Lecture Notes in Artificial Intelligence 1083

Subseries of Lecture Notes in Computer Science
Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science
Edited by G. Goos, J. Hartmanis and J. van Leeuwen

Karen Sparck Jones Julia R. Galliers

Evaluating Natural Language Processing Systems

An Analysis and Review

Springer

Series Editors

Jaime G. Carbonell
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213-3891, USA

Jörg Siekmann
University of Saarland
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany

Authors

Karen Sparck Jones
Julia Rose Galliers
Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street
Cambridge CB2 3QG, UK

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Sparck Jones, Karen:
Evaluating natural language processing systems : an analysis and review / K. Sparck Jones ; J. R. Galliers. - Berlin ; Heidelberg ; New York ; Barcelona ; Budapest ; Hong Kong ; London ; Milan ; Paris ; Santa Clara ; Singapore ; Tokyo : Springer, 1996
(Lecture notes in computer science ; 1083)
ISBN 3-540-61309-9
NE: Galliers, Julia R.:; GT

CR Subject Classification (1991): I.2.7, H.3, I.2

ISBN 3-540-61309-9 Springer-Verlag Berlin Heidelberg New York

Typesetting: Camera ready by author
SPIN 10513063 06/3142 – 5 4 3 2 1 0 Printed on acid-free paper

Preface

This book presents a detailed analysis and review of NLP evaluation, in principle and in practice. Chapter 1 examines evaluation concepts and establishes a framework for NLP system evaluation. This makes use of experience in the related area of information retrieval, and the analysis also refers to evaluation in speech processing. Chapter 2 surveys significant evaluation work done so far, in relation both to particular tasks, for instance machine translation, and to NLP evaluation methodology as such, for example as this bears on generic system evaluation. The conclusion is that evaluation strategies and techniques for NLP need much more development, in particular to take proper account of the influence of system tasks and settings. Chapter 3 develops a general approach to NLP evaluation, aimed at methodologically sound strategies for test and evaluation motivated by comprehensive performance factor identification. The analysis throughout the book is supported by extensive illustrative examples.

The book is a development and update of an earlier technical report (Galliers and Sparck Jones, 1993). The work for the report was carried out in the Computer Laboratory, University of Cambridge, under the UK Science and Engineering Research Council's Grant GR/F 35227, 'A Contextual Reasoning and Cooperative Response Framework for the Core Language Engine (CLARE)'. This grant was for a component of a much larger project, developing CLARE itself, which was conducted at SRI International's Cambridge Centre, and we are grateful to Dr S.G. Pulman, the Centre's Director, for providing us with the opportunity and push for our enterprise. We are also very grateful to many people who supplied us with reports, papers, and information about their evaluation activities, both for the original report and this successor book, and to our colleagues, notably Louis des Tombe, for thoughtful comments.

Computer Laboratory
University of Cambridge

Karen Sparck Jones
Julia Galliers

January 1996

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Glossary of terms

(assessment	= evaluation)
acceptability	class of performance criterion
activity	of user in setup
adequacy evaluation	ends-oriented evaluation
AI	artificial intelligence
aims	of user
angle	viewpoint in evaluation linked to subject's ends
annotated corpus	corpus with labels
answer	output system should supply
antecedent variable	usually environment variable
apparatus	equipment other than system in setup
application	(system for) task in specific domain
architecture	infrastructure specification for NLP system
argot	very restricted sublanguage
attribute	pertinent to quality characteristic
baseline	performance floor from simple system
behaviour	of user
benchmark	established performance norm
black box	input/output-only evaluation
bound	area of evaluation - wide or narrow
broad scope	of setup
catalogue	fact list on evaluation subject
category	of user e.g. casual
checking collection	controlled test collection
checklist	evaluators' aid for featurisation
class	of performance criterion - efficiency etc
complexity	of system
component	part of system
composite	evaluation subsuming several measures
constitution	of evaluation subject
consumer	of evaluation findings
context	of evaluation subject
corpus	of test material
coverage corpus	corpus with all phenomena
criterion	for evaluation
customer	interested/consuming party for evaluation
data sort	kind of test/evaluation material
data source	e.g. corpus
decomposition	of setup working or system operation
design	for/of evaluation
design goal	system specification for objective
development data	working data for whole community
(diagnosis	= evaluation)
diagnostic evaluation	analytical evaluation

dialogue	user-system interaction involving NL
distribution corpus	corpus giving phenomena distribution
division	between l-system and n-system
domain	area or field of task
dry run	of evaluation procedure to check out
eccentric	idiosyncratic system performance
effect	of system, including output
effectiveness	class of performance criterion
efficiency	class of performance criterion
ends	system objectives/functions or setup purposes/functions
environment	setup from system's point of view
environment factor	variable as factor affecting performance
evaluation	of system or setup performance
evaluation data	data used for evaluation
evaluation methodology	methodology for evaluation
evaluation procedure	for carrying out test
evaluation standard	requirements for evaluation criteria etc
exemplar	baseline or benchmark performance
exigent processing	thorough NLP
experiment	to explicate system/setup performance
extrinsic criterion	for evaluating wrt embedding setup
factor	see performance factor
feature	attribute, attribute value
featurisation	feature choice for evaluation
field evaluation	evaluation in real-life situation
form	of evaluation yardstick as attainable/ideal etc
full processing	complete NLP
function	role of system in setup (or one setup in another)
general-purpose system	system for any application
generic system	system independent of application
glass box	internal operation evaluation
goal	of evaluation
granularity	of parameters under evaluation
grid	design style for test/evaluation
guidelines	for evaluation
hybrid system	with both l- and n-subsystems
indicator	of performance, variable or parameter
informativeness	of evaluation about system or setup
interactive system	with user-system dialogue
interface	system for user interaction involving NLP
interest	category of evaluation requester e.g. developer

intervening variable	usually system parameter
instance	of input in test data
intrinsic criterion	for internal system/setup evaluation
investigation	to determine system performance
IR	information retrieval
kind	of evaluation as experiment/investigation
l-system	see language system
language	natural language
language system	(sub)system doing NLP
legitimacy	of data for test/evaluation use
linkage	of variables, parameters, objectives, effects, and measures
measure	of performance, instantiating criterion
mega-evaluation	large-scale, long-term, multi-task evaluation
method	of applying measure
methodology	for test/evaluation
metric	= measure
mode	of evaluation as qualitative/ quantitative/hybrid
motivation	stimulating reason for evaluation
n-system	(sub)system not doing NLP
narrow bound	of area of evaluation
narrow scope	of setup
NL	natural language
NLP	natural language processing
non-interactive	system without user dialogue
norm	performance requirement (benchmark or target)
objective	what system itself is for
observation	of system/setup, preceding evaluation
operation	of system
orientation	of evaluation as extrinsic/intrinsic
p-setup	setup for individual user
parameter	of system
partial processing	incomplete NLP
performance	of system/setup wrt objective/purpose
performance exemplar	from baseline or benchmark
performance factor	any system/setup element affecting performance
perspective	aspect under which evaluation subject seen e.g. financial
pretest	test of evaluation measures, methods
programme	set of related tests/evaluations
progress evaluation	improvement, development evaluation
pseudo-language	not actually natural language
purpose	what setup is for
qualitative	holistic, non-numeric performance measure

quality characteristic	(general) desired property of evaluation subject
quantitative	numeric performance measure
quasi-language	sublanguage with own life
range	of NLP system, especially generic
rationale	for performance comparison
reality	of data for test/evaluation use
reasonable	fair performance, given environment
references	data including answers supporting evaluation
reliability	consistency of performance measure
remit	of evaluation
reportable attribute	for evaluation customer
representativeness	of data for test/evaluation use
richness	of language
role	of user in setup e.g. data input
run	of system giving performance measure
scenario	test and evaluation plan
scope	of setup - broad or narrow
separation	of system from user
serious material	complex natural language material
setting	of parameter
setup	system plus operational context
simple processing	rudimentary NLP
sort	of test/evaluation data e.g. test suite
source	corpus for data
standards	requirements for test/evaluation conduct
status	of test/evaluation data for e.g. representativeness
strategy	for conducting evaluation
style	of evaluation as exhaustive/indicative etc
subject	of evaluation, i.e. component/(sub)system/ setup
sublanguage	of natural language
substitute	honed answer in reference data
subsystem	l- or n- part of system
system	software+hardware entity
system factor	parameter as factor affecting performance
target	for performance
task	what system does
terminal	I/O manifestation of interface
test	investigation or experiment
test bed	application for exploring system design
test collection	data, with references, especially for experiment
test data	data for tests

test methodology	methodology for tests
test program	for doing runs, scoring performance, etc
test set	data subset used for system testing
test suite	designed test material
tool	for evaluation, i.e. data or program
toolkit	processing tools e.g. software for test/ evaluation
training set	data subset used for system development
transcription	test collection of transcribed speech
transportability	of system
trivial material	simple natural language material
tuning	of system to application
tweaking	of system to evaluation conditions
type	of evaluation as black box/glass box
user	of system in setup
utility	operational setup-oriented criterion
validity	propriety of performance measure
value	of variable
variable	property of environment affecting system
wide bound	large area of evaluation
working	of setup
working data	system input material for testing
yardstick	nature of performance for comparison