Deliverable #3

Multi-document Summarization System
LING 573

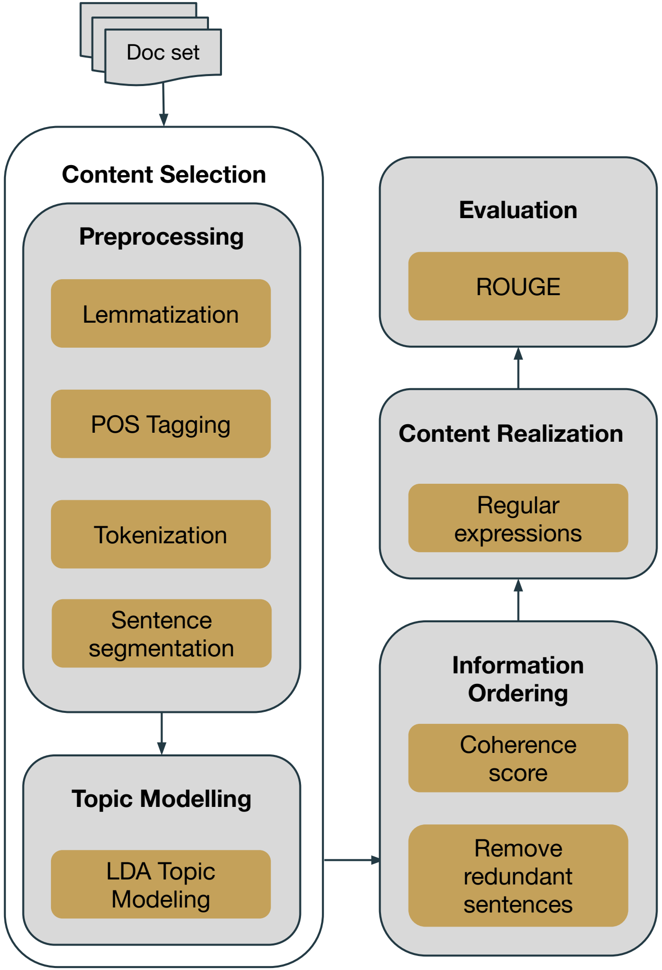
Group 1 - Elena Khasanova, Erica Gardner, Saumya Shah,
Sophia Chan and Vikash Kumar

Focus for deliverable #3

- R&D - experimentation
  - Content Selection - Graph-based methods
  - Information Ordering - Cosine similarity to filter out redundant sentences and find coherent order
- Content realization
  - Regular expressions to remove material

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 and tokenization
   - **LDA** - Produce a set of topics over document and a vocabulary over the topics
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



Content Selection - Literature Review

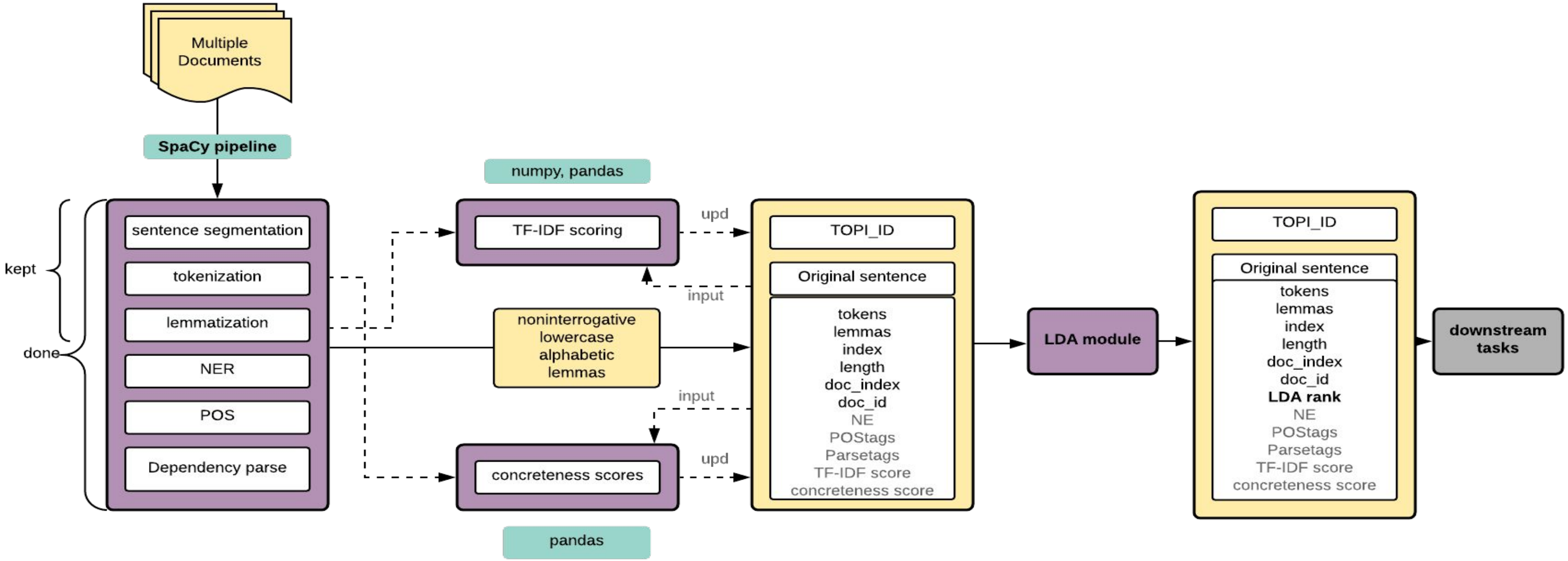
The methods we came across that perform extractive summarization are given below:

1. **“Latent Dirichlet Allocation”** - Topic modelling that estimate latent/hidden topics in any given set of document (Blei et al. (2003))
2. **“Variations of the Similarity Function of TextRank for Automated Summarization”** - Use a **TextRank matrix** that shows the **cosine similarity** between sentences to generate the top N sentences that will be used for the summary (Barrios et al. (2016))
3. **Skip-thought vectors -** The use of **skip-thought vectors** for sentence similarity and clustering to find topic clusters within a sentence (Kiros et al. (2015))

Content Selection R&D

- **TextRank Matrix Implementation**
  - **Networkx**
    - Advantage: Takes care of multi-document summarization
    - Disadvantage: If the graph does not converge at a node, it raises an Exception
  - **Gensim Summarizer**
    - Advantage: Renders summaries quickly and supports word limits
    - Disadvantage: Did not perform well with ROUGE metrics, does not support multi-document summarization
- **Skip-thought Vectors Implementation**
  - Still in progress, but the main implementation is in Python 2.

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.

LDA Experiments

- We experimented with bigrams as input to the LDA document term matrix but with much lower success. This may be due to the fact that bigrams have a lot lower distinct counts compared to unigrams so almost every bigram ends up getting almost equal weightage.
- We used the tf-idf variation to reorder the topic-term significance scores. The idea is to use the calculate ‘phi’ value for every word for every topic as a substitute for term frequency
- We tried to select sentences in the proportion of document-topic probabilities but this was discarded as most the documents had skewed distributions towards one topic.

LDA Limitations

- Topics are abstract entities which may not always be comprehensible or semantically meaningful and have no ordinal ranking.
- Document-Topic distribution is skewed towards one topic(even after controlling for the alpha parameter)
- Overlap of terms among topics, this makes it difficult to isolate words that uniquely represent a new theme.

Information Ordering and Content Realization - Literature Review

- For our approach, we drew inspiration from the following papers:
    - **Information Ordering/Coherence:**
        - **“Modeling Local Coherence: An Entity-Based Approach”** - Barzilay and Lapata (2008)
        - **“The measurement of textual coherence with latent semantic analysis”** - Foltz, Kintsch, and Landauer (1998)
    - **Content Realization**
        - **“Back to Basics: CLASSY 2006”** - Conroy et al. (2006)

Information Ordering - Removing redundant sentences

- Calculate cosine similarity across pairs of sentences
    - GloVe vectors via SpaCy
    - Pretrained sentence embeddings via **Hugging Face** and sentence-transformers
        - Base
        - Fine-tuned
            - NLI
            - MRPC (Microsoft Research Paraphrase Corpus)

Information Ordering - Removing redundant sentences

To select a **redundancy threshold** to filter out redundant sentences:

1. Annotate with labels **redundant** and **coherent**
    - Find average **cosine similarity** within each set
2. Review pairs of training sentence ordered based on cosine similarity
3. Observe changes in ROUGE scores and in outputs

Information Ordering - Coherence scores

- After removing redundant sentences, we generate different permutations of the summary sentences
- We generate 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):

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

Content Realization

- Kept heuristics from D2 submission
  a. Did not include sentences with **< 8 words, added max word count**
  b. Removed some parenthetical expressions, those with **(), [], added -- --**
  c. Removed adverbs, except for sentence-final adverbs
- Looked at CLASSY 2006 for additional sentence trimming techniques:
    - ■ Removed phrases like “as a matter of fact”, “however”, “also”
    - ■ Removed ages
    - ■ Removed gerund phrases
    - ■ Removed attributions
- Evaluated effect on ROUGE scores for each successive addition and only included heuristics that brought up the score
- Tracked changes in overall summary word counts

Results - D3

- **ROUGE-2**

Average Recall:	**0.05427**
Average Precision:	**0.06700**
Average F-1:	**0.05981**

Results - Comparison

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

Error Analysis: ordering + selection

D1001 - Columbine Massacre

1. ~~"The young killers of Columbine High School do stand for the spirit of America," Gore said.~~

2. The nation and the world have joined in grieving for the students of Columbine.

3. The day that Columbine High School students are to return to class has been delayed because many have been attending funerals for students killed in the April 20 massacre.

4. Two days, a massacre by two students at Columbine High, whose teams are called the Rebels, left 15 people dead and dozens wounded.

Better ordering: 4, 3, 2

Error Analysis: abundant named entities

D1004 - Papua tsunami

1. Igara said reports indicated that a community school Catholic mission station and the Nimas village in the Sissano area west of Aitape had been destroyed, 30 people were dead.

2. The Papua New Guinea Defense Force, the police and health services are on standby to help the victims of a tsunami that wiped out several villages on PNG's remote north-west coast Friday night.

3. Prime Minister Bill Skate, National Disaster Services Chairman Colin Travertz and National Disaster Services Director-General Ludwig Kembu were reported to inspect the affected areas Sunday.

Error Analysis: more attribution, redundant fragments

D1012

1. A Cypriot Helios Airways Boeing 737 jetliner with 115 passengers and six crew members on board crashed Sunday north of Athens, ~~Greek news media reported~~.

2. Airport officials in Larnaka said the pilot of the airliner reported problems with the air-conditioning system minutes before losing contact with Greek and Cypriot traffic controllers.

3. The Cypriot airliner that crashed Sunday in Greece may have experienced a catastrophic loss of cabin pressure that starved the pilots of oxygen.

4. Aviation experts were baffled, saying it was rare for a plane to crash because of depressurization.

Error Analysis: coreference

D1013

1. In July, a new civil law became effective, forcing credit agencies to clear the records of identity-theft victims who can present a police report documenting the fraud.

2. Last fall, she said, she watched an NBC News report on thieves needed a name and address to steal a person's identity, using that information to set up fraudulent credit accounts.

3. She had no idea what identity theft was until 1995, she received a call from the Bank of New York demanding that Frank pay up on her $11,000 credit card balance.

Error Analysis: connectors

D1030

But the revised guidelines allow companies to advertise the health effects of herbal supplements without having to prove their effectiveness as the wording does suggest that the products prevent, treat or cure disease.

The federal Food and Drug Administration says it has received hundreds of reports from physicians, health authorities and others about adverse reactions to ephedrine-based products.

But some local authorities have recognized the seriousness of the problem, and have done to crack down on drug-related crime.

Results: Missing fragments

D1014

Original:
Many spoiled teenagers from one-child families eat only a particular kind of food, such as meat or sweets, which might easily lead them to develop obesity, diabetes or heart disease.

Post-processed:
Many spoiled teenagers from one-child families eat a particular kind of food, such as meat or sweets diabetes **or heart disease.**

Error Analysis

- Why did it happen?
  - **Content**: several sentences receive the same LDA rank, content realization makes the most coherent but not the most relevant selection
  - **Redundancy**: hard to measure similarity of sentence fragments
  - **Coreference:** no specific treatment of coreference, no punishment for pronominal subjects
  - **Connectors**: not adapted to the selected context
  - **Missing fragments**: surface level postprocessing
  - **Attribution**: not enough diverse cases were considered
  - **Pipelines** are difficult: error propagation, fixing one issue breaks another

Potential Solutions

- What can be changed?
  - **Irrelevant content**: distribute sentences across LDA subtopics
  - **Redundancy**: hard to measure similarity of sentence fragments
  - **Coreference:** penalize pronominal subjects
  - **Missing fragments**: do post-processing on a phrase-structure level
  - **Attribution**: consider more cases
  - **Connectors**: apply a new heuristic
  - **Error propagation**: perform ablation study and select the best configurations of each model, then solve the resulting issues

Error Analysis: Summary

- Much less redundancy: sentence - 13 %, fragments - 22%
- About 46 % of summaries need minor improvements
- 26% of summaries have injections of short out-of-place sentences
- 44% can benefit from reordering of the selected sentences

Improvements for D4

- **Content Selection**
  - Selecting spans instead of entire sentences. We have re-ordered our topic term distribution and can now use this to extract spans around the most important terms.
  - Delexicalization of the input based on word vector similarity
  - Explore **skip-thought vectors** and clustering methods
- **Information Ordering**
  - Weight cosine similarity by LDA topic and document ID
  - Store pairwise cosine similarities instead of calculating on-the-fly so that we can get more sentence permutations
- **Content Realization**
  - Use phrase structure to ensure **cohesion**
  - Incorporate **entity linking** and **coreference resolution**
  - Select among **different versions of the same sentence** to get us closer to the word count limit

Conclusion

The end-to-end system achieved the ROUGE-1 and ROUGE-2 scores comparable with the baselines and outperformed the baseline in one out of three cases. The scores show that our approach has the potential to yield good results.

This deliverable is largely focused on experimenting information ordering and content realization methods and improving the concerns in the last deliverable such as redundancy. We also experimented with several content selection methods and built the scaffolding for a more refined versions of this module.

Thank you!