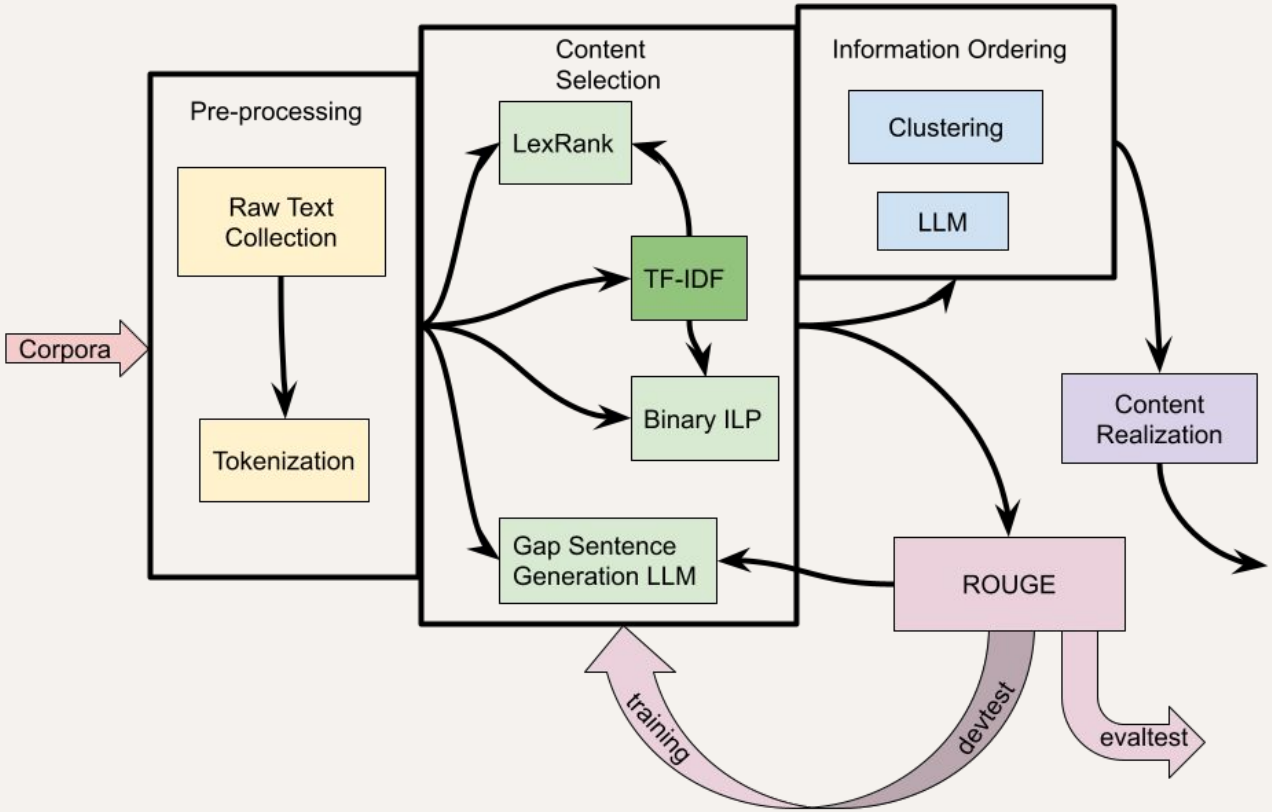
D4

Improved System

Anna Batra, Sam Briggs, Junyin Chen, Hilly Steinmetz

Table of Contents

System Architecture



Full Functionality TF-IDF

- Merged ILP and LexRanks TF-IDFs
- Created options for sentence (LexRank) vs. docset level (ILP)
- Option for idf to be over the entire data, while still letting tf be sentence or docs level
- Unigram, bigram, trigram functionality
- Options to use logged functions, change log_base, use smoothing (w/ deltas for tf and idf), remove punctuation, remove lowercase

Under the Hood: Vector Generator

- Created an interface to create sentence vectors based different methods
- Implemented for TF-IDF, Word2Vec, and DistilBERT with the help of abstract classes

```python
class VectorModel(ABC):

    @abstractmethod
    def vectorize_sentence(self) -> np.ndarray:
        pass

    def __call__(self, sentence: List[str]) -> np.ndarray:
        return self.vectorize_sentence(sentence)
```

Under the Hood: Vector Generator

TF-IDF

Sentence TF-IDF
(IDF from current docset or provided)

Word2Vec

Obtain centroid of vector from pretrained word2vec model

DistilBERT

Average pool of DistilBERT hidden layers for a sentence

LexRank – Improvements

Minimum sentence length

Avoid short clippings and prevents skewed IDF scores

Redundancy Detection

Do not include sentences with similar word sets

New Sentence Extractor

Extract top *k* sentences until reaches the token limit

Compared Different Vectors

Examined whether word2vec or DistilBERT sentence vectors improved scores

Redundancy Measure - Jaccard Distance

- Jaccard distance is a method of comparing sets
- Can be adapted for comparing sentences
- Used in the updated *sentence extractor*

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$J(s_i, s_j) = \frac{|\{w \mid w \in s_i \,\&\, w \in s_j\}|}{|\{w \mid w \in s_i \; or \; w \in s_j\}|}$$

Sentence extractor

- Like before, it picks the top *k* sentences until the maximum number of tokens is reached
- Skips sentences highly similar to a previously seen sentence
- Continues into lower ranked sentences to see if they can fit into summary
  - It ends if it goes beyond the top *k*
- Skips sentences that contain more tokens than the summary allows

Algorithm Sentence Extractor

Input: top_k, max_w, $min\_jaccard$
Output: Summary string

Array $S \leftarrow$ LexRank(doc)
Array $SeenSentences$
String $Results$
Int $Length \leftarrow 0$
Int $i \leftarrow 0$
while $i \leq top_k$ or $Length \leq max_w$ **do**
    Bool $CheckJaccard \leftarrow True$
    for $j \leftarrow 1$ to len($SeenSentences$) **do**
        if $D_j(s[i], s[j]) \geq min\_jaccard$ **then**
            $CheckJaccard \leftarrow False$
    if $CheckJaccard$ is $False$ **then**
        $i \leftarrow i + 1$
        Continue
    else if $\text{len}(s[i]) + Length > max_w$ **then**
        $i \leftarrow i + 1$
        Continue
    else$\text{len}(s[i]) + Length \leq max_w$
        $w \leftarrow w + \text{len}(s[i])$
        $Results \leftarrow Results \oplus s[i]$
        $SeenSentences \oplus s[i]$
        $Length \leftarrow \text{len}(s[i]) + Length$
        $i \leftarrow i + 1$

LexRank – Comparing Sentence Vectors

	ROUGE1 (F-Score)	ROUGE2 (F-Score)
TF-IDF	0.23101 [0.20691, 0.25529]	0.06025 [0.04748, 0.07397]
Word2Vec	0.17514 [0.15522, 0.19621]	0.03605 [0.02613, 0.04612]
DistilBERT	0.17789 [0.15117, 0.20498]	0.04157 [0.03074, 0.05313]

LexRank – Best Hyperparameters for TF-IDF

IDF	Document set (unsupervised)
ngrams	1
IDF δ-smoothing	0.7
Log-smoothed TF	No smoothing

- Other hyperparameters (minimum Jaccard distance, threshold, minimum error, minimum sentence length, ...) held constant

min. Jaccard distance	0.6
threshold	0.
min. error	1e-16
min. sentence length	7
Top *k* sentences	25

LLM - Content Selection

- Rank sentence based on Gap Sentences Generation introduced in Zhang et al. (2019) for training Pegasus
  - Select top 20% sentences based on the ROUGE-2's F1 score
    - For training data, we calculate the average ROUGE score between selected
    - For test and validation data, we calculate the ROUGE score based on all available documents
  - Discard lower 30% of the sentence based on the ROUGE score to truncate input sequence to 1024 token
- Use "google/pegasus-large" model for training with batch size of 6 and epoch of 12

Algorithm 1 Independent sentence selection

$D := \{x_i\}_n \leftarrow$ sentences in *document*
$S := \emptyset$
$I \leftarrow$ list contains index from 0 to n
for $j \leftarrow 1$ to n **do**
    $s_i := rouge(x_i, D \setminus \{x_i\})$
    $S := S \cup \{s_i\}$
$I := sort(I)$ Based on the value in S

LLM - Information Ordering

- Trained a zero-shot learning language model based on Reorder-BART's implementation by Chowdhury et al. (2021)
  - Take sets of shuffled sentences with sentence-specific markers as input
    - <S1> Sample sentence. <S2> Another sentence. ....
  - Generate a sequence of position markers as output
    - A list which contains sentences index
- We should be able to reconstruct the document with correct order based on output
  - The trained output does not produce such behavior
  - Will investigate in next delivery

LLM

- Use “google/pegasus-large” model for training with batch size of 6 and epoch of 12
- We increased batch size from 1 to 6 due to larger GPU memory
- We increased epoch from 6 to 12

Rouge-on	Epoch	Discard	Combine Masking	ROUGE1	ROUGE2
Single	6	50%	True	0.21037	0.06214
Multiple	12	50%	True	0.26419	0.05367
Multiple	12	30%	True	0.24330	0.04773
Multiple	12	30%	False	0.24263	0.05343

LLM Error Analysis - D1006

Gold

On Sept 30, Merck voluntarily recalled the pain killer Vioxx, used by almost 2 million, after clinical trials for its use in colon cancer showed unacceptable rates of stroke/heart attack.

Results corroborated earlier warnings that had not resulted in recalls by the Food and Drug Administration (FDA).

As a COX inhibitor, Vioxx was safer for digestive tracts, important for arthritis patients.

Merck's advertising campaigns did not clearly warn about side effects.

The case highlighted concerns about drug manufacturers' advertising and FDAs role in insuring safety of drugs on the market.

Safety of other COX inhibitors is now a concern."

D4

Merck recalled Vioxx in September 2004 after a study showed that it doubled the risk of heart attacks and strokes in older people taking it for at least three years.

The drug had been approved by the FDA in 1999 for arthritis.

Merck had promoted Vioxx as a way to lower blood pressure and cholesterol, but the study showed that it increased the risk of heart attacks and strokes.

Merck's decision to withdraw Vioxx from the market raised questions about aggressive marketing of the drug before its long-term safety had been proven.

D3

Vioxx should cast scrutiny on at least two problems inherent in the nation's system for assessing and monitoring drug safety.

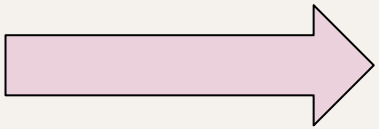
The number of warning letters has dropped precipitously since the Bush administration took power, from 82 in 2000 to 24 in 2003.

The FDA could help solve these problems not only by enforcing its own rules but by requiring doctors and hospitals to report adverse events'' when patients use drugs.

Improved ILP

- Added minimum sentence length as a hyper-parameter
- Updated to work with the new TF-IDF
- Hyperparameter tuning!

Trial-ID	Sent_length	Gram	δ_tf	δ_idf	Eliminate Punc	Lowercasing	log	Rouge-1 (Recall)	Rouge-2 (Recall)
F8	25	Unigram	0.01	0.7	No	Yes	Yes	33.697	7.437



Google Sheets

Clustering - What is Fractional Ordering?

Fractional Ordering of a sentence: The index of the sentence relative to the other sentences, divided by the number of sentences in the document it is from. Ex. the first sentence has index 1, the second has index 2... and so on.

Fractional_Ordering(i, d) = (Index of sentence i)/ (# of sentences in the document d)

Clustering - Information Ordering

1. Combine each document's sentences in the docset to be one document.
2. Convert each sentence to embeddings using Hilly's Word2Vec from his Vector-API (we may test on the other vectorizers for D5 later)
3. Then use k-means clustering to figure out blocks of related sentences (we used K=8 for now, and may improve this in D5)
4. We then ordered the blocks, by averaging the fractional ordering of every sentence in the block.
5. Afterwards, we map each summary sentence to their associated block and order them by the block ordering.
6. If there is more than one sentence from the same block, we use the fractional ordering of each sentence to break ties.

Referenced: (Barzilay, 2002)

Clustering Error Analysis

Unordered	Ordered
The school wanted to make sure there was enough to eat since students couldn't leave campus for lunch and get back in. So many forms of community, rippling outward from Columbine High and across the planet, have come together since last week's violence that it was difficult to tell. Graham praised the Columbine community for uniting under the pain of a tragedy that could have torn it apart. But Wells said he is more interested in simply trying to have fun and move beyond the tragedy that put his life on hold.	So many forms of community, rippling outward from Columbine High and across the planet, have come together since last week's violence that it was difficult to tell. Graham praised the Columbine community for uniting under the pain of a tragedy that could have torn it apart. The school wanted to make sure there was enough to eat since students couldn't leave campus for lunch and get back in. But Wells said he is more interested in simply trying to have fun and move beyond the tragedy that put his life on hold.
Several of the officers are said to have told associates that they continued firing because Diallo did not fall even after they had unleashed the fusillade. Police officers in criminal trials have often asked for a judge to decide their case, fearing that juries would be unsympathetic. While the trial date would come nearly a year after Diallo's death on the night of Feb. 4, it is not unusual in such high-publicity cases. They are accused of firing 41 times at Amadou Diallo while searching for a rape suspect on Feb. 4.	They are accused of firing 41 times at Amadou Diallo while searching for a rape suspect on Feb. 4. Police officers in criminal trials have often asked for a judge to decide their case, fearing that juries would be unsympathetic. Several of the officers are said to have told associates that they continued firing because Diallo did not fall even after they had unleashed the fusillade. While the trial date would come nearly a year after Diallo's death on the night of Feb. 4, it is not unusual in such high-publicity cases.

Improved Results

	ROUGE1 (D3)	ROUGE2 (D3)	ROUGE1 (D4)	ROUGE2 (D4)
Binary ILP	0.12085	0.01533	0.33697	0.07437
LexRank	0.13720	0.02341	0.21925	0.05966
GSG LLM	0.21037	0.06214	0.26419	0.05367

ROUGE Recall Scores

D3 Error Analysis

gold	On Sept 30, Merck voluntarily recalled the pain killer Vioxx, used by almost 2 million, after clinical trials for its use in colon cancer showed unacceptable rates of stroke/heart attack.Results corroborated earlier warnings that had not resulted in recalls by the Food and Drug Administration (FDA).As a COX inhibitor, Vioxx was safer for digestive tracts, important for arthritis patients.Merck's advertising campaigns did not clearly warn about side effects.The case highlighted concerns about drug manufacturers' advertising and FDAs role in insuring safety of drugs on the market.Safety of other COX inhibitors is now a concern."
Binary ILP	FDA urged to weigh in Then I gave her the facts , '' he said . said Lopez-Mendez , medical director of rehabilitation services at Winter Haven Hospital . Graham presented his findings in France Aug. 25 , but already had encountered the resistance from supervisors . The FDA said that Graham decided to revise his abstract conclusion . Over the next three days , Kim and his researchers convened three teleconferences with about 15 outside medical experts to get their advice . `` This morning Merck is announcing a voluntary worldwide withdrawal of Vioxx , '' Gilmartin began .
LexRank	\" That's the tragedy here . '' And if courts determine that Merck was negligent, the company will pay a heavy price in compensation. It is in the insurance industry's interest, the FDA's interest and the federal government's interest--because the federal government is a major provider of health insurance--either to require drug companies to conduct such comparative tests or to set up a neutral agency to do so. FDA: http: //www.fda.gov/
GSG LLM	Vioxx should cast scrutiny on at least two problems inherent in the nation's system for assessing and monitoring drug safety. The number of warning letters has dropped precipitously since the Bush administration took power, from 82 in 2000 to 24 in 2003. The FDA could help solve these problems not only by enforcing its own rules but by requiring doctors and hospitals to report adverse events'' when patients use drugs.

D4 Error Analysis

gold	On Sept 30, Merck voluntarily recalled the pain killer Vioxx, used by almost 2 million, after clinical trials for its use in colon cancer showed unacceptable rates of stroke/heart attack.Results corroborated earlier warnings that had not resulted in recalls by the Food and Drug Administration (FDA).As a COX inhibitor, Vioxx was safer for digestive tracts, important for arthritis patients.Merck's advertising campaigns did not clearly warn about side effects.The case highlighted concerns about drug manufacturers' advertising and FDAs role in insuring safety of drugs on the market.Safety of other COX inhibitors is now a concern."
Binary ILP	Merck officials said last week its latest research showed an increased risk of heart attack and other cardiovascular complications in patients who took Vioxx for at least 18 months. Heavily advertised as an arthritis drug, Vioxx was pulled from the market last week after its maker said a study showed it doubled the risk of heart attack and stroke. But some doctors say this group of drugs may work in a way that increases the risk of heart problems for some patients, and they point to this latest information as additional reason for concern.
LexRank	With Vioxx, researchers had been warning about the drug's possible cardiovascular risks since 2000, only a year after it was approved by the FDA . Data from a company study found then that users had four times as many heart attacks and strokes as those who used another painkiller . But the data was not definitive, and Merck, which even critics say is one of the most responsible drug companies, repeatedly reassured the medical and financial communities that Vioxx was safe.
GSG LLM	Merck recalled Vioxx in September 2004 after a study showed that it doubled the risk of heart attacks and strokes in older people taking it for at least three years. The drug had been approved by the FDA in 1999 for arthritis. Merck had promoted Vioxx as a way to lower blood pressure and cholesterol, but the study showed that it increased the risk of heart attacks and strokes. Merck's decision to withdraw Vioxx from the market raised questions about aggressive marketing of the drug before its long-term safety had been proven

References

- Luo, W., Liu, F., Liu, Z., & Litman, D. (2018). A novel ILP framework for summarizing content with high lexical variety. *Natural Language Engineering, 24*(6), 887-920. doi:10.1017/S1351324918000323
- Erkan, Günes, and Dragomir R. Radev. 2004. "Lexrank: Graph-based lexical centrality as salience in text summarization." in *Journal of artificial intelligence research* 22: 457-479.
- Zhang, Jingqing, et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." International Conference on Machine Learning. PMLR, 2020.
- Chowdhury, Somnath Basu Roy, Faeze Brahman, and Snigdha Chaturvedi. "Is Everything in Order? A Simple Way to Order Sentences." arXiv preprint arXiv:2104.07064 (2021).
- Barzilay, R., Elhadad, N., & McKeown, K. (2002). Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research, 17, 35–55. https://doi.org/10.1613/jair.991

Thank you