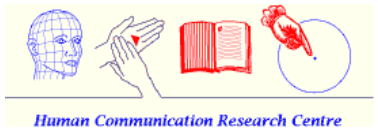


Some problems in evaluating multimodal systems

Jean Carletta

University of Edinburgh

AMI Project

Human Communication Research Centre

AMI Meeting Rooms

4 close- and 2 wide-view cameras, 4 head-set and 8 array microphones, presentation screen capture, whiteboard capture, pen devices, plus extra site-dependent devices.

TNO

Edinburgh

IDIAP

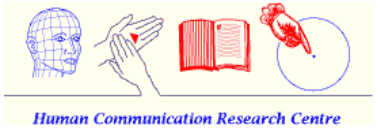
Human Communication Research Centre

Annotations

- location of person on video to aid tracking
- low level timestamping against signal
  - movement around room; emotion coding; some head and hand gesture; focus of attention
- orthographic transcription w/ timing at "segment" level and forced alignment
- discourse structure over orthography
  - dialogue acts w/ addressing, named entities, topic segments, linked abstractive and extractive summaries

Implications

- multimodal in using multiple capture devices of different types for the same basic events
  - even synchrony and determining what signals to read for what processes is hard
  - multiple people being recorded
- browser technologies themselves are multimodal
- many, many possible things to be evaluated
- highly interdisciplinary, but what binds the groups together is the vision of a system



Things we do know how to do (more or less)

Component evaluation

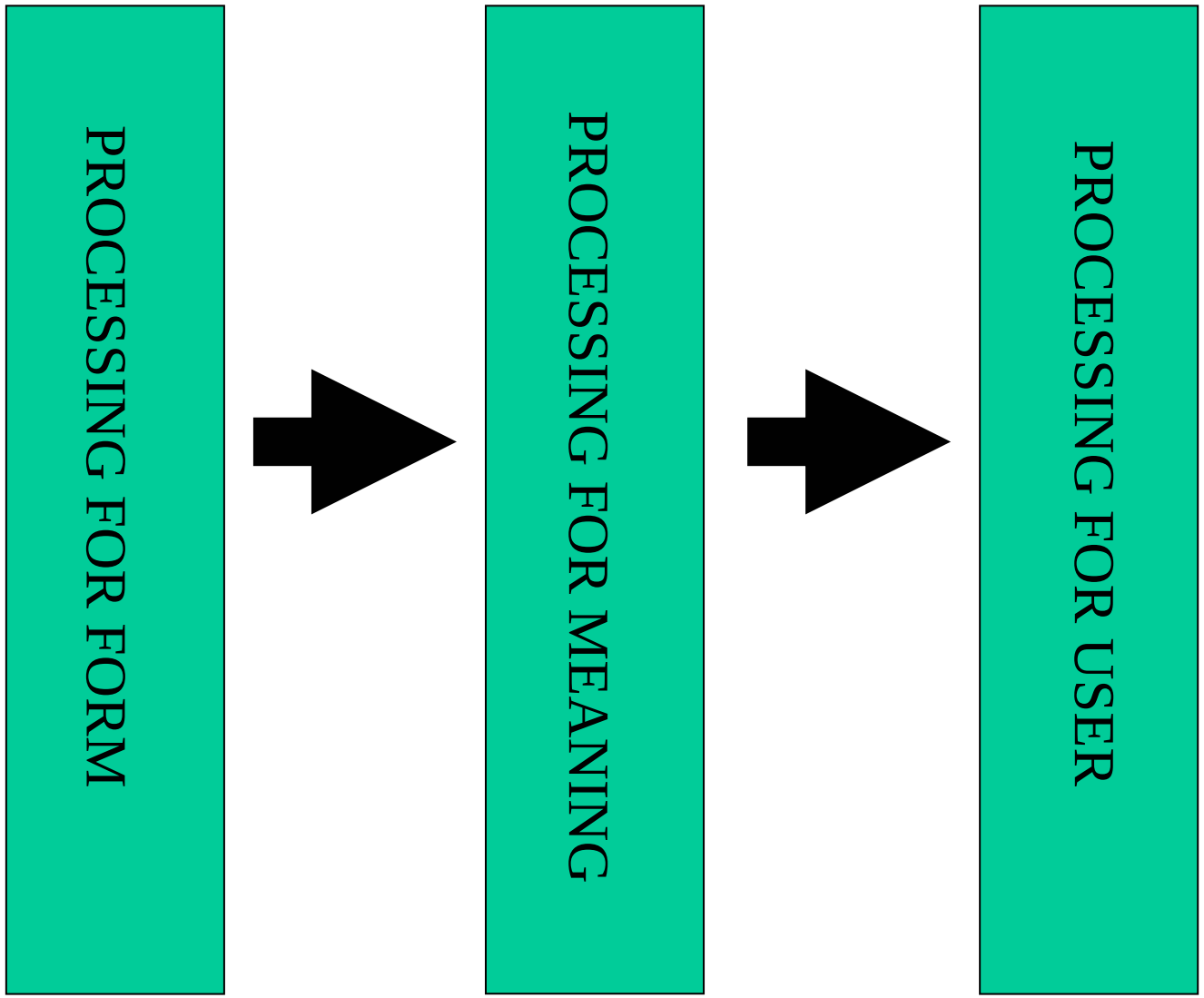
- Hand-annotate some data for "ground truth"
- Develop some statistical measure for differences between component output and the ground truth
- Improve component by improving on measure.
- Compete.
- In practice, community finds **any** performance improvement exciting, no matter how small.

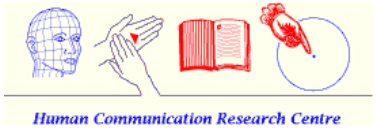
CHIL technologies: forthcoming benchmarking (March-May 2006)

- Audio technologies
  - CTM Speech Recognition
  - FF Speech Recognition
  - Acoustic Person Tracking (in space)
  - Acoustic Speaker Identification
  - Speech Activity Detection
  - Acoustic Event Detection
  - Acoustic Environment Classification
  - Acoustic Emotion Recognition
- Content processing
  - Questions Answering
  - Automatic Summarization
- Vision technologies
  - Face & Head tracking
  - Visual Person Tracking
  - Visual Speaker Identification
  - Head Pose Estimation
- Multimodal technologies
  - Audiovisual Speech Recognition
  - Multimodal Person Identification
  - Multimodal Person Tracking

Human Communication Research Centre

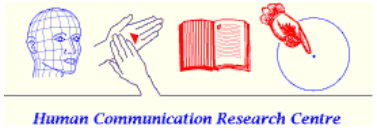
Typical processing flow





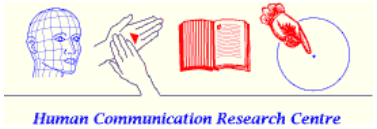
Low level processes for form

- ASR
- Affect ("is it good, or bad?")
- Head, Hand, and Body Movement
- Localization and Tracking
- Person Segmentation and Identification



High level processes tending towards meaning

- Topic segmentation and labelling
- (Meaningful) gesture
- Dialog act segmentation and classification
- Addressing
- Named entity recognition
- Extractive summarization
- Abstractive summarization
- Indexing/Retrieval
- Syntactic chunking
- Focus of attention



Problems for evaluation

Problem 1: Component performance ≠ system performance

- The only real evaluation is extrinsic - does the system work for its intended users doing the intended task?
- Need component evaluation to aid development, but users can't evaluate components
- Improving a component doesn't necessarily improve the overall system evaluation; don't know when to stop investing in improvements
- Need system evaluation to aid development, but using extrinsic evaluation "in the loop" is expensive.

Problem 2: Even evaluation of simple things can be hard

- Currently, the community only evaluates the simplest tasks that are easiest to measure
- Example: CLEAR evaluation of 3D tracking
  - metric is easy to devise, based on Euclidean distance; ground truth marks the location
  - BUT what about other (more useful?) tasks on the same type of data, like "where is the person looking"?

Problem 3: Knock on effects of low level processing errors

- In a working system, high level processes do not have ground truth for inputs, just the output of low level processing
- Estimating distance between ground truth and output of low level processes will allow high level processes to fit approaches to likely errors
- Errors will still exist and must be taken into account in the evaluation metrics for the high level processes

Example: using ASR

- Some processes (e.g., NER, chunking) have been previously applied to text and have existing evaluation techniques

BUT

- NER in multimodal systems is over ASR output, not "ground truth" text (human transcription)
  - can fail to recognize named entity because words are wrong
  - can recognize right named entity but wrong words

Problem 4: Different modalities use different evaluations

- Consider tracking
  - audio, video, or multimodal
- not just a matter of seeing whether combining information from different modalities improves results - communities conceive of as different tasks

Problem 5: No one ground truth

- for some components (e.g., summarizers), there is no one correct output
- can get human judges to look at component output and judge "goodness"
  - expensive, fixed cost per component/version
- can get a bunch of correct human-authored outputs and look at how well component output fits in
  - less expensive because fixed cost for any number of components/versions to test on same base material
  - need good measure of fit to judge goodness, but current measures don't correlate with human judgments - i.e., dangerous

Example: multi-document summarization

- two years ago, community thought objective evaluation would work
- been through a series of measurements that don't match human judgments
- discovered uncertainty even about what the task is
- BUT this is a relatively simple task compared to anything one would do with multi-modal sources like meetings

Problem 6: methodology for extrinsic evaluation

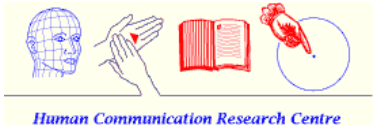
- field testing gives qualitative results, but is expensive and slow
- observational analysis of real users admits some quantification, but can't be sure how tasks compare so can't easily compare different interfaces
- need more control

Controlled system evaluation example

- Build multiple meeting browsers and a baseline system
- Have human observers play meetings in full, write true statements relating to content, complement them with false statements, and rank the set for importance
- Run subjects choosing which of the pair is true and false; give them some proportion of the running time of the meeting and score in terms of how many they answer, penalizing for wrong answers.

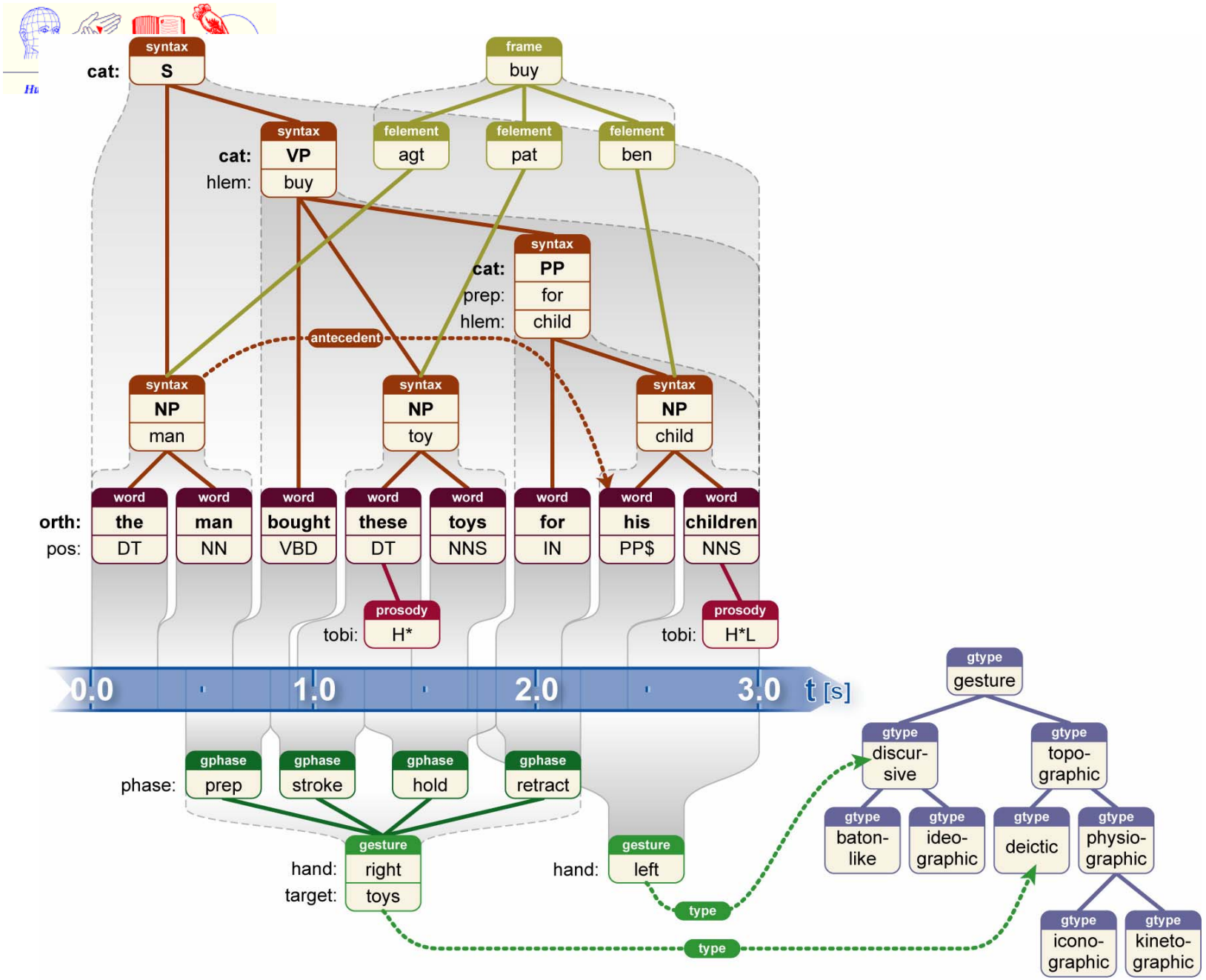
Issues for controlled approach

- Does the task bear any relationship to what users will actually do with the application?
  - Meeting browsing clearly isn't about truth/falsity of given statements. Is it even about question-answering?
  - user requirements for new technologies are hard to gather!
- What do you tell the annotators authoring the true/false statements to get good ones?



Problem 7: Data representation is hard, but important

- For most low level tasks, annotations are just timestamped labels drawn from an enumerated set
- Where the meaning of language is involved, structure is required
- Where input is from multiple modalities, they relate to each other


cat:
syntax
S
frame
buy
cat:
hlem:
syntax
VP
buy
felement
agt
felement
pat
felement
ben
cat:
prep:
hlem:
syntax
PP
for
child
antecedent
syntax
NP
man
syntax
NP
toy
syntax
NP
child
orth:
pos:
word
the
DT
word
man
NN
word
bought
VBD
word
these
DT
word
toys
NNS
word
for
IN
word
his
PP$
word
children
NNS
prosody
tobi:
H*
prosody
tobi:
H*L
0.0
1.0
2.0
3.0
t [s]
phase:
gphase
prep
gphase
stroke
gphase
hold
gphase
retract
hand:
target:
gesture
right
toys
hand:
gesture
left
type
type
gtype
gesture
gtype
discur-
sive
gtype
topo-
graphic
gtype
baton-
like
gtype
ideo-
graphic
gtype
deictic
gtype
physio-
graphic
gtype
icono-
graphic
gtype
kineto-
graphic

Summary

- Component performance ≠ system performance
- We don't know how to devise evaluation metrics even for some simple, intuitive tasks
- Processing errors in low level components affect components that use their output and change how we have to evaluate them.
- Different modalities use different evaluations
- For some components, there is no one ground truth upon which to base an evaluation metric.
- Multimodal inputs and trying to get at meaning make data representation important.

Basic tension for MM interfaces

- To get really good comparability even just for components, need whole community to work on:
  - same task (so performance issues are the same, and it's worth everyone putting same effort into a component)
  - same data set (because a component developed on a different one won't work)
  - same architecture (so individual components have same impact on end performance)
- Great way to stifle innovation, hit local minima for progress, and make it hard to see how to use our results for a wide range of tasks
- Would really help publication rates.

What will aid progress? (1)

- Freely available data, annotations, and evaluation metrics
  - lowers the bar for contributing
  - student groups can do surprisingly well in some community evaluations, and that's important
- Allowing reuse of same data for different tasks, outside community evaluations, so someone with a bright idea can try it out cheaply
- Even the infrastructure for annotating is too fragmented at present

What will aid progress? (2)

- Better understanding of the relationship between component and system performance
  - subassemblies common to several systems at least must have characteristics that can be known
  - could there be a rule of thumb about importance of ASR accuracy to systems of different kinds?
- Better understanding of how to adapt components to different data set and different genres