Evaluating Natural Language Processing Systems: An Analysis and Review

Karen Sparck Jones and Julia R. Galliers
(University of Cambridge)

Berlin: Springer-Verlag (Lecture Notes
in Computer Science 1083, edited by G.
Goos, J. Hartmanis and J. van Leeuwen,
Subseries Lecture Notes in Artificial
Intelligence, edited by J. G. Carbonell
and J. Siekmann), 1995, xv + 228 pp;
paperbound, ISBN 3-540-61309-9, $43.00

Reviewed by
Sharon M. Walter
Air Force Research Laboratory

Having matured to the point of utility for certain circumscribed applications, natural language processing (NLP) technology has a growing need for the formulation of evaluation methodologies. *Evaluating Natural Language Processing Systems: An Analysis and Review* provides a historical perspective and outline of efforts that have been made to provide meaningful and useful evaluations of NLP systems, describes the deficiencies of those efforts, and offers an evaluation approach intended to address those deficiencies.

Evaluating Natural Language Processing Systems is divided into three chapters. While Chapter 1 purports to introduce evaluation concepts and terminology, the esoteric presentation style detracts from its value for technology neophytes. The chapter does, however, provide a great deal of information and in the course of doing so, presents relevant, interesting issues such as the special problems associated with evaluating so-called **generic systems**, and the value of **qualitative** versus (or, in conjunction with) **quantitative** evaluation approaches.

Chapter 2 is a comprehensive and in-depth review of significant NLP evaluation research activities and resources, ranging across multiple modalities (e.g., speech), task domains (e.g., message processing), participatory activities (e.g., workshops, tutorials), evaluation methodologies and experiment techniques (e.g., Wizard-of-Oz experiments), resources (e.g., corpora, test suites, toolkits), and resource associations (e.g., the Linguistic Data Consortium). It is made up of three sections. Subdivisions within the first section focus on evaluation activities in each of the task areas of machine translation, message processing, database inquiry, speech understanding, summarizing and message categorization. The task area of text retrieval evaluation is also reported here, with the caveat that "though text retrieval is not defined here as an NLP task," recent evaluation activities "recognising the effects of environment and task or function complexity" are relevant to NLP system evaluation. Section 1 concludes with a brief, but excellent, investigation of whether "lessons" learned from task-specific evaluations provide any insight toward the development of general NLP evaluation tools, and lists seven basic, clear and concise, lessons relating to the set-up requirements for any (general *or* task-specific) NLP evaluation. These lessons are recommendations about evaluation parameters that must be clarified before an evaluation takes place.

Section 2 of Chapter 2 presents a survey of NLP evaluation activities and materials that do *not* focus on specific NLP tasks. Informative synopses of three evaluation workshops, two parsing evaluation workshops, and an evaluation tutorial are provided. An assessment of the general impact of DARPA-sponsored activities acknowledges DARPA's crucial role in NLP evaluation and presents a penetrating critique of those activities. Two evaluation methodologies, the Wizard-of-Oz method and the Neal-Montgomery System Evaluation Methodology, are reviewed. I was peripherally involved in the development of the latter of these and was astonished to read that it "was intended by its authors to be *the* standard evaluation tool for any NLP system." Unfortunately, I believe that I am the source of that misinformation, having misinterpreted, and responded to, a query at the DARPA Speech and Natural Language Workshop of February 1992 about the methodology as "the" standard. I hope my misstatement has not deterred researchers from investigating the value of that exceptional piece of work. As I stated in the preface of Neal et al. (1992), "The evaluation methodology is not presented here as a product to be accepted, in toto, by the NLP community as the standard for system evaluation, but rather as a basis for discussion, critique, and possible refinement towards standards development."

Sparck Jones and Galliers conclude Chapter 2 with a summary of four basic requirements for NLP evaluations that they have derived from their observation of existing approaches and methodologies:

1. Evaluations must be designed to address issues relevant to the specific task domain of the NLP system; therefore, NLP systems operating in different task domains require different evaluation criteria.
2. Evaluations must focus more attention on the "environmental" factors associated with NLP systems in actual use—for example, end-user characteristics.
3. Evaluations must identify all system elements that can figure as performance factors. Sparck Jones and Galliers recommend use of a process of "factor decomposition" to ensure that all relevant aspects of a system and its environment are considered in an evaluation.
4. Evaluation criteria for generic NLP systems are not, and probably cannot be, adequately defined. It is the authors' opinion that generic systems must be instantiated within a task to allow meaningful evaluation.

In Chapter 3, from conclusions presented in Chapter 2, the authors propose that evaluations be designed and conducted on the basis of an in-depth examination (by "decomposition") of all NLP system and environmental factors that may affect performance. The in-depth examination is performed comprehensively and systematically, in a top-down fashion, before the evaluation per se proceeds. In the first phase, answers to a series of general questions about the purpose and mode of evaluation define the evaluation **remit**. Information in the remit is then used to develop the evaluation **design**, the precise definition of the evaluation parameter settings. Together, the remit and design represent the fine detail of all aspects of the evaluation. Evaluations with minor variations on the entries in the remit or design represent individual **runs** of the evaluation, which can then be compared to each other by presenting them in a **grid** format. This evaluation methodology is proposed to address the four basic requirements of NLP evaluations (above, from the book's Chapter 2).

The NLP system evaluation methodology described by Sparck Jones and Galliers makes the circumstances of evaluations clear, comparable, and repeatable. Properly applied, it could ensure common, consistent ground conditions for NLP system evaluation comparisons, making it a very valuable resource. The evaluation methodology does not, however, appear to strike at the heart of the evaluation problem of defining specific criteria by which to describe and compare system capabilities, evading the issue in fact by proposing that general criteria cannot be defined due to the necessity of case-by-case specification of evaluation criteria.

I disagree with the authors' premise that generic NLP systems cannot be evaluated without being instantiated within a specific task domain. As a matter of fact, it seems that the proposed evaluation methodology goes a great distance towards providing a means to relay valuable information about the boundary capabilities of an uninstantiated generic system. Further, in doing so, the methodology may characterize **classes** of task domains to which a generic system could appropriately be applied and the data that is required for its instantiation.

Finally, on a book format matter, there is a glossary at the front of the book, but the definitions it provides are neither complete nor sufficiently descriptive of its terms to be of use. For example, *complexity, operation, parameter,* and *transportability* are all listed with the same vague glossary definition, "of system." Similarly, *broad scope, narrow scope,* and *working* are listed with the obscure definition, "of setup." A separate table of abbreviations would also have been worthwhile.

Reference

Neal, Jeannette G., Elissa L. Feit, Douglas J. Funke, and Christine A. Montgomery. 1992. An Evaluation Methodology for Natural Language Processing Systems. Rome Laboratory Technical Report 92-308.

Sharon M. Walter is a member of the Speech Processing Group at the Air Force Research Laboratory Rome Research Site in Rome, New York. Her technical interests include natural language processing and NLP evaluation, spoken language translation, and speech processing technology applications for law-enforcement domains. Walter's address is: AFRL/IFEC, 32 Hangar Road, Rome, New York 13441-4114; e-mail: walters@rl.af.mil