The Dangling Conversation: A Corpus and Algorithm for Conversation Disentanglement

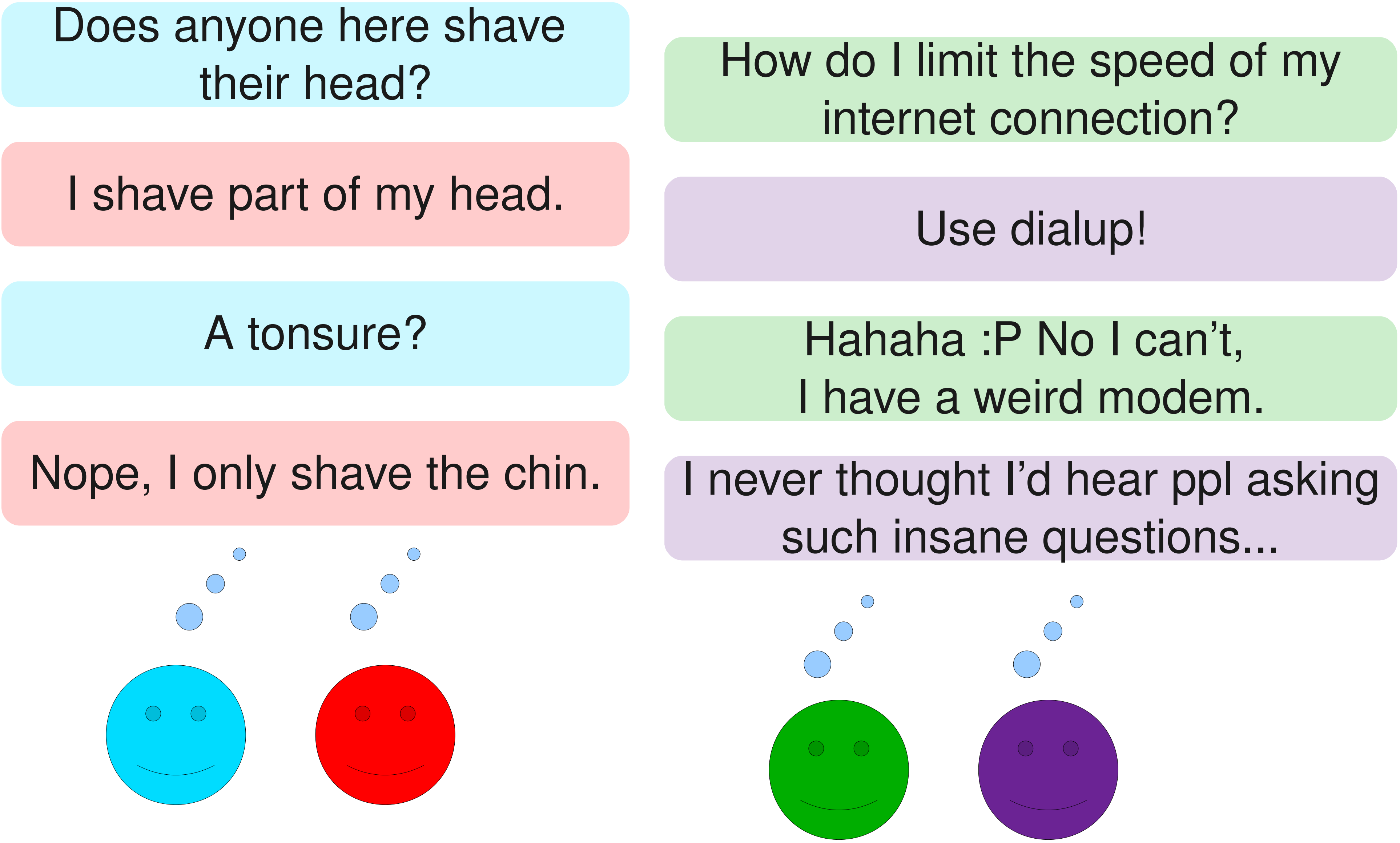
Micha Elsner and Eugene Charniak

Brown Laboratory for Linguistic Information Processing (BLLIP)

BROWN

21 Jan 2009, University of Maryland

Life in a Multi-User Channel



Does anyone here shave their head?

I shave part of my head.

A tonsure?

Nope, I only shave the chin.

How do I limit the speed of my internet connection?

Use dialup!

Hahaha :P No I can’t,
I have a weird modem.

I never thought I’d hear ppl asking such insane questions...

Real Life in a Multi-User Channel

Does anyone here shave their head?

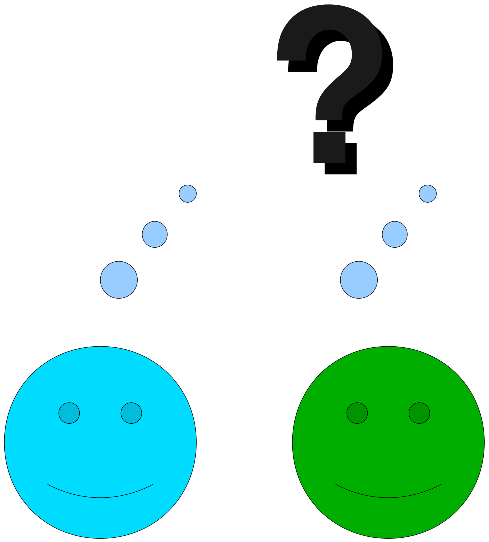
How do I limit the speed of my internet connection?

I shave part of my head.

A tonsure?

Use dialup!

Nope, I only shave the chin.



- A common situation:
  - Text chat
  - Push-to-talk
  - Cocktail party

Why Disentanglement?

- A natural discourse task.
  - Humans do it without any training.
- Preprocess for search, summary, QA.
  - Recover information buried in chat logs.
- Online help for users.
  - Highlight utterances of interest.
  - Already been tried manually: Smith et al ‘00.
  - And automatically: Aoki et al ‘03.

Outline

- Corpus
  - Annotations
  - Metrics
  - Agreement
  - Discussion
  - Features
- Modeling
  - Previous Work
  - Classifier
  - Inference
  - Baselines
  - Results
- Extensions
  - Specificity Tuning
  - Conversation Start Detection

Questions are welcome!

Outline

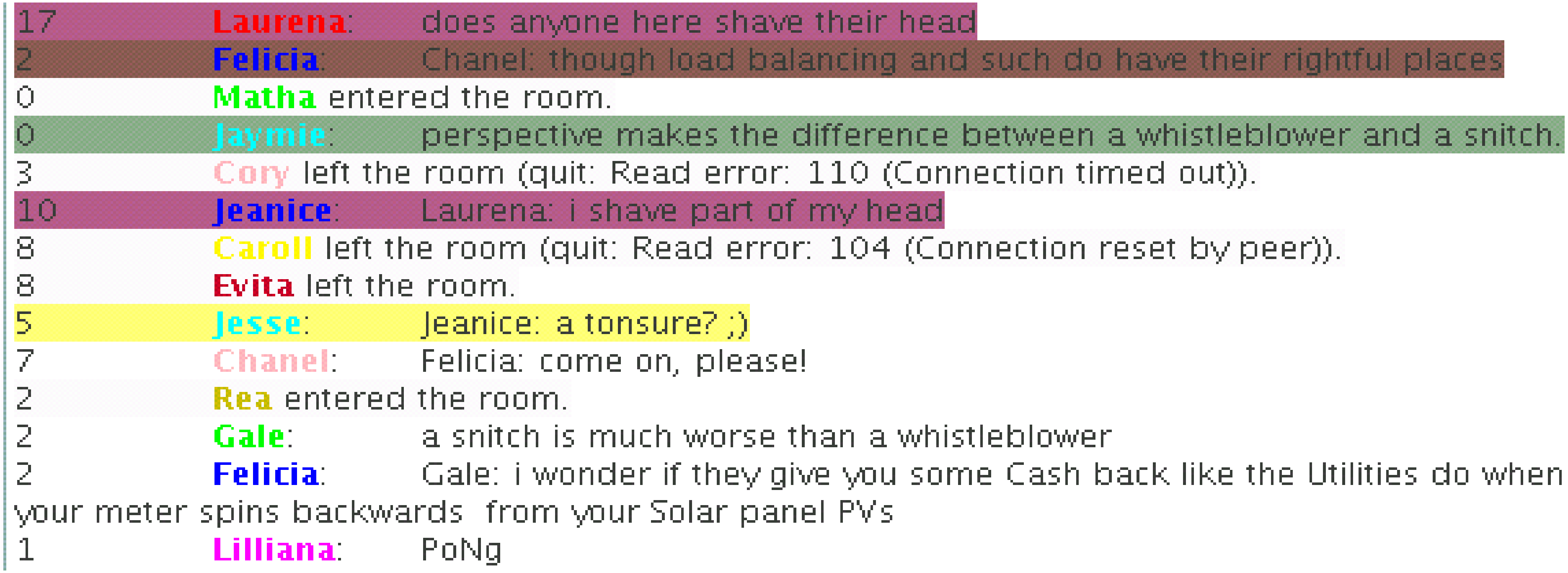
- Corpus
  - Annotations
  - Metrics
  - Agreement
  - Discussion
  - Features
- Modeling
  - Previous Work
  - Classifier
  - Inference
  - Baselines
  - Results
- Extensions
  - Specificity Tuning
  - Conversation Start Detection

Questions are welcome!

Dataset

- Recording of a Linux tech support chat room.
- 1:39 hour test section.
  - Six annotations.
  - College students, some Linux experience.
- Another 3 hours of annotated data for training and development.
  - Mostly only one annotation by experimenter.
  - A short pilot section with 3 more annotations.

Annotation



```
17          Laurena:      does anyone here shave their head
2           Felicia:      Chanel: though load balancing and such do have their rightful places
0           Matha entered the room.
0           Jaymie:       perspective makes the difference between a whistleblower and a snitch.
3           Cory left the room (quit: Read error: 110 (Connection timed out)).
10          Jeanice:      Laurena: i shave part of my head
8           Caroll left the room (quit: Read error: 104 (Connection reset by peer)).
8           Evita left the room.
5           Jesse:        Jeanice: a tonsure? ;)
7           Chanel:       Felicia: come on, please!
2           Rea entered the room.
2           Gale:         a snitch is much worse than a whistleblower
2           Felicia:      Gale: i wonder if they give you some Cash back like the Utilities do when
your meter spins backwards  from your Solar panel PVs
1           Lilliana:     PoNg
```

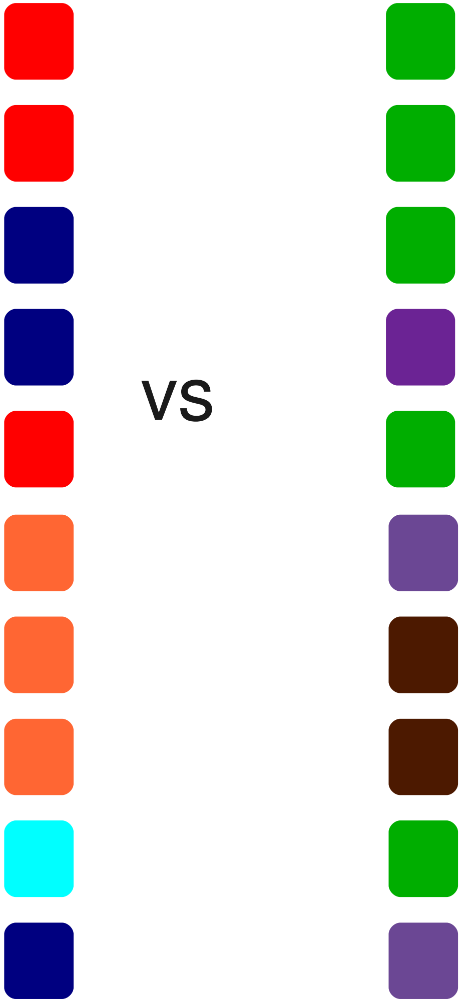
- Annotation program with simple click-and-drag interface.
- Conversations displayed as background colors.

Outline

- Corpus
  - Annotations
  - **Metrics**
  - Agreement
  - Discussion
  - Features
- Modeling
  - Previous Work
  - Classifier
  - Inference
  - Baselines
  - Results
- Extensions
  - Specificity Tuning
  - Conversation Start Detection

Questions are welcome!

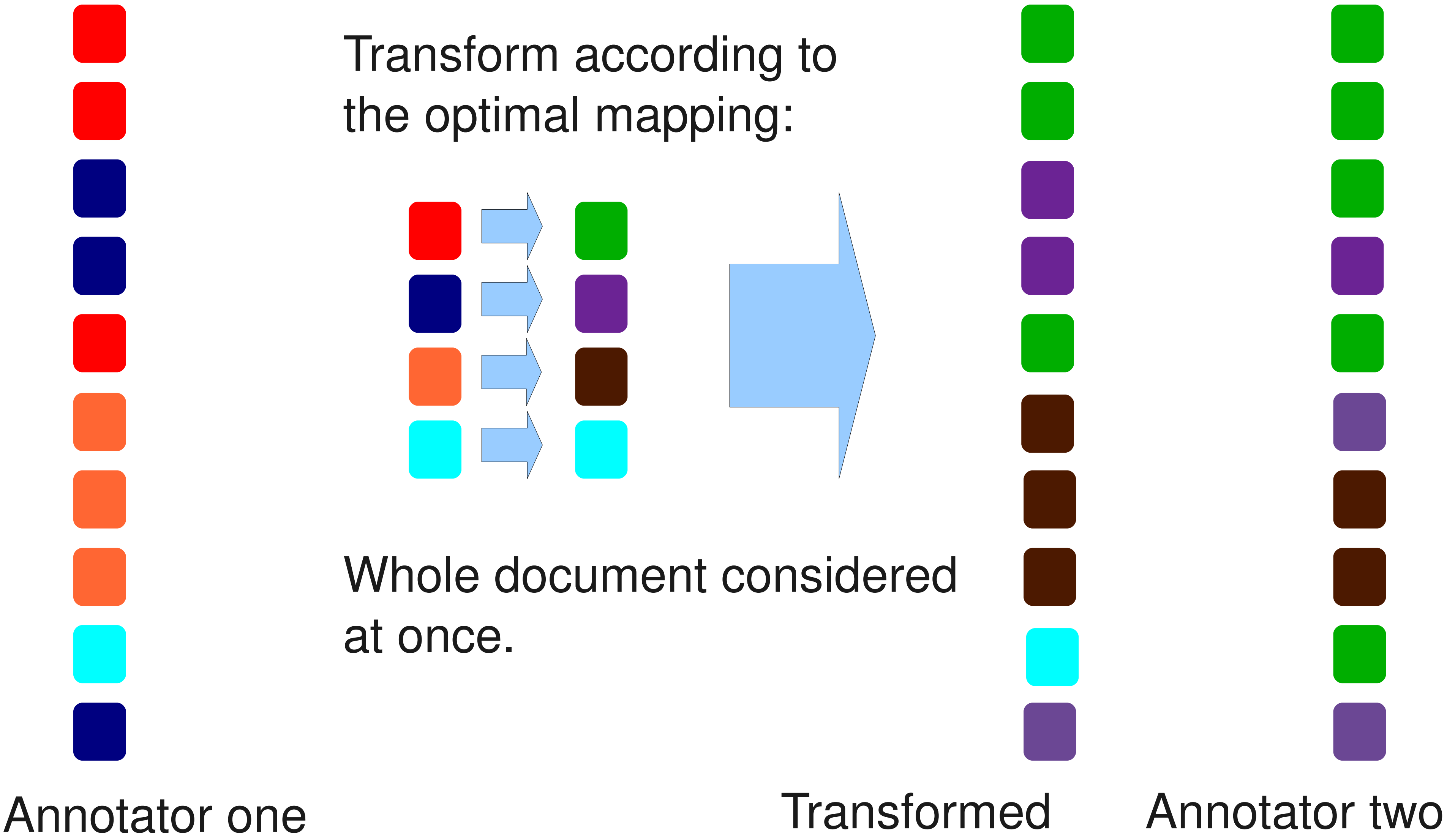
One-to-One Metric



vs

Two annotations of the same dataset.

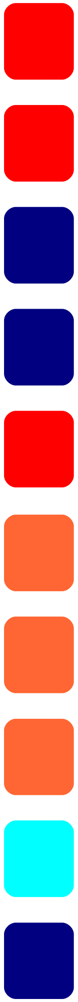
One-to-One Metric



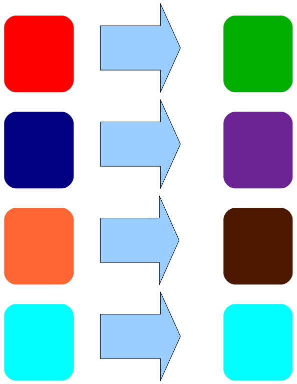
Transform according to the optimal mapping:

Whole document considered at once.

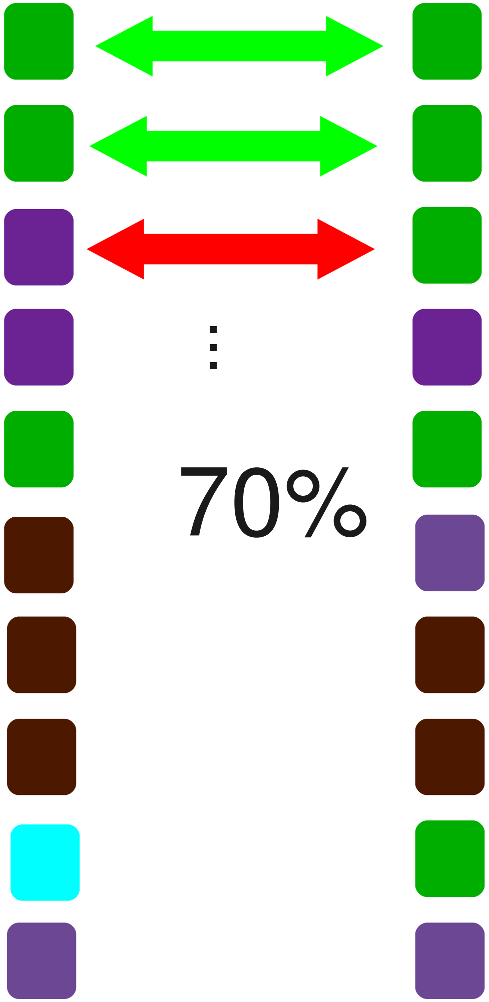
One-to-One Metric



Transform according to the optimal mapping:



Whole document considered at once.

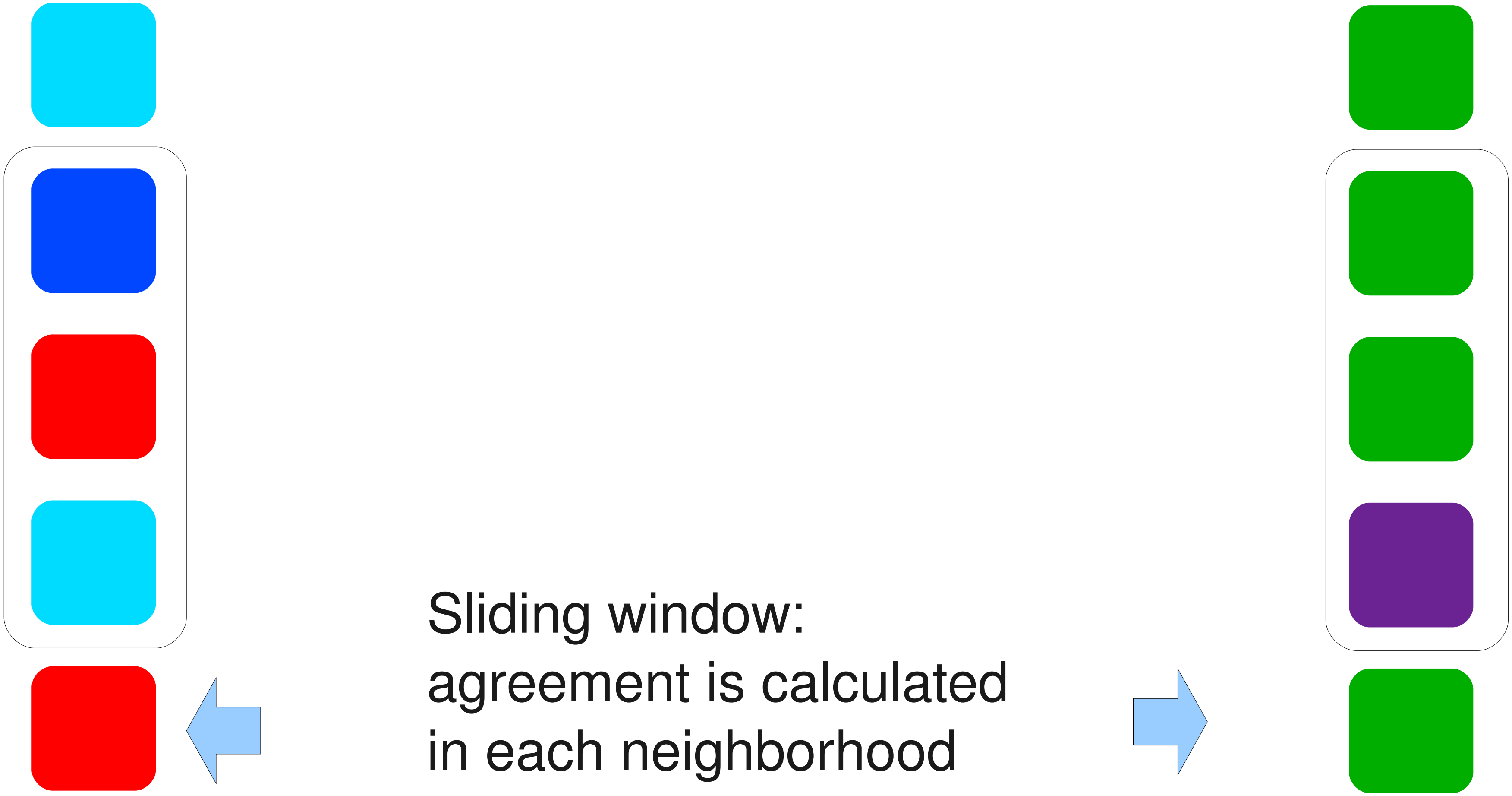


Annotator one

Transformed

Annotator two

Local Agreement Metric

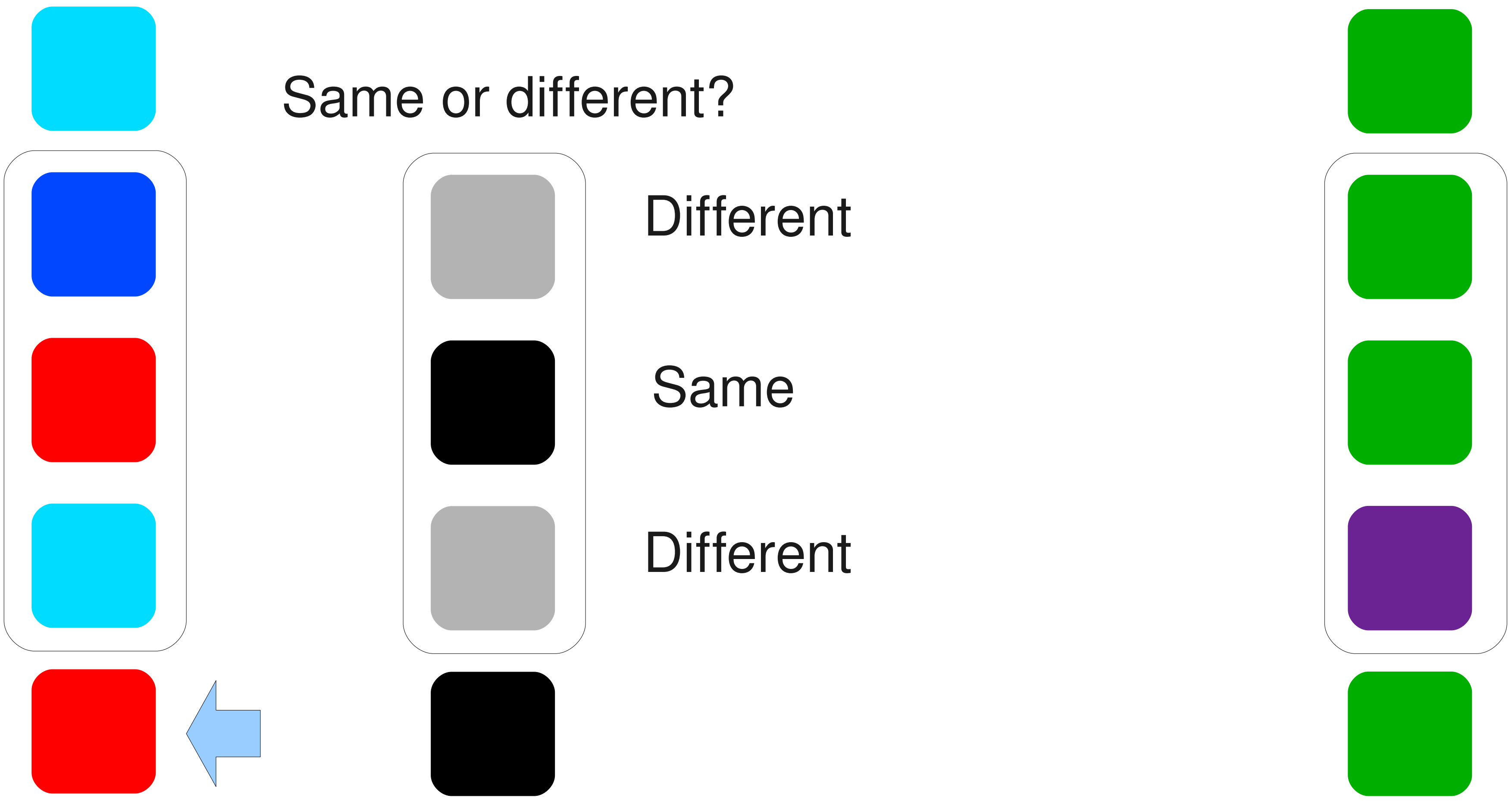


Sliding window:
agreement is calculated
in each neighborhood
of three utterances.

Annotator 1

Annotator 2

Local Agreement Metric



Same or different?

Different

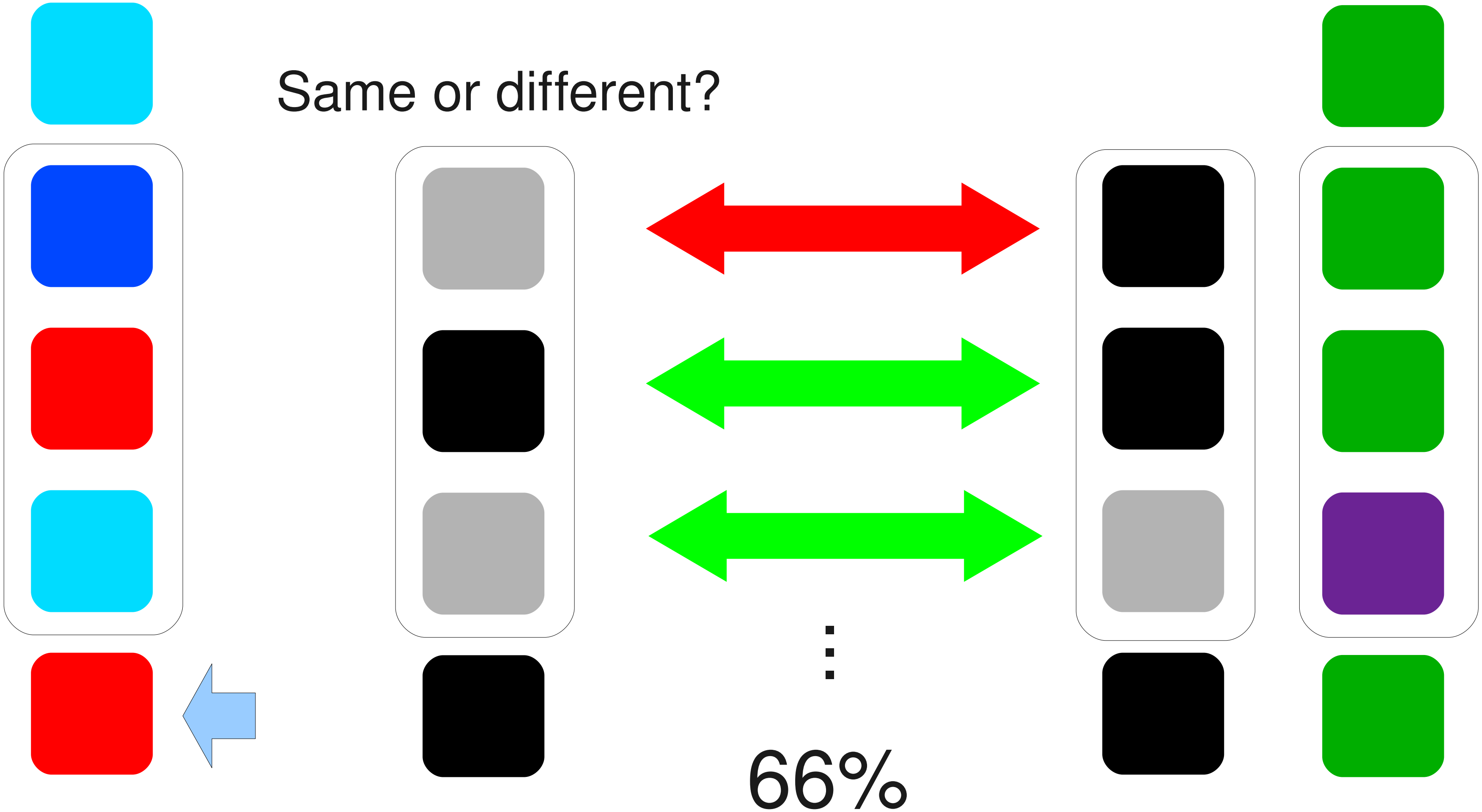
Same

Different

Annotator 1

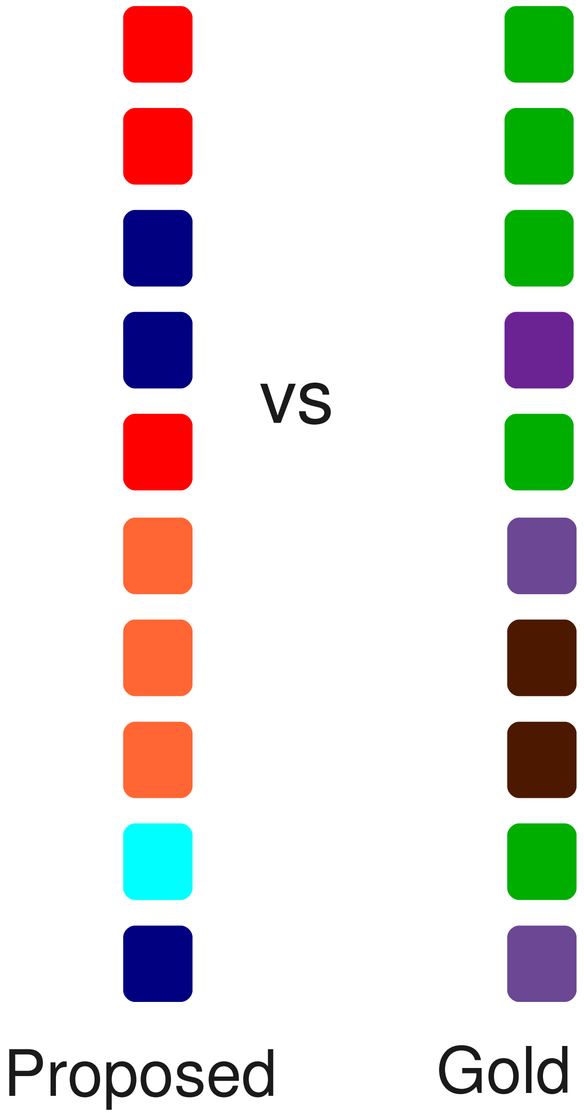
Annotator 2

Local Agreement Metric

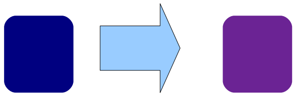


F-Score Metric

Shen et al '06
Adams + Martell '08



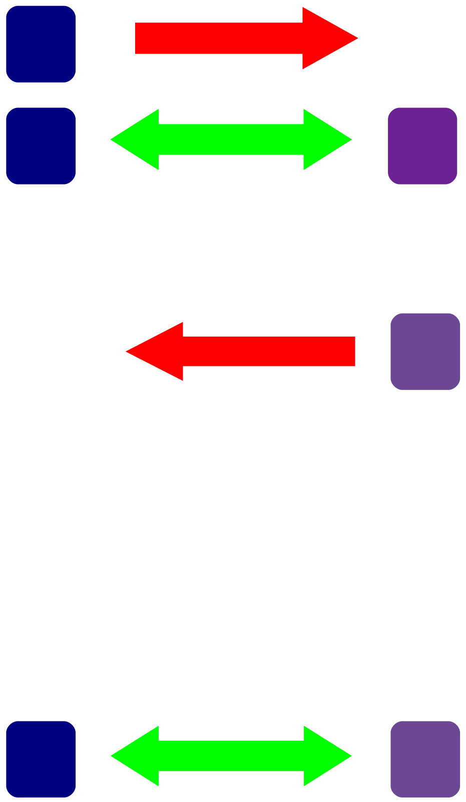
Define retrieval precision and recall for a single thread:



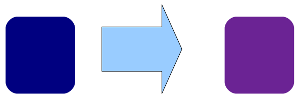
Not symmetric!

F-Score Metric

Shen et al '06
Adams + Martell '08



Define retrieval precision and recall for a single pair of threads:



$$\text{Prec} = \frac{\leftrightarrow}{\leftrightarrow + \rightarrow}$$

$$\text{Rec} = \frac{\leftrightarrow}{\leftrightarrow + \leftarrow}$$

F-Score Metric

- Defined by Shen for a whole transcript:
  - For every gold thread:
    - Match to best annotated thread.
  - Average weighted by thread size.
- Correlates well with one-to-one.

Interannotator Agreement

	Min	**Mean**	Max
One-to-One	36	**53**	64
Local Agreement	75	**81**	87

- Local agreement is good.
- One-to-one not so good!

How Annotators Disagree

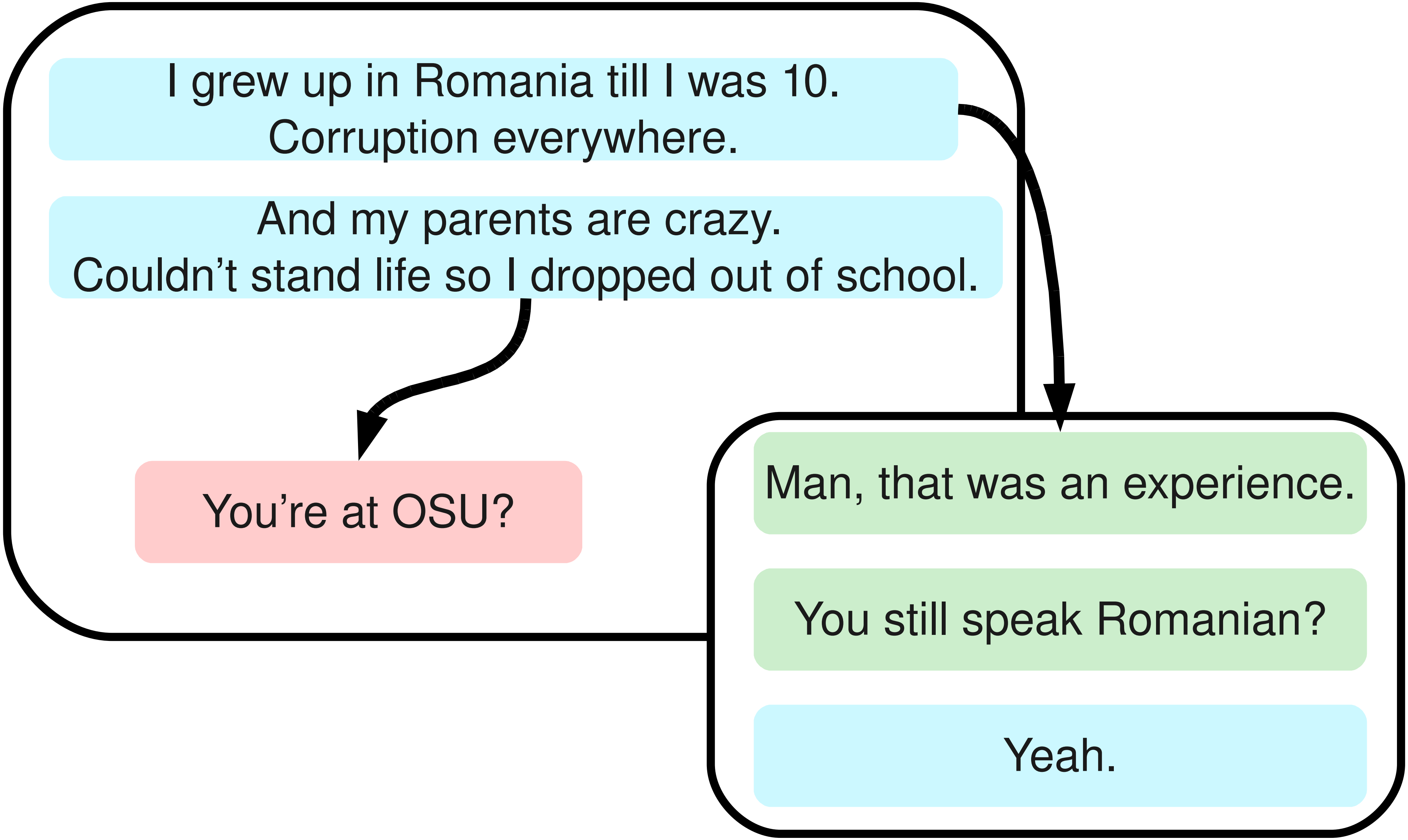
	Min	Mean	**Max**
# Conversations	**50**	81	**128**
Entropy	**3**	4.8	**6.2**

- Some annotations are much finer-grained than others.

Schisms

- Sacks et al ‘74: Formation of a new conversation.
- Explored by Aoki et al ‘06:
  - A speaker may start a new conversation on purpose...
  - Or unintentionally, as listeners react in different ways.
- Causes a problem for annotators...

To Split...



I grew up in Romania till I was 10.
Corruption everywhere.

And my parents are crazy.
Couldn’t stand life so I dropped out of school.

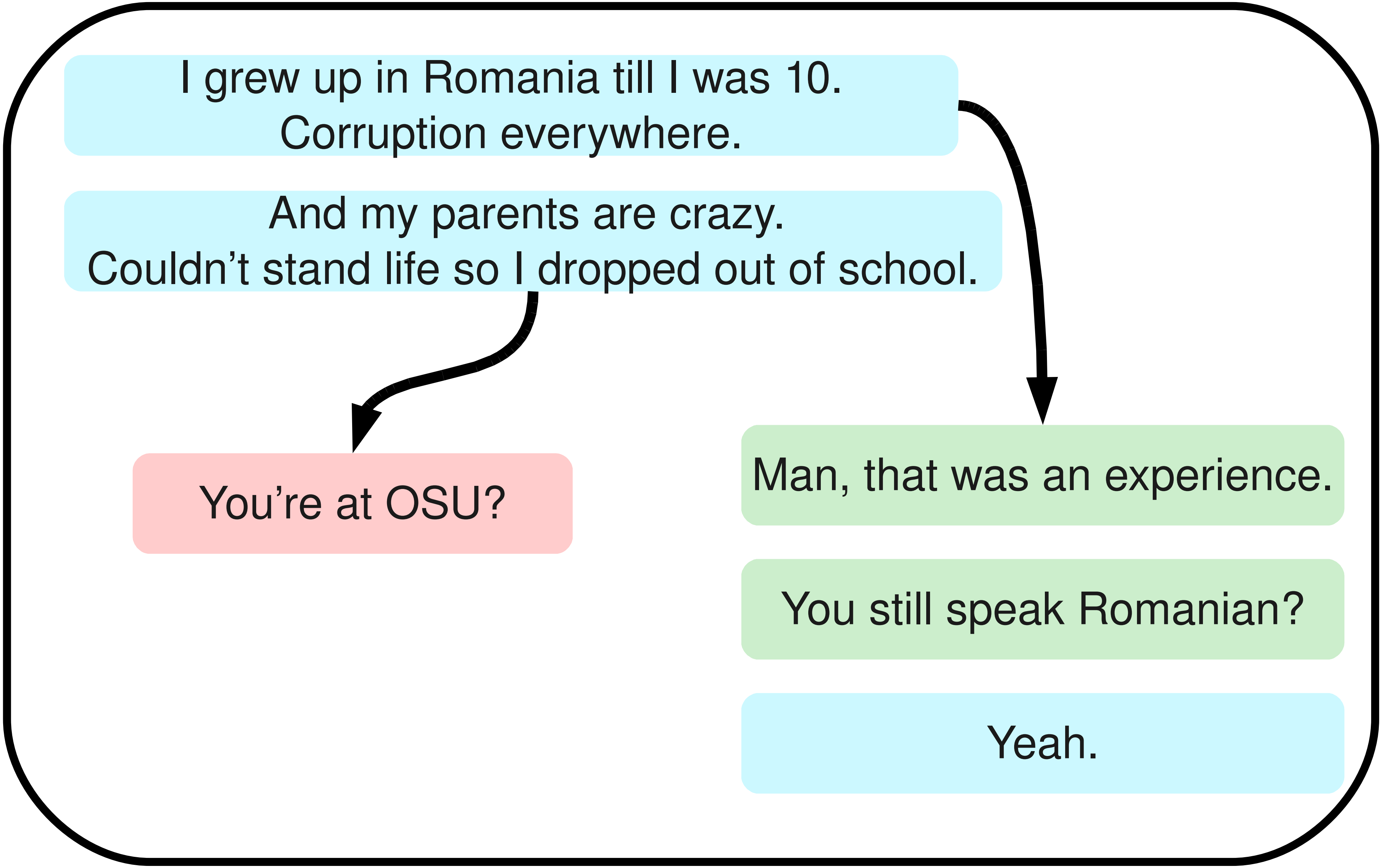
You’re at OSU?

Man, that was an experience.

You still speak Romanian?

Yeah.

Or Not to Split?



I grew up in Romania till I was 10.
Corruption everywhere.

And my parents are crazy.
Couldn’t stand life so I dropped out of school.

You’re at OSU?

Man, that was an experience.

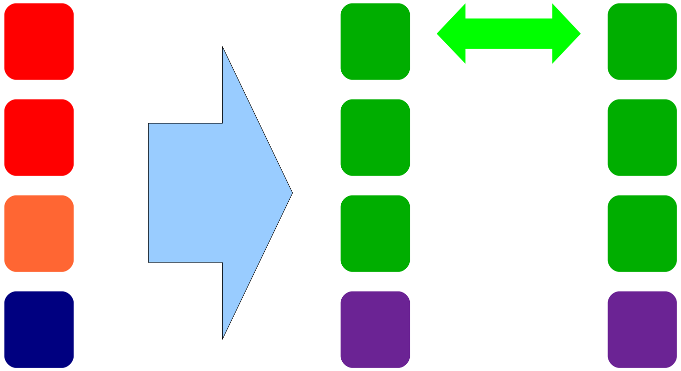
You still speak Romanian?

Yeah.

Accounting for Disagreements

	Min	**Mean**	Max
One-to-One	36	**53**	64
Many-to-One	76	**87**	94

Many-to-one mapping from high entropy to low:



First annotation is a strict refinement of the second.

One-to-one: only 75%
Many-to-one: 100%

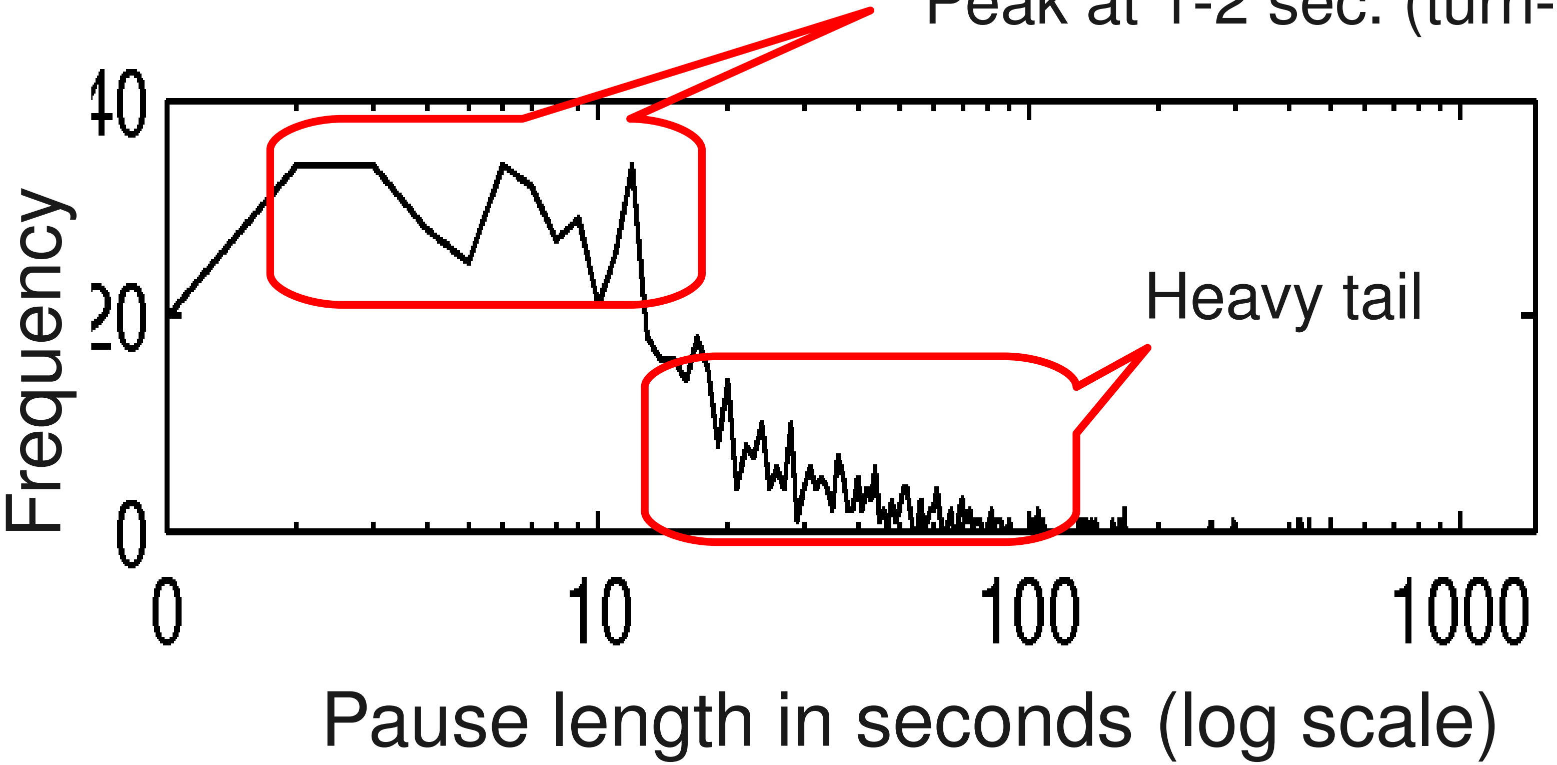
Outline

- Corpus
  - Annotations
  - Metrics
  - Agreement
  - Discussion
  - **Features**
- Modeling
  - Previous Work
  - Classifier
  - Inference
  - Baselines
  - Results
- Extensions
  - Specificity Tuning
  - Conversation Start Detection

Questions are welcome!

Pauses Between Utterances

A classic feature for models of multiparty conversation.



Peak at 1-2 sec. (turn-taking)

Heavy tail

Name Mentions

Sara Is there an easy way to extract files from a patch?

Carly Sara: No.

Carly Sara: Patches are diff deltas.

Sara Carly, duh, but this one is just adding entire files.

- Very frequent: about 36% of utterances.
- A coordination strategy used to make disentanglement easier.
  - O'Neill and Martin '03.
- Usually part of an ongoing conversation.

Outline

- Corpus
  - Annotations
  - Metrics
  - Agreement
  - Discussion
  - Features
- Modeling
  - **Previous Work**
  - Classifier
  - Inference
  - Baselines
  - Results
- Extensions
  - Specificity Tuning
  - Conversation Start Detection

Questions are welcome!

Previous Work

- Shen '06
  - Class discussion corpus
  - Unsupervised (geometric) clustering
  - TF-IDF features
  - ... and discourse features
- Adams + Martell '08
  - Discussion and Navy tactical chat
  - Geometric with TF-IDF

Previous Work

- Aoki et al ‘03, ‘06
  - Conversational speech
  - System makes speakers in the same thread louder
  - Evaluated qualitatively (user judgments)

- Camtepe ‘05, Acar ‘05
  - Simulated chat data
  - System intended to detect social groups

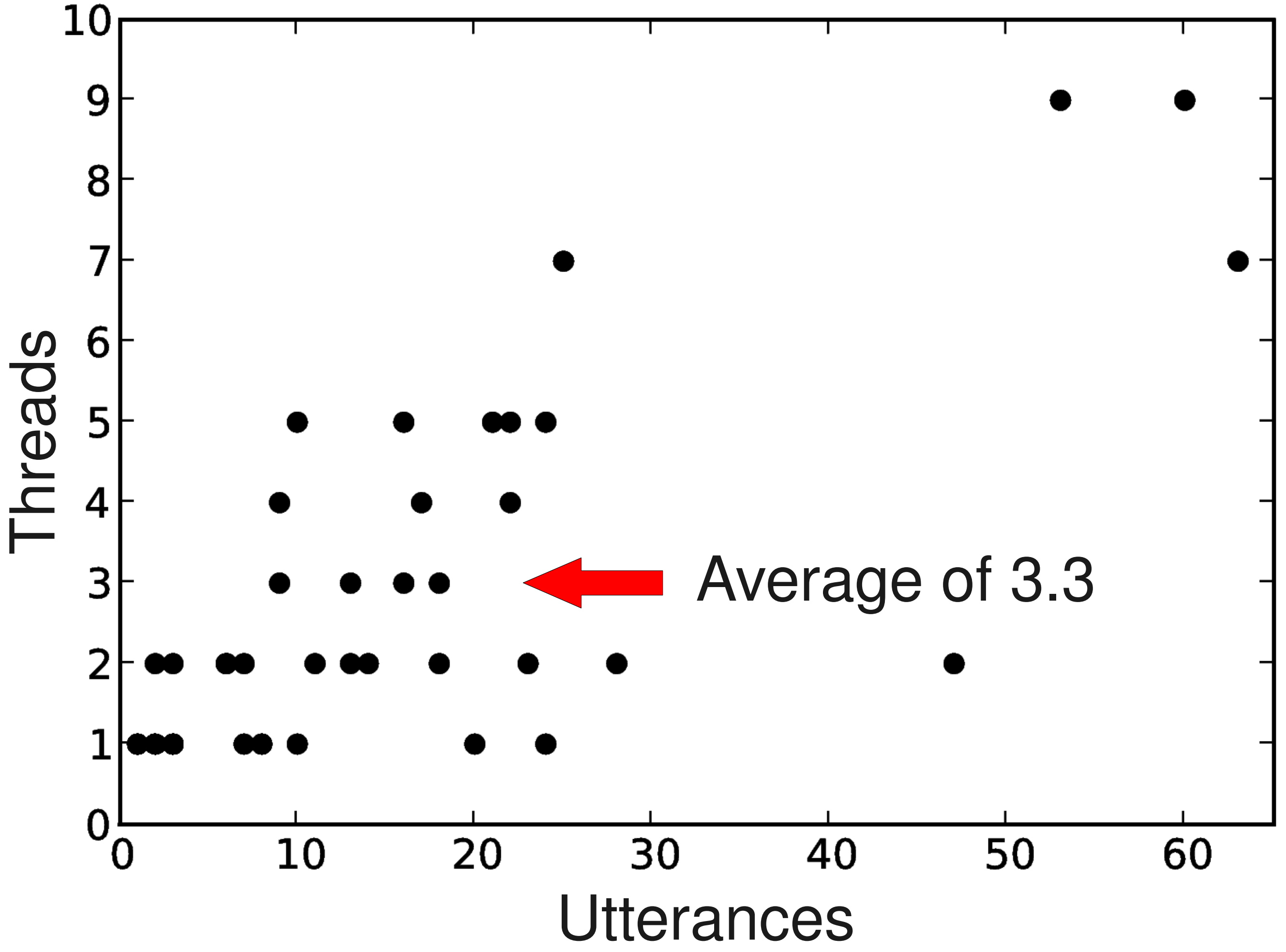
Previous Work

- Pause features critical for everyone.
- Lexical features:
  - Shen, Adams: very useful.
  - Acar '05: tries (badly), but no gain.
- Message speaker:
  - Adams: tries, no gain.
  - Key for Aoki, Camtepe, Acar.
- Semantics:
  - Adams: tries, no gain.

One Conversation Per Speaker?

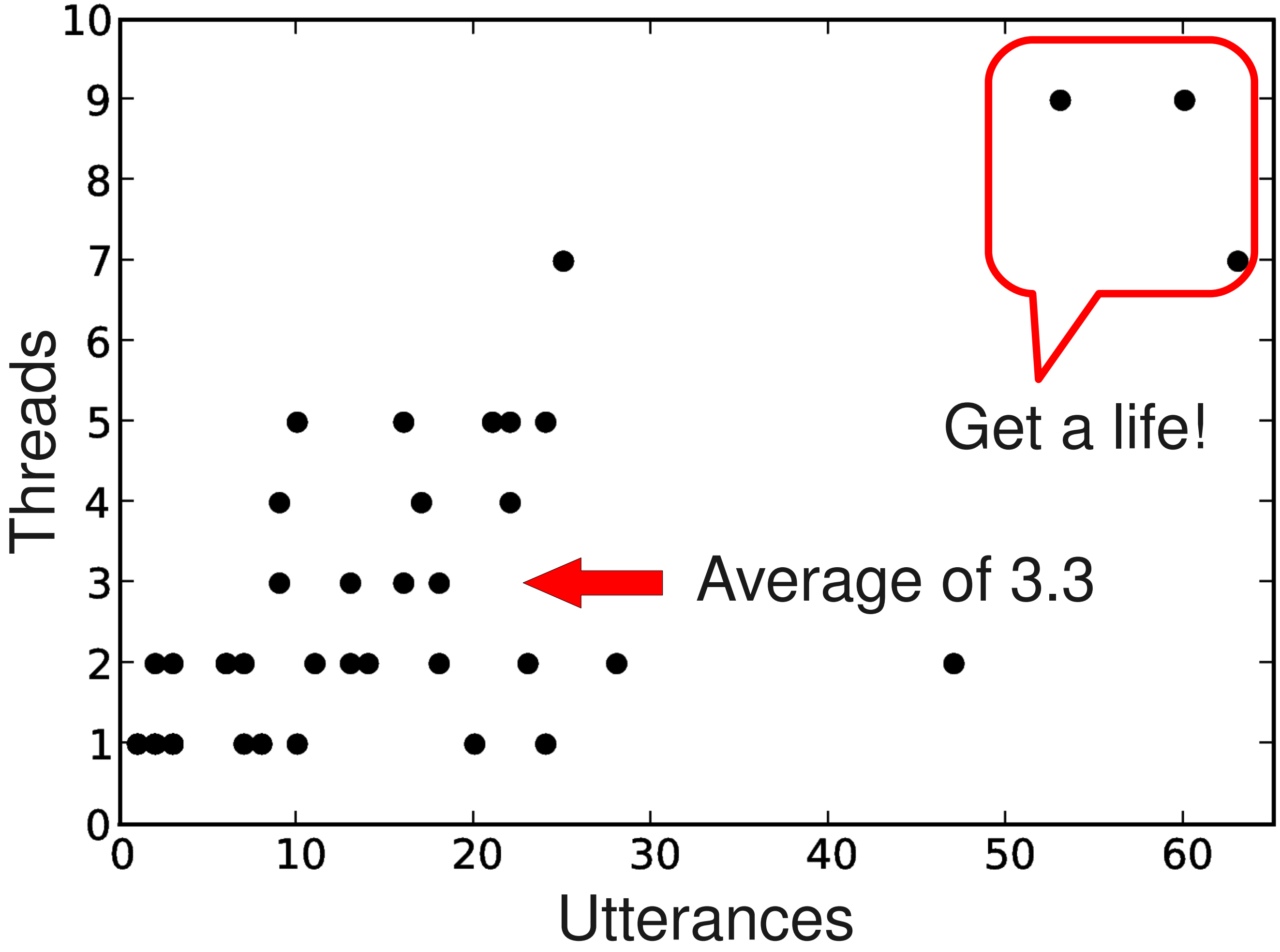
- Assumed by Camtepe, Acar:
  - Trying to detect social groups
- Aoki:
  - In 30-second window
  - Computational simplicity
- Legitimate assumption? No!

Conversations Per Speaker



Average of 3.3

Conversations Per Speaker



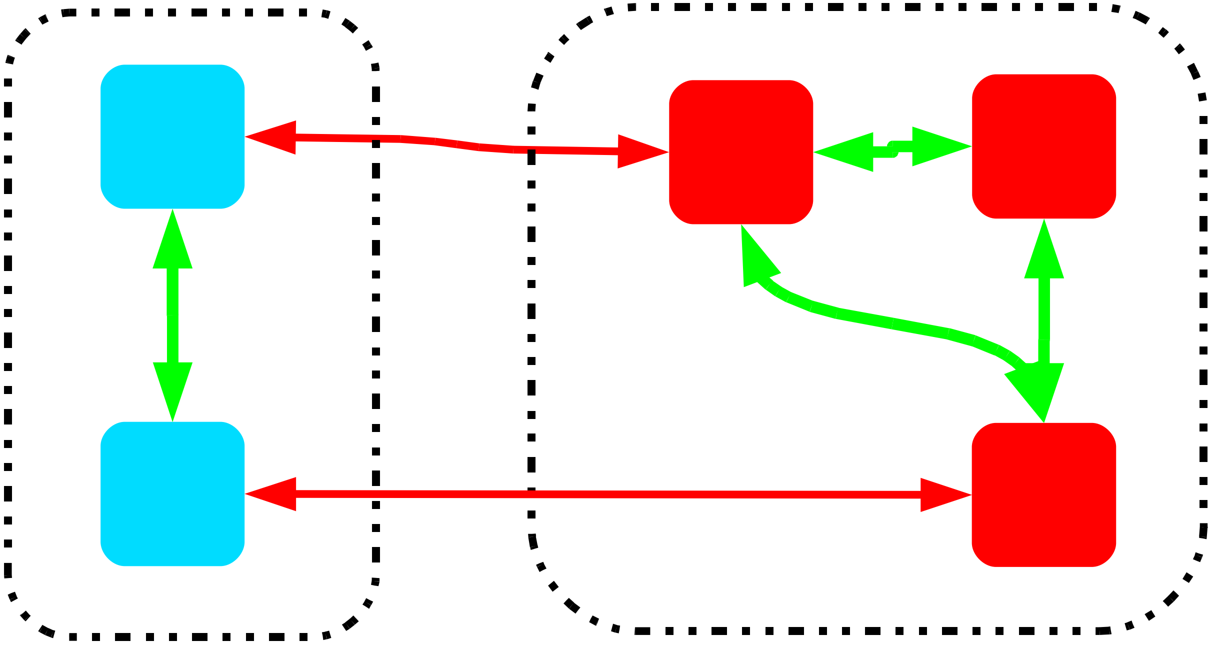
Outline

- Corpus
  - Annotations
  - Metrics
  - Agreement
  - Discussion
  - Features
- Modeling
  - Previous Work
  - **Classifier**
  - **Inference**
  - Baselines
  - Results
- Extensions
  - Specificity Tuning
  - Conversation Start Detection

Questions are welcome!

Our Method: Classify and Cut

- Common NLP method: Roth and Yih ‘04.
- Links based on max-ent classifier.
- Greedy cut algorithm.
  - Found optimal too difficult to compute.



Comparison

- Supervised method.
- Pros:
  - Easy feature combination.
  - All parameters tuned from training data.
- Cons:
  - Needs annotated data.
  - Less portable across corpora?

Classifier

- Pair of utterances: same conversation or different?

- Chat-based features (F 66%):
  - Time between utterances
  - Same speaker
  - Name mentions

- Most effective feature set.

Classifier

- Pair of utterances: same conversation or different?

- Chat-based features (F 66%)
- Discourse-based (F 58%):
  - Detect questions, answers, greetings &c
- Lexical (F 56%):
  - Repeated words
  - Technical terms

Classifier

- Pair of utterances: same conversation or different?

- Chat-based features (F 66%)
- Discourse-based (F 58%)
- Lexical (F 56%)

- Combined (F 71%)

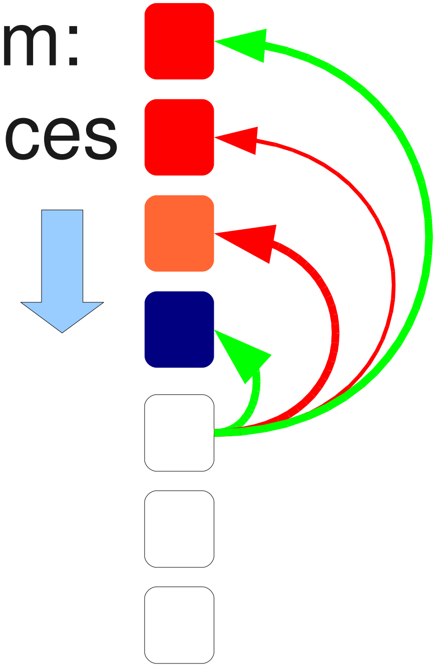
Technical Terms

- Tech support vs. idle chat:
  - Rarely in the same thread
- Detect “tech” keywords using a Linux manual.
- A light-weight semantic technique.
- Slight improvements.

- Open question: some way to use WordNet or LSA?

Inference

Greedy algorithm: process utterances in sequence

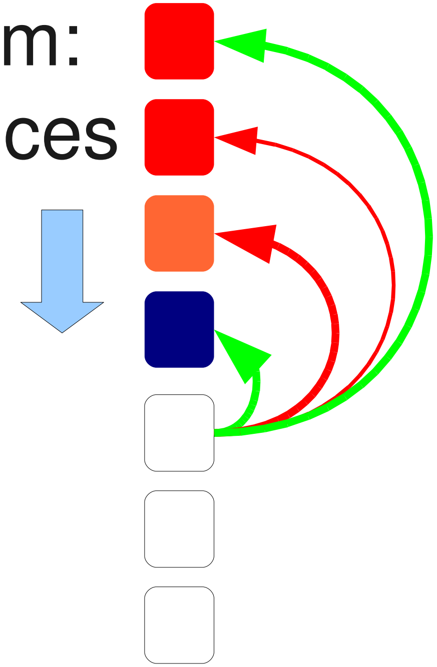


Classifier marks each pair "same" or "different" (with confidence scores).

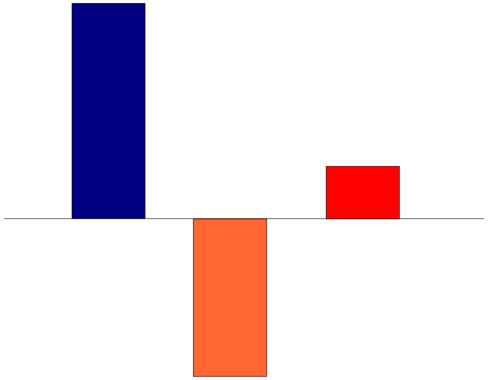
Pro: online inference
Con: not optimal

Inference

Greedy algorithm: process utterances in sequence



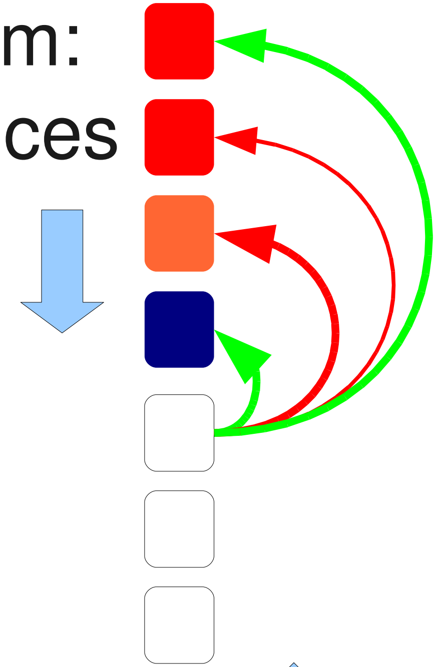
Treat classifier decisions as votes.



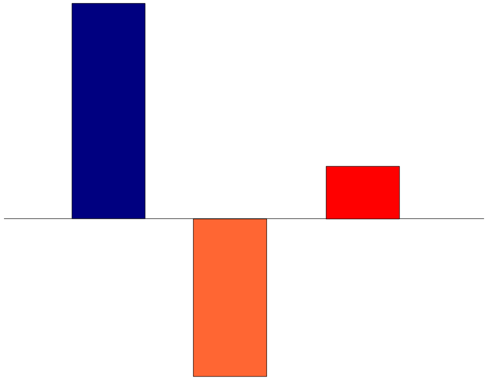
Pro: online inference
Con: not optimal

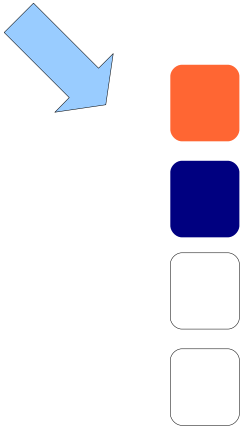
Inference

Greedy algorithm:
process utterances
in sequence



Treat classifier decisions
as votes.





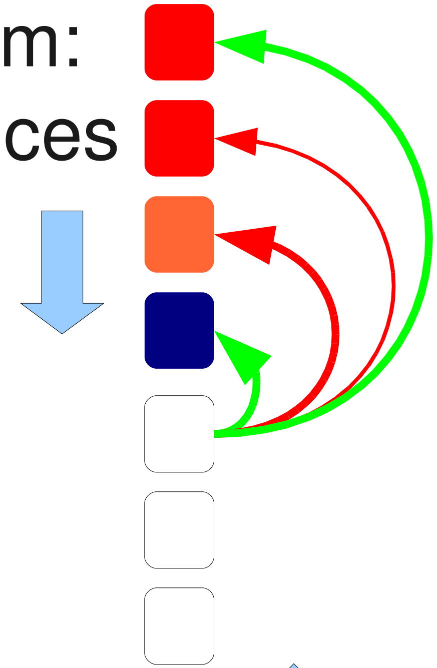
Color according to the
winning vote.

If no vote is positive,
begin a new thread.

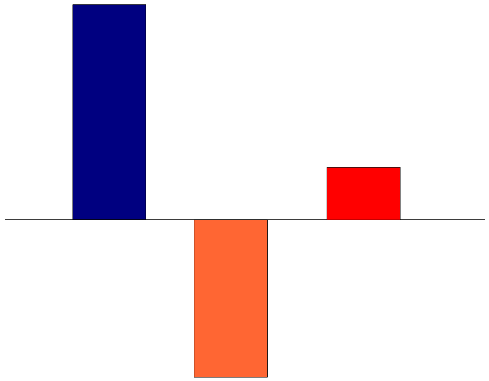
Pro: online inference
Con: not optimal

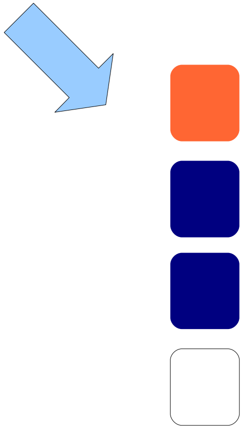
Inference

Greedy algorithm:
process utterances
in sequence



Treat classifier decisions as votes.





Color according to the winning vote.

If no vote is positive, begin a new thread.

Pro: online inference
Con: not optimal

Outline

- Corpus
  - Annotations
  - Metrics
  - Agreement
  - Discussion
  - Features
- Modeling
  - Previous Work
  - Classifier
  - Inference
  - **Baselines**
  - **Results**
- Extensions
  - Specificity Tuning
  - Conversation Start Detection

Questions are welcome!

Baseline Annotations

- All in same conversation
- All in different conversations
- Speaker's utterances are a monologue

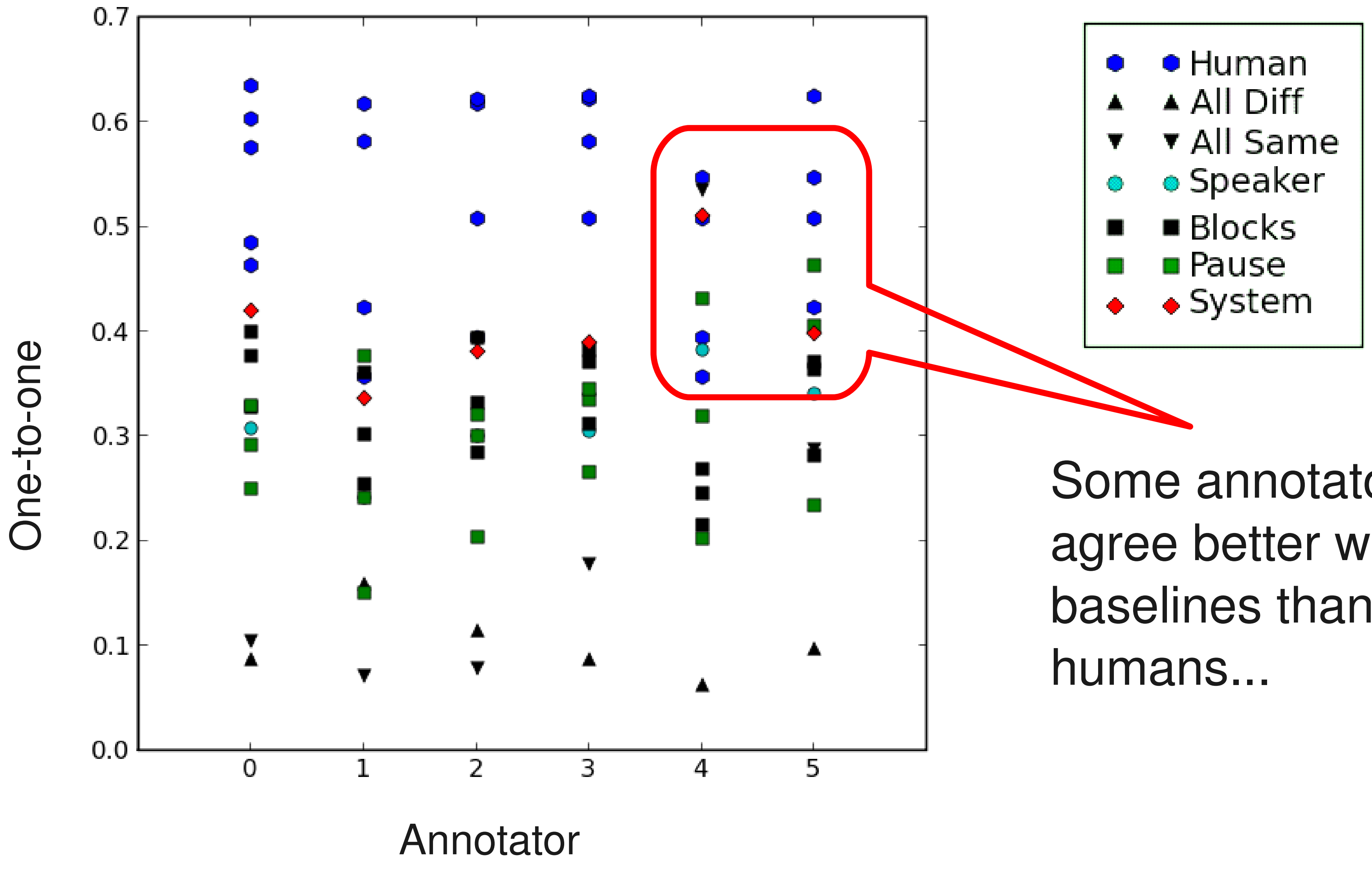
- Consecutive blocks of *k*
- Break at each pause of *k*
  - Upper-bound performance by optimizing *k* on the test data.

Results

	Humans	Model	Best Baseline	All Diff	All Same
Max 1-to-1	64	51	56 (Pause 65)	16	54
Mean 1-to-1	**53**	**41**	**35 (Blocks 40)**	**10**	**21**
Min 1-to-1	36	34	29 (Pause 25)	6	7

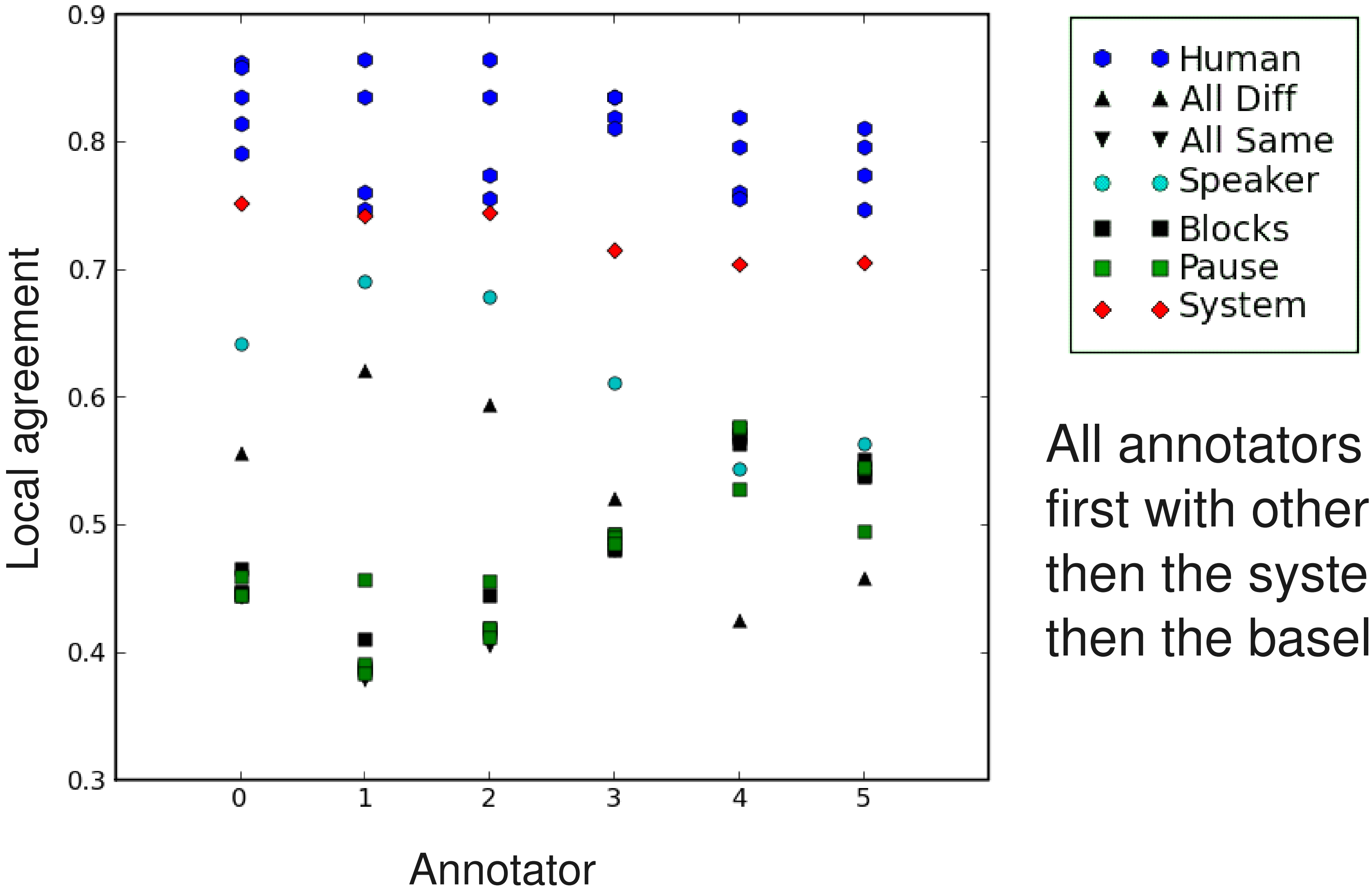
	Humans	Model	Best Baseline	All Diff	All Same
Max local	87	75	69 (Speaker)	62	57
Mean local	**81**	**73**	**62 (Speaker)**	**53**	**47**
Min local	75	70	54 (Speaker)	43	38

One-to-One Overlap Plot



Some annotators agree better with baselines than other humans...

Local Agreement Plot



All annotators agree first with other humans, then the system, then the baselines.

Mention Feature

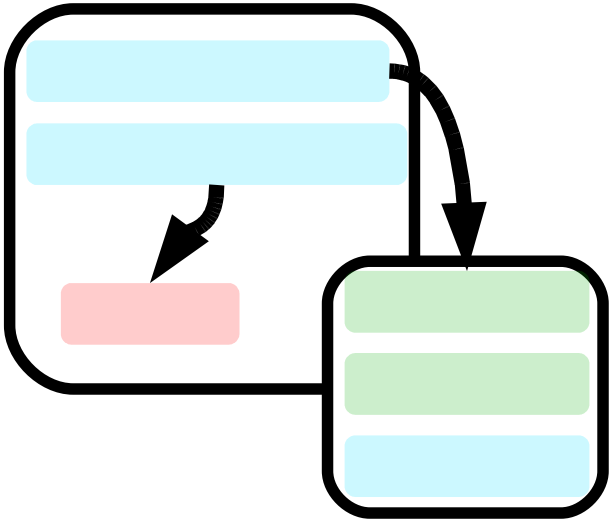
- Name mention features are critical.
  - When they are removed, system performance drops to baseline.
- But not sufficient.
  - With only name mention and time gap features, performance is midway between baseline and full system.

Outline

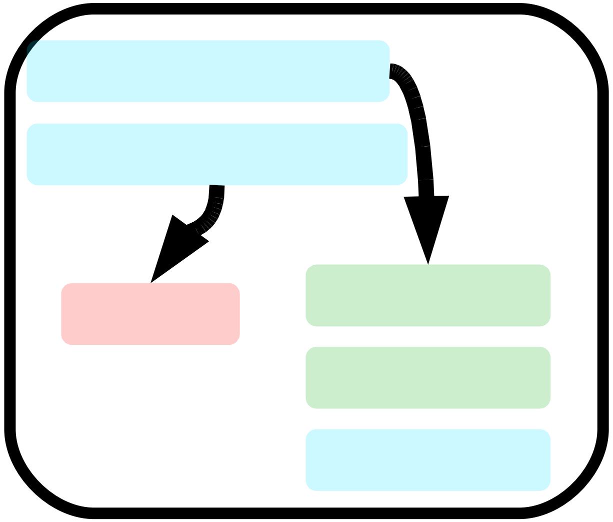
- Corpus
  - Annotations
  - Metrics
  - Agreement
  - Discussion
  - Features
- Modeling
  - Previous Work
  - Classifier
  - Inference
  - Baselines
  - Results
- Extensions
  - **Specificity Tuning**
  - Conversation Start Detection

Questions are welcome!

Coarser/Finer Annotation on Demand



vs



- Annotators disagree about specificity
- Can we meet different demands
  without retraining?

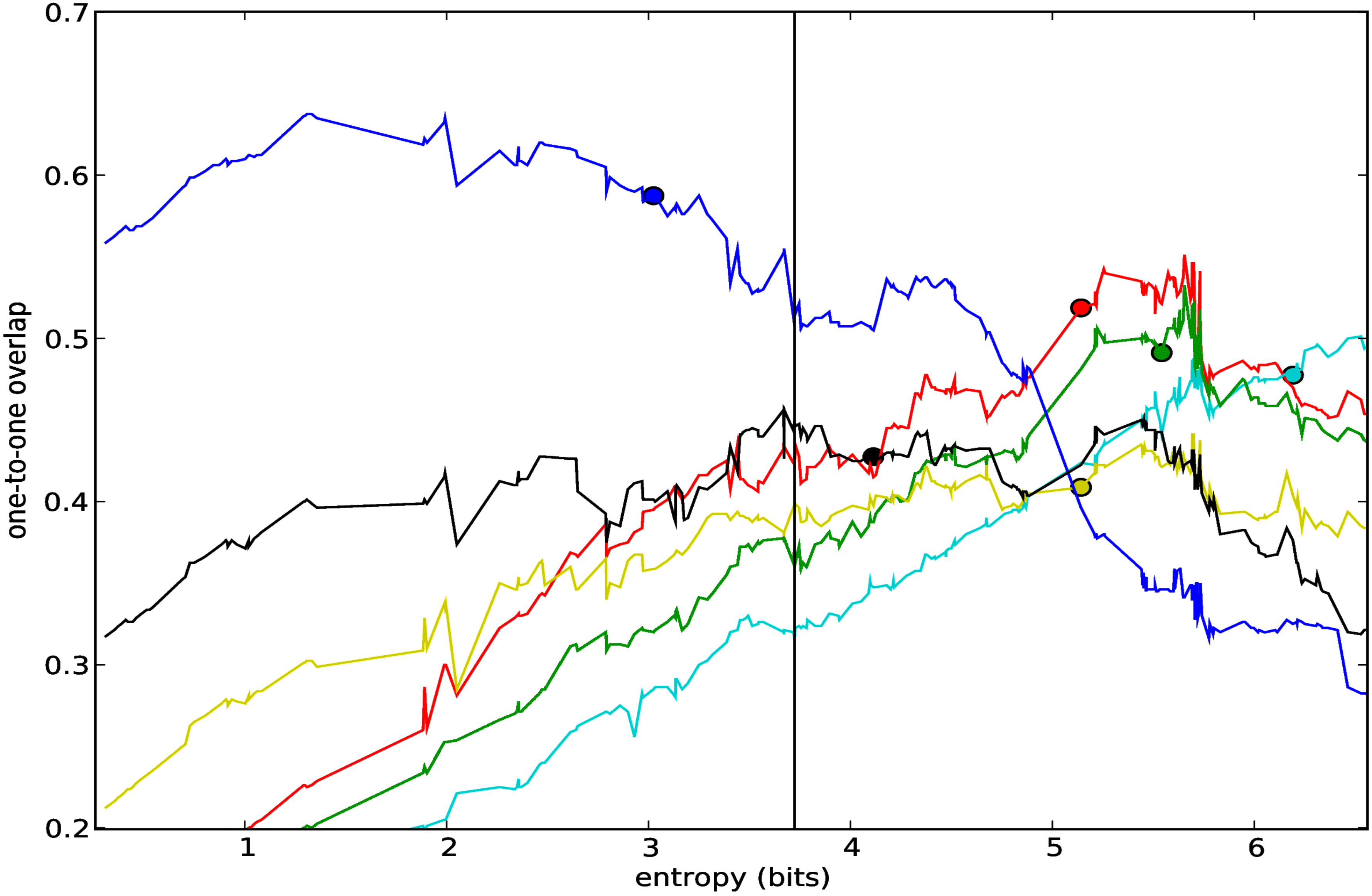
Bias Tuning

- Classifier:

$$\frac{1}{1 + \exp(- w \bullet x + \underbrace{\boxed{b}}_{\text{Bias}})}$$

- Assumption: know *exact* entropy annotator wants.
- Add or subtract from bias...
  until target entropy reached.

Results



Results

	Untuned	Tuned
Mean 1-to-1	41	49
Mean Loc3	73	73

- Specificity has little effect on local metric.
- Useful globally, but...
  - Assumption of exact entropy unrealistic.

- What *can* users tell us about what they want?

Outline

- Corpus
  - Annotations
  - Metrics
  - Agreement
  - Discussion
  - Features
- Modeling
  - Previous Work
  - Classifier
  - Inference
  - Baselines
  - Results
- Extensions
  - Specificity Tuning
  - **Conversation Start Detection**

Questions are welcome!

Where Conversations Start

- Current model:
  - Many pairwise decisions.
- Better?
  - One pointwise decision.
  - (like discourse-new classification in coref)
- Couldn't get much improvement...

Oracle Results

- If we *had* perfect detection:

	Normal	Oracle
Mean 1-to-1	41	47
Mean Loc3	73	74

- How good is “normal”?
  - Not very!
  - F-score ~ 50%.

- Can we build a better detector?

Plenty of Work Left

- Annotation standards:
  - Schemes with better agreement
  - Explicitly model splits/merges?
  - No partitioning, just link utterances? (Traum pc.)
- What metrics can we use for these schemes:
  - Graphs, not just clusterings.
- How can users express their preferences?

Plenty of Work Left

- Modeling:
  - Better classification/distance metrics.
  - Semi-supervised methods?
  - Conversation start detection.
  - Semantics.
- Applied settings:
  - Which metrics correlate with IR scores?
- Other domains? Speech?

Data and Software is Free

- Available at:

  www.cs.brown.edu/~melsner

- Dataset (text files)
- Annotation program (Java)
- Analysis and Model (Python)

Acknowledgements

- Suman Karumuri and Steve Sloman
- Matt Lease
- David McClosky
- Craig Martell
- David Traum
- 7 test and 3 pilot annotators
- 3 anonymous reviewers
- NSF PIRE grant