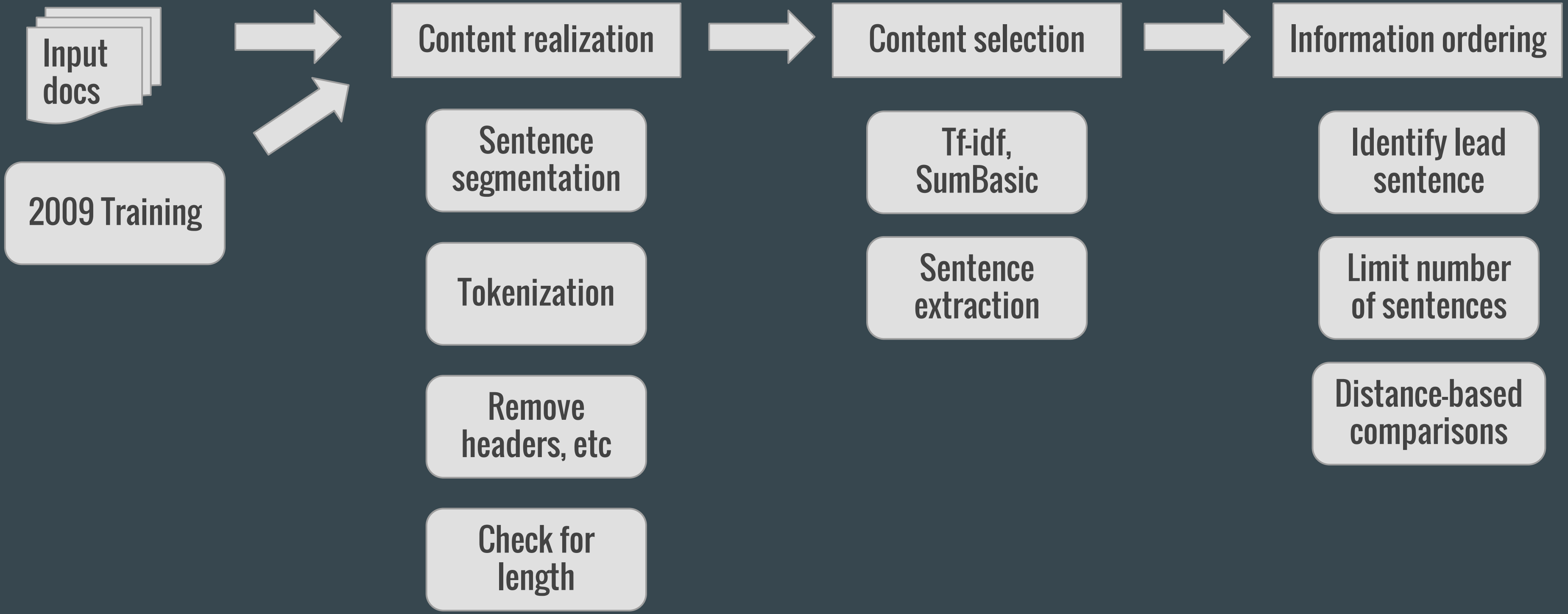
D3 - Multi-Document Summarization

•••

Maria Sumner, Micaela Tolliver, Elizabeth Cary

SYSTEM ARCHITECTURE



Input docs

2009 Training

Content realization

Sentence segmentation

Tokenization

Remove headers, etc

Check for length

Content selection

Tf-idf, SumBasic

Sentence extraction

Information ordering

Identify lead sentence

Limit number of sentences

Distance-based comparisons

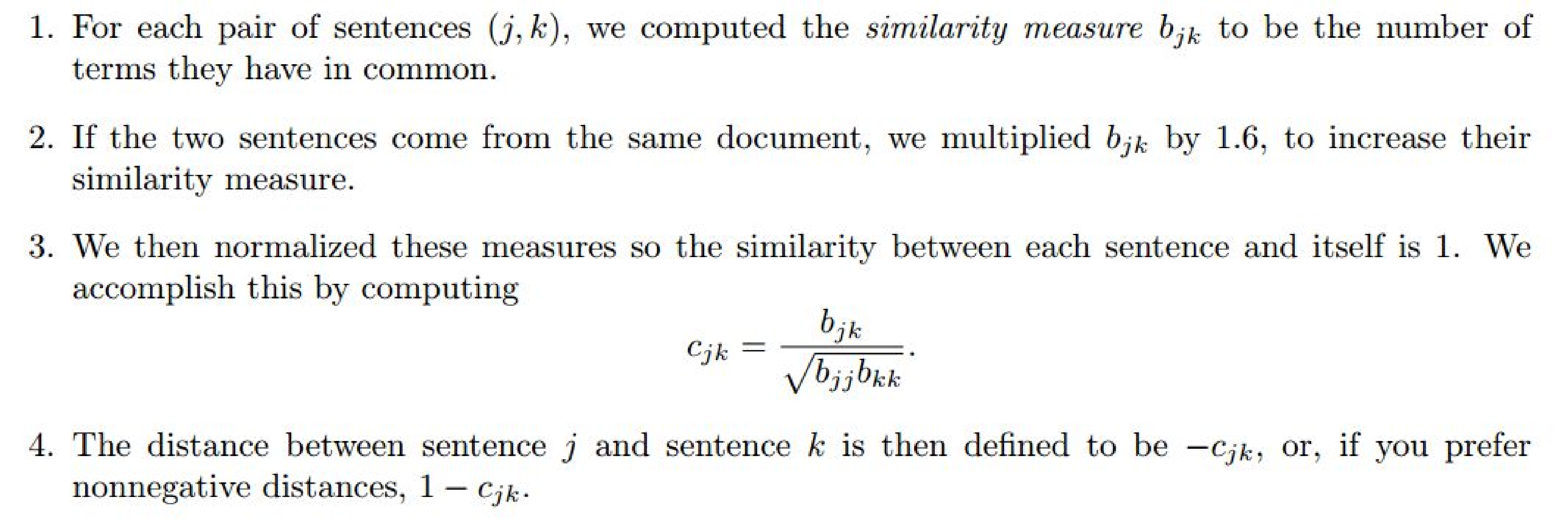
IMPROVEMENTS IN PRE-PROCESSING / CONTENT REALIZATION

- More header information is cut out
  - Time information: 10:55 a.m. (0755 GMT)
  - Location information: AUSTRA_AVALANCHE (Galtuer, Austria)
- Ignores sentences with phone numbers and URLs
- Initial whitespace and dashes are taken out
- Underscores are taken out
- Ignores sentences with quotations
- Ignores sentences with questions

IMPROVEMENTS IN CONTENT SELECTION

- When summing up tfidf values in sentence scoring, penalize repeating words to avoid redundancy in sentence
    - Similar approach to downweighting; update TFIDF score by a downweighting factor (0.8)
- Calculate sentence length differently
    - Originally used whitespace delimited sentence length
    - Now averages whitespace delimited sentence length and tokenized sentence length

INFORMATION ORDERING



1. For each pair of sentences (j, k), we computed the *similarity measure* b_{jk} to be the number of terms they have in common.

2. If the two sentences come from the same document, we multiplied b_{jk} by 1.6, to increase their similarity measure.

3. We then normalized these measures so the similarity between each sentence and itself is 1. We accomplish this by computing

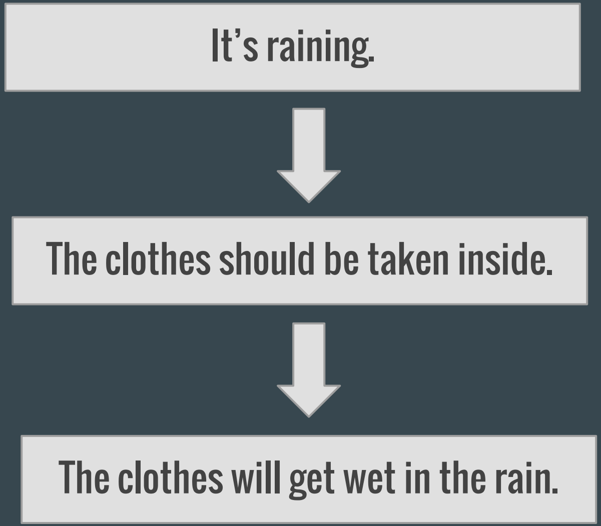
$$c_{jk} = \frac{b_{jk}}{\sqrt{b_{jj} b_{kk}}}.$$

4. The distance between sentence j and sentence k is then defined to be $-c_{jk}$, or, if you prefer nonnegative distances, $1 - c_{jk}$.

(Conroy et al, 2006)

INFORMATION ORDERING

- Precedence/Succession (Bollegala et al., 2012)
- Logical closeness (Zhu et al., 2012)



It’s raining.

The clothes should be taken inside.

The clothes will get wet in the rain.

INFORMATION ORDERING

SELECTED (A): There have been no arrests, although police have said JonBenet's parents, John and Patsy Ramsey, are under suspicion.

PRECEDING, ORIGINAL (B): There have been no arrests and authorities have said only that Patsy and John Ramsey are under suspicion.

SELECTED (C): The Ramseys have denied any involvement.

SYSTEM OUTPUT:

There have been no arrests, although police have said JonBenet's parents, John and Patsy Ramsey, are under suspicion.

The Ramseys have denied any involvement.

RESULTS

D3 - Average recall

ROUGE-1	0.29498
ROUGE-2	0.08520
ROUGE-3	0.03001
ROUGE-4	0.01209

D2- Average recall

ROUGE-1	0.27697
ROUGE-2	0.07920
ROUGE-3	0.02732
ROUGE-4	0.01145

ISSUES AND SUCCESSES

A judge ordered four police officers Wednesday to stand trial for the fatal shooting of an unarmed West African immigrant.
Diallo was hit 19 times.
The four officers fired 41 shots, hitting Diallo 19 times.
Officers Kenneth Boss, Sean Carroll, Edward McMellon and Richard Murphy left the courthouse without comment.
McMellon reportedly slipped and fell as the officers confronted Diallo.
Officers Kenneth Boss, Sean Carroll, Edward McMellon and Richard Murphy pleaded innocent in a Bronx courtroom to second-degree murder.
My client is innocent of all charges.
The officers in the Diallo case did not testify before the grand jury.

ISSUES AND SUCCESSES

A tsunami spawned by a 7.0 magnitude earthquake crashed into Papua New Guinea's north coast, crushing villages and leaving hundreds missing, officials said Sunday.

Australia will provide transport for relief supplies and a mobile hospital to Papua New Guinea (PNG) following Friday's tsunami tragedy.

A 10-meter tsunami engulfed the heavily populated villages near Aitape, 800 km north of PNG's capital city of Port Moresby.

Dalle said the Nimas village near the Sissano lagoon, the Warapu village and the Arop village had been wiped out and the Malol village had almost been completely destroyed.

Thirty people were confirmed dead.

FUTURE WORK

- Sentence simplification (Done in SumFocus)
- Stemming
- POS tagging?
- Generalizability

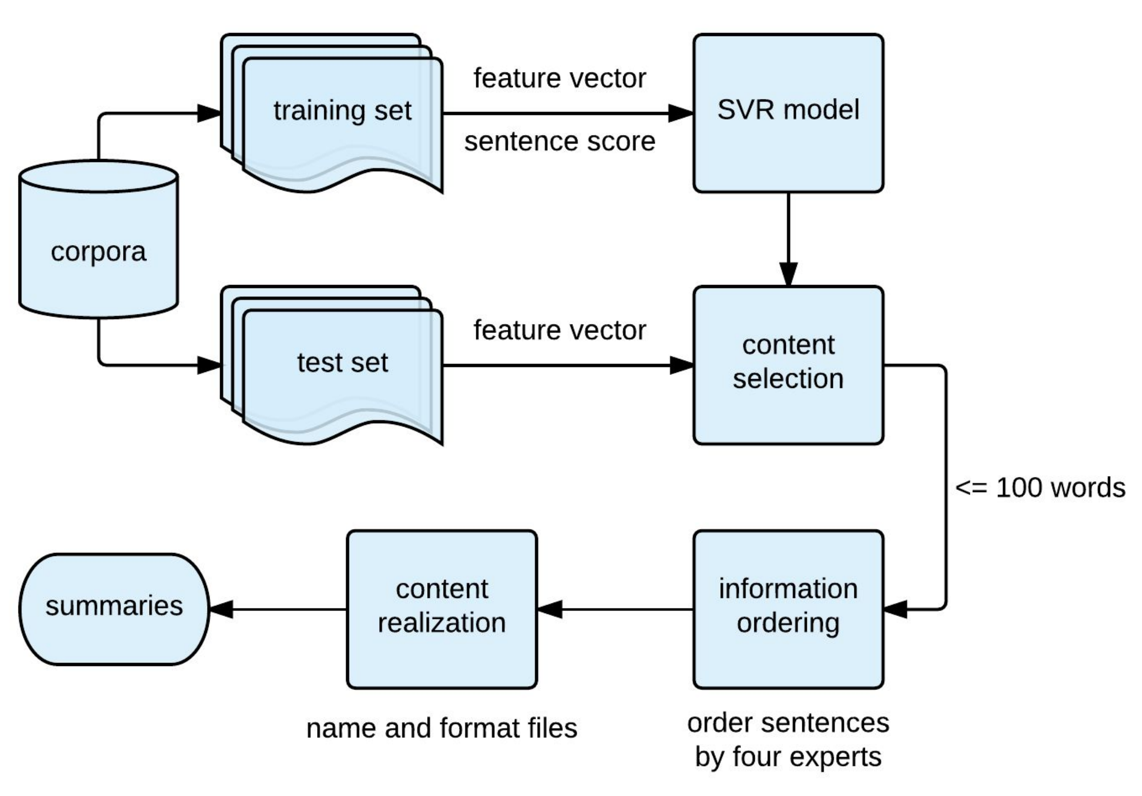
REFERENCES

- Bollegala, D., Okazaki, N., & Ishizuka, M. (2012). A preference learning approach to sentence ordering for multi-document summarization. Information Sciences, 217, 78-95. doi:10.1016/j.ins.2012.06.015
- Conroy, J. M., Schlesinger, J. D., O’Leary, D. P., & Goldstein, J. (2006). Back to Basics: CLASSY 2006.
- Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out: Proceedings of the ACL-04 workshop. Vol. 8. 2004.
- Vanderwende, Lucy, et al. "Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion." Information Processing & Management 43.6 (2007): 1606-1618.
- Zhu, Tiedan and Zhao, Xinxin. (2012). “An Improved Approach to Sentence Ordering For Multi-document Summarization.” In Proceedings of 2012 4th International Conference on Machine Learning and Computing.

Ling573 Project D3 System

Xiaosu Xue   Yveline Van Anh   Alex Cabral

System Architecture



Content Selection

- **Based on the SIEL algorithm**: iiit hyderabad at tac 2009
- **Training set:** TAC 2009 Update Summarization task data -- docset A
- **Test set:** TAC 2010 Guided Summarization task data -- docset A
- **Approach:** extract sentences with the highest predicted scores given by the SVR model (RBF kernel)
- **Avoid redundancy:**
  - cosine similarity: threshold 0.7

Content Selection (cont.)

- **Features:**
  - sentence position: 1-n/1000 if n <=3; n/1000 otherwise
  - query score
  - document frequency score
  - Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

- **Sentence score:** sentence-level ROUGE-2 precision score

Content Selection - features

Feature Name	ROUGE-1	ROUGE-2
sentence position	0.20607	0.05159
query score	0.21106	0.05505
document frequency score	0.20442	0.05675
KLD	0.17942	0.04431

Content Selection - output

	Mad Cow Disease
D3	The human form of mad cow disease is called variant Creutzfeldt-Jakob. It is the second case since March in which the disease, also known as bovine spongiform encephalopathy, or BSE, has been confirmed in a cow that died rather than having been slaughtered, the ministry said. However, Chen said, if there is any doubt over the quality of the beef, the ban will not be lifted at that time. Mad cow disease, or bovine spongiform encephalopathy, eats holes in the brains of cattle.
D2	Department of Health officials said Friday that there is no timetable for reintroducing the importation of U.S. beef to Taiwan after America was declared an area affected by mad cow disease late last year. (sentence #1) Canada, whose exports of beef products are affected by a single case of mad cow disease since may 2003, has exceeded its mad cow testing target for 2004, the Canadian Food Inspection Agency reported Sunday. (sentence #1)

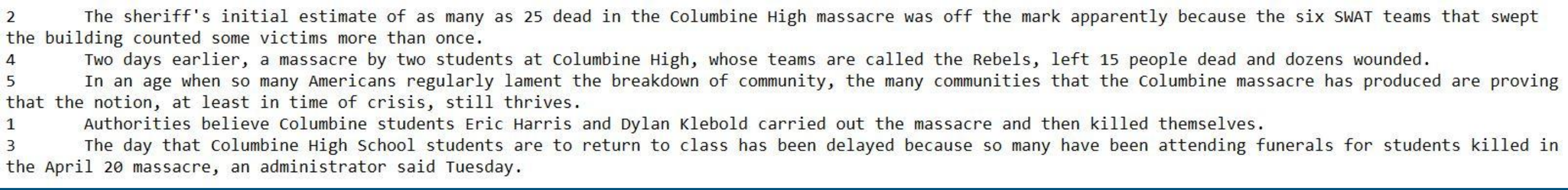
Information Ordering

- Sentences ordered by 4 experts in Bollegala et al.:
  - Chronological
  - Precedence
  - Succession
  - Topicality
- Removed probabilistic expert

- Output ordered sentences + rank from content selection portion

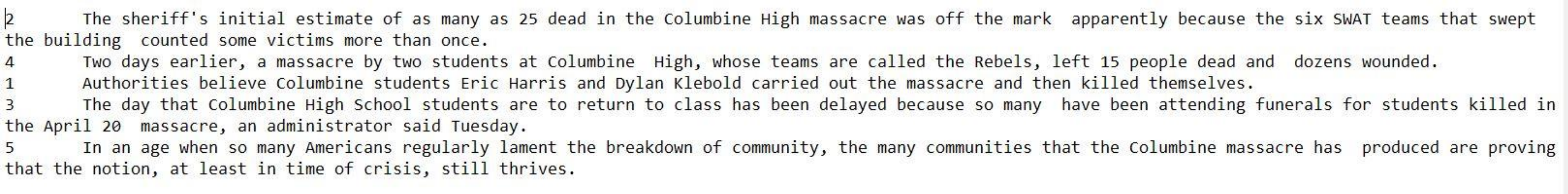
Information Ordering (Good Example)

- Chronological only:



```
2	The sheriff's initial estimate of as many as 25 dead in the Columbine High massacre was off the mark apparently because the six SWAT teams that swept
the building counted some victims more than once.
4	Two days earlier, a massacre by two students at Columbine High, whose teams are called the Rebels, left 15 people dead and dozens wounded.
5	In an age when so many Americans regularly lament the breakdown of community, the many communities that the Columbine massacre has produced are proving
that the notion, at least in time of crisis, still thrives.
1	Authorities believe Columbine students Eric Harris and Dylan Klebold carried out the massacre and then killed themselves.
3	The day that Columbine High School students are to return to class has been delayed because so many have been attending funerals for students killed in
the April 20 massacre, an administrator said Tuesday.
```

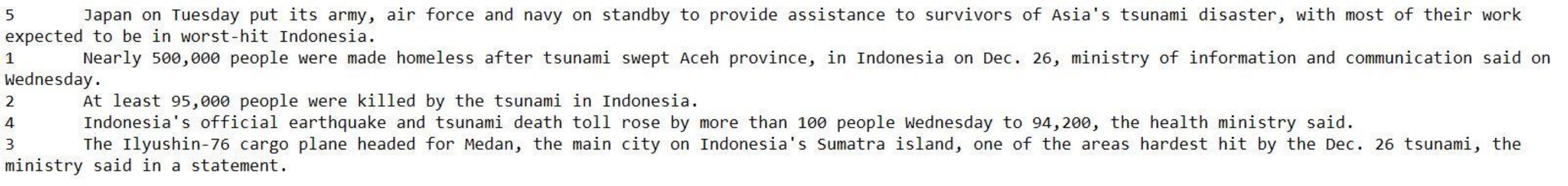
- Improved ordering:



```
2	The sheriff's initial estimate of as many as 25 dead in the Columbine High massacre was off the mark  apparently because the six SWAT teams that swept
the building  counted some victims more than once.
4	Two days earlier, a massacre by two students at Columbine  High, whose teams are called the Rebels, left 15 people dead and  dozens wounded.
1	Authorities believe Columbine students Eric Harris and Dylan Klebold carried out the massacre and then killed themselves.
3	The day that Columbine High School students are to return to class has been delayed because so many  have been attending funerals for students killed in
the April 20  massacre, an administrator said Tuesday.
5	In an age when so many Americans regularly lament the breakdown of community, the many communities that the Columbine massacre has  produced are proving
that the notion, at least in time of crisis, still thrives.
```

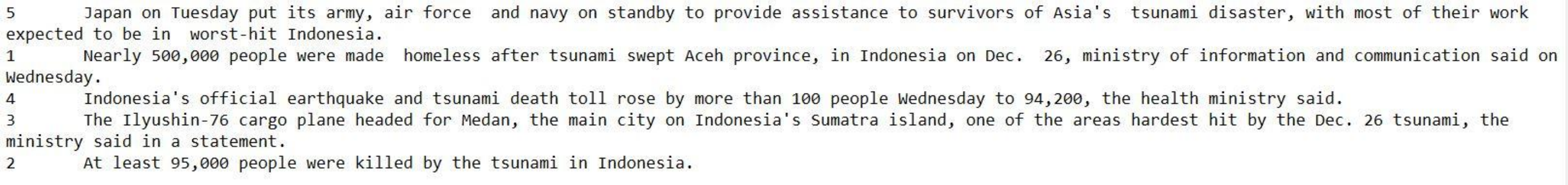
Information Ordering (Bad Example)

- Chronological only:



```
5	Japan on Tuesday put its army, air force and navy on standby to provide assistance to survivors of Asia's tsunami disaster, with most of their work
expected to be in worst-hit Indonesia.
1	Nearly 500,000 people were made homeless after tsunami swept Aceh province, in Indonesia on Dec. 26, ministry of information and communication said on
Wednesday.
2	At least 95,000 people were killed by the tsunami in Indonesia.
4	Indonesia's official earthquake and tsunami death toll rose by more than 100 people Wednesday to 94,200, the health ministry said.
3	The Ilyushin-76 cargo plane headed for Medan, the main city on Indonesia's Sumatra island, one of the areas hardest hit by the Dec. 26 tsunami, the
ministry said in a statement.
```

- Improved ordering:



```
5	Japan on Tuesday put its army, air force  and navy on standby to provide assistance to survivors of Asia's  tsunami disaster, with most of their work
expected to be in  worst-hit Indonesia.
1	Nearly 500,000 people were made  homeless after tsunami swept Aceh province, in Indonesia on Dec.  26, ministry of information and communication said on
Wednesday.
4	Indonesia's official earthquake and tsunami death toll rose by more than 100 people Wednesday to 94,200, the health ministry said.
3	The Ilyushin-76 cargo plane headed for Medan, the main city on Indonesia's Sumatra island, one of the areas hardest hit by the Dec. 26 tsunami, the
ministry said in a statement.
2	At least 95,000 people were killed by the tsunami in Indonesia.
```

Results

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
RANDOM	0.14563	0.02488	0.00557	0.00113
FIRST	0.18883	0.04752	0.01592	0.00586
MEAD (baseline)	0.22437	0.06144	0.01889	0.00668
SIEL (improved)	**0.24145**	**0.07059**	**0.02700**	**0.01299**

Content Realization

- Additional formatting of sentences
- Removal of temporal words

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
SIEL (improved)	**0.24145**	**0.07059**	**0.02700**	**0.01299**
SIEL with cont. realization	**0.23894**	**0.06908**	**0.02590**	**0.01158**

On Dec. 14 last year, Feng Shiliang, a farmer from Youfangzui Village, told the Fengxian County Wildlife Management Station that he had spotted an animal that looked very much like a giant panda and had seen giant panda dung while collecting bamboo leaves on a local mountain.

On, Feng Shiliang, a farmer from Youfangzui Village, told the Fengxian County Wildlife Management Station that he had spotted an animal that looked very much like a giant panda and had seen giant panda dung while collecting bamboo leaves on a local mountain.

Discussion

- Improved SIEL system performed better than our baseline
- Shorter sentences output by content selection
- Readability seemed to be improved by our information ordering
- First sentence of summary was always the same for original and improved ordering
    - Only expert to be considered at that time is that of chronology
- Improved content realization efforts actually hurt ROGUE scores
    - Not removing entire phrases

Future Work

- Perform more pruning in content realization
    - Remove preceding adjuncts
    - Remove ‘unnecessary’ clauses
    - Remove PPs without named entities
- Experiment with pruning sentences before content selection vs. after information ordering

Reference

Bollegala, Danushka, Naoaki Okazaki, and Mitsuru Ishizuka. "A preference learning approach to sentence ordering for multi-document summarization." *Information Sciences* 217 (2012): 78-95.

Varma, V., Bysani, P., Kranthi Reddy, V. B., Santosh GSK, K. K., Kovelamudi, S., Kiran Kumar, N., & Maganti, N. (2009, November). iiit hyderabad at tac 2009. In Proceedings of Test Analysis Conference 2009 (TAC 09).

Radev, D. R., Blair-Goldensohn, S., & Zhang, Z. (2001). Experiments in single and multi-document summarization using MEAD. Ann Arbor, 1001, 48109.

Radev, D. R., Jing, H., Styś, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing & Management, 40(6), 919-938.

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop (Vol. 8).

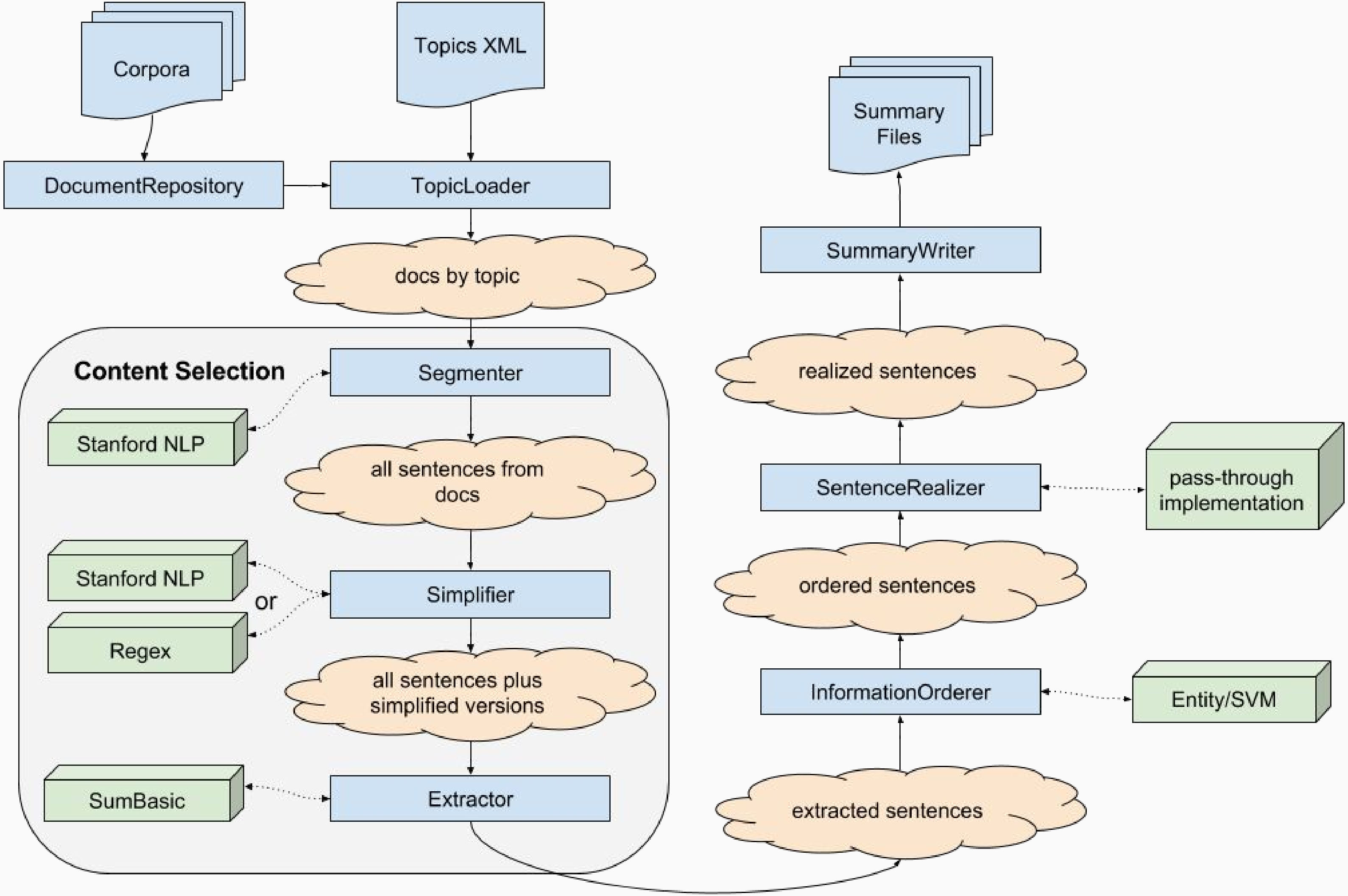
Thank you!

Summarization System Improvements

Alex Burrell, Robert Gale, and Chris LaTerza

Improvements for Deliverable #3

- Regex-based sentence simplification
- Sentence selection/extraction
- Sentence ordering
- Bug fixes in our scoring code!


Corpora
Topics XML
Summary Files
DocumentRepository
TopicLoader
SummaryWriter
docs by topic
Content Selection
Segmenter
Stanford NLP
all sentences from docs
Stanford NLP
or
Regex
Simplifier
all sentences plus simplified versions
SumBasic
Extractor
realized sentences
SentenceRealizer
pass-through implementation
ordered sentences
InformationOrderer
Entity/SVM
extracted sentences

Regex-based simplification

1. Remove newspaper-style headings (e.g. SHANGHAI JULY 20 --)
2. Remove all content between dashes
3. Remove all content in parentheses
4. Then split by commas...

Regex-based simplification

For each comma-separated clause, remove if it starts with

1. A cardinal number
2. A preposition
3. An adverb
4. A gerund verb

Sentence extraction

- Still using SumBasic as our baseline
- Began implementing a version of FastSum
- Challenges: general feature extraction, working with libSVM, choosing a way to handle redundancy

FastSum overview

- Training sentences receive a score based on word overlap with gold-standard summaries
- Sentence features include (a subset of) unigram features like content word frequency, document frequency, and topic title occurrence as well as sentence length, sentence position.
- Uses Support Vector Regression

Sentence ordering

- Barzilay & Lapata 2005/2008
- Our implementation: libsvm, linear kernel, default parameters
- Only “quick-and-dirty” versions of coref and salience processing so far
- For test example (“Microsoft” sentences), we got a 144-way tie out of 720 permutations! (Probably could be fixed with better syntax/salience)

Results

	Average-R score
ROUGE-1	0.23817
ROUGE-2	0.06159
ROUGE-3	0.01978
ROUGE-4	0.00759

Example summary

- The second problem is the nation's almost exclusive reliance on drug companies to police the safety and efficacy of their own drugs.
- These studies were designed to answer questions about cardiovascular risk raised by earlier less conclusive research.
- So why is Merck recalling the drug now?
- The FDA said last week after Vioxx was withdrawn that the problems were unique to that drug.
- The results of that study came on the heels of an earlier study that showed a greater number of heart attacks in patients taking Vioxx, although there were fewer stomach ulcers and bleeding.
- Did the 20 million Americans who used the drug since its launch in 1999 really have to spend that extra money and, incur a slight extra risk?

Example summary

- England became pregnant while in Iraq.
- England was one of seven members of the 372nd Military Police Company charged with humiliating and assaulting prisoners at Abu Ghraib.
- No senior officer at the prison, and no one higher up in the chain of command, has faced charges in the case.
- The Abu Ghraib abuse scandal went public in April 2004, after photographs showing American soldiers mistreating and humiliating Iraqi prisoners surfaced.
- Lynndie England on Monday pled guilty to charges of abusing Iraqi prisoners at the Abu Ghraib prison but told a court martial she did not believe she was doing wrong when photographed holding a leash on a naked inmate.

Our highest scoring summary

- The 14th case in Japan was confirmed last week.
- The human form of mad cow disease is called variant Creutzfeldt-Jakob.
- The fatal brain-wasting disease is believed to come from eating beef products from cows struck with mad cow disease.
- public health authorities have warned that people may catch vCJD from eating meat infected with mad cow disease, known as bovine spongiform encephalopathy, or from infected blood transfusions.
- Emory University Hospital has confirmed that a brain surgery patient does not have the human version of mad cow disease, but does have a rare, fatal disorder that resembles it.
- The target was 8,000 cattle tested by the end of 2004.

Our lowest scoring summary

- ASEAN Transport Ministers issued a ministerial declaration on ASEAN road safety Tuesday to enhance the road safety and reduce the traffic casualties in member countries.
- Federation officials compare the use of soccer headgear, which lack an industry safety standard, to the largely unregulated business of nutritional supplements.
- WHO calls on China to lower 680-a-day road accident death toll by Robert J. Saiget ATTENTION- INSERTS details, ADDS quotes / / / ss problems with the way transportation is organized, factors contributing to accidents, the need to create better safety devices for vehicles and passengers and to build a better mechanism to respond to accidents.

Next steps

- FastSum for extraction
- Improve sentence ordering (syntax/salience)
- Content realization (post-processing sentences after ranking)

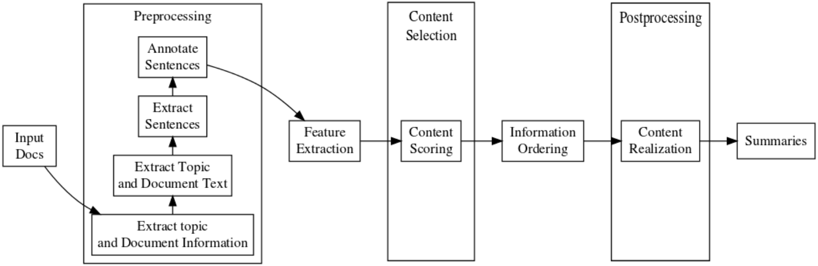
Thanks!

D3: Automatic Summarization with Neural Networks

Tony Princing and Ernie Chang and Jason Blum

May 19, 2016

System Architecture



Information Ordering

- Conceptually applies principles of single document summarization to multi-document summarization
  - Order by salience and then by position
  - Two ordering passes
- All topic sentences sorted first by saliency score
- Salience summary built from saliency sorted sentences limited by compression value (max sentences parameter) and redundancy threshold parameter
- This first pass has not changed for D3

Information Ordering

- Improved for D3, our position ordering (2nd pass) now uses more information from the input documents
- Inspired by Barzilay et. al., 2002 Majority Ordering
- Each sentence in salience summary is considered a theme
- Sentences left out of salience summary are clustered to these theme sentences
- Cluster members then use their document positions to vote on summary precedence between pairs of themes (i.e. salience summary sentences)

Information Ordering

- Overall votes determine path score between theme pairs
- Best (max) path through salience summary is then determined producing ordered summary
- If length of salience summary prevents exhaustive path calculation then a sliding lookahead window is used
- Exhaustive search within window
- Parameter setting for window size to keep computationally tractable
- Fixed starting point for window and only top new sentence is kept for each sliding window ordering

Content Realization

- Creates final summary from position summary
- Starting with top-ranked sentences adds sentences to final summary if the addition will not cause the final summary to exceed the summary word limit
- Attempts to add all position summary sentences to final summary. Potential to have a lower scoring, but short sentence added to final summary – because it fits
- New for D3, the final summary is re-ordered using cosine similarity on 3 by 4 skipgrams (tri-grams, 4 word skips) to improve coherence
- Again, a sliding lookahead window is used if exhaustive best path calculation is not computationally tractable

Content Selection

Training Data

Model	Rouge-1	Rouge-2	Rouge-3	Rouge-4
LDA+ngram	0.31014	0.08566	0.02967	0.01295
SumCNN	0.23118	0.05905	0.01898	0.00797

Moving Forward

- NER
- Neural Attention Model

ROUGE Results

Name	Average R	CI Lower	CI Upper
ROUGE-1	0.07118	0.05750	0.08601
ROUGE-2	0.01484	0.01011	0.01998
ROUGE-3	0.00359	0.00146	0.00600
ROUGE-4	0.00046	0.00000	0.00103

Table : D3 ROUGE results table

Name	Average R	CI Lower	CI Upper
ROUGE-1	0.19325	0.17105	0.21344
ROUGE-2	0.04657	0.03734	0.05547
ROUGE-3	0.01423	0.00989	0.01895
ROUGE-4	0.00436	0.00214	0.00684

Table : D2 ROUGE results table

Problems With N-Gram (continued)

```
amphibian experience
scientist compare frog
first vertebrate species
almost species
Gerardo de la Cruz
one third
amphibian experience precipitous decline across globe
    accord first comprehensive world survey creature
    include frog toad salamander
small frog
year
facilitate
```

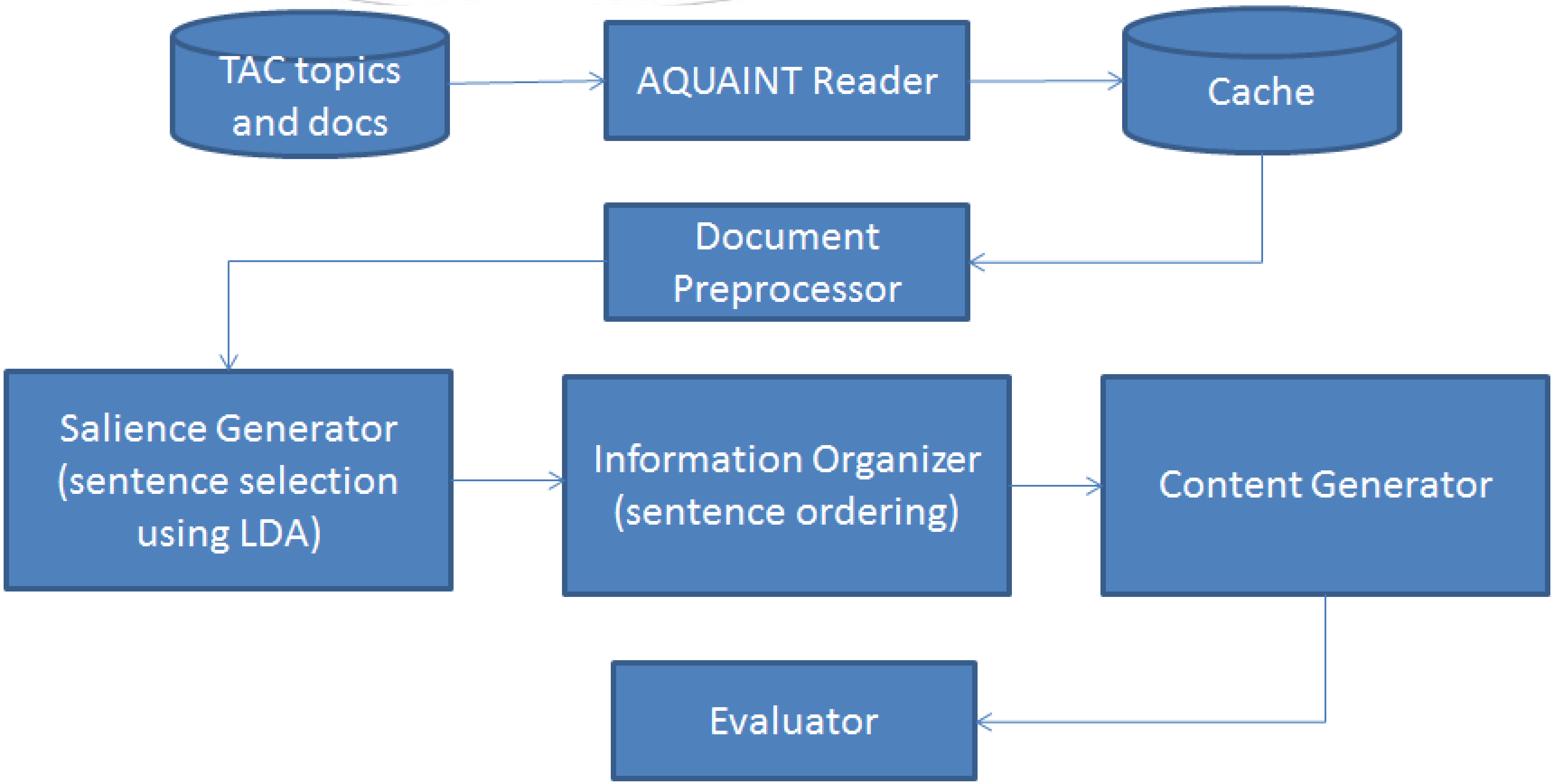
Results (Older Sentence Model with New Ordering)

- ROUGE-1 Average_R: 0.26900 (95%-conf.int. 0.24814 - 0.28852)
- ROUGE-2 Average_R: 0.06284 (95%-conf.int. 0.05342 - 0.07218)
- ROUGE-3 Average_R: 0.01992 (95%-conf.int. 0.01493 - 0.02567)
- ROUGE-4 Average_R: 0.00676 (95%-conf.int. 0.00361 - 0.01136)

MultiDocSummarizer

Kevin Wonus, Cade Bryant and Natalia Rodnova
Ling573-2016, UW

System Architecture



TAC topics and docs
AQUAINT Reader
Cache
Document Preprocessor
Salience Generator (sentence selection using LDA)
Information Organizer (sentence ordering)
Content Generator
Evaluator

Tools

- Python 3
- NLTK
- Gensim: “Topic modeling for humans” – by Radim Rehurek
  - Thoughtfully written
  - Well documented
  - Actively supported
  - Google forum
  - https://radimrehurek.com/gensim/

Approach

- Initial focus on making all pieces work together
- Select a well-known method as a base line, and later choose something more modern and less developed.
- Initially used LLR
- Choices: LSA -> pLSA -> LDA
- Winner: LDA

Latent Dirichlet Allocation

- First introduced by David Bleu, Andrew Ng and Michael Jordan in 2003. Paper is called “Latent Dirichlet Allocation”
- Algorithm used by gensim was created by Matthew Hoffman, David Bleu and Francis Bach in 2010. Paper is called “Online Learning for Latent Dirichlet Allocation”

Latent Dirichlet Allocation (cont'd)

- LDA represents documents as a mixture of topics that share words with certain probabilities
- It assumes that documents are written in the following fashion:
  - Choose number of words
  - Chose topic mixture (according to a Dirichlet distribution over a fixed set of K topics)
  - Generate each word by a) picking a topic and b) generate word using the topic (according to the topic's multinomial distribution)
- Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection.

Inspiration

- “Latent Dirichlet Allocation Based Multi-Document Summarization” by Rachit Arora and Balamaran Ravindran (2008). (*They also came up with the idea of using LDA + LSA combination.*)
- “Research On Multi-document Summarization Based On LDA Topic Model” by Jinqiang Bian, Zengru Jiang, Qian Chen (2014)
- “Comparative Summarization via Latent Dirichlet Allocation” by Michal Campr and Karel Jezek (2013)

Content Selection Using LDA

- Feed documents (related to a single TAC topic) to LDA model
- Get topic distribution and calculate topic probabilities
- For each sentence, calculate its probability to describe each topic
- For N most important topics, pick K most probable sentences

Results

	Our system	Peers (avg)	Peers (best)	Peers(worst)
ROUGE-1	0.15280	0.227089	0.30849	0.02188
ROUGE-2	0.03258	0.057298	0.08206	0.00470
ROUGE-3	0.00860	0.017914	0.03020	0.00135
ROUGE-4	0.00212	0.006188	0.01193	0.00019

Improvements

- Select optimal number of topics (using perplexity measure)
- Eliminate redundant sentences (using a similarity measure)
- Take into account sentence length
- Train LDA on a huge corpus with a lot of topics and then get the document distribution over those topics
- Combine LDA with LSA: first, run LDA model to get topics, then use SVD on each topic

Improvements in D3

Sentence Length

- **Sentences too long for effective ordering**
  - Therefore, split sentences based on:
    - Transition words *(and, or, although….)*
    - Keep split if both halves grammatical
    - Recurse as needed
  - Implemented in */D3/src/Preproc/Segmenter.cs*
  - Utilizes ERG/LOGON
  - Code communicates with service via */D3/src/Preproc/Poster.cs*
- **Order the resulting sentences (see next slide)**

Improvements in D3

Information Ordering

Chronological Ordering

- Based on publication date of document in corpus.
- Implemented in */D3/src/Ordering/ChronOrder.py*

Augmented Ordering (per *Barzilay et al, 2001*)

- Based on per-segment ratio of:
  - Count(themed sentence pairs in same document and segment)
  - Count(themed sentence pairs in same document)
  - Theme parsing discussed in next slide
- Keep pair if ratio >= predetermined threshold
  - 0.6 per *Barzilay*
- Implemented in */D3/src/Ordering/OrdAugmenter.py*

Improvements in D3

Topic Orientation

Theme-based Approach (per *Barzilay et al, 2001*)

- Sentences make up a *theme* if their content is similar
- Used Cosine Distance to determine similarity
- Additional code to remove stopwords/punctuation and vectorize sentences
- Implemented in */D3/src/ThemeBuilder/ThemeBuilder.py*

Conclusions for D3?

Sadly, personal emergencies on behalf of team members inhibited our testing efforts. The code has not been tested on the corpus, and the new portions of the code are not yet successfully integrated with each other.

Further Work

- To be discussed with team when we regroup