

AUTOMATIC INFORMATION AND
LANGUAGE PROCESSING :
RETHINKING EVALUATION

Karen Sparck Jones

Computer Laboratory
University of Cambridge

Current evaluation style :

good for technology

BUT

more limited than appears

underlying model :

system plugs into context

- defective

especially given IT evolution

MESSAGE :

Don't look at the system -

look at its role

TALK STRUCTURE

review task state :

Information Retrieval

Speech Recognition

Machine Translation

Information Extraction

Automatic Summarising

Intelligent Inquiry

with respect to

evaluation paradigm [DARPA]

task properties

paradigm assumptions

justification

evaluation deficiencies

consider

impact of current IT

==>> NECESSARY NEW DIRECTIONS

SOME TERMS

[[component] system] setup]
setup includes data, people ...

system

does task defined by objectives
has function in context supplied
by setup

evaluate using criteria
interpreted by measures
applied through methods
wrt system parameters
environment variables

for intrinsic effectiveness
relative to objectives
extrinsic usage
relative to function

INFORMATION RETRIEVAL [TREC]
(document, text)

evaluate on
relevant document delivery

measure effectiveness by
precision, recall

performance for
system as black box

simplify, abstract by
ignoring real detail

study through
controlled laboratory test

BUT
is the user really there ?

IR EXAMPLE

U: 'modern laser printing
technology'

S CASE 1 :

rank 5

'laser technology development
and applications'

including para on printing

rank 10

'modern printing technology'
including para on lasers

paras relevant and similar

what does user think by rank 10 ?

S CASE 2 :

ranks reversed

what does user think by rank 10 ?

(interactive evaluation is tough)

but

focus on relevant retrieval sound

ie JUSTIFIED BY

notion of task CORE

BUT should ask

1. can we get a solid core technology ?
2. how solid is it by now ?
3. how important is core in context ?

IR so far

has solid CORE technology
(statistical ...)

BUT

performance plateau

cf also Gordon & Pathak

==>> ATTEND TO THE CONTEXT

context properties ?

core/context relation ?

ie shift focus to usage

and user operations

SPEECH RECOGNITION

[DARPA]

(continuous)

core competence assumption :

recognition = transcription

=> WER fine as measure

black box evaluation

controlled laboratory testing

technology convergence and advance
(HMMs)

BUT (compared IR)

transcription is SUBtask

implications for core status :

SR necessary, not sufficient,
for task

there's no user there

hidden assumptions :

better subtask performance
means better task performance

want transcriptions
as long term records

evaluation on embedding tasks :

eg SPOKEN DOCUMENT RETRIEVAL

need long term records

BUT retrieval aspect dominates

eg SPOKEN INFORMATION INQUIRY

do not need records

AND dialogue

requires evaluation as whole
allows tolerant interaction

SR EXAMPLE

U: Do you have trains
after six ?

S: Yes, at 7.30 and 8.30.

U: Including Sundays ?

[Do you have drains
after six ?]

Do you have trains
<> ?

S: Sorry, what precisely is
it you want to know ?

U: Trains after six on Sundays.

SR so far

has some core technology

BUT very condition dependent

good enough for some tasks

presumably not for others

==>> ATTEND TO THE CONTEXT

especially :

where do the users go ?

MACHINE TRANSLATION

long evaluation tradition [non DARPA]

evaluate by

input accuracy, fidelity

output propriety, comprehensibility

(semi) objective counting, grading
measures

assumption :

usage-free core (like SR)

BUT

challenge of objective evaluation :

process very complex, open

literal/fair translation

not well-defined

- implies no clear core

SO

encourage evaluation by

acceptability in context

- implies variable core (like SR)

MT EXAMPLE

She borrowed the book and
failed to return it.

? She borrowed the book and
failed to return it.

? She borrowed the book but
failed to return it.

? She borrowed and failed to
return the book.

? She borrowed but failed to
return the book.

MT so far

has successful applications

but little generic technology

in theory could approximate core ?

in reality

==>> ATTEND TO THE CONTEXT

especially :

where do human users go ?

INFORMATION EXTRACTION [DARPA]

evaluation follows SR paradigm

specific fully-detailed answers

'recall', 'precision' as if WER

strong notion of correctness

no reference to context

BUT

IE is not really like SR at all

in conventional IE evaluation

no functional role stated

so

no rationale for answer style

no definition of usage constraints

(not even IR-type reference to
users)

ie no natural defaults

(essential, all facts)

what is the output for ??

eg alerting vs database entry

IE EXAMPLE

The company announced some developments. BoxCo, the packaging subsidiary, would merge with Bags Inc. The subsidiary would enter Bags Inc's Japanese market.

? BoxCo to market in Japan.

? BoxCo and Bags Inc to market in Japan.

IE so far
(has a few applications)

has some tools

overall performance only so-so

BUT

issue not system defects

the notion of core is weak :

performance best on
least controversial subtask (NamEnt)

limitation, disagreement on
subtler subtasks (CoRef, PredArg)

what is information to extract ?

no autonomous core evaluation
(even to extent in MT)

==>> ATTEND TO THE CONTEXT

and primarily the user

AUTOMATIC SUMMARISING

core task as in MT ?

evaluate analogously ?

input concept capture ?

output text coherence ?

task not tightly reducible :

what to capture ?

how to present ?

condensation is not extraction

intrinsically open task

evaluate by 'acceptability' ?

hidden assumptions :

default summary

as essence of source

for like community

BUT

no independent natural summary

so no core (even like MT)

AS EXAMPLE

The tests were conducted on a thousand wombats. They showed half preferred shredded wheat, half rice crispies.

? Wombats have no cereal preferences.

? Wombats will eat breakfast cereals.

? Wombats are like people.

? Huge test on wombats.

AS so far

barely begun

very hard

both practice and principle imply

==>> ATTEND TO CONTEXT

DARPA example good

BUT

only modest start

- undemanding task context

INTELLIGENT INQUIRY

eg SR ATIS [DARPA]

user in core (like IR)

BUT

cannot treat as question/answer :

interaction discourse dependent

communicative behaviour adaptive

ie q/a core too restricted

==>> ATTEND TO CONTEXT

focus on user there

II EXAMPLE

U: Do you have trains
after six ?

S: Yes, at 7.30 and 8.30.

U: Including Sundays ?

[Do you have drains
after six ?]

Do you have trains
<> ?

S: Sorry, what precisely is
it you want to know ?

U: Trains after six on Sundays.

... before ...

COMPARING THE TASKS

shared evaluation style misleading :

large differences -

nature of task core

status of core technology

contribution of user

balance of core and context

eg

MT more automomous core
than IE

IR current core better than
AS current core

IR core evaluation requires
user in way MT does not

AS, MT core competence likely
contributes more to overall
task than IR

etc

CONCLUSIONS FROM REVIEW

once have some core technology :

consider task context

once have established technology :

address task context

NOT just because helpful

BUT because I and L P tasks
demand it

EVALUATION IN FUTURE :

OF COURSE

want some core capabilities

so work on technology for them

BUT

bring in context sooner

at least as

CAREFUL, DETAILED, WIDER,
ENVIRONMENT SPECIFICATION

more properly through

REAL LIVE USERS

IMPACT OF IT EVOLUTION

eg/ie Web

tasks are not disjoint

information management is
seamless :

users mix information actions
treat tasks as subtasks
move between (sub)tasks
opportunisticly

ie leave (sub)tasks incomplete
redefine goals

(eg AS highlighting on the fly
for unique situations)

==>> DYNAMIC CONTEXT CRUCIAL

BUT

(sub)task specification harder

user-oriented evaluation harder

TWO JOKERS :

generalising over applications
hence defining tasks

concatenating less than perfect
systems (no users between)

? ?