APPLICATIONS OF SENTIMENT ANALYSIS

NICK CHEN, MAX KAUFMANN, JEREMY MCLAIN

SUMMARIZING EMAILS WITH CONVERSATIONAL COHESION AND SUBJECTIVITY

GIUSEPPE CARENINI, RAYMOND T. NG AND XIAODONG ZHOU

WHAT …?

What is it?
What’s the problem?

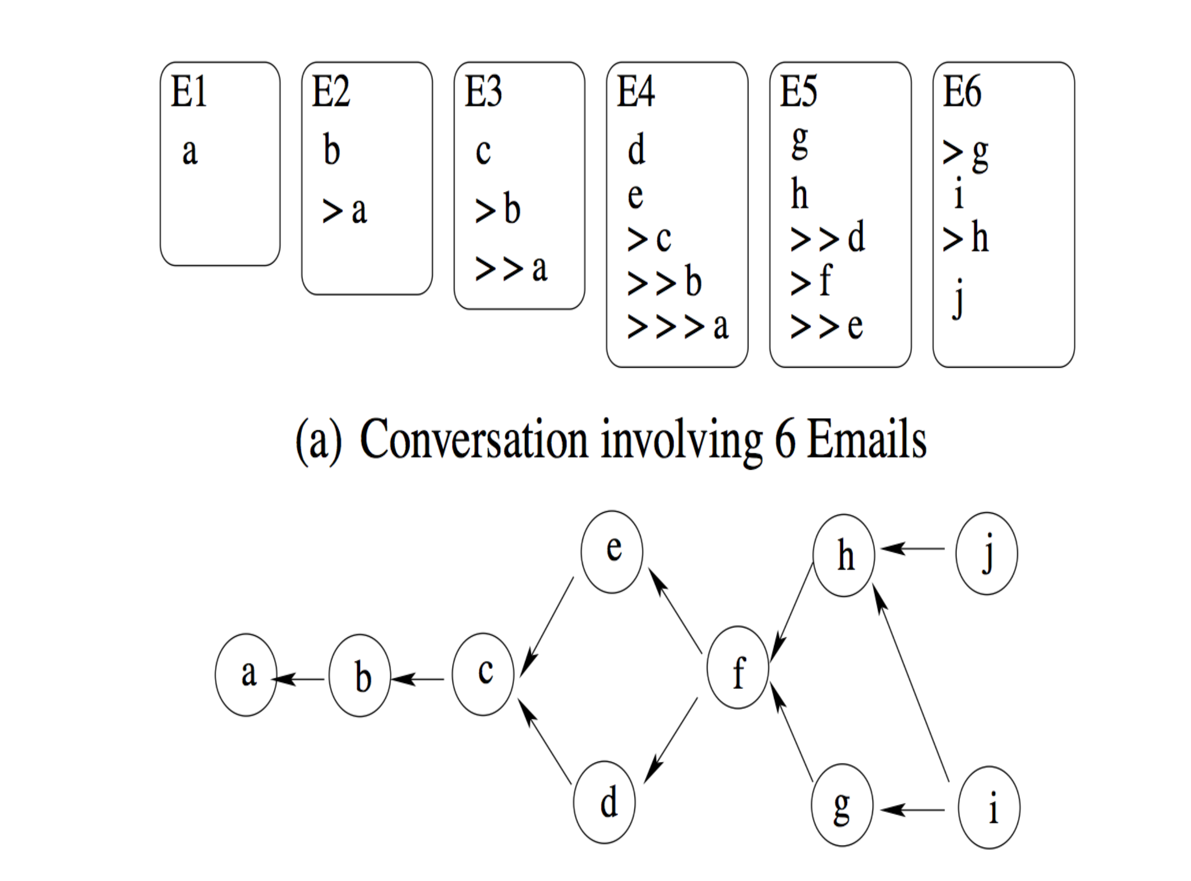
SUMMARIZING EMAILS WITH CONVERSATIONAL COHESION AND SUBJECTIVITY

Why emails?
What's the problem?
Data Set?
Setup?

APPROACH

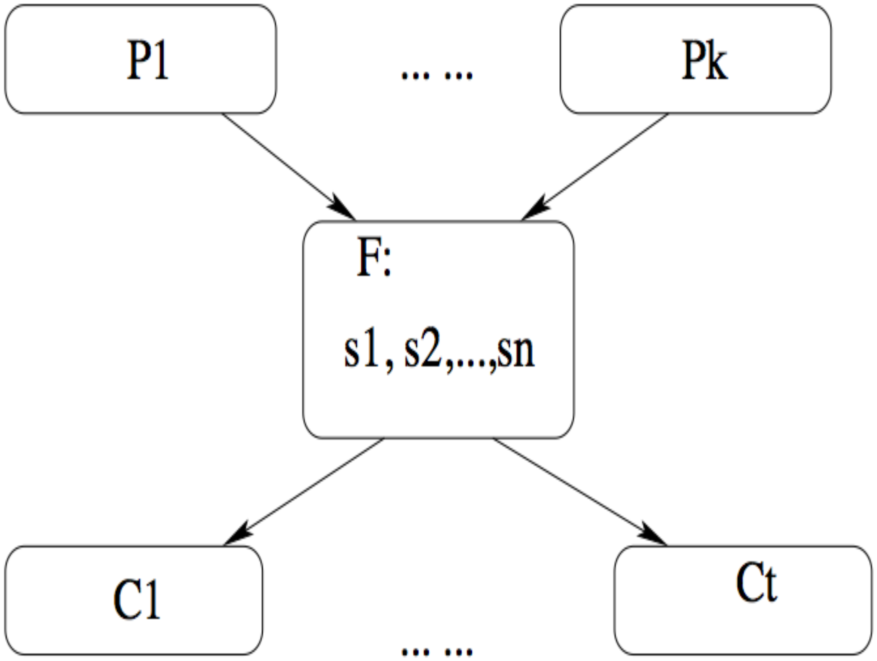
Sentence Quotation Graph
Sentence Relationships
Subjective Opinions

SENTENCE QUOTATION GRAPH



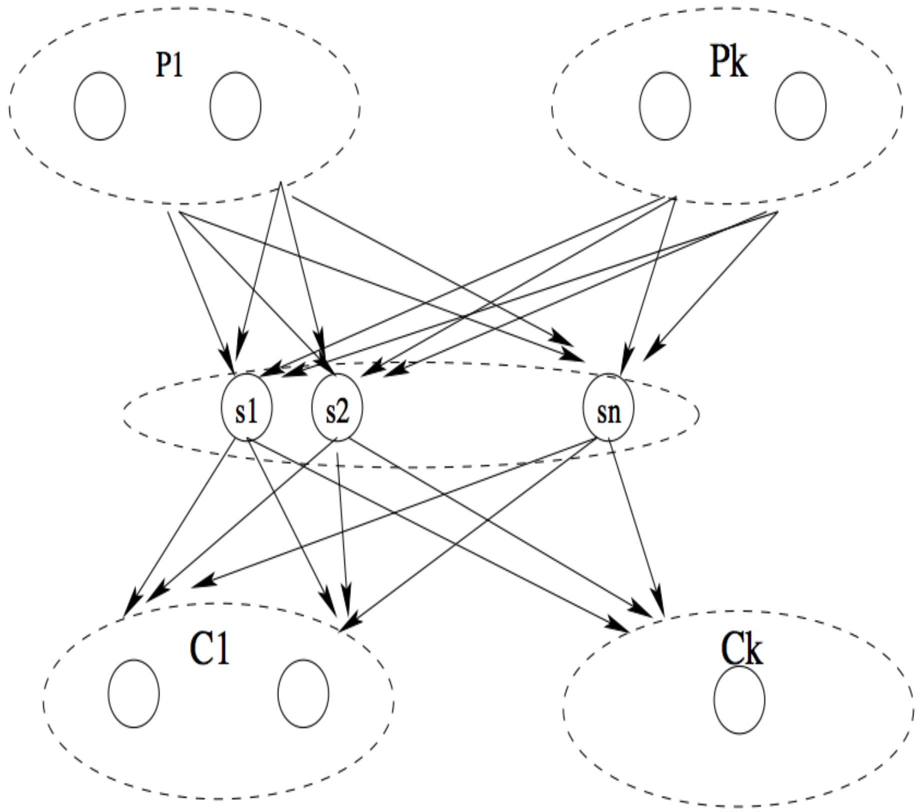
(a) Conversation involving 6 Emails

FRAGMENT QUOTATION GRAPH



(a) Fragment Quotation Graph

SENTENCE QUOTATION GRAPH



(b) Sentence Quotation Graph

SUMMARIZATION BASE ON SQG

ClueWordSummarizer algorithm

$$SentScore(s) = \sum_{(s,t)\in GS} weight(s,t) + \sum_{(p,s)\in GS} weight(p,s)$$

PageRank algorithm

$$PR(s) = (1-d) + d * \frac{\sum_{s_i \in child(s)} weight(s,s_i) * PR(s_i) + \sum_{s_j \in parent(s)} weight(s_j,s) * PR(s_j)}{\sum_{s_i \in child(s)} weight(s,s_i) + \sum_{s_j \in parent(s)} weight(s_j,s)}$$

SUBJECTIVE OPINION

Degree of subjectivity

RESULTS

Evaluation:

- Sentence Pyramid Precision
- ROGUE

	CWS	CWS-Cosine	CWS-lesk	CWS-jcn
Pyramid	0.6	0.39	0.57	0.57
p-value		<0.0001	0.02	0.005
ROUGE-2	0.46	0.31	0.39	0.35
p-value		<0.0001	<0.001	<0.001
ROUGE-L	0.54	0.43	0.49	0.45
p-value		<0.0001	<0.001	<0.001

CRITIQUE

Thoughts?

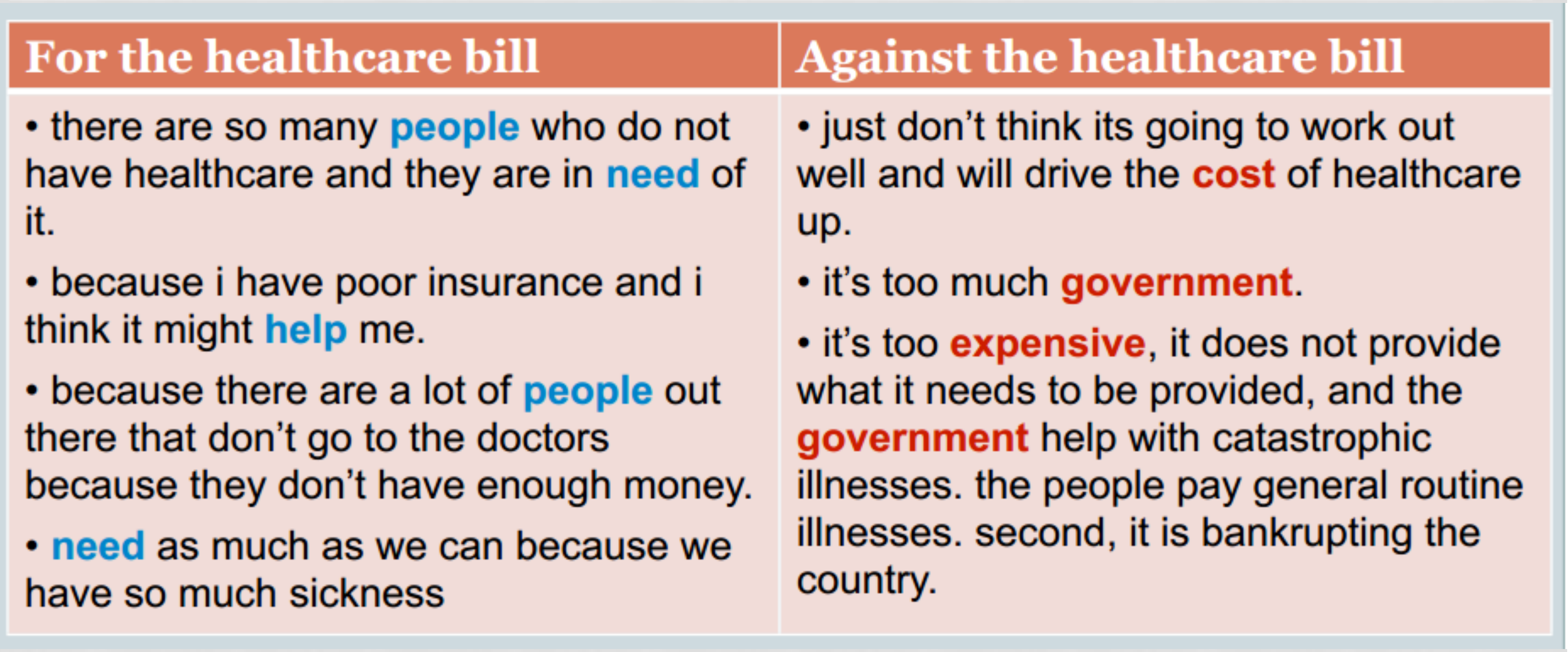
SUMMARIZING CONTRASTIVE VIEWPOINTS IN OPINIONATED TEXT

PAUL, MICHAEL AND ZHAI, CHENGXIANG AND GIRJU, ROXANA

SUMMARIZING CONTRASTIVE VIEWPOINTS IN OPINIONATED TEXT

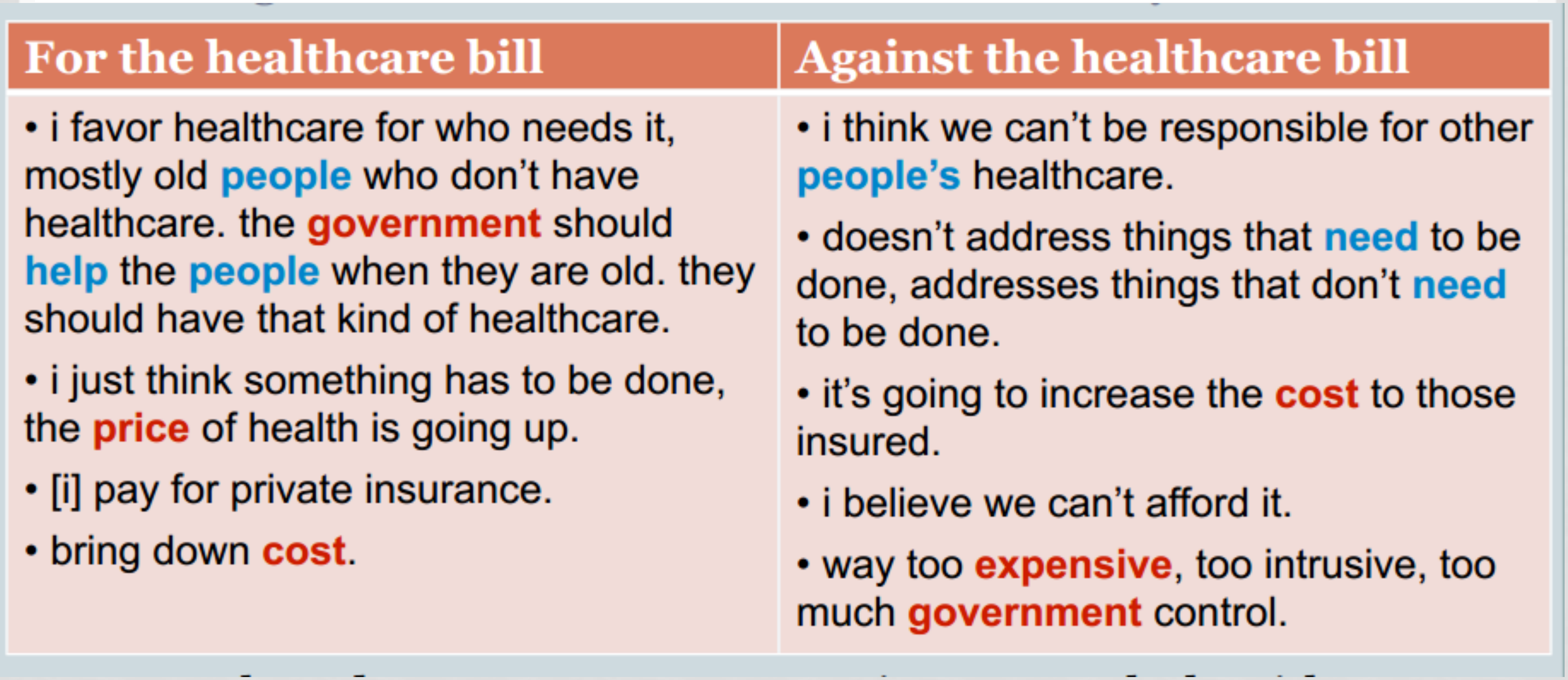
- Opinions in text are usually tied to a viewpoint
  - Sentiment + topic go together
- Task
  - Extract viewpoints from corpus
  - Summarize viewpoints

SUMMARIZATION



For the healthcare bill	Against the healthcare bill
• there are so many **people** who do not have healthcare and they are in **need** of it. • because i have poor insurance and i think it might **help** me. • because there are a lot of **people** out there that don't go to the doctors because they don't have enough money. • **need** as much as we can because we have so much sickness	• just don't think its going to work out well and will drive the **cost** of healthcare up. • it's too much **government**. • it's too **expensive**, it does not provide what it needs to be provided, and the **government** help with catastrophic illnesses. the people pay general routine illnesses. second, it is bankrupting the country.

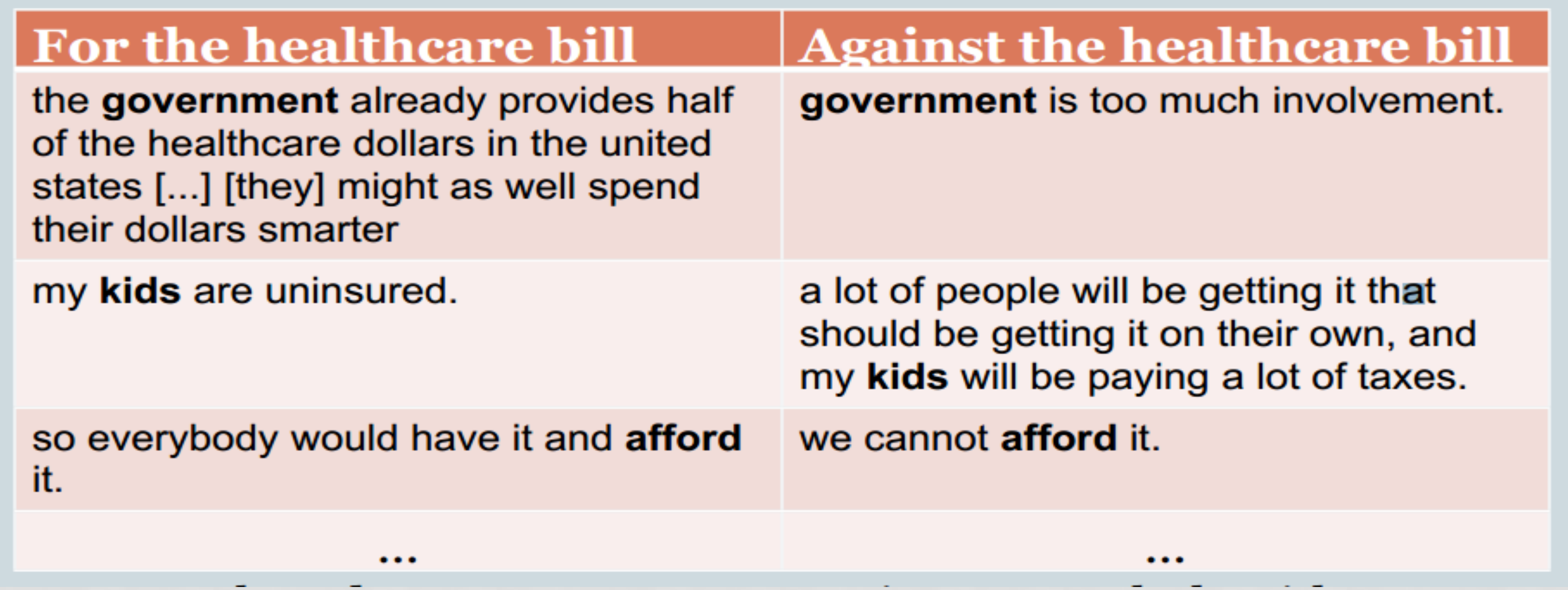
MACRO SUMMARIZATION



For the healthcare bill	Against the healthcare bill
• i favor healthcare for who needs it, mostly old **people** who don't have healthcare. the **government** should **help** the **people** when they are old. they should have that kind of healthcare. • i just think something has to be done, the **price** of health is going up. • [i] pay for private insurance. • bring down **cost**.	• i think we can't be responsible for other **people's** healthcare. • doesn't address things that **need** to be done, addresses things that don't **need** to be done. • it's going to increase the **cost** to those insured. • i believe we can't afford it. • way too **expensive**, too intrusive, too much **government** control.

- Multiple sentences summarizing one event
- Sentences are aligned to allow for easier contrast

MICRO SUMMARIZATION



For the healthcare bill	Against the healthcare bill
the **government** already provides half of the healthcare dollars in the united states [...] [they] might as well spend their dollars smarter	**government** is too much involvement.
my **kids** are uninsured.	a lot of people will be getting it that should be getting it on their own, and my **kids** will be paying a lot of taxes.
so everybody would have it and **afford** it.	we cannot **afford** it.
...	...

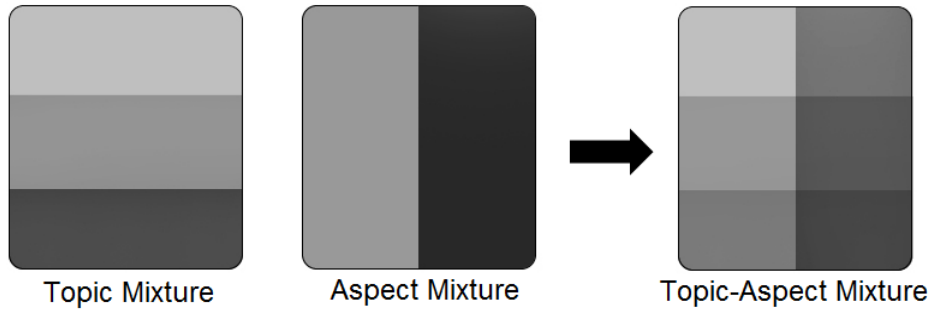
- Replace monolithic summary with sentence pairs (1 pro and 1 con)

PREVIOUS WORK

- Micro summaries have been done before
  - Based on the polarity of adjectives
- Macro summaries shave been done before
  - Modify LexRank to minimize the contrastiveness in 1 summary
- Nobody has attempted to do both at once
- Authors propose an integrated approach that does both

VIEWPOINT SUMMARIZATION

- Used Topic-Aspect Modeling



- Each document has
  - a multinomial topic mixture
  - a multinomial aspect mixture
- Words may depend on both!
- Run TAM with 2 topics to forcefully segregate text into viewpoints
- Supervised Training
  - Set P(Aspect | Document) = 1 if known that document is entirely one aspect

FEATURES

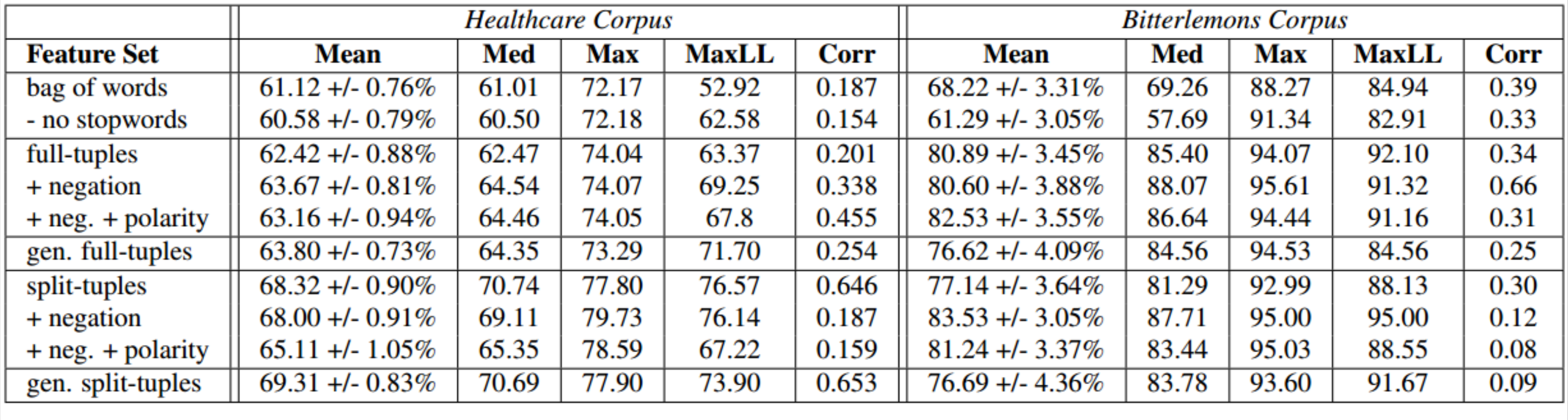
- Features are input to TAM
- Original TAM does not support any features

FEATURES

- Stanford dependency parses
  - ‘split-tuple’
    - rel(a,b) -> rel(a,*) and rel (*,b)
- Hiearchical dependencies
  - Dojb(a,b) -> obj(a,b)
  - Indrobj(a,b) -> obj(a,b)
- Polarity (from Wilson Subjectivity Clues lexicon)
  - Amod(idea,good)
    - Amod(idea,+) and amod (*,good)

RESULTS

- Clustered documents using results of Tam
  - Didn't say how they clustered!
  - Clustering accuracy only looked at documents where P(v | doc) > .8
  - Tinkering with TAM
    - Good: Gave parameters (reproducibility)
    - Bad: No explanation (5 topics for healthcare but 8 for bitter lemons??)



	Healthcare Corpus					*Bitterlemons Corpus*				
Feature Set	**Mean**	**Med**	**Max**	**MaxLL**	**Corr**	**Mean**	**Med**	**Max**	**MaxLL**	**Corr**
bag of words	61.12 +/- 0.76%	61.01	72.17	52.92	0.187	68.22 +/- 3.31%	69.26	88.27	84.94	0.39
- no stopwords	60.58 +/- 0.79%	60.50	72.18	62.58	0.154	61.29 +/- 3.05%	57.69	91.34	82.91	0.33
full-tuples	62.42 +/- 0.88%	62.47	74.04	63.37	0.201	80.89 +/- 3.45%	85.40	94.07	92.10	0.34
+ negation	63.67 +/- 0.81%	64.54	74.07	69.25	0.338	80.60 +/- 3.88%	88.07	95.61	91.32	0.66
+ neg. + polarity	63.16 +/- 0.94%	64.46	74.05	67.8	0.455	82.53 +/- 3.55%	86.64	94.44	91.16	0.31
gen. full-tuples	63.80 +/- 0.73%	64.35	73.29	71.70	0.254	76.62 +/- 4.09%	84.56	94.53	84.56	0.25
split-tuples	68.32 +/- 0.90%	70.74	77.80	76.57	0.646	77.14 +/- 3.64%	81.29	92.99	88.13	0.30
+ negation	68.00 +/- 0.91%	69.11	79.73	76.14	0.187	83.53 +/- 3.05%	87.71	95.00	95.00	0.12
+ neg. + polarity	65.11 +/- 1.05%	65.35	78.59	67.22	0.159	81.24 +/- 3.37%	83.44	95.03	88.55	0.08
gen. split-tuples	69.31 +/- 0.83%	70.69	77.90	73.90	0.653	76.69 +/- 4.36%	83.78	93.60	91.67	0.09

- Labels
    - Mean/Med/Max is because of multiple Gibbs Runs
    - MaxLL maximized log-likelihood with TAM
    - Corr is Pearson correlation coefficient

VIEWPOINT SUMMARIZATION

- TAM aligns text excerpts to viewpoints
  - But how do those become summaries?
- LexRank
  - Graph
    - Sentences = nodes
    - Edges = connect sentences
    - Weight of edges = sentence similarity

COMPARATIVE LEXRANK

- Bias the random walk to favor
  - excerpts that represent a viewpoint
  - Excerpts that represent a topic
- Jumping to sentences representing a viewpoint
  - Use P(V | X) from TAM
- Tunable parameter to control level of contrast

SUMMARY GENERATION

- Macro
  - Split excerpts into two sets, one for each viewpoint
  - Generate one summary for each viewpoint
    - Keep to *n* sentences above relevancy threshold
- Micro
  - Input: pair of sentences
  - Use TAM to see if they represent different viewpoints, but same topic
  - Keep to *n* sentences above relevancy threshold

DATA

- 948 Responses to Gallup phone survey about healthcare views
  - Terse responses of transcribed spoken sentences
  - Balanced
- Bitterlemons: 600 editorials about the Israel Palestine conflict
    - Long/verbose with actual sentences
    - balanced
- Pros
  - Available
  - Different domains

RESULTS

- Comparisons
  - LexRank
  - Lerman and McDonalds (2009)
    - LexRank + algorithm to minimize contrastiveness of sentences
- Metric
  - Rouge

EVALUATION

- Bitterlemons
  - Generate macro summaries for 2 viewpoints
  - Ask humans to label each summary as Israeli or Palestinian
  - 11/12 sentences places in correct summaries
  - Humans labeled 78% of the summaries correctly
  - Rouge scores .1 higher than baseline
- Healthcare
  - Microsummaries
  - Annotators identify contrastive pairs in gold summaries
  - No previous algorithms to compare against, but rouge scores ranged from from .3 to .35

SENTIMENT SUMMARIZATION

EVALUATING AND LEARNING USER PREFERENCES

KEVIN LERMAN, SASHA BLAIR-GOLDENSOHN, RYAN MCDONALD

GOALS

- Generate summaries of product reviews.
- Each summary should reflect the average opinion.
- It should contain opinions about the important aspects.
- They should consist of complete sentences extracted from the reviews.
- The total length of the summary should not exceed a predetermined length.

THREE PHASES

1. Create three hand-made models for summarizing reviews.
2. Use humans to rate the quality of the summaries and choose which ones they prefer.
3. Use the human ratings as the training data to learn, using SVM, which model is the best to use for any situation.

THE MODELS

- Sentiment Match (SM)
  - Pick a summary whose sentiment matches that of the star rating.
  - Disregards aspect.
- Sentiment Match + Aspect Coverage (SMAC)
  - Pick a summary with good sentiment match and has good diversity over the aspects.
  - It is possible to have good sentiment match and still pick sentences that are contrary to the true overall opinion of aspects so long as the sentiment balances out.
- Sentiment-Aspect Match (SAM)
  - Pick a summary that has a high probability of being representative of the sentiment of the entire entity with respect to aspects.
  - Attempts to solve the sentiment-aspect mismatch problem.
- Baseline
  - Pick first sentence of each review until the target summary length is satisfied.
  - Disregards both sentiment and aspect.

HUMAN EXPERIMENT

- Dataset
  - 165 electronics product reviews
  - 4 to 3000 reviews per product with an average of 148
  - Target length for summary is 650 characters
  - SM, SMAC, SAM, and baseline are compared
- Process
  - Raters are shown the original overall star rating and two summaries created using two different models.
  - Raters pick which one they prefer.
  - Raters are also asked to pick either no preference, strongly preferred, preferred, or slightly preferred for each review judgment
  - Over 100 raters and 1980 rater judgments

EXPERIMENT RESULTS

- No significant difference in user preference overall between the three sentiment aware models.
- Rater's prefer sentiment aware models over the non-sentiment aware summarization method (baseline).
- Analysis of results reveal that some models are preferred over others in certain circumstances.
- Authors decided to learn these circumstances with machine learning (SVM) and using the experiment results as the training data.
- The SVM model was able to choose the correct model 7.5%-13% more often than the baseline that had ~55% accuracy.

CRITIQUE

- The authors demonstrate a reasonable method of tuning a difficult to tune algorithm.
  - Create multiple systems
  - Get user feedback
  - Use user feedback to train new model
  - Wash, rinse, repeat…
- Raters did not directly rate the quality of the summarization. Instead they rated which summary they preferred (i.e. they didn't look at the original reviews).
- It isn't clear if the development dataset used to create the models was the same dataset as in the human experiment.