

# At Which Level Should We Extract? An Empirical Analysis on Extractive Document Summarization

**Qingyu Zhou\***

Tencent Cloud Xiaowei  
Beijing, China

qingyuzhou@tencent.com

**Furu Wei, Ming Zhou**

Microsoft Research Asia  
Beijing, China

{fuwei, mingzhou}@microsoft.com

## Abstract

Extractive methods have been proven effective in automatic document summarization. Previous works perform this task by identifying informative contents at sentence level. However, it is unclear whether performing extraction at sentence level is the best solution. In this work, we show that unnecessary and redundancy issues exist when extracting full sentences, and extracting sub-sentential units is a promising alternative. Specifically, we propose extracting sub-sentential units based on the constituency parsing tree. A neural extractive model which leverages the sub-sentential information and extracts them is presented. Extensive experiments and analyses show that extracting sub-sentential units performs competitively comparing to full sentence extraction under the evaluation of both automatic and human evaluations. Hopefully, our work could provide some inspiration of the basic extraction units in extractive summarization for future research.

## 1 Introduction

Automatic text summarization aims to produce a brief piece of text which can preserve the most important information in it. The important contents are identified and then extracted to form the output summary (Nenkova and McKeown, 2011). In recent decades, extractive methods have proven effective in many systems (Carbonell and Goldstein, 1998; Mihalcea and Tarau, 2004; McDonald, 2007; Cao et al., 2015; Cheng and Lapata, 2016; Zhou et al., 2018; Nallapati et al., 2017).

In previous works, extractive summarization systems perform extraction on the sentence level (Mihalcea and Tarau, 2004; Cheng and Lapata, 2016; Nallapati et al., 2017). As the extraction unit, a sentence is a grammatical unit of one or more words that express a statement, question, request, etc. There are several advantages of extracting sentences. First, extractive systems are simpler, easier to develop, and faster during run-time in real application scenarios, compared with abstractive systems. Moreover, original sentences in the input are naturally fluent and grammatically correct. Finally, extracted sentences are factually faithful to the input document, compared with abstractive methods (Cao et al., 2018).

Despite the success of extractive systems, from previous works, it is still not clear whether extracting at sentence level is the best solution for extractive methods. There are several drawbacks of extracting the full sentences. The most obvious issue is that the extracted sentences may contain unnecessary information. Some previous works have also noticed this problem and try to solve it by compressing or rewriting the extracted sentences (Martins and Smith, 2009; Chen and Bansal, 2018; Xu and Durrett, 2019). Furthermore, extracted sentences may contain duplicate contents. Thus, methods such as Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) and sentence fusion (Barzilay and McKeown, 2005; Lebanoff et al., 2019) are proposed to avoid or merge duplicate contents.

The redundancy and unnecessary issues might be caused by extracting full sentences since an important sentence may also contain unnecessary information. Besides, different importance sentences may have duplicate (un)important words. This inspires us that we can perform extraction at a finer granularity so that the important and unimportant contents could be separated. Therefore, we argue that extracting

\* Contribution done during internship at Microsoft Research Asia while pursuing PhD at Harbin Institute of Technology. This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

sub-sentential units in a sentence could be a solution. As for the sub-sentential units, we mainly focus on the non-terminal nodes in a constituency parsing tree in this paper. In a given parsing tree of a sentence, the root node represents the entire sentence while the leaf nodes represent each corresponding lexical token. An extractive system could perform extraction on the non-terminal nodes which can express more fine-grained information. To keep the advantages of extractive methods, we choose the nodes which can still express a full statement. Specifically, the nodes with the clause tag such as S and SBAR are used for creating extraction units.

In this paper, we conduct experiments and analyses to answer the following questions:

- Q1 Does extracting full sentences introduce unnecessary or duplicate information? (§3)
- Q2 Can extracting sub-sentential units solve these problems of full sentence extraction? How to perform sub-sentential unit extraction? (§4)
- Q3 Can extracting sub-sentential units improve the performance of extractive document summarization systems? (§5, §6.1, §6.2)
- Q4 Does extracting sub-sentential units cause any other issues? (§6.4)

## 2 Related Work

Extractive document summarization methods have proven effective and been extensively studied for decades. As an effective approach, extractive methods are popular and dominate the summarization research. As the very first work in 1950s, Luhn (1958) uses lexical frequency to determine the importance of a *sentence*. Many research works further develop extractive summarization methods. The most important step is to determine the sentence importance. Many different methods have been proposed and they can be roughly categorized from different perspectives.

From the perspective of having supervision or not, there are two major types: unsupervised methods and supervised methods. One of the difficulties in training an extractive system is the lack of extraction labels. The reason is that most of the reference summary is written by human experts, therefore, it is hard to find the exact appearance in the input document. Without natural training labels, unsupervised and supervised methods treat extractive summarization as different problems.

Graph-based methods (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Wan and Yang, 2006) are very useful unsupervised methods. In these methods, the input document is represented as a connected graph. The vertices represent the sentences, and the edges between vertices have attached weights that show the similarity of the two sentences. The score of a sentence is the importance of its corresponding vertex, which can be computed using graph algorithms.

Supervised methods for extractive summarization create training labels manually. (Cao et al., 2015; Ren et al., 2017) directly train regression models using ROUGE scores as the supervision. (Cheng and Lapata, 2016; Nallapati et al., 2017; Zhou et al., 2018; Zhang et al., 2019) search the oracle extracted sentences as the training labels. Cheng and Lapata (2016) propose treating document summarization as a sequence labeling task. They first encode the sentences in the document and then classify each sentence into two classes, i.e., extraction or not. Nallapati et al. (2017) propose a system called SummaRuNNer with more features, which also treat extractive document summarization as a sequence labeling task. Zhou et al. (2018) propose using pointer networks (Vinyals et al., 2015) to repeatedly extract sentences.

Recently, Reinforcement Learning (RL) is also introduced in unsupervised extractive summarization (Dong et al., 2018; Böhm et al., 2019). Dong et al. (2018) treat sentence extraction as a Bandit problem so that they can train a RL-based system whose reward is the ROUGE scores. Böhm et al. (2019) propose that using human judgments as reward is better than ROUGE. However, these methods still need reward as a signal, which differ from the previously introduced fully unsupervised methods.

## 3 Q1: Review of Extracting Full Sentence

Performing sentence extraction in summarization systems have proven effective in previous works (Luhn, 1958; Cao et al., 2015; Cheng and Lapata, 2016; Nallapati et al., 2017). Despite the success of these

systems, it is still unclear whether performing content extraction at the sentence level is the best solution. In this section, we will examine the drawbacks of extraction at the sentence level.

### 3.1 The Dataset

There are various datasets for text summarization, such as DUC/TAC, *CNN/Daily Mail*, New York Times, etc. In this paper, we take the most commonly used dataset in recent research works (Cheng and Lapata, 2016; Nallapati et al., 2017; Zhou et al., 2018; Xu and Durrett, 2019; Lebanoff et al., 2019), *CNN/Daily Mail*, as our testbed. The statistics of it can be found in Table 2. One of the most distinguishable features of this dataset is that the output summary is in the form of highlights written by the news editors. As shown in the example in Figure 1, the summary (highlights) is a list of bullets. Therefore, extractive methods perform well on this dataset (Grusky et al., 2018).

| Story highlights   | Marseille, France (CNN)   |
|--|---|
| Lufthansa CEO promises to help victims' families for as long as needed as he visits crash site   | — The French prosecutor leading an investigation into the crash of      |
| Marseille prosecutor says "so far no videos were used in the crash investigation" despite media reports  | Germanwings Flight 9525 insisted Wednesday that he was not aware of any |
| Journalists at Bild and Paris Match are "very confident" a video clip is real, an editor says  | video footage from on board the plane.                                  |
| Marseille prosecutor Brice Robin, in charge of the criminal inquiry into the crash, told CNN that "so far no videos were used in the crash investigation." |   |

Figure 1: A screenshot example of the document-summary pair in the *CNN/Daily Mail* dataset.

### 3.2 The Drawbacks

There are two main potential drawbacks of extracting sentences. First, unnecessary information is smuggled with the extracted sentences. Second, duplicate content may appear when extracting multiple sentences. To analyze whether the issues exist, we conduct experiments and analyses with both count-based statistics and human judgments. We consider two different settings to reach our final conclusion, i.e., the *extractive oracle* and a real extractive system.

First, we check the quality of the *sentence level extractive oracle*, since it is the upper bound of any extraction system. Two different methods are used in recent extractive summarization research for building the oracle training label. The first one is based on semantic correspondence (Woodsend and Lapata, 2010) of document sentences and reference summary, used in (Cheng and Lapata, 2016). The second one is heuristic, which maximizes the ROUGE score with respect to gold summaries. This one is more broadly used in many recent extractive systems (Nallapati et al., 2017; Zhou et al., 2018; Zhang et al., 2019; Liu, 2019). We adopt the second method since it is more widely used and easy to implement. The extractive oracle is computed with the metric of ROUGE-2 F1 score, which is also the metric used in the final automatic evaluation in these systems.

Second, we check the output of a BERT-based sentence level extraction method and denote it as BERT-SENT. Following previous works (Devlin et al., 2018; Liu, 2019; Zhang et al., 2019), BERT-SENT treats extractive document summarization as a sentence classification task. The model is borrowed from Liu (2019), but we remove the interval segment embeddings in it since it does not have obvious benefits.

#### 3.2.1 Unnecessary Information

In this section, we examine whether unnecessary information is introduced unavoidably when extracting full sentences.

|                    | REF   | Ora-sent | Ora-ss |
|--------------------|-------|----------|--------|
| # (Sent)           | 3.88  | 2.61     | N/A    |
| # (Word)           | 58.31 | 70.46    | 52.77  |
| ROUGE-1 P          | N/A   | 52.59    | 61.84  |
| ROUGE-2 P          | N/A   | 33.97    | 43.45  |
| 1-gram Overlap (%) | 15.77 | 19.24    | 16.75  |
| 2-gram Overlap (%) | 1.40  | 2.22     | 1.90   |
| 3-gram Overlap (%) | 0.21  | 0.51     | 0.45   |

Table 1: Statistics of the reference (REF), and the extractive oracle of sentence level (Ora-sent) and sub-sentential level (Ora-ss) on the *CNN/Daily Mail* test set.

**Oracle:** Table 1 shows the information of the extractive oracle on the *CNN/Daily Mail* test set. The ROUGE-1 precision of the extractive oracle is 52.59, which means that there are 47.41 percent of the unigrams are not in the reference summary. As to the ROUGE-2 scores, the precision drastically drops to 33.97. These two metrics show that large amount of unwanted lexical units, i.e. unigram and bigram, are extracted along with the desired contents. This indicates that there exists unnecessary information on the lexical level.

The surface lexical matching (ROUGE scores) has its limits, that it cannot fully express the semantic level. We also conduct human analysis to check whether unnecessary information is extracted at the same time. The labeling criteria of unnecessary information is whether a 5-token span is not needed, lexically or semantically, comparing to ground-truth summary. We randomly sampled 50 documents from the *CNN/Daily Mail* test set. Evaluation results show that, 48% of the extractive oracles contain unnecessary information.

**BERT-SENT:** Similar experiments and analyses are also conducted on BERT-SENT. Table 6 shows the count-based statistics. Results show that 63.07% percent of the unigrams and 82.73% bigrams are not in the reference summary. These rates are much higher than the sentence-level extractive oracle, and show that the unnecessary issue is quite severe. Human evaluation shows that 54% the outputs have unnecessary information.

### 3.2.2 Redundancy

In this section, we check whether redundancy problem exists in extractive summarization. Similarly, we conduct the experiments and analysis on the extractive oracle and the BERT-SENT summarization system. We first define a metric for redundancy, i.e., the n-gram overlap rate. We calculate the n-gram overlap between each pair of sentences. This overlap is calculated as:

$$\text{n-gram overlap} = 1 - \frac{\#(\text{unique n-gram})}{\#(\text{total n-gram})} \quad (1)$$

**Oracle** Table 1 shows the information of the extractive oracle on the *CNN/Daily Mail* test set. It can be observed that there are 19.24% unigram and 2.22% bigram are duplicated in the extractive oracle, which is much higher than the reference summary.

Beyond this lexical level statistics, human evaluation is also conducted. Results show that 12% of the extractive oracle has the redundancy issue. This result matches the n-gram overlap rates and shows that the redundancy issue even exists in oracle.

**BERT-SENT** Results in Table 6 shows that the BERT-SENT has high n-gram overlap rates, i.e., 27.18% 1-gram and 7.68% 2-gram overlap. Thus, the redundancy issue is more severe in a real system than the extractive oracle, even for a state-of-the-art BERT-based system. Human evaluation also shows that 49% of the BERT-SENT output has the redundancy issue.

## 4 Q2: Efficacy of Extracting Sub-sentential Units

In this section, we propose an alternative to performing extraction on full sentence for extractive document summarization. Instead, we perform extraction on sub-sentential units. Specifically, the extraction units are based on the clause nodes in the constituency parsing tree of a sentence. Figure 2 shows two simplified examples of the constituency tree. The root node in a constituency tree represents the entire sentence, and the leaf node represents its corresponding lexical token. Extracting on the root node is essentially extracting the full sentence, while extracting on the leaf node is doing compressing by extracting words (Filippova et al., 2015). We perform extraction on the non-terminal nodes which can both express a relatively complete meaning and be human-readable. Therefore, the clause nodes, such as S and SBAR, become a good choice. In this section, we introduce how to perform extraction on the sub-sentential units, and present a BERT-based model for it.

### 4.1 The Sub-Sentential Units

In order to perform extraction on the sub-sentential units, we need to determine what units can be extracted. The proposed method is based on the constituency parsing tree. The basic idea is based on the sub-sentential clauses in the tree. In our experiments, we adopt the syntactic tagset used in the Penn Treebank (PTB) (Marcus et al., 1993). There are two main types in the PTB tagset, phrase and clause. We use the clause tag since the information in a clause is more complete than a phrase.

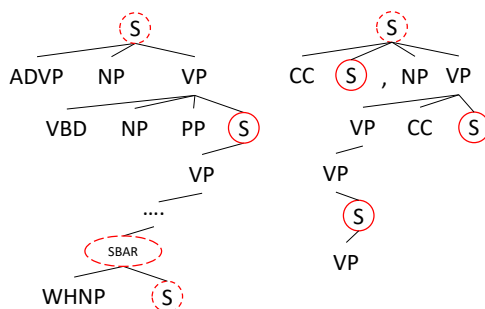


Figure 2: Two simplified constituency parsing trees. The nodes in circles are candidates. The final selected node is the one in red solid-lined circle.

Given the parsing tree  $t_i$  of sentence  $s_i$ , we traverse it to determine the boundary of extraction units. Specifically, every clause is treated as the extraction unit candidates. If one of its ancestors is a clause node, we choose the highest level ancestor clause node (except for the root node) as the extraction unit to include more complete information. This heuristic is visualized in Figure 2. If no sub-sentential clauses can be found in a sentence, we use the full sentence as the extraction unit. Finally, the input sentence is split into chunks using the selected clauses' boundaries.

### 4.2 Model

In this section, we present a BERT-based neural extractive summarization model for extracting sub-sentential units (SSE). We follow previous works (Cheng and Lapata, 2016; Nallapati et al., 2017; Xu and Durrett, 2019; Liu, 2019) to treat the document summarization as a sequence labeling task. Figure 3 shows the overview of the proposed model. It consists of two levels of encoders. The first level is the BERT-based document encoder, and the second level is the Transformer-based sub-sentential units encoder. The BERT-based document encoder reads the tokens in the document, and then the Transformer-based encoder constructs the final extraction unit representations.

#### 4.2.1 BERT-based Encoder

Following previous works (Liu, 2019; Zhang et al., 2019; Lebanoff et al., 2019) which use BERT and achieve state-of-the-art results, we use BERT as the first level encoder. The processed input document is denoted as  $\mathcal{D} = (S_1, S_2, \dots, S_n) = (w_1, w_2, \dots, w_m)$  with  $n$  sentences,  $m$  BPE tokens. The  $i$ -th

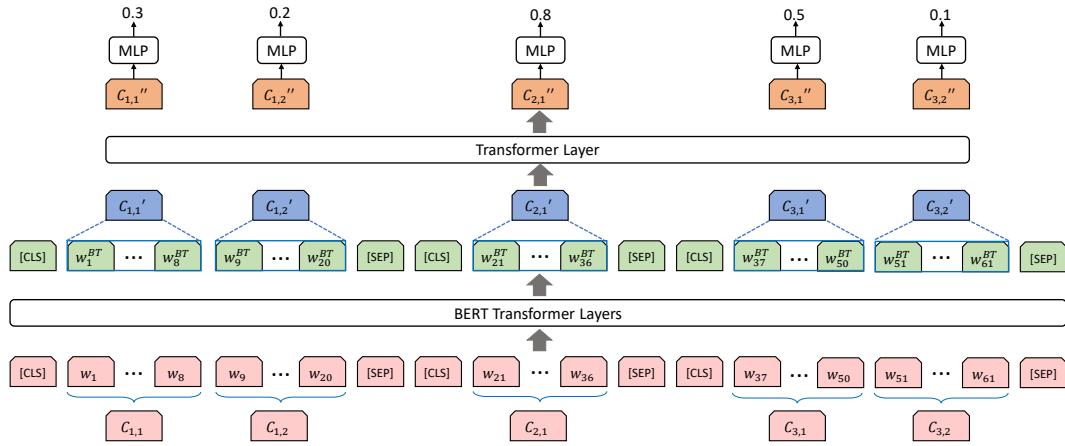


Figure 3: The overview of the BERT-based model for sub-sentential extraction (SSE). In this simplified example, the document has 3 sentences. The first and the third sentences have two extraction units and the second sentence has one. After encoding the document with pre-trained BERT encoder, an average pooling layer are used to aggregate information of each extraction unit. The final Transformer layer captures the document-level information and then the MLP predicts the extraction probability.

sentence contains  $l$  chunks  $S_i = (C_{i,1}, C_{i,2}, \dots, C_{i,l})$ . The  $j$ -th chunk with  $k$  words in  $S_i$  is denoted as  $C_{i,j} = (w_{i,j,1}, \dots, w_{i,j,k})$ . Following Liu (2019), we add additional [CLS] and [SEP] labels between sentences to separate them. However, since the extraction unit is not the full sentence, the vector of [CLS] is not used for classification in our model. After the BERT encoder, the vector of the  $m$  document tokens are represented as  $(w_1^{BT}, w_2^{BT}, \dots, w_m^{BT})$ .

#### 4.2.2 Transformer-based Encoder

The BERT-based encoder reads the entire document and builds the representation of each words. The Transformer-based encoder then constructs the final representation of each chunk. As shown in Figure 3, we first apply an average pooling on the chunk level. Specifically, given the BERT-based encoder output of chunk  $C_{i,j} = (w_{i,j,1}^{BT}, \dots, w_{i,j,k}^{BT})$ , the pooled representation  $C'_{i,j}$  is:

$$C'_{i,j} = \frac{1}{k} \sum_1^k w_{i,j}, \quad (2)$$

To note that, the [CLS] and [SEP] labels are not covered by the chunks, and thus not used in the average pooling.

After the average pooling, the document is represented as a sequence of chunk vectors:  $C' = (C'_{1,1}, \dots, C'_{1,l_1}, \dots, C'_{n,1}, \dots, C'_{n,l_n})$ . We then apply a chunk level Transformer to capture their relationship for extracting summaries:

$$\hat{C}' = \text{LN}(C' + \text{MultiHead}(C')) \quad (3)$$

$$C'' = \text{LN}(\hat{C}' + \text{FFN}(\hat{C}')) \quad (4)$$

where  $\text{MultiHead}(\cdot)$  is Multi Head Attention in Transformer (Vaswani et al., 2017),  $\text{LN}(\cdot)$  is Layer Normalization (Ba et al., 2016),  $\text{FFN}(\cdot)$  is a feed-forward network which consists of two linear transformations with a ReLU activation in between. In this paper, we simplify the  $\text{MultiHead}(Q, K, V)$  to  $\text{MultiHead}(\cdot)$  since we only use the self attention mechanism for encoding thus  $Q = K = V$ .

#### 4.2.3 Training Objective

With the chunk level representation vectors  $C''$ , the model predict the output probability of each chunk  $C''_{i,j}$ :

$$p(C''_{i,j}) = \sigma(\mathbf{W}_o C''_{i,j} + b_o) \quad (5)$$

where  $\sigma(\cdot)$  is the sigmoid function. The training objective of the model is the binary cross-entropy loss given the extractive oracle label  $y_{i,j}$  and the predicted probability  $p(C''_{i,j})$ .

## 5 Experiment

### 5.1 Dataset

Following previous extractive works (Zhou et al., 2018; Xu and Durrett, 2019; Lebanoff et al., 2019; Zhang et al., 2019; Dong et al., 2018), we conduct data preprocessing using the same method<sup>1</sup> in See et al. (2017), including sentence splitting and word tokenization. we preprocess the data as same as See et al. (2017). We then use a state-of-the-art BERT-based constituency parser (Kitaev and Klein, 2018) to process the input document whose performance is 95.17 F1 on WSJ test set. The statistic of the original *CNN/Daily Mail* dataset and the sub-sentential version are listed in Table 2.

| <i>CNN/Daily Mail</i>  | Training | Dev    | Test   |
|------------------------|----------|--------|--------|
| #(Document)            | 287,227  | 13,368 | 11,490 |
| #(Ref / Document)      | 1        | 1      | 1      |
| Doc Len (Sentence)     | 31.58    | 26.72  | 27.05  |
| Doc Len (Word)         | 791.36   | 769.26 | 778.24 |
| Ref Len (Sentence)     | 3.79     | 4.11   | 3.88   |
| Ref Len (Word)         | 55.17    | 61.43  | 58.31  |
| Doc Len (Sub-Sentence) | 52.84    | 51.37  | 52.02  |

Table 2: Data statistics of *CNN/Daily Mail* dataset.

### 5.2 Implementation Details

We found that the tokenizer used in the constituency parser is different from the one in BERT. Therefore, we apply some simple tokenization fix to process the text before feeding them into BERT. The input of BERT-based encoder is then processed with the BERT’s subword tokenizer. Since the maximum length in the BERT’s position embedding is 512, we truncated the document to 512 subwords. We use Adam (Kingma and Ba, 2015) as the optimizing algorithm. For the hyperparameters of Adam optimizer, we set the learning rate  $\alpha = 2e - 5$ , two momentum parameters  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$  respectively, and  $\epsilon = 10^{-8}$ . The model is implemented with PyTorch (Paszke et al., 2017) and PyTorch Transformer (Wolf et al., 2019). We use the `bert-base-uncased` version of BERT. We train the model using 4 NVIDIA P100 GPUs with a batch size of 40. The dropout (Srivastava et al., 2014) rates in all the Transformer layers are set to 0.1. We train the model for 4 epochs which takes about 6 hours. The final model is picked according to the performance on the development set among the 4 model checkpoints.

During inference, we rank the extraction units according to  $p(C''_{i,j})$  and select the top ones. Since the extraction unit in this paper is shorter than full sentence, we repeatedly select next sub-sentential unit until the summary length reaches the limit. The length limit is set to 60 words according to the statistics on the development set in Table 2.

### 5.3 Evaluation Metric

We employ ROUGE (Lin, 2004) as our evaluation metric. Following previous work, we use ROUGE-1 (unigram), ROUGE-2 (bigram) and ROUGE-L (LCS) as the evaluation metrics in the experimental results.

Additionally, we also conduct human evaluation on the output summaries. Following previous works (Cheng and Lapata, 2016; Nallapati et al., 2017; Liu, 2019; Zhang et al., 2019), we randomly sampled 50 documents from the *CNN/Daily Mail* test set, which is the same as in §3.

<sup>1</sup><https://github.com/abisee/cnn-dailymail>

## 6 Results

### 6.1 Automatic Evaluation

Table 3 shows the ROUGE evaluation results. We compare the SSE with the following systems:

**Abstractive Systems** Pointer-Generator Network (PGN) (See et al., 2017) is a sequence-to-sequence model with copy and coverage mechanisms. FastRewrite (Chen and Bansal, 2018) conducts extraction first then generation. JECS (Xu and Durrett, 2019) first extracts sentences then compresses them to reduce redundancy. Bottom-Up (Gehrmann et al., 2018) applies constraints on the copying probability.

**Extractive Systems** LEAD3 is a commonly used baseline which simply extracts the first three sentences. TEXTRANK (Mihalcea and Tarau, 2004) is a popular graph-based unsupervised system. SUMMARUNNER and NN-SE (Nallapati et al., 2017; Cheng and Lapata, 2016) use hierarchical structure for document encoding and predict sentence extraction probabilities. NEUSUM (Zhou et al., 2018) jointly model the sentence scoring and selection steps. BERTSUMEXT, BERTSUMEXT + TRIBLK (with trigram blocking) (Liu, 2019), SELF-SUPERVISED (Wang et al., 2019) and HIBERT (Zhang et al., 2019) use pre-training techniques in extractive document summarization. BERT-SENT is the sentence-level extractive baseline described in section 3.2 .

| Model             | ROUGE-1      | ROUGE-2      | ROUGE-L      |
|-------------------|--------------|--------------|--------------|
| PGN               | 39.53        | 17.28        | 36.38        |
| FastRewrite       | 40.88        | 17.80        | 38.54        |
| JECS              | 41.70        | 18.50        | 37.90        |
| Bottom-Up         | 41.22        | 18.68        | 38.34        |
| LEAD3             | 40.24        | 17.70        | 36.45        |
| TEXTRANK          | 40.20        | 17.56        | 36.44        |
| SUMMARUNNER       | 39.60        | 16.20        | 35.30        |
| NN-SE             | 41.13        | 18.59        | 37.40        |
| NEUSUM            | 41.59        | 19.01        | 37.98        |
| BERTSUMEXT        | 42.61        | 19.99        | 39.09        |
| BERTSUMEXT+TRIBLK | <b>43.25</b> | 20.24        | 39.63        |
| SELF-SUPERVISED   | 41.36        | 19.20        | 37.86        |
| HIBERT            | 42.10        | 19.70        | 38.53        |
| BERT-SENT         | 42.13        | 19.73        | 38.59        |
| SSE               | 42.72        | <b>20.29</b> | <b>39.98</b> |

Table 3: Full length ROUGE F1 evaluation (%) on CNN / Daily Mail test set. Results for comparison systems are taken from the authors’ respective papers or obtained by running publicly released software.

The proposed method SSE achieves the state-of-the-art results on the *CNN/Daily Mail* dataset. According to the output of the official ROUGE script, the difference between SSE and baselines are all statistically significant with a 0.95 confidence interval. Compared to our sentence-level extraction baseline system BERT-SENT, using sub-sentential unit extraction leads to a +0.56 ROUGE improvement. As for the other existing systems which leverage BERT or other pre-training techniques and perform extraction on sentence level, SSE still outperforms them statistically significantly in terms of ROUGE.

### 6.2 Human Evaluation

Human evaluations are also conducted on the same 50 randomly sampled documents as in Section 3. The BERT-SENT and SSE models are evaluated. The workers are asked to rank the outputs of these systems from best to worst by the overall quality (with ties allowed). In addition, we are also curious about how sub-sentential extraction solves the problems of full sentence extraction. Specifically, the workers are asked to identify whether redundant or unnecessary information exists.



|             | BERT-SENT | SSE  |
|-------------|-----------|------|
| Unnecessity | 54%       | 37%  |
| Redundancy  | 49%       | 29%  |
| Readability | 1.00      | 1.24 |
| Overall     | 1.50      | 1.35 |

Table 4: Human evaluation results. Unnecessity and Redundancy are reported as occurrence frequency, and lower is better. Readability and Overall are reported as ranking, and lower is better.

---

**Document:** (CNN) When *Etan Patz went missing in New York City at age 6*, hardly anyone in America could help but see his face at their breakfast table. His photo’s appearance on milk cartons after his May 1979 disappearance marked an era of heightened awareness of crimes against children. On Friday, more than 35 years after frenzied media coverage of his case horrified parents everywhere, *a New York jury will again deliberate over a possible verdict against the man charged in his killing, Pedro Hernandez*. He confessed to police three years ago. *Etan Patz’s parents have waited that long for justice*, but some have questioned whether that is at all possible in Hernandez’s case. .... He said he killed the boy and threw his body away in a plastic bag. Neither the child nor his remains have ever been recovered. *But Hernandez has been repeatedly diagnosed with schizophrenia and has an “IQ in the borderline-to-mild mental retardation range,”* his attorney Harvey Fishbein has said. Police interrogated Hernandez for 7 1/2 hours before he confessed. .... A judge found *Ramos responsible for the boy’s death* and ordered him to pay the family \$ 2 million – money the Patz family has never received. ....

---

**Reference:** The young boy ’s face appeared on milk cartons all across the United States . Patz ’s case marked a time of heightened awareness of crimes against children . Pedro Hernandez confessed three years ago to the 1979 killing in .

---

Table 5: An example document and gold summary in the *CNN/Daily Mail* test set. The words highlighted with red are extracted as a full sentence. The *italic words highlighted with cyan* are extracted as sub-sentential units.

Table 4 presents the human evaluation results. We compare the SSE with BERT-SENT. As shown in the results, SSE performs better than BERT-SENT for both redundancy and unnecessary. The frequency of having these issues drops 20% and 17% respectively. Thus the overall quality of SSE is also better than BERT-SENT.

### 6.3 Analysis

Table 5 shows an example of a document, gold summary and the output of SSE. It can be observed that by performing sub-sentential extraction, the full sentence is broken into more fine-grained semantic units. Therefore, the model can extract important parts without introducing unimportant contents.

Table 6 shows the statistics of the outputs of BERT-SENT and SSE. Similarly, we conduct experiments and analyses with both statistics and human judgments, on both the unnecessary information and redundancy issues as in Section 3. Compared to BERT-SENT, SSE performs significantly better in terms of ROUGE precision by a large margin. This shows that extracting sub-sentential units can bring less unimportant information. We also found that the n-gram overlap rate of SSE is also much lower than BERT-SENT, which shows that the output contains less redundant contents.

### 6.4 Q4: Readability of Sub-Sentential Units

Performing sub-sentential unit extraction improves the ROUGE scores and alleviates the problem of extracting full sentences. However, it is not clear whether this method introduces new issues. One possible issue is that the sub-sentential units are fragmented so the readability is poor. To investigate this problem, we also add an item about the readability of the produced summary in the human evaluation questionnaire. In detail, the workers are asked to rank the system outputs by the readability (with ties allowed).

Table 4 shows the results of readability. As shown in results, the BERT-SENT is always ranked as the best since the sentences are fully extracted. The readability of extracted sub-sentential units is slightly worse than the full sentences. We also manually checked the output of SSE whose readability is worse. We found that there are two reasons: 1) the sub-sentence is fragmented which affects the readability; 2)

|                    | BERT-SENT | SSE   |
|--------------------|-----------|-------|
| # (Sent)           | 3         | N/A   |
| # (Word)           | 82.83     | 68.74 |
| ROUGE-1 P          | 36.93     | 39.22 |
| ROUGE-2 P          | 17.27     | 18.66 |
| 1-gram Overlap (%) | 27.18     | 24.58 |
| 2-gram Overlap (%) | 7.68      | 6.00  |
| 3-gram Overlap (%) | 4.13      | 2.85  |

Table 6: Statistics of the BERT-SENT and SSE methods on the *CNN/Daily Mail* test set.

the sub-sentence is wrongly extracted due to the error of the constituency parser. Therefore, we hope that the readability of SSE could be improved if we can: 1) design better sub-sentential unit extraction algorithm; 2) have an even better syntactic parser.

## 7 Conclusion

In this paper, we investigate the problem of the extraction granularity for extractive document summarization. We observe that performing extraction at sentence level has the redundancy and unnecessary issues. We found that these problems can be alleviated by doing sub-sentential unit extraction. Both automatic and human evaluations show that sub-sentential extraction performs competitively compared to the full-sentence-extraction systems. Therefore, sub-sentential unit extraction could be a promising alternative to full-sentence extraction. Our experiments and analyses on revisiting the basic extraction unit could provide some hints for future research on this direction.

## References

- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Regina Barzilay and Kathleen R McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.
- Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3101–3111, Hong Kong, China, November. Association for Computational Linguistics.
- Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *AAAI*, pages 2153–2159.
- Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM.
- Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia, July. Association for Computational Linguistics.
- Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany, August. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

- Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium, October-November. Association for Computational Linguistics.
- Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.
- Katja Filippova, Enrique Alfonseca, Carlos A Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368.
- Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium, October-November. Association for Computational Linguistics.
- Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719.
- Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of 3rd International Conference for Learning Representations*, San Diego.
- Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July. Association for Computational Linguistics.
- Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy, July. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.
- Yang Liu. 2019. Fine-tune BERT for extractive summarization. *arXiv preprint arXiv:1903.10318*.
- Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.
- Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June.
- André Martins and Noah A. Smith. 2009. Summarization with a joint model for sentence extraction and compression. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.
- Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *European Conference on Information Retrieval*, pages 557–564. Springer.
- Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.
- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*, pages 3075–3081.
- Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.
- Pengjie Ren, Zhumin Chen, Zhaochun Ren, Furu Wei, Jun Ma, and Maarten de Rijke. 2017. Leveraging contextual sentence relations for extractive summarization using a neural attention model. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 95–104, New York, NY, USA. ACM.

- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July. Association for Computational Linguistics.
- Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.
- Xiaojun Wan and Jianwu Yang. 2006. Improved affinity graph based multi-document summarization. In *Proceedings of the human language technology conference of the NAACL, Companion volume: Short papers*, pages 181–184. Association for Computational Linguistics.
- Hong Wang, Xin Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Self-supervised learning for contextualized extractive summarization.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 565–574, Uppsala, Sweden, July. Association for Computational Linguistics.
- Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3290–3301, Hong Kong, China, November. Association for Computational Linguistics.
- Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy, July. Association for Computational Linguistics.
- Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia, July. Association for Computational Linguistics.