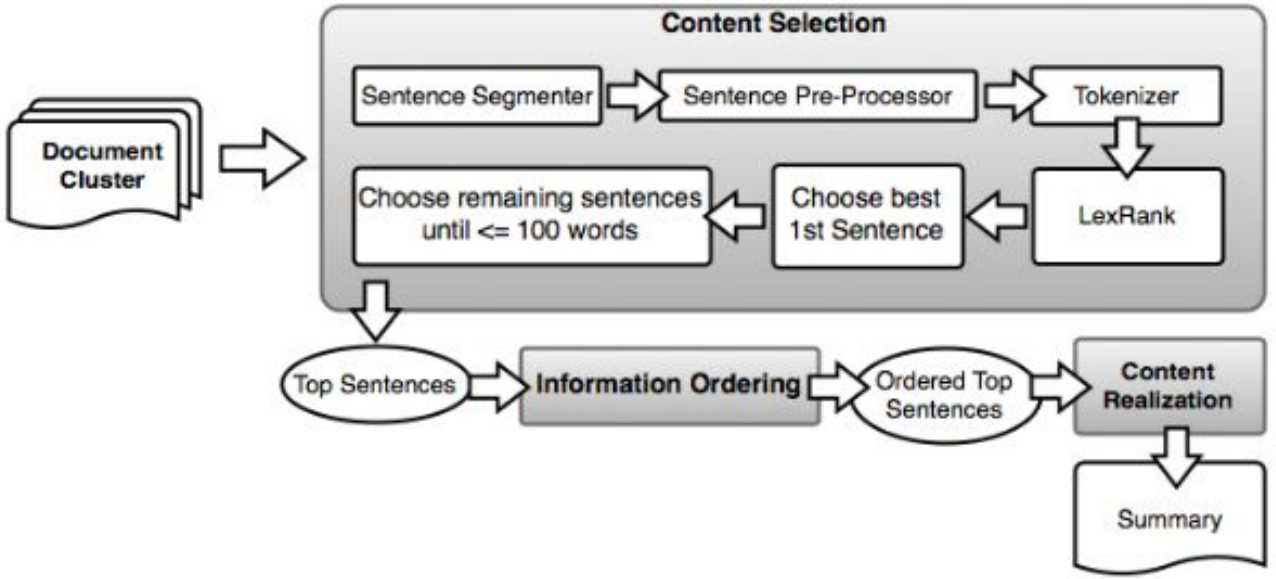
Ling 573 - Multi-document Summarization Baseline System

Martin Horn, William Lane, Ryan Lish, Spencer Morris

Roadmap

- System Architecture
- Content Selection
  - LexRank
- Information Ordering & Content Realization
- Results
- Issues and Successes
- Related Reading

System Architecture



Content Selection

- Sentence segmenter
- Sentence pre-processing
- Tokenization
- LexRank
- Choose best 1st sentence
- Choose best sentence until 100 words

LexRank

- Based on algorithms in Erkan and Radev, 2004
- Idea of random walk through graph of sentences
- Compute tf-idf for each sentence - used Brown for idf corpus
- Create matrix of idf-modified cosine similarities
- Find degree centrality d for each sentence
- Power method simulates random walk
- Highly ranked sentences are considered central to topic

Information Ordering & Content Realization

Current System:

- Maintains the order and realization determined previously in the system
- Exist as modules in the pipeline

Results

ROUGE1:

```
1 ROUGE-1 Average_R: 0.17167 (95%-conf.int. 0.15250 - 0.18959)
1 ROUGE-1 Average_P: 0.24481 (95%-conf.int. 0.22117 - 0.26838)
1 ROUGE-1 Average_F: 0.20010 (95%-conf.int. 0.17923 - 0.21994)
```

ROUGE2:

```
1 ROUGE-2 Average_R: 0.04023 (95%-conf.int. 0.03289 - 0.04779)
1 ROUGE-2 Average_P: 0.05831 (95%-conf.int. 0.04820 - 0.06939)
1 ROUGE-2 Average_F: 0.04717 (95%-conf.int. 0.03883 - 0.05573)
```

ROUGE3:

```
1 ROUGE-3 Average_R: 0.01149 (95%-conf.int. 0.00810 - 0.01527)
1 ROUGE-3 Average_P: 0.01644 (95%-conf.int. 0.01164 - 0.02167)
1 ROUGE-3 Average_F: 0.01342 (95%-conf.int. 0.00959 - 0.01780)
```

ROUGE4:

```
1 ROUGE-4 Average_R: 0.00311 (95%-conf.int. 0.00163 - 0.00473)
1 ROUGE-4 Average_P: 0.00425 (95%-conf.int. 0.00226 - 0.00655)
1 ROUGE-4 Average_F: 0.00357 (95%-conf.int. 0.00189 - 0.00547)
```

Issues and Successes

Issues:

- Lower results than we would have expected based on the LexRank literature
- Limited sentence simplification processing means that long “informative” sentences eat up a lot of space, and cause <100 word summaries
- No semantic overlap checking causes redundancy in summaries

Issues and Successes

Successes:

- Altering the LexRank algorithm to include the best first sentence significantly improved scores
- Some useful regexes allow us to filter out certain types of clauses that are unlikely to be useful: “according to…”, etc
- Adjusting the LexRank cosine threshold from .1 to .2 (as suggested in class) thinned out graph edges and brought modest improvement

Related Reading

Main Inspiration:

- Erkan and Radev 2004 - LexRank

Other:

- Radev et al. 2004 - MEAD

Summarization System D2

Katherine Topping

Stephanie Peterson

Laurie Dermer

Preprocessing and the Corpus

- Parsing XML is always a pain
  - ...but you still have to do it (we used lxml)
  - Accounted for mal-formed xml and documents using diff conventions
- Started preprocessing the whole corpus. Went "woops." Then stopped.
- For now preprocessing just:
  - Takes doc_ids from topic group and breaks doc text into original sentences
  - Tokenizes
  - Removes standard nltk stopwords
  - Applies snowball stemming to remaining words
  - Retains original and processed text via parallel doc_id->s_id->sentence dictionaries
- Corpus doc naming conventions -> folder paths was a complex conversion
  - Kind of threw a wrench in corpus navigation
  - But the efficiency improvement we got from deciphering them was worth it

Sentence Selection

- tf*idf based
  - We used raw word frequencies, not averaged, for tf
  - We'll probably change that
- Calculate sentence scores
- Select 5 highest scoring sentences
  - We figure this is plenty of sentences to reach 100 words
  - We could be wrong
- Send those sentences to Information Ordering as a list

Sentence Selection Ctd.

- Next steps:
  - Hook up llr() implementation and compare results
  - Implement down weighting strategy to avoid redundancy
  - Try out average tf versus raw count tf
  - Maybe try out some graph-based approaches?

Information Ordering

- Right now, super basic
- Just returns the sentences in the same order that the Content Selection spits them out
- So sentences are ordered from best score -> slightly less-best score
- Really doesn't do anything
- Next steps:
  - Make it do something
  - Try to improve on the Content Selection order

Content Realization

- Similar to Ordering, pretty basic for now
- Makes sure summary does not exceed 100 words (by too much)
- Otherwise doesn't really do anything besides printing the summaries
- Next steps:
  - Coreference resolution
  - Getting rid of spurious nonsense
  - Otherwise working on making summaries more coherent

Results

- ROUGE-1 : 0.12271
- ROUGE-2 : 0.02196
  - (compare to MEAD at 0.05927)
- ROUGE-3 : 0.00522
- ROUGE-4 : 0.00183

- Overall, not super great yet
- ...But that means we have lots of room for improvement!

Discussion

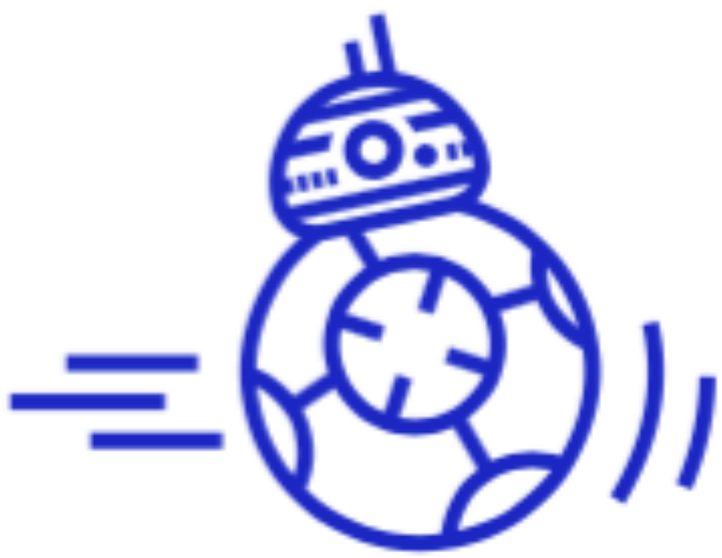
- We still have a long way to go
- But we also have a working system!
- Some challenges we ran into:
  - Some XML metadata snuck into our final summaries
  - We've already started cleaning this up
  - This may have impacted ROUGE scores
  - Just XML in general
  - Now hopefully we can focus on other challenges moving forward

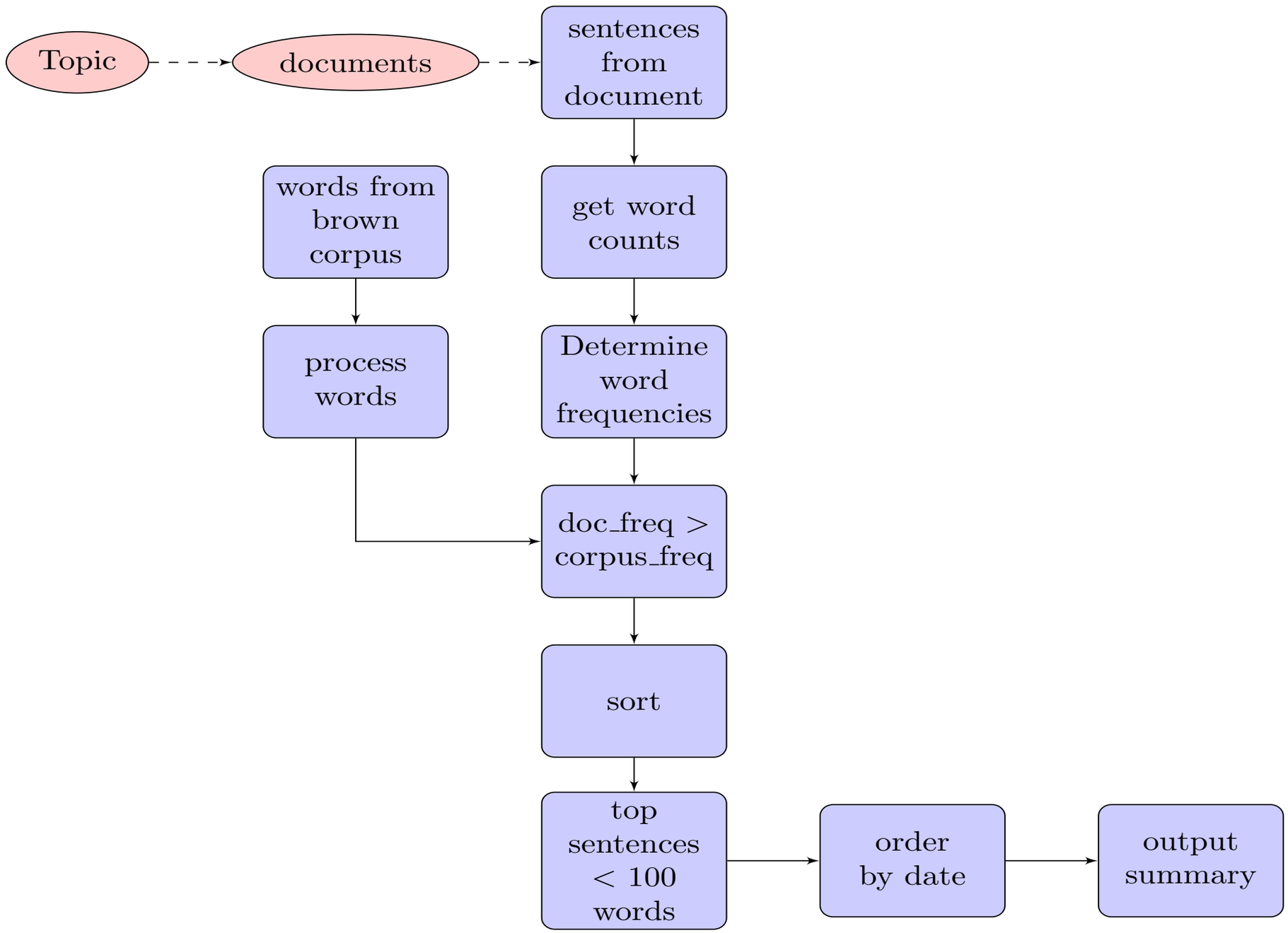
Deliverables 2

Matt Calderwood

Kirk LaBuda

Nick Monaco





System Architecture Diagram

System Overview

- Content Selection - after preprocessing, tf*idf comparison with lemmatized Brown corpus.
- Choose sentences with highest score (thematizing sentences)

System Overview

- Information Ordering - Order the sentences according to date of publication. (Ad hoc heuristic.)
- Content Realization - keep summaries under 100 words.

Issues and Successes

1 ROUGE-1 Average_R: 0.10987 (95%-conf.int. 0.09229 - 0.12813)

1 ROUGE-2 Average_R: 0.01891 (95%-conf.int. 0.01412 - 0.02389)

1 ROUGE-3 Average_R: 0.00502 (95%-conf.int. 0.00317 - 0.00720)

1 ROUGE-4 Average_R: 0.00129 (95%-conf.int. 0.00039 - 0.00242)

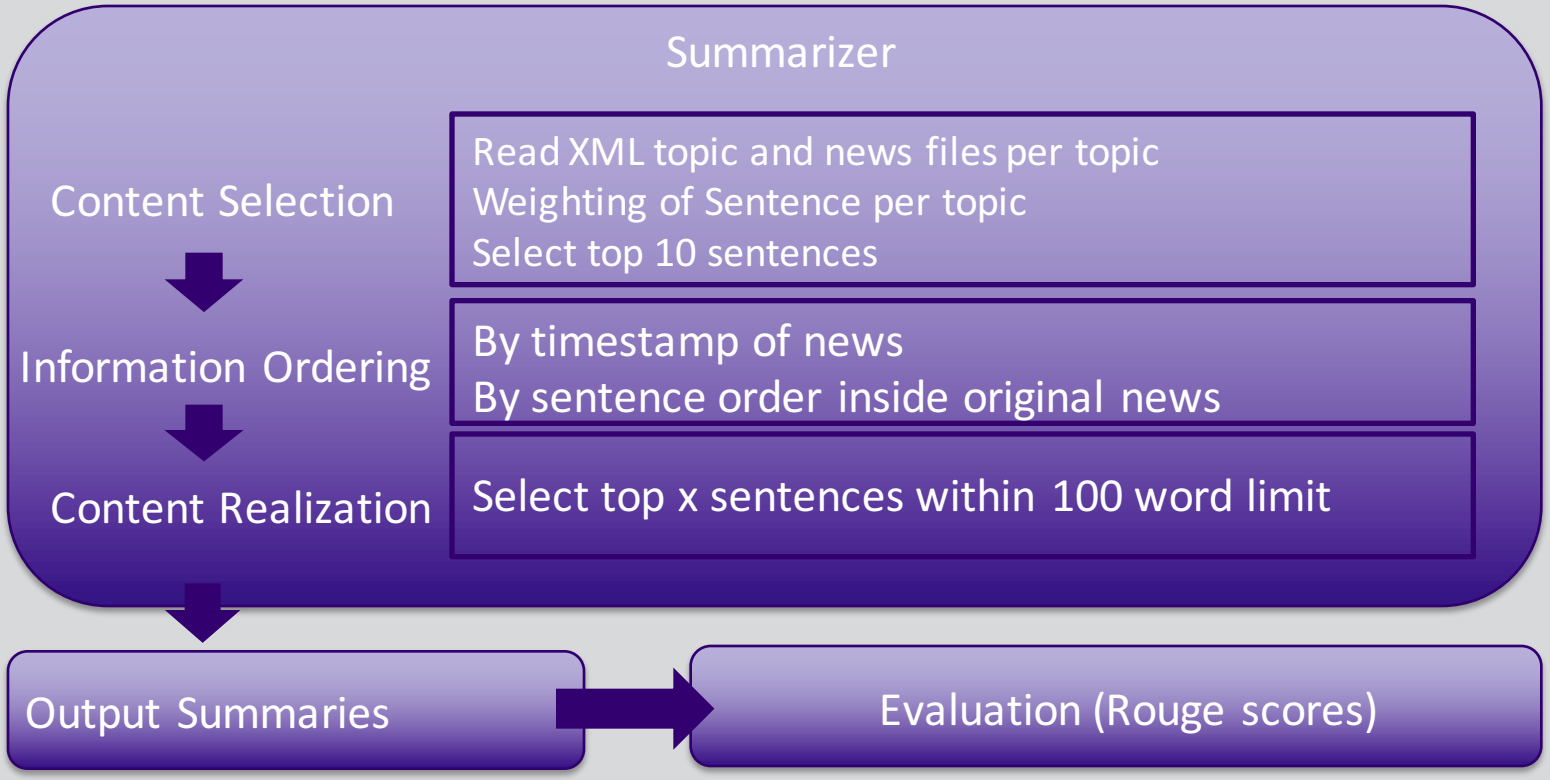
UNIVERSITY of WASHINGTON

Summarization Task

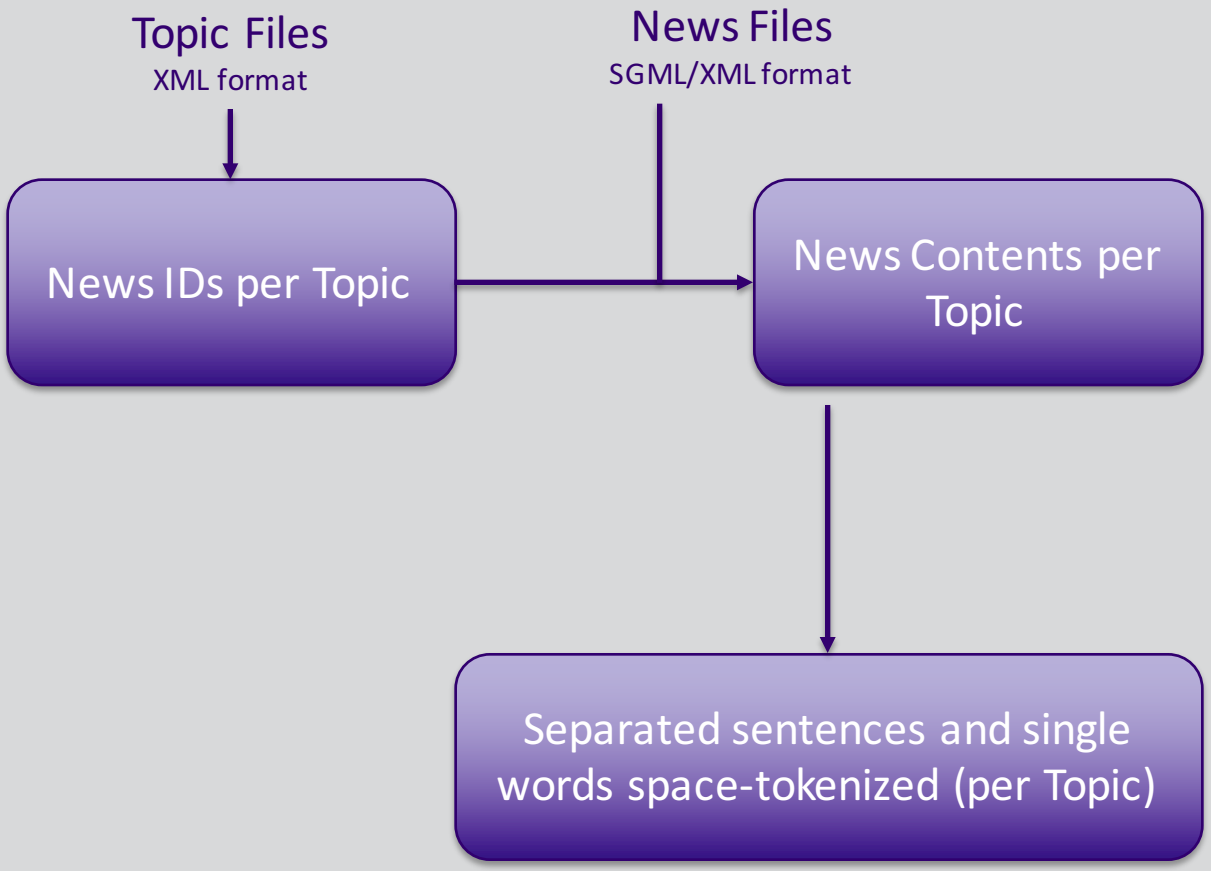
Deliverable 2

LING 573 – Spring 2016

System Architecture – Diagram



Content Selection – Information Extraction



Content Selection
Weighting of Sentences

- Simple weighting method based on Word Probability as described by Hong and Nenkova
- $p(w) = c(w)$
- $\mathrm{p}(S) = \frac{\sum_{w \in S} p(w)}{|S|}$
- Create a list of all sentences under a topic ordered by weight

Content Selection
Stop Words

- Taken from http://www.ranks.nl
- Relatively long
- Includes determiners, punctuations, common adjectives and adverbs, etc.

678 lines (677 sloc) | 4.76 KB

```
a
able
about
above
abst
accordance
according
accordingly
across
act
actually
added
adj
affected
affecting
affects
after
afterwards
again
against
ah
all
almost
alone
along
already
also
although
always
am
among
amongst
an
and
announce
another
any
```

Evaluation of Results

ROUGE scores for 1,2,3 and 4-ngrams.
Here average scores over all 46 summary topics:

	R-1	R-2	R-3	R-4
Average	0.18020	0.04338	0.01398	0.00575

- Highest value was 0.33184, lowest 0.00000
- The longer the n-gram the lower the ROUGE score

Suggestions:
- A better content selection strategy ?
- Sentences with similar content should be avoided (cosine similarity)