D2 - Multi-Document Summarization

Maria Sumner, Micaela Tolliver, Elizabeth Cary

GOAL / MOTIVATION

- Implement a simple base system
- FREQUENCY
- Luhn (1958), Nenkova & Vanderwende (2005)
- “the high frequency words from the input are very likely to appear in the human models”
- Role in LexRank and MEAD
- SumBasic - based on frequency and accounts for redundancy

SUMBASIC

Step 1 Compute the probability distribution over the words w_i appearing in the input, $p(w_i)$ for every i; $p(w_i) = \frac{n}{N}$, where n is the number of times the word appeared in the input, and N is the total number of content word tokens in the input.

Step 2 For each sentence S_j in the input, assign a weight equal to the average probability of the words in the sentence, i.e.,

$$Weight(S_j) = \sum_{w_i \in S_j} \frac{p(w_i)}{|\{w_i | w_i \in S_j\}|}$$

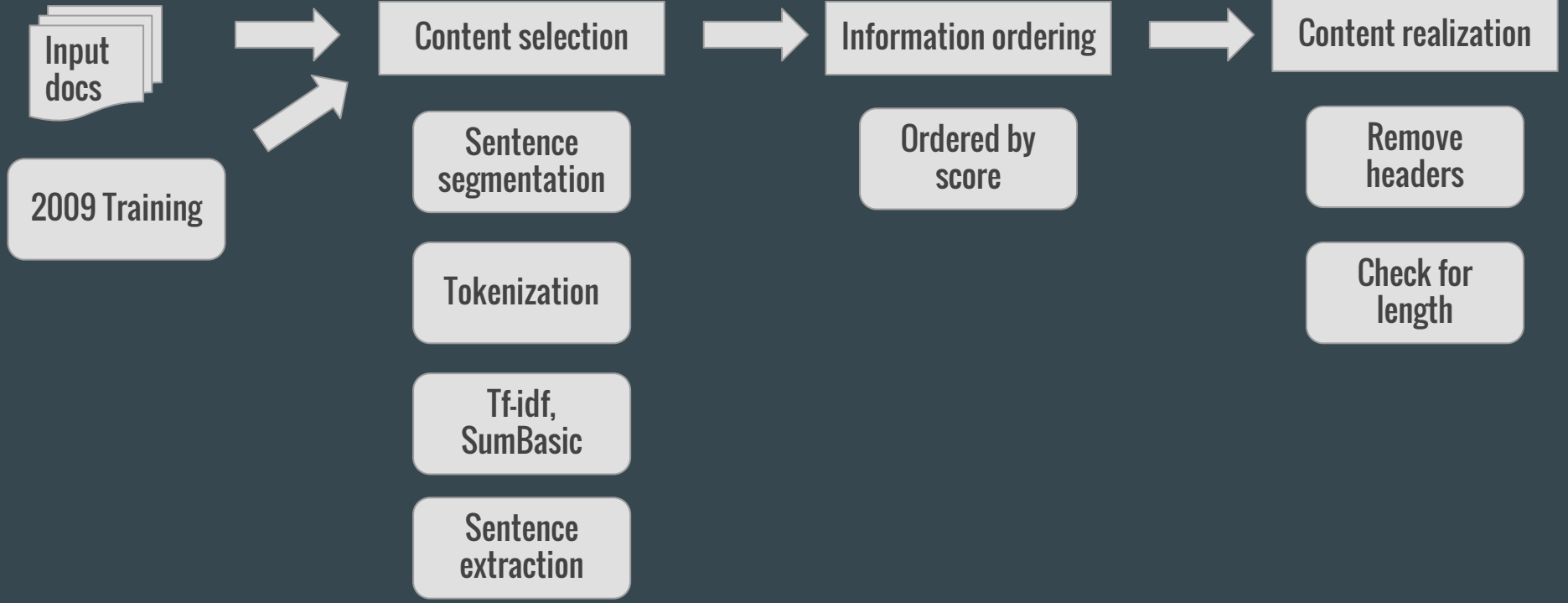
Step 3 Pick the best scoring sentence that contains the highest probability word.

Step 4 For each word w_i in the sentence chosen at step 3, update their probability:

$$p_{new}(w_i) = p_{old}(w_i) \cdot p_{old}(w_i)$$

Step 5 If the desired summary length has not been reached, go back to Step 2

SYSTEM ARCHITECTURE



TF-IDF

- Emphasises salient words for each document cluster
- Calculated TF-IDF for the cluster compared to other document clusters
- Utilized TF-IDF values in a similar algorithm to SumBasic
    - Cut off heading information for each sentence
    - Calculated a score for each sentence based off of the sum of TF-IDF scores for tokenized words
    - Normalize this score by the sentence length
    - After selecting a sentence, down-weight the TF-IDF scores of all the tokenized words in the sentence
- Fill the summary until it hits 100 words

RESULTS

Average recall, with 2009 training data

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

RESULTS

Average recall, with 2009 training data*

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

Average recall, with 2010 training data

ROUGE-1	0.28013
ROUGE-2	0.07950
ROUGE-3	0.02811
ROUGE-4	0.01163

*Denotes the current system

ISSUES AND SUCCESSES

- Issues
  - Inclusion of contact information, including phone numbers, URLs, and email addresses
  - Presence of irrelevant attributives, unresolved referents, questions, incomplete quotes
- Successes
  - Removal of sentences under 5 words eliminated uninformative sentences such as exclamations: “Avalanche!”
  - Downweighting has reduced redundancy
    - Without downweighting
      - Diallo was hit 19 times.
      - The four officers fired 41 shots, hitting Diallo 19 times.
    - With downweighting
      - The four officers fired 41 shots, hitting Diallo 19 times.

SAMPLE SUMMARIES

By contrast, Vioxx made $2.5 billion for Merck last year.

__ On the Net: FDA: http://www.fda.gov/

So why is Merck recalling the drug now?

FDA urged to weigh in Vioxx, Celebrex and Bextra are the only three drugs in a class known as Cox-2 inhibitors.

(On Friday, Pfizer Inc. issued a warning that its Cox-2 drug Bextra may increase cardiovascular risk for some patients.)

The FDA's own study of the Vioxx safety issue has become mired in controversy.

FitzGerald also challenged Pfizer's contention that no science shows increased risk from Celebrex.

The community outpouring has touched some Columbine students.

Denver's newscasters have donned blue Columbine ribbons.

Students returned to classes Thursday at Chatfield High School, but the bloodbath at rival Columbine High haunted the halls.

in Jonesboro, Ark., scene of an earlier school shooting, reach out to those in Littleton, Colo.

Authorities believe Columbine students Eric Harris and Dylan Klebold carried out the massacre and then killed themselves.

Wells, a 16-year-old catcher on Columbine's varsity baseball team, watched the junior varsity play Arvada West High School on Wednesday.

FUTURE WORK

- Further combine elements of SumBasic and tf*idf
- Remove stopwords
- Sentence simplification
- Get closer to 100 words
- Optimize choice of downweighting factor

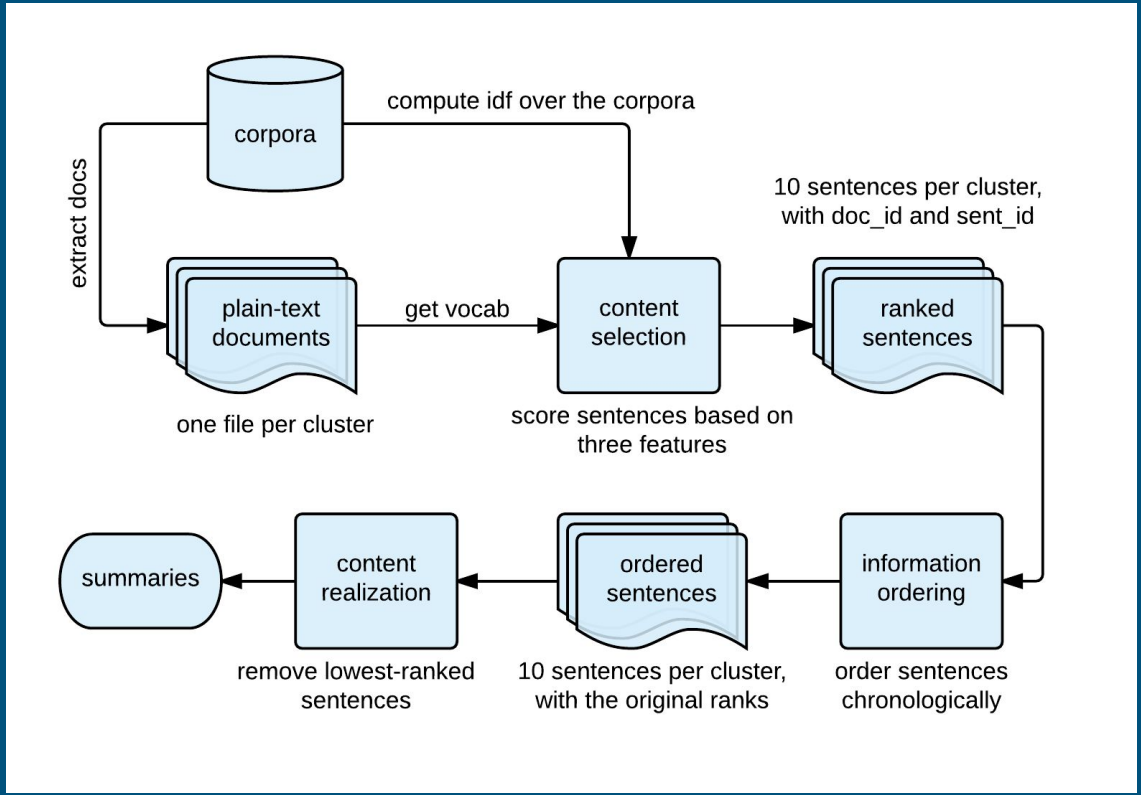
REFERENCES

- Daumé III, H., and D. Marcu. 2005. Bayesian Multi-Document Summarization at MSE. In Proceedings of MSE 2005.
- Jones, Karen Spärck. "Automatic summarising: The state of the art." Information Processing & Management 43.6 (2007): 1449-1481.
- Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out: Proceedings of the ACL-04 workshop. Vol. 8. 2004.
- Luhn, Hans Peter. (1958). The automatic creation of literature abstracts. IBM Journal.
- Nenkova, Ari and Lucy Vanderwende. 2005. The impact of frequency on summarization. Technical report, Microsoft Research.
- Rajaraman, A.; Ullman, J. D. (2011). "Data Mining". Mining of Massive Datasets (PDF). pp. 1–17.
- Vanderwende, Lucy, et al. "Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion." Information Processing & Management 43.6 (2007): 1606-1618.

Ling573 Project Baseline System

Xiaosu Xue Yveline Van Anh Alex Cabral

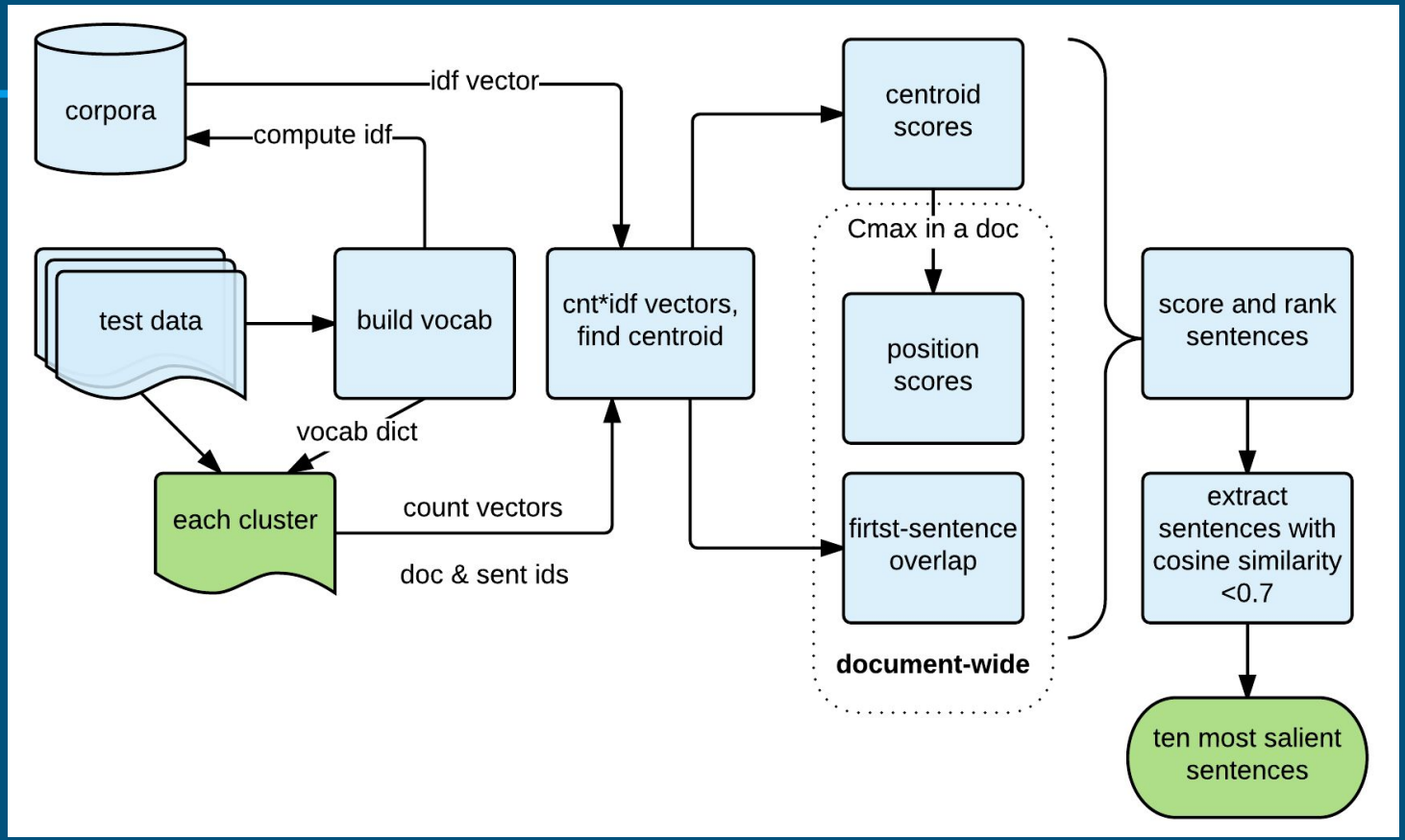
System Architecture



Content Selection

- **Based on the MEAD algorithm**: Radev, D. R., Jing, H., Styś, M., & Tam, D. (2004)
- **Goal**: extract the ten most salient sentences from a document set
- **Saliency**:
  - centroid score: the sum of centroid values in a sentence
  - position score: $P = (n-i+1)/n*Cmax$
  - first-sentence overlap: the inner product of sentence vectors
- **Avoid redundancy:**
  - cosine similarity: threshold 0.7

Content Selection

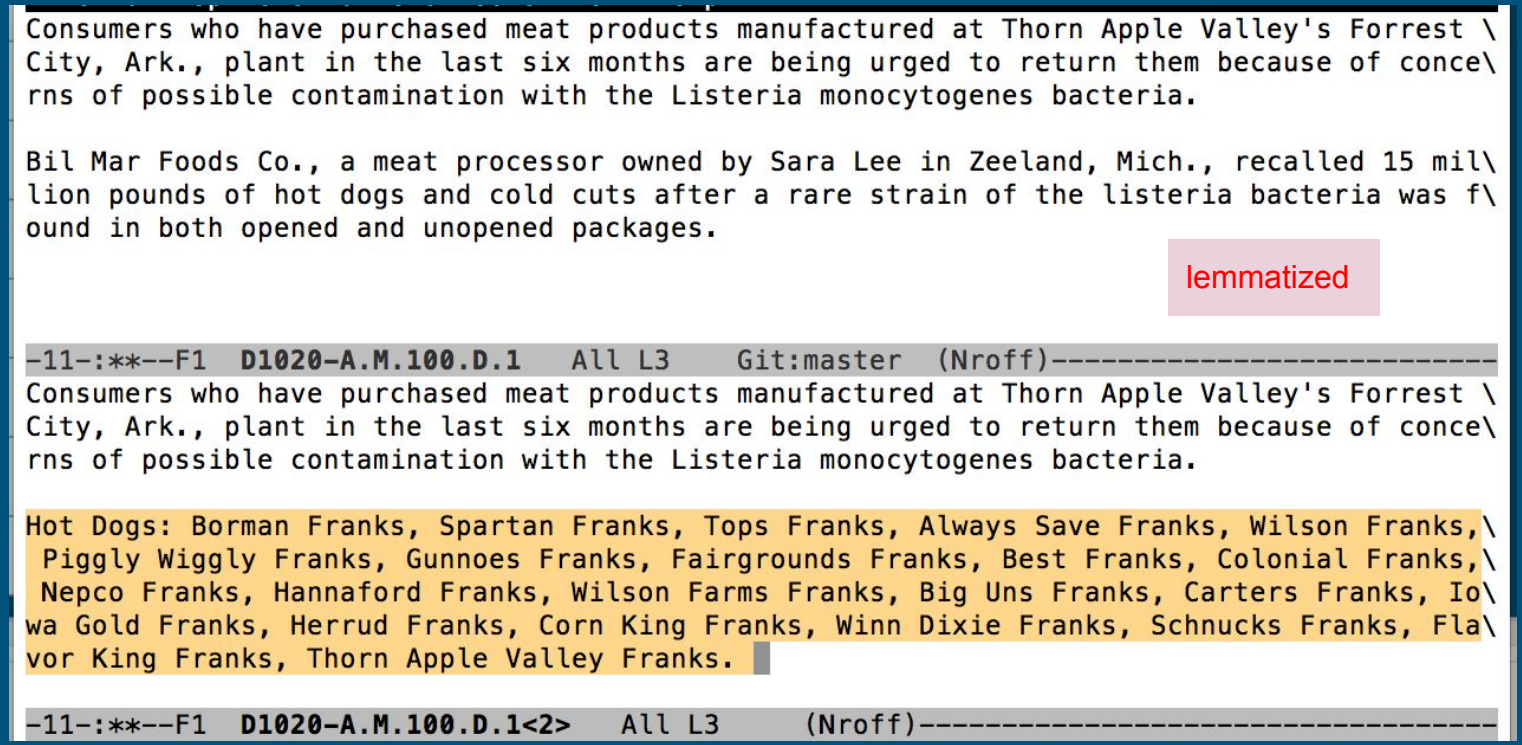


Content Selection - the effect of lemmatization

centroid(raw)	tc-idf
listeria	162.2537
bil	79.9813
recall	75.5636
franks	72.4694
listeriosis	55.6909
food-borne	55.0369
food	54.7183
mar	52.9917
meats	52.9641
dogs	51.2142

centroid(lemma)	tc-idf
listeria	162.2537
meat	82.8555
recall	80.4009
bil	79.9813
food	74.5046
listeriosis	55.6909
food-borne	55.0369
cheese	54.5179
bacteria	49.6131
outbreak	45.1247

Content Selection - the effect of lemmatization (cont.)



Content Selection - normalization of feature scores?

APW19990123.0111_1 Consumers who have purchased meat products manufactured at Thorn Apple Valley's Forrest City, Ark., plant in the last six months are being urged to return them because of concerns of possible contamination with the Listeria monocytogenes bacteria.

sentence score	centroid score	position score	first-sent overlap
1924.8574	162.2537	1645.5405	117.0632

cluster-wide mean and max feature scores

C mean	C max	P mean	P max	F mean	F max
79.0346	1645.5405	240.9706	1645.5405	5.9757	117.0632

Information Ordering

- Sentences output in chronological order
  - Date and time
  - Order within article
- Output sentence + rank from content selection portion
- Issues & future directions
  - Some sentences from later articles should be earlier in the summary
  - Chronological ordering combined with methods to increase coherence
    - Cosine similarity for adjacent sentences
    - Probabilistic component

Content Realization

- Limited the summary to 100 words
- Output sentences in same order as input
- Attempted to remove ‘unnecessary’ parts of speech, but ran into issues
  - Readability severely decreased
  - Not as straightforward as it seems
- Next Steps
  - Co-references
  - Eliminating quotations

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**

Discussion

- Baseline performed better than random and first sentence, but still not as well as we would like
- Hoping that further work on information ordering and content realization will improve results
  - Shorter, pruned sentences
  - More sentences included in final summary
  - More summary-like in nature
- Results were slightly improved after testing different weights, but rank of sentences changed, and seemingly not always for the better
- Further work is needed for content selection

Original Weighting-normalized (3,2,1)

1 China's State Environmental Protection Administration (SEPA) said here on Wednesday that the Songhua River in northeast China suffered a major water pollution incident owing to the explosion of a petrochemical plant at the upper reaches.

3 "After the explosion at the Jilin Petrochemical Company under China National Petroleum Corporation, our observation showed pollutants containing benzene had flown into the Songhua River and caused water pollution," said an official with SEPA.

4 The massive floating pollutants traversing the Harbin section of northeast China's Songhua River is unlikely to cause big troubles to the city's rural areas, according to a senior hydrologist.

2 Two reservoirs on the Songhua River on Thursday reinforced water discharge to dilute a massive slick of chemical pollutants floating through this capital of northeast China's Heilongjiang.

5 He said the polluted water in the Songhua River, which reached Heilongjiang provincial capital Harbin on Thursday morning, is expected to flow into the Heilongjiang River (called the Amur River in Russia) on the Sino-Russian border in around 14 days judging from the current flow speed.

New Weighting-raw (1,1,1)

3 China's State Environmental Protection Administration (SEPA) said here on Wednesday that the Songhua River in northeast China suffered a major water pollution incident owing to the explosion of a petrochemical plant at the upper reaches.

2 "After the explosion at the Jilin Petrochemical Company under China National Petroleum Corporation, our observation showed pollutants containing benzene had flown into the Songhua River and caused water pollution," said an official with SEPA.

5 The front of the polluted water of Songhua River in northeast China reached Harbin, capital of Heilongjiang Province, on early Thursday morning, local environment authority said.

4 China pledged on Thursday it will try its utmost to minimize the impact of northeast China's Songhua River pollution on the neighboring Russia, including intensifying monitoring and water quality control measures.

1 He said the polluted water in the Songhua River, which reached Heilongjiang provincial capital Harbin on Thursday morning, is expected to flow into the Heilongjiang River (called the Amur River in Russia) on the Sino-Russian border in around 14 days judging from the current flow speed.

System Output

"After the explosion at the Jilin Petrochemical Company under China National Petroleum Corporation, our observation showed pollutants containing benzene had flown into the Songhua River and caused water pollution," said an official with SEPA.

He said the polluted water in the Songhua River, which reached Heilongjiang provincial capital Harbin on Thursday morning, is expected to flow into the Heilongjiang River (called the Amur River in Russia) on the Sino-Russian border in around 14 days judging from the current flow speed.

Model File

On November 13, 2005, an explosion at the Jilin Petrochemical Company released 100 tons of pollutants, including high-density benzene and nitrobenzene, into the Songhua River in NE China, to create a 50-mile slick of floating chemicals.

Water quality was monitored constantly and active carbon added.

Patrols enforced non-use by humans or animals.

China kept Russia informed since the Songhua joins the Amur border river.

The Songhua supplies drinking water for Harbin, a city of over 3 million.

Harbin cut off its water supply for four days.

Pollutant density declined with sedimentation, adsorbtion, and dilution as tributaries and reservoirs added water.

Reference

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!

Base End-to-End Summarization System

Alex Burrell, Robert Gale, and Chris LaTerza

System Approach

- XML parsing
- Caching
- Parallel processing
- Inversion of Control pattern
- Extract-based

Caching

- Made use of Java generics to allow for strongly typed, reusable caching

```
public interface Cacher {
        <TCachedItem extends Serializable> TCachedItem getFromCache(
                CacheKey cacheKey, Supplier<TCachedItem> getter);
}
```

- Lambdas kept cache retrieval syntax slim

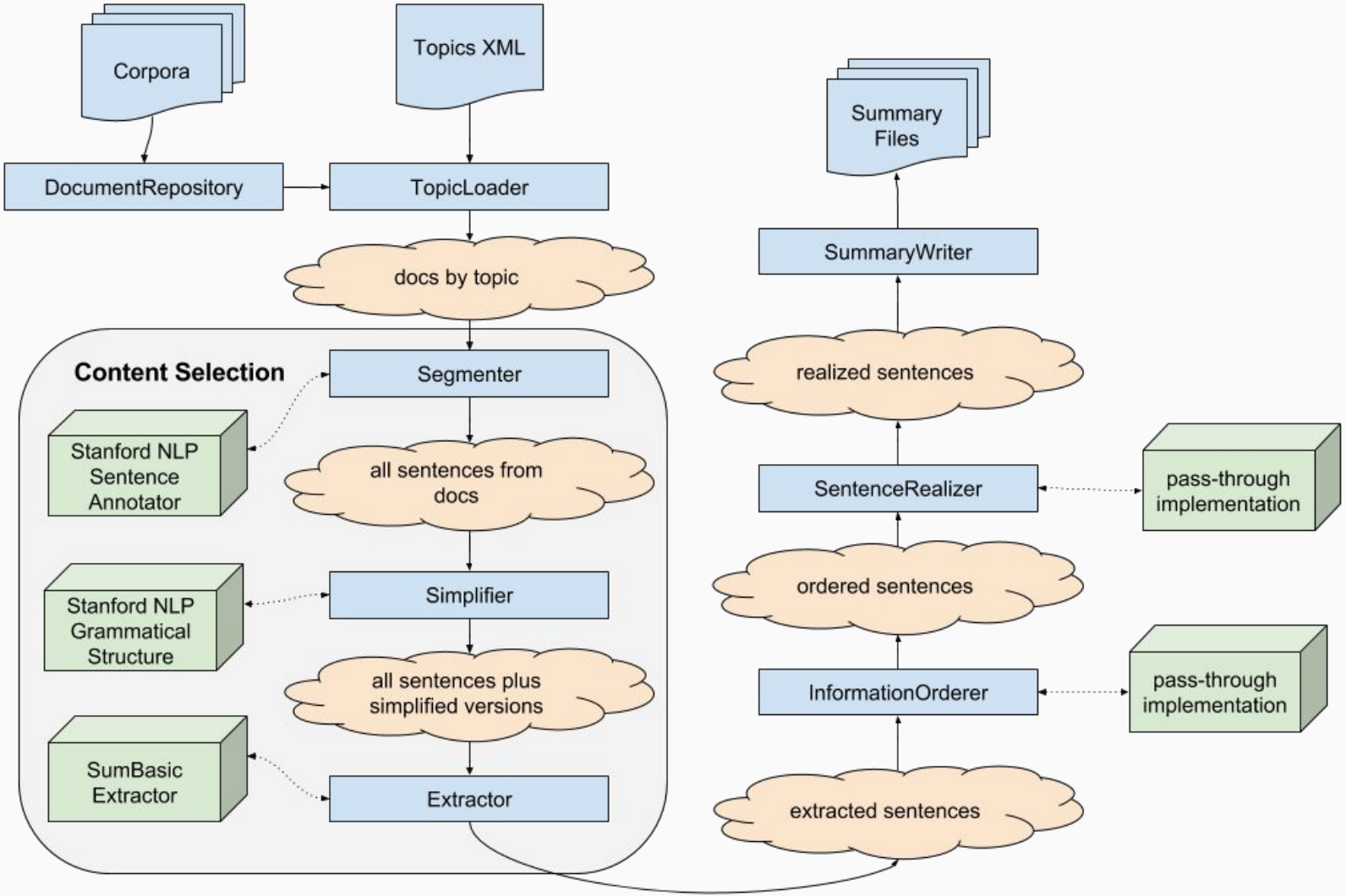
```
CacheKey cacheKey = new CacheKey(CACHE_BUCKET_NAME, uniqueId);
Thing myThing = cacher.getFromCache(cacheKey, () -> {
        // Cacher only uses this lambda if it hasn't been cached already.
        return thingRetriever.getThingBySomeExpensiveMeans();
});
```

Parallel processing

- Needed the speed boost, reduce bottlenecks
- Java 8’s .parallelStream() made asynchronous implementation easier
- Encountered ConcurrentModificationException, had to accommodate for non-thread-safety of third party tools

Inversion of Control pattern

- Rely on Java `interfaces` (e.g. `Simplifier` or `Extractor`) to make each component interchangeable at run-time
- Hot-swap modules by passing a command line argument
- E.g. `--simplifier AlternativeSimplifier` will use this simplifier module instead of the default `StanfordNLPDependencySimplifier`


Corpora
Topics XML
Summary Files
DocumentRepository
TopicLoader
SummaryWriter
docs by topic
realized sentences
Content Selection
Segmenter
Stanford NLP Sentence Annotator
all sentences from docs
SentenceRealizer
pass-through implementation
Stanford NLP Grammatical Structure
Simplifier
ordered sentences
all sentences plus simplified versions
InformationOrderer
pass-through implementation
SumBasic Extractor
Extractor
extracted sentences

Content Selection

- Sentence segmentation (Stanford CoreNLP sentence annotator)
- Sentence simplification
- Sentence extraction

Sentence Simplification

- Stanford CoreNLP dependency parser to remove less useful sentence fragments
- Does not work well! Ungrammatical, mislabeled, too simplistic...
- Shorter sentences means more sentences in the final summary
    - Average 5.44 sentences/summary

Sentence extraction

- SumBasic algorithm: Nenkova and Vanderwende (2005)
- A simple (too simple?) frequency based approach
- Dynamically modifies word probabilities as a way of avoiding redundancy in the summary.

Sentence extraction

1. Create a probability distribution of non-stopwords in the cluster.
2. Weigh each sentence: average the probabilities of words in the sentence.

$$weight(S_j) = \sum_{w_i \in S_j} \frac{p(w_i)}{|\{w_i | w_i \in S_j\}|}$$

3. Pick the best scoring sentence that contains the word in the distribution with the highest probability and add it to the summary.

Sentence extraction

4. Update probabilities for each word that appears in the chosen sentence.

$$p_{new}(w_i) = p_{old}(w_i) \cdot p_{old}(w_i)$$

5. Until desired summary length is reached, go back to step 2 and continue.

Sentence ordering and realization

- Both are just pass-through implementations
- Order matches the order within the documents and topics
- No modifications made for sentence realization

Results

- Super not impressive right now

Next steps

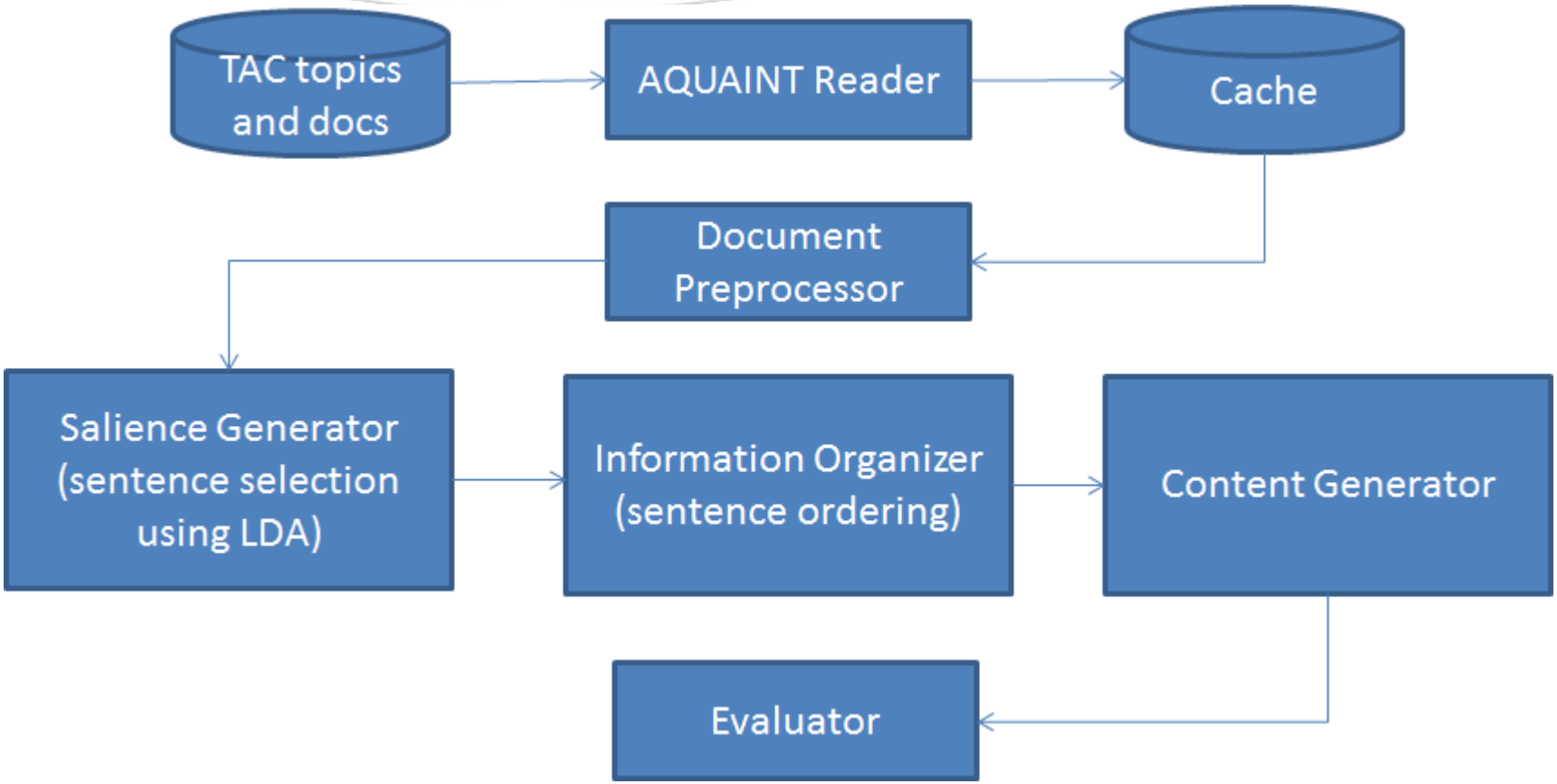
- Improve simplification (Conroy et al., 2004)
  - Regex instead of dependency parsing
- Improve extraction
  - Try more sophisticated approaches; both unsupervised and supervised (but keep SumBasic as a baseline)

Thanks!

MultiDocSummarizer

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

System Architecture



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 Blei, Andrew Ng and Michael Jordan in 2003. Paper is called “Latent Dirichlet Allocation”
- Algorithm used by gensim was created by Matthew Hoffman, David Blei 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

Further Work

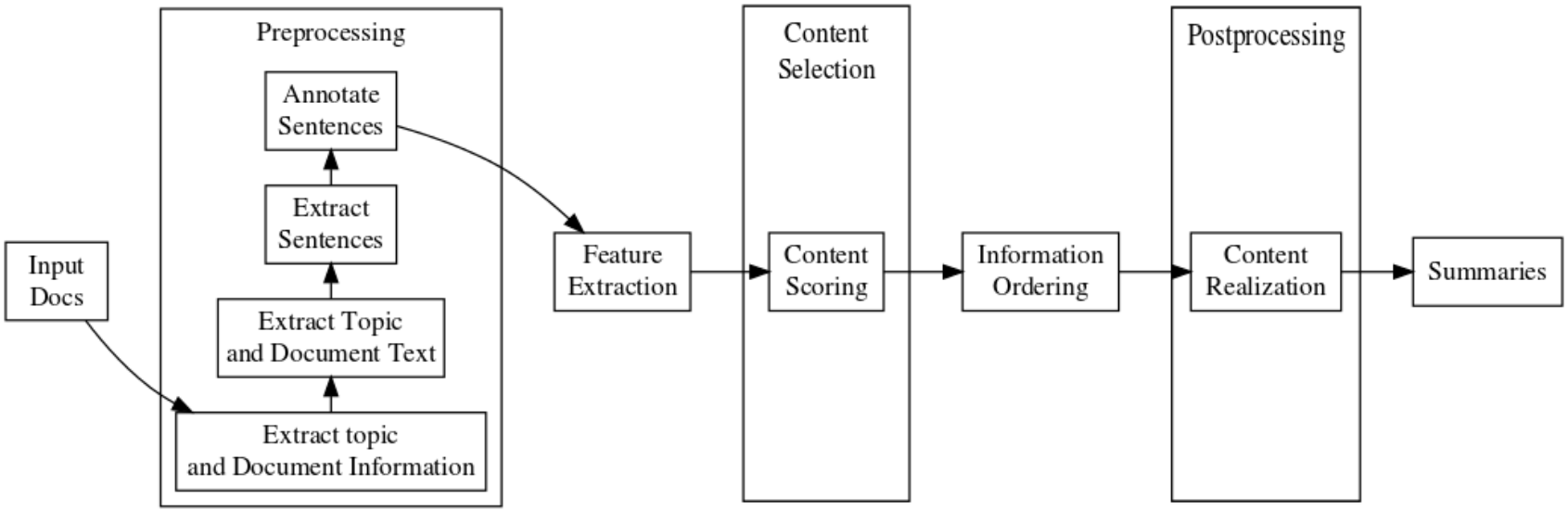
- Use machine learning for sentence ordering (based on "A preference learning approach to sentence ordering for multi-document summarization" by D. Bollegala, N. Okazaki and M. Ishizuka). Or cluster adjacency method proposed by J. Donghong and N. Yu in "Sentence Ordering based on Cluster Adjacency in Multi-Document Summarization" since we already have sentences clustered around topics by LDA.
- Use sentence compression and/or fusion
- Use anaphora resolution for better coherence

Automatic Summarization with Neural Networks

Tony Princing and Ernie Chang and Jason Blum

April 28, 2016

System Architecture



Information Ordering

- Processes all topic sentences.
- Sorted first by saliency score.
- Interim summary built from saliency sorted sentences limited by compression value (max sentences parameter) and redundancy threshold parameter.
- Interim summary is then sorted by sentence position (location in original document).

Content Realization

- Creates final summary from interim 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 interim sentences to final summary. Potential to have a lower scoring, but short sentence added to final summary – because it fits.

Content Selection

- Propose a model of convolutional neural network (SumCNN) -> An extractive framework
- Convolutional (CONV) layer -> Pooling Layer -> Input Concatenation Layer -> Fully Connected (FC) Layer -> Loss Layer

CONV layer

- Word vectors as input
- Filters of size 3,4,5
- Implemented but not included for training yet

Input Concatenation Layer

- concat((CONV + POOLING), feater_vector)

FC Layer

- 3 layers
  - each layer has 2/3 number of nodes of the previous layer
- Tanh function as the activation function

Loss Layer

- Loss calculation
  - Rouge for 1,2,3,4-grams used as target score
  - each percentage point is treated as one class
- Multiclass Logistic Regression
  - negative log likelihood

Bag of Features

- Document-dependent features
  - sentence positions
- Linguistic features
  - binary feature values
- Tf-idf scores
  - tf-idf scores of all words

ROUGE Results

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

Results (continued)

the Nixon administration do it. wondering why you do not get a job offer after that last interview? SAN FRANCISCO _ John Santner like to collect name. SAN FRANCISCO _ Marvin Young Jr. get home from work one evening and find a certified letter from the Sacramento police that change he life. LOS ANGELES _ Sandy Darby be furious. WASHINGTON _ the Clinton administration announce a new privacy initiative Friday aim at protect child and medical patient, among other.

South Korea's envoy to the United States have offer he resignation, bow to pressure over he alleged involvement in a major political slush fund scandal, official say Tuesday. south korean President Roh Moo-Hyun call Monday for a inquiry into allegation that intelligence official have illegally tape conversation between the man who be now ambassador to the United States and a Samsung Group executive. South Korea's spy agency publicly apologize Friday for illegally wiretap telephone call year ago, reveal by the leakage of a tape 1997 conversation of a alleged payment by the Samsung Group to presidential candidate.

Moving Forward

- Word vectors as input
- Concatenate CONV layers output with pre-defined document-dependent features
- Parameter tuning
- Linear regression instead of logistic regression
- Topic and narrative as query terms