

Deliverable #4

Multi-document Summarization System
LING 573

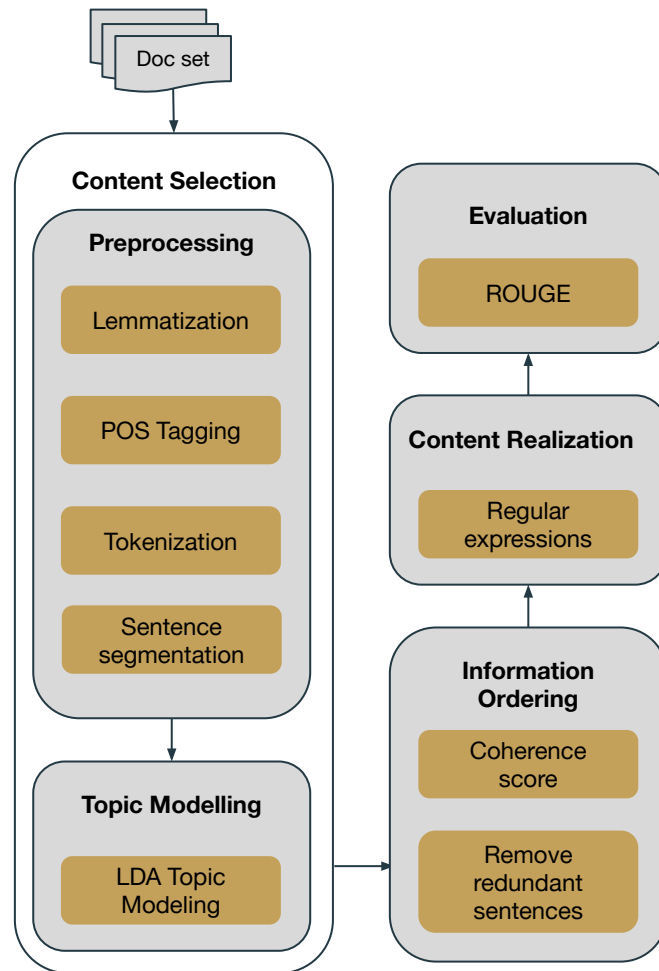
Group 1 - Elena Khasanova, Erica Gardner, Saumya Shah,
Sophia Chan and Vikash Kumar

Focus for deliverable #4

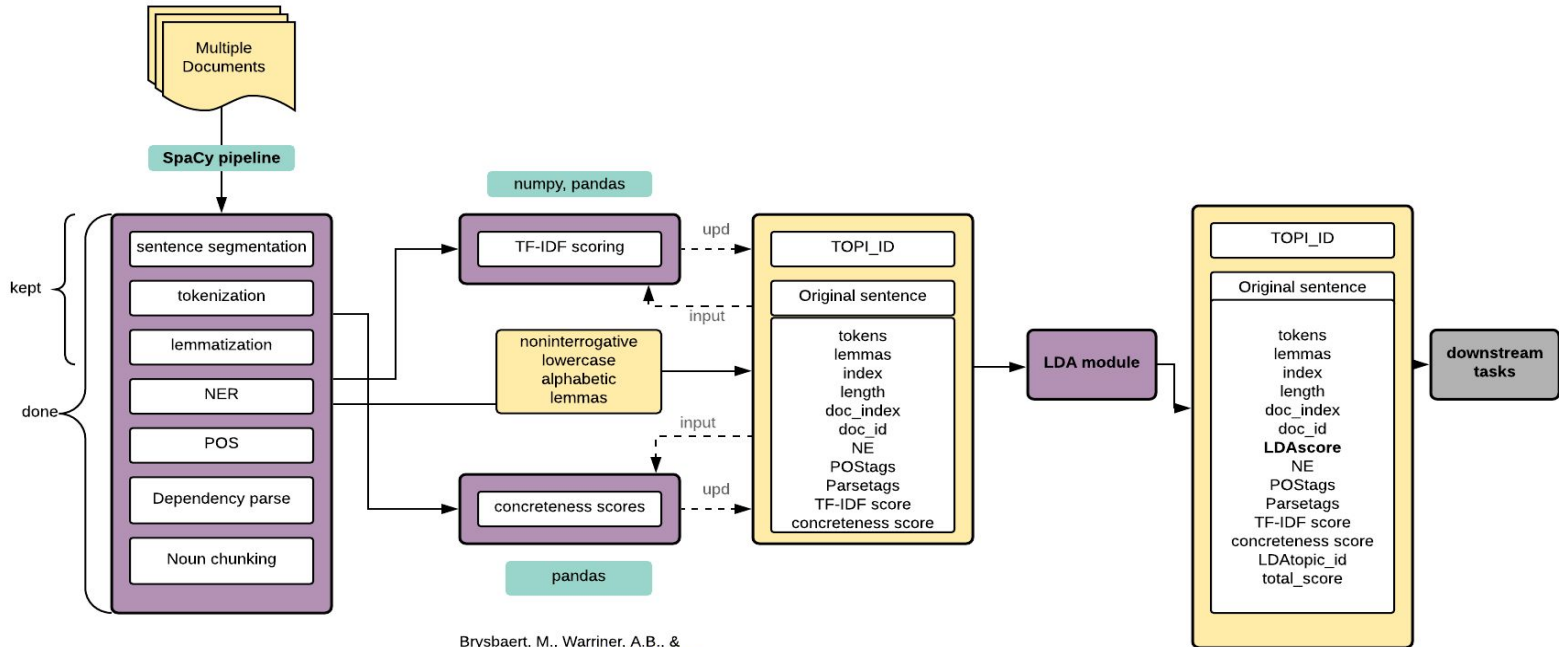
- R&D - experimentation
 - Information Ordering - use LDA topic ids to order the information
 - Content realization - coreference resolution
- Content Selection - update LDA, find a good combination of scores and parameters
- Content realization: more regular expressions and heuristics

System Overview

1. **Data Extraction** - Arrange documents by topic id and extract documents in XML format
2. **Content Selection**
 - **Preprocessing** - Stop word removal, lemmatization, sentence segmentation, POS tagging, dependency parsing, noun chunking, NER
 - **LDA with TF-IDF and concreteness** - produce a set of topics over document and a vocabulary over the topics, rank sentences
3. **Information Ordering**
 - **Removing redundant sentences** - Use cosine similarity with sentence embeddings
 - **Compute coherence score** - Use pairwise cosine similarity to find the most coherent ordering
4. **Content Realization**
 - Removing parentheses, removing adverbs, eliminating sentences shorter than 8 words



Preprocessing



Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). [Concreteness ratings for 40 thousand generally known English word lemmas](#). Behavior Research Methods, 46, 904-911.

Content selection: Revised version

For **document set** : for each **sentence**:

Compute scores:

1. Compute TF-IDF score, normalize
2. Compute concreteness
3. Perform LDA and return LDA score for each of 3 topics, take the average
4. Choose the topic with the highest score, assign this topic id (0, 1, 2) to the sentence
5. Normalize TF-IDF and concreteness scores (divide by max in the set)
6. Compute the overall score:

$$\text{TF-IDF} * \text{LDA} * \text{Concreteness} / \text{sentence length}$$

Content selection: Revised version

For document set:

Choose sentences:

1. Set num_sent parameter - desired number of sentences per document set (arbitrary - 20)
2. Divide the sentences into three subsets by the dominating LDA topic
3. Get $\text{int}(\text{num_sent}/\text{num_topics})$ top scored sentences for each topic, each set

(in our case 6 sentences per topic)

OUTPUT: a nested dictionary with keys=document set id, mapped to 18 sentences with all preprocessing and scoring information

Information Ordering

No changes to information ordering from D3

Remove redundant sentences

- Sentence representation: pretrained BERT embeddings
 - Pretrained on MRPC (Microsoft Research Paraphrase Corpus)
- Set a **redundancy threshold**
 - Do not add a sentence to a summary if it is too similar to a sentence that is already in the summary

Information Ordering

Compute coherence scores

- Generate different permutations of the summary sentences
- Compute a **coherence score** for each ordering
 - Average pairwise cosine similarity across the sentences
- Pick the ordering that gives us the highest coherence score
- Barzilay & Lapata, 2008 (from Foltz et al., 1998):

$$\text{coherence}(T) = \frac{\sum_{i=1}^{n-1} \cos(S_i, S_{i+1})}{n - 1}$$

Experiments

For each **document**:

- access LDA topic ID
- take the order of topics (0, 1, 2)
- get the most frequent order of topics

For each **document set**:

- get the most frequent order of topics

When choosing a summary:

- make sure all three topics are present
- put sentences in order of topics

- no significant improvement or drop in scores
- less coherent summaries

Content Realization

Heuristics from D3

- Lower bound and upper bound for length of sentences
- Parenthetical expressions (those with `()`, `[]`, `-- --`)
- Removed phrases like “as a matter of fact”, “however”, “also”
- Removed ages, gerund phrases, and attributions

Content Realization

Updates to heuristics based on manual evaluation of devtest summaries

- Added an “adverb exception list” to augment the removal of tokens tagged “RB”
- Remove sentences containing any quotes ('"', '"', '"') and/or capslocked words
- Remove days of the week
 - **Tuesday**’s incident → **the** incident
 - **Tuesday** night → **one** night
- Fixes to punctuation
- Remove “, which...” clauses at the end of the sentence
- Remove “but” at the beginning of a sentence

Experiments: named entities

Problem: pronominal subjects, names with no explanation

Solution: use coreference, penalize pronominal subjects, insert appositions

1. Use libraries for coreference: **SpaCy neuralcoref**
 - longer preprocessing, have to do on each document, often not accurate
2. **Construct a “knowledge graph” or “knowledge map”:**

Step 1: Use dependency parses to extract <Agent, Predicate, Theme> | <entity, apposition> lemmas

Assumption: predicates and objects would be repetitive in documents

```
'despise': {'agent': 'Blake', 'theme': 'Bakley'} | 'Levenson': {'apposition': 'prosecutor'}
```

Step 2. Replace subjects, objects, entities, appositions with corresponding noun chunks
or named entities in the mapping

Step 3: Postprocess the summaries: replace pronouns with entities based on matching predicates and
objects, add appositions to named entities mentioned for the first time

Result: potentially promising, needs more experimentation in dealing with sparsity & reconstruction of constituents

Results - D4

- **ROUGE-2**

- Average Recall: **0.06179**
- Average Precision: **0.09365**
- Average F-1: **0.07301**

Parameter tuning

LDA:

- number of terms:
going from 50 to 100 increased the scores, 150 - decreased
- random_state : 1

Sentence selection:

- weighting the scores differently
- varying the number of selected sentences: 45 sentences brings **0.068** ROUGE-2 Average Recall

Similarity thresholds:

- current: 0.094
- scores up to ROUGE-2 Average_R: 0.068 with increased threshold -> more redundancy

Results - Comparison

	ROUGE-2 (Average Recall)
LEAD Baseline	0.05376
MEAD Baseline	0.05927
D2 System	0.04340
D3 System	0.05427
D4 System	0.06179

Error Analysis - Devtest

D1003

- Three giant pandas living in Beijing Zoo will be sent to Wolong of Sichuan Province, southwest China, and three others in Sichuan will fly to Beijing, under an exchange program aimed to maintain the biodiversity of the giant panda population.
- **China to complete country's first blood bank for pandas.**

Short, out-of-context sentences

D1005

- Three years **[ago]**, Bush limited federal funding of embryonic stem cell research to the 78 stem cell lines in existence.
- A new robotic microscope that follows a brain cell in the lab from a normal state to its death has led researchers to a surprising finding that appears to debunk a long-standing theory about Huntington's disease.

“Ago” stripped by adverb removal process

Error Analysis - Devtest

D1010

- According to statistics released by the U.S. Department of Agriculture, testing for mad cow disease from **June 1 to Oct. 17** on 85,812 cows came negative, indicating that there is little chance that American cows have mad cow disease, or bovine spongiform encephalopathy.

Dates - either remove or include full year

D1022

- Humane Society International was denied permission by **Australia's** Federal Court last week to sue the Japanese whaling company for killing hundreds of whales in Antarctic waters the **Australian** government has declared a whale sanctuary.
- **Australia** vows unprecedented drive to curb Japan whale kill.
- **Australia** last week warned Tokyo it was embarking on an unprecedented diplomatic campaign to dissuade Japan from trying to escalate its killing of whales.

Repetition of referring expressions

Error Analysis - Evaltest

- Remaining issues with named entities
 - **D1002:**
 - **They** pointed to hackers' powerful attacks this week on the computers that manage global Internet traffic as evidence of the mounting threat from online criminals.
 - The effort would not improve response time, **the company** said, but would make it possible to recognize and contain Internet attacks more quickly.
 - **D1120:**
 - **It** allows the Southern Nevada Water Authority to pipe water from the Muddy, a short river about 60 miles northeast of Las Vegas; the Virgin, a long tributary of the Colorado; and the Coyote Spring Basin northeast of the city.
 - With water levels in reservoirs dropping, a record eight-year drought, the prospect that climate change could bring more dry spells and new scientific analyses suggesting the West could be drier than has been believed.

Error Analysis - Evaltest

- Short summaries
 - **D1111:**
 - An 18-year-old student opened fire in a high school in southern Finland, killing eight people including the principal before turning the gun on himself.
 - **D1125:**
 - If the polar bear were listed as a threatened species, all federal agencies would have to ensure that anything they authorize that might affect polar bears will not jeopardize their survival or the sea ice where they live.
- Missing context
 - **D1140:**
 - At least eight people died in floodwater or flood-related incidents in three northern districts, and about 1 million people were marooned by floodwater in about ten districts in northern, northeastern and central parts of the country.
 - The government's flood forecasting and warning centre said the situation could worsen in the next few days as major monsoon and Himalayan-glacier fed rivers are expected to crest.

Conclusion

We focused on improving content selection by taking TF-IDF and concreteness into account, and on manually assessing summaries.

Our D4 summaries are more coherent, less redundant, and more readable compared to our summaries in D4, but not all outputs are acceptable. Our ROUGE-2 recall scores increased moderately above the baseline.

We have investigated several different avenues for tackling a multi-document summarization system, and provided a sense of the types of issues that can arise at each part of the pipeline (content selection, information ordering, and sentence realization) and how these issues may interact.



Thank you!