Targeted Meeting Understanding at CSLI

Matthew Purver
Patrick Ehlen
John Niekrasz
John Dowding
Surabhi Gupta
Stanley Peters
Dan Jurafsky

The CALO Meeting Assistant

- Observe human-human meetings
  - Audio recording & speech recognition
  - Video recording & gesture/face recognition
  - Written and typed notes
  - Paper & whiteboard sketches

- Produce a useful record of the interaction …

A Hard Problem

A Hard Problem

- Human-human speech is hard
  - Informal, ungrammatical conversation
  - Overlapping, fragmented speech
  - High speech recognition error rates (20-30% WER)
- Overhearing is hard
  - Don’t necessarily know the vocabulary
  - … the concepts
  - … the context
- No point trying to understand *everything*
  - Target some useful things that we can understand

Speech Recognition Errors

- But remember: the real input is from ASR:
  - **do you have the comments cetera and uh the the other is**
  - **you don't have**
  - **i do you want**
  - **oh we of the time align said is that**
  - **i you**
  - **well fifty comfortable with the computer**
  - **mmm**
  - **oh yeah that's the yeah that**
  - **sorry like we're set**
  - **make sure we captive that so this deviates**
- Usually better than this, but 20-30% WER

What would be useful?

- Banerjee et al. (2005) survey of 12 academics:
  - *Missed meeting - what do you want to know?*
  - Topics: which were discussed, what was said?
  - Decisions: what decisions were made?
  - Action items/tasks: was I assigned something?
- Lisowska et al. (2004) survey of 28 people:
  - *What would you ask a meeting reporter system?*
  - Similar questions about topics, decisions
  - People: who attended, who asked/decided what?
  - Did they talk about me?

Overview

- Topic Identification
  - Shallow understanding
  - Producing topics and segmentation for browsing, IR
- Action Item Identification
  - Targeted understanding
  - Producing to-do lists for user review
- User interface & feedback
  - Presenting information to users
  - Using user interaction to improve over time

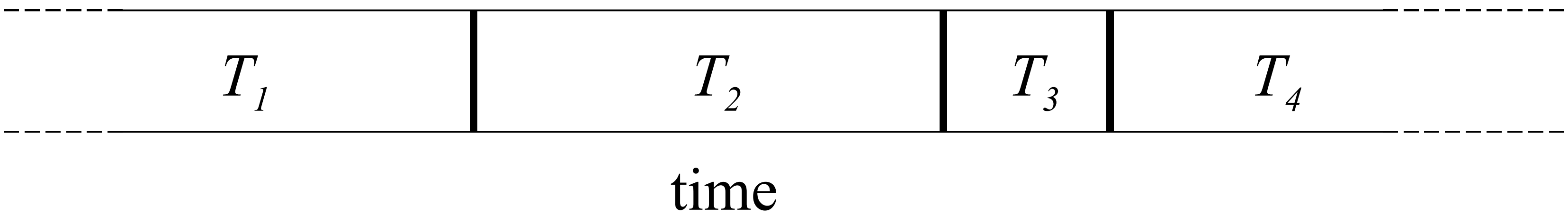
Topic Identification

Topic Identification

- Problem(s):
  - (1) Identify the topics discussed (*identification*)
  - (2) Find them/find a given topic (*segmentation/localization*)
    - Effectively summarize meetings
    - Search/browse for topics
    - Relate meetings to each other
- Neither (1) or (2) are new, but:
  - Not usually done simultaneously
  - Not done over speech recognition output
- Joint work with MIT/Berkeley (Tenenbaum/Griffiths)
  - Unsupervised generative modelling, joint inference

Segmentation vs. Identification

- Segmentation: dividing the discourse into a series of topically coherent segments
- Identification: producing a model of the topics discussed in those segments



  - Both useful/required for browsing, summary
  - Joint problems: try to solve them jointly

Topic Subjectivity

- Both segmentation & identification depend on your conception of topic …
- Given the job of simultaneously segmenting & identifying, humans don't agree:
  - Kappa metric ~0.50 (Gruenstein et al., 2005)
  - Given more constraints (e.g. identify agenda items), they agree much better (Banerjee & Rudnicky, 2007)
  - But people often want different things …
- If we can model the underlying topics, we can allow people to search for the ones they're interested in
- We'd also like to make a "best guess" at unsupervised segmentation, but it'll never be ideal
  - Adapt a state-of-the-art unsupervised algorithm to discourse

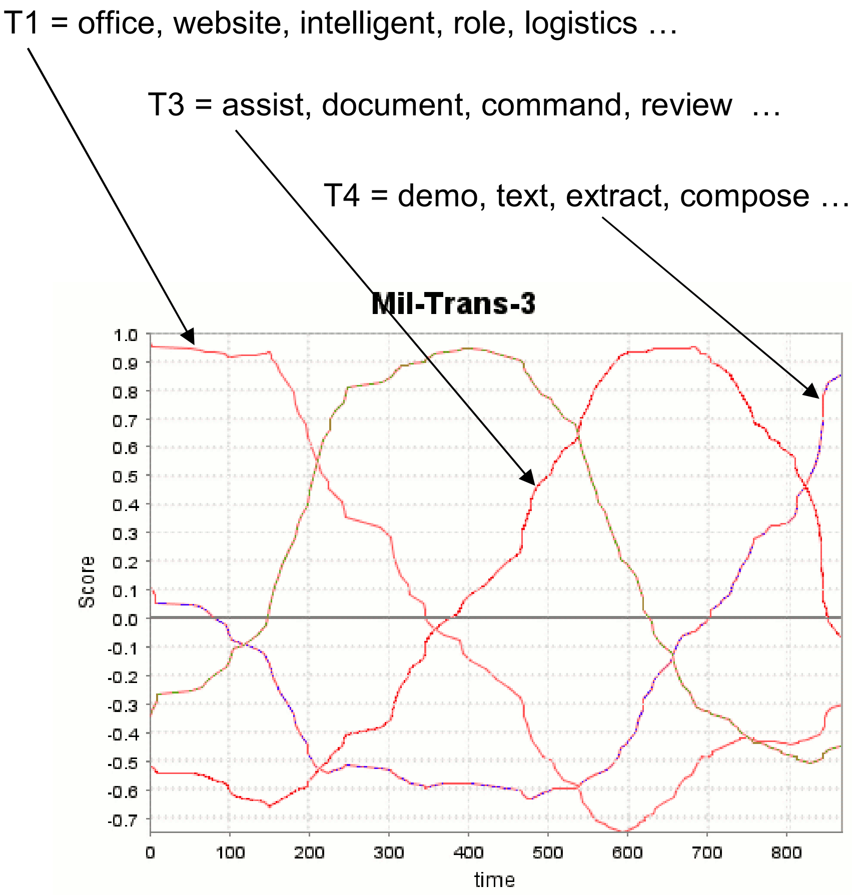
Related Work

- Segmentation for text/monologue (broadcast news, weather reports, etc.)
  - (Beeferman et al., Choi, Hearst, Reynar, …)
- Identification for document clustering
  - (Blei et al., 2003; Griffiths & Steyvers, 2004)
- Joint models for text & monologue (HMMs)
  - (Barzilay & Lee, 2004; Imai et al;, 1997)
- Little precedent with spoken multi/dialogue …
  - Less structured, more "noisy", interruptions, fragments
  - Less restricted domain
  - Worse speech recognition accuracy
- (Galley et al., 2003) lexical cohesion on ICSI meeting corpus (*"LCSeg"*)
  - Segmentation only (no topic identification)
  - Manual transcripts only (not ASR output)

What are we trying to do?

- Get amazing segmentation? Not really.
  - Human-human agreement only 0.23 P_k, 0.29 W_D
- 1. Add topic identification:
  - Segmentation on its own may not be that much help
  - User study results focus on topic identification
  - Would like to present topics, summarize, understand relations between segments
- 2. Investigate performance on noisier data:
  - Off-topic discussion; speech recognition (ASR) output

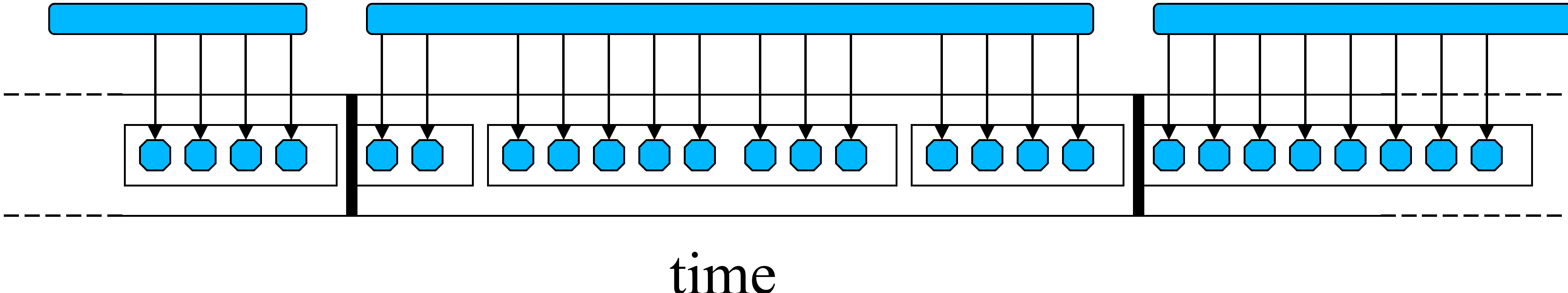
Topic Modelling



- Model topics as probabilistic word vectors
  - Can find most relevant topic for a given time/segment
  - … or likely times/segments for a given topic
  - … or both
- Learn the vectors unsupervised
  - Latent Dirichlet Allocation
    - Assume words generated by mixtures of fixed "micro-topics"
    - Basic assumptions about model distributions
    - Random initialization, statistical sampling
  - Joint inference for topics/segments
  - Extend models over time/data

A Generative Model for Topics

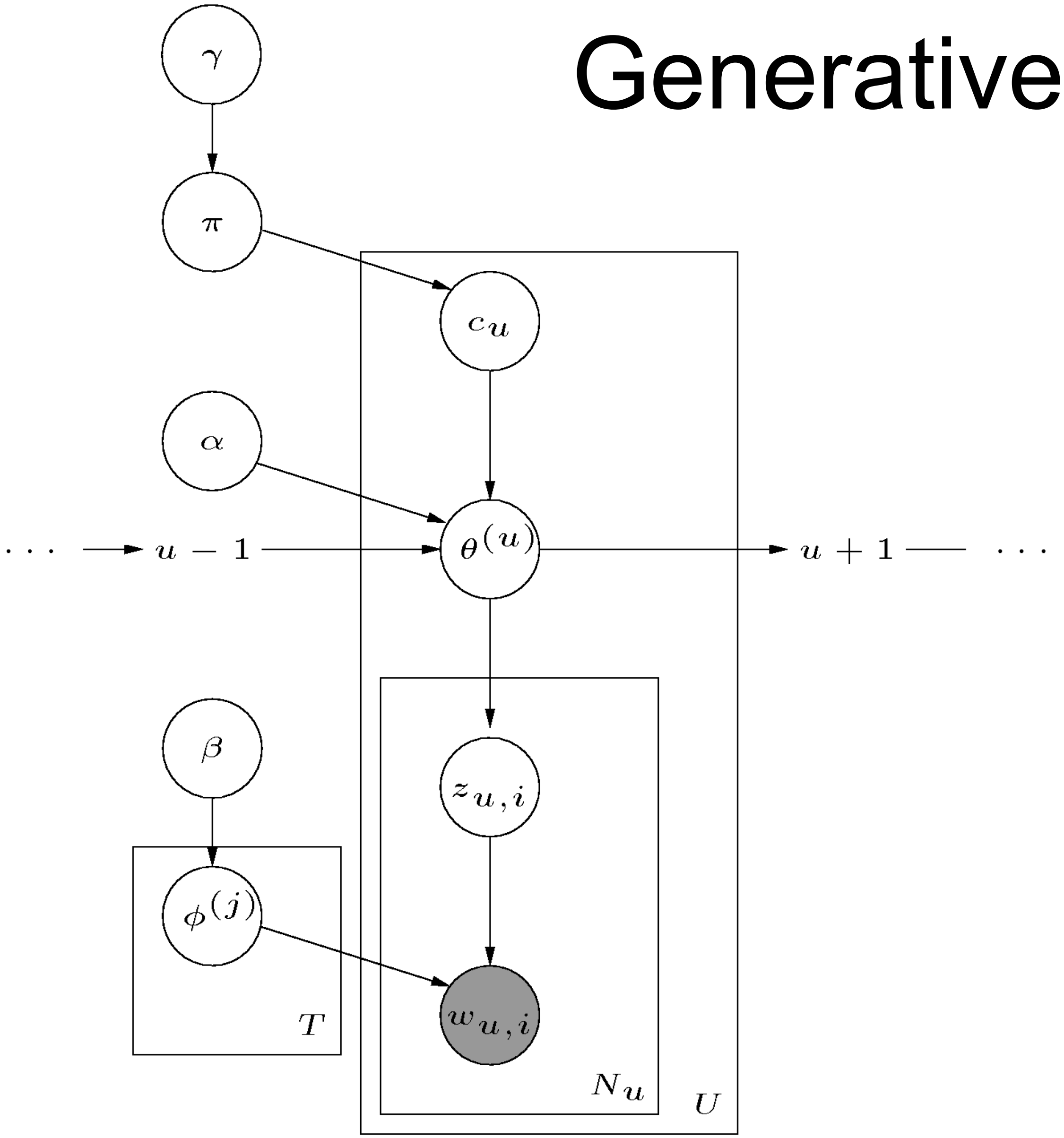
- A discourse as a linear sequence of utterances
  - Utterances as linear sequences of word tokens
- Words as generated by "topics"
- Discourse segments have fixed "topics"
  - Assume utterances have fixed "topics"
  - Assume segments only shift at utterance starts



A Bit More Detail

- Topics: probability distributions over word types
  - A fixed set of these "micro-topics"
- Segments: fixed weighted mixtures of micro-topics
  - An infinite possible set of these "macro-topics"
  - A "topic shift" or "segment boundary" means moving to a new weighted mixture
- We will try to jointly infer micro-topics, macro-topics and segment boundaries …
- Extension of Latent Dirichlet Allocation (Blei et al., 2003)
  - General model for inferring structure from data
  - Used for document clustering, hand movements etc.

Generative Model



- T per micro-topic
- U per utterance
- N_u per word

- θ macro-topic mixture
- $z_{u,i}$ micro-topic assignment
- φ micro-topic
- $w_{u,i}$ observed word
- c_u segment switch

Segmentation accuracy

- Segmentation compares well with previous work:
  - Pk = 0.33 (vs. 0.32 for *LCSeg*) on ICSI meeting corpus
- Improves if number of topics is known (from agenda)
  - Pk = 0.29 (vs. 0.26 for *LCSeg*)
- Robust in the face of ASR inaccuracy
  - Pk = 0.27 to 0.29 (vs. 0.29 to 0.38 for *LCSeg*)
- Robust to data variability
  - Tested on 10-meeting CMU corpus (Banerjee & Rudnicky)
  - Pk = 0.26 to 0.28, robust to ASR output
- But importantly, we are identifying topics too:
  - Word lists fit with known ICSI discussion topics
  - Lists rated as coherent by human judges

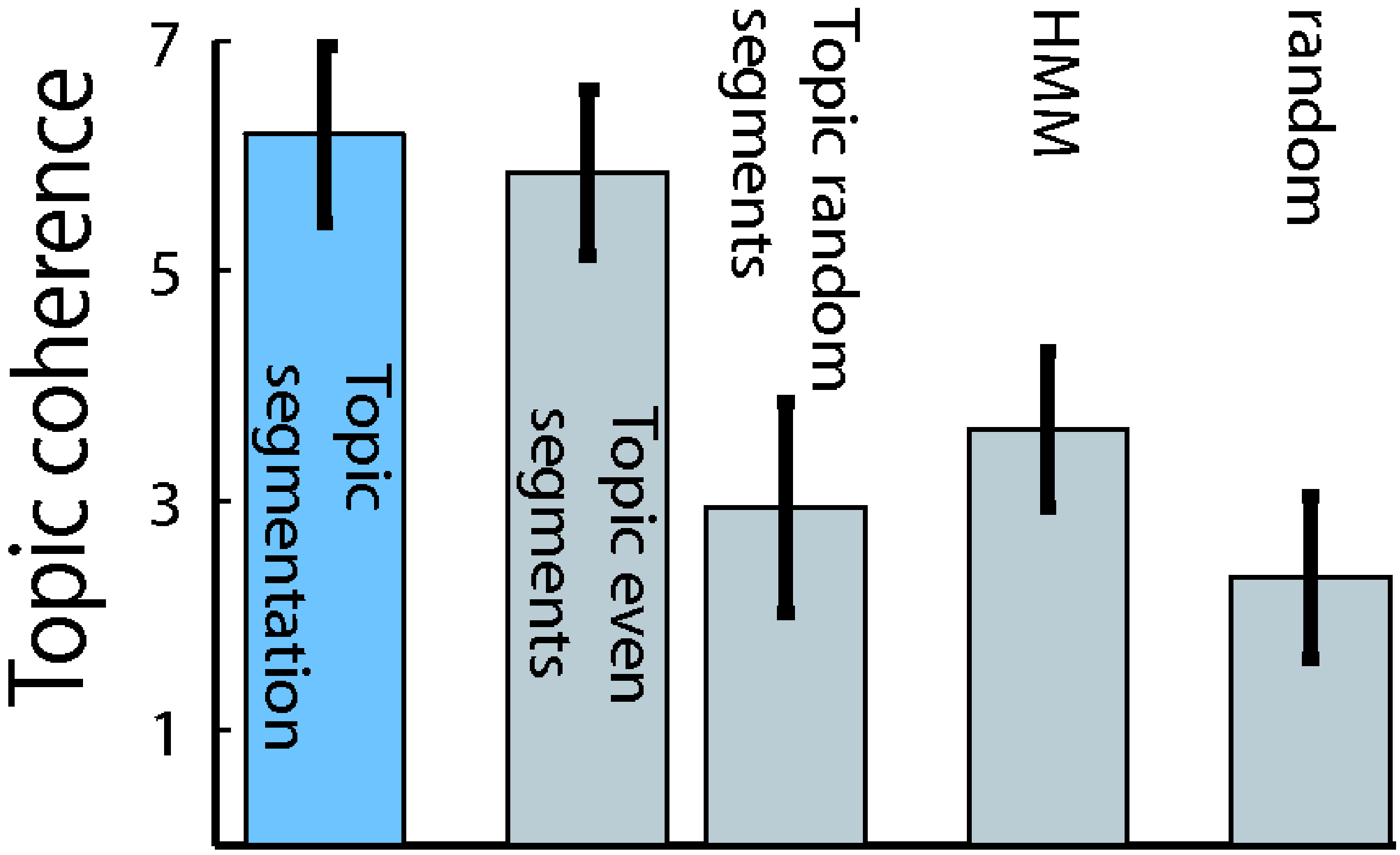
ICSI Topic Identification

Topic

Word	1	2	3	4	5	6	7	8
	technology	models	speakers	wouldn't	v_a_d	mikes	enter	disk
	u_m_t_s	reverberation	overlaps	you'd	worse	microphones	construction	beep
	routing	voicing	alignment	agree	t_i-digits	record	constructions	beeps
	transmission	multi-band	region	matter	baseline	collection	belief-net	gig
	i_p	targets	breath	depends	l_d_a	subjects	object	display
	mobile	phonemes	laugh	open	percent	wizard	ontology	disks
	packet	effects	native	others	italian	notes	schema	linux
	university	echo	backchannels	feeling	improvement	brian	parser	dollars
	concerning	combining	laughing	term	adaptation	u_w	bayes-net	laptop
	networking	insertions	marks	opposed	latency	age	deep	p_c

- Meetings of ICSI research groups
  - Speech recognition, dialogue act tagging, hardware setup, meeting recording
  - General "syntactic" topic

ICSI Topic Ratings



Where to go from here?

- Improvements in topic model robustness
  - Interaction with multiple ASR hypotheses
- Improvements in segmentation quality
  - Interaction with discourse structure
- Relating topics to other sources
  - Relation between meetings and documents/emails
- Learning user preferences

Action Item Identification

Action Item Identification

- Problem(s):
  - (1) Detect action item discussions
  - (2) Extract salient “to-do” properties
    - Task description
    - Responsible party
    - Deadline
- (1) is difficult enough!
  - **Never done before on human-human dialogue**
  - **Never done before on speech recognition output**
- New approach: use (2) to help (1)
  - Discussion of action items has characteristic patterns
    - Partly due to (semi-independent) discussion of each salient property
    - Partly due to nature of decisions as group actions
  - Improve accuracy while getting useful information

Action Item Detection in Email

- Corston-Oliver et al., 2004
  - Marked a corpus of email with "dialogue acts"
  - *Task* act: "items appropriate to add to an ongoing to-do list"
- Bennett & Carbonell, 2005
  - Explicitly detecting "action items"
- Good inter-annotator agreement ($\kappa > 0.8$)
- Per-sentence classification using SVMs
  - lexical features e.g. n-grams; punctuation; syntactic parse features; named entities; email-specific features (e.g. headers)
  - f-scores around 0.6 for sentences
  - f-scores around 0.8 for messages

Can we apply this to dialogue?

- 65 meetings annotated from:
  - ICSI Meeting Corpus (Janin et al., 2003)
  - ISL Meeting Corpus (Burger et al., 2002)
  - Reported at SIGdial (Gruenstein et al, 2005)
- Two human annotators
- “Mark utterances relating to action items”
  - create groups of utterances for each AI
  - made no distinction between utterance type/role
  - Annotators identified 921 / 1267 (respectively) action item-related utterances
- Try binary classification
  - Different classifier types (SVMs, maxent)
  - Different features available (no email features; prosody, time)

Problems with Flat Annotation

- Human agreement poor ($\kappa < 0.4$)
- Classification accuracy poor (Morgan et al., SIGdial 2006)
  - Try a restricted set of the data where the agreement was best
  - F-scores 0.32
  - Interesting findings on useful features: lexical, prosodic, fine-grained dialogue acts)
- Try a small set of easy data?
  - Sequence of 5 (related) CALO meetings
  - Simulated with given scenarios, very little interruption, repair, disagreement
  - Improved f-scores (0.30 - 0.38), but still poor
- This was all on gold-standard manual transcripts
  - ASR inaccuracy will make all this worse, of course

What’s going on?

- Discussion tends to be split/shared across utterances & people
  - Contrast to email, where sentences are complete, tasks described in single sentences
- Difficult for humans to decide which utterances are “relevant”
  - Kappa metric 0.36 on ICSI corpus (Gruenstein et al., 2005)
  - Doesn’t make for very consistent training/test data
- Utterances form a very heterogeneous set
- Automatic classification performance is correspondingly poor

Should we be surprised?

- DAMSL schema has dialogue acts **Commit, Action-directive**
  - annotator agreement poor (κ ~ 0.15)
  - (Core & Allen, 1997)
- ICSI MRDA dialogue act **commit**
  - Automatic tagging arruracy poor
  - Most DA tagging work concentrates on 5 broad DA classes
- Perhaps “action items” comprise a more heterogeneous set of utterances

A Dialogue Example

SAQ	not really. **the there was the uh notion of the preliminary patent, that uh**
FDH	yeah, it is a cheap patent.
SAQ	yeah.
CYA	okay.
SAQ	which is
FDH	so, it is only seventy five dollars.
SAQ	and it is it is e an e
CYA	hm, that is good.
HHI	talk to
SAQ	yeah and and it is really broad, you don't really have to define it as w as much as in in a you know, a uh
FDH	yeah.
HHI	**I actually think we should apply for that right away.**
CYA	**yeah, I think that is a good idea.**
HHI	**I think you should, I mean, like, this week, s start moving in that direction**. just 'cause that is actually good to say, when you present your product to the it gives you some instant credibility.
SAQ	[Noise]
SAQ	**mhm**.
CYA	**right.**

Rethinking Action Item Acts

- Maybe action items are not aptly described as singular "dialogue acts"
- Rather: multiple people making multiple contributions of several types
- Action item-related utterances represent a form of group action, or *social action*
- That social action has several components, giving rise to a heterogeneous set of utterances
- What are those components?

Action Item
Dialogue Moves

- Four types of dialogue moves:

Action Item Dialogue Moves

- Four types of dialogue moves:
  - Description of task

Somebody needs to fill out this report!

Action Item Dialogue Moves

- Four types of dialogue moves:
  - Description of task
  - Owner

Action Item Dialogue Moves

- Four types of dialogue moves:
  - Description of task
  - Owner
  - Timeframe

Can you do it by tomorrow?

Action Item Dialogue Moves

- Four types of dialogue m
  - Description of task
  - Owner
  - Timeframe
  - Agreement

Action Item Dialogue Moves

- Four types of dialogue m
  - Description of task
  - Owner
  - Timeframe
  - Agreement

A Dialogue Example

SAQ	not really. **the there was the uh notion of the preliminary patent, that uh**
FDH	yeah, it is a cheap patent.
SAQ	yeah.
CYA	okay.
SAQ	which is
FDH	so, it is only seventy five dollars.
SAQ	and it is it is e an e
CYA	hm, that is good.
HHI	talk to
SAQ	yeah and and it is really broad, you don't really have to define it as w as much as in in a you know, a uh
FDH	yeah.
HHI	**I actually think we should apply for that right away.**
CYA	**yeah, I think that is a good idea.**
HHI	**I think you should, I mean, like, this week, s start moving in that direction**. just 'cause that is actually good to say, when you present your product to the it gives you some instant credibility.
SAQ	[Noise]
SAQ	**mhm.**
CYA	**right.**

Define task

Define timeframe

Assign owner

Agree

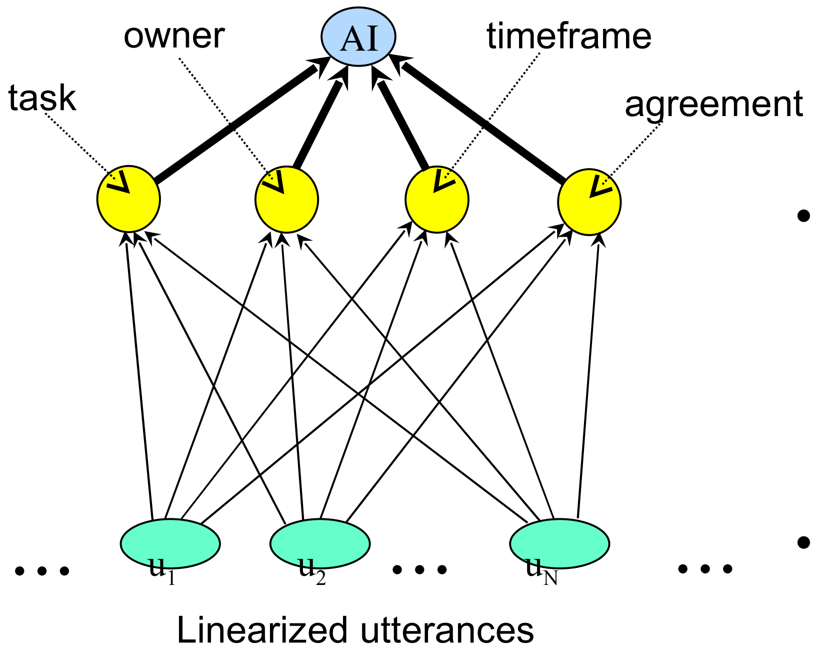
Exploiting discourse structure

- Action item utterances can play different roles
  - Proposing, discussing the action item properties
    - (semantically distinct properties: task, timeframe)
  - Assigning ownership, agreeing/committing
- These subclasses may be more homogeneous & distinct than looking for just "action item" utts.
  - Could improve classification performance
- The subclasses may be more-or-less independent
  - Combining information could improve overall accuracy
- Different roles associated with different properties
  - Could help us extract summaries of action items

New annotation schema

- Annotate utterances according to their role in the action item discourse
  - can play more than one role simultaneously
- Improved inter-annotator agreement
  - Timeframe: $\kappa = 0.86$
  - Owner 0.77, agreement & description 0.73
- Between-class distinction (cosine distances)
  - Agreement vs. any other is good: 0.05 to 0.12
  - Timeframe vs. description is OK: 0.25
  - Owner/timeframe/description: 0.36 to 0.47

Structured Classifier



- Individual "dialogue act" classifiers
  - Support vector machines
  - Lexical (n-gram) features
  - Investigating prosody, dialogue act tags, syntactic & semantic parse features
- Sub-dialogue "super-classifier"
  - Features are the sub-classifier outputs over a window of N utterances
  - Classes & confidence scores
  - Currently SVM, N=10 (but under investigation)
- Performance for each "act" type compares to previous overall performance
  - ICSI data: f-scores 0.1-0.3
  - CALO data: f-scores 0.3-0.5
  - (with a basic set of features)

Subdialogue Detection Results

- Evaluation at the utterance level isn't quite what we want
  - Are agreement utterances important? Ownership?
  - Look at overall discussion f-scores, requiring overlap by 50%
- 20 ICSI meetings, 10% cross-validation
  - Recall 0.64, precision 0.44, f-score 0.52
  - With simple unigram features only
  - Predict significant improvement …
- CALO project unseen test data f-scores 0 – 0.6
  - ASR output rather than manual transcripts
  - Little related training data, though …

Does it really help?

- Don't have much overlapping data
  - Structured annotation is slow, costly
  - Set of utterances isn't necessarily the same
  - Hard to compare directly with (Morgan et al.) results
- Can compare directly with a flat binary classifier
  - Set of ICSI meetings, simple unigram features
- Subdialogue level:
  - Structured approach f-score 0.52 vs. flat approach 0.16
- Utterance level:
  - Flat approach f-scores 0.05-0.20
  - Structured approach f-scores 0.12-0.31
  - (Morgan et al. f-scores 0.14 with these features)
- Can also look at sub-classifier correction: f-score improvements ~0.05

Extracting Summaries

- Structured classifier gives us the relevant utterances
  - Hypothesizes which utterances contain which information
- Extract the useful entities/phrases for descriptive text
  - Task description: event-containing fragments
  - Timeframe: temporal NP fragments
- Semantic fragment parsing (*Gemini* – joint work with John Dowding (UCSC))
  - Small grammar, large vocabulary built from *Net
  - Extract many potential phrases of particular semantic types
  - Use word confusion networks to allow n-best word hyps
- Experimenting with regression models for selection
  - Useful features seem to be acoustic probability and semantic class

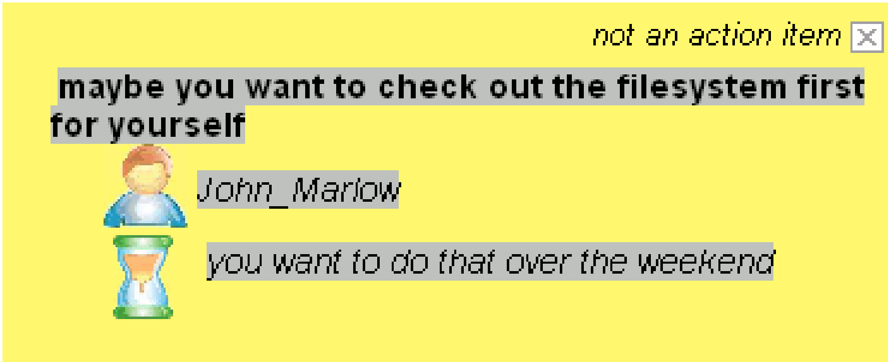
Extracting Ownership

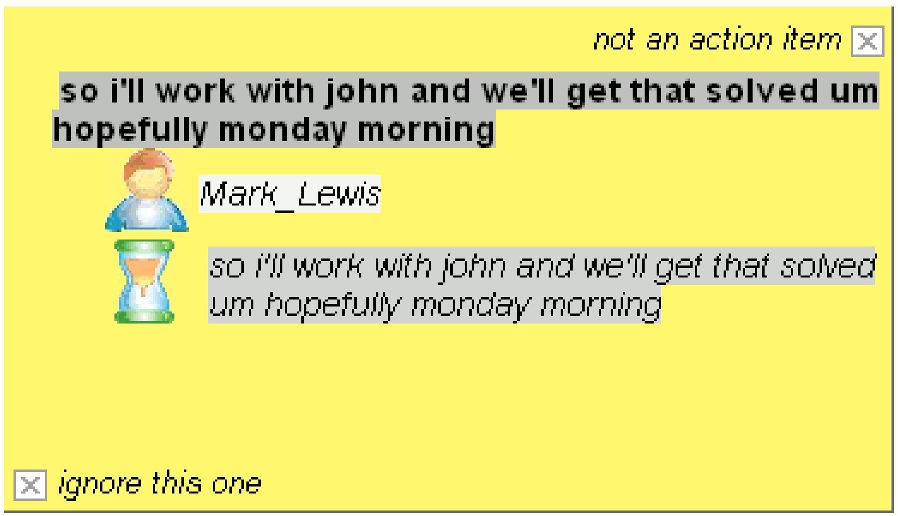
- Sometimes people use names, but only < 5% of cases
- Much more common to volunteer yourself (*"I'll do X …"*) or suggest someone else (*"Maybe you could …"*)
- Self-assignments: speaker
  - Individual microphones, login names (otherwise, it's a speaker ID problem)
- Other-assignments: addressee
  - Addressee ID is hard, but approachable (Katzenmaier et al., 2004; Jovanovic et al., 2006 about 80% accuracy)
  - Also investigating a discourse-only approach
- Need to distinguishing between the two, though
  - Presence of "I" vs. "you" gets us a lot of the way
  - Need to know when "you" refers to the addressee

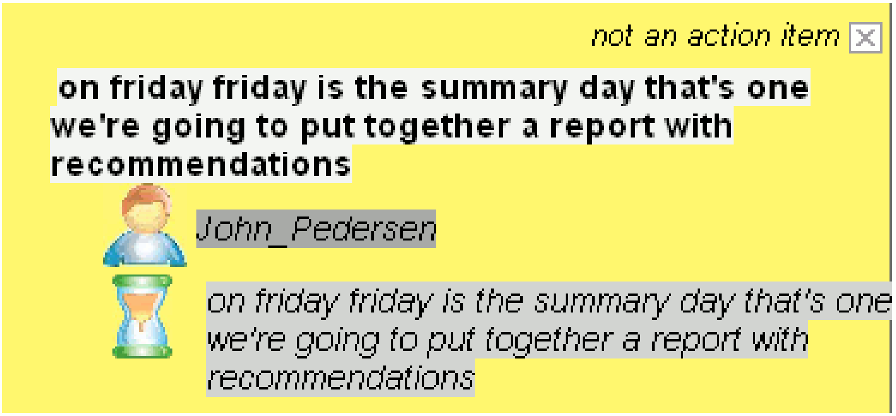
Addressee-referring "you"

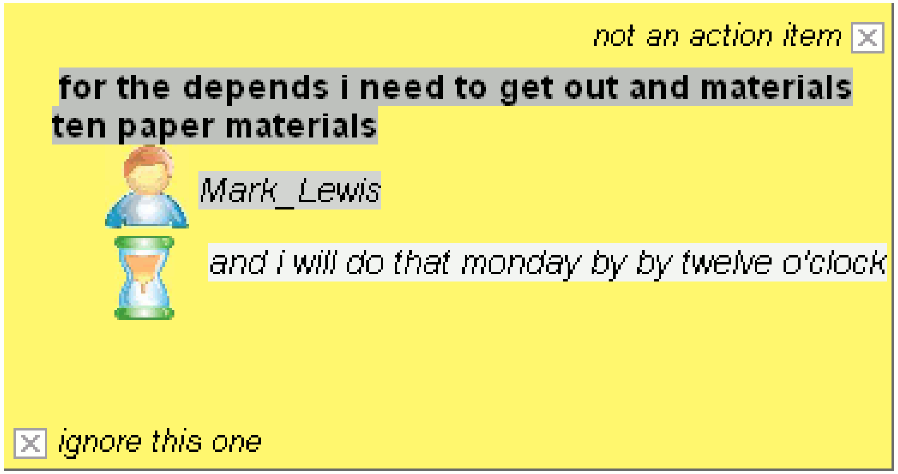
- An interesting sub-problem of ownership detection
- Some *"you"*s refer to the addressee
  - *"Could you maybe send me an email"*
- Some are generic
  - *"When you send an email they ignore it"*
- Investigation in two- and multi-party dialogue
  - Only about 50% of "you" uses are addressee-referring
  - Can detect them with about 85% f-score using lexical & contextual features
    - Some dialogue acts are very useful (question vs. statement)
    - Some classes of verb are very useful (communication)
  - ACL poster (Gupta et al., 2007)

Some Good Examples

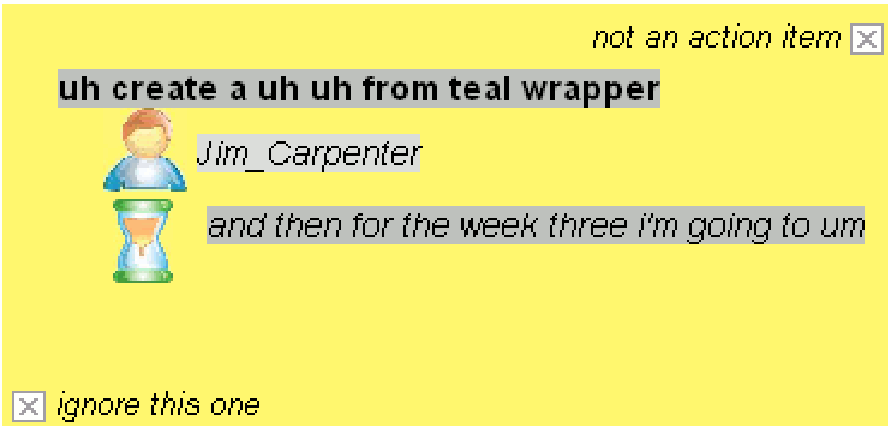


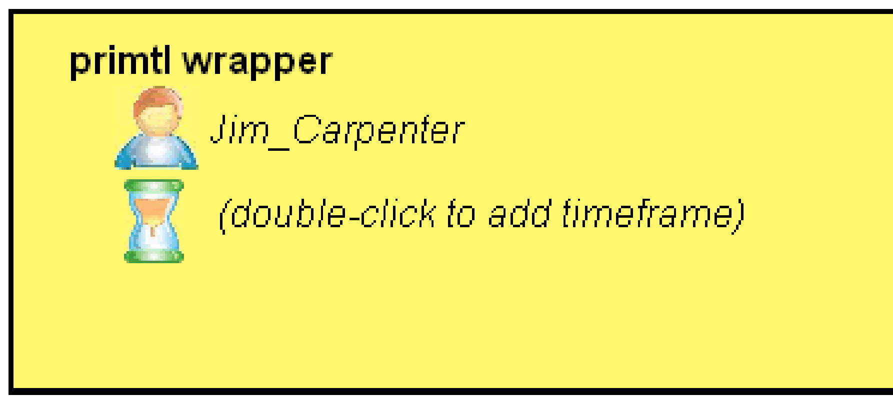




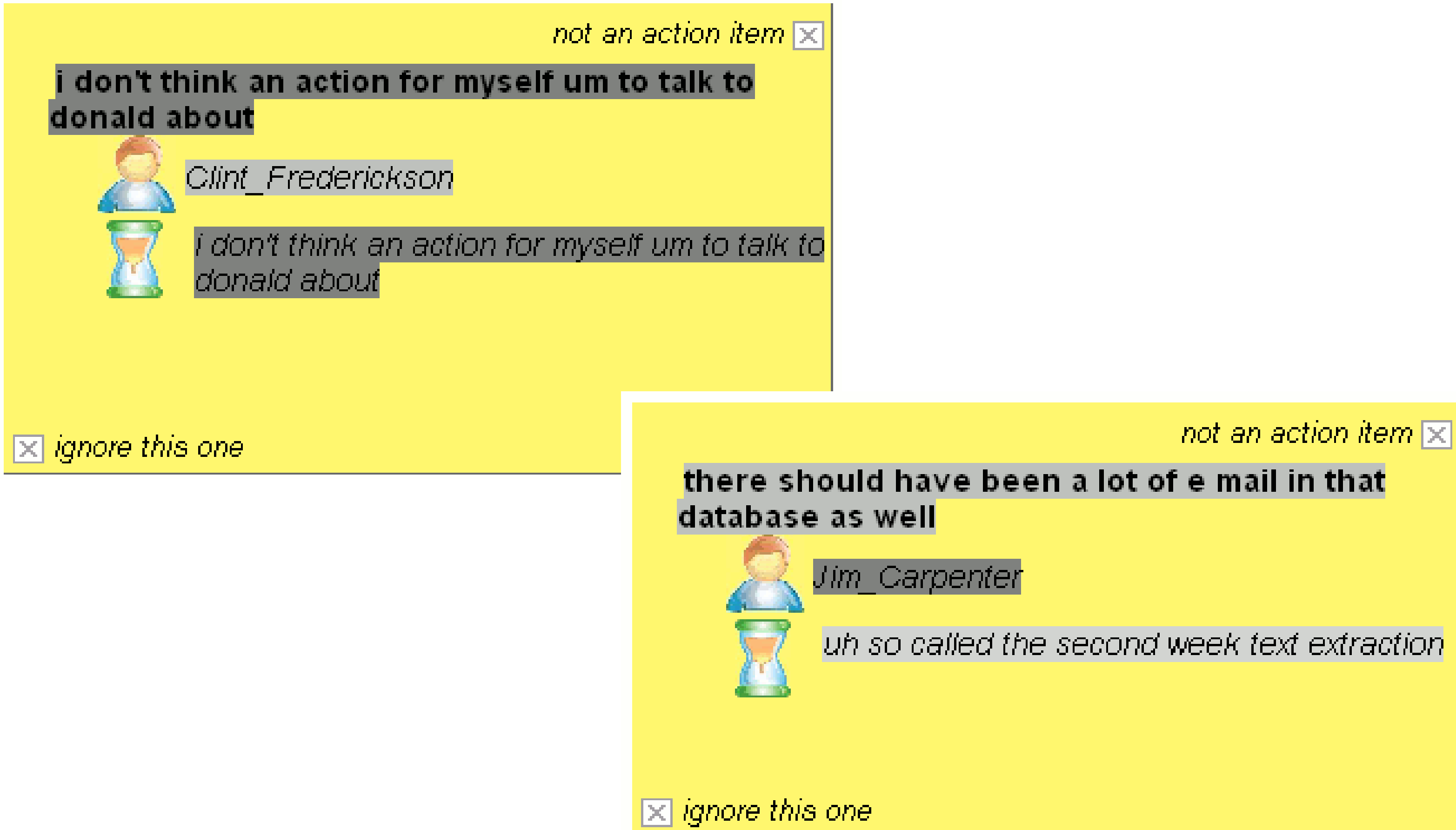


A Great Example





Some Bad Examples



Where to go from here?

- Further semantic property extraction
- Tracking action items between meetings
  - Modification vs. proposal
- Extension to other characteristic discourse “patterns”
  - (including general decision-making)
- Learning for improved accuracy
- Learning user preferences

Feedback & Learning

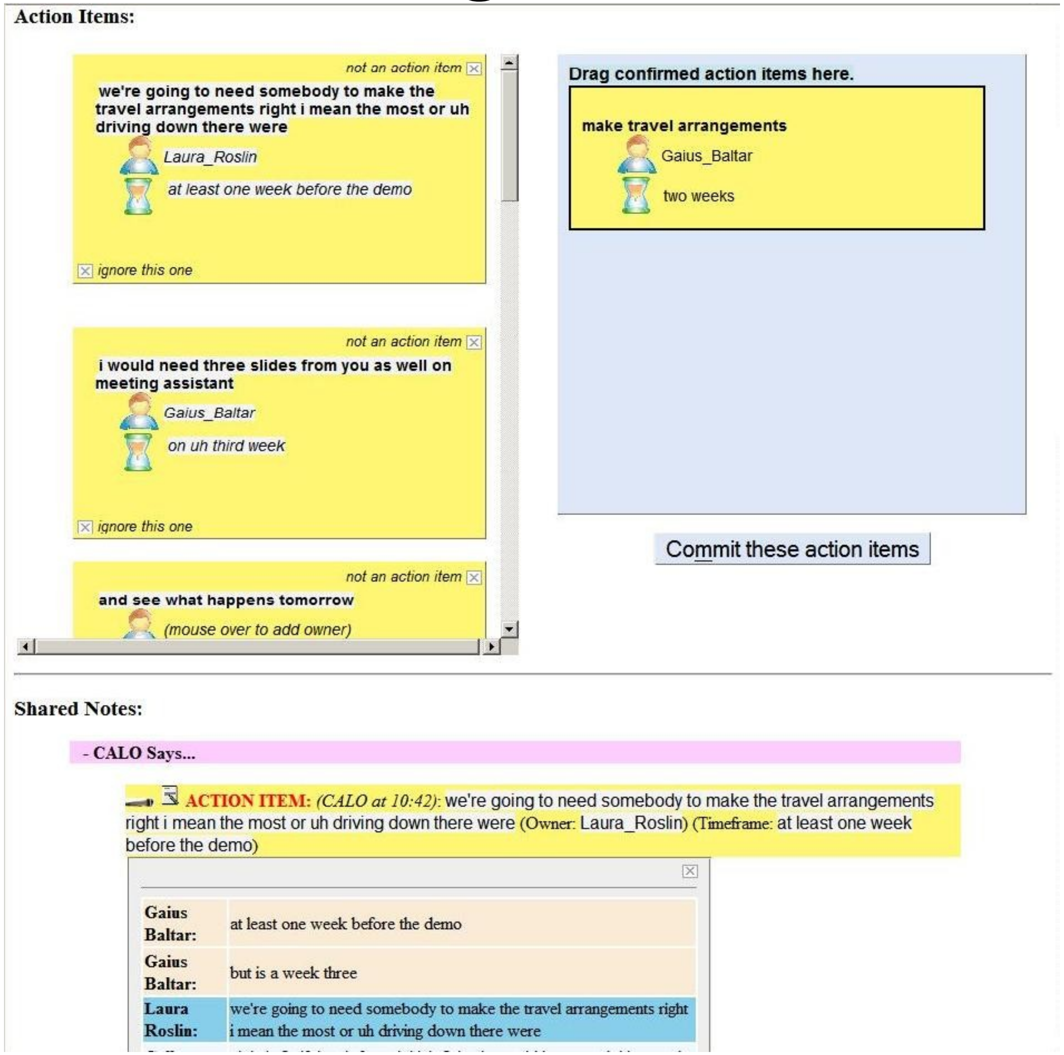
Two Challenges:

- A machine learning challenge:
  - Supervised approach, with costly annotation
  - Want classifiers to improve over time
  - How can we generate training data cheaply?
- A user interface challenge:
  - How do we present users with data of dubious accuracy?
  - How do we make it useful to them?
- Users should see our meeting data results while doing something that's valuable to *them*
- And, from those user actions, give us feedback we can use as *implicit supervision*

Feedback Interface Solution

- Need a system to obtain feedback from users that is:
  - light-weight and usable
  - valuable to users (so they will use it!)
  - can obtain different types of feedback in a non-intrusive, almost invisible way
- Developed a meeting browser
  - based on SmartNotes, a shared note-taking tool already integral to the CALO MA system (Banerjee & CMU team)
- While many “meeting browser” tools are developed for research, ours:
  - has end user in mind
  - is designed to gather feedback to retrain our models
  - two types of feedback: top-level and property-level

Meeting Browser



Action Items

not an action item ☒

maybe you want to check out the filesystem first for yourself

John_Marlow

you want to do that over the weekend

not an action item ☒

on friday friday is the summary day that's one we're going to put together a report with recommendations

John_Pedersen

on friday friday is the summary day that's one we're going to put together a report with recommendations

not an action item ☒

so i'll work with john and we'll get that solved um hopefully monday morning

Mark_Lewis

so i'll work with john and we'll get that solved um hopefully monday morning

☒ *ignore this one*

not an action item ☒

for the depends i need to get out and materials ten paper materials

Mark_Lewis

and i will do that monday by by twelve o'clock

☒ *ignore this one*

Action Items

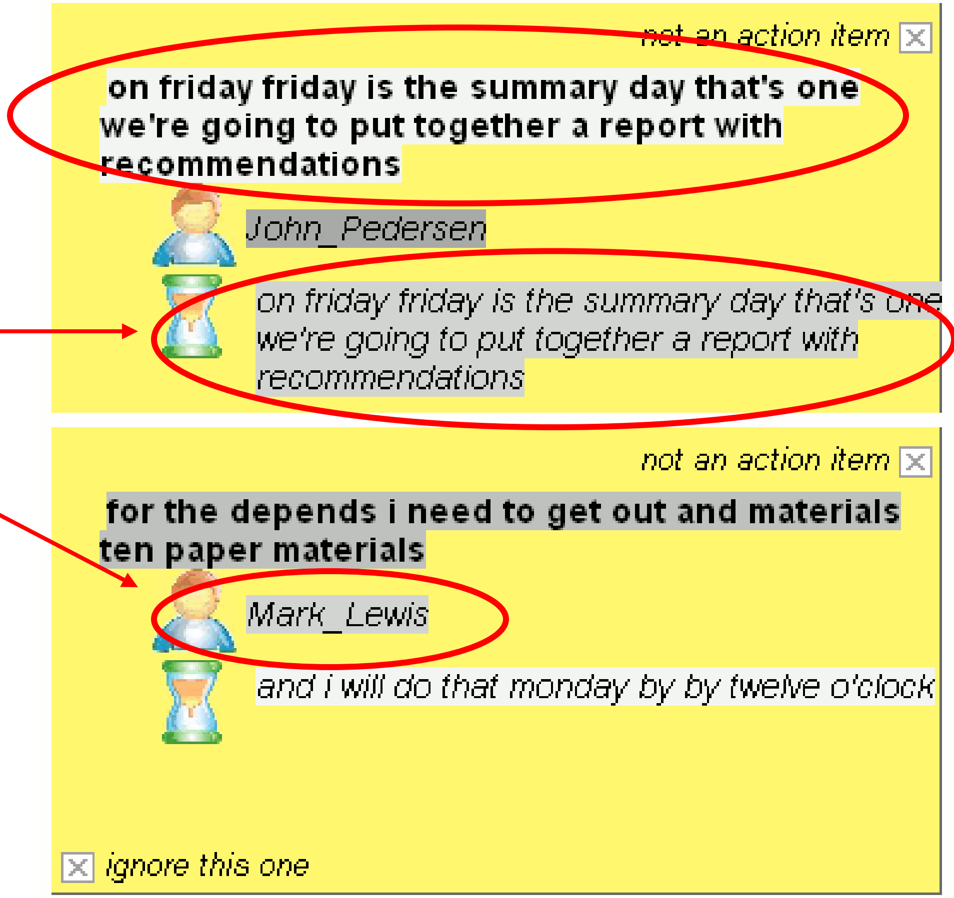
Subclass hypotheses

Top hyp is highlighted

Mouse-over hyps to change them

Click to edit them

(confirm, reject, replace, create)



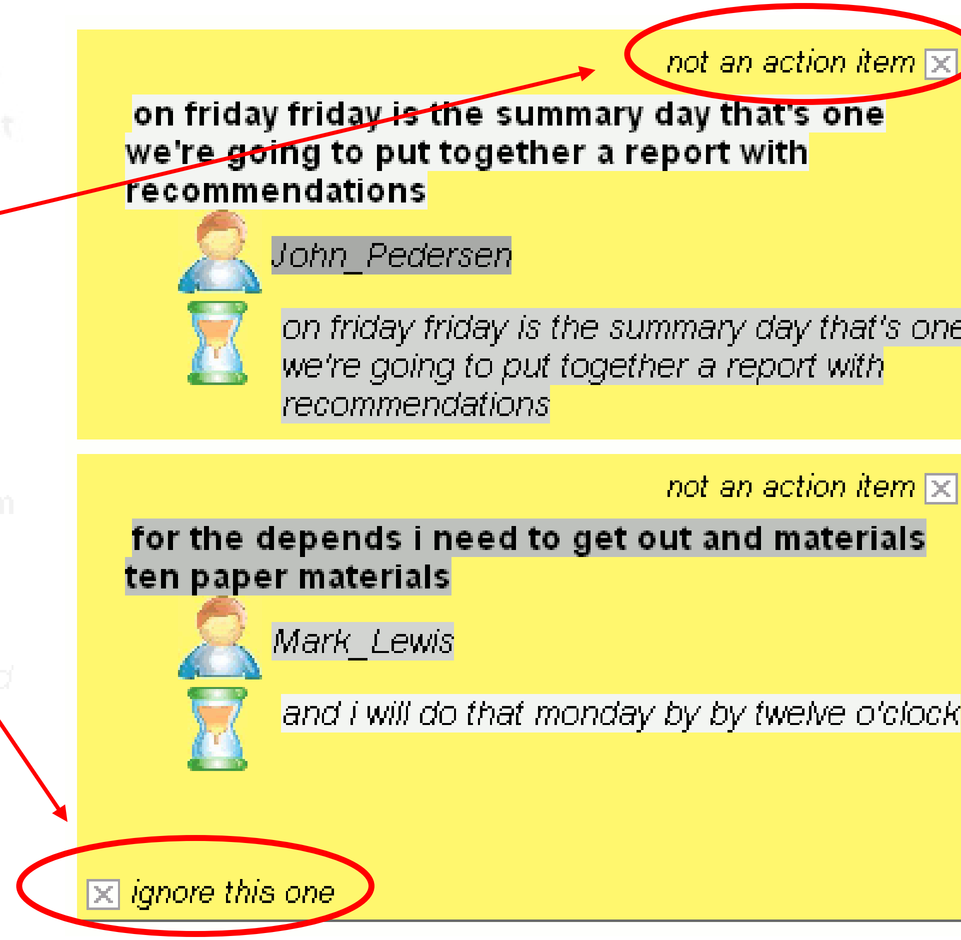
Action Items

Superclass hypothesis

delete = neg. feedback

commit = pos. feedback

merge, ignore

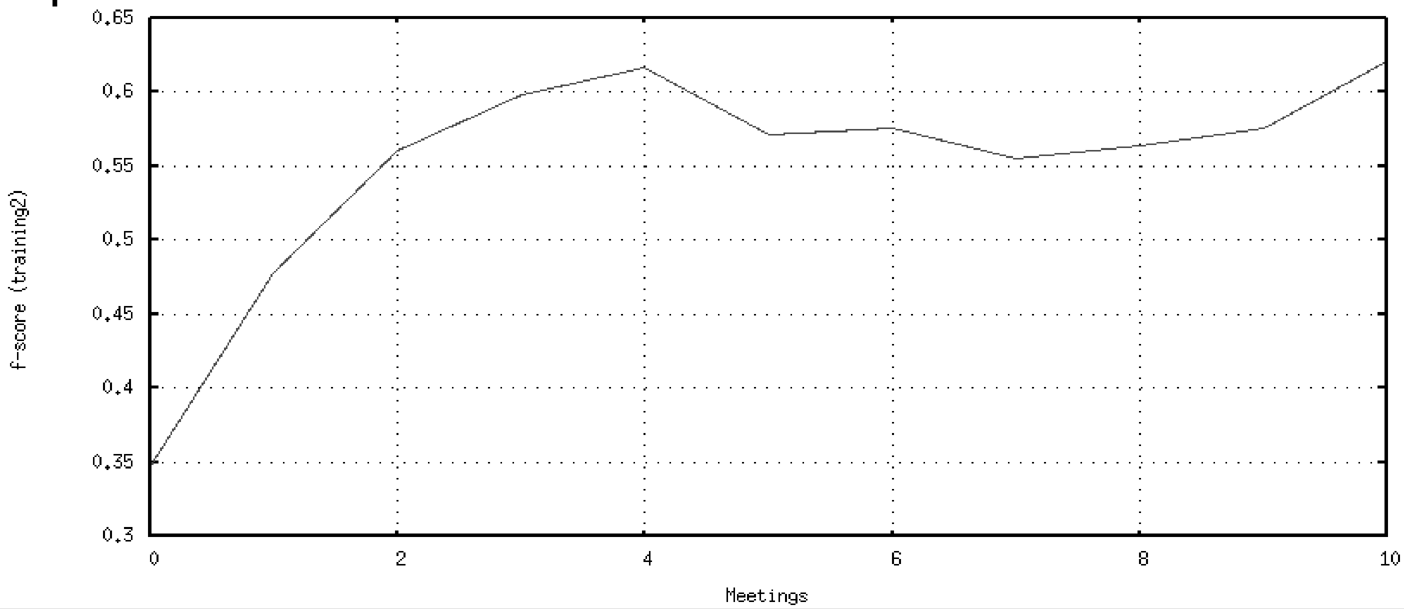


Feedback Loop

- Each participant's implicit feedback for a meeting is stored as an "overlay" to the original meeting data
  - Overlay is reapplied when participant views meeting data again
  - Same implicit feedback also retrains models
  - Creates a personalized representation of meeting for each participant, and personalized classification models

Implicitly Supervised Learning

- Feedback from meeting browser converted to new training data instances
  - Deletion/confirmation = negative/positive instances
  - Addition/editing = new positive instances
  - Applies to overall action items and sub-properties
- Improvement with "ideal" feedback:



What kind of feedback?

- Many different possible kinds of user feedback
- One dimension: time vs. text
  - Information about the *time* an event (like discussion of an action item) happened
  - Information about the *text* that describes aspects of the event (task description, owner, and timeframe)
- Another dimension: user vs. system initiative
  - Information provided when the *user* decides to give it
  - Information provided when the *system* decides to ask for it
- Which kind of information is more useful?
  - Will depend on dialogue act type, ASR accuracy
- Which kind of information is less annoying?
  - During vs. after meeting, Clippy factor

Experiments

- To evaluate user factors, we need to experiment directly
  - Wizard-of-Oz experiment about to start
- To evaluate theoretical effectiveness, can use idealized data
  - Turn gold-standard human annotations of meeting data into posited "ideal" human feedback
- For *text* feedback, use annotators' chosen descriptions
  - Use string/semantic similarity to find candidate utterances
- For *time* feedback, assume 30-second window
  - Use existing sub-classifiers to predict most likely candidates
- For *system* initiative, use existing classifiers to elicit corrections
- Determine which dimensions (time, text, initiative) contribute most to improving classifiers

Ideal Feedback Experiment

- Compare inferred annotations directly
  - Well below human agreement: average 0.6 for best interface
  - Some dialogue act classes do better: owner/task > 0.7
- Compare effects on classifier accuracy
  - F-score improvements very close to ideal data
- Results:
  - both *time* and *text* dimensions alone improve accuracy over raw classifier
  - using both time and text together performs best
  - textual information is more useful than temporal
  - user initiative provides extra information not gained by system-initiative

Wizard-of-Oz Experiment

- Create different Meeting Assistant interfaces and feedback devices (including our Meeting Rapporteur)
- See how real-world feedback data compares to the ideal feedback described above
- Assess how the tools affect and change behavior during meetings

WOZ Experiment Rationale

- Eventual goal: A system that recognizes and extracts important information from many different types of multi-party interactions, but doesn’t require saving entire transcript
  - Meetings may contain sensitive information
  - People’s behaviors will change when they know a complete record is kept of things they say
  - May often be better to extract certain types of information and discard the rest
- To deploy an actual system, also need to know how people will actually use it
  - Especially for a system that relies on language, people’s speech behavior changes in the presence of different technologies

WOZ Experiment Goals

- Provide a corpus of multi-party, task-oriented speech from speakers using different meeting-assistant technologies (does not currently exist)
- Allow us to analyze how verbal and written conceptions of tasks evolve as they progress in time and across different media (speech, e-mail, IM)
- Assess different ways of obtaining user feedback

WOZ Experiment

- Conduct a “Wizard-of-Oz” experiment designed to test how people interact in groups given different kinds of meeting assistant interfaces
  - private, post-meeting interface (individuals interact with it after the meeting, like our current system)
  - private online interface (individuals interact with it during meeting)
  - shared online interface (group interacts with it during meeting)