Supervised Topic Segmentation of Email Conversations

Shafiq Joty, Giuseppe Carenini, Gabriel Murray, and Raymond Ng

Motivation

Email conversations often discuss multiple topics

e.g. a conversation about arranging **a conference** may cover:

Location and time, Registration, Food menu, Workshops

Two subtasks:

- **Segmentation:** Grouping sentences into coherent clusters
- **Identification:** Assigning topic labels to the clusters

Prerequisite for:

- Higher-level conversation analysis (e.g., speech act tagging).
- Text summarization and Automatic question answering.
- Intelligent user interfaces for emails.

Challenge

Topics in emails do not change in a sequential way

Models in monolog and synchronous dialog not so effective

Our Supervised Graph-theoretic Approach

(1) Sentence Pair Classification

(2) Graph Construction

(3) Graph Partitioning

- Integrates lexical and topic features with **conversational** ones.

Results

- Our sup approach achieves better accuracy than unsupervised method of [Joty et al. 2010] with very limited amount of training data.

Step1 Sentence Pair Classification

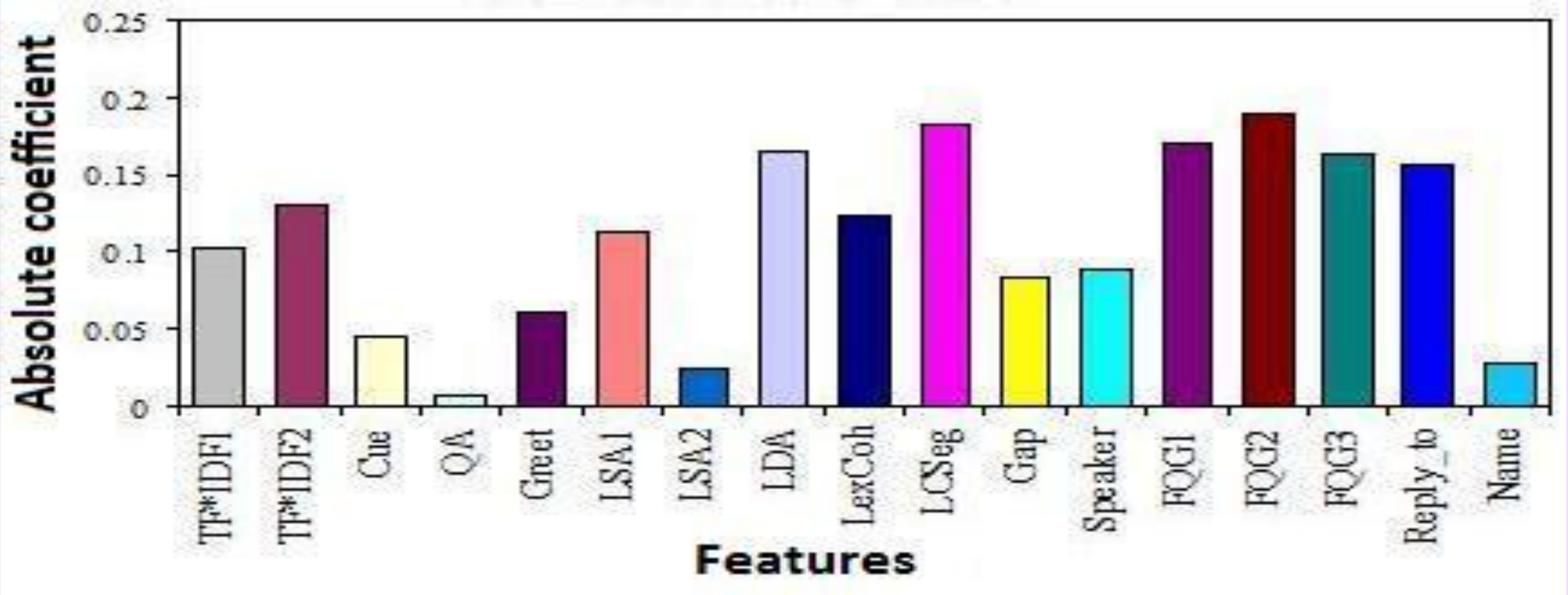
- A binary classifier marks each pair of sentences of a conversation as 'same' or 'different' topics.
- A conversation of n sentences produces O(n^2) training examples.
- **Comparison of classifiers:**

Classifier	Regularizer	Train error	Test error
KNN	-	47.7%	46.7%
SVM (lin)	-	33.2%	32.6%
SVM (rbf)	-	26.4%	34.3%
LR	l_2	30.6%	30.9%
LR	l_1	32.1%	33.3%
RMLR (rbf)	l_2	10.8%	38.9%

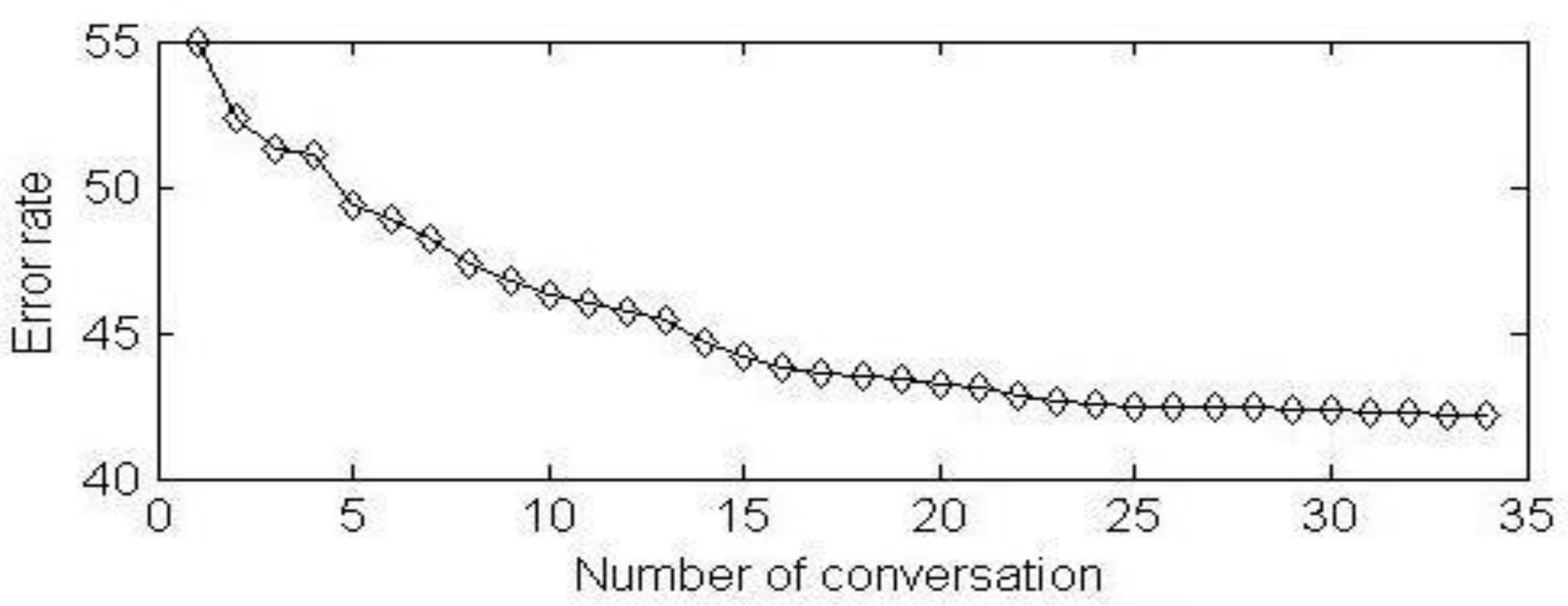
- **Features with average performance:**

Lexical	**Acc: 59.6 Pre: 59.7 Rec: 99.8**
TF.IDF1	TF.IDF similarity (k=1).
TF.IDF2	TF.IDF similarity (k=2).
Cue Words	Either one contains a cue word.
QA	x asks a question explicitly using '?' & y contains any of (yes, yea, ok, etc.)
Greet	Either one has a greeting word.
Topic	**Acc: 65.2 Pre: 64.4 Rec: 79.6**
LSA1	LSA function for x & y (k=1).
LSA2	LSA function for x & y (k=2).
LDA	LDA decision on x & y.
LCSeg	LCSeg decision on x & y.
LexCoh	Lexical cohesion function of x & y.
Conv	**Acc: 65.3 Pre: 66.7 Rec: 85.1**
Gap	The gap between y & x in # of sent.
Speaker	x & y have the same sender.
FQG1	Dist. between x & y in Dir. FQG (frag. Id).
FQG2	Dist. between x & y in Dir. FQG (#edges).
FQG3	Dist. between x & y in Undir. FQG (#edges)
Reply-to	Both are in the same email or one is a reply
Name	x mentions y or vice versa.
All	**Acc: 69.1 Pre: 68.4 Rec: 81.5**

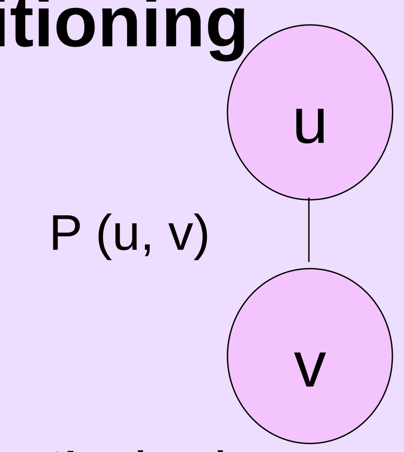
- **Relative importance of the features:**



- **Accuracy vs. amount of labeled data:**



Step2&3 Graph Construction and Partitioning



➢ Construct the graph:

- Nodes => Sentences
- Edge-weights => Probability ('same' class)

➢ Partition the graph by optimizing the *'normalized cut'* criterion.

Evaluation of our Sup. Topic Segmenter

Dataset: *BC3* email corpus. See [Joty et al. 2010] for corpus stats.

	Baseline		Models					Human
			Unsupervised				Super.	
Scores	Speaker	Block 5	LDA	LDA+FQG	LCSeg	LCSeg+FQG		
Mean 1-1	0.52	0.38	0.57	0.62	0.62	0.68	**0.70**	0.80
Mean loc_3	0.64	0.57	0.54	0.61	0.72	0.71	**0.75**	0.83

Reference Joty, S.; Carenini, G.; Murray, G.; Ng, R. Exploiting conversation structure in unsupervised topic segmentation for emails. In *EMNLP-2010*.