Text Summarization WG:

Personalized Extractive Text Summaries

5 August 2022

Group Members

Blake Centini

John M. Conroy

Priscilla C.

Nick Gawron

James Hardaway

Matthew K.

Joseph Kepler

Cas Laskowski

Neil Molino

Liz Richerson

Josh Sheinberg

Josh T.

Sophie Trotto

Julia Y.

Outline

- Summarization Problem Overview
- EDA/Preprocessing
- Use Cases
  - Topic Specific
  - Term Weights
  - Audio
  - Sentence Embeddings
  - Sentiment Analysis
- QA Evaluations
- Future Work
- Demonstration

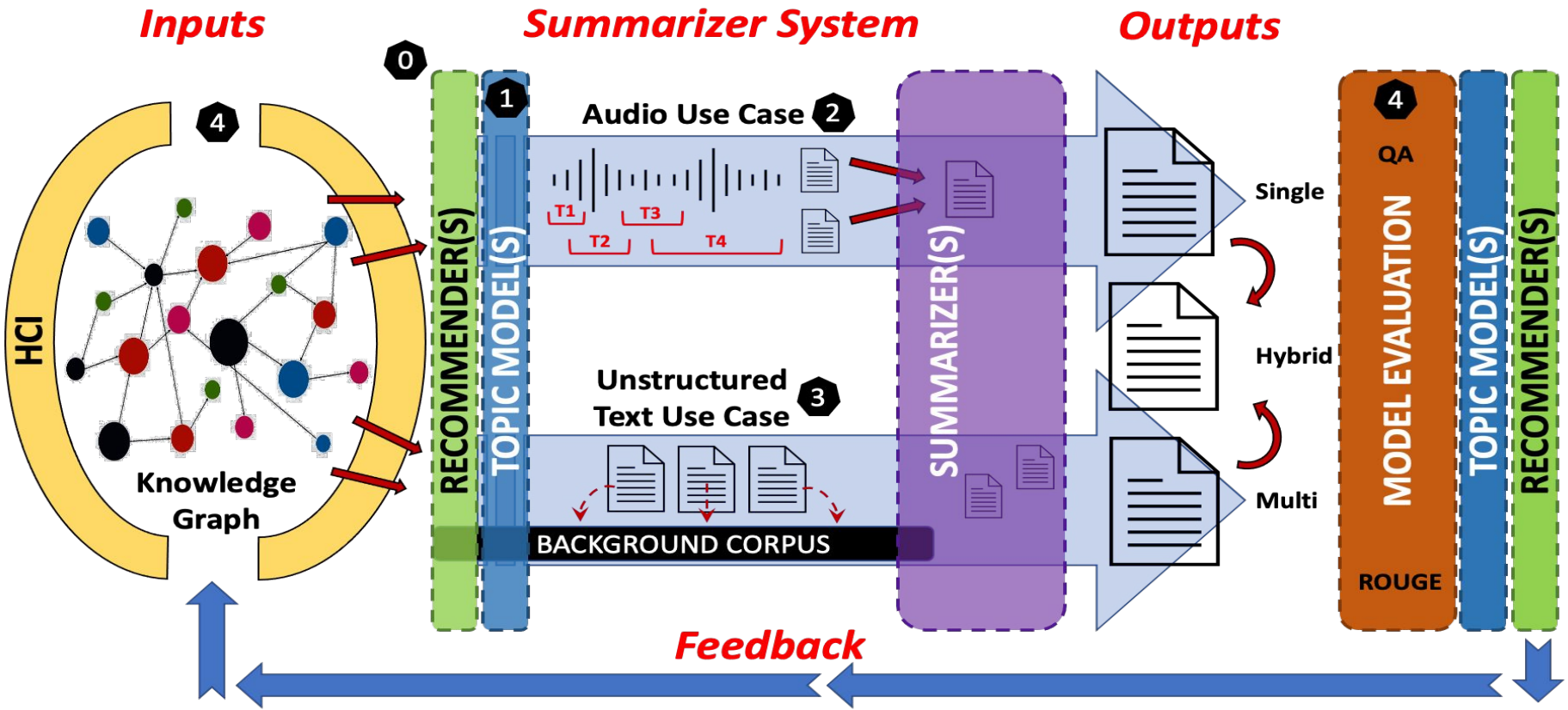
* Please hold questions until the end time permitting

Summarization Problem Statement

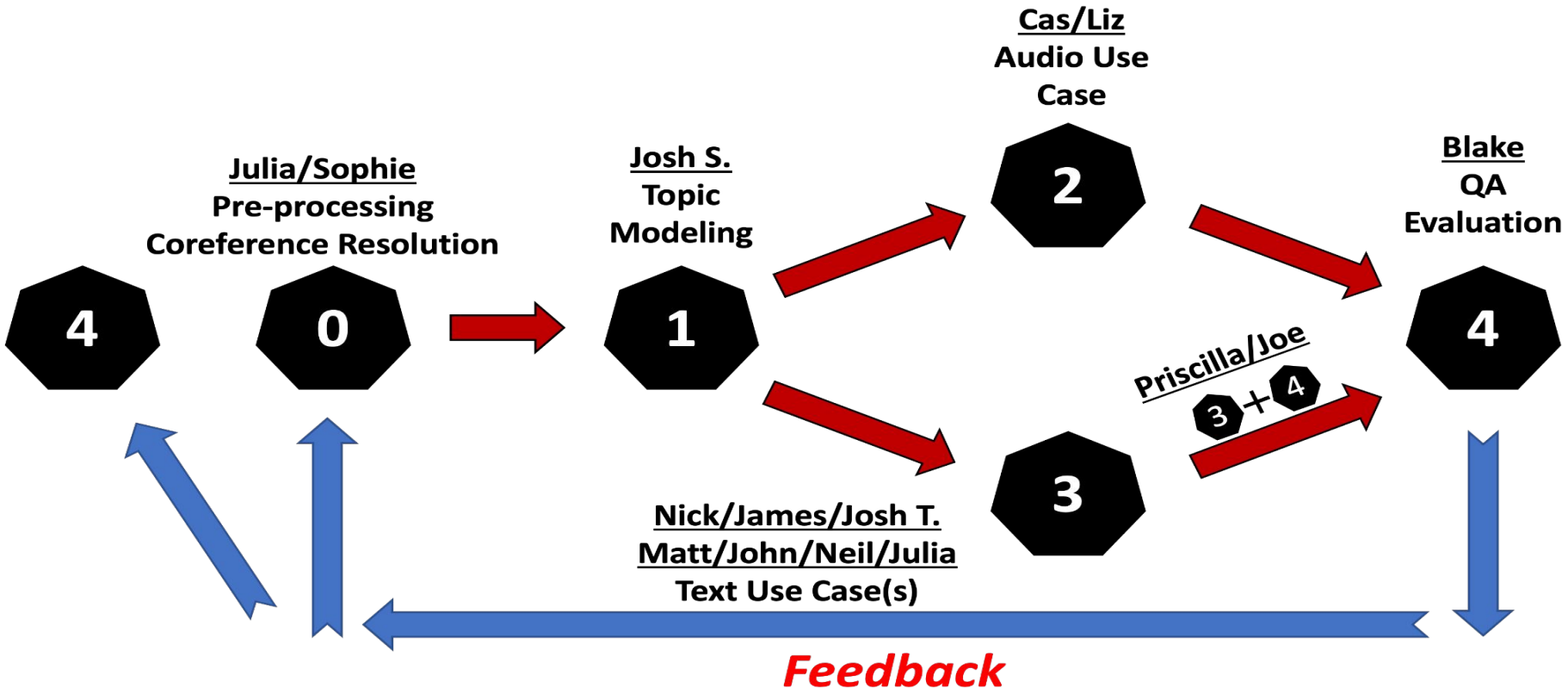
Develop a text summarization capability employing an analyst-specific **background corpus** to drive **term weighting** and **selection**. This challenge was addressed through 3 primary research questions:

1. Can we pre-train the language model to produce more insightful summaries?
2. How can pre-processing improve summarization accuracy or information value?
3. Without reference summaries, how do we determine machine summarization quality?

Text Summarization Process View



Text Summarization Workflow



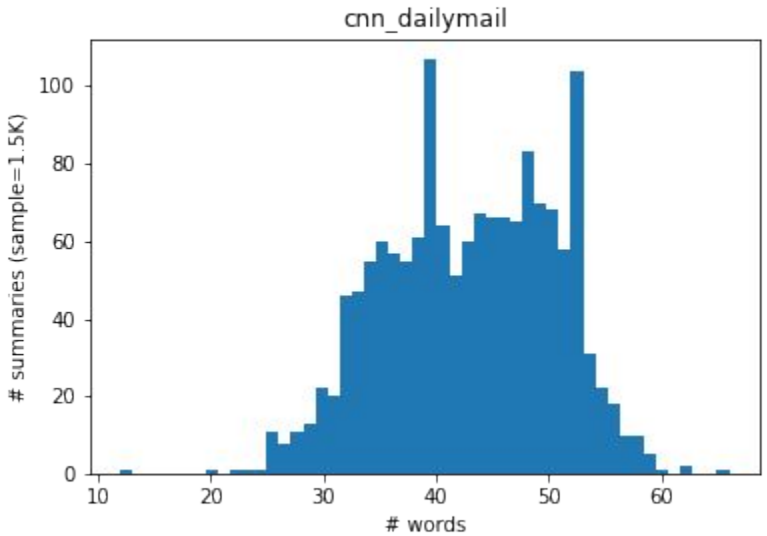
An EDA of the MIND Dataset

- Large-scale dataset for news recommendation research.
- English documents with summaries, titles, categories,...

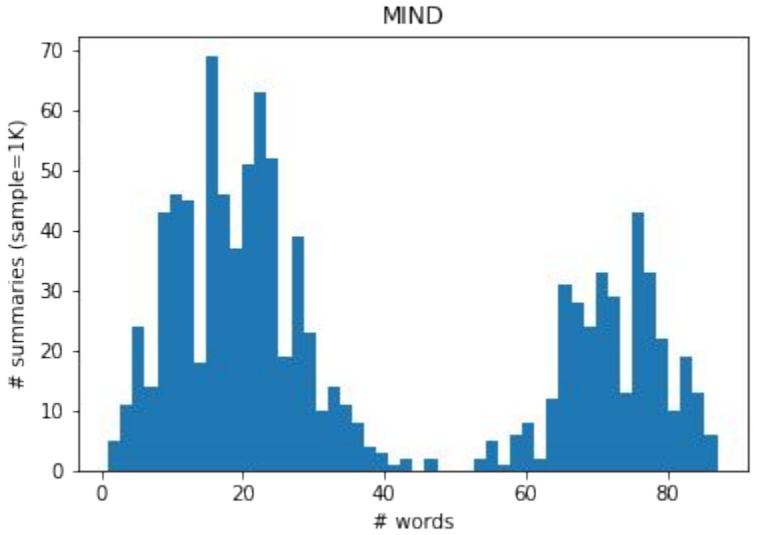
Can we use MIND for text summarization research?

Histograms

CNN/Daily Mail
1 camel hump



MIND
2 camel humps???



Histograms: *x*-axis: # words in a summary; *y*-axis: # summaries (sample size).
Note: EDA restricted to MIND training data.

Warning: Right Camel Hump Full of LEAD!

LEAD	30192	LEAD-prefix	562	OTHER	1697	TOTAL	32451

- **The longer right camel hump summaries are ~95% LEAD (leading sentences of the document)** after text cleaning (due to inconsistent punctuation, spacing) and adjusting for prefixes (dateline) and summary suffixes.
- ***Warning:*** need human summaries for text summ development (benchmarking). LEAD is not human. **MIND is bad for multi-sentence text summ dev.**

Left Camel Hump + SEO

LEAD	6022	LEAD-prefix	1980	OTHER	40938	TOTAL	48940

- **The shorter left camel hump summaries are ~16% LEAD** after text cleaning and prefix/suffix adjustment.
- Manually inspected a handful of summaries:
  - **Meta description** (HTML element for a short (~1 sent) webpage summary), likely human for news. A **search result snippet** (summary part of a search engine query) is based on a meta description, other sources.
    - Unlikely to use bad desc: keywords, irrelevant to content/query, spam.

MIND: Final Remarks

Look at the data. Look at the data. Look at the data.

- This is an ill-mannered dataset for text summarization.
- ***Warning***: MIND is **not** appropriate for multi-sentence text summarization algorithm development. It may be good for single-sentence models. Be careful w/word length: avg misleading ("dip" between camel humps).
- OK: build models w/other news datasets, apply MIND.

Coreference Resolution in Text Summarization

- Extractive text summarizers value sentences less when they contain more pronoun references than proper noun references
- If a sentence included in a summary and lacks the appropriate referential context, the reader may not understand or potentially misinterpret the identity of the referent
- Increases accuracy of abstractive summaries by improving tracing of information flow

SCADS Approach to Coreference Resolution

- Used AllenNLP model as a basis to generate coreference clusters, and then fine-tuned these clusters with custom rules
- Goal to minimize the necessary coreference replacements made
- Some of the most important rules included are:
  - Only one coreference replacement would be made per sentence
  - By default, AllenNLP chooses the first span of each cluster as its referent. Instead, we use spaCy's part-of-speech and entity tagging across the cluster to determine the "best" referent for each cluster

Results for Single Document

Original document	AllenNLP resolution	SCADS resolution
International human rights groups on Saturday urged **Sri Lanka's new president** to immediately order security forces to cease use of force against **protesters** after troops and police cleared **their main camp** following months of demonstrations over the country's economic meltdown. A day after **President Ranil Wickremesinghe** was sworn, hundreds of armed troops raided **a protest camp outside the president's office** in the early hours of Friday, attacking demonstrators with batons.	International human rights groups on Saturday urged **Sri Lanka's new president** to immediately order security forces to cease use of force against **protesters** after troops and police cleared **protesters's main camp** following months of demonstrations over Sri Lanka's economic meltdown. A day after **Sri Lanka's new president** was sworn, hundreds of armed troops raided **their main camp** in the early hours of Friday, attacking demonstrators with batons.	International human rights groups on Saturday urged **President Ranil Wickremesinghe** to immediately order security forces to cease use of force against **protesters** after troops and police cleared a **protest camp outside the president's office** following months of demonstrations over the country's economic meltdown. A day after **President Ranil Wickremesinghe** was sworn, hundreds of armed troops raided a **protest camp outside the president's office** in the early hours of Friday, attacking demonstrators with batons.

Suggestions for SCADS Coref Resolution

- Need for metrics, improvements only observed visually for a small sample size
- Choice of rules could be modified. Dependence on accuracy (or lack thereof) of spaCy's POS and NER tagging
- AllenNLP model could be modified under the hood, or consider another model entirely once one is available
- Coreferences as tooltips in the TLDR UI

Topic Applicable Backgrounds

- **General Background Corpus (GB):** a large set of documents representative of a language
- **Topic Applicable Background (TAB):** a background corpus with documents of a similar topic (SAS) to a document that we wish to summarize.

G^2-Test Formulation

- Helps determine *common* words in a Document

For Document $\mathbf{D}$, a Background Corpus $\mathbf{B}$, & term t:

$$H_0 : P(t|\mathbf{D}) = P(t|\mathbf{B})$$

$$H_\alpha : P(t|\mathbf{D}) > P(t|\mathbf{B})$$

Methods for Sequential Testing

- **1-Test**: G^2-Test on General Background
  - Input: Document Bigrams
  - Output: Keeping terms with $p<0.0001$

- **2-Test**: G^2-Test on TAB
  - Input: Significant Terms from 1-Test
  - Output: Keeping terms with $p>0.01$

Percentile Boosting Term Weight Update

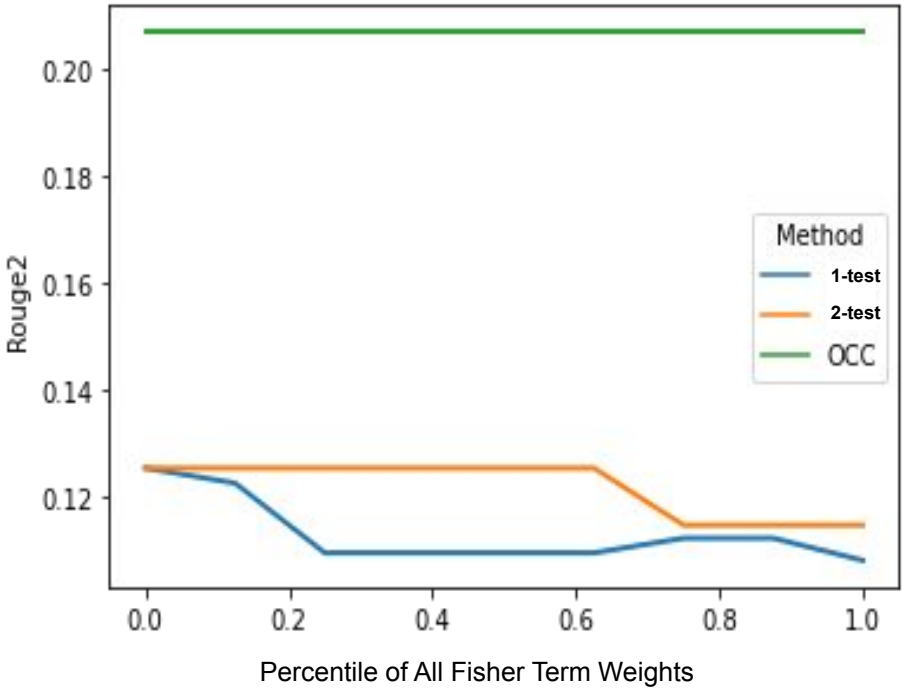
- Significant Terms ⇒ Increased Term Weight

- Each Term Given a Weight
  - Compute the percentile `p`
- Add constant to weight of Significant Terms

```
twU ← %(twD,p)*Sig + twD
```

Rouge-2 Results for TABs

- Sweep to Find Optimal *p*
- GB Rouge-2 Bounded Above
- Lack of Position ⇒ Lower ROUGE
  - Leading Summary



What's New, TLDR?

- The TLDR should be able to pick up on new information and summarize it for the analyst

- We start with background information
- New information comes in
- We would like to summarize just the new information

2 Days In The Life Of An Analyst

- Based on data from Text Analysis Conference (‘08-’11)
- We had 10 articles on a particular topic at each of two points in time (A then B)
  - About 50 topics in each year’s data set
  - The B set is always an update to the A set
- We can compute A and B summaries for each topic
- Can we focus the B summaries on just the new info?

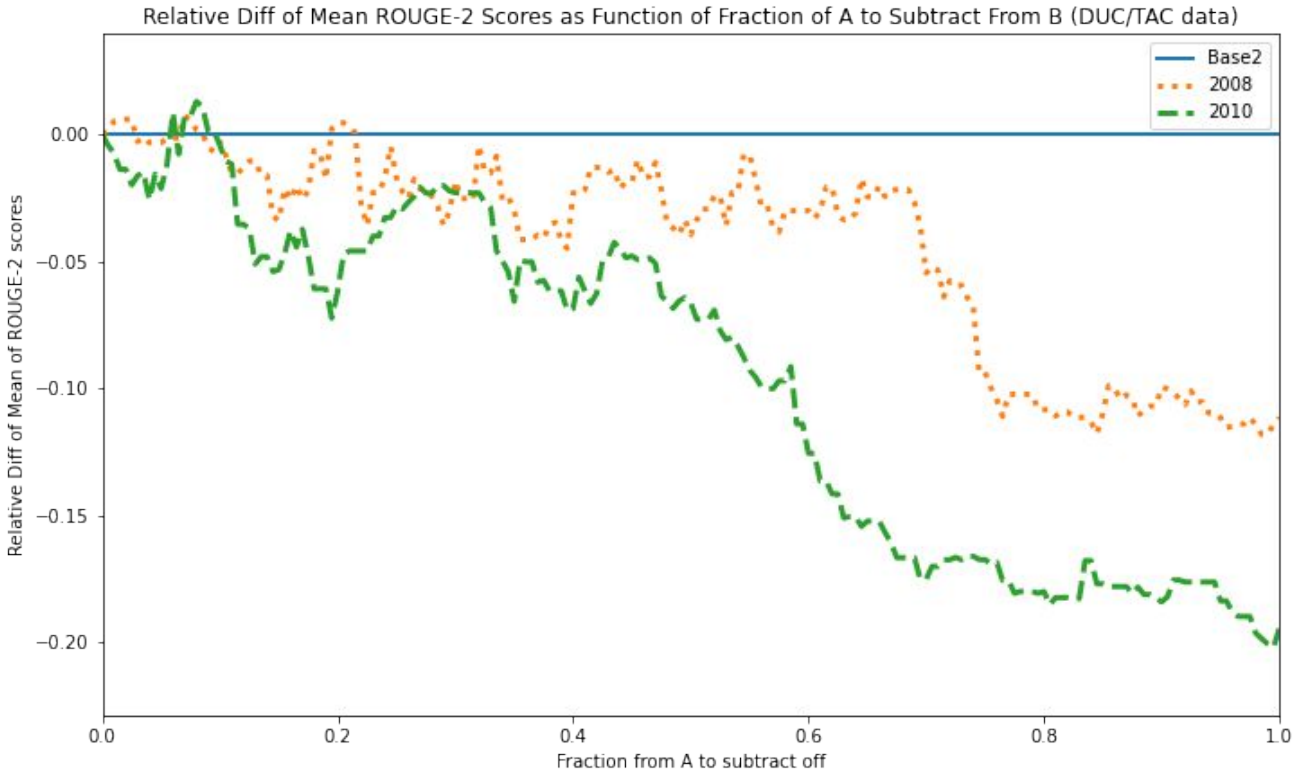
occams

- **occams**: provides extractive multi-doc text summaries
- It estimates latent **term weights** from a non-negative matrix factorization derived from the documents
- It **extracts sentences** to optimize a combinatorial covering of the documents **using the term weights**

`occams` Term Weights Experiment #1

1. Compute term weights for both Updates A and B
2. Subtract a specified fraction of the Update A term weights from the Update B term weights
   - **`twB ← twB - x*twA`**
3. Generate extractive summaries for Update B (**`occams`**)
4. Compare these with human-generated summaries (ROUGE-2 scores)

occams Term Weights Experiment #1



occams Term Weights Experiment #1

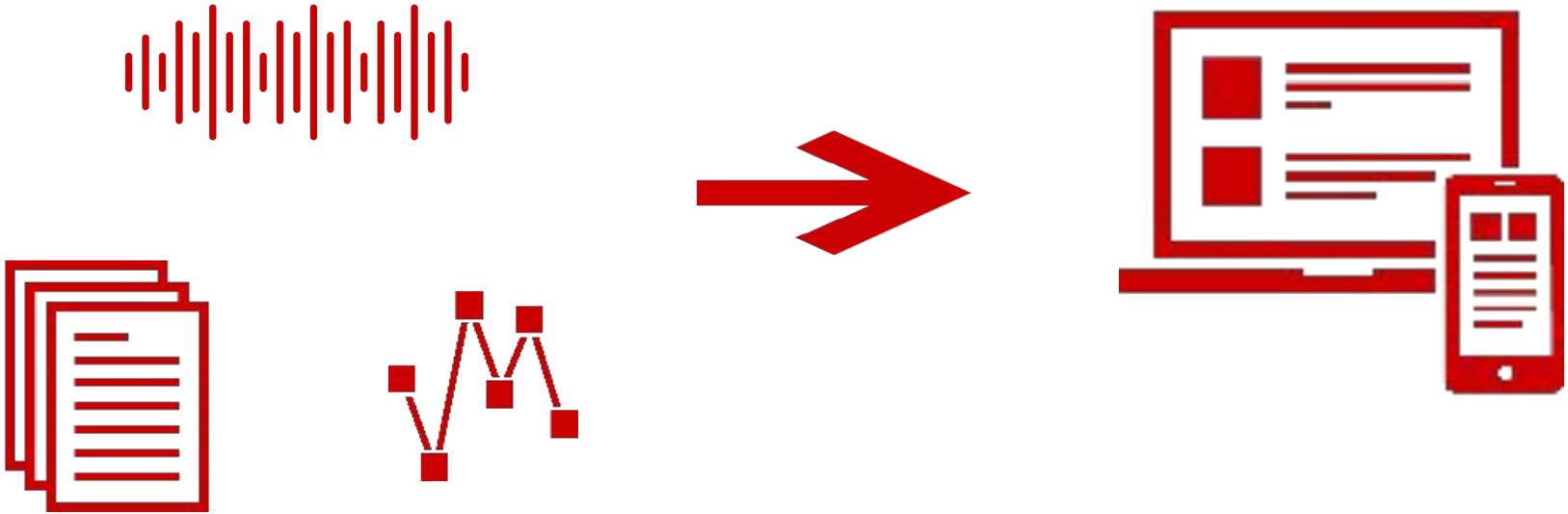
- We conclude that subtracting the term weights from the earlier update does not help focus the later update on the new material

- But the output really liked a few “generic” sentences
  - *“That would have been great for the country.”*

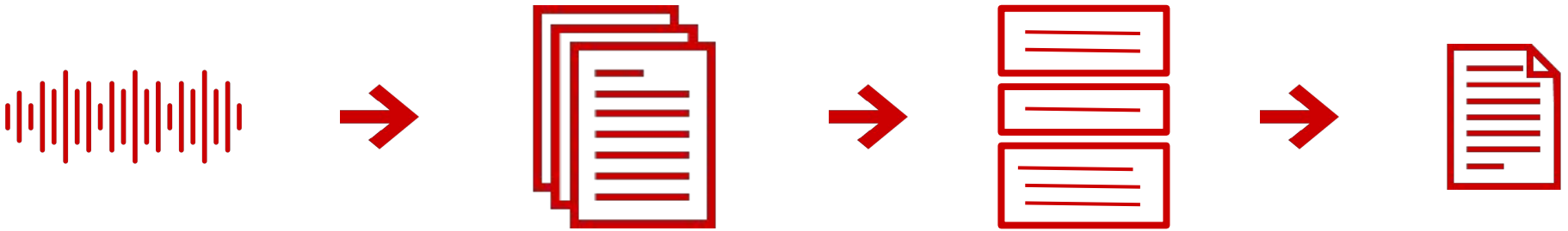
occams Term Weights Experiment #2

- Could removing the most common terms help **occams** select more significant sentences?
- We conducted experiments using a background corpus to remove several variations of “common” terms
- The results were mixed:
  - This is likely too blunt a technique
  - It raised promising questions for future research

Audio Data Summarization



Audio Summarization Workflow



No Segmentation

Transcript

Transcript,
no metadata

Speaker Turn Segmentation

Transcript

Segment 1:
speaker 1, utterance 1

Segment 2:
speaker 2, utterance 2

Segment 3:
speaker 1, utterance 3

Semantic Segmentation

Transcript

Topic 1:
00:01:08-00:02:42

Topic 2:
00:02:27-00:04:16

Sample Segments

Speaker Turn Segmentation

the point you raised about the **profits**. The word **profits** in the name of top of page two on Uhh. After considerable discussion, I think this decision was to And if we can tow work **profits** as a kind of uh huh. Many of us played of his credit card to start with this issue. That doesn't really her enough to protect my double tomorrow. But, I mean, let me argue it the other way for a fact, if I can. First Secretary and I are way out there on this subject. I mean, I

Semantic Segmentation

The point you raised about the **profits**. After considerable discussion. I think this decision was to And if we can tow work **profits** as a kind of. Small **profits**, and that's something else again are put in there. **Control of profits**. But there have been talking about **profits**, **windfall profits**. And that is a briefing, say **profits** or control. No one only **control profits** in the sense that we make promises to high as we can.

Sample Summaries: Speaker Turn Segmentation

Extractive

You talk about that. I want to see. But I don't know. Do it. There is the pressure. You know what I mean? We could do that. So that's the problem. You've got a lot of lacking. I have. Do you feel that we are on the way now. That's what we have to do this take. We're gonna be a partner. I think it is. You get a job. All right. You think that a big company? But you have it out there. This is wrong. I'm trying to get out of one thing. So we've got to go.

Abstractive

Roger Lee says the story was bad, and it made them, but if they want was when he read it . China doesn't see he could go in the staff meeting this morning and Ziegler shot them, which was paying Good run stepped in and he said, Well, maybe a prominent. But in all fairness to the medical centers is ripping Rogers, You know, staffing that you shouldn't dio, uh, brought civil . The Hildebrand, this place in view of the fact that we would have such a short time. It's good for the people here right here.

Sample Summaries: Semantic Segmentation

Extractive

The point you raised about the **profits**. I think this decision was to And if we can tow work **profits** as a kind of. And that is a briefing, say **profits** or control. No one only **control profits** in the sense that we make promises to high as we can.

Abstractive

There have been talking about **profits**, windfall **profits** and **control of profits** . No one only **control profits** in the sense that we make promises to high as we can . And if we can tow work profits as a kind of . **Small profits**, and that's something else again are put in there .

Semantic Segmentation

Transcripts

Topic 1:
File1, File 3

Topic 2:
File 1, File 4

Noise/Disfluencies

It's great. All right, way. Okay, okay. That's what i. A Yeah. I think you're right. Oh, well, right. Very good. A lot of people. Sure, sure. Uh oh, yeah. Yeah, that's fine. Yeah, a Okay. Yeah, that. I mean,. it's true. Uh Okay. Right, but okay. there, Yeah. Okay, Alright. So Okay,. A Okay. You know. Which way. In a way. You mhm. a, of course. Oh, Yeah. Well, all right. Oh, right, right, right, right. Uh, really. a a got that. You know,. I know that. You Yeah, Okay. Yeah, well, I. Oh, Yeah, yeah, right. Yeah, Okay.

Artifacts of Transcription

Well, I'll plan **toe**. Well, then we'll plan. He wants it **toe** to be a little bit later, though. You have to go **toe** hand. Just try **toe**. For example, I'm going **toe**. The point is **toe**. I'll have **toe**. think it's just well, **toe**. you have **toe** to put it this way. Toe what. I don't want to **toe**. **Toe** sort of put the spotlight of attention out there on if something could come out of it. Well, I'm just going **toe**. Well, anyway, they will try **toe**. I'm going **toe**. I'm going **toe**. I mean, I don't want **toe**

Clustering semantic segments

- Clustered similar segments across multiple file on subset of full corpus
- Resulting clusters highlight similar content in different files

Sample Cluster

It's a national issue and will be here long after **Vietnam**. Hurt the **North Vietnamese** for us. Get **Vietnam** and everything. **Saigon**, I suppose. **Laos and Cambodia**. What the response in **Saigon** is. You know, the same thing on **Laos**, that kind of argument, see. Nothing more we could do on the **Vietnam** side, except they're not going to get the way, way, way. **North Vietnam**, this is your credibility that we wanna be.

Sample Summaries: Clustered Segments

Extractive

Yes, **Vietnam**, that. **Vietnam** started there for us. This is altogether **Vietnam**. Oh, you in **Vietnam**. **North Vietnam**, this is your credibility that we wanna be. **North Vietnam** is, Ah, just as a dragon. assuming no action in **Vietnam**. A few problems with **Vietnam** and others. So I talked mostly about **Vietnam**. And the **North Vietnamese** are bantry.

Abstractive

CNN.com will feature iReporter photos in a weekly Travel Snapshots gallery . Please submit your best shots of our featured destinations for next week . Visit CNN iReport.com/Travel next Wednesday for a new gallery of snapshots . Visit www.dailyimpact.com for a gallery next week for snapshots .

Findings

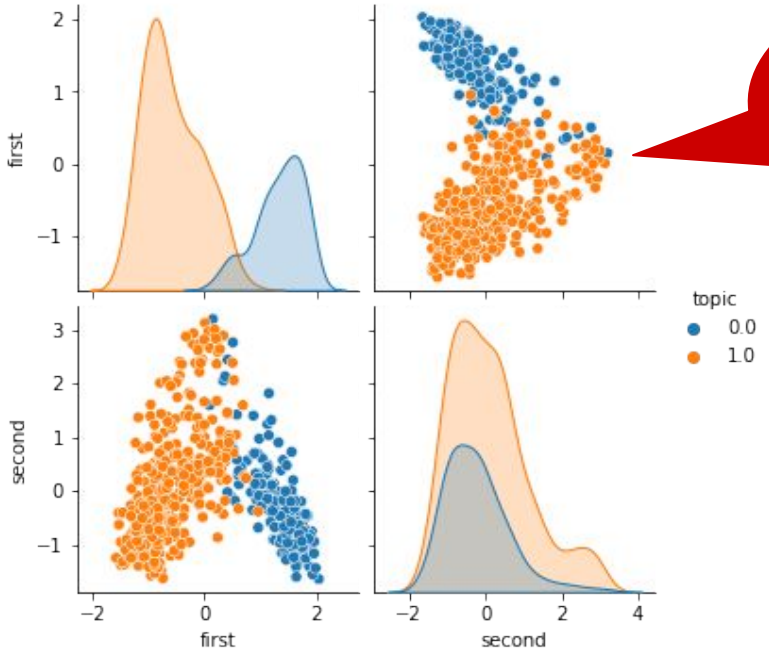
- Audio data summarization requires separate pipeline
- Extractive summaries are of limited value
- Abstractive summaries are unreliable
- Semantic clustering has promise

- Extend semantic clustering work
- Use additional preprocessing, such as CRR & NER
- Explore other description methods

Future Work

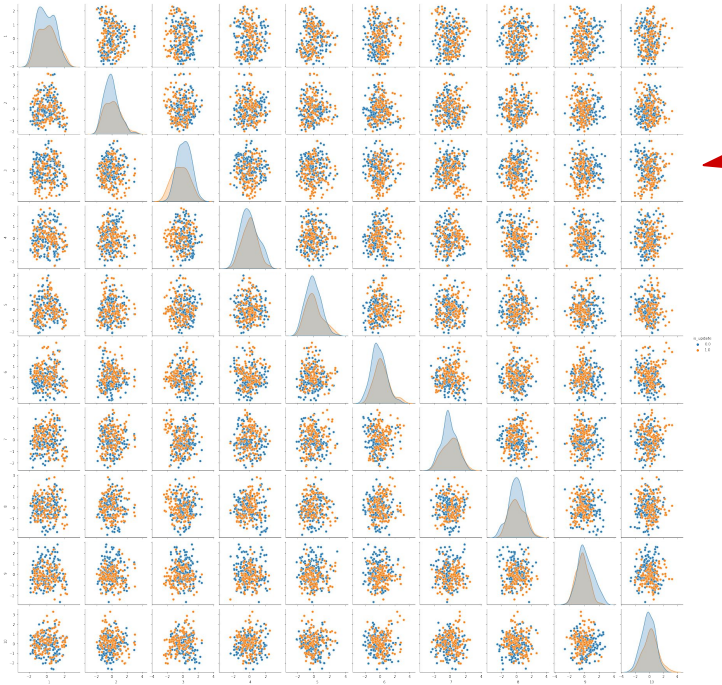
- Extend semantic clustering work to explore utility for content triage and navigation
- Run question-answering evaluation on summaries to objectively compare summaries and quantify hallucinations
- Use additional preprocessing, such as coreference resolution on transcripts and background corpora for summaries
- Provide additional information in summary, such as by incorporating additional metadata into summary

Sentence Embeddings from 2 topics of Day 1



Very Clean Separation!

Sentence Embeddings from corresponding topics

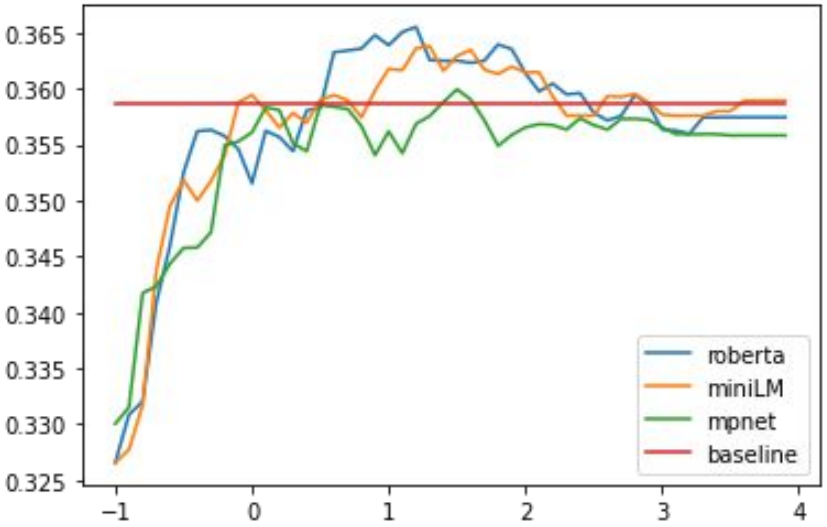


Less Clean Separation

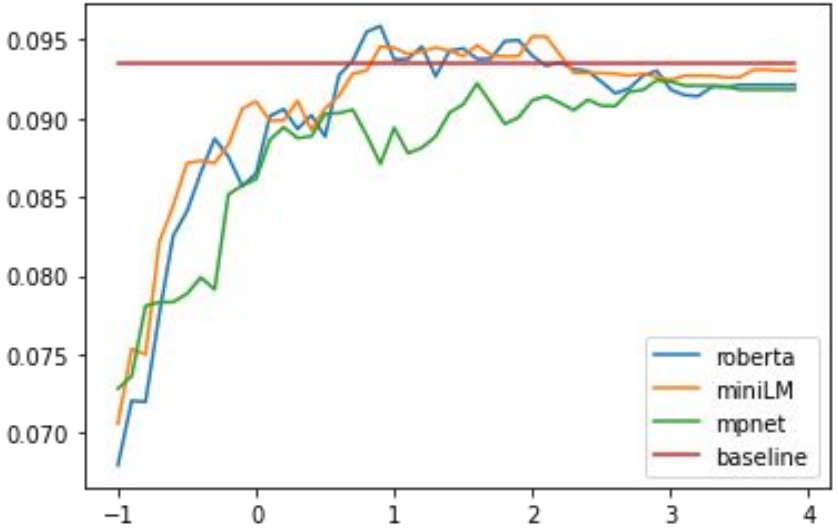
Quadratic Discriminant Analysis is ~70% accurate

Use a classifier to exclude sentences that look like from day 1?

ROUGE results are so-so



ROUGE-1 RoBERTa & MiniLM vs Baseline by QDA Threshold



ROUGE-2 RoBERTa & MiniLM vs Baseline by QDA Threshold

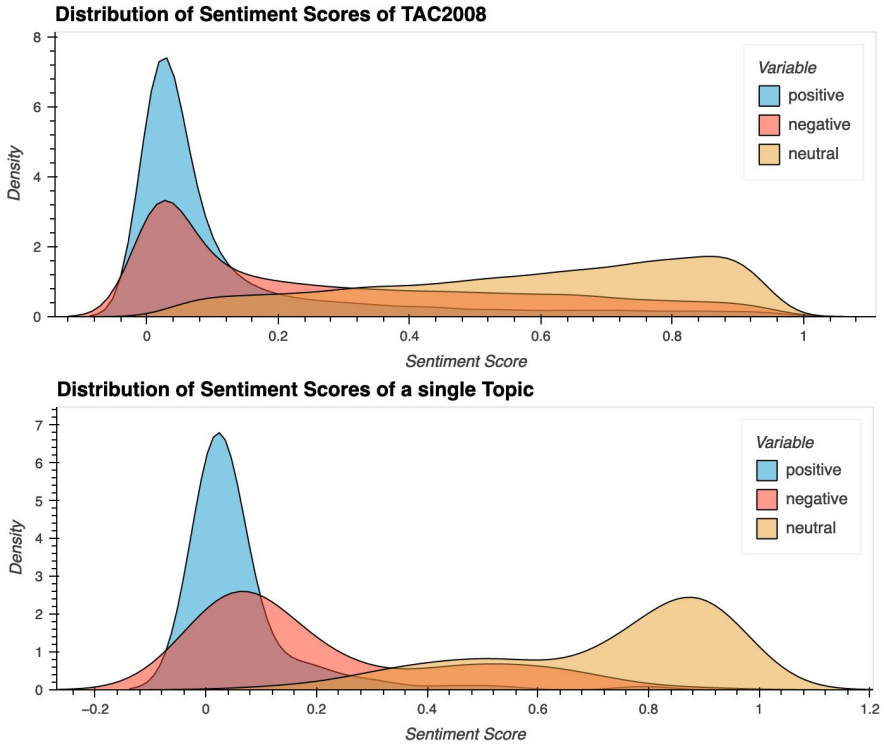
modest improvement, but highly sensitive and not obviously statistically significant

Sentiment analysis to aid summarization

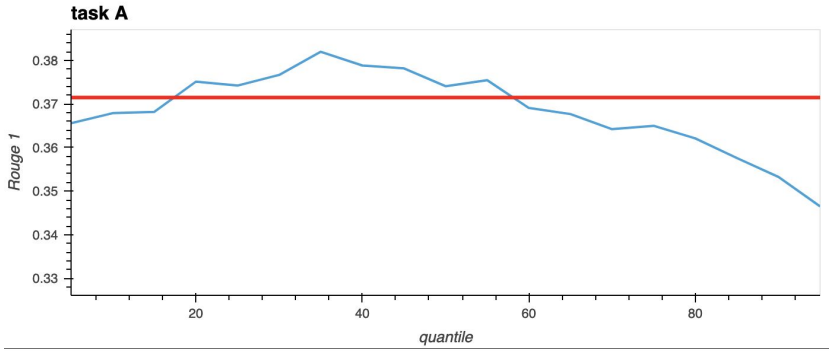
- Simple hypothesis - sentiment of extracted sentences should match sentiment of document corpus
- Simple experiment - use neural sentiment models to classify sentences and filter poor matches
- Use TweetEval to evaluate sentence sentiments (Trained on short, sentence like text, most models trained to classify reviews)

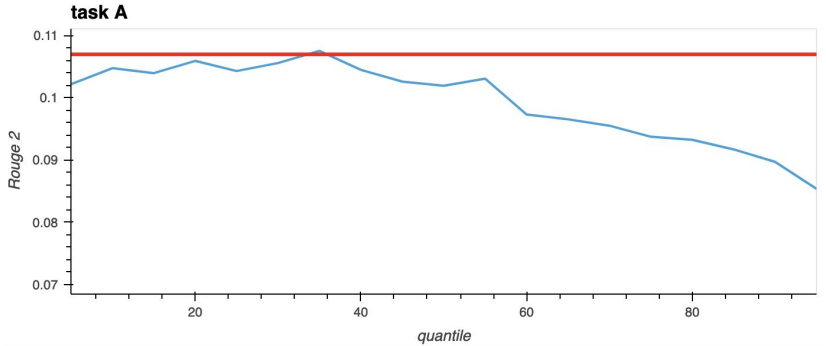
Most news articles are neutral to negative

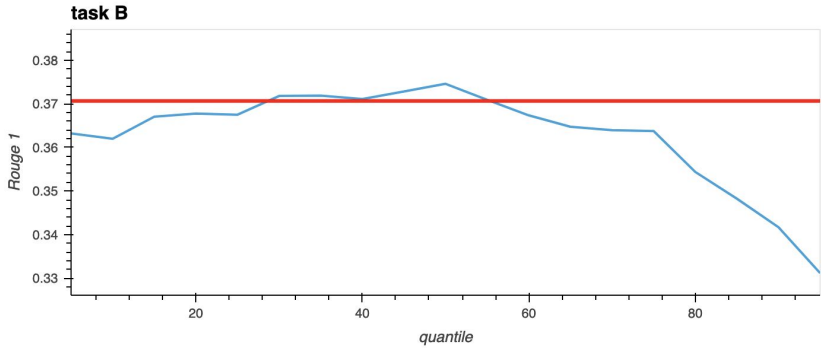
- Overall TAC2008 is neutral with negative tail
- Individual topics can have interesting distributions
- Given the sentiment bias, filter out sentences that do not match the bias for each topic

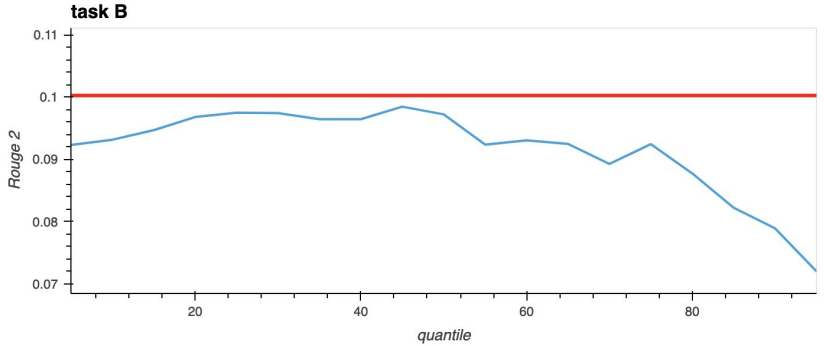


Filter sentences based on quantiles of negativity to positivity ratio



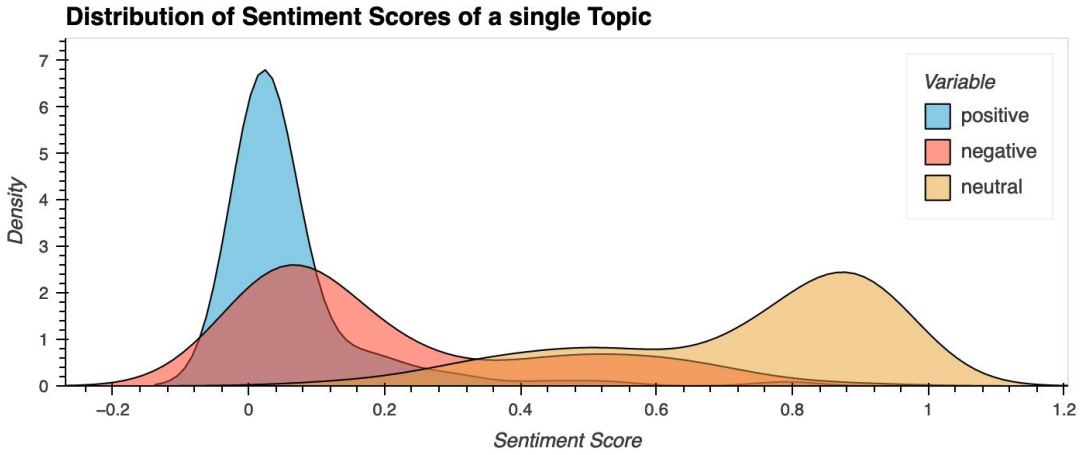






Future experiments should be more sophisticated

- Current experiment allows no nuance (KL divergence of resulting summarized distribution?)
- Implement sentence weighting feature in `occams`
- Explore sentiment scoring as auxiliary metric



QA Evaluations 1

- General methodology and benefits
  - Question generation
    - To pre-filter or not to pre-filter
  - Answer overlap evaluation

QA Evaluations 2

- FEQA
  - Not technologically compatible with OCCAMS
- QAFactEval
  - Works with Extractive and abstractive
  - Uses multiple metrics

QAFactEval Results

```
Out[42]: [({'qa-eval': {'f1': 0.0,
            'lerc_quip': 0.03752674425349516,
            'is_answered': 0.029411764705882353,
            'em': 0.0}},
          [{'question': {'question_id': '633e46b10e15d2667eb184c416ef3750',
              'question': 'What type of devices are there?',
              'answer': 'Explosive Devices',
              'sent_start': 0,
              'sent_end': 17,
              'answer_start': 0,
              'answer_end': 17},
             'prediction': {'prediction_id': 'cfcd208495d565ef66e7dff9f98764da',
              'prediction': 'gun laws',
              'probability': 5.5734798653532126e-11,
              'null_probability': 0.9999999996922013,
              'start': 653,
              'end': 661,
              'f1': 0.0,
              'lerc_quip': 0.0,

In [41]: score
Out[41]: 0.03752674425349516
```

DEMONSTRATION

Key Insights

- NIST dataset most appropriate for evaluating text summaries
- Coreference resolution useful for pre-processing, but more methods can be applied
- Audio transcript quality matters
- TABs improves GB ROUGE
- Sentiment scoring a viable alternative to term stats


THANK YOU

Icons

Use NC State's on-brand icons to add visual interest and illustrate important facts and figures within your content.

brand.ncsu.edu

BACKUP AFTER THIS

"Two days in the life of an Analyst"

Information Technology Laboratory
Text Analysis Conference
NIST National Institute of Standards and Technology

Day1 ~ updateA

~50 topics, each with 10 documents each
⇒ **multidocument** summarization

4 different human summaries per topic

Day2 ~ updateB

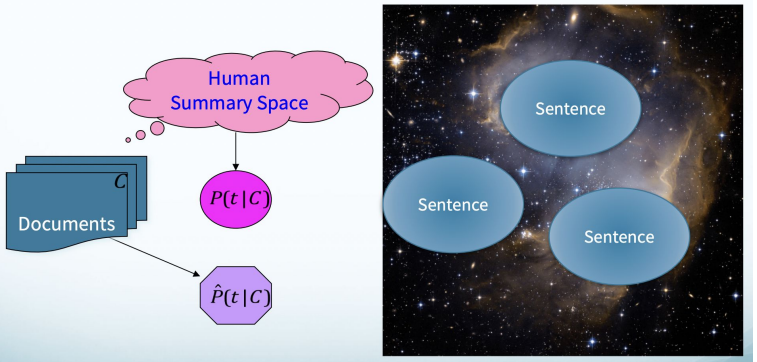
same 50 topics, each with 10 documents each

humans summaries aim to convery only the **new** information.

OCCAMS for abstractive summarization

OCCAMS conceptually

We can have the computer summarize by covering important terms with sentences



OCCAMS concretely

OCCAMS approximates the below binary integer linear program.

$$y^* = \text{argmax}_y \sum_j y_j \hat{\Pr}(j|\tau)$$

$$\text{subject to}$$

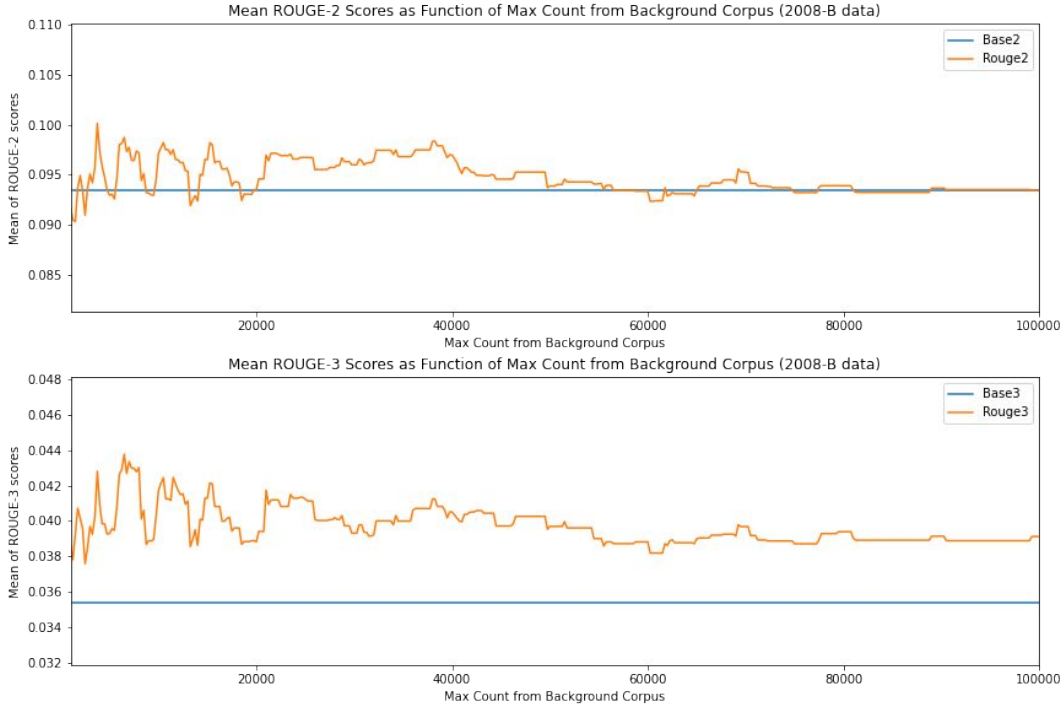
$$\sum_j x_j n_j \leq 100$$

$$y_i - \sum_j a_{ij} x_j \leq 0$$

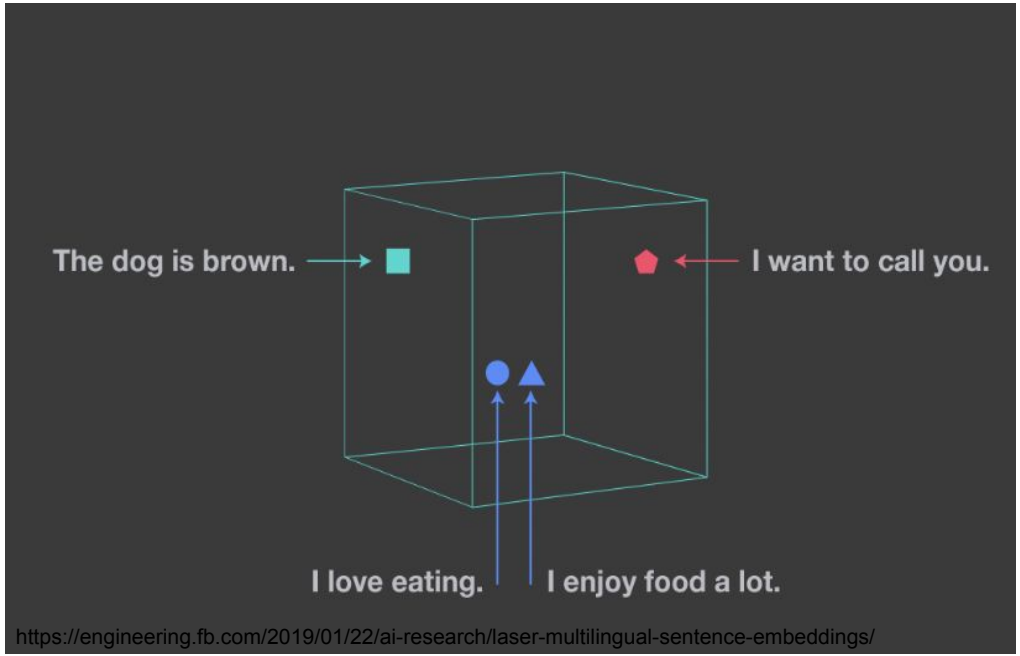
$$a_{ij} x_j - y_i \leq 0$$

Can influence machine generated summary by (1) adjusting term weights or (2) excluding sentences

Josh T. say something here?

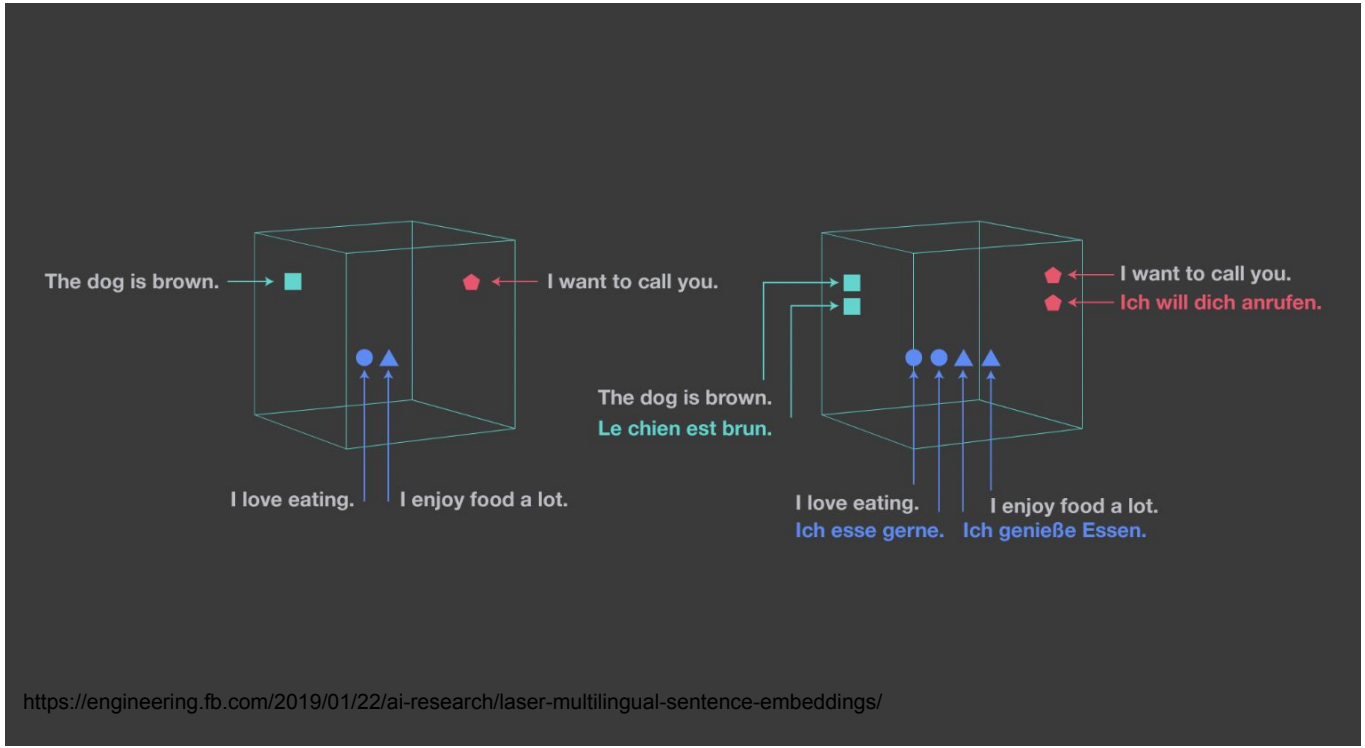


Sentence Embeddings



Semantically similar sentences land near each other!

Sentence Embeddings in Any Language!



This will be useful when we extend to many domains of interest!