

Understanding Multi-Party Interaction: Who Decided What For Who?

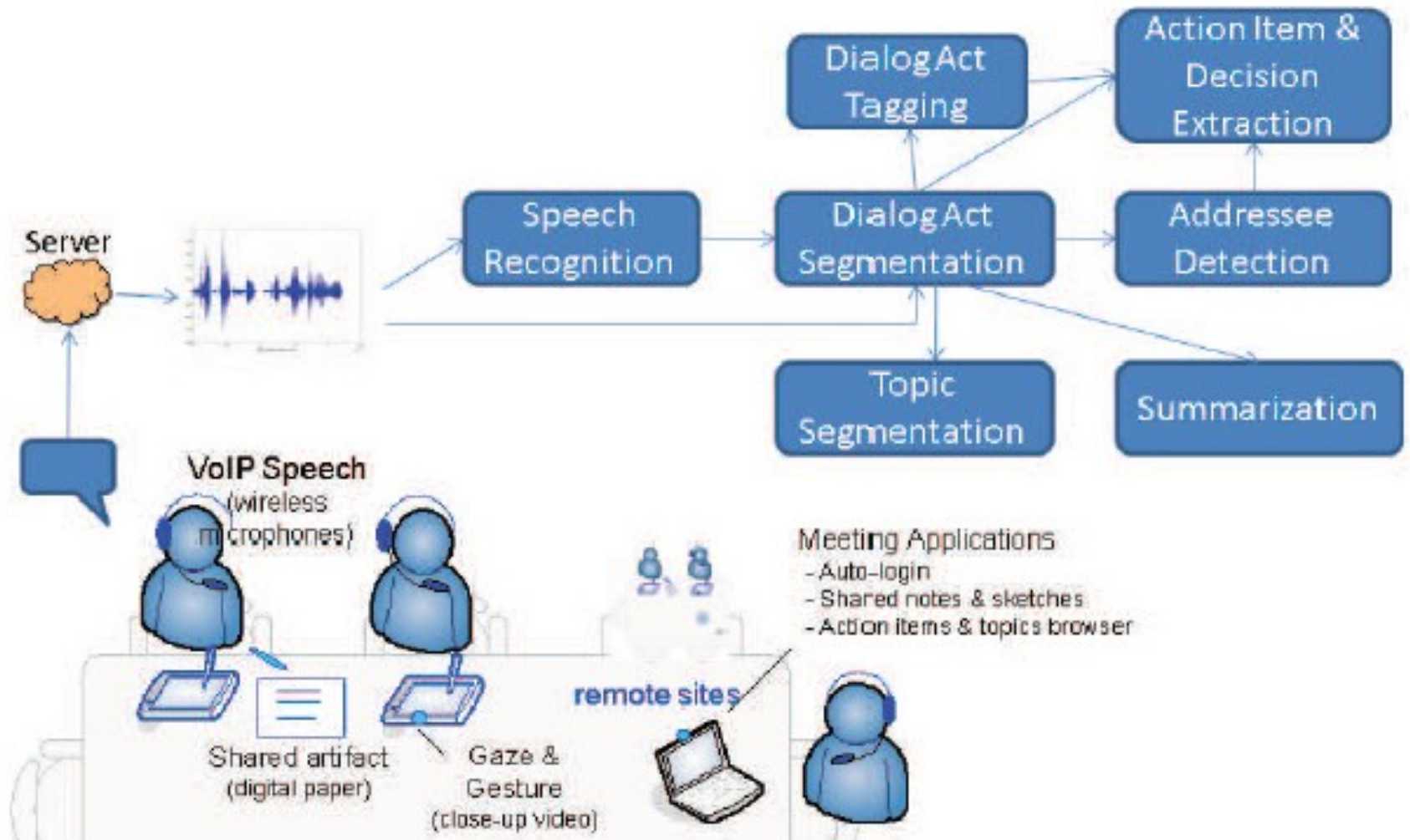
Matthew Purver

Stanley Peters, Matthew Frampton, Raquel Fernandez,
Trung Bui, John Niekrasz, John Dowding,
Patrick Ehlen, Surabhi Gupta, 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 ...

The CALO Meeting Assistant



A Hard Problem

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

Speech Recognition Errors

- 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

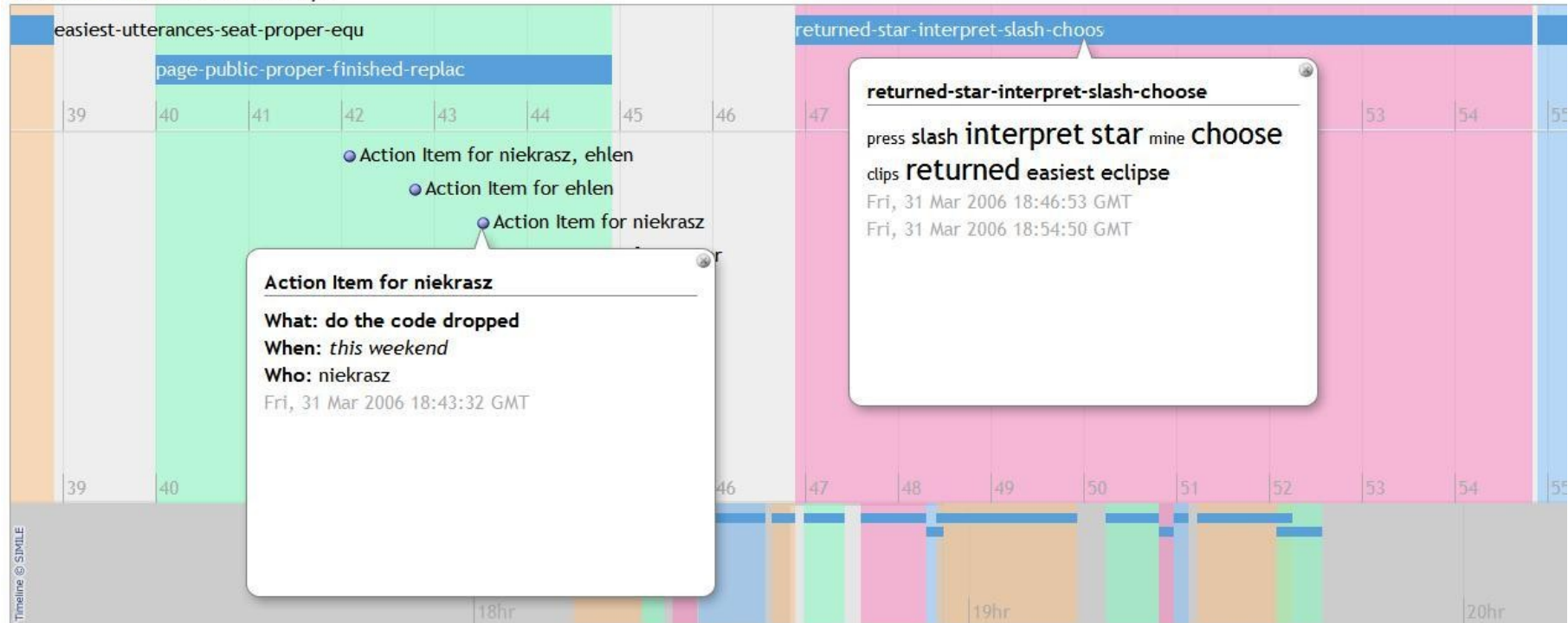
What would be useful?

- 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?

Meeting Browser

CSLI-N01-31-03-2006 Meeting Timeline

Click on timeline to view transcript.



niekrasz: because i'll probably do the code dropped like this weekend like
and i'd like to send a short couple paragraphs tilman about where stuff is
how to how to get started um but it's not a big deal if there are some holes it's totally not
going to matter

ehlen: okay

niekrasz: it would be nice yeah

ehlen: this but on so for the browser java doctoring we really just need

Meeting Browser

Summary	Transcript	Action Items	Topics	QA Pairs	Ink	Meeting Notes	Mark Meeting
What To Do	When To Do It	Who Should Do It	My Actions				
organizing committee	the first week of november	Melinda_ [redacted]					
we should probably get a report from her so	but the plan is that	Pauline_E [redacted]					
an action item for me would be to try to find and i'm not three to wave you know and	come back to you tomorrow	Thierry_D [redacted]					
set up a meeting	next week		<div style="border: 1px solid gray; padding: 5px;"> <p>next week</p> <p>will try</p> <p>oh yeah so kelly was very excited i did and on factor and so we'll try and set up something for next week as an action item</p> <p>you get as an action item to schedule a meeting for sometime next week</p> <p>transition</p> </div>				
to look							
the hero	be interesting	Tim_ [redacted]					

Overview

- Topic Identification
 - Shallow understanding
 - Interactional features can help
- Action Item & Decision Identification
 - Targeted understanding
 - Exploiting interaction structure is crucial
- “You” resolution
 - Specific reference understanding
 - Context of interaction (including vision) is crucial

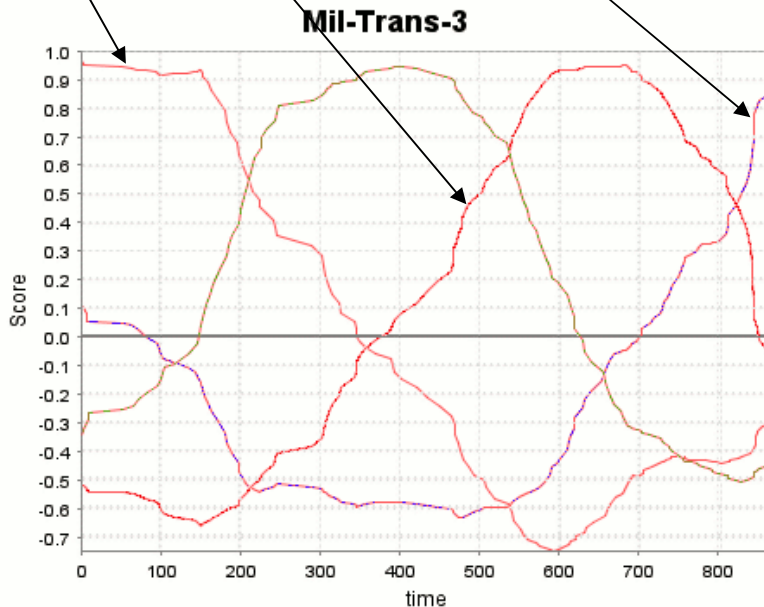
Topic Identification

Topic Modelling

T1 = office, website, intelligent, role, logistics ...

T3 = assist, document, command, review ...

T4 = demo, text, extract, compose ...



- Model topics as probabilistic word vectors
 - Can find most relevant topic for a given time/segment
 - ... or likely times/segments for a given topic
- 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
 - Segmentation accuracy $P_k \sim 0.33$
 - Joint work w/ MIT/Berkeley
 - (Purver et al., 2006)

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

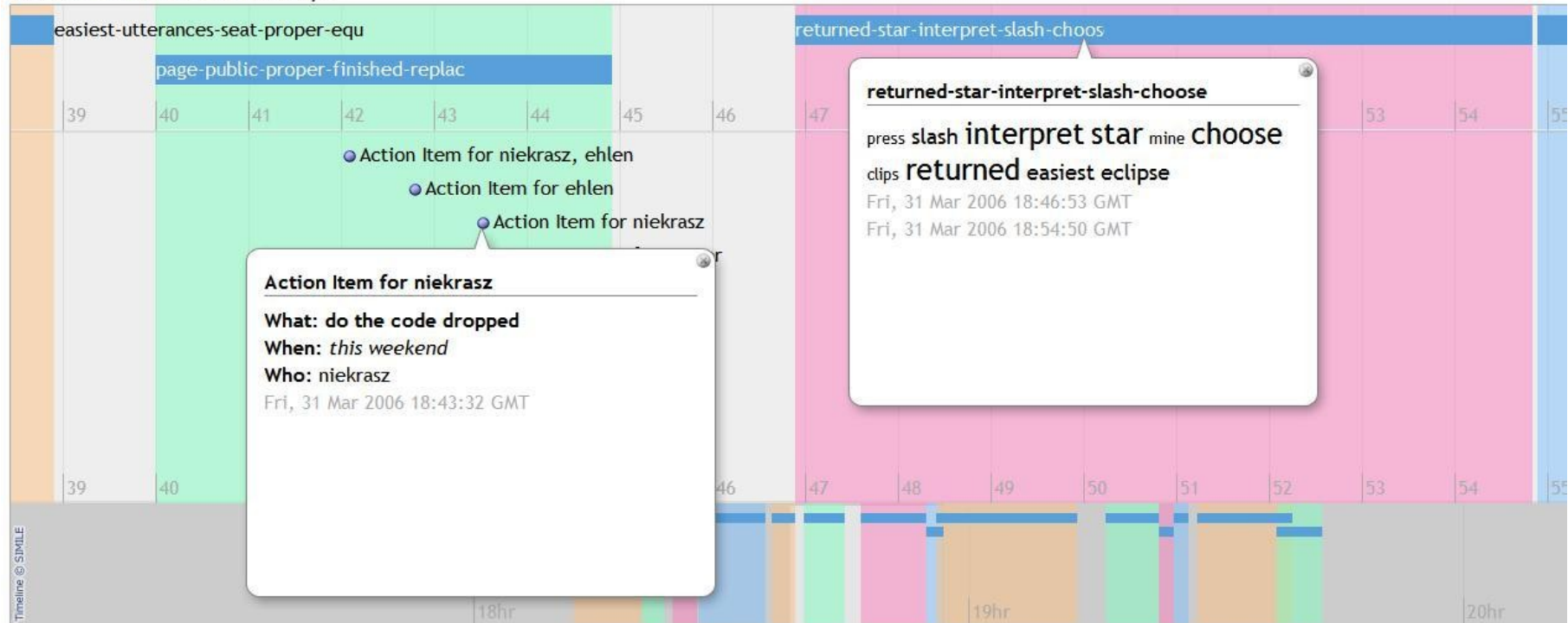
Discourse Features

- Topic shifts aren't just indicated by vocabulary change
 - Changes in speaker activity
 - Presence of silence
 - Presence of discourse markers “So”, “Anyway”, ...
 - (see TDT, AMI etc.)
- Including these helps segmentation accuracy
 - $P_k = 0.33 \rightarrow 0.26$
 - $WD = 0.36 \rightarrow 0.33$
- (Dowman et al., 2008)

Meeting Browser

CSLI-N01-31-03-2006 Meeting Timeline

Click on timeline to view transcript.



niekrasz: because i'll probably do the code dropped like this weekend like
and i'd like to send a short couple paragraphs tilman about where stuff is
how to how to get started um but it's not a big deal if there are some holes it's totally not
going to matter

ehlen: okay

niekrasz: it would be nice yeah

ehlen: this but on so for the browser java doctoring we really just need

Meeting Browser

Summary	Transcript	Action Items	Topics	QA Pairs	Ink	Meeting Notes	Mark Meeting
What To Do	When To Do It	Who Should Do It	My Actions				
organizing committee	the first week of november	Melinda_ [REDACTED]					
we should probably get a report from her so	but the plan is that	Pauline_E [REDACTED]					
an action item for me would be to try to find and i'm not three to wave you know and	come back to you tomorrow	Thierry_D [REDACTED]					
set up a meeting	next week		<div style="border: 1px solid gray; padding: 5px;"> <p>next week</p> <p>will try</p> <p>oh yeah so kelly was very excited i did and on factor and so we'll try and set up something for next week as an action item</p> <p>you get as an action item to schedule a meeting for sometime next week</p> <p>transition</p> </div>				
to look							
the hero	be interesting	Tim_ [REDACTED]					

Conversational Event Detection

Action Items & Decisions

- Problem 1: detect the event regions
 - Action item assignment (public commitment to a given task)
 - Decision-making (public agreement to a course of action)
- Problem 2: extract useful description (properties)
 - Task, responsible party, deadline
 - Issue, agreed resolution
- (1) is difficult enough!
- Our approach: use (2) to help (1)
 - Discussion regions have 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

- (No precedent in dialogue)
- 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 (> 0.8)
- Per-sentence/message 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?

- Annotated 65 meetings for action item utterances
 - (Gruenstein et al, 2005)
 - ICSI Meeting Corpus (Janin et al., 2003)
 - ISL Meeting Corpus (Burger et al., 2002)
- Annotator agreement poor ($\kappa = 0.36$)
- Try binary classification (Morgan et al., 2006):
 - Different classifier types (SVMs, MaxEnt)
 - Different features available (no email features; prosody, time, dialogue acts)
- Classification accuracy poor
 - F-score = 0.32 even on subset of data with best agreement

Decision Detection in Dialogue

- AMI Project tried similar approach for decisions
 - (Hsueh & Moore, 2007)
- Mark utterances as “decision-related”
 - Based on whether they belong in an extractive summary
- Binary classification using MaxEnt
 - Lexical, prosodic, dialogue act, topic features
- Classification accuracy poor
 - F-score = 0.35

What's going on?

- Sparse phenomena (1 – 5% of utterances)
- Discussion tends to be split across utterances & people
 - Contrast to email, where sentences are complete, tasks described in single sentences
- Utterances form a very heterogeneous set
 - Perform different dialogue functions
 - Address different event attributes
- Difficult for humans to decide which utterances are “relevant”
 - cf. Core & Allen (1997) DAMSL 'commit' tag $\kappa = 0.15$
 - Doesn't make for very consistent training/test data
- Automatic classification performance is correspondingly poor

Discussing an Action Item

- 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 Items

- 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 Acts

- Four types of dialogue act:



Action Item Dialogue Acts

- Four types of dialogue act:
 - Description of task

Somebody needs
to fill out this
report!



Action Item Dialogue Acts

- Four types of dialogue act:
 - Description of task
 - **Owner**

Somebody needs
to fill out this
report!

I guess I could
do that.



Action Item Dialogue Acts

- Four types of dialogue act:
 - Description of task
 - Owner
 - **Timeframe**



Can you do it
by tomorrow?

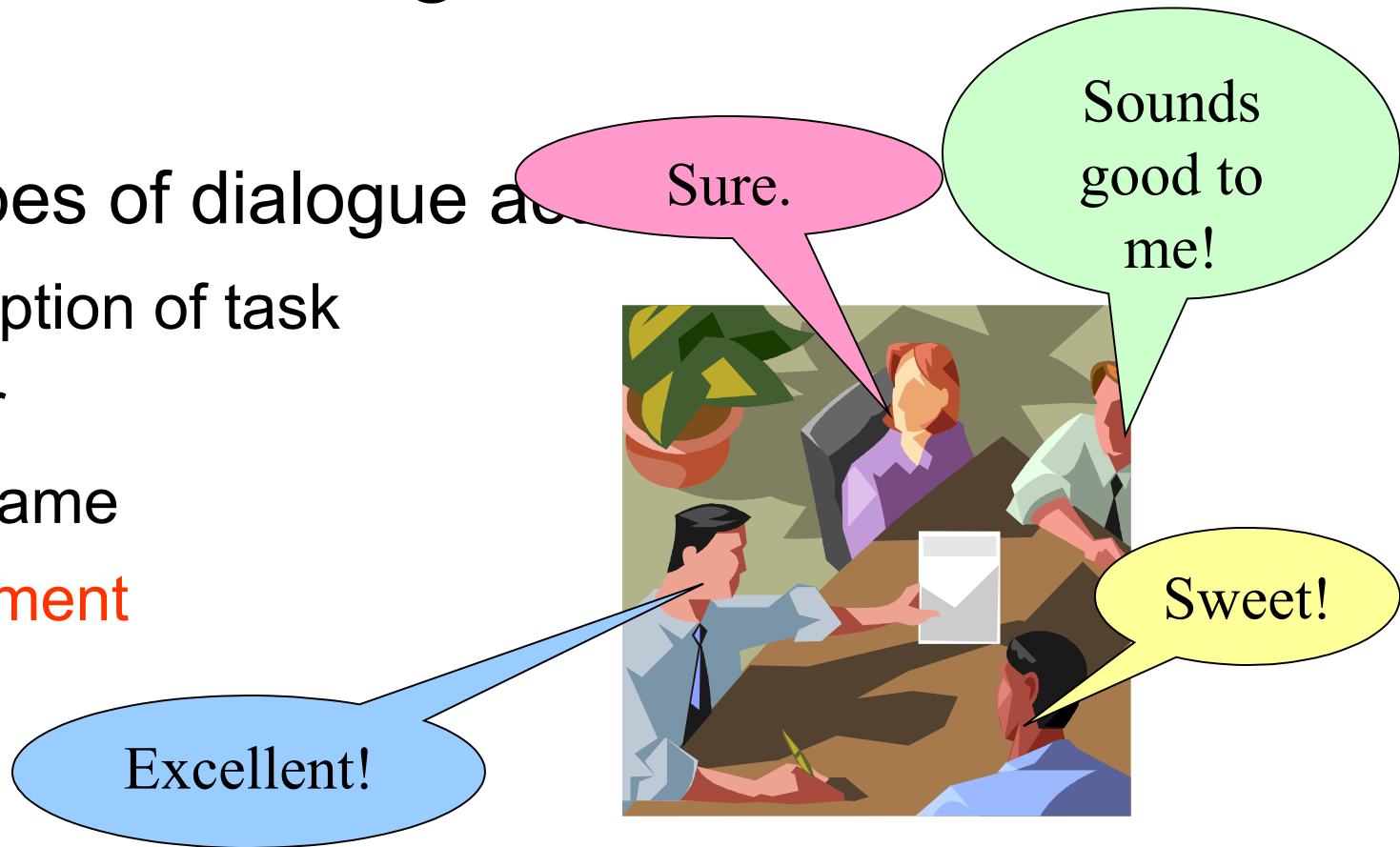
Action Item Dialogue Acts

- Four types of dialogue acts
 - Description of task
 - Owner
 - Timeframe
 - **Agreement**

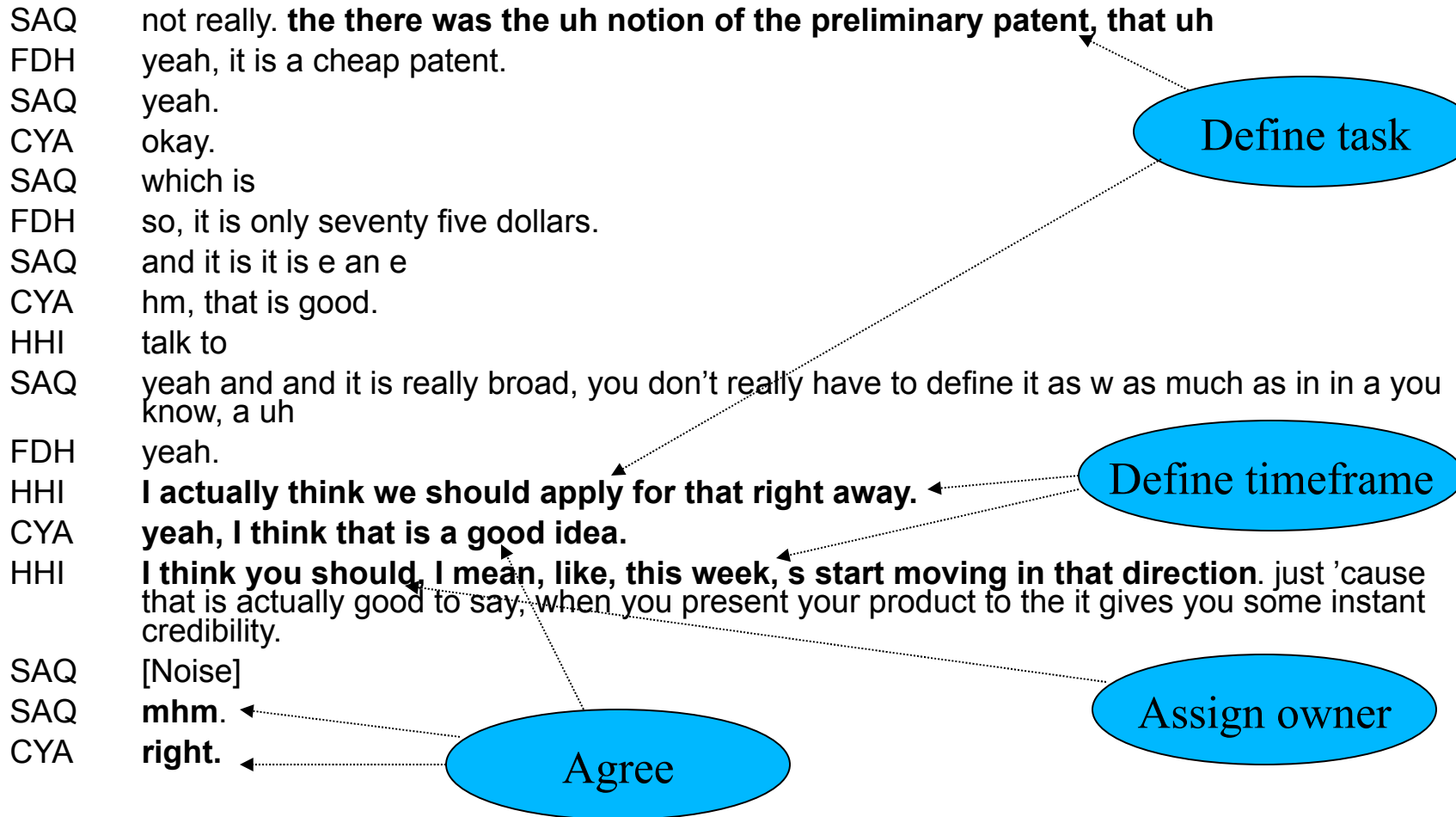


Action Item Dialogue Acts

- Four types of dialogue acts
 - Description of task
 - Owner
 - Timeframe
 - **Agreement**



Discussing an Action Item



Exploiting discourse structure

- Utterances can play different roles
 - Proposing, discussing 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” utterances
 - 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

Discussing a Decision

A: **Are we going to have a backup?**

A: Or we do just—

B: **But would a backup really be necessary?**

A: I think maybe **we could just go for the kinetic energy** and be bold and innovative.

C: **Yeah.**

B: I think— yeah.

A: It could even be one of our selling points.

C: Yeah —laugh—.

D: Environmentally conscious or something.

A: **Yeah.**

B: **Okay, fully kinetic energy.**

D: **Good.**

Discussing a Decision

A: **Are we going to have a backup?**

Define issue

A: Or we do just—

B: **But would a backup really be necessary?**

A: I think maybe **we could just go for the kinetic energy** and be bold and innovative.

Propose resolution

C: **Yeah.**

B: I think— yeah.

A: It could even be one of our selling points.

C: Yeah —laugh—.

D: Environmentally conscious or something.

Restate resolution

A: **Yeah.**

B: **Okay, fully kinetic energy.**

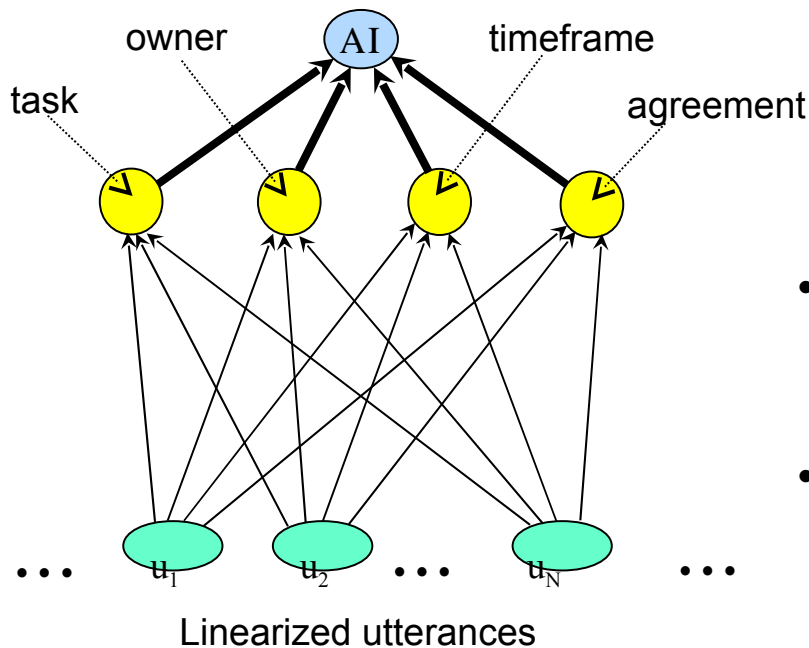
D: **Good.**

Agree

Structured Annotation

- Annotate utterances according to their role in the action item/decision discourse
 - Assign AIDA/DDA “dialogue act” tags
 - Allow multiple tags
 - Not all tags required (although always present for decisions)
- Improved inter-annotator agreement
 - AIDA = 0.86 (timeframe) → 0.73 (agreement/description)
 - DDA = 0.73 (resolution prop) → 0.63 (agreement)
- Between-class distinction (cosine distances)
 - Agreement vs. any other is good: 0.05 to 0.12
 - Owner/timeframe/description: 0.36 to 0.47

Hierarchical Classifier



- Individual “dialogue act” classifiers
 - Support vector machines
 - Lexical (n-gram) features
 - Utterance, speaker features
 - Prosody, dialogue act tags, syntactic & semantic parse features not so much help
- Event region “super-classifier”
 - Features are the sub-classifier outputs over a window of N utterances
- Performance for each “act” type compares to previous overall performance
 - ICSI AIDAs: f-scores 0.1-0.3
 - CALO AIDAs: f-scores 0.3-0.5
 - AMI DDAs: f-scores 0.2-0.4
 - (with a basic set of features)

Event Detection Results

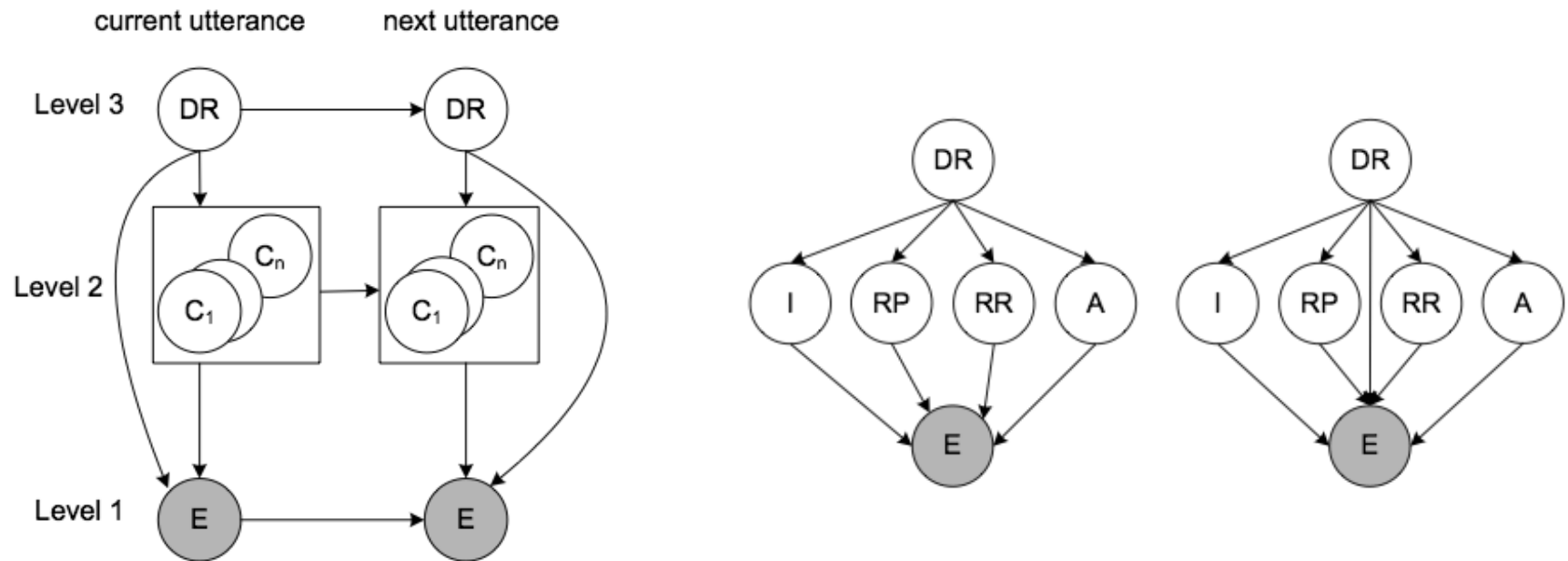
- Evaluation at the utterance level not quite what we want
 - Are agreement utterances important? Ownership? Near misses?
 - Look at overall discussion f-scores, requiring overlap by 50%
- Action items: 20 ICSI meetings, cross-validated (Purver et al 2007)
 - Recall 0.64, precision 0.44, f-score 0.52 (flat baseline 0.35)
 - (better on CALO test meetings → 0.67)
- Decisions: 17 AMI meetings, cross-validated (Fernandez et al 2008)
 - Recall 0.88, precision 0.43, f-score 0.63 (flat baseline 0.45)
- Slight sub-classifier improvements (but small)
- Robustness to ASR output reasonable
 - Use word confusion networks from SRI's Decipher
 - Absolute f-score drop AIs 7%, decisions 7%

The Wrong Classifier!

- Super-classifier is naïve
 - SVM: feature array = $N \times D$ sub-classifier outputs
 - DA “sequence” expressed as SVM feature index
- Sequence models (HMM, CRF) don't help
 - DA sequence is flexible
 - Utterances can perform multiple functions
 - DDA RR follows RP, but optional and can repeat
 - AIDA owner/timeframe unordered
 - Agreement can be distributed
- We need something more flexible ...

The Right Classifier?

- Hierarchical directed graphical model (DBN)



- DDA f-scores 0.7 – 0.8 (Bui & Peters, to appear)

Meeting Browser

Summary	Transcript	Action Items	Topics	QA Pairs	Ink	Meeting Notes	Mark Meeting
What To Do	When To Do It	Who Should Do It	My Actions				
organizing committee	the first week of november	Melinda_ [REDACTED]					
we should probably get a report from her so	but the plan is that	Pauline_E [REDACTED]					
an action item for me would be to try to find and i'm not three to wave you know and	come back to you tomorrow	Thierry_D [REDACTED]					
set up a meeting	next week		<div style="border: 1px solid gray; padding: 5px;"> <p>next week</p> <p>will try</p> <p>oh yeah so kelly was very excited i did and on factor and so we'll try and set up something for next week as an action item</p> <p>you get as an action item to schedule a meeting for sometime next week</p> <p>transition</p> </div>				
to look							
the hero	be interesting	Tim_ [REDACTED]					

Extracting Summaries

- Structured classifier gives us the relevant utterances
 - Hypothesizes which utterances contain which information
 - Action item: task, timeframe, owner
 - Decision: issue, agreed resolution
- Extract the useful entities/phrases for descriptive text?
 - Semantic parsing over WCNs can help in some cases
 - e.g. action item timeframe (Purver et al., 2007)
 - Time expressions, names, event structures
 - But hard to beat presenting 1-best ASR hypothesis
- Can learn to improve given user feedback
 - Infer training data, re-learn
 - (Purver et al., 2008; Ehlen et al., 2008)

Some Good Examples

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

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


ignore this one

Some Bad Examples

not an action item ✕

i don't think an action for myself um to talk to donald about


 *Clint_Frederickson*


 *I don't think an action for myself um to talk to donald about*

✕ *ignore this one*

not an action item ✕

there should have been a lot of e mail in that database as well

 *Jim_Carpenter*

 *uh so called the second week text extraction*

✕ *ignore this one*

Extracting Ownership

- What do we do with personal reference?
 - Action item ownership in particular
 - 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 via “I”: speaker
 - Individual microphones, login names
 - (otherwise, it’s a speaker ID problem)
- Other-assignments “you”: addressee
 - Addressee ID is hard, but approachable
 - (Katzenmaier et al., 2004; Jovanovic et al., 2006, 70-80% accuracy)
 - Need to know when “you” refers to the addressee ...

“You”-resolution

Second-Person Pronouns

- What does “you” refer to?
 - (the addressee, right?)
 - In two-party dialogue, this seems trivial ...
- In multi-party dialogue, it's not
 - Who is the addressee?
 - Who are the addressees?
- Actually, in two-party dialogue, it's not either
 - “You” is often not addressee-referring at all

Does “you” refer?

- *Deictic* “you” refers to the addressee:
 - “I’m not going to do it. *You* do it.”
 - “Could you send me an email?”
- *Generic* “you” refers to no-one in particular:
 - “On entering the church you are struck by the stained glass windows”
 - “If you send an email they just ignore it”
 - cf. French “on”?
- *Discourse markers* aren't referential at all:
 - “It's just, you know, noises or something.”

How many addressees?

- Sometimes just one:
 - “Tim, I think you should do it.”
- Sometimes more than one:
 - “Do you guys have any questions?”
 - “Tim and Tina, you should do this.”
 - “How are y'all doing?”
 - cf. French “vous”?
- A subset of those present? Everyone?



“You” categories

- 10 meetings from AMI corpus (4 participants)
 - 948 “you”-containing utterances
 - Good inter-annotator agreement ($\kappa = 0.84$)

Discourse marker	72			8%	
Generic	431			45%	
Deictic	445			47%	
Singular		290			68%
Plural		137			32%
Everyone			130		
Subgroup			7		(< 2%)

“You” addressees

- Can treat people as individuals (4 classes)
 - Designer, marketer, manager, UI person

ID	ME	PM	UI	Total
80	70	51	89	290
28%	24%	18%	31%	

- Or by position relative to speaker (3 class)
 - L1 = opposite, L2 = diagonal, L3 = next to

L1	L2	L3	Total
102	88	100	290
35%	30%	35%	

“You” categories

- No examples with multiple classes/addressees
 - although it certainly seems possible:
 - “Do you think that if you do something generic ...”
 - “Tom, you do this, and Tina, you do that”
- Except: discourse marker + something else
 - “You know, I think you should do that, Tom.”
- But: discourse markers are *easy*
 - “ \neg (do|as) you know \neg (how|that)” → 99% accuracy
- So remove them, and work per utterance

“You” resolution

- 5-way or 6-way problem
 - generic, L1, L2, L3, plural
 - generic, ID, ME, PM, UI, plural
- Not quite the same as general addressee ID
 - See (Katzenmaier et al., 2004; Jovanovic et al., 2007)
- Multiple sources of information:
 - Transcripts (manual/ASR)
 - Dialogue act tags (manual/automatic)
 - Speaker diary (close-talking mics)
 - Video? (webcams with relative positioning)

Intra-utterance Features

- Sentential patterns
 - “if you”, “whenever”, etc. → generic?
 - “do you”, “have you”, “you said” → deictic?
 - “Tom”, “Tina” etc. → addressee? “you guys” → plural?
- Prosody
 - Pitch, intensity of “you” (more stressed → deictic?)
 - Duration, speech rate (generic → faster?)
- Dialogue acts
 - Commands, questions → deictic?
- Lexical
 - All words, ngrams

Context Features

- Speaker activity
 - Who speaks next?
 - (and after that ...)
 - Who spoke last?
 - (and before that ...)
- Utterance context
 - Overlaps, long pauses → not addressee?
 - Common material → addressee?
 - Dialogue act combinations → addressee?

Visual Features

- Who's the speaker looking at?
 - During the “you”
 - During the start/end of the utterance
- Who's looking at the speaker?
 - At various points
- Is there mutual gaze?
- Is there an (un)equal distribution?

Chuck it all in?

- Bayesian Network classifier
 - 10-fold cross-validation
- Fairly poor performance
 - Baseline 51% accuracy (always generic)
 - Best 62% accuracy (→ 56% with ASR/auto features)
 - F-scores:
 - 75% for generic
 - 41% for plural
 - 38-60% for individual addressees

Different problems ...

- Some aspects seem sentential
 - generic vs deictic distinction
- Some aspects seem interactional
 - individual addressee reference
- Some might be in between
 - singular vs plural distinction
- Should we treat them differently?
 - reduced feature space, optimal classifiers ...

Generic vs Deictic

- Sentential features do well (79%)
 - Generic words “always” etc.; multiple “you”s
 - Names, first-person pronouns
- Fully lexicalizing does best (88% → 85% with ASR)
 - Dialogue act-relevant n-grams
 - Vocabulary: meeting topic → generic; management → deictic
- Context features don't help at all
- Visual features beat the baseline (60%)
 - But don't help above anything else
- Perhaps this problem is really sentential?

Singular vs Plural

- Baseline 69% accuracy (singular)
- Sentential features (lexicalizing) best (83% → 77% ASR)
 - Plural reference (“we”)
 - Questioning (singular) vs statement (plural)
- Context features beat the baseline
 - Speaker activity (one speaker → plural)
 - Utterance similarity (higher → singular)
 - But don't help above lexicalizing
- Visual features don't help at all (66%)
 - (although gaze at whiteboard → plural)

Singular Reference

- Context features excellent
 - Baseline = “next speaker”: 71% (MC baseline 35%)
- Sentential features alone poor (49%)
- Context features good (72%)
 - Next speaker(s), previous speaker(s)
 - Intervening time/utterances, utterance similarities
 - Backward-looking only (online) 59%
- Visual features good (74%)
- Combining them even better (84% → 74% ASR/auto)
- Perhaps this problem *isn't* really sentential?

Cascaded Classification

- Treat the problems separately (pipeline)
- Generic vs. deictic
 - plural vs singular
 - individual reference
- Use context where it's helpful
- Use vision where it's helpful
- Optimise techniques/features for each
 - 78% accuracy (→ 72% ASR/auto)
 - Individual speaker f-scores 0.64 → 0.83

Conclusions?

- Some problems are really about language
- Some problems really aren't ...
 - take context (interactional, visual) into account!
- Multi-party dialogue is complicated
 - even “simple” problems get hard
 - take context (interactional, visual) into account!
 - (even if you can't model it fully.)