Evaluation in Human Language Technology

Maghi King

ISSCO/TIM/ETI

University of Geneva

Two traditions

- Different viewpoints
- Different aims
- Different focus
- Different problems

But sharing

- Common interests
- Common problems
- At least one common dilemma

Different viewpoints

- Define what the software ought to be able to do
  - – investigate how closely it gets to being able to do it

the research tradition

typified by evaluation campaigns

Different viewpoints

- Describe a task which a human wants to achieve
  - – investigate to what extent the software actually helps him in accomplishing the task

the industrial tradition

typified by ISO 9126 and 14598, EAGLES

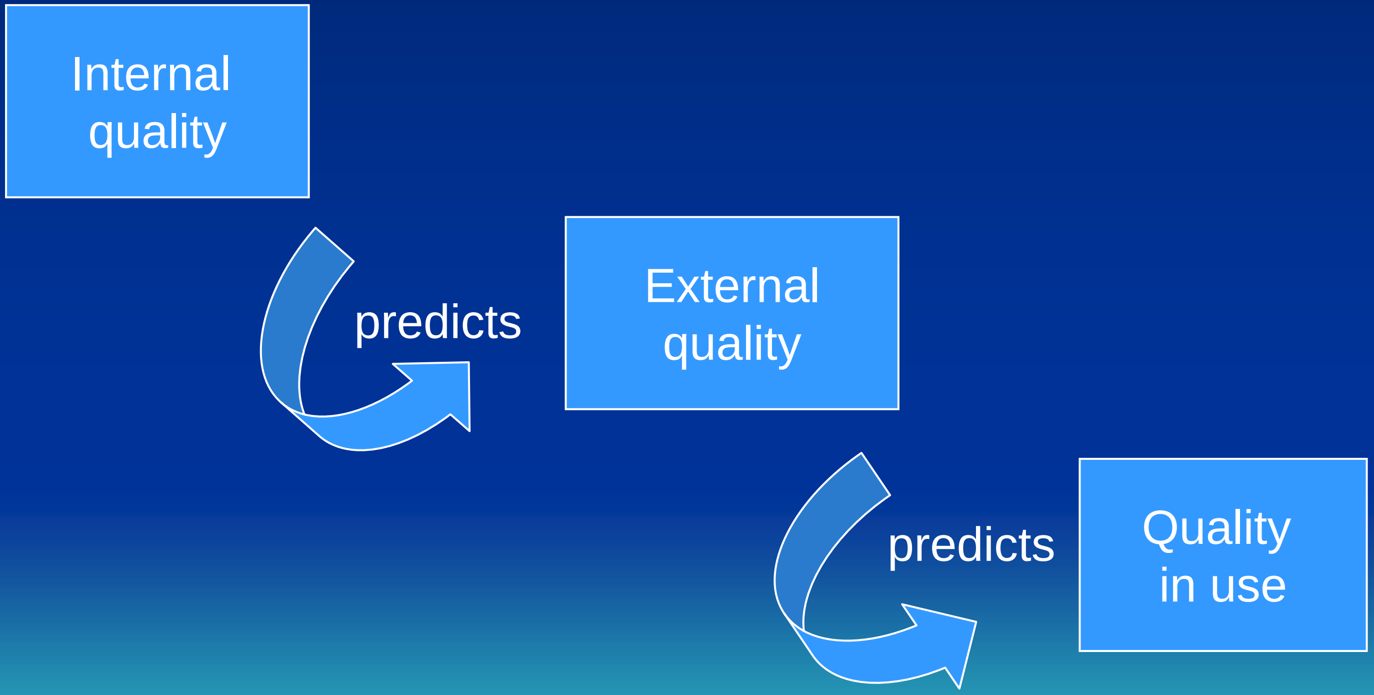
Different aims

- The research tradition
  - Advancing the core technology
- The industrial tradition
  - Quality assurance during production
  - Minimizing investment risk
  - Maximizing return on investment

Different focus

- The research tradition
  - – Concentrate on functionality, and within that on accuracy
    - (do the results meet the specifications)
- The industrial tradition
  - – Concentrate on describing software quality
    - (what does ‘a good software’ mean?)

Good software: the quality chain



Internal quality

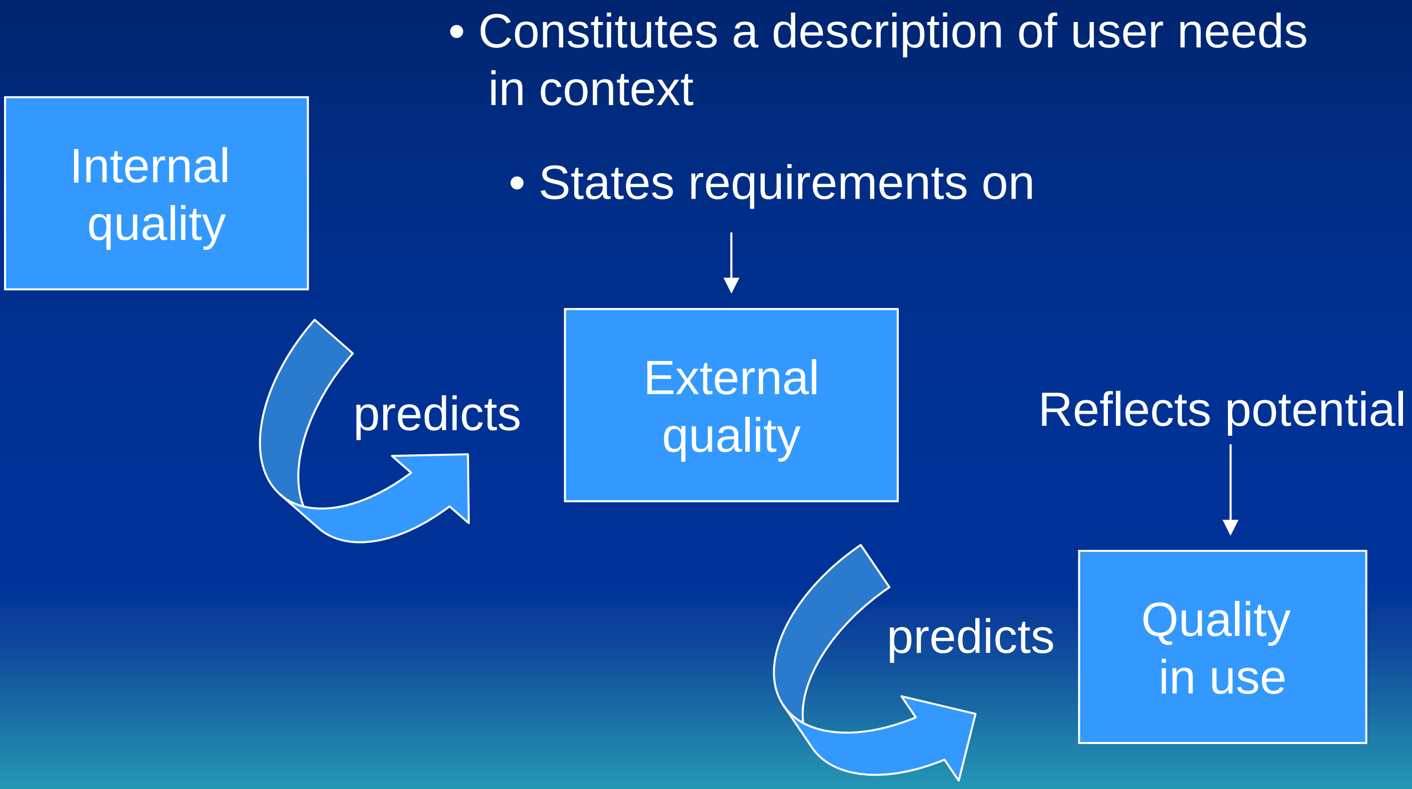
predicts

External quality

predicts

Quality in use

A quality model



- Constitutes a description of user needs in context
- States requirements on

Internal quality

predicts

External quality

Reflects potential

predicts

Quality in use

Different problems

- The research tradition
  - Comparing apples and pears: finding acceptable metrics
- The industrial tradition
  - Generalizing away from a mass of specific and particular contexts: avoiding unacceptable expense

In slogan form

- The research tradition seeks to advance technology

- The industrial tradition seeks to minimize risk and maximize profit in using technology

So are they poles apart?

- Common interests
- Shared problems

Common interests

- The ISO quality characteristics
  - Functionality
  - Reliability
  - Usability
  - Efficiency
  - Maintainability
  - Portability

Relevant to research evaluation

- The ISO quality characteristics
  - – Functionality
  - – Reliability
  - – Usability ?
  - – Efficiency
  - – Maintainability
  - – Portability ?

However:

- Reliability, efficiency are pre-requisites:
  - Only tested indirectly

- Maintainability

  (analysability, changeability, stability, testability)

  - Tested directly, but between evaluations

So the difference is a task to be done?

- Can’t be true!
  - Choice of what to evaluate in the research tradition depends on what is assumed to contribute to achieving a generically useful task
  - Industrial tradition starts from a specifically useful task

So the difference is including the user?

- Can’t be true!
  - A task – generic or specific - implies a user
    - The research tradition makes assumptions about the user
    - The industrial tradition uses knowledge about specific users

So, is there any real difference?

- Only that:
  - The research tradition (rightly) works on the level of what would be useful at a very general level
  - The industrial tradition works on the level of what would be useful in a particular situation

So, is there any real difference?

- And that:
  - The research tradition directly tests functionality (accuracy)
    - Evaluation campaigns typically allow for improvement cycles, so
    - other quality characteristics are tested indirectly
  - The industrial tradition thinks in terms of one-off evaluations taking account of a particular context
    - All relevant quality characteristics have to be tested for explicitly

And just one fundamental difference

- Questions of suitability (sub-characteristic of functionality) are not pertinent in the research tradition
  - And therein lie the roots of a shared dilemma

The roots of the dilemma: metrics

Both traditions rely critically on being able to find good metrics

Good metrics

- Valid
- Reliable
- Objective
- Economical
- Informative

Comfortable cases

- The task is (relatively) simple, accuracy and suitability co-incide, e.g.
  - Word error rate in a dictation system
    - Modulo vocabulary known to the system
  - Precision/recall in a document retrieval system
    - Modulo a manageable pool of documents
    - Modulo agreement on relevance judgements

Increasing discomfort

- Suitability begins to outweigh accuracy, e.g.
  - Word error rate in dialogue systems
  - Lexical/terminology coverage in translation systems
  - String extraction in term extraction systems
    - (not all words are equal)

Increasing discomfort

- Metrics become heavily resource dependent, e.g.
  - Creating relevance judgements for document retrieval systems working over a large document collection
  - Creating templates for fact extraction systems
    - Making gold standards is expensive
    - Expense prevents change of focus (research tradition)
    - Evaluation becomes unacceptably expensive (industrial tradition)

Common problems

- Objectivity becomes suspect, e.g.
  - Relevance judgements obtained by pooling results of several systems

And yet more common problems

- Validity becomes suspect, e.g.
  - Gold standard material does not match intended real application (BLEU, NIST …)
  - Metric is executed over a finite and stable data collection when real application works over much larger and unstable data collection (using a ‘snapshot’ of the web …)

More validity problems

- Humans get involved
  - In defining the gold standard (e.g. reference translations)
  - In executing the metric (e.g. information retrieval through web searching)

The shared dilemma: extreme discomfort

- Systems where
  - system performance and human performance cannot be separated out
  - the application by definition works over vast amounts of data which no human could master or analyse
  - the data is by definition constantly shifting

Symbiotic systems: some examples

- Document retrieval on the web
- Information retrieval on the web
- Data mining systems
- Text mining systems

  – i.e. most of the emerging human language technologies!

Summary

- We have learnt a great deal
- We have a much better understanding of what we want
- We are faced with new and difficult challenges

A question for this workshop:

- How can we build on what we have learnt in order to
  - deploy effectively knowledge and experience gained
  - share experience and insights as they develop
  - build bridges to other evaluation communities
  - meet new challenges