Applying Extractive Summarization using Global and Local Context to Long Government Documents

Danielle Sim simd@usc.edu
Erin Szeto erinszet@usc.edu
Haley Massa hmassa@usc.edu
Hee Ji Park heejipar@usc.edu
Surya Solanki suryasol@usc.edu

Abstract

In this paper, we apply an extractive summarization model (referred to in this paper as the GLCES model), introduced in (Xiao and Carenini, 2019), to a dataset containing long government documents, GovReport, to analyze challenges with respect to document length and differing domain. We compare four summaries: (1) main summarization model GLCES, (2) baseline summarization model SumBasic, (3) ground truth extractive summaries from the generated extractive labels, and (4) GovReport's ground truth abstractive summaries. The summaries are evaluated using ROUGE scores and human evaluation on 25 sampled documents. We find that the GLCES model outperforms the baseline model and maintains relevancy, but when compared to the ground truth extractive and abstractive summaries, the GLCES summaries perform poorly in regards to consistency and coherence.

1 Introduction

Long document summarization is a task that involves generating a shorter summary of a document that provides a quick and coherent understanding of the text. Reading long documents is a time-consuming and difficult task, and summarizing documents on domains such as medicine, history, and law is challenging since these domains contain distinct and advanced vocabulary.

In Natural Language Processing there are two main methods for long document summarization: abstractive and extractive. Abstractive summarization methods generate brand new summary sentences to imitate how a domain expert would manually generate a summary of a document. Extractive summarization methods aggregate chosen sentences from the main document based on metrics to evaluate importance of those sentences to generate a summary. Abstractive and extractive methods have their respective strengths and weaknesses - abstractive methods are perform better than extractive methods in terms of avoiding redundancy but struggle with fluency. On the other hand, extractive methods may struggle with redundancy, but still produce summaries that read better in terms of fluency and grammar and avoid false information. One specific research area in extractive summarization methods involve using documents that are particularly long (longer than 1000 words).

A model published by Xiao and Carenini 'Extractive Summarization of Long Documents by Combining Global and Local Context' (Xiao and Carenini, 2019) works to use section information in the documents to generate extractive summaries of particularly long sentences and documents using both global and local context. Xiao and Carenini specifically apply their model to scientific papers (PubMed and arXiv datasets) which contain structures of multiple subsections throughout the document, all following similar structures that are commonly found in research publications (introduction, methods, results, discussion, related words, etc).

In this work we aim to apply the model of (Xiao and Carenini, 2019) (referred to in this paper as the GLCES model - global and local context extractive summarization model) to a different dataset, GovReport, to evaluate performance in light of two challenges: documents of longer length and of a different domain.

The GovReport dataset comes from (Huang et al., 2021), a collection of nearly 19.5k reports written by government agencies such as the Congressional Research Service and U.S. Government Accountability Office. The average document length is around 9.4k words and the average summary length is 553 words, both of which are significantly longer than the average documents in the arXiv data set (4938 words/document, 220 words/summary) and the PubMed data

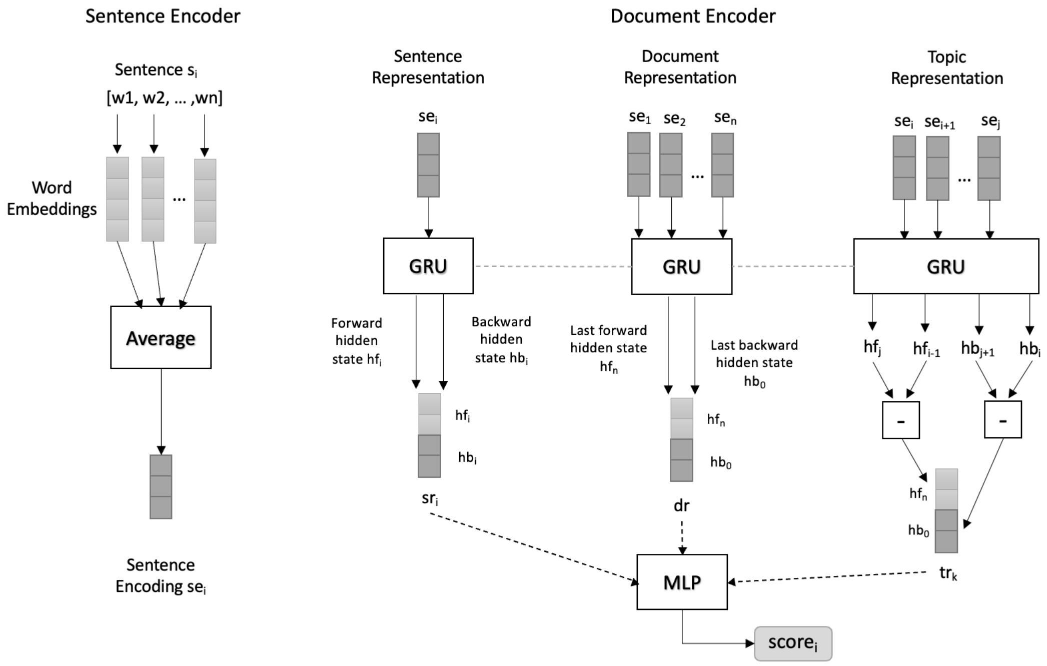


Figure 1: GLCES Model Architecture

set (3016 words/document, 203 words/summary) used in (Xiao and Carenini, 2019). A challenge the longer government reports brings is that the summary quality from GLCES may be poor due to the model architecture. The GLCES model's decoder layer consists of a multi-layer perceptron and does not consider whether previous sentences have been selected as part of the extracted summary; this could lead to higher redundancy in the extracted summaries of lengthy documents. A second challenge of the GovReport dataset is that government documents follow structures that vary and are specific to each paper (i.e. the subsections no longer follow similar structures as introduction, methods, results, etc as described earlier) and some documents feature topics that are harder to summarize, such as financial reports.

To this effect, our project focuses on analysis of four summaries: (1) the main summarization model GLCES summary (2) the baseline summarization model SumBasic summary, (3) the ground truth extractive summary from the generated extractive labels, and (4) GovReport's ground truth abstractive summary. We find that the GLCES model outperforms the baseline model and maintains relevancy, but when compared to the ground truth extractive and abstractive summaries, the GLCES summaries perform poorly in regards to consistency and coherence.

2 Methods

2.1 Extractive Summarization of Long Documents Combining Global and Local Context (GLCES)

(Xiao and Carenini, 2019) proposed a novel neural extractive summarization model, GLCES, for single long documents that exploits section information by capturing local context (e.g. sections, chapters) and global context (e.g. whole document) of documents.

The model's architecture, shown in Figure 1, contains three components: a sentence encoder, document encoder, and sentence classifier. Average Word Embedding is used as the sentence encoder and bi-directional gated recurrent units (GRU) encode the document sequentially forward and backward. To capture the local context of each sentence, the model applies the LSTM-Minus method (Wang and Chang, 2016), originally proposed to help with segment embedding. Once obtained, these three representations (sentence, topic, and document) are combined using attentive context to make a prediction on whether the sentence should be included in the extractive

Model	Compare against ground truth extractive			Compare against ground truth abstractive		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE1	ROUGE-2	ROUGE-L
SumBasic	47.64	20.63	22.63	43.67	14.07	18.45
GLCES	62.00	40.06	56.52	56.36	27.60	47.58

Table 1: Automatic evaluation metric scores of the SumBasic and GLCES models' summaries compared against the ground truth extractive and abstractive summaries.

summary.

3 Experiments

3.1 Training Details

In order to train the GLCES model, we needed to reformat the GovReport data to follow the example data in (Xiao and Carenini, 2019) and generate extractive summary labels. Because GLCES is an extractive summarization model and GovReport dataset contains abstractive summaries, we used an algorithm from (Kedzie et al., 2018) to construct extractive summary labels, in which sentences with the highest ROUGE-1 (Lin, 2004) score when compared to the gold-standard abstractive summaries of each report are labeled as part of the ground truth extractive summary. Details of the data preprocessing and restructuring of the GovReport data are included in Appendix A, and details of the extractive label generation algorithm are shown in Appendix B.

3.2 Compared Summaries

We compare four different summaries and outline implementation details for the GLCES model and SumBasic model in Appendix C.

- **GLCES** (Xiao and Carenini, 2019), the novel neural extractive summarization model that combines global and local context.

- **SumBasic** (Vanderwende et al., 2007), the baseline model, which is a traditional extractive summarization model.

- **Ground truth extractive**, the extractive summary that comes from the generated extractive labeled sentences of the GovReport dataset.

- **Ground truth abstractive**, the abstractive summary included in the GovReport dataset, handwritten by experts.

3.3 Automatic Evaluation Metrics

Automated evaluation aids system development and avoids labor-intensive and potentially inconsistent human evaluation (Liu and Liu, 2010). There are several automated evaluation methods, but in our paper, we use ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which is a standard evaluation measure.

- **ROUGE-N**: This method has been shown to be effective in capturing n-gram overlap between the automatic summaries and the reference summaries, such as unigram, bigram, and trigram (Lin, 2004), (Ganesan, 2018).

- **ROUGE-L**: This method measures the longest matching string using the LCS technique. The advantage of LCS is that it does not require consecutive matching of words like ROUGE-2, but rather measures the matching that occurs within a string, allowing more flexible performance comparison (Lin, 2004).

3.4 Experimental Results

We use ROUGE-1, ROUGE-2, and ROUGE-L scores to compare the GLCES and SumBasic summaries against the ground truth extractive and ground truth abstractive summaries. The scores are displayed in Table 1.

First, we compare the GLCES and SumBasic summaries against the ground truth extractive. In Table 1, the GLCES model has higher ROUGE-1 and ROUGE-2 scores than SumBasic. This indicates that the summaries generated by the GLCES contain many keywords included in the ground truth summaries. The ROUGE-L score is also higher in GLCES, showing that the summaries generated by GLCES not only have many of the same words but also many of the same sentences as the ground truth summaries.

Second, we compare the GLCES and SumBasic summaries against the ground truth abstractive

Summary	Relevance	Consistency	Non-redundancy	Fluency	Coherence
SumBasic	2.24	1.68	2.44	2.32	2.00
GLCES	3.00	1.88	3.60	3.40	2.60
Ground truth extractive	4.00	3.24	4.0	3.60	3.20
Ground truth abstractive	4.80	4.64	4.52	4.68	4.52

Table 2: Average scores of human evaluation across the four summaries

summaries. All ROUGE scores of the GLCES model summaries are higher than those of SumBasic. This demonstrates that the GLCES model outperforms the baseline model and is more likely to produce higher-quality summaries that are more similar to the ground truth abstractive summaries.

4 Human Evaluation & Analysis

4.1 Limitations of Automatic Evaluation Metrics

The ROUGE score measures how many words appear in the generated summary and the reference summary and how the order of the words match. This is a limitation because if, for example, the generated summary contains synonymous words to the reference summary and has a different, but coherent order of words, the summary may get a lower ROUGE score even though it was not of poor quality. Thus, trying to increase the ROUGE score may result in harming the expressive diversity of the summary. For this reason, many papers provide not only the ROUGE score, but also the expensive human evaluation results.

4.2 Details of Human Evaluation

Human evaluation has been the most trusted evaluation method and used as the gold standard for summarization evaluation (Gatt and Krahmer, 2017). In order to reduce subjectivity and evaluate summaries consistently, absolute evaluation criteria should be established. We set the following five criteria to judge a good summary, referring (Fabbri et al., 2020), (Bražinskas et al., 2020) and (Jia et al., 2021). We rate each summary on a scale from 1 (worst) to 5 (best).

- **Relevance:** The rating measures whether the summary captures the key point of the article. Consider whether some or all of the crucial information is included in the summary.
- **Consistency:** The rating measures whether each sentence is well-placed for each information. Consider whether the sentence flows well.
- **Non-redundancy:** The rating measures whether the summary contains unnecessary repetition sentences. Redundant sentences do not mean only sentences with the same word composition. Where words have different uses but have the same meaning, we define them as duplicate sentences.
- **Fluency:** The rating measures whether the summary is overall easy to read and understand. Consider the quality of the summary as a whole.
- **Coherence:** The rating measures whether the summary are well structured and well organized overall.

4.3 Results

We randomly sampled 25 government reports and scored each reports' four summaries on relevance, consistency, non-redundancy, fluency, and coherence. The average scores of these five criteria across the four summaries are shown in Table 2.

From these results, we find that the ground truth abstractive summaries are ranked first across the five criteria and are deemed the summaries with the highest quality. The ground truth extractive summaries are ranked second, the GLCES summaries are third, and the SumBasic summaries are fourth.

4.4 Error Analysis

When performing the human evaluation on the sampled GovReport documents, we found that the GLCES model summaries contain relevant and easy to read sentences. The summaries were not very redundant, indicating that the GLCES model's architecture did not lead to high redundancy. Although the GLCES summaries achieved scores of 3 or greater in relevancy and fluency, the summaries performed poorly in terms of consistency and coherence. A large issue with the

GLCES summaries was while individual sentences flowed well, the overall ideas and topics that are expressed in the documents are not well reflected, as well as inconsistent ordering of sentences. A sample summary is shown in Appendix E.

Additionally, depending on the report topic, the GLCES summaries did not capture pertinent details of the report. For instance, for a report dedicated to understanding the budgets for stockpile stewardship, the GLCES summary had 9 sentences on budget details while the ground truth abstractive summary had around 25 sentences related to budgeting. The GLCES model did not seem to grasp the importance of the financial numbers in the document, and this may be due to the nature of the government domain. This could also be due to the challenge of generating extractive summary labels. Because the generated labels were constructed from an algorithm with no human validation, the quality of the labels may be poor and the model may not have learned well.

Comparing the SumBasic, GLCES, and ground truth extractive summaries with the ground truth abstractive summaries, the ground truth abstractive summaries had a better flow and summarized the report well from start to finish. Because the ground truth abstractive summaries are written by domain experts, these summaries were found to be the strongest overall, especially from the perspective of being a reader with no domain knowledge and reading the summaries with the intent to get a quick and coherent understanding of the documents. The ground truth extractive summaries were also found to be strong with good fluency and relevancy, but they were relatively weaker compared to the ground truth abstractive summaries and often missed the overview of the document and lacked coherence. This could be because not only are the ground truth abstractive summaries written by domain experts, but they also are written with specific intent to capture document ideas and topics as a whole, whereas the ground truth extractive summaries contain selected sentences from specific portions of the documents.

5 Conclusion

In this project we apply Xiao and Carenini's GLCES model (global and local context extractive summarization) to a new dataset to analyze its performance with respect to two different challenges: a different domain and longer documents. To analyze the scope of this model we compare it to a baseline summarization model SumBasic, as well as two ground truth summaries, one extractive and abstractive. Evaluation is done using automatic evaluation metrics of ROUGE-1, ROUGE-2 and ROUGE-L scores and human evaluation on summary relevance, consistency, non-redundancy, fluency and coherence.

Our results from both of these evaluation methods indicate that the GLCES model routinely outperforms the SumBasic model, but its output summaries are not quite on par with the ground truth extractive and ground truth abstractive summaries. We concluded that this outcome was because although the GLCES model consistently selected relevant, informative sentences from the original document, it failed to aggregate them in a coherent and understandable manner. Overall, the results from our experiments are promising that successful long document summarization is achievable, but there is still significant progress to be made in the field.

Code

The code can be found at `https://github.com/erinszeto/csci544-project`.

Dataset

The GovReport dataset can be found at `https://gov-report-data.github.io/`.

Video

The demo video with a brief discussion of our project can be found at `https://www.youtube.com/watch?v=pwKDke-jchE`

References

[Bražinskas et al.2020] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-Shot Learning for Opinion Summarization. *arXiv e-prints*, page arXiv:2004.14884, April.

[Fabbri et al.2020] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. SummEval: Re-evaluating Summarization Evaluation. *arXiv e-prints*, page arXiv:2007.12626, July.

[Ganesan2018] Kavita Ganesan. 2018. ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks. *arXiv e-prints*, page arXiv:1803.01937, March.

[Gatt and Krahmer2017] Albert Gatt and Emiel Krahmer. 2017. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *arXiv e-prints*, page arXiv:1703.09902, March.

[Huang et al.2021] Luyang Huang, Shuyang Cao, Nikolaus Nova Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *CoRR*, abs/2104.02112.

[Jia et al.2021] Ruipeng Jia, Yanan Cao, Fang Fang, Yuchen Zhou, Zheng Fang, Yanbing Liu, and Shi Wang. 2021. Deep differential amplifier for extractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 366–376, Online, August. Association for Computational Linguistics.

[Kedzie et al.2018] Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium, October-November. Association for Computational Linguistics.

[Lin2004] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

[Liu and Liu2010] Feifan Liu and Yang Liu. 2010. Exploring correlation between rouge and human evaluation on meeting summaries. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18:187 – 196, 02.

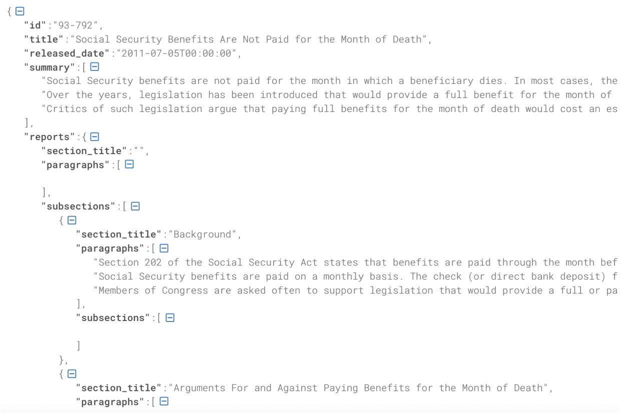
[Vanderwende et al.2007] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing Management*, 43.

[Wang and Chang2016] Wenhui Wang and Baobao Chang. 2016. Graph-based dependency parsing with bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2315, Berlin, Germany, August. Association for Computational Linguistics.

[Xiao and Carenini2019] Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context.

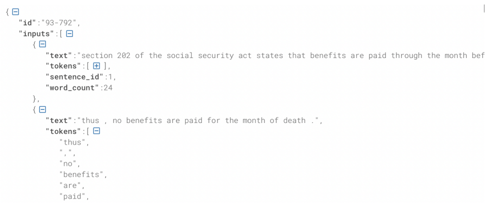
A Data Preprocessing

Data preprocessing steps were modeled after the example input for the GLCES model. Documents are converted to lowercase, and sentences are listed one-by-one. The number of words for each sentence are counted after word-level tokenziation. Subsections in the documents are noted and number of sentences per subsection (section length) are noted as well. Tokenziation on punctuation such as periods, commas and dashes and grammar structure such as apostrophes are all made to follow the GLCES model. The Figures 2, 3 and 4 show a sample document both before and after preprocessing steps.



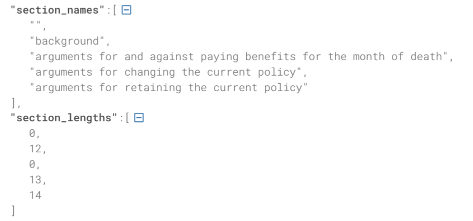
```
{ ⊟
   "id":"93-792",
   "title":"Social Security Benefits Are Not Paid for the Month of Death",
   "released_date":"2011-07-05T00:00:00",
   "summary":[ ⊟
      "Social Security benefits are not paid for the month in which a beneficiary dies. In most cases, the
      "Over the years, legislation has been introduced that would provide a full benefit for the month of
      "Critics of such legislation argue that paying full benefits for the month of death would cost an es
   ],
   "reports":{ ⊟
      "section_title":"",
      "paragraphs":[ ⊟

      ],
      "subsections":[ ⊟
         { ⊟
            "section_title":"Background",
            "paragraphs":[ ⊟
               "Section 202 of the Social Security Act states that benefits are paid through the month bef
               "Social Security benefits are paid on a monthly basis. The check (or direct bank deposit) f
               "Members of Congress are asked often to support legislation that would provide a full or pa
            ],
            "subsections":[ ⊟

            ]
         },
         { ⊟
            "section_title":"Arguments For and Against Paying Benefits for the Month of Death",
            "paragraphs":[ ⊟
```

Figure 2: Sample GovReport document before preprocessing



```
{ ⊟
   "id":"93-792",
   "inputs":[ ⊟
      { ⊟
         "text":"section 202 of the social security act states that benefits are paid through the month bef
         "tokens":[ ⊞ ],
         "sentence_id":1,
         "word_count":24
      },
      { ⊟
         "text":"thus , no benefits are paid for the month of death .",
         "tokens":[ ⊟
            "thus",
            ",",
            "no",
            "benefits",
            "are",
            "paid",
```

Figure 3: Sample GovReport document after preprocessing (head)



```
   "section_names":[ ⊟
      "",
      "background",
      "arguments for and against paying benefits for the month of death",
      "arguments for changing the current policy",
      "arguments for retaining the current policy"
   ],
   "section_lengths":[ ⊟
      0,
      12,
      0,
      13,
      14
   ]
```

Figure 4: Sample GovReport document after preprocessing (tail)

B Extractive Label Generation

Algorithm 1 is used to generate extractive labels from abstractive summaries of government report data. It is a modified version of the algorithm used in (Xiao and Carenini, 2019) and (Kedzie et al., 2018). For the word budget argument, 10% of the report's total word count is used. If at least five sentences have been picked and the past three sentences chosen are the same sentence, which is redundant, then the algorithm will break out of the loop and return the list of unique sentences that are chosen as the extractive labels. This modification is made because running the algorithm to obtain the labels is very time-consuming.

Algorithm 1: Extractive label generation

```
def labelGeneration(ref,sentences,budget):
   hyp = ''
   wc = 0
   picked = []
   highest_r1 = 0
   sid = -1

   while wc <= budget:
      for i in range(len(sentences)):
         ##fmeasure of ROUGE1
         score = scorer.score(hyp+sentences[i],ref)['rouge1'][2]
         if score > highest_r1:
            highest_r1 = score
            sid = i

      if (len(picked) > 5) and (picked[-1] == sid) and (picked[-2] == sid):
         break
      elif sid != -1:
         picked.append(sid)
         hyp = hyp+sentences[sid]
         wc += wordCount(sentences[sid])
      else:
         break

   ## unique sentences
   picked = list(set(picked))

   return picked
```

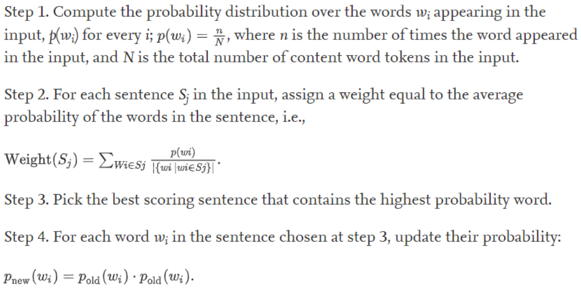
C Implementation Details

Listed below are the Hyperparameters used while training the GLCES Model:

- Decoding Method: Attentive Context
- Epochs: 20
- Summary word count: 553

- Batch size: 32
- Embedding dimension: 300
- MLP dimension: 100
- Hidden dimension: 300

SumBasic, the baseline model we used, is an extractive summarization system designed for topic-focused multi-document summarization. The motivation behind the algorithm is that words occurring frequently in the document cluster occur in summaries with higher probability. Figure 5 details the steps of the algorithm. Due to the small word probabilities, we had to calculate the probabilities in the log space.



Step 1. Compute the probability distribution over the words w_i appearing in the input, $p(w_i)$ for every i; $p(w_i) = \frac{n}{N}$, where n is the number of times the word appeared in the input, and N is the total number of content word tokens in the input.

Step 2. For each sentence S_j in the input, assign a weight equal to the average probability of the words in the sentence, i.e.,

$\text{Weight}(S_j) = \sum_{Wi \in Sj} \frac{p(wi)}{|\{wi \, | wi \in Sj\}|}$.

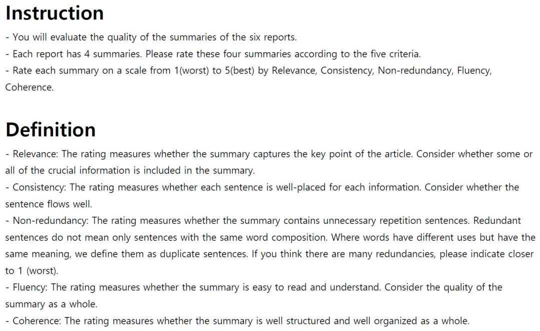
Step 3. Pick the best scoring sentence that contains the highest probability word.

Step 4. For each word w_i in the sentence chosen at step 3, update their probability:

$p_{\text{new}}(w_i) = p_{\text{old}}(w_i) \cdot p_{\text{old}}(w_i)$.

Figure 5: SumBasic Implementation

D Human evaluation



Instruction

- You will evaluate the quality of the summaries of the six reports.
- Each report has 4 summaries. Please rate these four summaries according to the five criteria.
- Rate each summary on a scale from 1(worst) to 5(best) by Relevance, Consistency, Non-redundancy, Fluency, Coherence.

Definition

- Relevance: The rating measures whether the summary captures the key point of the article. Consider whether some or all of the crucial information is included in the summary.
- Consistency: The rating measures whether each sentence is well-placed for each information. Consider whether the sentence flows well.
- Non-redundancy: The rating measures whether the summary contains unnecessary repetition sentences. Redundant sentences do not mean only sentences with the same word composition. Where words have different uses but have the same meaning, we define them as duplicate sentences. If you think there are many redundancies, please indicate closer to 1 (worst).
- Fluency: The rating measures whether the summary is easy to read and understand. Consider the quality of the summary as a whole.
- Coherence: The rating measures whether the summary is well structured and well organized as a whole.

Figure 6: Sample Human Evaluation

E Sample GLCES Summary

Sample GLCES summary for The Independent Payment Advisory Board: 111-148 , as amended) created the independent payment advisory board (ipab , or the board) to ”reduce the per capita rate of growth in medicare spending. . the board’s proposals will be implemented by the secretary of health and human services (the secretary) unless congress acts either by formulating its own proposal to achieve the same savings or by discontinuing the automatic implementation process defined in the statute . the patient protection and affordable care act (ppaca , p.l . the report then describes the structure of the board , the calculations and determinations required to be made by the office of the chief actuary (the chief actuary) in the centers for medicare medicaid services (cms) that trigger a board proposal , and the content of and constraints on board proposals — including the medicare productivity exemptions under section 3401 of ppaca . among some proponents of health care reform , a major impetus for reform , in addition to improving quality and increasing access , has been the rising cost of the medicare program . in addition , the report reviews the expedited and other parliamentary procedures that relate to congressional consideration of board proposals and other board - related activities , and concludes with a description of how the board’s proposals are to be implemented and their possible impact . this report , which provides an overview of the board , begins with a discussion of the rationale behind the creation of an independent medicare board and briefly reviews prior proposals for similar boards and commissions . appendix a details key dates for ipab implementation and various reports required by the law , and appendix b compares the ipab with the medicare payment advisory commission (medpac) . the explicit charge given by ppaca to the board in section 3403 (b) is to ”reduce the per capita rate of growth in medicare expenditures. . recommendations relating to payments to plans under medicare parts c and d and recommendations relating to payment rate changes that take effect on a calendar year basis take effect on january 1 of the iy . in addition , the government accountability office (gao) is directed , as described below , to undertake a review of the board’s initial recommendations and report to congress by july 1 , 2015 . this joint resolution requires a super - majority vote of both chambers and either the signature of the president or overriding his veto by a two - thirds vote in each house to enter into force . since some providers and suppliers of services will receive a reduction in payments beyond their productivity adjustment in some years , section 3403 (c) (2) (iii) prohibits , as described below , the board from recommending in some years further reduction in payment rates to those providers and suppliers . 452 was combined with the help efficient , accessible , low - cost , timely healthcare (health) act of 2011 (h.r . the national commission on fiscal responsibility and reform , popularly referred to as the simpson - bowles deficit commission , proposed two sets of recommendations (recommendations 3.5 and 3.6) regarding ipab .