Where are we with evaluation?

Maghi King

TIM/ISSCO
School of Translation and Interpretation
University of Geneva

“Building the LR&E Roadmap"
29th May,
LREC 2004

UNIVERSITÉ DE GENÈVE

ISSCO
University of Geneva

Foresaking all others, lets concentrate on

Functionality

- Accuracy
  - does the system meet its own specifications
- Suitability
  - does the system give results the user wants

An easy example: terminology extractors

- Accuracy: list all sequences of two words or more repeating two times or more

- Suitability: useless (gives far too much noise)

Accuracy will do

- If some core functionality coincides with needs of a wide range of users
- If there's a way to define the 'right' answers

Examples

- Speech recognition
- Spelling checkers
- Fact extraction
- Tagging
- …

When accuracy isn't enough

- And suitability starts to be an issue
  - E.g. with document retrieval

Sometimes possible to construct right answers anyway, although subjectivity and dispute begin to creep in

New applications

- Accuracy isn't enough
- It's hard (impossible?) to construct right answers because
  - The corpus being treated is too large
  - The corpus being treated is inherently unstable
  - Users' interests don't coincide

    Example: research engines

So where are we?

- We’re quite good at
  - Applications where user satisfaction can be made to hang on accuracy
- We’re not totally bad at
  - Applications where we can supply a definition of suitability by constructing a gold standard

Where are we?

- Only just beginning to think about
  - applications where
    - accuracy is not enough,
    - we can’t find a gold standard,
    - we can’t identify groups of users whose interests coincide

Time scales?

- No idea – much too early to tell