Evaluation of NLP systems

Diana Santos
SINTEF Telecom and Informatics

Plan

1. Introduction
2. Basic concepts

[3. NLP seen through applications]

4. Evaluation in detail in several domains
5. Further reading

Motivation

- *Evaluation is itself a first-class research activity: creation of effective evaluation methods drives more rapid progress and better communication within a research community* (Hirschman, 1998:302f)
- *[Before] there were no common measures and no shared data. As a consequence, systems and approaches could not be precisely compared and results could not be replicated* (Gaizauskas, 1998:249)

Lack of evaluation history

- In linguistics and the humanities in general
- In computer science
  - *There are plenty of computer science theories that haven't been tested* (Tichy, 1998:33)
  - *I'm not aware of any solid evidence showing that C++ is superior to C with respect to programmer productivity or software quality* (Tichy, 1998:35)
  - *the standard of empirical research in software engineering [...] is poor* (Kitchenham et al., 1999:1)

Uses of evaluation

- Internal progress measuring
- Internal "scientific" project management
- External progress reporting (state of the art)
- For requesting funding
- Funding rationale of funding agencies
- For marketing purposes
- For customer protection

Kinds of evaluation

- Pure research setup (verification of an hypothesis)
- Technical evaluation
- User testing
- Market testing

Forms of evaluating

- Absolute: A (set of) meaningful number(s) or other consensual objects
- Relative: Comparison with other systems
- User-related: Proof of satisfaction by the user community
- Financial: Analysis of profits

Methods of evaluating

Given an hypothesis,

- Analytical (hypothetic-deductive)
  logically proving its correctness
- Empirical
  - experiments
  - measures
  - decision treshold

What to evaluate

- the theory(ies) underlying the system
- the system
- the setup of the system (system plus its social context)

Relative evaluation: Comparing with...

- all systems
- the best system
- previous versions of the same system

- the baseline case
- upper and lower cases

- a control case
- training and test

Evaluation points

Gaizauskas (1998:252) distinguishes between *user-visible* vs. *user-transparent* tasks :

User-transparent	**User-visible**
• sentence alignment	• translation
• parsing	• grammar checking
• query expansion	• question-answer pair
• generation	• translation
• world model	• actions by a robot

The “box” view

- Black box evaluation : assess the system in terms of inputs/outputs
- Glass box evaluation : look at the contributions of parts of the system to its overall performance

Evaluation decisions

- Conflicting analyses
- Indeterminacy
- Non-mutually-exclusive classification

We prefer acceptable tags to correct tags … We do not consider wrong the cases that are hard to decide between linguists, we will only count consensually flagrant errors (Valli & Véronis, to appear, p.18)

100% correct?

- 70% to 85% interannotator agreement between two expert annotators in the English micro-electronics domain (Will, 1993)
- Theoretical disagreement
- Interpretation disagreement

Indeterminacy / vagueness

- an intrinsic property of natural language as a communicative system (Santos, 1997)
- how to encode it?
  - how to evaluate a choice not required?
  - accept two or more classifications?
  - require two or more classifications?

Non-mutually-exclusive classes

- portmanteau tags
- hyponym/hyperonym relations
- dependencies between the results
- classification of words and of sentences - what is the unit?

Evaluation related concepts

- Measures
- Resources
- Evaluation contests
- Test setups
- User involvement
- Publication

Measures for evaluation

- Unambiguously defined
- Context-independent
- Replicable
- Computable

- Consensual
- Related to the main goal of the system

Resources for evaluation

- Public
- Available

- (Annotated) corpora (*object* corpora)
- Test suites
- Treebanks
- Manually created solutions (*meta* corpora)

What is a treebank?

- A syntactically analysed corpus (generally in the form of trees)
- Revised and corrected
- Reflecting a consensus
- Publicly available
- Well documented; maintained; with associated tools
- With an evaluation purpose

History of treebanks

- The PENN treebank
  - SUSANNE
  - The Lancaster treebank
- The Prague dependency treebank
- The NEGRA project
- The TALANA annotated corpus
- Spanish, Chinese, Polish, Italian ...

What is a "test suite"?

- Prepared materials (not "real" text)
- systematic coverage (for a particular purpose)
- non-redundant
- including negative data
- control over test data

Oepen et al. (1998)

Evaluation contests

- Conferences organized around the test of a particular application
- Participant systems compete on the same set of tasks
- Participants agree and take part on the evaluation procedure
- There is a tight schedule for the whole setup

History of evaluation contests

1987 Speech Recognition Benchmark tasks

1987 MUC (message understanding conference)

1989 ATIS (Air Travel Information System) - spoken language understanding

1992 TREC (text retrieval conferences)

1992 Parseval (evaluation of parsers)

History of evaluation contests (cont.)

1998 SENSEVAL (word sense disambiguation)

MUC's

- 1st MUC (1987)
  - common corpus with real message traffic
- MUC II (1989)
  - introduction of a template
  - training data annotated with templates
- MUC 3 (1991)
  - newswire text
  - semiautomatic scoring mechanism
  - collective creation of a large training corpus

MUC's (cont.)

- MUC 6 (1995)
  - domain independent metrics
  - introduction of tracks
    - named entity
    - co-reference
    - template elements
    - template relation
  - emphasis on portability

Hirschman (1998)

(Partially) manual solutions

- human judges and/or annotators (as in MUC)
- a pooling of the results of several systems (as in TREC)
- some artificial automatic change to natural text (ablation studies)
- natural speech in artificially controlled settings
- human actors (Wizard of Oz)

Evolution / maintenance

- *it is necessary to evolve benchmarks to prevent overfitting* (Tichy, 1998:36)
- *evaluation must keep pace with technology developments* (Hirschman, 1998:302)

Test setups

- Devising a series of experiments to get at non-trivial results
- Setting up detailed test conditions, separate from the development process

- Benchmarks
- Wizard of Oz; User observation/logging
- Judge management

Rationale for introducing usability

- *Productivity within the service sector would rise 4% - 9% annually if every software program were designed for usability* (Landauer, 1995)
- According to a 1995 study by the Standish Group the three main factors in a successful software project are: user involvement, support from executive management and clearly formulated requirements. (http://www.standishgroup.com/chaos.html)
- *The implementation of usability engineering techniques has demonstrated a reduction in the product development cycle by 33% - 50%* (Bosert, 1991)
- *Every $1 invested in user-centred design returns between $2 and $100* (Pressman, 1992)

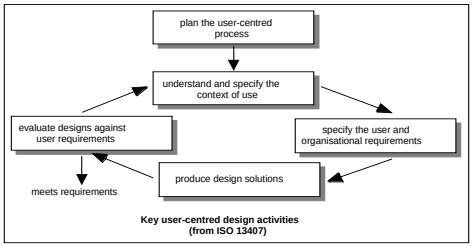
SINTEF's HCI group

User-centred design implies

- early focus on users, tasks and environment
- the active involvement of users;
- an appropriate allocation of function between user and system;
- the incorporation of user-derived feedback into system design;
- iterative design whereby a prototype is designed, tested and modified;

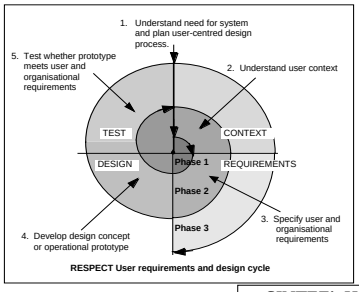
SINTEF's HCI group

Key user-centred design activities



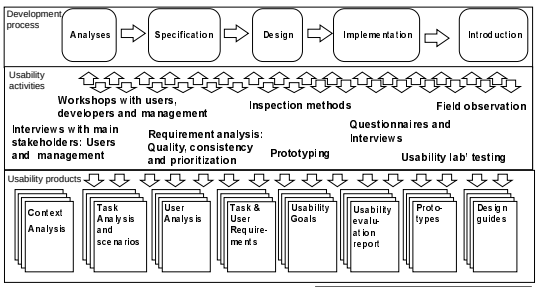
SINTEF's HCI group

User-centred design process



SINTEF's HCI group

Usability methods applied in system development



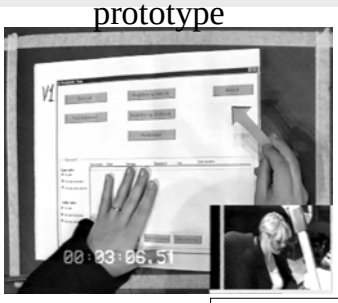
SINTEF's HCI group

Methods and techniques

- Brainstorm
- Card sorting
- Controlled testing
- Diary keeping
- Focus groups
- Functionality matrix
- Group discussion
- Interviews
- Modelling
- Observation
- Paper prototyping
- Parallel design
- Rapid prototyping
- Scenario building
- Storyboarding
- Survey
- Task analysis
- Task allocation
- Walkthrough
- Wizard of Oz prototyping

SINTEF's HCI group

Usability testing of paper prototype



SINTEF's HCI group

User involvement

- Elicitation of user requirements
- Evaluation of user friendliness (e.g. with the help of paper prototyping)
- In spoken dialogue systems, WOZ
  - tests behaviour implementable
  - collects sample dialogues

(Evaluation results can often be used later in system development)

Evaluation of publication

Armstrong (1982) studied
objectivity, replicability, importance, competence, intelligibility and efficiency
as criteria for evaluating publication (policy)

Frequently, authors select unimportant problems, advocate a dominant hypothesis, avoid challenges to current beliefs, provide less than full disclosure, use complex methods, or write in a obscure manner (Armstrong, 1992:97ff)

Armstrong (1982)'s findings

- Referees are seriously biased by knowledge of
  - who wrote the artice
  - whether the conclusions of the paper conform with their beliefs (p.85)
- 12 out of 30 replications yielded results that conflicted with the original study (p.88)
- Journals do not seem to give preference to important problems (p.90)

Publications in the NLP discipline

- Is publication of research around an NLP system a good indicator about its quality?
- Does quality of a paper reflects the quality of a system?
- Does the citation outcome provide indirect evidence for its importance or relevance?
- Do NLP publications focus on evaluation?

NLP seen through applications

Part II

Based on Santos (to appear)

An overview of NLP

- Text production
- Reading & browsing
- Translation
- Teaching & learning
- Information systems
- Interactive systems
- Indexing
- Data input
- Security and identification
- Others

Writing / creating text

- (spelling, grammar, style) checking
- authoring environments
- controlled language verifiers
- automatic documentation
- digitation help
- dictation machines
- formaters
- communication with seriously handicapped

Reading & browsing

- Information extraction
- Summarization
- Text mining
- Automatic reading aloud
- Guided tours
- Multilingual information retrieval

Translation

- Machine translation
- Translation helpers
- Gisting browsers
- Translator's workstations
- Speech translation
- Real-time captioning

Teaching and learning

- Tutoring systems
- Intelligent enciclopedias
- Didactical games
- Automatic exercise creation

Access to information

- Telephone information systems
- Questions to enciclopædias
- Web search
- Automatic advisors (expert systems)
- Personalized publication

Interactive systems

- Spoken commands
- Interaction with automatic assistants
- Animated simulation of instructions
- Automatic reporting
- Interactive games

Indexing

- semi-automatic thesauri creation
- indexing large quantities of material
- image cataloguing
- semi-automatic creation of (bilingual) terminology

Data input

- Automatic information extraction as input to internal databases
- Optical reading
- Data input by speech
- Handwriting recognition

Security and identification

- Signature identification
- Obedience to “his master’s voice”
- Language /dialect recognition
- Author identification

Other

- Sociological studies
- Vocabulary analysis of a particular science
- Measures of the “weight” of a culture
- Metaphor selection for a good interface
- Language teaching planning

Examples of evaluation

Part III

Spelling checkers

- Processing speed (verification speed; average answer time; suggestions / second)
- Result accuracy
  - suggestion dispersion (sugg./word; sugg. factor)
  - suggestion ordering (ordering factor)
  - failure (missing, zero, completeness, robustness)

$F_i=N_f/N_s$ $F_z = N_z/N_e$ $F_c=1-N_d/N$ $F_r=1-N_{in}/N_e$

- Functionality

Medeiros (1996)

Spelling checkers (cont.)

- Correction index $\frac{(F_i + F_z)}{(F_d + F_o + 2F_r)}$
- Test materials
  - Free text
  - List of errors
  - List of pairs (error, correction)
  - List of error-free words ordered by frequency in real-text corpora (Underwood et al., 1995)

Style checker: CRITIQUE

- 10 randomly selected essays from each of the four groups: Freshman, Business, ESL and Professional writing
- classified the critiques as *correct*, *useful* and *wrong*
- did not consider errors that were missed
- adjusted the analysis in both directions

Richardson & Braden-Harder (1993)

Ranking alternative parses

- 80 multiple parse sentences (2 to 6 parses)
  - right in 65% of the cases
  - if wrong parses were taken out of consideration, right in 80% of the cases
  - more than 10% of the failures required semantic information for accurate decision

Heidorn (1993)

Disambiguating verb definitions

a criterion by which to measure the accuracy of judgements of correctness

- 58 randomly selected definition strings whose interpretation [by the program] we judged correct and assigned each of them to 3 or 4 different participants
- 44 almost unanimous agreement (41 agreed with the program); 11 split (10 agreed); 3 large variation

Ravin (1993)

Information retrieval

- Precision: how many hits were right?
- Recall: which percentage was recovered?
- F-measure: $\frac{2*P*R}{P+R}$

N.B. PR measures presuppose a universe where there are more candidates to hits than items selected/classified by a system

PR for all?

- Sentence alignment (Simard et al., 2000)
  P: given the pairings produced by an aligner, how many are right
  R: how many sentences are aligned with their translations
- Anaphora resolution (Mitkov, 2000)
  P: correctly resolved anaphors / anaphors identified by the system OR all anaphors
  R: correctly resolved anaphors / anaphors attempted to be resolved
- Parsing accuracy
  100% recall in CG parsers ... (all units receive a parse)

Summarization methods

- task based: information retrieval and text categorization, 8 summarization methods
- 30 subjects; 10 queries; 20 summaries per query (subjects can consult the full text)
- (subjective) relevance; time spent per query; (four-graded) readibility
- study significance of agreement among judges

Mochizuki & Okumura (2000)

Parser evaluation

- Listing linguistic constructs
- (Corpus) Coverage, Average Parse Base, Entropy
- (Annotated corpus) POS assignment accuracy, structural consistency, ranked consistency, tree similarity, GEIG scheme, dependency structure scheme, GR scheme

Carroll et al. (1998)

Evaluation of PP attachment

- Given a surface parser for English (Fidditch) and 13 million words AP (2,661,872 NPs)
  - 467,920 objects
  - 753,843 followed by preposition
  - 223,666 were in V NP PP class
- Lexical association score (v,n,p)

$$\frac{P(\text{prep assoc. noun} \mid \text{verb, noun})}{P(\text{prep assoc. verb} \mid \text{verb, noun})}$$

Hindle & Rooth (1993)

Speech synthesis (TTS)

- Text analysis
- Intelligibility
- Syllable articulation (nonsense words)
- Word intelligibility (semantically unpredictable sentences)
- Sentence intelligibility (first read the questions)
- Overall quality

Itahashi (2000)

Analysis-Modification-Synthesis systems (AMSS)

- Glassbox evaluation, human performance as ultimate target (recording speakers)
- elementary tasks
  - manipulate melody, tempo, intensity
  - modify spectral quality
- several distortion measures
  - principal component analysis to select the best combination of results

Bailly et al. (2000)

Dictation systems

- systems that continuously evolve towards the needs and properties of user and the documents produced
- researchers must use off-line accuracy; users are concerned about time and effort it takes to create a finished document
- 15 professional EU translators (5 languages), 75% of translations generated by voice in a 3-months period with a commercial discrete-utterance SRS
- subjective response: 2/3 "Good" or "Very Good", microphone problems, special EU vocabulary, prefer typing for editing, …

Hunt (1997)

NL Generation

accuracy | readibility | task-based | glass box

- simplicity a la controlled languages should not be used as measure - use human subjects
  - comprehension time
  - time taken to post-edit the text
  - direct grading
- for particular tasks inside an NLG
  - create inputs from (processed) real text

Mellish & Dale (1998)

Spoken dialogue systems

- Objective measures
  - % correct answers (set of reference answers)
  - % successful transactions (completed tasks)
  - number of turns
  - dialogue time
  - mean user/system dialogue time
  - % diagnostic error messages
  - % of non-trivial (more than 1 word) utterances

Walker et al. (1998)

Spoken dialogue systems (cont.)

- Subjective measures
  - % implicit/explicit recovery utterances
  - % contextually appropriate system utterances
  - cooperativity
  - % (partially) correct answers
  - % appropriate system directive/diagnostic utterances
  - user satisfaction

Sentence alignment

- What is the baseline?
- What are the units that are being measured?
- What is the result of the annotation?
- What is the contribution to the application the aligner is embedded in?
- What is the import of the languages in question? Evaluation relative to language!

Santos & Oksefjell (2000)

Machine translation

Linguistic, operational and economic assessment

- post-editor monitoring
- task performance by a human reading the source or the translation
- back-translation
- direct and indirect product costs, setup in operation

Sparck Jones & Galliers (1996)

Multilingual gisting

- Evaluate the extent to which a gisting method helps the user to make decisions
  - Creating topical categories (from 32/73 listings)
  - Categorization task
    - control (English cards)
    - experimental (gisted cards)
    - random baseline
  - Analysis (differences between subjects)

Resnik (1997)

Question answering

TREC-8 (domain-independent Q&A systems)

- selecting test questions (127 orgs, 49 judges, 24 FF)
  - straightforward (no list, no ambiguous or tricky, solution in the document collection)
- judging answer strings

  .641 mean overlap across all 3 judges for the 193 test questions that had at least 1 correct string found
- stability of QA evaluation

Voorhes & Tice (2000)

User expectations

- What is the best tool for CALL?
- A talking machine (Atwell, 1999)

Food for thought

- The more complex evaluation procedure(s)
  - ? the more mature is the (sub)area
  - ? the easier is to progress
  - ? the easier not to reinvent the wheel
  - ? the easier to know what one is doing
- Start developing/describing your system by suggesting a way to evaluate it ...

References

The references for the present tutorial can be found on the Web at

http://www.portugues.mct.pt/Diana/bitTutEval.html