Computer Speech and Language (1998) **12,** 249–262
Article No. la980100

Evaluation in language and speech technology

Robert Gaizauskas
Department of Computer Science, University of Sheffield, U.K.

> I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be.
> Lord Kelvin, *Popular Lectures and Addresses*, (1889), vol. 1, p. 73.

1. Introduction

A little over a decade ago it was common to read descriptions of natural language processing systems that discussed the theoretical underpinnings of the system, supplied an architecture diagram and perhaps illustrated the behaviour of the system on some carefully chosen texts. While this sort of activity suggested that automatic language processing might indeed be possible, it was pre-scientific in the sense of the Kelvin quotation above: nothing was measured, because there were no common measures and no shared data. As a consequence, systems and approaches could not be precisely compared and results could not be replicated.

Now all this has changed, and while it is perhaps premature to claim that automatic language processing has emerged from the shadowy valley of pre-science onto the sunny uplands of mature science, where there is a general consensus about measures and empirical methodology, there has been remarkable progress. It is now rare to read a paper, either concerning an algorithm for language processing or concerning a system for some applied task, which does not contain a section on quantitative evaluation. Resources, both annotated and unannotated, in a range of languages are now available from language resource providers like LDC and ELRA.[1] Comparative evaluation exercises in which all comers may run their systems on blind test data have emerged in areas ranging from speech recognition to word sense disambiguation.

No doubt the reasons for this sea change are complex, and it would be foolish to claim a full understanding of the process. But amongst the factors leading to this change one can cite: the appearance of large, shared, electronic corpora of spoken and written language; frustration of researchers with the “pre-scientific” state of the field; demand of consumers of the research—especially funding agencies in the U.S.—for

[1] Linguistic Data Consortium—http://www.ldc.upenn.edu/; European Language Resources Association—http://www.icp.grenet.fr/ELRA/home.html

0885–2308/98/040249+14 $30.00/0

measurable results; higher speed, lower cost computer processing and data storage devices.

This special issue is a reflection of these wider trends. It has come about as a consequence of a workshop on "Evaluation in Speech and Language Technology" held in Sheffield in June, 1997. The workshop was held under the aegis of the British Speech and Language Technology Club (SALT) and sponsored by the U.K. Department of Trade and Industry and the Engineering and Physical Sciences Research Council. As such, it was part of an ongoing series of SALT workshops which have served as focal events for the U.K. speech and language communities for quite a few years. While these events have never been closed to non-British participants, they have usually attracted a primarily British audience. This event, however, was strikingly different: participants came from many European countries (France, Germany, Denmark, The Netherlands) and a significant contingent came from the United States. Now, this could simply have been the consequence of using the Internet to promote the workshop—reaching an international pool of potential attendees. But I think not: the real reason, confirmed by talking to many of the attendees, is that evaluation in speech and language technology has become a central issue. If anything, this trend has continued since the workshop—witness the phenomenal scale of the 1st International Conference on Language Resources and Evaluation held at Granada in May of 1998.

Following the Sheffield SALT workshop, at which a general desire was expressed for a form of publication less ephemeral than the workshop proceedings (SALT, 1997), a call was announced for this special issue, aimed at both participants of the workshop and at others interested in the theme of evaluation in language and speech technology. The response was overwhelming and this issue is the result.

2. The role of evaluation in speech and language technology

Undeniably, evaluation has become a central topic in speech and language (S & L) research. But what is meant by "evaluation"? Any such general concept, especially one laden with associated emotive and rhetorical potential, needs to be carefully unpacked if it is to play a genuinely useful role in advancing research. This activity of clearing the complex terminological underbrush surrounding the concept of evaluation has been carried out extensively by Sparck Jones & Galliers (1996), Crouch, Gaizauskas & Netter (1995) and EAGLES (1995). These analyses, given the complexity of the topic, cannot even be adequately summarized here. However, a few conceptual distinctions must be mentioned, briefly, to set this volume in context. In particular, it is important to distinguish different perspectives one may adopt towards evaluation in S & L technology, and from the primary perspective adopted in this issue, that of technology evaluation, to clarify what are appropriate subjects for evaluation.

2.1. Perspectives on evaluation

First, what is meant by evaluation depends on the *perspective* one adopts towards S & L technology. As Hirschman's paper in this issue so clearly observes, there are three major stakeholders in the evaluation process: users, researchers and funders. These stakeholders have differing, though overlapping, perspectives.

If one's perspective is that of a potential *user* concerned to accomplish a task for which the technology is simply a tool then one must consider evaluating a system in its actual operational context or environment (the system + environment forming what Sparck Jones & Galliers (1996) call the *setup*). Thus, it is no use having a brilliant piece of technology if it cannot help you to do want you want to do in the actual context of use in which you will deploy it. Of course one should not overlook the fact that new technology can radically alter the environment in which it is deployed, or may require such alteration if it is to succeed. One should also not overlook the fact that any system deployed in a working environment will quite possibly contain non-language processing components as well as language processing components (Sparck Jones & Galliers (1996) n-system and 1-system); for example, a spoken language interface to a database will contain both speech recognition and database components. Thus, assigning credit or blame to solely the language processing portion of the overall system is tricky, since there may be faults in both components, as well as awkward interaction effects. Nevertheless, the role of evaluation in user-centred evaluation is to enable users to see whether and how S & L technology may be of use to them.

If one's perspective is that of a *researcher* or *technology developer* who wishes to understand better and improve the techniques and models he uses then different measures and evaluation scenarios will be appropriate, though, of course, user evaluations may provide relevant information to technologists, just as technology evaluations may provide relevant information to users. For the researcher/technologist, evaluation plays the role of a crucial part of the empirical method: a system embodies a hypothesis about how certain input may be transformed into a certain output, and the evaluation is hypothesis testing. Based on the outcome, the hypothesis—or the implementation, algorithm or theory on which it is based—may be modified, the system revised and further testing undertaken (see Walker & Moore (1997) for a further discussion of the role of evaluation as a component of the empirical method in language research, and Cohen (1995) for a more general discussion of empirical methods in AI research).

Finally, if one's perspective is that of a *funder* trying to decide whether R & D money has been well spent then evaluation will mean something else again and may involve a complex calculation that takes into account technology and user evaluation, as well as broader issues such as social impact. The role of evaluation here is to account for and to plan for the allocation of limited resources to achieve valued social and technological goals.

In this special issue the perspective adopted is primarily that of the researcher/technologist. This reflects the interests of the readership of *Computer Speech and Language* and the character of the bulk of submissions to the issue. Issues facing funders evaluating S & L programmes are best left to wiser heads than ours; and issues in user-centred evaluation, while not totally ignored herein (see, particularly, the paper by Walker *et al.* in this issue), do not feature centre-stage (see EAGLES (1995) for discussion and examples of user-centred evaluation). We make no apology for this focus. As this issue shows, the problems facing technology evaluation are more than ample in scope and bear fruit when pursued.

2.2. "Technology" evaluation

Given that the focus of this issue is on what is generally, but somewhat inappropriately, termed "technology evaluation", it is useful to make several further distinctions. Following Crouch, Gaizauskas & Netter (1995) one can distinguish language processing *tasks* from language processing *systems*. Systems carry out tasks or functions, and while tasks may be broken down into subtasks, and systems into subsystems, there is no requirement that there be an isomorphism between these decompositions. That is, task structure and system architecture need not be the same, and designers of evaluations need to take this into account. The task, specified independently of any particular system or class of systems, is the appropriate subject of evaluation.

Further, one can distinguish what Crouch, Gaizauskas & Netter (1995) call, in somewhat cumbersome terminology, *user-visible* from *user-transparent* tasks. The former are tasks where both input and output have some functional significance for the user; the latter are tasks where the input and/or output do not have such significance. So, for example, both machine translation (viewed as text in, text out) and parsing are language processing tasks. The former is user-visible, while the latter is not, since most users have no interest in parser output (parse trees or dependency structures). Usually, a user-transparent task is a subtask of a higher level user-visible task.[2]

Both user-visible and user-transparent tasks are suitable candidates for defining evaluation scenarios. Note that an evaluation with respect to a user-visible task is not a user-centred evaluation, as no specific environment is assumed in performing the evaluation. The criteria applied will be purely *intrinsic* (in Sparck Jones & Galliers (1996) terms): how well does the system meet some objectively defined functional specification of the task it is intended to carry out on some specific test data set? This contrasts with a user-centred evaluation which applies *extrinsic* criteria: how well does the system enable the user to complete a goal in the environment in which the system is deployed. Thus, for example, evaluating speech recognition systems by measuring divergence between system- and human-generated transcriptions of an agreed corpus of spoken data is technology evaluation of a user-visible task; measuring how much time is saved by a human post-editing the output of a speech transcribing dictation device vs. a human typing the entirety of the input in an actual office environment is part of a user-centred evaluation.

When thinking about technology evaluation, the distinction between user-visible and user-transparent tasks is important for at least three reasons.

(1) Most user-transparent tasks rest on some theoretical assumptions about the modularization of language processing and about the content of intermediate representations. Since very little theory about language processing is universally shared, finding a community which shares these assumptions about modularization and, if so, shares assumptions about the informational content of the representations the intermediate module consumes or produces, is difficult. Finding agreement concerning user-visible tasks is easier, though by no means without difficulty.

(2) Creation of resources to carry out user-transparent task evaluation is expensive, both because these resources cannot be found (by definition, as either the input or output of a user-transparent task is not part of the common user world) but must be manually encoded, and because they must be created by experts. Further,

[2] The distinction between user-visible and user-transparent tasks might be thought to duplicate the well-known *black-box/glass-box* distinction, but this is not the case. The latter distinction is made without reference to the user—it simply distinguishes looking at the overall input/output behaviour of a module without reference to its internal functioning vs. looking at its internal manipulations. This distinction holds at any level of system decomposition, without reference to the significance to the user of the inputs or outputs of the component or subcomponents under discussion.

as the theoretical assumptions required to posit a user-transparent task may shift or be abandoned altogether, the value of the resources accumulated to do this sort of evaluation may be of limited duration.

(3) Since user-visible task evaluation is more easily comprehensible to users and funders, and closer to user-centred evaluation, it is easier to convince these other stakeholders in the evaluation process of its value.

While these observations seem to argue against user-transparent task evaluation, it is clear that without it empirical investigation of language processing theories cannot take place. Advances in theories and algorithms take place because existing theories and algorithms cannot account for data, which can only be determined by testing/evaluating them against appropriate data; and if they change, the direction in which they change should be motivated by the discrepancies observed in the evaluation. So, while creating the resources for user-transparent task evaluation is expensive, this is simply the inevitable cost of doing research (which is not to propose the mindless proliferation of such resources—clearly each case needs to be justified and must merit support). Finally, though users and funders may find user-transparent task evaluation more comprehensible, the research community must show, through argument and results, that the expense involved in user-transparent task evaluation is worthwhile.

Thus, there is a role in technology evaluation for both user-visible and user-transparent task evaluation and examples of both are included in this issue.

3. Overview of the issue

Any journal special issue has limited space and hard decisions must be made about what to include and exclude. Setting criteria for selecting papers in as broad an area as evaluation in speech and language technology is extremely challenging. In this issue there has been an attempt to meet the following goals:

- to include work on evaluation across a broad range of areas in speech (recognition and synthesis) and language (understanding and generation) in order to reflect accurately the genuine breadth and diversity of work on evaluation at present;
- to review evaluation exercises that have been carried out and are of historical importance as well as to include proposals for, or initial work on, new approaches to evaluation and new areas to be evaluated;
- to include evaluations at the level of user-visible tasks (e.g. spoken dialogue, natural language generation, information extraction or message understanding), as well as at the level of user-transparent tasks (word sense disambiguation, parsing, grapheme-to-phoneme conversion).

In the rest of this section each paper in the issue is briefly introduced. The first three papers (Young & Chase; Hirschman; Mariani) review existing comparative technology evaluation programmes, primarily of user-visible tasks (speech recognition and information extraction and retrieval). The next two (Walker, Litman, Kamm & Abella; Mellish & Dale) make proposals concerning the evaluation of two user-visible tasks—spoken dialogue interaction and natural language generation—that have proved less tractable as subjects for large-scale comparative evaluations, but the evaluation of which is now under serious discussion. Cox, Linford, Hill & Johnston's paper concentrates on the metric for speech recognition evaluation, and proposes, effectively, a shift in the paradigm for speech recognition evaluation. The final four papers (Yvon *et al.*; Oepen & Flickinger; Sonntag & Portele; Kilgarriff) discuss evaluation of user-transparent tasks—grapheme-to-phoneme conversion, grammar profiling, prosodic content production and word sense disambiguation.

3.1. Young and Chase

The DARPA Continuous Speech Recognition (CSR) and Large Vocabulary Conversational Speech Recognition (LVCSR) evaluation programmes are in some sense the "canonical" S & L evaluations, as well as being the oldest. So it makes sense to start here.

Young and Chase review these programmes, supplying the original historical setting and motivation for them and chronicling how they have developed over their 10–15 year history. In particular they discuss the emergence of the "Hub and Spoke" paradigm, whereby an evaluation is split into a core test, the "hub", which all participants undertake, and a number of independent, optional tasks, the "spokes". The authors also provide details on the mechanics of the evaluations—the provision of data (for training and testing), the scoring metrics, the evolution of transcription and scoring practices for "found" speech and LVCSR, statistical analysis techniques for determining the significance of the results, the recent interest in confidence annotation and aspects of the organizational infrastructure. They close by laying out the positive and negative aspects of the evaluation experience, and conclude that the positive features outweigh the negative. While cautious about the extent to which these evaluations can serve as models for all areas of language processing, they firmly believe in the utility of component evaluations like these, in contrast with application-level "black-box" evaluations (user-centred evaluations in the terminology of Section 2), as the only way to promote systematic development in S & L technology and, ultimately, to produce a "mature science".

3.2. Hirschman

The Message Understanding Conference evaluation programme, or MUC, for short, is one of the written language counterparts to the DARPA CSR and LVCSR programme. Designed to evaluate language understanding technology, MUC started at roughly the same time as the first speech recognition evaluation in the late 1980s, and has evolved through a series of seven exercises till the most recent, MUC-7, held in the spring of 1998.

Hirschman's paper describes the progression of MUC from its informal, grassroots beginnings amongst U.S. NLP groups eager to compare their systems on common, real-world data, to the international, multi-task event it has become in recent years. Along the way many difficulties needed to be resolved (metrics, template design, automatic scoring software, suitable corpora) and new challenges needed to be invented to drive development (e.g. multilinguality, rapid domain portability). In recent years, the model of a single, scenario-based extraction task that characterized earlier MUCs (e.g. extracting information from newswires about terrorist attacks or business joint ventures) has been superseded by a model in which there are a number of less complex, domain-independent tasks (identifying named entities, linking certain co-referring expressions, filling small-scale templates about certain entities and entity relations), as well as the traditional domain-dependent scenario task. This has provided greater diagnostic insights for developers, has allowed high performance to be shown on simpler tasks while core extraction tasks remain stubbornly difficult and has led to the spin-off of commercial products (e.g. for named entity spotting).

The paper explains these developments in detail, providing the motivations for changes over the course of the programme and reporting the results obtained by the best systems at all stages. Following the historical review, Hirschman gives an overall assessment of the impact of the programme on information extraction technology, asking why systems appear to have hit a recall ceiling of around 60%. This observation raises additional questions about what performance levels are acceptable for information extraction, whether the scenario task is realistic, what the significance is of certain levels of human disagreement on the tasks and where MUC-style evaluation should go next. Hirschman goes on to compare the MUC evaluations with some other evaluation exercises—speech recognition evaluation (CSR), text retrieval (TREC), spoken language understanding (ATIS) and parsing (PARSEVAL). She closes by offering a set of "lessons learned", concerning the costs and benefits of MUC and related evaluations; here as elsewhere in the paper, she emphasizes the differing perspectives of the multiple stakeholders in the evaluation process: funders, developers, users.

3.3. Mariani

Mariani's paper is a report on the current state of evaluation in language engineering in the French-speaking world. The best known comparative evaluation exercises have been the US DARPA/NIST human language technology evaluations (reviewed in the Young & Chase and Hirschman articles in this issue), but as this article makes clear there is a wide range of evaluation activity underway in both speech and language in the French-speaking world. Mariani describes four actions pertaining to the evaluation of written language (text retrieval, text alignment, terminology extraction, message understanding) and three pertaining to spoken language (voice dictation, vocal dialogue, text-to-speech synthesis), all of which are underway and involve, in total, 69 research laboratories in seven French-speaking countries. He also describes various other activities relating to evaluation, including the provision of language resources and a morphosyntactic tagging evaluation exercise. For those who may have thought of evaluation in LE as an exclusively American enterprise, here is evidence of a dynamic and committed culture of evaluation in a neighbouring research community.

3.4. Walker, Litman, Kamm and Abella

Unlike areas where there is a "gold standard" (*the* correct transcription of a speech signal, *the* correct filled template in an extraction task), dialogue systems have proved difficult to evaluate because there is no unique correct dialogue. Clearly, to be successful a dialogue must minimally convey the information whose transfer is the purpose of the dialogue; but while achievement of this goal narrows the set of dialogues which can be construed to be successful, there are many other criteria which enable us to judge one dialogue as better than another, such as total number of exchanges, number of repair exchanges, etc.

Walker *et al.* describe a framework for evaluating spoken language dialogue agents that addresses this problem by showing how to define a single evaluation performance function that incorporates measures of both task success (core information transfer) and dialogue costs (efficiency and other qualitative measures). First, a task and a representation for task success are defined (the latter takes the form of an attribute-value matrix containing the information that must be exchanged between user and agent in order for the task to be carried out). Second, a wide range of features likely to influence dialogue cost are specified (number of utterances, number of repairs, mean recognition score). Experiments are then carried out with alternate dialogue agents and a range of users, and the task success and dialogue costs measured for each. Finally, user satisfaction is measured independently using surveys. Underlying the approach is the assumption that external indicators, such as user satisfaction, measure usability and that usability is correlated with performance. Thus, since overall performance is assumed to be a weighted linear combination of task success and dialogue costs, the independent user satisfaction measures can be used to solve for the weights in the performance function using linear regression.

This methodology is illustrated in the paper by using it to derive performance equations for two actual spoken dialogue systems—a system for voice retrieval of e-mail and a system for accessing train timetable information by telephone. The data from these two separate studies is then combined to seek generalizations about spoken dialogue agents across applications.

Having shown how to derive the performance function, the authors go on to show how it may be used (1) to learn optimal dialogue strategies and (2) to make predictions about what changes to an agent are likely to lead to increased performance.

The ideas presented in this paper have relevance well beyond spoken dialogue agents. They can be applied to any area of human-machine interaction where the overall value of the interaction must take into account both task success and interaction costs.

3.5. Mellish and Dale

Mellish and Dale address the role of evaluation in the area of natural language generation (NLG). As they observe, evaluation has to date played much less of a role in this area than in the complementary area natural language understanding (NLU)—for example, there have been no DARPA-style exercises for NLG. In their paper they review what work has been done in this area (accuracy evaluation; fluency/intelligibility evaluation; task evaluation), noting some overlap with machine translation evaluation, since machine translation systems also contain a generation component. They also attempt to characterize those aspects of NLG that make it so challenging to evaluate. Perhaps the most significant of these is lack of consensus about the inputs and outputs of the process. This is less of a concern for NLU since, for input, there are lots of exemplars in the form of real texts, while for output, once a target representation has been agreed—such as a MUC-style template—humans can perform the task and generally agree a unique, correct answer. Such is not the case for NLG, since many "right" generated texts may exist—the same problem as was observed for spoken dialogues in the preceding section. Other difficulties facing the evaluation of NLG systems include: uncertainty about what to measure; what controls to use (humans, or other systems?); how to acquire adequate training and test data; and how to handle disagreement among human judges.

Despite the difficulty of these challenges, the authors are bullish about both the need and prospects for evaluation in NLG. They close by proposing a scenario—generating summaries from monthly meteorological data—that illustrates how a fine-grained evaluation of an NLG system with respect to this task might be carried out. Such an evaluation would require reference to six component subtasks that are broadly representative of research areas within NLG: content determination, document structuring, lexicalization, aggregation, referring expression generation, surface realization. By working through the example scenario and considering what might go wrong with respect to each of these subtasks, we get a clearer idea of what we would want an evaluation to tell us. This breakdown of the NLG process into separably evaluable subcomponents is, as the authors point out, just a starting point (and even this breakdown, as they acknowledge, is not without controversy): precisely specifying metrics for each subcomponent, and getting agreement about these, is another matter altogether. Here, indeed, is a challenge for the NLG community.

3.6. Cox, Linford, Hill and Johnston

Unlike other papers in this issue which have described comparative exercises in evaluation (Young & Chase; Hirschman; Mariani; Yvon *et al.*) or the creation and use of carefully crafted corpora for diagnostic evaluation (Oepen & Flickinger; Kilgarriff), Cox *et al.* focus on the *metric* used for evaluation and report on experimentation which has as its goal the introduction of a new metric for speech recognition evaluation.

Speech recognizers have traditionally been assessed using the single measure of word error rate (WER) against a specific database. This involves the alignment of a proposed transcription of a segment of spoken data with a human-generated transcription of the same data, followed by the calculation of the number of deletions, substitutions and insertions required to transform the system's response into the reference transcription provided by the human. Cox *et al.* instead propose to use as a measure the amount of distortion which when added to a speech signal leads human recognition performance on the signal to match that of the recognizer. This approach depends, among other things, on the identification of a means of progressively distorting the speech signal which leads to a monotonic decrease in human recognition performance. This paper describes controlled experiments with a range of speakers (ages, genders, etc.) to see if one such means—time frequency modulation of isolated words—has the appropriate characteristics. These experiments are the first, but necessary, steps in a programme of exploring the viability of an alternative error measure.

This paper is provocative, not just in challenging the orthodox measure in speech recognition evaluation, but because it suggests a fundamentally different notion of evaluation (equivalence to human performance under some percentage distortion of input) to the standard one (percentage closeness to idealized human performance). This notion may have application in other areas as well.

3.7. Yvon et al.

One of the evaluation exercises coordinated by the Francophone Language Engineering network described in the Mariani paper was an evaluation for French language text-to-speech (TTS) systems, specifically, an evaluation of grapheme-to-phoneme conversion capability. This exercise is described in detail by Yvon *et al.*, in their paper in this volume. As far as I am aware, this is the first large-scale, comparative evaluation of grapheme-to-phoneme conversion on running text, and as such it makes a significant contribution to the spread of rigorous evaluation exercises in S & L technology. This paper details the methodology of the evaluation: the task—to transcribe phonemically excerpts from the newspaper, *Le Monde*; the preparation of reference transcriptions—the phonemic alphabet selected and the encoding of alternative pronunciations; and, the evaluation protocol and scoring procedures. The paper also summarizes the overall results of the eight systems participating in the exercise and attempts some error analysis. It concludes with a critical analysis of the evaluation methodology and an assessment of what the exercise can tell us about the state-of-the-art in French grapheme-to-phoneme conversion.

Grapheme-to-phoneme conversion is a good example of a user-transparent task—users do not want phonemes, they want synthesized speech. One issue that emerges clearly in this paper is the tension between finding an appropriate level of task granularity/specificity at which a common objective measure can be agreed across systems and wanting to assess the overall utility of systems for the high-level task for which they were designed. To be more specific: for TTS synthesis what matters overall is the comprehensibility of the synthesized speech. However, this is a difficult thing to assess, and further, even if it could be, it is not clear that such assessments would help system developers to improve their systems. So, as in this case, researchers look at the key components of which their systems are composed (e.g. in TTS, grapheme-to-phoneme conversion) and then attempt to define an objective measure for the output of these components (in this case, a reference phonemic transcript in an agreed common phonemic alphabet with an encoding for alternatives). The appropriate level of component grain and the appropriate precision of output have now been achieved to conduct an objective evaluation. But the danger is that maximizing scores in this evaluation will not necessarily lead to better speech synthesis systems: the chosen modularization for the task may be inappropriate (for example, key difficulties may have been exported into other components) or the agreed common output representation may be inadequate. The authors are well aware of these difficulties and their article contains a reflective assessment of the methodology.

3.8. Oepen and Flickinger

Oepen and Flickinger's paper presents a very different model of evaluation. Like Yvon *et al.* they are concerned with the evaluation of a user-transparent task, in this case the evaluation of broad-coverage computational grammars in the HPSG framework. However, rather than the using an annotated "real-world" *corpus*, their work involves the use of systematically constructed *test suites* of positive and negative examples of English sentences (which the parser/grammar should parse and fail to parse, respectively).

The work they report has taken place in the context of an international consortium of research groups aiming to produce a "multi-purpose broad-coverage, precise and re-usable computational grammar of English". By necessity the work takes place at multiple sites, involves multiple grammar writers and evolves over considerable time. These constraints mean that regular evaluation that is easy to carry out and informative at the appropriate level of granularity is a central concern. Above all, evaluation in this context must serve a *diagnostic* function, enabling developers to see what is wrong, and not merely benchmarking performance.

The paper first sets the context by describing the grammar development effort and the software environment in which it is carried out. The test suites are not simply a set of positive and negative examples of English sentences, but rather form a database of annotated examples which is integrated into the grammar engineering environment in such a fashion as to allow the grammar developer sophisticated control over which data is to be selected for evaluation, as well as multiple views of the results of evaluation. The authors illustrate how the test suite system they have developed allows for analysis of coverage, overgeneration, progress evaluation and computational performance, at various levels of detail. They compare the current test suite approach with the original Hewlett-Packard test suite of 1987 and show how progress has occurred. Finally, they offer a critical assessment of the approach, raising interesting points about the impossibility of separating application development and test suite construction and the difficulties of working with test items that test multiple phenomena at the same time.

3.9. Sonntag and Portele

Sonntag and Portele describe work to investigate and assess the prosodic quality of generated (natural or synthesized) speech—an important topic given advances still required to make synthesized speech sound natural. As with natural language generation and spoken dialogues, there is no single "correct" prosodic realization—many different realizations can have the desired effect. Yet, human assessors can consistently distinguish "good" from "bad" prosody, as they can reliably distinguish good from bad generated text and good from bad dialogues. How can prosodic quality be measured?

In order to focus solely on the prosodic content of spoken language, Sonntag and Portele propose transforming a speech signal so as to eliminate its lexical content, while preserving its key prosodic information. The first part of their paper reports experiments to assess different techniques for delexicalization. Six techniques were examined and experiments on subjects carried out to see if the delexicalized signals retained sufficient prosodic information to permit syllable recognition, phrase accent assignment, phrase boundary detection and phrase modality recognition. An overall subjective "pleasantness" assessment was also recorded. Based on the outcome of these experiments, a signal manipulation method for delexicalization was selected.

Next, they carried out experiments to ensure that the delexicalized signals retained enough prosodic information to allow listeners to associate aurally presented manipulated signals with written sentences of equivalent syntactic structure, but different lexical content.

Finally, the authors carried out experiments whereby listeners rated the prosodic naturalness of manipulated signals originating from multiple sources (human speakers or synthesizers) in relation to a written version of the sentences used to generate the initial speech signal. The outcome showed significant differences between the voices and offers interesting diagnostic information.

This paper offers a specific proposal for how to isolate and evaluate prosodic content of spoken language. However, the underlying method is potentially of wider applicability: can selective masking of certain features of the input to an S & L system allow other features to be isolated and studied independently?

3.10. Kilgarriff

Kilgarriff's paper discusses the evaluation of a user-transparent task—word sense disambiguation (WSD). The WSD task may be stated in general terms as follows: given a set of word types, each with one or more senses specified in associated word sense definitions, and a set of corpus instances of these word types, select the appropriate sense, or senses, for each instance. Evaluations which involve instantiating this general scheme have been carried out to date by various individual researchers, each choosing particular corpus data to be sense tagged and particular dictionaries to serve as reference repositories of word senses and word sense definitions. Kilgarriff reviews these previous efforts, but the chief concern of his paper is to document the emerging consensus in the WSD community about how this earlier work should be synthesized into a single comparative evaluation exercise. This exercise, dubbed SENSEVAL, will for the first time, in September 1988, invite researchers across the field to participate in a common evaluation in the general style of the DARPA evaluations in other areas of speech and language. Designing this common exercise has involved resolving issues such as whether to tag all words in a test corpus or only certain selected word types, whether to allow multiple senses to be assigned to a word instance (in the system response, or even in the key), which dictionary to use as a reference, how to sample the corpus for word types and instances to be tagged, and so on. These issues and the history and reasoning behind the choices made for SENSEVAL are fully presented in the paper.

4. Future directions

As this issue demonstrates, there is tremendous diversity and vitality in efforts relating to the evaluation of S & L technology. This ranges from well-established programmes, like the DARPA CSR and LVCSR exercises, to first time exercises like SENSEVAL, which are just getting under way. Clearly these exercises will continue and have the potential to offer insights to researchers for years to come. Other areas, such as spoken language dialogue and natural language generation, are still wrestling with the problems involved in agreeing acceptable task definitions and metrics for evaluation. But, as the Walker *et al.* and Mellish and Dale papers in this issue indicate, ways forward in these areas are being found, and given the general climate it seems likely that evaluation will play an ever increasing role in these areas too. Of course evaluation activities are also ongoing or planned in many areas not touched on directly in this issue, including (text and spoken language) information retrieval (Harman, 1998; TREC, 1998), machine translation (White & Taylor, 1998), summarization (TIPSTER, 1998; Johnson, 1998), topic detection and tracking (Wayne, 1998), controlled language checking (Rodier, 1998), corpus-based parsing (Carroll, 1998), and part-of-speech tagging (Adda, Mariani, Lecomte, Paroubek & Rajman, 1998).

Most work on evaluation in S & L technology to date has been in English. However, as the Mariani paper in this issue shows, there is now a strong programme of evaluation underway for French language systems. As this issue goes to press a first call has been issued for a Japanese information retrieval and extraction evaluation exercise—IREX (IREX, 1998). We can expect that evaluation exercises involving other languages will emerge in the near future, both for uni-lingual and cross-lingual tasks.

It might therefore seem that all is rosy with respect to the future of technology evaluation in S & L. However, this is far from the case. S & L evaluation is expensive: creating the annotated resources that most exercises require demands significant amounts of manpower and participating in the exercises also requires considerable effort. Funders need to be continually convinced that an evaluation regime is genuinely leading somewhere. This means demonstrating progress against agreed metrics, increasing task complexity or diversity over time, and showing that the resulting technology does, in fact, end up in real applications.

With the U.S. TIPSTER programme ending in the autumn of 1998, DARPA have announced that MUC-7, run in the spring of 1998, will be the last MUC. Their view is that it has been a very successful programme, but that it is time for a review to be undertaken and something new to take its place. Thus, at present, a question mark is hanging over the future of evaluation in the important area of text understanding.

Meanwhile, in the European context, agreement still needs to be achieved concerning the inclusion of an evaluation regime within the European Commission's Language Engineering programme. The ELSE project (ELSE, 1998) is investigating how an infrastructure for language and speech evaluation could be set up in Europe as part of the Fifth Framework programme. Amongst the scenarios being considered is a "grand challenge" exercise involving multi-lingual, spoken language retrieval of information from distributed textual and spoken language sources. Such an exercise might be set up along the lines of the "braided chain" model discussed in Crouch, Gaizauskas & Netter (1995) and Sparck Jones & Galliers (1996). In this model there are multiple evaluation points and participants choose at which points they wish to be evaluated; they may be supplied with "vanilla" components which perform aspects of the task other than those on which they choose to be evaluated. This scenario and evaluation model has much to recommend it, but is still only a proposal at this stage.

Regardless of the direction that centrally funded, large-scale evaluation exercises take, the empirical trend in S & L seems set to continue. The setting up of SENSEVAL shows just how much can be achieved with the enthusiasm of the research community and very modest amounts of funding. Were he here to see the direction S & L research is heading, with its growing emphasis on quantitative evaluation, I think Kelvin would be pleased.

Thanks to Steve Young for suggesting the idea of a special issue in *Computer Speech and Language* and for holding my hand through the process of being a guest editor. Thanks also to all those who reviewed papers for the issue and whose comments improved the issue and contributed greatly to all those who submitted papers. Thanks to Lyn Walker who finally tracked down the Kelvin quotation for which I'd been hunting for years, and to Mark Hepple for comments on this introduction to the issue. And last, but certainly not least, many thanks to Gillian Callaghan for secretarial support without which the issue never would have made it out the door.

References

Adda, G., Mariani, J., Lecomte, J., Paroubek, P. & Rajman, M. (1998). The GRACE French part-of-speech tagging evaluation task. *Proceedings of the First International Conference on Language Resources & Evaluation.* Granada, pp. 433–446.

Carroll, J. (ed.) (1998). *Proceedings of the Workshop on Parsing Evaluation, Granada.* At the First International Conference on Language Resources & Evaluation.

Cohen, P. R. (1995). *Empirical Methods for Artificial Intelligence.* MIT Press, Cambridge, MA.

Crouch, R., Gaizauskas, R. J. & Netter, K. (1995). Interim report of the study group on assessment and evaluation. Technical report, EAGLES project, Language Engineering Programme, European Commission. Available at http://xxx.lanl.gov/ps/9601003.

EAGLES: Expert Advisory Group on Language Engineering Standards. (1995). Evaluation of natural language processing systems. Available at http//:issco-www.unige.ch/ewg95 or http://www.ilc.pi.cnr.it/EAGLES/home.html. EAG-EWG-PR.2.

ELSE: Evaluation in Language and Speech Engineering. http://www2.echo.lu/langeng/en/le4-/else/else.html. Site visited July 21, 1998.

Harman, D. (1998). The Text REtrieval Conferences (TRECs) and the Cross-Language Track. *Proceedings of the First International Conference on Language Resources & Evaluation.* Granada, pp. 517–522.

IREX: IR and IE Contest for Japanese Language. http://cs.nyu.edu/cs/projects/proteus/irex/index-e.html. Site visited August 19, 1998.

Johnson, F. C. (1998). Evaluation: a perspective on automatic abstracting research. *ELRA Newsletter* **3**(2).

Rodier, E. (1998). Semi automatic generation of reference diagnostics in an evaluation tool for simplified English. *Proceedings of the First International Conference on Language Resources & Evaluation*. Granada, pp. 283–287.

SALT (U.K. Speech and Language Technology Club). (1997). *Proceedings of the Workshop on Evaluation in Speech and Language Technology.* Sheffield.

Sparck Jones, K. & Galliers, J. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review.* Springer-Verlag, Berlin.

TIPSTER Text Program. http://www.tipster.org. Site visited July 21, 1998.

TREC: Text REtrieval Conference. http://trec.nist.gove. Site visited July 24, 1998.

Walker, M. A. & Moore, J. D. (1997). Empirical studies in discourse. *Computational Linguistics* **23**(1), 1–12.

Wayne, C. (1998). Topic detection and tracking: a case study in corpus creation and evaluation methodologies. *Proceedings of the First International Conference on Language Resources & Evaluation.* Granada, pp. 111–115.

White, J. S. & Taylor, K. B. (1998). A task-oriented evaluation metric for machine translation. *Proceedings of the First International Conference on Language Resources & Evaluation.* Granada, pp. 21–25.