Towards the Evaluation of Natural Language Generation

Robert Dale[†] **and Chris Mellish**[‡]

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

Abstract

What role should evaluation play in the development of natural language generation (NLG) techniques and systems? In this paper we begin by characterising the NLG task, and in the light of this characterisation we discuss how the problems of NLG evaluation are different from the problems of evaluating natural language *understanding*. We then suggest a way forward for evaluation of work in natural language generation that proceeds by looking more closely at the component problems that are addressed in research in the field.

1. Introduction

In this paper, we consider how evaluation can play a role in the development of natural language generation (NLG) systems and techniques. We suggest that there are differences between the kinds of evaluation currently carried out in the natural language understanding community and the kinds of evaluation that are relevant to work in NLG, and that in many regards it is still unclear what evaluation means in the context of NLG. As a way of making progress on clarifying the issues here, we suggest a decomposition of the problem of NLG that may make it easier to see how evaluation can be carried out.

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 we look at the question of what it might mean to carry out evaluation in the context of NLG. We distinguish the evaluation of systems from the evaluation of their underlying theories, and distinguish both of these from task evaluation. Section 4 we then identify a number of questions that anyone attempting evaluation in the context of NLG must face.

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 and considering how these might be subjected to evaluation. Finally, in Section 6 we draw some conclusions.

2. What is Natural Language Generation?

Natural language generation (NLG) 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. Individual pieces of work in NLG are however generally concerned with what are viewed as component parts of this overall problem. The most common decomposition of the 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. This distinction often finds itself manifested in an architectural distinction within implemented systems, which in many cases consist of a TEXT PLANNING component and a LINGUISTIC REALISATION component. Although this breakdown of the task is widespread—and, in fact, in more recent years a tripartite distinction into TEXT PLANNING, SENTENCE PLANNING and LINGUISTIC REALISATION has achieved something of a consensus in the field (see Reiter(1994))—when we look at specific details of these systems it is less clear that they agree on the nature of the component tasks. There is no universally accepted view of the architecture of the generation task. However, 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 some differences of view as to how these are best combined into components into a system.

Reiter and Dale (1997) categorise current work in the field as being concerned with six main categories of problems, as follows:

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 should the text be organised 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 organistion for that information.

Lexicalisation: 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, as we noted above, 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 should information 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 Realisation: determining how the underlying content of a text should be mapped into a sequence of grammatically correct sentences that express the desired meaning. A generally held view is that the same propositional content can often be realised 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, all are issues which must be addressed one way or another by a complete NLG system.[1] 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.

Natural language generation is then, in some sense, the inverse of natural language understanding (NLU): whereas in NLU the ultimate concern is to map from text to some representation of the meaning embodied in that text, NLG is concerned with going in the other direction and mapping from some representation of meaning to a text that embodies that meaning. 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 what we characterised above as surface realisation—only one small (but still very complex) part of what is involved in NLG. 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.

Natural language generation and natural language understanding are thus the two halves of the puzzle of natural language processing. This symmetry has not so far been reflected in the balance of effort expended in the research arena: although interest in NLG has increased considerably in the last 10–15 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 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. We will return to this point below; for the moment, we need to stress 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 much 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 with larger multi-sentential discourses.

3. What Does It Mean to Evaluate NLG?

It is clear that some notion of evaluation is important for NLG, as it is for NLU. It is necessary to be able to demonstrate progress (both to fellow researchers and to potential funders), and the field needs to be able support possible

[1] Of course, addressing an issue doesn't necessarily mean providing an acceptable solution to it. Building a complete system is a large and complex task, and as a result in many systems the the solutions adopted for specific problems may be relatively trivial because they are hard-wired into the operation of the system. This is legitimate when the particular problem in question is not a focus of the work being carried out, and is no different to the situation in NLU.

users of the technology by helping make them decisions about whether it is good enough for their purposes. Specific areas of NLU appear to be devising useful evaluation metrics, and the current empiricist emphasis of much NLP is justifiably asking *which techniques work how well?* Against this background, it is not surprising that researchers in NLG are beginning to take evaluation seriously. In 1990, a AAAI workshop was held on the theme of Evaluating NLG Systems, and there has been a noticeable increase in empirical work since then; a common request from reviewers of NLG submissions to journals and conferences is for some material on evaluation to be included. However, there is no extant wisdom as to how NLG should be evaluated: we all ask the question, but no one is quite sure what the answer is.

As many commentators on evaluation have pointed out (e.g. Sparck Jones and Galliers 1995), evaluation can have many objectives and can consider many different dimensions of a system or theory. We consider the evaluation of NLG techniques to break down into the following main categories, with the order reflecting 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 extension of an implementation to a new domain, will be productive.

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 application potential for 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).

The last of these needs to consider 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. For instance, how cost effective is the system? what is its impact on the existing (e.g. social) environment? it the system maintainable? is it fast enough, robust enough? how does it compare to its rivals (not necessarily all other NLG systems)? We will largely ignore these issues here, focussing on the second category of evaluation; see (Reiter and Dale 1997) for some discussion of issues and alternatives to the use of NLG techniques that should be considered in any applications context.

4. Issues and Problems in Evaluating Natural Language Generation

Evaluation methods for NLG are still very much in their infancy, and many researchers are feeling their way in terms of how to evaluate their systems. This section attempts to summarise some of the general problems that arise in evaluating NLG, as revealed by the small number of evaluative exercises reported in the literature. Each of these problems is to some extent more serious for NLG than for NLU.

4.1. What Should the Input be?

Whereas it is reasonably clear what an 'input text' is, 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 tasks and applications seem to require knowledge and goals of different kinds, and even a relatively superficial scan of the literature reveals that a very broad range of assumptions have been adopted by researchers in the field. It is not always easy to judge whether an NLG system is actually 'cheating' by including in its input very direct guidance about how to handle problems that the NLG system will have (for instance, whether a system whose input is supposedly language-independent is actually anticipating the structures that will be generated in a specific language). Natural language analysis systems at least largely agree on the sequential structure and tokenisation of written natural language texts, which permits successful architectures like GATE (Gaizauskas *et al* 1996), where the results of processing can be expressed as annotations on text spans. With NLG, there is no such agreed basis.

The only way to be sure of having realistic input is probably to take input that is provided by another system developed for purposes unrelated to NLG. This was achieved by Lester and Porter (1997) by making use of an independently developed knowledge base in the domain of biology; in certain domains (such as those the reporting of stock market behaviour or weather forecasting), numerical input may be available from independent sources. But experience shows that an application that wants to use NLG often needs to be built with this in mind if it is to produce material from which NLG is possible (Swartout 1983; Mellish and Evans 1989).

4.2. What Should the Output be?

The second 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 satisfactorily achieve more or less the same goals. Evaluation therefore has to somehow consider the *quality* of the output; notions like 'accuracy' (does the text convey the information it is supposed to) and 'fluency' (does the text present the information in a readable manner) are two informal measures of

text quality, but there is a general uncertainty about what the key dimensions of measurement here 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, and may hinge on other aspects of the context of use that have little to do with NLG.

4.3. What to Measure?

It is not always obvious what one needs to measure about the performance of an NLG system or what measurable attributes best indicate this. For instance, in some owork on the automatic generation of hypertext, Carter (1996) 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. In exploring different algorithms for the generation of anaphoric referring expressions in Chinese, Yeh and Mellish (1997) measured the extent to which the generated references agreed with those selected by humans. However, 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. Researchers in NLG have much to learn here from their colleagues in experimental sciences like psychology.

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 to Use?

For many NLG tasks we have no idea how well it is possible to do how easily, which means that absolute performance figures are not very useful. For instance, in some user trials testing of the IDAS system (Levine and Mellish 1995), an attempt was made to determine how many generated hypertext nodes visited by the users were considered irrelevant to their needs; the results were that '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 method was used in the evaluation work of Acker and Porter (1994) 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 and Mellish (1997).

Sometimes upper and lower bounds can be determined by analysis or simple experimentation. Knight and Chander (1994) pointed out that, given a text with articles missing, 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% faced with the same task. This puts Knight and Chander's achievement of 78% into perspective.

4.5. How to Get Adequate Training and Test Data?

As Robin (1994) points out, such is the complexity of NLG that many people have focussed on particular subtasks which have been evaluated using hand-coded (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 it 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 been resulted in a system that was over-tuned to this data set. A separate evaluation involving comparison to humans on different data was essential to discount this possibility.

4.6. How to Handle Disagreement among Human Judges?

In evaluating the output of NLG, we must face the problem that humans will not always agree about subjective qualities like 'good style', 'coherence' and 'readability'. 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 found significant disagreement about what the preferred referring expressions should be. Faced with results of this kind, a natural strategy is to invesigate 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 and Castellan 1988) is very useful for this.

5. Ways Forward

In the preceding section, we have raised a number of questions that arise in any evaluation exercise, and along the way we have surveyed, albeit briefly, the various approaches to evaluation that have been attempted in work on NLG. Clearly the use of some notion of evaluation in NLG is intuitively appealing. However, it is our view that 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, 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. Similarly, the evaluation of a part-of-speech tagger without some independent motivation for the part-of-speech tags to be used is questionable. Outside of these rather narrowly defined contexts, evaluation becomes much less easy to manage in NLU. For example, there is no obvious way of determining whether one has extracted the 'meaning' from a given text—unless one chooses to take the target data records in a MUC-style system as representations of meaning. An argument can be made for such a view, but the position is not a widely held one.

It is hardly surprising, then, that it is unclear how we might carry out evaluation in the context of natural language generation. Because of some of the asymmetries between NLU and NLG that we identified earlier, 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. It is sometimes said that whereas NLU is fundamentally about HYPOTHESIS MANAGEMENT—ruling out possible interpretations at various levels of processing—NLG is more fundamentally concerned with CHOICE: choosing between different ways of doing things in a context dependent manner. 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.

At the end of the day, leaving aside task evaluation, any evaluation we carry out is focussed on the output of the NLG system: the text that is generated. This, however, is fraught with problems if viewed as a 'black box' evaluation task. We suggested at the outset that a finer-grained decomposition of the problems involved in NLG is widely accepted as useful, and in Section 2 we identified six aspects of the overall task of generating natural language. Our view is that, if progress is to be made here, we have to switch to a more fundamentally "glass box" perspective. That is, we believe that only through considering the nature of the *component tasks* of NLG can we gain an appreciation of what it makes sense to measure about NLG systems as a whole. Focussing 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 reseachers. It could be, of course, that considering evaluation from this perspective will lead us to reassess this particular way of analysing the task of NLG: we may as a result identify sub-problems in NLG that are not encompassed by this decomposition, or we may decide that some of the distinctions made here are inappropriate.

Below, we offer some thoughts and comments on evaluation from the perspective of the decomposition that we have presented. We focus on *how these components can go wrong* as a way to start thinking about what evaluation might highlight.

5.1. 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 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: there might be errors in the underlying information source; it might select true facts to be expressed, but express them using misleading language or even the wrong words; or it might perform some analysis of the underlying information source in order to determine higher order facts, and it might do this wrongly. The first of these problems cannot be laid at the door of the NLG system; the second may indeed be a problem for the NLG system, but is not concerned with content determination.

The third case is indeed a place where a content determination process could fall down: this would be an instance of the data analysis algorithms being incorrect. So, for example, imagine that we have a system which generates text from some numeric source—this might be stock market prices or meteorological data. Such a system might be tasked with identifying trends in the data: periods over which prices or temperatures have increased. A failure to identify trends that a human expert would identify, or the incorrect identification of trends that are not present in the data, would then be a failure in content determination.

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 examples 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 characterisation 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 NLU process that involves reasoning and inference using world knowledge. There may be things we can learn here from work in the evaluation of expert systems.

5.2. 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 operationalisations 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. In such approaches, the rhetorical relations are typically modelled as plan operators in the traditional AI sense. Although this is a simplification, one can think of the first approach as being in some sense top-down, and the second as being bottom-up; it has often been noted that schemas be viewed as compilations of conventionalised 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 sentence grammar which over-generates can be repaired.

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.

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 relation definitions—reflects the true nature of coherence in text. We therefore have to be very careful about what exactly is being evaluated here: is it the underlying theory of coherence, or is it the implementation of that theory in a particular system? In the latter case, for example, the issues often come down to questions such as determining what it would mean to elaborate upon a topic given some knowledge base as input. The theory-based definitions of these concepts are generally insufficiently specific to provide hard answers here.

5.3. Lexicalisation

Lexicalisation 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. Such errors, we might argue, are not errors of lexicalisation but mistakes in the lexical data source used.

More sophisticated problems in lexicalisation 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; but it should also be noted that the 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. Again, we are faced with determining whether a deficiency in performance here is due to a deficiency in the underlying conceptual base (and perhaps the theory it embodies) or the context-dependent mapping from that base to lexical items.

5.4. 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; when faced with real data sources, any such assumption easily leads to choppy texts consisting of sequences of short sentences. 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 lexicalisation and referring expression generation that confuse the issues here: lexicalisation (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, lexicalisation 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.

5.5. 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): 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 as 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 in the NLU literature 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 and 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.

5.6. Surface Realisation

Surface realisation is concerned with taking some specification of the content of a sentence and realising 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 realisation really involves; as a result, each system tends to adopt a different model of what the appropriate input to surface realisation should be. Although terms like sentence specification, realisation specification and sentence plan are in quite common use, they are typically taken to mean different things by different people. Thus, some realisation systems expect a very high level semantic specification as input, with the realisation 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 linearise and add appropriate morphological inflections to what is to all intents and purposes a completely specified syntactic structure.

For any researcher evaluating an existing realisation 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.

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 NLG systems and techniques, and suggested that evaluation in NLG faces difficulties not present in the popular paradigms of evaluation found in some areas 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 catalog of phenomena is not without controversy. Nonetheless, it is our view that a decomposition of this kind is essential for making progress on the evaluation of NLG techniques and systems. It is only by moving towards a clearer view of the parameters and dimensions of the NLG task that we will establish appropriate grounds for comparison; without such common ground, we are comparing apples with oranges. Our catalog of phenomena and the considerations that arise from each should only, therefore, be seen as a very provisional agenda. The

real work is in refining these ideas so that we may gain a better appreciation of the issues.

In the interim, we will, no doubt, see increasing attempts to evaluate work in NLG; this effort is 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.

7. Bibliography

Acker, L. and Porter, B. (1994) "Extracting viewpoints from knowledge bases". In *Procs of the Twelfth National Conference on Artificial Intelligence*, pages 547–552. AAAI Press/MIT Press.

Carletta, J. (1996) "Assessing Agreement on Classification Tasks: The Kappa Statistic", *Computational Linguistics*, Vol 22, No 2.

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

Cohen, P. R. (1996) 'Getting what you deserve from data", *IEEE Expert*, October 1996.

Gaizauskas, R., Cunningham, H., Wilks, Y., Rodgers, P. and 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*, **63**: 341–386.

Jameson, A. and Wahlster, W. (1982) User modelling in anaphora generation: ellipsis and definite description. In *Proceedings of ECAI-82*, pages 222-227.

Knight, K. and Chander, I. (1994) "Automated postediting of documents". In *Procs of the Twelfth National Conference on Artificial Intelligence*, pages 779–784. AAAI Press/MIT Press.

Lester, J. C. and Porter, B. W. (1997) "Developing and Empirically Evaluating Robust Explanation Generators: The KNIGHT Experiments", *Computational Linguistics* Vol 23, No 1.

Levine, J. and 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*, pages 75–94, RijksUniversiteit Leiden, 1995.

Mellish, C. and Evans, R. (1989) "Natural Language Generation from Plans", *Computational Linguistics* Vol 15, No 4, pp233-249.

Reiter, E. (1994) Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the 7th International Workshop on Natural Language Generation*, pages 163–170.

Reiter, E. and Dale, R. (1997) "Building Applied Natural Language Generation Systems", *Natural Language Engineering* Vol 3, Part 1, pp57-87.

Robin, J. (1994) "Revision-Based Generation of Natural Language Summaries Providing Historical Background: Corpus-Based Analysis, Design, Implementation and Evaluation", PhD thesis, Columbia University, 1994; also available as Technical Report CUCS-034-94.

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

Sparck Jones, K. and Galliers, J. (1995) *Evaluating Natural Language Processing Systems*. Springer Verlag.

Swartout, W. R. (1983) "XPLAIN: A System for Creating and Explaining Expert Consulting Programs", *Artificial Intelligence* Vol 21.

Thompson, H. (1977) "Strategy and Tactics: A Model for Language Production", Papers from the Thirteenth Regional Meeting of the Chicago Linguistics Society, Chicago.

Yeh, C-L. and Mellish, C. (1997) "An Empirical Study on the Generation of Anaphora in Chinese", *Computational Linguistics*, Vol 23, No 1.