Exploiting Conversation Structure in Unsupervised Topic Segmentation for Emails

Shafiq Joty, Giuseppe Carenini, Gabriel Murray, Raymond Ng

University of British Columbia
Vancouver, Canada

“Topic” Segmentation

- “Topic” is something about which the participants of a conversation discuss or argue.
- Email thread about arranging a conference can have topics:
  - ‘location and time’,
  - ‘registration’,
  - ‘food menu’,
  - ‘workshops’
- Topic assignment: Clustering the sentences of an email thread into a set of coherent topical clusters.

Example

From: Charles ***To:*** WAI AU Guidelines ***Date:*** Thu May ***Subj:*** Phone connection to ftof meeting.

It is probable that we can arrange a telephone connection, to call in via a US bridge.

<Topic id = 1>

Are there people who are unable to make the face to face meeting, but would like us to have this facility?

<Topic id = 1>

From: William ***To:*** Charles ***Date:***Thu May ***Subj:*** Re: Phone connection to ftof meeting.

- Are there people who are unable to make the face to face meeting, but would like us to have this facility?

At least one "people" would.

<Topic id = 1>

……..

From: Charles ***To:*** WAI AU Guidelines ***Date:*** Mon Jun *Subj:* RE: Phone connection to ftof meeting.

Please note the time zone difference, and if you intend to only be there for part of the time let us know which part of the time.

<Topic id = 2>

9am - 5pm Amsterdam time is 3am - 11am US Eastern time which is midnight to 8am pacific time.

<Topic id = 2>

Until now we have got 12 people who want to have a ptop connection.

<Topic id = 1>

Motivation

- Our main research goal (on asynchronous conversation):
  - *Information extraction*
  - *Summarization*

- Topic segmentation is often considered a prerequisite for other higher-level conversation analysis.

- Applications:
  - Text summarization,
  - Information ordering,
  - Automatic QA,
  - Information extraction and retrieval,
  - Intelligent user interfaces.

Challenges

- Emails are different from written monologue and dialog:
  - Asynchronous and distributed.
  - Informal.
  - Different styles of writing.
  - Short sentences.
- Same topic can reappear.
- Relying on headers are often inadequate.
- No reliable annotation scheme, no standard corpus, and no agreed upon metrics available.

Example of Challenges

……

From: William *To:* Charles *Date:* Thu May *Subj:* Re: Phone connection to ftof meeting.

- Are there people who are unable to [illegible] to face meeting, but would like us to have this facility?

Short and informal

At least one "people" would. <Topic id = 1>

……

Header is misleading

From: Charles *To:* WAI AU Guidelines *Date:* Mon Jun *Subj:* RE: Phone connection to ftof meeting.

Please note the time zone difference, and if you intend to only be there for part of the time let us know which part of the time. <Topic id = 2>

9am - 5pm Amsterdam time is 3am - 11am US Eastern time w[illegible] to 8am pacific time. <Topic id = 2>

Topics reappear

Until now we have got 12 people who want to have a ptop connection <Topic id = 1>

Contributions: Outline of the Rest of the Talk

- Corpus:
  - Dataset
  - Annotations
  - Metrics
  - Agreement

- Segmentation Models
  - Existing Models
    - LCSeg
    - LDA
  - Extensions
    - LCSeg+FQG
    - LDA+FQG
  - Evaluation
  - Future work

Dataset

- BC3 email corpus
  - 40 email threads from W3C corpus.
  - 3222 sentences.
  - On average five emails per thread.
  - Previously annotated with:
    - Speech acts and meta sentences,
    - Subjectivity,
    - Extractive and abstractive summaries.
  - New topic annotations will be made publicly available: http://www.cs.ubc.ca/labs/lci/bc3.html

Topic Annotation Process

- Two phase pilot study:
  - Five randomly picked email threads.
  - Five UBC graduate students in the first phase.
  - One postdoc in the second phase.
- Actual topic annotation:
  - Three 4th year undergraduates (CS major and native speaker).
- Participants were also given a human written summary.

Annotation Tasks

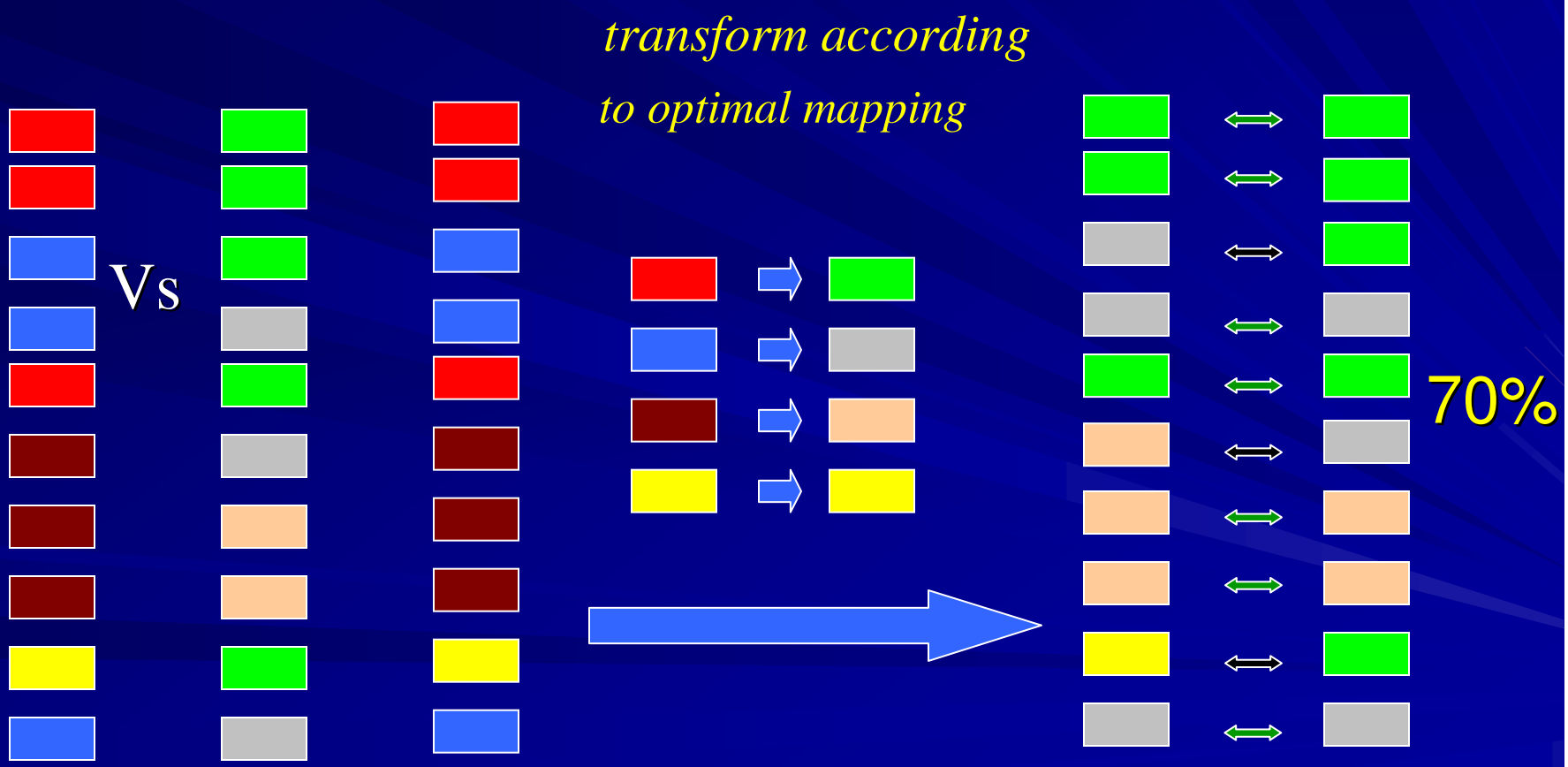
- First task:
  - Read an email thread and a human written summary.
  - List the topics discussed.
  - Example:
    - <Topic id 1, “location and time of the ftof mtg.”>
    - <Topic id 2, “phone connection to the mtg.”>
- Second task:
  - Annotate each sentence with the most appropriate topic (id).
  - Multiple topics were allowed.
  - Predefined topics: OFF-TOPIC, INTRO, END
  - 100% agreement on the predefined topics.

Agreement/Evaluation Metrics

- Number of topics varies across annotations.
  - "Kappa" not applicable.
- Segmentation in conversation not sequential.
  - "WindowDiff (WD)" and "P_k" also not applicable.
- More appropriate metrics (Elsner and Charniak, ACL-08):
  - One-to-One.
  - Loc_k.
  - M-to-One.

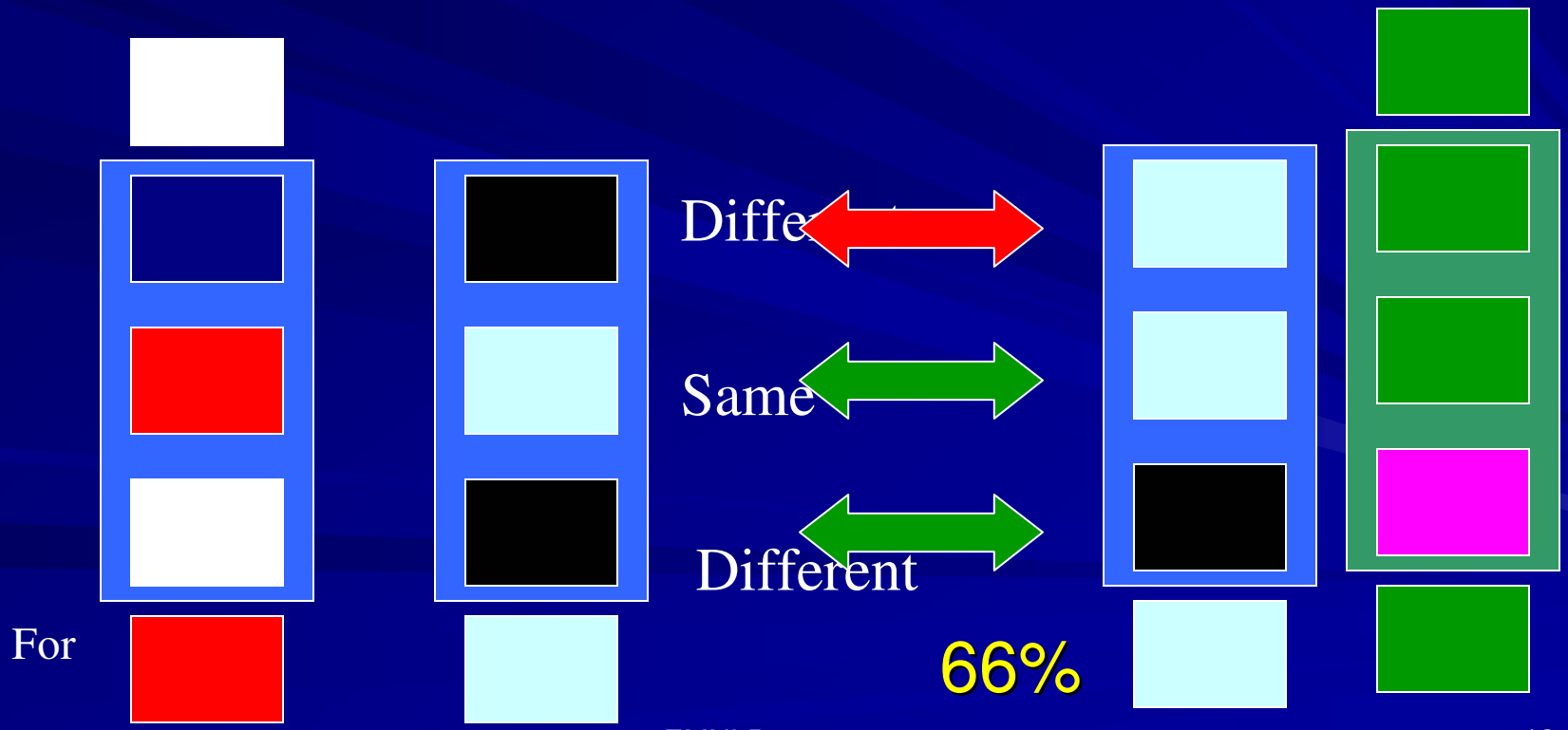
Metrics (1-to-1)

1-to-1 measures the global similarity by pairing up the clusters of 2 annotations to maximize the total overlap.



Metrics (loc_k)

loc_k measures the local agreement between two annotations within a context of k sentences.



Inter-annotator Agreement

	Mean	Max	Min
1-to-1	**0.804**	1	0.31
loc_3	**0.831**	1	0.43

- Agreements are pretty good!
- How annotators disagree:
  - Some are much finer-grained than others.

	Mean	Max	Min
# of Topics	2.5	7	1
Entropy	0.94	2.7	0

  - M-to-1 gives an intuition of annotator's specificity.

Metrics (M-to-1)

- M-to-1 maps each of the clusters of the 1st annotation to the single cluster in the 2nd annotation with which it has the greatest overlap, then computes the percentage of overlap.

	Mean	Max	Min
M-to-1	0.949	1	0.61

- To compare models we should use 1-to-1 and loc_k.

Outline of the Rest of the Talk

- Corpus:
  - Dataset
  - Annotations
  - Metrics
  - Agreement

- Segmentation Models
  - Existing Models
    - LCSeg
    - LDA
  - Extensions
    - LCSeg+FQG
    - LDA+FQG
  - Evaluation
  - Future work

Related Work: Existing Segmentation Models

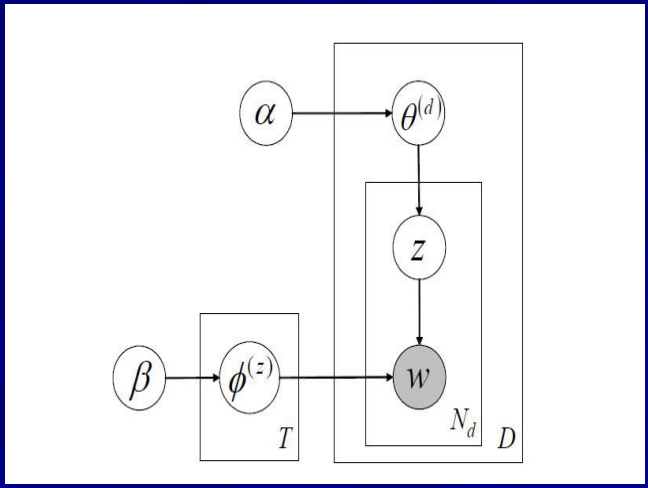
- Segmentation in monolog and sync. dialog:
  - Supervised: Binary classification with features.
  - Unsupervised:
    - LCSeg (Galley et al., ACL'03).
    - LDA (Georgescul et al., ACL'08).

- Multi-party chat (Conversation disentanglement):
  - Graph-based clustering (Elsner and Charniak, ACL'08).

- Asynchronous conversations (emails, blogs):
  - To our knowledge no work.

LDA on Email Corpus

- **Latent Dirichlet Allocation (Blei et al., 03):**



  - Generative model.
  - Generation process:
    - Choose a topic.
    - Choose a word.
  - Each email is a document.
  - Inference gives distributions of words over the topics.
  - Assuming the words in a sentence occur independently, we compute distributions of sentences over topics.
  - Assign topic by taking argmax over the topics.

LCSeg of Email Corpus

- **Lexical Chain Segmenter (Galley et al., 03):**
  - Order the emails based on their **temporal relation**.
  - Compute “lexical chains” based on word repetition.
  - Rank the chains according to two measures:
    - Number of repetition.
    - Compactness of the chain.
  - Score of the words in a chain is same as the rank of the chain.
  - Measure similarity between two consecutive windows of sentences.
  - Assign a boundary if the measure falls below a threshold.

Limitations of the Two Models

- Both LDA and LCSeg make BOW assumptions.
- Ignore important conversation features:
  - Reply-to relation.
  - Usage of quotations.
- In our corpus people use quotations to talk about the same topic.

- Example:

  >Are there people who are unable to make the face to face meeting, but would like us to have this facility?

  At least one “people” would. <Topic id = 1>

- In BC3, usage of quotations per thread is: 6.44.

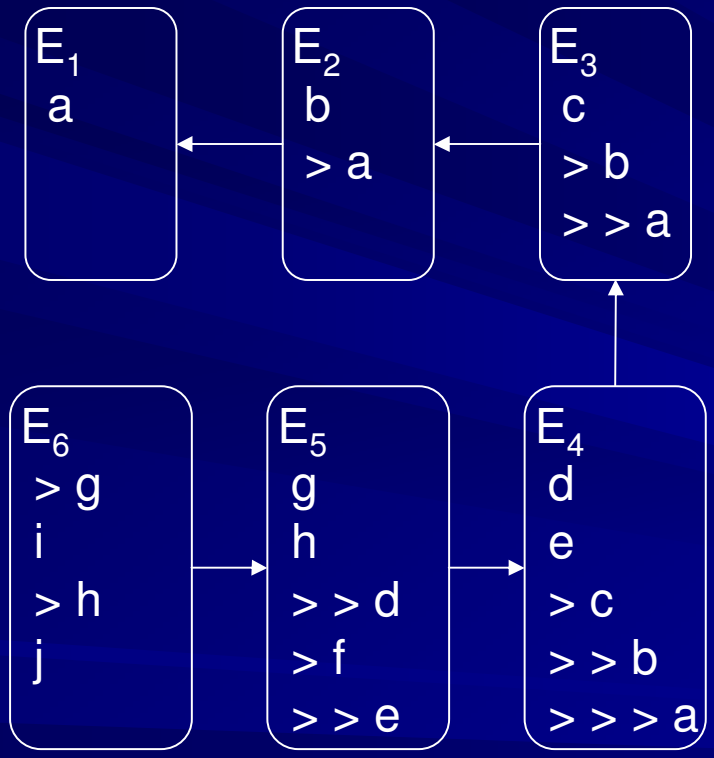
What We Need

- We need to:
  - Capture the conversation structure at the quotation level.
  - Incorporate this structure into the models.

Extracting Conversation Structure (Carenini et al., ACL'08)

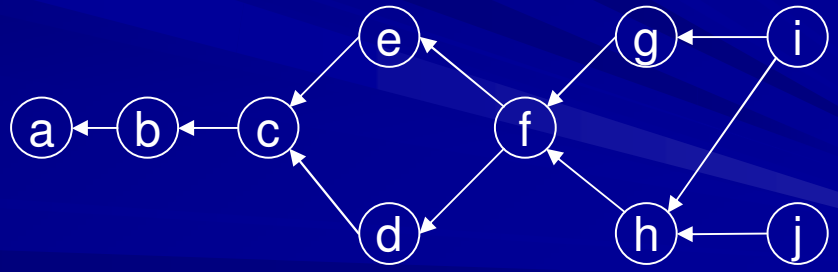
- We analyze the actual body of the emails.
- We find two kinds of fragments:
  - New fragment (depth level 0)
  - Quoted fragment (depth level > 0)
- We form a fragment quotation graph (FQG):
  - Nodes represent fragments.
  - Edges represent referential relations.

Fragment Quotation Graph



An email conversation with 6 emails.

- Nodes
  - Identify quoted and new fragments
- Edges
  - Neighbouring quotations

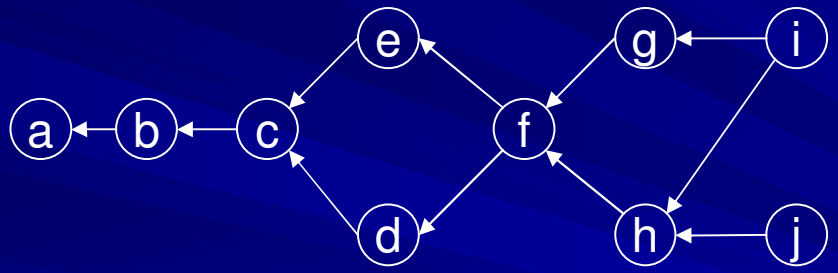


LDA with FQG

- Our primary goal is to regularize LDA so that sentences in nearby fragments fall in the same topical cluster.
- Regularize the topic-word distr. with a word network.
- Standard Dirichlet prior doesn't allow this.
- Andrzejewski et al., (2009) describes how to encode domain knowledge using Dirichlet Forest prior.
- We re-implemented this model (only "must link").
- We construct word network by connecting words in the same or adjacent fragments.

LCSeg with FQG

- Extract the paths (sub-conversations) of FQG.
- On each path run LCSeg.



- Sentences in common fragments fall in multiple segments.

LCSeg with FQG (Cont..)

- Consolidate different segments:
  - Form graph where
    - Nodes represent sentences.
    - Edge weight w(u,v) represents the number of cases, sentences u and v fall in the same segment.
  - Find optimal clusters using normalized cut criteria (Shi & Malik, 2000).

Outline of the Rest of the Talk

- Corpus:
  - Dataset
  - Annotations
  - Metrics
  - Agreement

- Segmentation Models
  - Existing Models
    - LCSeg
    - LDA
  - Extensions
    - LCSeg+FQG
    - LDA+FQG
  - Evaluation
  - Future work

Evaluation

- **Baselines:**
  - All different: Each sentence a separate topic.
  - All same: Whole thread is a single topic.
  - Speaker: Sentences from each participant constitute a separate topic.
  - Blocks of k(= 5, 10, 15): Consecutive group of k sentences a separate topic.

- Speaker and Blocks of 5 are two strong baselines.

Results

Scores	Baselines		Systems				Human
	Speaker	Block 5	LDA	LDA+FQG	LCSeg	LCSeg+FQG	
Mean 1-1	0.52	0.38	0.57	0.62	0.62	0.68	**0.80**
Mean loc_3	0.64	0.57	0.54	0.61	0.72	0.71	**0.83**

- Our systems performs better than baselines but worse than humans.

Results

Scores	Baselines		Systems				Human
	Speaker	Block 5	LDA	LDA+FQG	LCSeg	LCSeg+FQG	
Mean 1-1	0.52	0.38	0.57	0.62	0.62	0.68	**0.80**
Mean loc_3	0.64	0.57	0.54	0.61	0.72	0.71	**0.83**

- LDA performs very disappointingly.
- FQG helps LDA.

Results

Scores	Baselines		Systems				Human
	Speaker	Block 5	LDA	LDA+FQG	LCSeg	LCSeg+FQG	
Mean 1-1	0.52	0.38	0.57	0.62	0.62	0.68	**0.80**
Mean loc_3	0.64	0.57	0.54	0.61	0.72	0.71	**0.83**

- LCSeg is a better model than LDA.

Results

Scores	Baselines		Systems				Human
	Speaker	Block 5	LDA	LDA+FQG	LCSeg	**LCSeg+FQG**	
Mean 1-1	0.52	0.38	0.57	0.62	0.62	**0.68**	**0.80**
Mean loc_3	0.64	0.57	0.54	0.61	0.72	**0.71**	**0.83**

- FQG helps LCSeg in 1-1 metric.
- Loc3 suffers a bit but not significantly.
- LCSeg+FQG is the best model.

Future Work

- Consider other important features:
  - Speaker.
  - Mention of names.
  - Subject of the email.
  - Topic shift cue words.
- Transfer our approach to other similar domains
  - Synchronous domains (chats, meetings).
  - Asynchronous domains (blogs).

Questions?

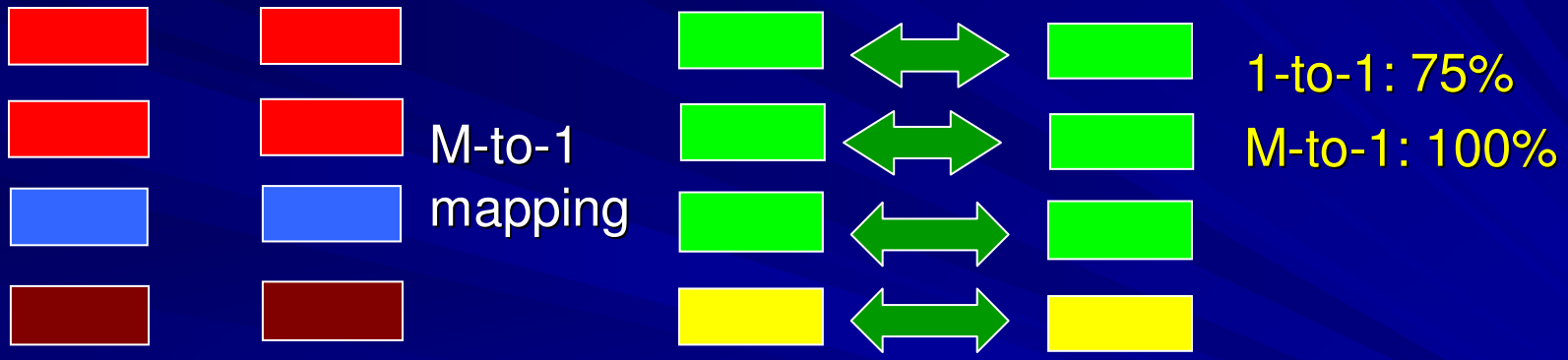
Thanks

Acknowledgements

- 6 pilot annotators.
- 3 test annotators.
- 3 anonymous reviewers.
- NSERC PGS award.
- NSERC BIN project.
- NSERC discovery grant.
- ICICS at UBC.

Metrics (M-to-1)

- M-to-1 maps each of the clusters of the 1st annotation to the single cluster in the 2nd annotation with which it has the greatest overlap, then computes the percentage of overlap.



	Mean	Max	Min
M-to-1	0.949	1	0.61

- To compare models we should use 1-to-1 and loc_k.

Results

Scores	Baselines			Systems			Human
	Speaker	Block 5	LDA	LDA+FQG	LCSeg	LCSeg+FQG	
Mean 1-1	0.52	0.38	0.57	**0.62**	0.62	**0.68**	**0.80**
Max 1-1	0.94	0.77	1.00	1.00	1.00	1.00	**1.00**
Min 1-1	0.23	0.14	0.24	0.24	0.33	0.33	**0.31**
Mean loc_k	0.64	0.57	0.54	**0.61**	**0.72**	**0.71**	**0.83**
Max loc_k	0.97	0.73	1.00	1.00	1.00	1.00	**1.00**
Min loc_k	0.27	0.42	0.38	0.38	0.40	0.40	**0.43**