

## *A Model of Evaluation in Natural Language Processing*

---

Andrei Popescu-Belis  
ISSCO - ETI - Université de Genève

April 19, 2001

## *Plan*

---

- 1. The problem
- 2. A formal framework
- 3. Coherence of quality measures

2

### *1. The problem: plan*

---

- 1. The problem
  - 1.1 What is evaluation
  - 1.2 Evaluation vs. verification
  - 1.3 Evaluation projects and campaigns
- 2. A formal framework
- 3. Coherence of quality measures

3

### *1.1 What is evaluation*

---

- EVALUATE v.t. = to determine or set the value of.  
VALUE n. = the relative worth, merit or importance.  
[Webster's]
- What is susceptible or evaluation?
  - A research direction
  - A technology
  - The application of a technology to a domain
  - The impact of a theory
  - A system
  - A language resource

4

### *1.2 Evaluation vs. verification (1)*

---

- Software engineering
  - specification / implementation / "verification & validation"
  - specifications must be precise, complete, verifiable
  - verification: does the result conform to the specifications?
- Language engineering / NLP systems
  - problems to be solved are hard to specify formally
  - the *format* of the output is specifiable
  - the *content* of the output is described using natural language

5

### *1.2 Evaluation vs. verification (2)*

---

- Exemple of an MT system or module
- Verification : determine whether the output conforms to the specifications, e.g., consists only of words from the target language – « always »
- Evaluation : determine whether the output is an acceptable translation of the input – « in general »

⇒ Evaluation of NLP systems : estimate to what extent non formalizable specifications are satisfied by a system

6

### 1.3 Evaluation projects and campaigns

---

- Projects about evaluation methods
  - aim at defining techniques and standards
  - mostly European: ELSE, EAGLES, TEMAA, TSNLP, DIET, FRACAS
- Evaluation campaigns
  - USA: MUC, TREC, TDT, SUMMAC
  - Europe : GRACE, SENSEVAL/ROMANSEVAL, JST-FRANCIL (French)
- Some www links on the workshop's website

7

### 2. A formal framework: plan

---

- 1. The problem
- 2. A formal framework
  - 2.1 Context
  - 2.2 The ISO 9126 standard
  - 2.3 A three stage evaluation model
  - 2.4 Computing measure M
  - 2.5 Framework for measure m
  - 2.6 Applications
- 3. Coherence of quality measures

8

### 2.1 Context

---

- Estimate the quality of an NLP system
- Quality with respect to...
  - a given task (*to be defined*)
  - a given category of users (*to be defined*)
  - economic factors: ~price
- Preferences
  - automatic measure
  - numeric measure

9

### Goal: be more formal than...

---



10

### 2.2 The ISO 9126 standard (1)

---

- Standard of definitions for software evaluation
- Software quality: decomposed into six characteristics
  - functionality
  - reliability
  - user-friendliness
  - efficiency
  - maintenance
  - portability

11

### 2.2 The ISO 9126 standard (2)

---

- Three phases defined for the evaluation process
  - definition of the required qualities, preparation, procedure
- Three stages in phase II / III
  - ▲ For each attribute
    - definition of a metrics / measurement
    - definition of scores / scoring (or rating)
  - ▲ For all the relevant attributes
    - definition of synthesis criteria / assessment (or summarization)

12

### 2.3 Three stage evaluation model (1)

- Formalize the ISO stages for NLP systems evaluation
- Define and criticize evaluation protocol
- Evaluation of capacities  $C_1, C_2, \dots, C_n$  of a given system to process input data

13

### 2.3 Three stage evaluation model (2)

#### 1. Measure



#### 2. Rating (scoring)



#### 3. Synthesis



14

### 2.3 Three stage evaluation model (3)

- **Measure**  $M : \{C_i \mid C_i \text{ capacity}(S)\} \rightarrow [0; 1]$   
 $C \rightarrow \mu(C) = V_C$
- **Rating**  $R_C : [0; 1] \rightarrow \{s_1, s_2, \dots, s_p\}$   
 $V_C \rightarrow R_C(V_C) = S_C$
- **Synthesis**  $S : \{s_1, s_2, \dots, s_p\}^k \rightarrow \{s_1, s_2, \dots, s_p\}$   
 $(S_{C_1}, \dots, S_{C_k}) \rightarrow S(S_{C_1}, \dots, S_{C_k}) = S_S$

15

### 2.4 Computing measure M

- Direct measure of a capacity is generally not possible (exception, e.g., dictionary size)  $\boxtimes$  other method?
  - Input data  $\Delta = \{D_1, D_2, \dots, D_n, \dots\}$
  - **Theoretical measure:** average on all possible data  
 $M(C) = \phi [ m(D_1), m(D_2), \dots, m(D_n), \dots ]$
  - **Estimated measure:** average on  $\Delta_{test} = \{D_1, D_2, \dots, D_k\}$   
 $M(C) \approx \phi [ m(res(D_1)), m(res(D_2)), \dots, m(res(D_k)) ], k \text{ is small}$
- ↙ We must compute  $m(res(D_k))$  and choose  $\Delta_{test}$

16

### 2.5 Framework for measure m

- Direct computing of  $m(res(D_k))$  is impossible - otherwise the system could use it as a clue
- Generally,  $m(rep(D_k))$  is computed from the distance between  $rep(D_k)$  and a correct / desired response
- Conditions for measure
  - describe representative input data  $D_1, D_2, \dots, D_n$
  - define the correct / desired response  $K(D_i)$  for each  $D_i$
  - describe all possible answers  $RES(D)$ ; let  $RES$  be the total set
  - define a quality measure, i.e. a computable function  
 $m : RES \rightarrow [0; 1]$   
 where  $m(res(D_i))$  measures the distance between  $res(D_i)$  &  $K(D_i)$

17

### 3. Coherence of quality measures: plan

1. The problem
2. Proposition d'un cadre théorique
3. Coherence of quality measures
  - 3.1 Ideal view of mesures
  - 3.2 Graphic representation
  - 3.3 Formal coherence criteria
  - 3.4 Comparison of measures

18

### 3.1 Ideal view of measures

- Hypothesis
  - each characteristic or capacity has an "objective quality" in the evaluation conditions (user, task, etc.)
  - therefore, each response  $res(D)$  also has an "objective quality"

➡ An evaluation measure  $m$  determines the "objective quality" using a set of scores  
 $m : \{ \text{"perfect", "average", "null", ...} \} \rightarrow [0; 1]$

19

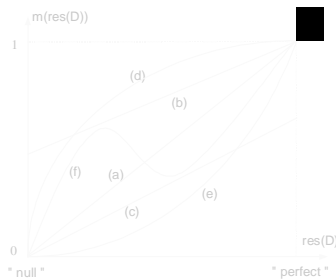
### 3.2 Graphic representation (1)

- Horizontal axis  $x$ 
  - objective quality of a system's response
- Vertical axis  $y$ 
  - numeric scores
- Criteria about the  $x$  axis
  - subject to *argumentation*
- Criteria about the  $y$  axis
  - subject to *proof*
- The following representation is purely orientative

20

### 3.2 Graphic representation (2)

- Possible aspects of a measure:
  - (a) ideal
  - (b) incomplete +indulgent
  - (c) incomplete +severe
  - (d) indulgent
  - (e) severe
  - (f) incoherent



21

### 3.3 Formal coherence criteria

- Upper limit (subject to argumentation)
  - (UL)  $m(res(D)) = 100\% \Leftrightarrow res(D) \in K(D)$
- Lower limit (subject to argumentation)
  - (LL)  $m(rep(D)) = 0\% \Leftrightarrow rep(D) \text{ is very bad}$
  - (LL-1)  $m^{-1}(0\%) \subset REP_M(D)$
  - (LL-2)  $m^{-1}(0\%) \supset REP_M(D)$
  - (LL-2-fuzzy) *bad responses must receive low scores*
  - (LL-3)  $m^{-1}(0\%) \neq \emptyset$
  - (LL-3-fuzzy) *minimal scores must be low*

22

### 3.4 Comparing measures

- Uniformity of a measure (argumentation)
  - (UN)  $m$  is uniform iff  $\forall D, \forall res_1(D), \forall res_2(D)$ ,  
 $[res_1(D) \text{ "better" than } res_2(D)]$   
 entails  $[m(res_1(D)) \geq m(res_2(D))]$
- Indulgence/severity of two measures (provable)
  - (IS)  $m_1$  is more indulgent (lenient) than  $m_2$  iff  
 $\forall D, \forall res(D), m_1(res(D)) \geq m_2(res(D))$
- Absolute indulgence/severity?

23

### Conclusion

- Theoretical framework open to discussion
- Difficulties
  - find a way to derive measurable characteristics (e.g., relate them to measurable ones, maybe related to other tasks)
  - find measures for various characteristics (same thing?)
- Meta-evaluation criteria: useful to choose among measures
- For MT systems: combine this with the ISLE taxonomy

24