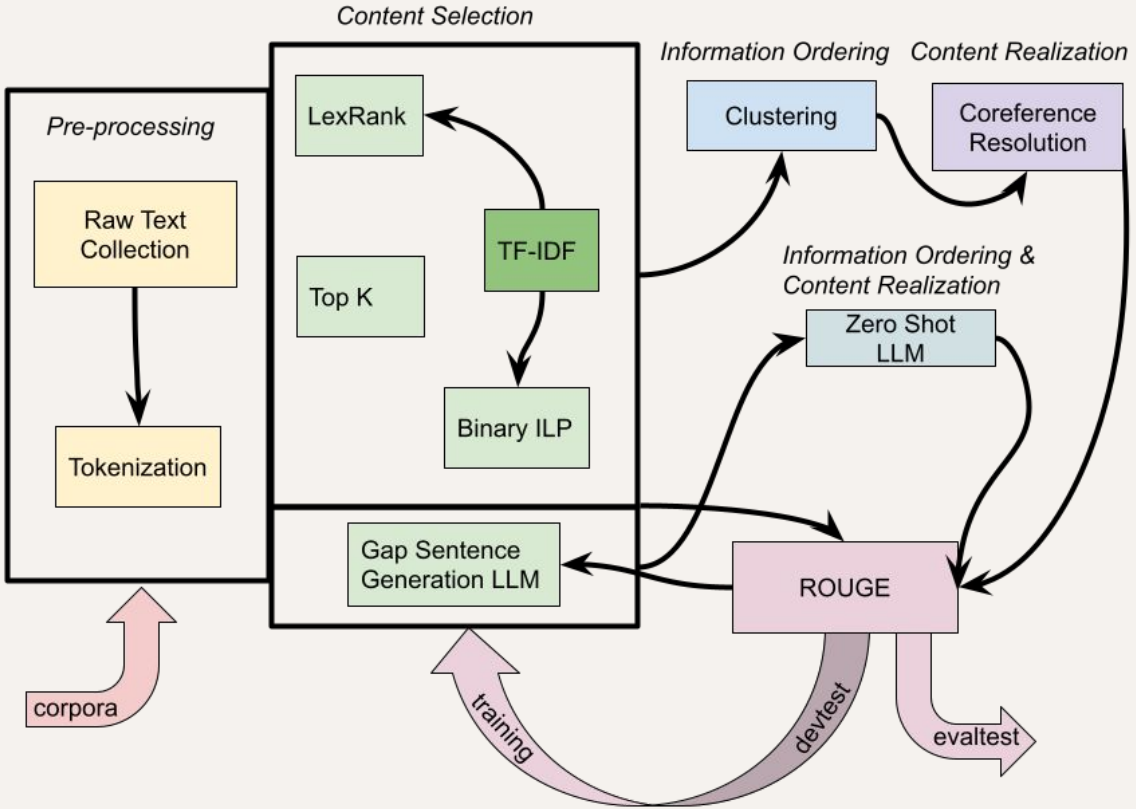
Exploring Extractive and Abstractive Approaches for Multi-Document Summarization: An End-to-End System with Benchmarking and Error Analysis

Anna Batra, Sam Briggs, Junyin Chen, Hilly Steinmetz

Table of Contents - #TODO

System Architecture



Clustering Hyperparameters

- Choosing one of the vectorizers: TFIDF, Word2Vec, DistilBert
- Choosing mean vs. median in terms of ordering the blocks
- Future Work: Test different numbers for K (the number of clusters/ blocks created)

Error Analysis - Mean (Experiment G2 on D1001, D1002)

TFIDF	Word2Vec	DistilBert
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. 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. 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.	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. The school wanted to make sure there was enough to eat since students couldn't leave campus for lunch and get back in.
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. They are accused of firing 41 times at Amadou Diallo while searching for a rape suspect on Feb. 4. 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. Police officers in criminal trials have often asked for a judge to decide their case, fearing that juries would be unsympathetic.	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.	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. 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. 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.

Error Analysis - Median (Experiment G2 on D1001, D1002)

TFIDF	Word2Vec	DistilBert
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. 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. 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.	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. The school wanted to make sure there was enough to eat since students couldn't leave campus for lunch and get back in.
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. They are accused of firing 41 times at Amadou Diallo while searching for a rape suspect on Feb. 4. 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. Police officers in criminal trials have often asked for a judge to decide their case, fearing that juries would be unsympathetic.	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.	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. 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. 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.

Baseline - Top K

- For our baseline we implemented taking the first k sentences of the first document in the docset.
- If the first sentence is too long (over 100 words), we keep skipping the first few sentences until we find one sentence less than 100 words.
- Then, we continue adding more sentences one by one until the next sentence makes the summary over 100 words.
    - Possibly we could end up with a one sentence summary if the proceeding sentence already makes it over 100 words.

LLM – Improvements and Issue

- We investigated on the zero-shot learning language method based on Reorder-BART's implementation by Chowdhury et al. (2021).
  - We increase the training epoch
  - We increase the data coverage
    - Include all the sentences from training, devtest, and evaltest.
    - Do not throw the sentences away when reaching input size limit. Instead, we put the sentences in another group with new index starting from zero.
    - We decrease the input size from 1024 token to no more than six sentence.

LLM – Improvements and Issue

- Input: [shuffled] <S0> LITTLETON, Colo. (AP) -- The sheriff's initial estimate of as many as 25 dead in the Columbine High massacre was off the mark apparently because the six SWAT teams that swept the building counted some victims more than once. <S1> Sheriff John Stone said Tuesday afternoon that there could be as many as 25 dead. <S2> By early Wednesday, his deputies said the death toll was 15, including the two gunmen. <S3> The discrepancy occurred because the SWAT teams that picked their way past bombs and bodies in an effort to secure building covered overlapping areas, said sheriff's spokesman Steve Davis. <S4> ``There were so many different SWAT teams in there, we were constantly getting different counts,'' Davis said. <S5> As they gave periodic updates through the night, Davis and Stone emphasized the death toll was unconfirmed. <S6> They said their priority was making sure the school was safe. [orig]
- Expected output: 0, 1, 3, 3, 4, 5, 6 [eos]
- Actual output: <s>LLE,.AP -- sheriff initial of many 25 in Columb High was the mark apparently the's estimate as Many 25 In Columb high wasthe in inine because six teams swept building counted victims than.Sher John said afternoon there be Many 25.By Wednesday his said death toll 15 including two. By WednesdayHis saiddeath toll15 includingTwo.The occurred the teams picked their past and in effort secure overlapping, sheriff spokesman Davis <S4> `` were many SWAT in, were constantly different,'' said said <S5> they periodic through night Davis Stone emphasized death wasconfirmed <S6> They their toll un.They his was death Toll 15including two gunmen <S3> The occurring because SWAT that his past bombs bodies an effort securing overlap, Sheriff spokesmanDavis <S4>

LLM – Improvements and Issue

- We investigated on the zero-shot learning language method based on Reorder-BART's implementation by Chowdhury et al. (2021).
  - We increase the training epoch
  - We increase the data coverage
    - Include all the sentences from training, devtest, and evaltest.
    - Do not throw the sentences away when reaching input size limit. Instead, we put the sentences in another group with new index starting from zero.
    - We decrease the input size from 1024 token to no more than six sentence.
- Unfortunately, none of the above experiment improve the result, and we choose not to include this method in out final end-to-end system.

LLM – Improvements and Issue

- We used the parameters from the last best model in D4 and increase the training epoch.
  - The highest ROUGE 1 score achieved on epoch 14.

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	24	50%	True	0.28415	0.06464
Multiple	12	30%	True	0.24330	0.04773
Multiple	12	30%	False	0.24263	0.05343

LLM – Improvements and Issue

- We conduct experiments on different checkpoints for PEGASUS

Model name	ROUGE1	ROUGE2
pegasus-large	0.26419	0.05367
pegasus-cnn_dailymail	0.31355	0.08191

Content Realization

- Adapted from algorithms and observations made by Siddharthan et al. (2011).

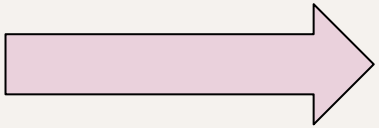
1. Obtained all coreferent clusters from the document set.
2. Went through all entities in the summary.
3. Tried to obtain:
   a. The longest premodifying noun phrase
   b. The longest noun phrase
4. Replaced the NP in the summary with the coreferring NP. Indexed the the cluster.
5. Upon iterating,
   If the NP is in an unseen cluster,
   replace it with the longest (pre-)modifying phrase
   Else,
   replace it with the shortest non-pronominal phrase

Content Realization - Error Analysis?

Ablation on Top Model - ILP

- Ran ablation testing on all of of the hyperparameters. Our final hyperparameters are shown in experiment F8.
- Discovered sentence length doesn't improve too much actually. IDF also gives the biggest jump of ROUGE-1 11% and ROUGE-2 4%. (See experiment J6 below)

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	33.32	7.41
J6	25	Unigram	0.01	0.001	No	Yes	Yes	22.592	3.107



Google Sheets

Last Time: 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

Test Results

	ROUGE1 (devtest)	ROUGE2 (devtest)	ROUGE1 (evaltest)	ROUGE2 (evaltest)
Baseline (top K)				
Binary ILP	0.3332	0.07415		
LexRank	0.21925	0.05966		
GSG LLM	0.31355	0.08191	0.29880	0.06799

ROUGE Recall Scores

D5 Error Analysis on devtest 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.
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	In September 2004, Merck \& Co. recalled its arthritis drug Vioxx after a clinical trial showed it doubled the risk of heart attacks and strokes. The drug had been used by 20 million Americans since its approval in 1999 and was the company's top-selling product. Merck had spent \$195 million to promote Vioxx as a wonder drug for the aging baby boomers. The FDA, which approved Vioxx for use, had been concerned about the drug's cardiovascular risks since at least 2000 but did not issue a warning until 2004.

D5 Error Analysis on eval D1105

gold	Boeing 737-400 plane with 102 people on board crashed into a mountain in the West Sulawesi province of Indonesia, on Monday, January 01, 2007, killing at least 90 passengers, with 12 possible survivors. The plane was Adam Air flight KI-574, departing at 12:59 pm from Surabaya on Java bound for Manado in northeast Sulawesi. The plane crashed in a mountainous region in Polewali, west Sulawesi province. There were three Americans on board, it is not know if they survived. The cause of the crash is not known at this time but it is possible bad weather was a factor
Binary ILP	
LexRank	
GSG LLM	The Indonesian Navy (TNI AL) has sent two Cassa planes to carry the bodies of five of its members who were killed in a plane crash in Sulawesi late Monday. An Adam Air Boeing 737-400 plane with 102 people on board crashed in a mountainous area near the town of Polewali late Monday on its way from Surabaya to Manado. At least 90 people, including five TNI AL members, were killed in the crash.

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