Machine Translation **12:** 375–379, 1997.

Book Review

Karen Sparck Jones & Julia R. Galliers, *Evaluating Natural Language Processing Systems: An Analysis and Review*. Lecture Notes in Artificial Intelligence 1083, Springer, Berlin, 1995, xv + 228 pp. ISBN 3-540-61309-9.

Although this book grew out of a report (Galliers and Sparck Jones, 1993), it can actually be seen as a useful manual on how to conduct NLP evaluation as well as a major reference book on the topic. The two authors present a thorough analysis of the sometimes extremely difficult concepts involved in NLP evaluation. Because the book contains a very useful glossary at the beginning and clear definitions are given in the text itself as well, the terminology which has often been very diversely interpreted in the field (see EAGLES, 1995) is never misleading in this text.

A large part of the book, especially in Chapter 1, consists in discussion of case studies, and examples of evaluation, both invented and adapted from real experience, lead the reader through increasing levels of complexity and bring out the issues in a very immediate fashion. For instance, the crucial distinctions between *environment*, *set-up* and *system* are established at the outset, through working out the intricate relations between these concepts in the evaluation of a Database Management System (DBMS) which contains an NL database query system.

The book is divided into three chapters of unequal length. Chapter 1 (62 pages) gives the general framework and defines the concepts used throughout the rest of the volume. Chapter 2 (128 pages) is mainly historical, but presents work done in NLP evaluation in the framework defined in Chapter 1. Chapter 3 (30 pages) proposes the authors' methodological and practical recommendations for performing an evaluation.

For the readers of this journal, there will be several sections of interest throughout those chapters.

1. Concepts and Definitions

In the first chapter, MT is used fairly often as an example of an NLP application under testing. This indeed has historically been the case, as is pointed out later in Chapter 2: evaluation was first done on systems for which there was a commercial and financial incentive in doing some form of testing or evaluation. For instance,

as was found in the study of test-suites done in the TSNLP project (see Estival et al., 1994), more test-suites have been developed for MT systems than for any other NLP applications.

However, the main example of testing in Chapter 1 is that of an Information Retrieval (IR) application. This example allows the authors to give an extensive description of the different combinations of setups, parameters, environments, users, user needs, etc., needed for different evaluations. The authors also draw parallels between IR evaluation and NL evaluation throughout the book, because a number of the concepts used in NL evaluation have been inherited from the earlier experience of IR evaluation.

Also drawing parallels with another type of applications, namely expert systems, the authors make a very careful distinction between *generic systems* and *general purpose systems*, which leads them to a number of observations (some of these may sound rather self-evident, but they are often ignored by developers and evaluators alike):

– "[T]he problem with generic shells is that it is impossible to evaluate them without a specific instantiating application" (p. 50)
  This is indeed one of the main problems of evaluation: balancing the intrinsic qualities or shortcomings of a system against the effects of a particular instantiation. This is always the case when evaluating an MT system which requires users to develop their own domain dictionaries before being able to use it on real texts.
– "[T]he more limited the range of a generic system, the more likely it is to be able to function as a general-purpose component" (p. 50)
  This point is in fact one of the lessons from the work done during this past decade on reusable systems and components: generic software will only save effort in the development of new applications if it doesn't require too much adaptation and modification when it is plugged into different applications and when it must interface with different components.

Furthermore, the authors point to the consequence of the assumption about modularity which lies behind generic systems: i.e. the danger that such a system will be rather shallow and that the more challenging aspects of processing will then be dealt with by other parts of the application. In this context, a number of interesting side issues are taken up and discussed, e.g. " How much is it the job of a *language* processor to indicate what might be wrong with an input it cannot handle?" (p. 54)

One of the most important distinctions that must be drawn when performing an evaluation of a system is that between *intrinsic* criteria, i.e. those concerned with the system's own objectives, and *extrinsic* criteria, i.e. those concerned with the function of the system in relation to its setup. This distinction is clearly defined and exemplified. Unfortunately, the same cannot be said for an equally important distinction, that between *reliability* (how dependable the evaluation measure being

used is) and *validity* (whether the measure used actually measures what is being tested), which is crucial for the design of an evaluation and the interpretation of its results but which, in spite of a couple of examples, will probably not be transparent to readers not already familiar with it.

2. Historical Perspective

While Chapter 1 is concerned with giving definitions and examples, Chapter 2 reviews the actual state of the art in NLP evaluation and starts with a history of evaluation. The authors make the rather negative point that the main reason there hadn't been much evaluation in NLP until recently was that there really is no point in evaluating a system when you already know it does not perform to your expectations, and that unfortunately has been the state of affairs during most of NLP early history.

The authors then go on to describe in a quite detailed way the various DARPA, ARPA, MUC, EAGLES, etc. efforts at NLP evaluation, and this chapter will be very valuable to anyone interested in the topic (although it is worth pointing out to the readers of this journal that while the authors only refer to (EAGLES, 1994), a more recent version of that report (EAGLES,1995) is available at the EAGLES ftp site.* It is as comprehensive as one might wish it to be, and (unlike Chapter 1, where most of the typos and text processing errors occur, see below) very readable.

2.1. MT ASSESSMENT

Chapter 2 contains much material which will be of direct relevance to MT researchers, developers and users, in particular section 2.1.1 (pp. 70–87), which focuses on the evaluation of MT systems. The sections entitled "Stages for evaluation", "Criteria for MT assessment" (linguistic assessment, operational assessment and economic assessment) are self-explanatory and should be required reading for anyone involved in MT development and/or assessment. The section "Some important MT evaluations" describes, as one would expect,the ALPAC, TAUM-AVIATION, SYSTRAN, JEIDA and JTEC, and the DARPA MT evaluations. These descriptions are couched in the terms defined and used so far in the book and thus provide a more unified picture of the field than can usually be found, albeit much shorter than the report of Falkedal (1991) from which some of the information seems to be drawn.

A separate section presents the EAGLES approach to evaluation of Translation Aids (EAGLES, 1994) and in particular the *checklist mechanism* which the authors consider would be appropriate not only for *product assessment* (or *adequacy evaluation*, which is what the EAGLES work was restricted to) but for evaluation in general, i.e. also for *progress* and *diagnostic* evaluations.

* ftp://ftp.ilc.pi.cnr.it/pub/eagles/evaluation/

The other sections of Chapter 2 which are devoted to reports on past evaluations of other types of NLP applications (Message Understanding, Data Base Query, Speech Understanding, Text Retrieval, etc.) will be of historical interest to some readers, but not as relevant as the sections reporting on the new developments of the past ten years (pp. 125–189). These include not only evaluation workshops and tutorials, but more importantly the building of methodologies and the setting of standards (see EAGLES, 1995), as well as the development of test corpora, test suites, test collections and toolkits.

3. Practical Recommendations

The final chapter is actually a guide for evaluation and sets out to outline procedures the authors recommend, by showing how they would be applied using a series of examples similar to those presented in Chapter 1. This chapter will be of practical interest to both developers and users.

4. Problems and Quibbles

Unfortunately, this book suffers from poor editing. There are a number of typographical errors and, more annoyingly, of text processing errors of the type "wrong cut-and-paste" which crop up when revising or merging documents. This is all the more regrettable because the prose is in general rather dense and not always easy to follow and those mistakes tend to affect the comprehension of the text. For instance: "Hanks, Pollack and Cohen's points about maintaining evaluation proprieties, in the form and [sic] of test beds, are thus relevant here." (p. 58) or ". . . or because, when one setup is included in another, there are transformation effects as criteria are reinterpreted are [sic] applicable to a particular setup, for instance when . . . " (p. 30)

While the glossary constitutes a very useful resource and is quite complete, the index is not as reliable. For instance, there is only one entry for 'EAGLES', pointing to the beginning of Chapter 2, while there is actually a whole section devoted to the work of the EAGLES group (section 2.2.3) and numerous other mentions of it throughout that chapter. For a book which should serve as reference on past work and a guide to such a complex field, the weakness of the index is an unfortunate shortcoming, which will make the information it contains less accessible to researchers.

It is also regrettable that, although test suites are described as an "obvious possibility" (p. 161) for test material, there is no mention of the TSNLP project (Lehmann et al., 1996) which addressed precisely the issue of standardization of testing material and the methodology of creating and using test suites.

In general, probably because this book is basically a revised version of the authors' earlier report, the reader will feel that some of the more recent work in the area is not reported.

5. Conclusion

Nevertheless, this book will be extremely useful to anyone involved in any sort of evaluation procedure for NLP applications, which means all NLP system developers and many users who have always had to devise their own evaluation procedures on an *ad hoc* basis and without guidelines as to what to expect from the time-consuming and labour-intensive process. This book is difficult to read at times and suffers from the drawbacks mentioned above but is definitely to be recommended. It would also serve as good supplementary reading for an advanced course in NLP, and will certainly provide interesting reading material even for an introductory NLP course. For a course in NLP evaluation, it would have to supplemented by other material but could be used as the main text.

References

EAGLES: 1994, *EAGLES: Evaluation of Natural Language Processing Systems*, Draft Interim Report, EAGLES Document EAG-EWG-IR.2, Eagles Secretariat, Istituto di Linguistica Computazionale, Pisa.

EAGLES: 1995, *EAGLES: Evaluation of Natural Language Processing Systems*, Final Report, EAGLES Document EAG-EWG-PR.2, Eagles Secretariat, Istituto di Linguistica Computazionale, Pisa. ftp://ftp.ilc.pi.cnr.it/pub/eagles/evaluation/

Estival, Dominique, Falkedal, Kirsten et al.: 1994, *Analysis of Existing Test Suites*, Report to LRE 62-089 (TSNLP, D-WP1), University of Essex.

Falkedal, Kirsten: 1991, *Evaluation Methods for machine Translation Systems. An Historical Overview and Critical Account*. Report, ISSCO, Université de Genève.

Galliers, J.R., and Sparck Jones, K.: 1993, *Evaluating Natural Language Processing Systems*. Technical Report 291. Computer Laboratory, University of Cambridge.

Sabine, Lehmann, Oepen, Stephan, Regnier-Prost, Sylvie, Netter, Klaus, Lux, Veronika, Klein, Judith, Falkedal, Kirsten, Fouvry, Frederik, Estival, Dominique, Dauphin, Eva, Compagnion, Hervé, Baur, Judith, Balkan, Lorna, and Arnold, Doug: 1996, 'TSNLP – Test Suites for Natural Language Processing', in *COLING Proceedings*, Copenhagen, **2**, 711–716.

DOMINIQUE ESTIVAL
Department of Linguistics and Applied Linguistics
University of Melbourne