applied sciences

Article

Diverse Decoding for Abstractive Document Summarization

Xu-Wang Han [1], Hai-Tao Zheng [1,*], Jin-Yuan Chen [1] and Cong-Zhi Zhao [2]

[1] Tsinghua-Southampton Web Science Laboratory, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China; hxw16@mails.tsinghua.edu.cn (X.-W.H.); jy-chen13@mails.tsinghua.edu.cn (J.-Y.C.)

[2] Giiso Information Technology Co., Ltd., Shenzhen 518055, China; zhaocz@giiso.com

[*] Correspondence: zheng.haitao@sz.tsinghua.edu.cn; Tel.: +86-180-3815-3089

Received: 13 December 2018; Accepted: 18 January 2019; Published: date

Featured Application: Automatic summarization is widely used in the news area. Due to the overload of news information, people are eager to have a tool to help them learn about the most useful information in the shortest time. In addition, search engines also comprise one of the applications. Automatic summarization based on query will help users find their content of interest as soon as possible. There are also many other research fields like data mining, chat bot, narrative generation, etc., having a strong connection to the summarization task.

Abstract: Recently, neural sequence-to-sequence models have made impressive progress in abstractive document summarization. Unfortunately, as neural abstractive summarization research is in a primitive stage, the performance of these models is still far from ideal. In this paper, we propose a novel method called Neural Abstractive Summarization with Diverse Decoding (NASDD). This method augments the standard attentional sequence-to-sequence model in two aspects. First, we introduce a diversity-promoting beam search approach in the decoding process, which alleviates the serious diversity issue caused by standard beam search and hence increases the possibility of generating summary sequences that are more informative. Second, we creatively utilize the attention mechanism combined with the key information of the input document as an estimation of the salient information coverage, which aids in finding the optimal summary sequence. We carry out the experimental evaluation with state-of-the-art methods on the CNN/Daily Mail summarization dataset, and the results demonstrate the superiority of our proposed method.

Keywords: neural abstractive summarization; sequence-to-sequence neural model; beam search; diverse decoding; optimal sequence selection

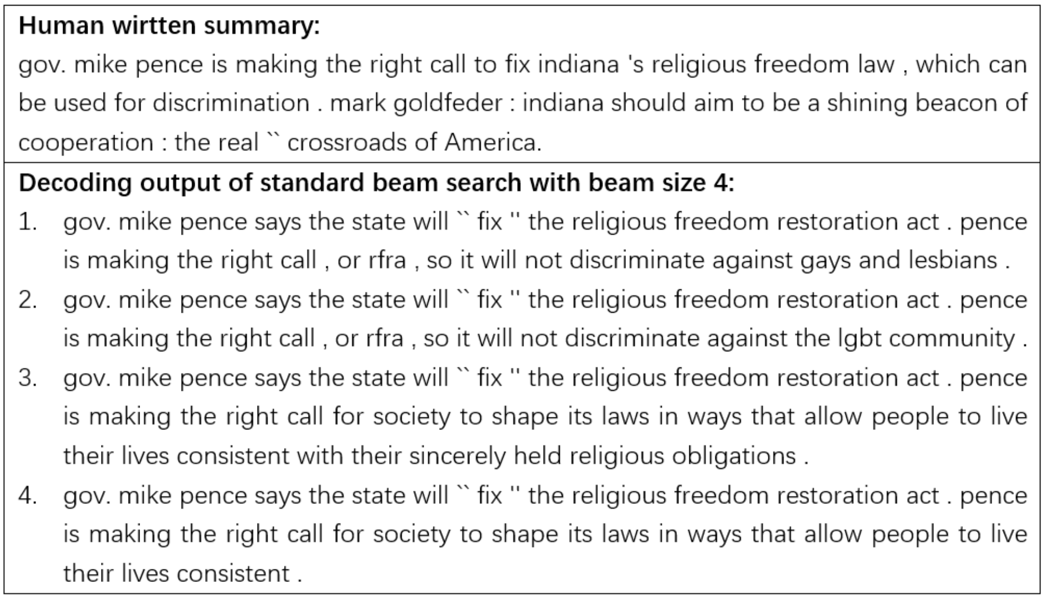
1. Introduction

Document summarization is the task of generating a condensed and coherent summary of a document while retaining the salient information. There are two broad types of summarization: extractive and abstractive. Extractive summarization systems generate summaries by extracting sentences from the original documents, while abstractive summarization systems use words or phrases that may not appear in the original document like the human-written summaries. Abstractive summarization is considered to be much more difficult, for it involves sophisticated techniques like natural language understanding, knowledge representation, natural language generation, etc.

Over the past few years, deep neural network approaches have shown encouraging results in various natural language generation (NLG) tasks like machine translation [1–3], image captioning [4], and abstractive summarization [5]. In particular, the attention-based sequence-to-sequence (seq2seq) framework with recurrent neural networks (RNNs) [5–7] prevails at the task of abstractive summarization. While promising, these methods still fail to provide satisfactory performance when faced with a long input sequence of multiple sentences, with a considerable gap with respect to extractive approaches.

In this work, we study the key factors that keep abstractive models from achieving better performance. The sequence-to-sequence framework consists of two parts: an encoder that processes the input and a decoder that generates the output. There have been extensive works on the design and alignment of the encoder, like the choice of the neural network structure [5–7], the specific design of the attention mechanism [8–14], etc. In contrast, the decoding process, which directly affects the results of the summarization task, has received relatively much less attention.

In the decoding stage, nearly all these seq2seq models use a heuristic algorithm called beam search, which generates sequences word-by-word in a greedy left-to-right fashion while keeping a fixed number (top N) of candidates at each time step. At the end of the beam search, the top sequence with the highest conditional probability would be selected as the summary of the input document. However, in practice, with the increasing length of generated sequences, candidate sequences expanded from a single beam would gradually take up the positions of top N. As a result, the beam search decoder always produces nearly identical sequences that only differ by several words in the endings. Figure 1 shows an example of the outputs of the neural decoder applying the standard beam search method. As the decoding process is in fact a process of iterative word prediction, the lack of diversity of beam search means that this widely-applied method only covers a small portion of the search space. Therefore, the standard beam search is not only a waste of computational resources, but more importantly, eliminates the chance of generating more distinct and maybe better sequences for the summarization systems.



Human wirtten summary:

gov. mike pence is making the right call to fix indiana 's religious freedom law , which can be used for discrimination . mark goldfeder : indiana should aim to be a shining beacon of cooperation : the real `` crossroads of America.

Decoding output of standard beam search with beam size 4:

1. gov. mike pence says the state will `` fix '' the religious freedom restoration act . pence is making the right call , or rfra , so it will not discriminate against gays and lesbians .
2. gov. mike pence says the state will `` fix '' the religious freedom restoration act . pence is making the right call , or rfra , so it will not discriminate against the lgbt community .
3. gov. mike pence says the state will `` fix '' the religious freedom restoration act . pence is making the right call for society to shape its laws in ways that allow people to live their lives consistent with their sincerely held religious obligations .
4. gov. mike pence says the state will `` fix '' the religious freedom restoration act . pence is making the right call for society to shape its laws in ways that allow people to live their lives consistent .

Figure 1. Example from the CNN/Daily Mail test dataset showing the outputs of the standard beam search decoder.

Another common weakness of previous works is the lack of an explicit method for estimating the salient information preservation of the input document, as saliency is one of the core requests for the summarization system [15,16]. During the whole process of generating the sequences, the only feature that neural decoders rely on is the conditional probability of the trained language model, which may result in the output summary being fluent in language, yet missing most of the salient points of the original document.

In this paper, we study how to improve abstractive document summarization by focusing on addressing the issues of decoding and propose a method called Neural Abstractive Summarization with Diverse Decoding (NASDD). In the process of decoding, the NASDD model introduces a new objective and the middle-term grouping approach to discourage sequences from sharing common roots, thus improving diversity. With the improvement of diversity, the selection of the optimal sequence as the final summary becomes a problem. The standard beam search process selects the one with the highest conditional probability, which does not make much sense because of the high similarity of candidate sequences. The probability based on the trained language model is more of an estimation of language fluency, and the participation of salient information of the input document is relatively limited. To address this problem and inspired by the unsupervised keyword extraction method, we locate salient points of the original document by such an approach. At each time step of the decoding process, the neural decoder produces an attention vector. This vector indicates the participation of every word in the document in this time step's word prediction process. Using the attention vector as a bridge to the input document and combining it with the scored words allow us to estimate the salient information coverage. In sum, the probability values of the language model represent the fluency, the connection between the generated words, and the original document reflects the saliency. Our proposed selection method considers both factors to find the optimal summary sequence.

We conduct extensive experiments on the popular CNN/Daily Mail dataset. Experiments show that our method outperforms state-of-the-art neural models in terms of both ROUGEscore and diversity statistics.

The main contributions of this paper are as follows:

1. We introduce a Diversity-Promoting Beam Search approach (DPBS) in the sequence-to-sequence neural model to generate multiple diversified candidate sequences for abstractive document summarization. This search method covers a much larger search space than the standard beam search, thus increasing the probability of generating better summary sequences.
2. We design a selection algorithm that considers extensive factors for the outputs of diverse beam search to locate the optimal sequence. The essential part of this method is using the attention mechanism as a bridge between the input document and the generated summary to estimate the salient information preservation, providing a novel and viable approach for saliency estimation.
3. We combine these methods in a unified neural summarization system and achieve state-of-the-art performance.

The paper is organized as follows. Section 2 introduces related works in abstractive document summarization. Section 3 describes our method. Section 4 presents the experiments and discussion. In Section 5, we conclude this paper.

2. Related Work

Early summarization works mostly focused on extractive methods with human-engineered features, for example parts of speech [17] and term frequency [18]. Classical approaches include graph-based methods [19], integer linear programming [20] and classifier-based methods [21,22]. On the other hand, there has been much less research on abstractive summarization. Early works on abstractive summarization were mostly restricted to the domain of sentence compression, using methods like syntactic tree pruning [23,24] and machine translation [25]. A more systematic review of these classical approaches can be found in [15].

In recent years, neural networks have been widely investigated both on extractive and abstractive summarization tasks. In terms of neural extractive models, impressive performance gains have been made by applying deep learning techniques in the traditional framework [26–29] or using a more data-driven way, i.e., the encoder-decoder approach [30,31]. In the meantime, neural sequence-to-sequence models have provided a viable new approach for abstractive summarization. As this paper is about abstractive summarization at the document level, the following sections will mainly focus on the related works of neural abstractive summarization.

Rush et al. [5] were the first to apply modern neural networks in the task of abstractive text summarization and achieved state-of-the-art performance on the sentence-level summarization datasets DUC-2004 and Gigaword. Based on the neural encoder-decoder architecture with the attention mechanism, their model used a Convolutional Neural Network (CNN) encoder and a neural language model decoder. Chopra et al. [6] extended this work by replacing the decoder with a Recurrent Neural Network (RNN). Nallapati et al. [7] further improved the performance by using a full RNN encoder-decoder framework. These neural models often use a fixed vocabulary, which leads to the Out-Of-Vocabulary (OOV) problem when there are new words in the input. One way to fix this is to enable the decoder network to copy these OOV words into the output sequence. Based on this copying idea, Gu et al. [32] proposed CopyNet and Vinyals et al. [33] proposed the pointer network.

Another challenge for abstractive summarization is the lack of large-scale datasets for longer document-level summarization. Nallapati et al. solved this problem by introducing the CNN/Daily Mail dataset, which was originally the DeepMind question-answering dataset [34]. In their work, they provided the first abstractive baseline for this dataset and illustrated a key problem in this document-level summarization: these attentional neural models often generate unnatural summaries with repeated phrases. Tan et al. [16] proposed a graph-based attention mechanism to address saliency and a hierarchical beam search algorithm to alleviate the repeated phrases problem. Paulus et al. [35] proposed the intra-attention mechanism and a new training method with reinforcement learning to generate more readable long sequences. See et al. [36] augmented the standard seq2seq model with a pointer network [37] for the OOV problem and the coverage mechanism [38,39] to discourage repetition. These models were able to achieve state-of-the-art performance on the CNN/Daily Mail dataset, generating summaries with basic readability. Unfortunately, these state-of-the-art models have overlooked the potential value of improving beam search, which keeps these models from getting better performance.

There have been some works on producing diverse decoding sequences from recurrent models in other Natural Language Processing (NLP) tasks like machine translation and conversation modeling. Li et al. [40] were the first to address the issue of output diversity in the neural sequence generation framework; they proposed using maximum mutual information as the objective function in neural models to produce more diverse and interesting responses in conversation modeling. Li and Jurafsky [41] proposed a beam search diversification method, which discourages sequences from sharing common roots to generate diverse lists for neural machine translation. Vijayakuma et al. [42] proposed a universal diverse beam search method, which can be applied to any model where the original beam search is applicable and outperformed the previously-proposed diverse decoding techniques. However, these methods only focus on generating different and fluent sequences, and they lack considerations of the content of the sequences. In abstractive summarization, diversity is meaningless if most of the generated lists have no connection to the input document. In addition, all of the above methods are for single-sentence-level generation tasks and are not applicable to abstractive summarization tasks.

Compared to previous works, our method focuses on the task of document-level abstractive summarization using the neural attention model. The proposed model combines a collection of previously-proven effective methods on the standard neural attention framework, including the pointer network for the OOV problem [37] and the converge mechanism [36] for the repetition problem. Moreover, instead of using the traditional decoder, we modify the standard beam search with our new objective and introduce new selection.

3. Neural Abstractive Summarization with Diverse Decoding

In this section, we introduce the framework and implementation details of the proposed NASDD approach. We chose a neural framework of summarization called pointer generator [36] as the base model to apply the diverse decoding method and sequence selection method. This framework itself is the standard sequence-to-sequence attentional model combined with a collection of proven effective measures and achieves the state-of-the-art performance on the document-level abstractive summarization. The original model then naturally becomes the baseline of our experiments. For the sequence-to-sequence attentional model, we used a bi-directional LSTM encoder and a single-layer LSTM decoder, and we used the pointer mechanism to handle OOV problems and the coverage mechanism to handle repetition problems. Based on the neural model, we applied the Diversity-Promoting Beam Search method (DPBS) and the optimal sequence selection method.

3.1. Structure of the Neural Network Model

The base neural model uses a bi-directional LSTM encoder and a single-layer LSTM decoder, as shown in Figure 2. The encoder takes the source text $x = \{w_1, w_2, \ldots, w_n\}$ as input and produces a sequence of encoder hidden states $\{h_1,\ h_2, \ldots, h_n\}$.

$$e_i^t = v^T \tanh\left(w_h h_i + w_s s_t + b_{attn}\right) \tag{1}$$

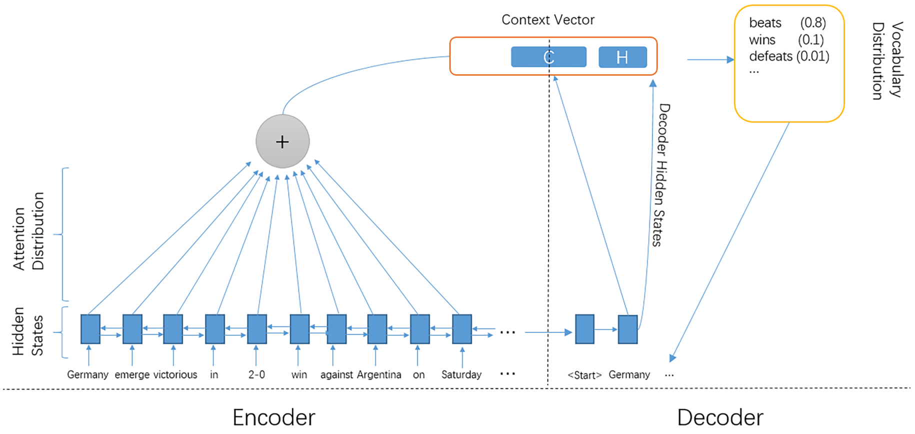


Figure 2. Illustration of the attentional sequence-to-sequence model. The model attends to relevant words in the source text to generate words of the output sequence.

The task is to predict the summary sequence $y = \{y_1,\ y_2, \ldots, y_n\}$. At each step t, the decoder receives the word embedding of the previous word and produces the decoder state s_t. Notice that all the following probability expressions are conditional probabilities. The attention distribution at each decoding time step is:

$$\mathrm{a}^t = \operatorname{softmax}\left(e^t\right) \tag{2}$$

where v, w_h, w_s, and b_{attn} are learnable parameters. The attention distribution can help the decoder to focus on a specific part of the input when predicting the next word. Next, the attention distribution is used to calculate the context vector, which is the weighted sum of encoder hidden states and can be used to produce the vocabulary distribution:

$$h_t^* = \sum_i a_i^t h_i \tag{3}$$

$$P_{vocab} = \operatorname{softmax}\left(V'\left(V\left[s_t,\ h_t^*\right] + b\right) + b'\right) \tag{4}$$

where V', V, b, b' are learnable parameters. P_{vocab} is the probability over all words in the vocabulary. Then, we add the pointer mechanism, which works as a switch, to choose whether to generate a word from the fixed vocabulary or to copy OOV words from the input. The generation probability P_{gen} for time step t is calculated as:

$$P_{gen} = \text{œ}\left(w_{h^*}^T + w_s^T s_t + w_x^T x_t + b_{ptr}\right) \tag{5}$$

where vectors w_{h*}, w_s, w_x, and scalar b_{ptr} are learnable parameters and œ is the sigmoid function. We finally obtain the following probability distribution:

$$P(y_t) = P_{gen} P_{vocab}(y_t) + (1 - P_{gen}) \sum_{i:y_i = y_t} a_i^t \tag{6}$$

During training, we use the human-written summary as the target sequence $y^* = \{y_1^*, y_2^*, \ldots, y_n^*\}$; the loss for time step t is the negative log likelihood of the target word y_t^*; and the overall loss for the whole sequence is:

$$\text{loss} = \frac{1}{T} \sum_{t=0}^{T} -\log P(y_t^*) \tag{7}$$

3.2. Diversity-Promoting Beam Search Method

In this section, we introduce our DPBS method, which is an improved decoding procedure for the standard decoder. We talk about the task of decoding in Section 3.2.1 and reveal the problem caused by the huge search space. In Section 3.2.1 we introduce the beam search method, which alleviates the search problem in a greedy way, and discuss its shortcomings when facing long sequence-generation tasks like summarization. Then, in Section 3.2.3, we introduce our DPBS method, which is composed of a novel objective and a middle-term grouping approach.

3.2.1. Decoding Problem

The neural model is trained to estimate the likelihood of sequences of generated words given the input x. When generating summaries, as Equation (6) expresses, the decoder produces the conditional probability distribution for the next output considering the input and all previous generated words. Let $\theta(y_t) = \log P(y_t | y_{t-1}, \ldots, y_1, x)$ be the conditional probability distribution over the vocabulary V at time step t. The log probability of a sequence of generated words can now be written as $\Theta\left(Y_{[t]}\right) = \sum_{i \in [t]} \theta(y_i)$. Then, the decoding task is to find a sequence y that maximizes $\Theta(Y)$.

The decoder produces the probability distribution over the vocabulary V, which means the decoder has as many choices as the vocabulary size $|V|$. Every one time step forwards means the whole search space is expanded $|V|$ times. At time step t, the search space for finding the global optimum solution is $|V|^t$, which is computationally intractable. An alternative approach is to choose the word with the highest probability each time, which is a strictly greedy approach and is not necessarily going to provide the sentence with the highest probability.

3.2.2. Standard Beam Search and Its Problem

A compromise between exact and greedy decoding is to use the beam search approach, which keeps the top K high-scoring candidates at each time; where K is known as the beam size. The beam search is the standard approach for neural decoding. At time step $t-1$ in the decoding process, we keep record of K hypotheses based on score $S(Y_{t-1} \mid x) = logP(y_{t-1}, \ldots, y_1 | x) = \Theta\left(Y_{[t-1]}\right)$. As the decoding process moves on to time step t, each of the K hypotheses (denoted as $Y_{t-1}^k = \{y_1^k, y_2^k, ..., y_{t-1}^k\}, k \in [1, K]$) is expanded individually by selecting the top K candidates, denoted as $Y_{t-1}^{k,k'}$, leading to the construction of $K \times K$ new hypotheses:

$$S\left(Y_{t-1}^k, y_t^{k,k'} \mid x\right), k' \in [1, K], k \in [1, K] \tag{8}$$

The goal is to find the top K hypotheses from the newly-generated $K \times K$ hypotheses; we rank theme with scores computed as as follows:

$$S\left(Y_{t-1}^k, y_t^{k,k'} \mid x\right) = S\left(Y_{t-1}^k \mid x\right) + logP(y_t^{k,k'} | x, Y_{t-1}^k) \tag{9}$$

Based on the score $S\left(Y_{t-1}^{k}, y_t^{k,k'} \mid x\right)$, the top K hypotheses are selected, and the remaining hypotheses are ignored as the decoder proceeds to next time step.

The beam search process can be viewed as a search algorithm that explores a graph by expanding the most promising node in a limited set. However, each time step when we are expanding the top K nodes, the first node's successors can easily outcompete other nodes' successors. This is because their parental node is already the one with the highest score, and the beam search selects candidates by the sum of the words' score. This parental advantage accumulates with the increasing of decoding length, resulting in the final candidates generate mostly coming from a single beam, with minor variations in the tail.

3.2.3. Diversity-Promoting Beam Search

To overcome this shortcoming and inspired by the work of [41] in machine translation, we propose to introduce the ranking information in the calculation of candidates' score $S\left(Y_{t-1}^{k}, y_t^{k,k'} \mid x\right)$ to discourage sequences from sharing common roots and thus promoting diversity. For each of the K candidates Y_{t-1}^{k}, we rank the newly-generated top K nodes $Y_{t-1}^{k,k'}$, based on their conditional probability $P(y_t^{k,k'} | x, Y_{t-1}^{k})$ in descending order. Therefore, now, k' denotes the ranking of the current node among its siblings. Then, we add an additional part $\gamma k'$ to rewrite the score function for $\left[Y_{t-1}^{k}, y_t^{k,k'}\right]$:

$$\hat{S}\left(Y_{t-1}^{k}, y_t^{k,k'} \mid x\right) = S\left(Y_{t-1}^{k}, y_t^{k,k'} \mid x\right) - \gamma k' \tag{10}$$

where scalar γ can be tuned based on the validation set. The top K candidates are then selected based on $\hat{S}\left(Y_{t-1}^{k}, y_t^{k,k'} \mid x\right)$ as the time step moves on. This additional term $\gamma k'$ works by punishing bottom ranked nodes among siblings (directly descended from the same parental node), thus making the whole decoding process tend to select candidates descended from different root nodes. The diversity-promoting process is shown in Figure 3.

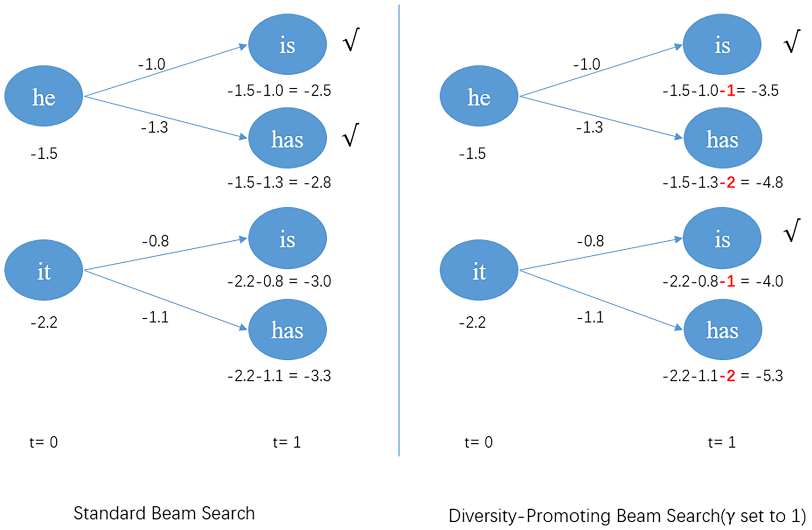


Figure 3. Illustration of standard beam search and the proposed diversity-promoting beam search. Note that this is only the initial part of the diversity-promoting process. After the candidates have diverged sufficiently, we turn to using the grouping strategy.

With the length of decoding sequences growing, the "accumulated parental advantage" of the first beam would gradually undermine the effect of $\gamma k'$ and make the selection still concentrate on the first beam in the last several time steps of decoding. To avoid this, we use the new score function for beam search at the beginning of decoding. Next, we divide the K unfinished candidates into groups G at time step t' when the candidates have diverged sufficiently. Then, we limit the selection of candidates into each of the groups, i.e., we use the standard beam search with beam size K/G inside each of these groups until the end of decoding. As the sequences have already diverged, the subsequent predicted nodes from different groups are unlikely to overlap, and thus, the grouping strategy can help to preserve the diverged beam.

3.3. Optimal Sequence Selection Method

At the end of the Diverse-Promoting Beam Search method (DPBS), there are K final candidate sequences to be the summary of the input document. As the diversity improves, the original selection method becomes inappropriate. We propose a novel selection method that considers more about the candidates' connection to the salient information of input documents, instead of only choosing the one with the highest probability as the standard beam search does.

Recall that the attention distribution a^t at time step t is a vector with a size equal to the number of words in the input document, and the distribution demonstrates the degree of participation of each input word in decoding at this time step. Along with each of the K candidate sequences, there is a list of attention vectors; we propose to use these vectors as a measurement for salient information preservation. We are inspired by the unsupervised keyword extraction method, which could be used for importance evaluation of input words without harming the full data-driven pattern of neural abstractive summarization. We use the strong baseline tf-idfweighting to score every word in the input document ($x = \{w_1, w_2, \ldots, w_n\}$) and combine the list of attention vectors, and we now have a term for saliency estimation: $\sum\limits_t \sum\limits_i a_i^t \omega_i$, where ω_i is the Tf-idf score of the i^{th} word of the input document. We also introduce an additional term that considers the extractive words (words appear both in generated sequences and the input document; once again, we use the Tf-idf score value) $\sum\limits_{\omega_j \in x} \omega_j$. The score for a given candidate sequence y is then calculated as follows:

$$Score\,(y) = \frac{1}{L_y}\left(\log\,(y \mid x) + \lambda \sum_t \sum_i a_i^t \omega_i + \eta \sum_{\omega_j \in x} \omega_j\right) \tag{11}$$

where L_y is the length of sequence y; we optimize λ and η on the validation set.

4. Experiment

In this section, we introduce our experimental setup and result analysis. We conduct experiments to validate the effectiveness of the NASDD approach at improving abstractive summarization.

4.1. Experimental Setup

In this section, we will describe our experimental setup in detail. The following subsections will present our dataset, implementation detail, evaluation metrics, and baseline methods, respectively.

4.1.1. Dataset

We used the CNN/Daily Mail dataset [7,34], which has been widely used in neural document summarization. The corpora was constructed by collecting human-written highlights from news articles on the website of CNN and Daily Mail, and it contains 287,226 train pairs, 12,268 validation pairs, and 11,490 test pairs.

4.1.2. Implementation

This model was implemented using Tensorflow. We used a one-layer biLSTM encoder and a one-layer LSTM decoder, with 256-dimensional hidden states and 128-dimensional word embeddings. For this model, we chose the 50k most frequently-used words in the dataset as the fixed vocabulary, and the word embeddings were learned from scratch during training. We trained the model using

Adagrad [43] with a learning rate of 0.15 and an initial accumulator value of 0.2. Gradient clipping was used with a maximum gradient norm of two. The number of epochs was determined by early stopping on the validation set.

4.1.3. Performance Measurement

We adopted the widely-used ROUGEmetric [44] for the evaluation of the summarization performance. ROUGE stands for Re-call-Oriented Understudy for Gisting Evaluation. Formally, ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-N is computed as follows:

$$ROUGE - N = \frac{\sum_{S \in \{\text{Referencesummaries}\}} \sum_{gram_n \in S} Count_{match}\left(gram_n\right)}{\sum_{S \in \{\text{Referencesummaries}\}} \sum_{gram_n \in S} Count\left(gram_n\right)} \tag{12}$$

where n stands for the length of the n-gram, $gram_n$, and $Count\left(gram_n\right)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. In this paper, we use full-length F1 score on ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (Longest common substring).

We also report the diversity statistics with the evaluation method proposed by [40], which counts the number of distinct n-grams present in the list of generated candidates of the beam search, and then divide these counts by the sentence length to bias against long sentences.

4.1.4. Baseline Methods

In order to verify the effectiveness of the proposed method, we compared our method with the results of state-of-the-art neural summarization methods reported in recent papers. Extractive models include a sentence ranking model called REFRESH [45], and SummaRuNNer [31], which is a recurrent neural network based sequence model. In addition, lead-3 is a strong extractive baseline that uses the first three sentences as the summary. Abstractive models include the original Pointer Generator Network (PGN) [36] and the Graph-Abs model [16] which applys the graph-based attention mechanism to seq2seq neural model.

4.2. Experiment Results

In this section, we first analyze the effectiveness of the two key components of the NASDD method and then evaluate the overall summarization performance of NASDD and baseline methods.

In particular, the experiments aim to answer three main questions:

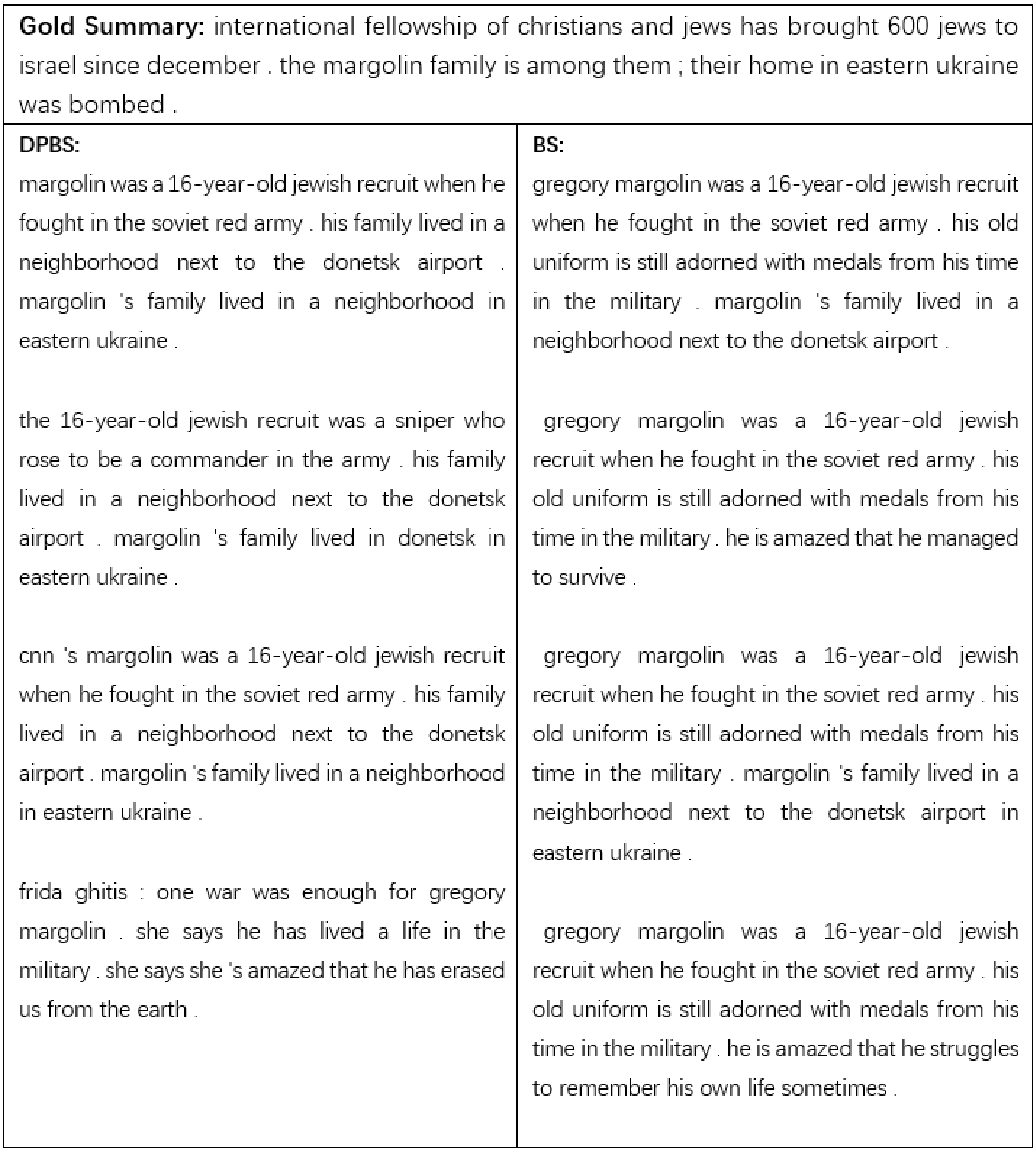
- Can the DPBS method effectively improve diversity?
- Can the optimal sequence-selection method locate the ideal sequences?
- Does the NASDD method achieve state-of-the-art performance on the abstractive summarization task?

4.2.1. DPBS Evaluation

In Table 1, we report the diversity statistics calculated according to the number of distinct n-grams produced by the standard beam search and the DPBS method. These two methods are applied to the same test set with the same beam size of four. As discussed in the previous section, the diversity statistics present how diverse the parallel generated sequences are. The standard beam search achieved 0.1714 (Distinct-1), 0.3627 (Distinct-2), and 0.4344 (Distinct-3). The DPBS exceeded this with 0.1953 (Distinct-1), 0.4476 (Distinct-2), and 0.6389 (Distinct-3). Distinct-1 measures how different the multiple generated sequences are at the single-word level, and Distinct-2 and Distinct-3 are more about the phrase level. The results in Table 2 show that DPBS had improvements on both sides. As the diversity was calculated using distinct n-grams divided by sequence length, the total number of the generated distinct n-grams of the method was considerably bigger than these figures suggest. To have an intuitive view of the diversity provided by our DPBS method, we used a sample news article from the test set and obtained the decoding outputs by DPBS and beam search. The results are show in Figure 4. Due to the original article being too long, we do not show it here. We used the human-written summary to have an overview of its content. As shown by this example, our DPBS effectively increased the diversity without much harm to the readability. We also provide the different decoding results of a battery of articles in this paper's Supplementary Material.

Table 1. Diversity statistics of the standard beam search and the Diversity-Promoting Beam Search (DPBS) method.

	Distinct-1	**Distinct-2**	**Distinct-3**
Standard Beam Search	0.1714	0.3627	0.4344
DPBS	0.1953	0.4476	0.6389



Gold Summary: international fellowship of christians and jews has brought 600 jews to israel since december . the margolin family is among them ; their home in eastern ukraine was bombed .

DPBS:	**BS:**
margolin was a 16-year-old jewish recruit when he fought in the soviet red army . his family lived in a neighborhood next to the donetsk airport . margolin 's family lived in a neighborhood in eastern ukraine .	gregory margolin was a 16-year-old jewish recruit when he fought in the soviet red army . his old uniform is still adorned with medals from his time in the military . margolin 's family lived in a neighborhood next to the donetsk airport .
the 16-year-old jewish recruit was a sniper who rose to be a commander in the army . his family lived in a neighborhood next to the donetsk airport . margolin 's family lived in donetsk in eastern ukraine .	gregory margolin was a 16-year-old jewish recruit when he fought in the soviet red army . his old uniform is still adorned with medals from his time in the military . he is amazed that he managed to survive .
cnn 's margolin was a 16-year-old jewish recruit when he fought in the soviet red army . his family lived in a neighborhood next to the donetsk airport . margolin 's family lived in a neighborhood in eastern ukraine .	gregory margolin was a 16-year-old jewish recruit when he fought in the soviet red army . his old uniform is still adorned with medals from his time in the military . margolin 's family lived in a neighborhood next to the donetsk airport in eastern ukraine .
frida ghitis : one war was enough for gregory margolin . she says he has lived a life in the military . she says she 's amazed that he has erased us from the earth .	gregory margolin was a 16-year-old jewish recruit when he fought in the soviet red army . his old uniform is still adorned with medals from his time in the military . he is amazed that he struggles to remember his own life sometimes .

Figure 4. Example from the CNN/Daily Mail test dataset showing the outputs generated by both DBPS and standard Beam Search (BS).

Table 2. ROUGE F1 scores on the test set. All of these ROUGE scores have a 95% confidence interval of at most ± 0.25 as reported by the official ROUGE script. PGN, Pointer Generator Network; NASDD, Neural Abstractive Summarization with Diverse Decoding.

	Rouge		
	1	**2**	**L**
lead-3	40.34	17.70	36.57
REFRESH	40.0	18.2	36.6
SummaRuNNer	39.6	16.2	35.3
Graph-Abs	38.1	13.9	34.0
PGN	38.15	16.46	35.37
NASDD (w/o selection)	39.03	16.02	35.71
NASDD	39.97	17.58	37.91

4.2.2. Selection Method Evaluation

Figure 5 shows the different results achieved by twenty different combinations of λ and η for the selection method. These experiments were all performed on the same validation set. As shown in the score function (11), $\lambda \sum_t \sum_i a_i^t \omega_i$ indicates the salient information preservation and $\eta \sum_{\omega_j \in x} \omega_j$ indicates the coverage of extractive keywords. Figure 5 has verified this by demonstrating the positive association between λ and ROUGE-L in the left part and the positive association between η and ROUGE-1 in the right part. ROUGE-l measures single-word overlapping, and ROUGE-L is more about sentence-level structure similarity. Although there are declines in both parts of Figure 5, like ROUGE-2 in the left part, the decline is in a small range, and proper choices of λ and η can reinforce each other. In NASDD, we tuned the value of λ and η on the validation set and finally set $\lambda = 0.4$ $\eta = 0.2$. In addition, the effectiveness of the two parameters has also proven that our DPBS method provided sufficient diversity, thus providing ample room for these factors to work.

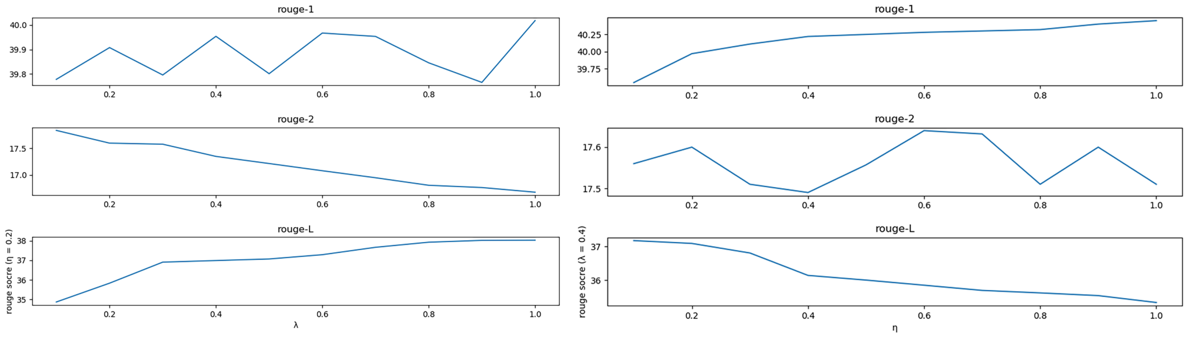


Figure 5. ROUGEscores achieved on the validation set with various value combinations of λ and η. The two parameters affected the result from different aspects.

4.2.3. Overall Summarization Performance

The results in Table 2 show the overall performance evaluation of baseline models and our proposed NASDD model. Firstly, our proposed method exceeded the state-of-the-art neural summarization model (PGN) by over one ROUGE point (+1.82 ROUGE-1, +1.12 ROUGE-2, + 2.54 ROUGE-L). Then, the extractive methods were still tough to beat, but our method achieved competitive performance. The sixth row of Table 2 is NASDD without the proposed selection method, which only uses the DPBS algorithm in decoding and adopts the conventional maximal possibility selection method for the final candidates of DPBS. The results were 39.53 (ROUGE-1), 16.02 (ROUGE-2), and 35.71 (ROUGE-L). This shows that without the selection method, only using the original maximal possibility selection method cannot achieve better, sometimes even worse, results than the baseline.

As the DPBS method is to improve the diversity of the beam search method to increase the possibility of generating better sequences, at the same time, worse sequences may also be generated, so the selection method capable of filtering bad candidate sequences becomes necessary. Moreover, the ROUGE evaluation has its own limitation: the evaluation is only performed based on the n-gram overlap between the generated sequences and the ground truth summaries written by a human. This means sequences containing insufficient words appeared in the ground truth summaries and will get very low ROUGE scores even when they are expressing the same idea. The selection method augmented the important words' influence in the selection process, and these words were also highly likely to appear in the ground truth summaries. In summary, the DPBS method can generate better sequences as summaries with the increase of diversity, and the selection method can effectively choose these sequences.

5. Conclusions and Future Work

In neural abstractive summarization, beam search is the most prevalent approximate inference algorithm for the decoding process; however, it suffers from a lack of diversity. The highly-similar sequences generated by standard beam search means the waste of computational resources and the elimination of potential better choice. In this paper, for the task of neural summarization, we present the Diversity-Promoting Beam search (DPBS) to encourage sequences generated from multiple roots and thus generate more diverse outputs. We also design an effective method to select better sequences based on saliency estimation. The DPBS method is applicable to any sequence model that uses the standard beam search and brings considerable improvement of performance. Our novel selection method creatively uses the attention vector for saliency estimation, providing a new promising solution for sequence evaluation in neural language generation. Experimental results show that the proposed NASDD method improves the performance with respect to the state-of-the-art neural summarization methods on the CNN/Daily Mail summarization task.

This work suggests several interesting directions for future research. We will try to integrate other methods to improve the coherence of the more diverse generated sequences. Moreover, we will further investigate methods to improve diversity from the sentence level to the discourse level, as a reasonable solution for generating multiple summaries with a single input document.

Supplementary Materials: The following are available online at http://www.mdpi.com/2076-3417/9/3/386/s1

Author Contributions: X.-W.H., H.-T.Z., J.-Y.C., and C.-Z.Z.; Conceptualization, X.-W.H. and H.-T.Z.; methodology, X.-W.H., H.-T.Z., and J.-Y.C.; software, X.-W.H.; validation, X.-W.H., H.-T.Z., and C.-Z.Z.; formal analysis, X.-W.H. and J.-Y.C.; investigation, X.-W.H. and H.-T.Z.; resources, H.-T.Z. and C.-Z.Z.; data curation, X.-W.H. and J.-Y.C.; writing, original draft preparation, X.-W.H.; writing, review and editing, X.-W.H., H.-T.Z., J.-Y.C., and C.-Z.Z.; visualization, X.-W.H.; supervision, H.-T.Z.; project administration, C.-Z.Z.; funding acquisition, H.-T.Z.

Acknowledgments: This research is supported by the National Natural Science Foundation of China (Grant No. 61773229), the Basic Scientific Research Program of Shenzhen City (Grant No. JCYJ20160331184440545), and the Overseas Cooperation Research Fund of the Graduate School at Shenzhen, Tsinghua University (Grant No. HW2018002).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schwenk, H. Continuous space translation models for phrase-based statistical machine translation. In Proceedings of the COLING 2012: Posters, Mumbai, India, 8–15 December 2012; pp. 1071–1080.
2. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.
3. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.

4. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
5. Rush, A.M.; Chopra, S.; Weston, J. A Neural Attention Model for Abstractive Sentence Summarization. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 379–389.
6. Chopra, S.; Auli, M.; Rush, A.M. Abstractive sentence summarization with attentive recurrent neural networks. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2016; pp. 93–98.
7. Nallapati, R.; Zhou, B.; Santos, C.N.D.; Gulcehre, C.; Xiang, B. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. In Proceedings of the Conference on Computational Natural Language Learning, Berlin, Germany, 11–12 August 2016; pp. 280–290.
8. Sankaran, B.; Mi, H.; Al-Onaizan, Y.; Ittycheriah, A. Temporal attention model for neural machine translation. *arXiv* **2016**, arXiv:1608.02927.
9. Cao, Z.; Li, W.; Li, S.; Wei, F.; Li, Y. AttSum: Joint Learning of Focusing and Summarization with Neural Attention. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 547–556.
10. Zeng, W.; Luo, W.; Fidler, S.; Urtasun, R. Efficient summarization with read-again and copy mechanism. *arXiv* **2016**, arXiv:1611.03382.
11. Chen, Q.; Zhu, X.; Ling, Z.; Wei, S.; Jiang, H. Distraction-based neural networks for modeling documents. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 2754–2760.
12. Zhou, Q.; Yang, N.; Wei, F.; Zhou, M. Selective Encoding for Abstractive Sentence Summarization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 1095–1104.
13. Nema, P.; Khapra, M.M.; Laha, A.; Ravindran, B. Diversity driven attention model for query-based abstractive summarization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 1063–1072.
14. Khan, A.; Salim, N.; Farman, H.; Khan, M.; Jan, B.; Ahmad, A.; Ahmed, I.; Paul, A. Abstractive Text Summarization based on Improved Semantic Graph Approach. *Int. J. Parallel Programm.* **2018**, *46*, 1–25. [CrossRef]
15. Torres-Moreno, J.M. *Automatic Text Summarization*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
16. Tan, J.; Wan, X.; Xiao, J. Abstractive document summarization with a graph-based attentional neural model. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 1171–1181.
17. Erkan, G.; Radev, D.R. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* **2004**, *22*, 457–479. [CrossRef]
18. Nenkova, A.; Vanderwende, L.; McKeown, K. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In Proceedings of the 29th Annual International ACM SIGIR Conference On Research and Development in Information Retrieval, Seattle, DC, USA, 6–11 August 2006; pp. 573–580.
19. Mihalcea, R. Language independent extractive summarization. In Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions, Ann Arbor, MI, USA, 25–30 June 2005; pp. 49–52.
20. Gillick, D.; Favre, B. A scalable global model for summarization. In Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing, Boulder, CO, USA, 4 June 2009; pp. 10–18.
21. Shen, D.; Sun, J.T.; Li, H.; Yang, Q.; Chen, Z. Document Summarization Using Conditional Random Fields. In Proceedings of the IJCAI 2007, Hyderabad, India, 6–12 January 2007; pp. 2862–2867.
22. Conroy, J.M.; O'leary, D.P. Text summarization via hidden markov models. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, USA, 9–13 September 2001; pp. 406–407.
23. Knight, K.; Marcu, D. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artif. Intell.* **2002**, *139*, 91–107. [CrossRef]

24. Clarke, J.; Lapata, M. Global inference for sentence compression: An integer linear programming approach. *J. Artif. Intell. Res.* **2008**, *31*, 399–429. [CrossRef]
25. Banko, M.; Mittal, V.O.; Witbrock, M.J. Headline generation based on statistical translation. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, China, 3–6 October 2000; pp. 318–325.
26. Kobayashi, H.; Noguchi, M.; Yatsuka, T. Summarization based on embedding distributions. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1984–1989.
27. Yin, W.; Pei, Y. Optimizing Sentence Modeling and Selection for Document Summarization. In Proceedings of the IJCAI 2015, Buenos Aires, Argentina, 25–31 July 2015; pp. 1383–1389.
28. Cao, Z.; Wei, F.; Dong, L.; Li, S.; Zhou, M. *Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization*; AAAI: Menlo Park, CA, USA, 2015; pp. 2153–2159.
29. Cao, Z.; Wei, F.; Li, S.; Li, W.; Zhou, M.; Houfeng, W. Learning summary prior representation for extractive summarization. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; Volume 2, pp. 829–833.
30. Cheng, J.; Lapata, M. Neural summarization by extracting sentences and words. *arXiv* **2016**, arXiv:1603.07252.
31. Nallapati, R.; Zhai, F.; Zhou, B. *SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents*; AAAI: Menlo Park, CA, USA, 2017; pp. 3075–3081.
32. Gu, J.; Lu, Z.; Li, H.; Li, V.O.K. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In Proceedings of the Meeting of the Association for Computational Linguistics, Toulouse, France, 6–11 July 2001; pp. 1631–1640.
33. Vinyals, O.; Fortunato, M.; Jaitly, N. Pointer networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 2692–2700.
34. Hermann, K.M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; Blunsom, P. Teaching machines to read and comprehend. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 1693–1701.
35. Paulus, R.; Xiong, C.; Socher, R. A deep reinforced model for abstractive summarization. *arXiv* **2017**, arXiv:1705.04304.
36. See, A.; Liu, P.J.; Manning, C.D. Get To The Point: Summarization with Pointer-Generator Networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 1073–1083.
37. Gulcehre, C.; Ahn, S.; Nallapati, R.; Zhou, B.; Bengio, Y. Pointing the Unknown Words. In Proceedings of the Meeting of the Association for Computational Linguistics, Toulouse, France, 6–11 July 2001; pp. 140–149.
38. Tu, Z.; Lu, Z.; Liu, Y.; Liu, X.; Li, H. Modeling Coverage for Neural Machine Translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 1, pp. 76–85.
39. Mi, H.; Sankaran, B.; Wang, Z.; Ittycheriah, A. Coverage Embedding Models for Neural Machine Translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 955–960.
40. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B. A Diversity-Promoting Objective Function for Neural Conversation Models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 110–119.
41. Li, J.; Jurafsky, D. Mutual information and diverse decoding improve neural machine translation. *arXiv* **2016**, arXiv:1601.00372.
42. Vijayakumar, A.K.; Cogswell, M.; Selvaraju, R.R.; Sun, Q.; Lee, S.; Crandall, D.; Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv* **2016**, arXiv:1610.02424 .
43. Duchi, J.; Hazan, E.; Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **2011**, *12*, 2121–2159.

44. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out, Post Conference Workshop of ACL 2004, Barcelona, Spain, 25–26 July 2004.
45. Narayan, S.; Cohen, S.B.; Lapata, M. Ranking Sentences for Extractive Summarization with Reinforcement Learning. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 1, pp. 1747–1759.