Two Stage Multi-Document Summarization with BERT

CS 886 (Winter 2020) Final Report

Archit Rakesh Shah*
University of Waterloo
Student ID: 20773927

Abstract

Automatic text summarization is a well known problem in the field of Natural Language Processing. It refers to the problem of creating a fluent and accurate summary from a given text. Recently, both BERT and GPT architectures have been fine-tuned to coerce them into generating summary from given text document. There has however been little focus in the field of multi-document summarization task. Multi-document summarization refers to the problem of summarizing information from multiple documents. This task is challenging for a number of obvious reason. In this paper, we first take a look at existing architectures that deal with multi-document summarization task, understand the inherent challenges associated and then conduct an experiment inspired by Subramanian et al[1] (2019). We approach the problem of multi-document summarization with a two staged summarization, by first summarizing individual text documents within a topic and then summarizing the combined resultant text to generate a final summary per topic. We conduct this experiment completely on Google Cloud Platform. Most of the existing architectures on Multi Document summarization have been performed on Wikipedia dataset which is quite huge and diverse in its topic coverage. Recently, a new dataset named ”Multi News Dataset” was published which contains news articles from same real-world events and a golden summary article as target. The code and results from our experiment is available here.

1 Introduction

Text summarization has been one of the central tasks of Natural Language Processing over the years and with the recent popularity of deep learning and transformers it has garnered even more attention. Most of the new architectures and experiments in academic papers have been focused on single document summarization - using newsroom dataset by Grusky et al.[2] (2018) as the benchmark. Many of these architectures have shown really promising results for single document summarization with high rouge scores. For instance, the current leaderboard maintained and verified by the authors of this dataset, is topped by modified pointer generator network Shi et al.[3] (2019) with R1 F score of 39.91. More recently BERT, a pre-trained transformer model, has been very popular in solving all kinds of Natural Language Processing problems. It has also been applied to single document text summarization task with architectures such as BERTSum by Liu et al.[4] (2019) achieving R1 F1 score as high as 43. Even GPT has been employed to perform summarization task by Subramnian et al.[1] (2019) on with R1 F score of 43.

Multi document summarization - task of generating summary from multiple documents that are related as they convey thematically similar information with a different style. As reported by Liu et al.[5] (2019), this task has not received as much attention as single document summarization due to lack of datasets. Liu, Peter J., et al.[6] (2018) have provided WikiSum datasets in the past which has been used as the baseline for quite some time now. However, this dataset too comes with a lot of challenges as it is extremely huge and has a wide variety of Wikipedia articles. More recently, a new dataset called Multi-News was released by Fabbri et al.[7] (2019) which also claims to be helpful in multi-document. The dataset has been evaluated on a couple of Pointer Generator networks with ROUGE score of 43 in the same paper.

*e-mail: archit.shah@uwaterloo.ca

With this project, we aim to conduct our own experiments around multi-document summarization task by leveraging the multi news dataset. There are quite a few pre-trained BERT summarization models available to be used, and some of them are even trained readily on newsroom dataset so they would be perfect to be applied to multi-news dataset - as both consist of news articles.

2 Related Work

The earliest automatic text summarization task focused on extractive frameworks. Erkan et al. [8] (2004) employed similarity metrics between a text document and its summary. This comparison was primarily limited to sentence features. More recently, further complex sentence features have been leveraged to extract sentence features. With the advent of encoder decoder architectures in neural networks, they have been put to solve the task of text summarization. Most of the work can be categorized as either extractive or abstractive summary generation.

Extractive summarization refers to frameworks that identify important sentences in the text and generate them verbatim producing a subset of sentences from the original text document. Because they are extarctive in nature, these techniques may lack the fluency and coherency of human generated gold summaries.Cheng et al.[9] (2016) and Nallapati et al.[10] (2016) proposed training encoder-decoder neural network architecture as a binary classifier to identify the important sentences in a given text, that could make up the summary. They proposed an extractive encoder-decoder architecture that can pick an unordered set of sentences from the text document to assemble the extractive summary. Chen et al. [11] (2018) use a pointer network (by Vinyals et al.[12] 2015) to sequentially select sentences from the text document forming the extractive summary.

Abstractive summarization refers to the frameworks that reproduces important material from the text in a new way after understanding the given text document. It can generate sentences and words that are not present in the original text. Attention based encoder-decoder architectures have been widely studied to understand abstractive summarization. One of the early works are from Rush et al.[13] (2015) indicate ROUGE scores of just 30 on the old DUC-2004 dataset. Neural architectures have come a long way since then. In 2017 See et al.[14] (2017) Pointer Generator networks, can copy words from the source text while retaining ability to generate new words which achieved the benchmark ROUGE score of 39 on newsroom dataset. In 2019, Liu et al.[4] (2019) proposed pre-training of BERT models to summarize text in abstracive manner with a ROUGE score of 43 on Newsroom dataset. In this project we will be using this BERT model to employ our summarization tasks in this two-staged experiment.

The above work has been solely covered for single document summarization. There has been somewhat limited work in terms of multi-document summarization tasks comparatively. In 2018, Liu, Peter J., et al.[6] (2018) proposed a two stage architecture where an extractive model first extracts important sentences from individual articles and then an abstractive model to combine those summaries with ROUGE F1 of 43. In the same paper they also introduced the new dataset WikiSum for multi-document summarization task. More recently in 2019, Liu et al.[5] (2019) had proposed a hierarchical document encoding with transformers to represent cross document relationships via an attention mechanism which allows to share information. The model achieved a ROUGE F1 score of 41 on WikiSum dataset.

In 2019, Fabbri et al.[7] (2019) introduced a new large scale multi-document summarization dataset. The authors also applied a pointer generator network to achieve a ROUGE F1 score of 43. In this paper we will utilize this dataset to conduct our experiment.

3 Data

The dataset released by Fabbri et al.[7] (2019) is called Multi-News. It consists of multiple articles for same news event and human written gold summaries of these articles from the site newser.com. Each summary has been written by professional editors and the sources have been cited in the articles. Through this citation links, the authors have extracted the citation material through stable wayback links. We will be using a subset of this dataset with about 45,000 news events where each event has between 2 to 10 citations for summary. In total we will be dealing with more than 125,000 individual news articles. The summaries have been written by more than 20 different editors. The summaries are notably longer than other summaries, as they have about 260 words on average. We suspect this will lead to a very poor recall on normal summary from BERT which could generate less than 100 words per summary.

3.1 Data Preparation

Each news event consists of multiple news articles from different sources. The dataset entails a separate news event on each new line. Each individual line consists of separate articles for the same event, and those articles are separated by a story separator tag word. A sentence separator token is further used to separate sentences in each article.

4 Experiment

Through this project, we conduct an experiment to perform multi-document summarization on multi-news dataset. The experiment will be conducted in two stage as it is inspired by Liu, Peter J., et al.[6] (2018) and Subramanian et al[1] (2019) as how they conducted a two staged summarization. In our experiment we will be performing abstractive sumarization on both stages. First abstractive summariation will be at local document level and second will be a global abstractive summarization on the results. We will be using a pre-trained BERT model for abstractive summarization from Liu et al.[4] (2019) that was already fine-tuned to abstractively summarize newsroom dataset in two steps.

4.1 Pre-trained models and BERT

Pre-trained models have recently gained much popularity as they can perform a wide variety of tasks. Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. [15] (2018)) is a language model that was trained with a masked language modeling and a "next sentence prediction" task on a corpus of 3,300M words. Pre-trained language models have been used to enhance performance in various NLP tasks, and we will also be using this pre-trained BERT model for our two-staged experiment.

In Figure 1, we can see the original architecture of BERT and the proposed BERTSUM by Liu et al.[4] (2019). In BERT, input text is first processed by inserting [CLS] and [SEP] tokens to notify the beginning of text and separation of sentences respectively. Each word/token is given three types of embedding: token embeddings which indicate meaning of each word, segment embedding which indicate meaning of sentences and position embeddings which indicate position of the words in a text. The three embeddings are summed and sent into a bi-directional transformer with layers identified as below:

$$\begin{aligned}\tilde{h}^l &= \mathrm{LN}(h^{l-1} + \mathrm{MHAtt}(h^{l-1})) \\ h^l &= \mathrm{LN}(\tilde{h}^l + \mathrm{FFN}(\tilde{h}^l))\end{aligned}$$

Here h is the input vector, LN is the Layer Normalization operation and MHAtt is the multi-head attention operation. We change this original BERT to summarize text by inserting external [CLS] tokens at the start of each sentence and each [CLS] will then store features from preceding sentence. BERTSUM also uses alternate interval segmentation embeddings (illustrated in Figure 1) to distinguish multiple sentences.

4.2 Fine-tuning BERT for Abstractive Summarization

In this project, we employ a fine-tuned BERT model that was specifically created to perform summarization task. Application of BERT to summarization task is not straight-forward. We refer to the BERTSum model by Liu et al.[4] (2019) where *interval segment* embeddings are used to distinguish between different sentences in the text. This allows document representations to be learnt hierarchically where Transformer layers represent adjacent sentences, while higher layers represent multi-sentence discourse in combination with self-attention.

For abstractive summarization in our project we use standard encoder-decoder framework where encoder is pre-trained BERTSUM and decoder is 6 layer Transformer. The 6 layer Transformer decoder has been trained on CNN/Dailymail dataset with a ROUGE F1 of 43 from Liu et al.[4] (2019). To ensure that fine-tuning is stable, optimizers have been separated for encoder and decoder. The two Adam optimizer $\beta_e = 0.9$ for encoder and $\beta_d = 0.999$ for decoder with separate warm-up steps, i.e. 20,000 for encoder and 10,000 for decoder. Pre-trained encoder is already fine-tuned with small learning rate and smooth decay.

The architecture itself was fine-tuned on CNN/Dailymail dataset in two phases. The encoder was fine-tuned on extractive summarization task and decoder was fine-tuned on abstractive summarization task.

4.3 Implementation and Testing

Liu et al.[4] (2019) have published this fine-tuned model online. We will use this model to perform our experiments on Google Cloud Platform (GCP). We first setup a computing instance on GCP with *2 vCPUs, 13 GB memory, 100 GB disk* and *1 x NVIDIA Tesla P100 GPU.* We install pytorch on this machine alongside py-rouge and NVIDIA libraries required to use GPU. We also download the Multi-News dataset on this machine and pre-process the data. We clone the git repository from Liu et al.[4] (2019) and add our own modifications in this repo to implement our own two staged multi-document summarization task.

At stage 1, we preprocess the data and abstractively summarize each news article for every news event from the dataset. We use a news event separator tag to keep track of news articles that belong to same news event. In total, we abstractively summarize 125,000 news articles in stage 1 and then tie them back to 45,000 original news events. This process took 70 hours to complete on GCP while utilizing the GPU. We accumulate the summaries per news event for later ROUGE analysis and stage 2 as well.

At stage 2, we further apply abstractive summarization on the accumulated results from stage 1. We generate 1 summary per news event, the source text for which comes from previous stage. We finally generate 45,000 abstractive summaries which are ready to be compared with their original hand-written gold summary in the dataset. We post-process the data to remove the story separator and sentence separator tags, so it is in a better shape to be compared with gold summaries. Stage 2 takes about 12 hours to complete.

5 Results and Analysis

Multi-Document summarization task is more challenging than single document summarization tasks. Especially when dealing news articles this task is even more challenging as different media outlets can report the same news event in different tones and different opinions. It is rather challenging to summarize on competing opinions and boiling it down to a few sentences.

For single document summarization task, CNN/Dailymail news dataset has been tried and tested to be the benchmark dataset for evaluation. This is however not the case for multi-document summarization task. Neverthless, most papers in the past 2 years have been reporting results on WikiSum dataset as the benchmark test for Multi-Document summarization task. In this project, we have picked one of those models, tweaked it to work on a two-staged mechanism and a new dataset called Multi-News.

In this section we report quantitative evaluation of our experiment with ROUGE scores. We use py-rouge library in Python to evaluate our system generated summaries. ROUGE scores have been the primary evaluation method for summary tasks as they compare overlapping n-grams between the target summaries and system generated summaries. However, the very definition of abstractive summarization means re-phrasing content from original text without using verbatim text, which renders ROUGE as not quite efficient as an evaluation method. Neverthless, in lack of any other accepted evaluation method for summarization tasks we report our ROUGE results in the following two tables and further analyse the results.

	Precision	*Recall*	*F1*
Rouge-1	36.55	35.66	35.68
Rouge-2	10.65	10.32	10.36
Rouge-3	4.88	4.68	4.71
Rouge-L	25.81	25.10	25.23

Table 1: Results from Rouge analysis for Stage 1 summaries, when compared with golden summaries in the dataset

Stage 1 summaries are individual abstractive summaries of each news article combined with all the news articles (mostly 2 or 3) in the same news event topic. The Rouge-1 F1 score of 35.68 is quite good as it is balanced between Precision and Recall with a pre-trained BERT model.

	Precision	*Recall*	*F1*
Rouge-1	48.63	15.28	22.69
Rouge-2	15.11	4.57	6.83
Rouge-3	6.83	2.00	3.01
Rouge-L	39.16	14.55	20.82

Table 2: Results from Rouge analysis for Stage 2 summaries, when compared with golden summaries in the dataset

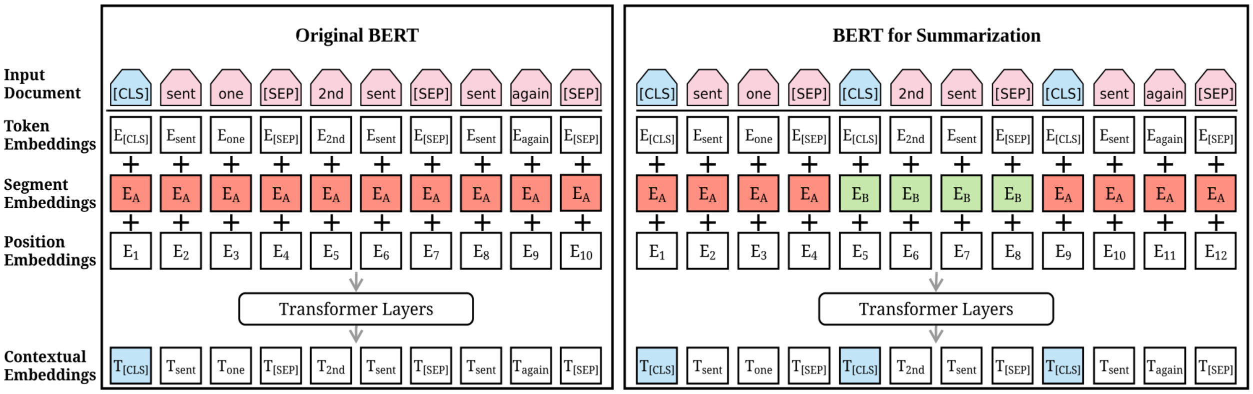
Original BERT
BERT for Summarization
Input Document
Token Embeddings
Segment Embeddings
Position Embeddings
Contextual Embeddings
[CLS] sent one [SEP] 2nd sent [SEP] sent again [SEP]
[CLS] sent one [SEP] [CLS] 2nd sent [SEP] [CLS] sent again [SEP]
$E_{[CLS]}$ E_{sent} E_{one} $E_{[SEP]}$ E_{2nd} E_{sent} $E_{[SEP]}$ E_{sent} E_{again} $E_{[SEP]}$
$E_{[CLS]}$ E_{sent} E_{one} $E_{[SEP]}$ $E_{[CLS]}$ E_{2nd} E_{sent} $E_{[SEP]}$ $E_{[CLS]}$ E_{sent} E_{again} $E_{[SEP]}$
+
E_A E_A E_A E_A E_A E_A E_A E_A E_A E_A
E_A E_A E_A E_A E_B E_B E_B E_B E_A E_A E_A E_A
E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}
E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10} E_{11} E_{12}
Transformer Layers
$T_{[CLS]}$ T_{sent} T_{one} $T_{[SEP]}$ T_{2nd} T_{sent} $T_{[SEP]}$ T_{sent} T_{again} $T_{[SEP]}$
$T_{[CLS]}$ T_{sent} T_{one} $T_{[SEP]}$ $T_{[CLS]}$ T_{2nd} T_{sent} $T_{[SEP]}$ $T_{[CLS]}$ T_{sent} T_{again} $T_{[SEP]}$

Figure 1: Comparison of original BERT with BERTSum proposed by Liu et al. (2019) for summarization task

Stage 2 summaries are abstractive summarization of stage 1 results. This stage yeilds exceptionally high Precision and quite lower recall as compared with Stage 1.

5.1 Comparison with existing models

Some of the existing prominent multi-document summarization models that we are comparing with are Transformer with attention (DMCA) by Liu, Peter J., et al.[6] (2018) on WikiSum dataset, Hierarchical Transformer by Liu et al.[5] (2019) on WikiSum dataset and Hi-Map Pointer Generator network by Fabbri et al.[7] (2019) on multi-news dataset. The below table shows a quick comparison of Rouge-1 F1 scores between these models and our model on multi-news dataset.

Model	*Dataset*	*Rouge F1*
Transformer-DMCA	WikiSum	46
Hierarchical Transformer	WikiSum	41.53
Hi-Map	Multi-News	43.47
Fine-tuned BERT (Our)	Multi-News	22.69

Table 3: Comparison with other Models

5.2 Analysis

Our model has scored quite low as compared to other models on multi-document summarization task. However, there are several points to be taken into account to put the performance of our model in context.

First, the task at hand is summarizing the text abstractively and using ROUGE to evaluate its performance can be problematic, since it looks for overlapping words. It does not really capture the qualitative aspect of summaries. Second, upon closer inspection we see our model has a pretty decent Precision but a very poor Recall. Precision captures the percentage of words from our model generated summary, which also present in the golden summary. Recall captures the percentage of words from golden summary that are also present in model generated summary. Average length of golden summaries is 260 words and for our model generated summary the average length is about 40 words. Given this huge gap, recall is bound to be lower in our model as it generates way shorter summaries.

One of the shortcomings of our model is that the summaries are too short in length. If we can increase the beam size and reduce the penalty on sentence length we could probably generate summaries that are longer in length and that might give a better Recall and overall F1 score. We have attached a few model generated summaries in table 4, after References section.

6 Conclusion

Through this project, we dive into the task of Multi-Document Summarization, its current status and conduct an experiment with a relatively new dataset in the domain. We work with we a standard encoder-decoder framework where encoder is pre-trained BERT and decoder is 6 layer Transformer which has been trained on CNN/Dailymail dataset. Out of the box, the encoder was fine-tuned for extractive summarization and decoder was fine-tuned for abstractive summarization. We apply this model for multi-document summarization on Multi-News dataset. We first summarize each individual news articles and then combine the results of each group of same event from the results. Due to the limited length of our summaries our model performs poorly on Recall but does give a decent Precision. This model can be further worked upon by training the decoder on Multi-News dataset instead of CNN/DailyMail and decreasing the penalty on sentence length. Our experiment also shows that taking the two-staged approach for multi-document datasets is not as beneficial as it is for single document summarization task.

References

[1] Subramanian, Sandeep, et al. ”On extractive and abstractive neural document summarization with transformer language models.” arXiv preprint arXiv:1909.03186 (2019).

[2] Grusky, Max, Mor Naaman, and Yoav Artzi. ”Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies.” arXiv preprint arXiv:1804.11283 (2018).

[3] Shi, Tian, Ping Wang, and Chandan K. Reddy. ”LeafNATS: An Open-Source Toolkit and Live Demo System for Neural Abstractive Text Summarization.” arXiv preprint arXiv:1906.01512 (2019).

[4] Liu, Yang, and Mirella Lapata. ”Text summarization with pretrained encoders.” arXiv preprint arXiv:1908.08345 (2019).

[5] Liu, Yang, and Mirella Lapata. ”Hierarchical transformers for multi-document summarization.” arXiv preprint arXiv:1905.13164 (2019).

[6] Liu, Peter J., et al. ”Generating wikipedia by summarizing long sequences.” arXiv preprint arXiv:1801.10198 (2018).

[7] Fabbri, Alexander R., et al. ”Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model.” arXiv preprint arXiv:1906.01749 (2019).

[8] Erkan, Günes, and Dragomir R. Radev. ”Lexrank: Graph-based lexical centrality as salience in text summarization.” Journal of artificial intelligence research 22 (2004): 457-479.

[9] Cheng, Jianpeng, and Mirella Lapata. ”Neural summarization by extracting sentences and words.” arXiv preprint arXiv:1603.07252 (2016).

[10] Nallapati, Ramesh, Bowen Zhou, and Mingbo Ma. ”Classify or select: Neural architectures for extractive document summarization.” arXiv preprint arXiv:1611.04244 (2016).

[11] Chen, Yen-Chun, and Mohit Bansal. ”Fast abstractive summarization with reinforce-selected sentence rewriting.” arXiv preprint arXiv:1805.11080 (2018).

[12] Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer networks. 2692–2700.

[13] Rush, Alexander M., Sumit Chopra, and Jason Weston. ”A neural attention model for abstractive sentence summarization.” arXiv preprint arXiv:1509.00685 (2015).

[14] See, Abigail, Peter J. Liu, and Christopher D. Manning. ”Get to the point: Summarization with pointer-generator networks.” arXiv preprint arXiv:1704.04368 (2017).

[15] Devlin, Jacob, et al. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).

QUALITATIVE RESULTS = EXAMPLES

Golden Summary	Model Generated Summary
fans of the simpsons who thought they ’ d be yelling ” doh ! ” last night were ranting ” dud ! ” this morning . viewers had been warned for months that what was rumored to be a major character would be killed off during the premiere of the show ’ s 26th season last night , the los angeles times reports . there was a character who kicked the cartoon bucket , but it wasn ’ t exactly a major one : it was rabbi hyman krustofski , krusty the clown ’ s dad , voiced by comedian jackie mason , a character that only appeared in ” a handful ” of episodes , the times notes . reaction to the ” clown in the dumps ” episode ranged from mild , npr-style disappointment to outright irritation : tim donnelly writes in the new york post that the death (and preceding teasers) were ” a lame play for attention by a show desperate to stay relevant . ” one twitter user quoted in the times complained , ” krusty ’ s dad died ... um , krusty had a dad ? never heard of him . wasted anticipation . ” producer al jean insists he has always said the untimely death was ” overhyped ” and that he never promised it would be one of the more-popular characters . ” i never said it ’ s an iconic character — i never used those words , ” he tells entertainment weekly . in fact , he assures fans that favorites will never be purposely annihilated before series ’ end . ” we ’ re never going to kill off homer , or even krusty , ” he tells tvline . ” this show is always running in syndication , and we don ’ t want you to feel bad every time you see an old character that you loved . ”	the simpsons ’ 26th season premiere saw the highly anticipated demise of a-beloved resident on sunday. a-beloved resident has been a part of the simpsons universe since reuniting with his estranged , red-nosed son in 1991
the us stands by the ” one-china ” policy , but that doesn ’ t mean it can ’ t sell weapons directly to taiwan , citing ithe taiwan relations act to ensure taiwan can adequately defend itself — and china isn ’ t happy about it . the obama administration announced a $ 1.8 billion arms package sale to congress on wednesday , reuters reports , including guided-missile frigates , anti-tank missiles , amphibious assault vehicles , and $ 416 million worth of guns , ammo , and other supplies . the announcement came amid reports that the us had stalled the sale to avoid hearing about it from china , which still claims taiwan as a territory , per the wall street journal . reuters notes the sale comes as us-china relations simmer over the latter ’ s man-made islands in the south china sea and us patrols in those waters . china notes it ’ s going to sanction the companies involved in the sale (including lockheed martin and raytheon) , with a foreign ministry official telling xinhua that the sale flouts international rules and ” severely ” damages china ’ s sovereignty . ” china ’ s government and companies will not carry out cooperation and commercial dealings with these types of companies , ” a ministry spokesman says . a pentagon spokesman gave the equivalent of an eyeroll wednesday , per the new york times , noting , ” the chinese can react to this as they see fit . . . . it ’ s a [clear-eyed] , sober view of an assessment of taiwan ’ s defense needs . . . . there ’ s no need for it to have any derogatory effect on our relationship with china . ” meanwhile , the ap notes that china has issued similar threats before , with ” no evidence they ’ ve had any meaningful effect . ” (all this despite a lengthy handshake last month.)	the obama administration announced a $ 1.83 billion arms sale on wednesday. the sale includes two decommissioned navy frigates , air and ground missiles , amphibious vehicles and communications systems.

QUALITATIVE RESULTS = EXAMPLES

Golden Summary	Model Generated Summary
a day after hundreds of thousands of young people took to the streets to call for gun control , an old man used his bully pulpit to urge them to keep shouting , reports reuters . speaking at his palm sunday mass , 81-year-old pope francis warned that ” the temptation to silence young people has always existed , ” along with ways ” to sedate them , to keep them from getting involved , to make their dreams flat and dreary , petty and plaintive . ” but , reports the ap , he told young people that ” it is up to you not to keep quiet . even if others keep quiet , if we older people and leaders , some corrupt , keep quiet , if the whole world keeps quiet and loses its joy , i ask you : will you cry out ? ” the response from the crowd : ” yes ! ”	pope francis urges young people to keep shouting and not allow older generations to silence their voices. the 81-year-old led a long and solemn palm sunday service before tens of thousands in the square pope
a big win for samsung in its long-running patent feud with apple : the us international trade commission has banned imports of the at&t models of older apple products including the iphone 4 and ipad 2 3g after deciding apple violated a samsung patent , the wall street journal reports . newer apple products like the iphone 5 are not affected by the ruling , which apple says it is ” disappointed ” by and will appeal . the ruling will take effect in 60 days unless it is vetoed by president obama , a move analysts say is nearly as unlikely as the two companies deciding to settle their difference amicably . ” there ’ s too much skin in the game now , ” a spokesman for technology research firm idc tells bloomberg . ” it ’ s almost so ugly i don ’ t think they ’ ll come to any agreement . both companies have a lot of cash and are generating a lot of money . it ’ s not like they have to worry about paying the legal bills . ”	the u.s. international trade commission ruled that apple violated a patent covering technology used to send information over wireless networks. the ruling would bar the importation of certain iphones and ipads made to work on at&t ’s network
remember the boundary-busting french satirical newspaper that was firebombed for making the prophet mohammed a ” guest editor ” ? journalists toned down the controversy this time around — not . in fact , right on the cover , a muslim is planting a big , slobbery kiss on a figure representing the publication , charlie hebdo . above the embrace are the words : ” love is stronger than hate . ” the guardian says the paper ” isn ’ t holding back , ” while gawker — convinced the muslim is a ” gay mohammad ” (though he ’ s not in the garb of the prophet) — calls it the ” ballsiest paper in the world . ” charlie hebdo ’ s editor said after the firebombing that ” freedom to have a good laugh is as important as freedom of speech . ” the french , including some muslim leaders , have strongly supported the publication , which is now operating out of the offices of the left-wing paris newspaper liberation . ” i am extremely attached to the freedom of the press , even if the press is not always tender with muslims , islam , or the paris mosque , ” said the head of the paris mosque .	charlie hebdo ’s offices have been firebombed , its website hacked , its facebook page suspended for 24 hours and its staff targeted with death threats