the quality of generated text, rather than the extent to which in conveys the desired information. This is related to the notion of "readability" and will include notions such a syntactic correctness, stylistic appropriateness, organization and coherence.

One might hope that there could be some way to directly measure the fluency of a text. However, although there was a flurry of interest in "readability formulae" in the 1950s (Flesch, 1948; Lorge, 1959) and some of these have made their way into modern word-processors, it seems unlikely that these have much potential for evaluating the output of NLG systems. The first problem is that it is very unclear what the relevance is of what these formulae measure. The second problem is that, because the formulae are very precise and simple, most NLG systems could do well in these terms if it was so wished (for instance, by not using particular words or syntactic constructions). There are some interesting questions here with regard to the trade-offs between different means for achieving the same readability values: see Dras (1997). However, these do not make it any easier to assess the real quality of a generated text. More recent work on controlled languages (Controlled Language Applications, 1996) has developed, with some empirical support, principles for making text in certain restricted domains "simpler" (for example, to permit easier comprehension by non-native speakers or badly-educated users, or to permit easier MT). Again, such principles are usually of a kind that could be implemented fairly straightforwardly in an NLG system and so might be of dubious use for evaluation. Moreover many NLG systems need to produce text which is attractive as well as functional, and many researchers are interested in the problem of producing coherent text exploiting the possibilities offered by real natural languages. So, to assess readability in general there seems to be no alternative to using human subjects.

We have found three approaches to assessing readability:

(1) Measuring comprehension time.
(2) Measuring the time taken to post-edit the text to a state of fluency.
(3) Asking human subjects directly to assess the readability of the text.

Minnis (1993) proposes the first two as ways of evaluating MT systems and suggests that there may be a correlation between them. He also suggests that errors can be classified and that it may be possible automatically to learn the contributions that the different classes of errors make to post-editing time. This would enable one to assess fluency directly by counting and classifying errors.

Eliciting direct human judgements of fluency is standard practice in domains such as text-to-speech systems (Johnston, 1996). This approach was used by Acker and Porter (1994) in the evaluation of a "view retriever" which retrieved facts from a knowledge base in order to describe a concept from a particular perspective. The facts were translated manually into simple English and the subjects were asked to rate the texts from 1 ("looks like a random set of facts") to 5 ("as coherent as a good textbook"). To give a basis for comparison, the same experiment was run with textbook viewpoints, degraded viewpoints and random collections of facts (Table I). A *t*-test showed that there was a significant difference between "view retriever" viewpoints and degraded or random viewpoints. Because the actual English was hand-generated and the subjects were unaware of the desired content, this evaluation shed light primarily on the organization of the texts.

This work was significantly extended by Lester and Porter (1997) into a system called KNIGHT, which was a complete NLG system generating explanations from a very large, and independently developed, knowledge base on Biology. For this, the human evaluation was developed into a *two-panel* methodology, where texts produced by the system are assessed together with texts generated by a first human panel, the assessment being done by a second panel. The judges (who were unaware that some texts were computer-generated) rated texts (using letter grades A, B, C, D and F) according to overall quality and coherence, content, organization, writing style, and correctness. For content and correctness (which relate to accuracy) they used their expert knowledge, rather than having access to the information provided by the knowledge base. The differences between the NLG texts and those for the humans were statistically significant for overall coherence and writing style. The NLG system, however, produced a better performance overall than one of the human writers.

One can take the notion of fluency beyond monologue and consider to what extent it is displayed by a dialogue system. Cawsey (1992) used human subjects to inform the development of a tutoring system using NLG to teach students about electronic circuits. The system allowed flexible dialogues between system and user, and subjects in the evaluation were asked to use the system (for as long as they needed) to find out about four different circuits. The sessions were observed (and recorded) and obvious problems with the interface were detected by inspection. The subjects were asked to fill in a questionnaire about their problems with the system and their subjective assessment of it. This revealed problems, for instance, with the coherence of the dialogue. An additional task evaluation might have revealed content problems that the users were not aware of.

In spite of its weaknesses, the approach of using subjective human judgements still seems to be the most popular way of assessing fluency/intelligibility.

3.2.2. Task evaluation

Many NLG systems are intended to be used by humans to assist them in performing some task or achieving some goal: for example, learning about a subject, making decisions, writing better computer programs, or even changing their lifestyle in some way. Although considering a system's appropriateness for some independently-motivated task in the real world involves taking into account many pragmatic factors of applications potential such as cost and social impact, just considering a system's raw performance in support of a more abstract task can also be revealing. Task evaluation involves

TABLE I. Evaluation of "view retriever" texts

Source	Coherence	
	Mean	σ
(1) Textbook viewpoints	4·23	0·56
(2) View retriever's viewpoints	3·76	0·74
(3) Degraded viewpoints	2·86	0·94
(4) Random collections of facts	2·62	0·86

TABLE II. IDAS: Analysis of nodes visited

Question	Contributing directly to answer	Necessary to move through	Repeat visits to the previous categories	Other	Total
Q1	26	23	2	98	149
Q2	33	14	9	24	80
Q3	45	22	6	50	123
Q4	44	11	4	41	100
Q5	28	47	6	22	103
Totals	176	117	27	235	555

observing how well a (possibly contrived) task is performed using the NLG system. Since the task may have its own independent criteria of success—the user either does or doesn't learn, he or she makes better or worse decisions, and so on—the advantage is that one comes to an assessment of the NLG system without directly having to consider either its input or its output.

Task evaluation has been used in the evaluation of MT systems. Suppose that the purpose of the translated texts is to inform: one can then attempt to measure the success of the translation process by determing how well human readers can answer questions on the basis of reading the target texts. This measures, for this task, how well the translation has succeeded (by whatever means) in conveying the underlying content of the source text.

Another task associated with reading documents is that of assessing and classifying the documents according to their subject area. Jorden *et al.* (1993) addressed the problem of MT for people scanning for texts in particular subject areas. A number of MT systems were compared in terms of how well human readers could state the subject matter of a document by reading its translation.

The IDAS user trials (Levine & Mellish, 1995) assessed a system that used natural language generation techniques in the automatic generation of hypertexts. Here the user task was to retrieve relevant information to answer specific questions, and in the trials subjects were asked to use the system to complete a test paper, where they had to answer factual questions from single and multiple hypertext nodes and to relate the textual descriptions to a picture. Success was measured not just in terms of their ability to answer the questions correctly but also in terms of how effectively they could do so. One measure of the effectiveness of the system was produced by an analysis of the hypertext nodes visited (Table II). The hypertext nodes visited were classified as contributing directly to the question answers, necessary nodes to move through, repeat visits to useful nodes and other nodes. Only 42% of nodes visited were not useful. Users got better as they progressed through a series of five questions (Q1 to Q5), and they did not seem to get "lost".

The advantage of task evaluation in avoiding consideration of the detailed role of the NLG system is, of course, also the root of its main disadvantage. If the IDAS system was indeed successful (or not), it is hard to tell which parts were reponsible for this. For instance, it could have been specific aspects of the user interface that had nothing to do with NLG. In practice, one might hope that other types of evaluation (here, usability questionnaires completed by the subjects) would shed light on this.

3.2.4. Deriving glass box techniques

Of the evaluation methods considered so far, task evaluation appears to be the ultimate in black box methods. We now consider ways in which glass box evaluation can be achieved. Glass box evaluation techniques can often be derived from black box techniques, at least in the sense that one can concentrate on evaluating one part of an NLG system at a time. We have observed this happening in the following forms:

(1) Task evaluation being performed using a task that highlights the performance of a particular module.
(2) Accuracy or fluency evaluation being performed when only the one module is operating, all other NLG functions being carried out by a human.
(3) Accuracy or fluency evaluation being performed with different implementations of the module concerned.

We will now briefly describe an example of each.

Carter (1996) needed to evaluate the results of various automatic methods of introducing hypertext links between pieces of text (for instance, methods which linked documents with similar distributions of word frequencies). Unfortunately, most ways of evaluating hypertext systems (e.g. Nielsen, 1990) treat them as black boxes. Objective measures of hypertexts (e.g. measurable properties of the link structures such as reachability) are not clearly related to usability, acceptability measures (e.g. via questionnaires) are subjective and can be distracted by irrelevant features of the system, error measures are difficult because it is hard to determine what is a genuine error and learning/memory measures (e.g. via questions asked afterwards) will be affected by both text and link quality. The solution was to develop a new task for subjects, which involved them recalling aspects of their traversed paths after performing another task using the system.

Knight and Chander (1994) wanted to evaluate an algorithm for selecting articles for English noun phrases. Their intention was to use this as a post-editing system for MT into English from languages which do not have articles. Their approach was to take existing (unseen) text, remove the articles and then compare the results of the system with the original text. This method could be used for other tasks that can be thought of as completing partial texts (as long as the task does not require any input apart from the provided text). It could be extended to other situations where a human can simulate some aspects of the text generation process and hence where an evaluation of the generated text can be assumed to reflect directly on the modules that are operating automatically.

Yeh and Mellish (1997) evaluated different versions of their algorithm for generating Chinese referring expressions by running them in the context of the same automatically generated text. Since the other parts of the system did not depend on the referring expression generation, this gave a fair basis for comparison. The actual comparison was done in a manner somewhat reminiscent of Knight and Chander's (1994) approach. Referring expressions were removed from the generated text, and human subjects were asked to select from a list of possible expressions to be inserted at each point. The similarity (or otherwise) between the human and the machine selections gave some basis for preferring algorithms taking into account discourse structure and syntactic salience from an algorithm with neither.

4. Issues and problems in evaluating natural language generation

In this section, we summarize some of the general problems that arise in evaluating work in natural language generation. As we will see, all of these problems are to some extent more serious for NLG than they are for NLU.

4.1. What should the input be?

One of the most fundamental obstacles to the evaluation of NLG systems is that, whereas it is fairly clear what an "input text" is for NLU, and examples can be observed directly in the world, there is no simple consensus as to what the input to an NLG system should be. Different researchers make different assumptions about what should be included in the input to generation, and different tasks require different knowledge sources and goals of different kinds. As a result, it is not always easy to judge whether an NLG system is actually avoiding difficult questions by including in its input very direct guidance about how to handle problems that it will have; for instance, a system whose input is supposedly language-independent may actually anticipate the structures that will be generated in a specific language. Natural language analysis systems at least largely agree on the nature of their inputs, characterized via the sequential structure and tokenization of written natural language texts. This kind of agreement makes it possible to adopt architectures like GATE (Gaizauskas, Cunningham, Wilks, Rodgers & Humphreys, 1996), where the results of processing can be expressed as annotations on text spans. With NLG, there is no such agreed basis.

One way to be sure of having realistic and relatively unbiased input is to take input that is provided by another system developed for purposes unrelated to NLG. This was the motivation in (Dale, 1992) for using the output of a hypothesized planning system as the input to language generation; on a larger scale, the same motivation caused Acker, Lester and Porter (Acker & Porter, 1994; Lester & Porter, 1997) to adopt an independently developed biology knowledge base as their source of content for NLG. A significant problem, however, is the scarcity of independently constructed underlying information sources that are rich enough to motivate the kinds of issues that researchers in generation consider. There are many domains where large numerical data sets are available, as in weather forecasting and stock market data, and these can very usefully be adopted as sources for the generation of summaries or reports. However, experience shows that an application that wants to use NLG as a means of producing output often needs to be built with this in mind if the underlying data source is to contain the appropriate kinds of information (see, for example, Swartout, 1983; Mellish & Evans, 1989); presented only with tables of numbers as a source of data for generation, inevitably the NLG system has to be provided with other substantial sources of knowledge to be able to utilize this data.

4.2. What should the output be?

The second major problem with evaluating NLG is that of assessing the output. There is no agreed objective criterion for comparing the "goodness" of texts. Other possible notions like "appropriateness" and "correctness" depend on the task performed and what constitutes success. Once again, a problem is that NLG systems span a wide range of domains and application types, and so a basis for comparison is usually lacking.

There is no "right answer" for an NLG system, because almost always there are many texts that will achieve the same goals, more or less. Evaluation therefore has to consider the quality of the output somehow. Accuracy and fluency are two measures of quality. As we have seen, they are distinct but often related. Some of the work surveyed has not made this distinction clearly, and this reflects a general uncertainty about what the key dimensions are.

Task evaluation attempts to avoid directly asking the question of what the NLG output (and, to some extent, also input) should be. But then its results only reflect indirectly the properties of the NLG system.

4.3. What should be measured?

It is not always obvious what, about the performance of an NLG system, one needs to measure, or what measurable attributes best indicate this. For instance, Carter considered measuring "mean time per node" as an indication of how easy it was to navigate through a hypertext system (low mean time indicating easy navigability). However, it turned out that certain subjects, when lost, resorted to a strategy of rapid random clicking and this, initially unexpected, phenomenon rendered the measure useless.

If possible, it seems best to choose measurements which degrade gracefully. Yeh and Mellish (1997) measured the extent to which generated referring expressions agreed with those selected by humans. However, as we discuss below, the humans could not always agree. It could be that in some circumstances two different referring expressions really are equally good. If the subjects had been asked to *rank* their selections (rather than just indicate the best), a more subtle measure of agreement (robust with respect to certain random variations) between system and humans might have arisen.

Human subjects are a valuable resource and carrying out experiments with humans is relatively time-consuming. One does not want to have to repeat experiments. If in doubt, it is probably worth measuring everything that could possibly be relevant when the subjects are there.

It is also worth bearing in mind from the start the kinds of statistical tests that one is going to use and what they can and cannot show. For instance, standard statistical tests can only reject an identity hypothesis, never accept one (Cohen, 1996).

4.4. What controls should be used?

For many NLG tasks we have no idea either how well it is possible to do or how easily the task can be done, which means that absolute performance figures are not very useful. For instance, in the IDAS trials “only” 42% of nodes visited were irrelevant. Is this good or bad? The only way to tell is to compare it to something else.

Figures obtained for humans performing the same task may give a good upper bound for an NLG system’s performance. Creating random or degraded texts (or hypertexts, or links) may give a basis for establishing a lower bound [this was done by Acker & Porter (1944) and by Carter (1996)]. In between, it may be possible to get figures for different versions of the NLG system [as was done by Yeh & Mellish (1997)].

Sometimes upper and lower bounds can be determined by analysis or simple experimentation. Knight and Chander (1994) pointed out that an article selection algorithm can achieve 67% correctness simply by guessing “the” every time. On the other hand, human experts cannot achieve much better than 95% (given purely the text without the articles). This puts Knight and Chander’s achievement of 78% into perspective.

4.5. How do we get adequate training and test data?

As Robin (1994) points out, such is the complexity of NLG that many people have focused on particular subtasks which have been evaluated using handcoded (and hence few in number) inputs. Few projects have had the benefit of a significant knowledge base of the kind used by Acker, Lester and Porter from which to generate and from which new inputs (uninfluenced by the developers of the NLG system) can be obtained in large numbers. Thus we have rarely had the luxury of having separate training and test sets of the kind that are traditionally required for evaluating systems.

Nevertheless, even if data is not available in large amounts, methodologically it is necessary to separate that which is used to develop the system from that which is used to evaluate the result. The latter needs to be put aside right from the very start, so that it plays no role in system development. Yeh and Mellish’s approach of using a single set of data to motivate a sequence of increasingly “better” algorithms could have resulted in a system that was overtuned to this data set. A separate evaluation involving comparison to humans on different data was essential to discount this possibility.

4.6. How do we handle disagreement among human judges?

In evaluating the output of NLG, the problem that humans will not always agree about subjective qualities like “good style”, “coherence” and “readability” must be faced. Some of the evaluation methods discussed here avoid consulting explicit human judgements for this reason. Others hope to avoid problems by having sufficient independent judgements that disagreements will become invisible as a result of the averaging process.

Although they consulted 12 different native speakers of Chinese, Yeh and Mellish (1997) found significant disagreement about what the preferred referring expressions should be. Faced with results of this kind, a natural strategy is to investigate whether the overall disagreement is being caused by, for instance, one speaker having very unusual intuitions or particular examples being unusually challenging. Unfortunately even these measures did not produce really significant agreement in the evaluations in this case.

As we discussed above, a well-designed experiment will attempt to plan for how to react to disagreement and will attempt to measure things in a way that finds a basis for agreement amongst the human subjects, if only on a slightly different version of the problem. Regardless of how well this succeeds, it is essential to measure the level of agreement between the speakers, and the kappa statistic (Carletta, 1996; Siegel & Castellan, 1988) is very useful for this.

4.7. Summary

We have surveyed the various approaches to evaluation that have been attempted in work on NLG and related areas in recent years, and we have identified some of the problems that arise in attempting some evaluations. In light of these problems, we suggest that while the use of some notion of evaluation in NLG is intuitively appealing, we should be wary of following too closely models of evaluation that have become common and popular in work in NLU. It should be borne in mind that even in NLU not all work is amenable to easy evaluation: MUC-style evaluation (Sundheim & Chinchor, 1993), where a well-defined target data structure can be determined for each input text, is only applicable in very specific circumstances. Other NLU tasks which look at first sight easily evaluable are not necessarily so: for example, although it might seem that the performance of a word sense disambiguation system is easily quantifiable, the results of such experiments are of dubious value unless we have independent motivations for the sets of possible senses that can be chosen—the senses of a word listed in a dictionary may be quite irrelevant for many real tasks where some notion of sense disambiguation is required.

It is hardly surprising, then, that it is unclear how we might evaluate in the context of natural language generation. Because of some of the asymmetries mentioned above, it is not clear how one could find the NLG analogues of the paradigm evaluation scenarios so often discussed in NLU research; intuitively, there is something about judging of the quality of a generated text that makes it a different task from that of determining how many database fields have been filled correctly. This basic difference may require a somewhat different approach to the evaluation task, and we should not be too hasty to assume that we can find quantitative measures similar to those that work in certain narrowly-defined NLU tasks.

5. Assessing overall performance in terms of component tasks

At the end of the day, leaving aside task evaluation, any evaluation that is carried out is focused on the output of the NLG system: the text that is generated. At the same time, it was suggested at the outset that a finer-grained decomposition of the problems involved in NLG is widely accepted as useful, and in Section 2 six aspects of the overall task of generating natural language were identified. Now that the attempts at evaluation in NLG found in the literature have been surveyed, this section suggests a somewhat hybrid approach to evaluation of NLG systems, where we consider the overall output of the system but attempt to assess this in terms of the contributions made by specific component tasks. This is motivated by our belief that it is only through exploring the nature of the component tasks of NLG that an appreciation of what it makes sense to measure about NLG systems as a whole can be gained. Focusing on evaluation as applied to these components is also likely to be more productive than thinking about NLG as an atomic task, because these better reflect the immediate focus of research activity in the field. It could be, of course, that considering evaluation from this perspective will lead to a reassessment of this particular way of decomposing the task

TABLE III. The weather summary for February 1995

The month was cooler and drier than average, with the average number of rain days. The total rain for the year so far is well below average. There was rain on every day for eight days from the 11th to the 18th, with mist and fog patches on the 16th and 17th. Rainfall amounts were mostly small, with light winds.

of NLG: as a result sub-problems in NLG that are not encompassed by this decomposition may be identified, or it may be decided that some of the distinctions made here are inappropriate.

To enable exploration of this idea, in the remainder of this section we look at how aspects of the quality of a generated text by reference to the component tasks identified at the outset of the paper might be evaluated. We do this by examining a particular instance of natural language generation: we explore what might be involved in the generation of weather summaries from automatically collected meteorological data. This scenario has the advantage we alluded to earlier that the data source used exists for reasons quite independent of NLG and thus helps us avoid the kinds of problems that bespoke representations can lead to. Our scenario is not necessarily representative of all NLG tasks, but, we would suggest, it is characteristic enough to allow the exploration of what can go wrong in the different components of an NLG system and hence indicate what an evaluation should highlight.

5.1. The scenario

We have a collection of meteorological data that includes information about temperature and precipitation over an extended period of time, and we want to generate monthly summaries of the weather. This is, in fact, a generation scenario being explored by one of the current authors; at Macquarie University, the Staff Newsletter each month carries a human-authored report on the weather conditions over the preceding month. Table III shows an example human-authored weather summary. At the same time, there are automatic weather stations on campus that collect around two dozen meteorological parameters—ground temperature, air temperature, wind speed, wind direction and so on—sampled at 15 min intervals. Since the human-authored weather summaries do not discuss weather phenomena at a granularity below that of single days, we will assume the corresponding raw data gathered by the automatic weather stations is preprocessed and aggregated to produce what we will call DAILY WEATHER RECORDS; an example is shown in Figure 2.

In the remainder of this discussion we look at how various aspects of the NLG task might play a role in generating something like the human-authored text using these daily weather records as input.

Our concern here is with how performance of an NLG system might be evaluated by looking at how specific tasks within the overall process can contribute to the overall goodness or badness of the generated text. This means that there are a number of potential aspects of evaluation that are ignored by the authors. We list these here in order to clarify our main concerns:

$$\left[\begin{array}{ll}
\text{type: dailyweatherrecord} & \\
\text{date:} & \left[\begin{array}{l}\text{day: 17}\\ \text{month: 05}\\ \text{year: 1995}\end{array}\right] \\
\text{temperature:} & \left[\begin{array}{ll}\text{minimum:} & \left[\begin{array}{l}\text{unit: degrees-centigrade}\\ \text{number: 10.1}\end{array}\right]\\ \text{maximum:} & \left[\begin{array}{l}\text{unit: degrees-centigrade}\\ \text{number: 21.3}\end{array}\right]\end{array}\right] \\
\text{rainfall:} & \left[\begin{array}{l}\text{unit: millimetres}\\ \text{number: 12}\end{array}\right]
\end{array}\right]$$

Figure 2. A daily weather record.

Black box evaluation is too coarse: Given the scenario we have sketched it might appear that we have a situation that would be very amenable to some kind of block box evaluation. For example, since there is an archive of previously written human-authored weather summaries, and a corresponding source of underlying meteorological data, we could simply aim to build a system that would replicate the human-authored texts in a selected training set, and then see how the system performed on a test set. We might evaluate the results by developing some metric to compare the machine generated texts with the human authored texts. The point, however, is that any such metric would be too coarse to give much in the way of useful information unless it is considered how the component tasks in the generation process contribute to the generated texts.

Applications potential is not relevant: We are not concerned here with the question of whether NLG really is the most appropriate solution to this problem: for example, since these summaries are generated only once a month, the construction cost of an NLG system is likely to be far greater than the value it returns.

Unavailable data is not an issue: As Reiter and Dale (1997) point out, it is important to establish before building an NLG system whether it is even possible to generate the required texts given the available data. For this scenario, we are assuming that the requisite input data is indeed available. It turns out that some of the weather summaries written by the human author refer to phenomena for which there is no data available that a program could access; for example, the author in question often mentions whether a particular creek on the University campus is dry. This is information available to the author (perhaps because he walks past the creek each day) that is not available to an NLG system, and so it would be quite inappropriate to evaluate the system on its failure to report on such phenomena.

With these issues dispensed with, we can move on to consider how we might assess the performance of the NLG system at a more fine-grained level.

5.2. Content determination

As noted earlier, content determination is concerned with the selection of the information to be conveyed in a text. There are a number of questions that we might ask of the content determination capabilities of a given NLG system:

- Does it state only information which is true?
- Does it say enough?
- Does it say too much?
- Does it provide information which is inappropriate to the context?

An NLG system might express incorrect or false information in a number of ways:

(1) There might be errors in the underlying information source. In the case of our weather summary scenario, some problem with the automatic weather station might result in incorrect temperature values being recorded. We would clearly not want to say that this was a problem with the NLG system, and so some independent validation of the input data is required.

(2) The system might perform some analysis of the underlying information source in order to determine higher order facts, and it might do this wrongly. It seems reasonable to lay the blame here at the door of the content determination process. One instance of a problem here would be where the data analysis algorithms are incorrect. If the content determination component is tasked with analysing the source data and identifying trends in the data, such as periods over which temperatures have steadily increased, and it fails to identify trends that a human expert would identify, or identified trends that are not present in the data, then this would be a failure in content determination. The problem here is the same as that which arises in an expert system which reaches a conclusion on the basis of rules which do not correspond to expert knowledge.

(3) Similarly, suppose we have a system which is intended to generate different texts for different user groups, perhaps varying the content expressed depending upon the assumed level of experience or knowledge of the audience. If the system's model of the audience is incorrect, or if the theory that determines what content is appropriate in specific circumstances is flawed, then the system may select content that is not appropriate, either saying too much, too little, or the wrong things.

As these last two examples in particular demonstrate, content selection is closest in nature to a standard expert system task: it requires the development of code that models "what an expert would say" in the given circumstances. This characterization of the task of content determination means that it is not particularly linguistic in nature, and this may lead some to consider it not to be part of the task of NLG. Similar concerns might be raised, of course, of any part of an NL understanding process that involves reasoning and inference using world knowledge. In short, if an NLG system says the wrong things, then to put it simply, its theory of what to say may be incorrect.

5.3. Document structuring

Document structuring is concerned with the overall coherence of the generated text. In current NLG systems, there are basically two approaches that are used to address document or text structuring. One is to use SCHEMAS (McKeown, 1985), which can be thought of as scripts or text grammars that stipulate the overall structure of a text. A schema-based approach uses a set of rules that dictate the order and content of a text; as with a sentence grammar, arbitrarily complex structures can be produced given an appropriate set of generating rules. The other popular approach is to use operationalizations of RHETORICAL RELATIONS, these being the basic theoretical constructs in theories such as Rhetorical Structure Theory which dictate how the component parts of a text can be combined. It has often been noted that schemas can be viewed as compilations of conventionalized combinations of rhetorical relations.

In each case the aim is to capture something of the coherence relations that reside in real texts so as to avoid the generation of jumbled texts. The quality of what results clearly depends on how well the underlying model of text structure mirrors reality. In the case of schemas this is quite straightforward: since in general these approaches are quite limited, with a typical text grammar only containing a relatively small number of rules, the outcomes are quite predictable, and it is unlikely that incoherent texts would be generated if the domain is appropriately narrow. Any such problems encountered are easily fixed, in the same way that a simple sentence grammar which over-generates can be repaired.

In a schema-based system, a bad model of text structure might show itself in our current scenario via inappropriate orderings of information. For example, the most common pattern adopted in the human-authored texts is to first talk about the overall temperature and rainfall, and then to go on to mention specific details about rainy spells or spells of extreme temperature if any are there to be reported. However, if a month is particularly unusual in some regard—for example, if it is the hottest September on record—then the human-authored text tends to stray from this pattern, instead commencing by noting this significant fact. If such exceptions are not catered for by the part of the system that is responsible for structuring the text, then we might characterize this as a case of bad DOMAIN COMMUNICATION KNOWLEDGE.

In the case of approaches based on rhetorical relations, there is generally a considerable increase in the variety of textual structures that can be generated; this brings with it a lessening of the predictability of outcomes and an increase in the chances of incoherent texts being generated. Work by Hovy (1993) in this area has suggested that the incorporation of other aspects of coherence, using notions such as focus or theme development, may be required in order to ensure that the generated texts remain coherent; otherwise the possible textual structures that can be imposed over a given set of informational elements may be quite underconstrained.

In both cases, the resulting coherence of the texts generated depends on the extent to which the model of text structure used—either embodied in the patterns over elements permitted by the schemas, or the combinations of elements permitted by the rhetorical definitions—reflects the true nature of coherence in text.

5.4. Lexicalization

Lexicalization is concerned with choosing the right word or phrase. This can go wrong in a number of ways. The most obvious and trivial case is where words are incorrectly associated with underlying concepts: a system suffering such a fault might, for example, state that the temperature today is warmer than it was yesterday when in fact the temperature has decreased.

More sophisticated problems in lexicalization arise where a small set of canonical concepts are used, but these have to be mapped into a larger set of words and phrases with the choices being determined by collocational factors. For example, the underlying conceptual base might express a notion of increase or decrease over both temperature and rainfall; however, although we can use terms like greater and higher to express an increase in either case, the terms warmer and wetter are specific to one or the other. An incorrect representation of the appropriate collocational data will lead to problems here. We might reasonably see this as a problem with the theory of lexicalization embodied in the system. It should be noted that such problems can only arise relative to a chosen set of underlying concepts, and so any attempt to evaluate here must take account of this dependence in some way. If the appropriate lexical distinctions are already made in the underlying concept set then there is no real possibility of error at the level of lexicalization; however, if an inappropriate conceptual structure is built from some more basic data, then we might see this as a problem with the system's content determination. Once more we see that the distinction between these two aspects of NLG is not clear cut, and clearly depends very much on the representational assumptions embodied within the system.

5.5. Aggregation

The lack of appropriate aggregation strategies is perhaps the most obvious source of disfluent texts in simple generation systems. Although many early NLG systems assumed that the input elements of information to be expressed corresponded one-to-one to sentences, this is not in general true; any such assumption easily leads to choppy texts consisting of sequences of short sentences. So, for example, we want to generate something like the second of the two following texts rather than the first:

(1) The month was cooler than average.
The month was drier than average.
There were the average number of rain days.
(2) The month was cooler and drier than average, with the average number of rain days.

This area has received more attention in recent years as researchers move away from hand-crafted knowledge bases to those developed from some other source where the granularity may not be sentence-sized. There is real scope for evaluation here, in the sense that different aggregation strategies may be adopted: the same collections of informational elements can be put together into sentences in different ways. At the same time, there are relationships between aggregation and lexicalization and referring expression generation that confuse the issues: lexicalization (for example, in the choice of which element to express as the head verb of a sentence, and which verb to use) dictates the sentential complementation structure and thus may restrict aggregation possibilities; and the construction of referring expressions itself permits aggregation by allowing sites where information may be folded in (for example, as a post-modifying relative clause). For these reasons, aggregation, lexicalization and referring expression generation are sometimes seen as the domain of a distinct SENTENCE PLANNING task; the nature of the interactions between the three phenomena are very much ill-understood, and so it may be somewhat premature to attempt to evaluate approaches to any one of the three in isolation.

5.6. Referring expression generation

We have already noted in our survey of existing approaches to evaluation a number of cases where different aspects of referring expression generation algorithms have been evaluated. Some aspects of this problem lie properly in the domain of content determination (and so here we see another hint that the elements in this catalog of NLG phenomena are not necessarily distinct is seen): in particular, when an entity is introduced into a discourse, the properties used in this initial reference need to be selected in line with similar considerations to those that apply in the selection of content for a text as a whole.

The scope for evaluation in the generation of subsequent anaphoric references is perhaps easier to identify. In particular, if a system is able to choose to use pronouns to refer to entities, there is a tension between producing unambiguous text and producing text that does not contain redundant information. Many of the problems that plague approaches to pronoun resolution have analogues here, although evaluation is again more difficult in the case of NLG: in evaluating a pronoun resolution algorithm, we can always compare against a marked-up version of a text that contains co-reference annotations, but pronoun generation can only be evaluated by comparison with how people might use pronouns in similar circumstances, or by asking subjects if the machine-generated usages are ambiguous: in each case, there is considerable scope for inter-subject variability.

We have talked here simply of pronouns as the paradigm instances of anaphoric referring expressions, but all the same issues arise for any forms of anaphoric reference, be they proper names, reduced definite noun phrases or other forms. In the HAM-ANS system (Jameson & Wahlster, 1982), the appropriateness of anaphoric referring expressions generated by the system was tested at run-time by having the NLU component of the system determine whether it could disambiguate the references successfully: this self-evaluation is of course limited by the characteristics of the NLU component.

5.7. Surface realization

Surface realization is concerned with taking some specification of the content of a sentence and realizing this as a morphologically and grammatically correct sentence. This is the one area of NLG where there are a number of off-the-shelf components available for the researcher to use, and so one might expect that some evaluation of these components against each other might be possible. But this is not so, for a number of reasons.

The most important of these, at least from a theoretical perspective, is that no two systems share a common view as to what surface realization really involves; as a result, each system tends to adopt a different model of what the appropriate input to surface realization should be. Although terms like sentence specification, realization specification and sentence plan are in quite common use, they are typically taken to mean different things by different people. Thus, some realization systems expect a very high level semantic specification as input, with the realization component performing quite sophisticated reasoning as to how this content should be expressed; other systems expect input specified in terms of a broad grammatical structure that already dictates what the major constituents of the sentence should be; and, in the extreme case, some systems do no more than linearize and add appropriate morphological inflections to what is to all intents and purposes a completely specified syntactic structure.

For any researcher evaluating an existing realization component, a key question is the grammatical coverage of that system: does it generate all the required surface forms? Again, comparison of systems in this regard is made difficult by the different expectations of what the input should be like.

There is another possibility of failure, of course: a surface realization component might actually generate ungrammatical sentences.

6. Conclusion

In this paper, we have tried to give a picture of what is involved in NLG, looked at some previous attempts to perform evaluation in the context of the NLG systems and techniques, and suggested that evaluation in NLG faces difficulties not present in the popular paradigms of evaluation found in some area of NLU. There are, as yet, no clear answers as to how NLG systems should be evaluated; however, as a starting point we have suggested that a breakdown into the phenomena that NLG needs to consider provides one way of clarifying the questions that need to be asked. We have provided above some starting points in this regard, but as we have noted, even the catalogue of phenomena is not without controversy.

In the interim, we will no doubt see increasing attempts to evaluate work in NLG; these efforts are to be praised, and will surely lead to insights that will help to shape the debate. As we have seen, evaluating an NLG system demands resources such as time and human subjects, as well as good sources of data. The requirements for evaluation need to be taken into account right from the start of a project, and arguably even in the stages when a research project is being chosen. If one is about to embark on a project that seeks to bring advances in the techniques of NLG and yet no clear path to an evaluation of the results can be seen, one may want to ask whether one should be doing the project at all. A closer questioning and scrutiny at these stages will help to firm up the questions that need to be asked from inside the NLG task, so that we may move closer to a greater understanding and a broader consensus as to what evaluation in the context of NLG really means.

The authors would like to thank the members of their respective research groups for discussions which have helped develop the ideas and intuitions expressed in this paper. Particular thanks are due to Sarah Boyd, Cécile Paris and Ehud Reiter for motivating some of the questions raised in this paper.

References

Acker, L. & Porter, B. (1994). Extracting viewpoints from knowledge bases. *Proceedings of the Twelfth National Conference on Artificial Intelligence*. pp. 547–552. AAAI Press/MIT Press, Cambridge, MA.

Arnold, D., Humphreys, R. L. & Sadler, L. (1993). Special issue on evaluation of MT systems. *Machine Translation* **8**, 1–26.

Arnold, D., Sadler, L. & Humphreys, R. L. (1993). Evaluation: an assessment. *Machine Translation* **8**, 1–24.

Brandow, R., Mitze, K. & Rau, L. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing and Management* **31**, 675–685.

Carletta, D. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics* **22**, 249–254.

Carter, E. (1996). Quantitative Analysis of hypertext Generation and Organisation Techniques. PhD thesis, University of Edinburgh.

Cawsey, A. (1992). *Explanation and Interaction*, MIT Press, Cambridge, MA.

Cohen, P. R. (1996). Getting what you deserve from data. *IEEE Expert* **October**, 12–14.

The First International Workshop on Controlled Language Applications, Centre for Computational Linguistics, Katholieke Univ. Leuven, Leuven, Belgium, March 26–27, 1996.

Dale, R. (1992). *Generating Referring Expressions: Building Descriptions in a Domain of Objects and Processes*. MIT Press, Cambridge, MA.

Dras, M. (1997). Representing paraphrases using STAGs. *Proceedings of ACL-EACL97*. Madrid, Spain.

Flesch, R. F. (1948). A new readability yardstick. *Journal of Applied Psychology* **32**.

Gaizauskas, R., Cunningham, H., Wilks, Y., Rodgers, P. & Humphreys, K. (1996). GATE: An environment to support research and development in natural language engineering. *Proceedings of the 8th IEEE International Conference on Tools with Artificial Intelligence*, Toulouse, France.

Hovy, E. (1993). Automated discourse generation using discourse structure relations. *Artificial Intelligence* **65**, 341–386.

Johnston, R. D. (1996). Beyond intelligibility—The performance of text-to-speech synthesisers. *BT Technology Journal* **14**.

Jameson, A. & Wahlster, W. (1982). User modelling in anaphora generation: ellipsis and definite description. *Proceedings of ECAI-82*, Orsay, France, pp. 222–227.

Jordan, P. W., Dorr, B. J. & Benoit, J. W. (1993). A first-pass approach for evaluating machine translation systems. *Machine Translation* **8**, Kluwer Publishers, Dordrecht, Netherlands, pp. 49–58.

Knight, K. & Chander, I. (1994). Automated postediting of documents. *Proceedings of the Twelfth National Conference on Artificial Intelligence*. pp. 779–784. AAAI Press/MIT Press, Cambridge, MA.

Lester, J. C. & Porter, B. W. (1997). Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics* **23**. MIT Press, Cambridge, MA, pp. 65–102.

Levine, J. & Mellish, C. (1995). The IDAS user trials: Quantitative evaluation of an applied natural language generation system. *Proceedings of the Fifth European Workshop on Natural Language Generation*. pp. 75–94, Rijks Universiteit Leiden.

Lorge, I. (1959). *The Lorge Formula for Estimating the Difficulty of Reading Materials.* Teachers College Press, Columbia University, New York.

McKeown, K. (1985). Discourse strategies for generating natural-language text. *Artificial Intelligence* **27**, 1–42.

Mann, W. & Thompson, S. (1987). Rhetorical structure theory: description and construction of text structures. In *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics* (G. Kempen, ed.), Dordrecht: Nijhoff.

Mellish, C. & Evans, J. R. (1989). Natural language generation from plans. *Computational Linguistics* **15**, 233–249.

Meteer, M. & McDonald, D. (1991). Evaluation for generation. In *Natural Language Processing Systems Evaluation Workshop* (J. G. Neal & S. M. Wlater, eds.), pp. 127–131, NY. Rome Laboratory.

Minnis, S. (1998). Constructive machine translation evaluation. *Machine Translation* **8**, 67–76.

Moore, J. (1991). Evaluating natural language generation facilities in intelligent systems. In *Natural Language Processing Systems Evaluation Workshop* (J. G. Neal & S. M. Wlater, eds), pp. 133–140, NY. Rome Laboratory.

Nielsen, J. (1990). Evaluating hypertext usability. In *Designing Hypermedia for Learning* (D. Jonassen & H. Mandl, eds), Springer Verlag, Berlin/Heidelberg.

Pierce, J. R. & Carroll, J. B. (1966). *Language and Machines—Computers in Translation and Linguistics (ALPAC Report)*, Washington DC.

Reiter, E. & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering* **3**, 57–87.

Robin, J. (1994). Revision-based generation of natural language summaries providing historical background—corpus-based analysis, design, implementation and evaluation. PhD thesis, Columbia University. (Technical Report CUCS-034-94).

Salton, G., Singhal, A., Mitra, M. & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing and Management* **33**, 193–207.

Shiwen, Y. (1993). Automatic evaluation of output quality for machine translation systems. *Machine Translation* **8**.

Siegel, S. & Castellan, N. J. Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York.

Sparck Jones, K. & Galliers, J. (1996). *Evaluating Natural Language Processing Systems.* Springer Verlag, Berlin/Heidelberg.

Sundheim, B. M. & Chinchor, N. A. (1993). Survey of the message understanding conferences. *Human Language Technology: Proceedings of a Workshop*, Morgan Kaufmann, Palo Alto, CA.

Swartout, W. R. (1983). XPLAIN: A system for creating and explaining expert consulting programs. *Artificial Intelligence* **21**. Elsevier, Amsterdam, pp. 285–326.

Thompson, H. (1977). Strategy and tactics: A model for language production. *Papers from the Thirteenth Regional Meeting of the Chicago Linguistics Society* (W. A. Beach, S. E. Fox & S. Philosoph, eds) Chicago, 651–668.

Yeh, C.-L. & Mellish, C. (1997). An empirical study on the generation of anaphora in Chinese. *Computational Linguistics* **23**.