

# Neural Headline Generation with Sentence-wise Optimization

Ayana, Shiqi Shen, Yu Zhao, Zhiyuan Liu\*, Maosong Sun

Department of Computer Science and Technology,  
State Key Lab on Intelligent Technology and Systems,  
National Lab for Information Science and Technology, Tsinghua University, Beijing, China

## Abstract

Recently, neural models have been proposed for headline generation by learning to map documents to headlines with recurrent neural networks. Nevertheless, as traditional neural network utilizes maximum likelihood estimation for parameter optimization, it essentially constrains the expected training objective within word level rather than sentence level. Moreover, the performance of model prediction significantly relies on training data distribution. To overcome these drawbacks, we employ minimum risk training strategy in this paper, which directly optimizes model parameters in sentence level with respect to evaluation metrics and leads to significant improvements for headline generation. Experiment results show that our models outperforms state-of-the-art systems on both English and Chinese headline generation tasks.

## 1 Introduction

Automatic text summarization is the process of creating a coherent, informative and brief summary for a document. Text summarization is expected to understand the main theme of the documents and then output a condensed summary contains as many key points of the original document as it can within a length limit. Text summarization approaches can be divided into two typical categories: extractive and generative. Most extractive summarization systems simply select a subset of existing sentences from original documents as summary. Despite of its simplicity, extractive summarization has some intrinsic drawbacks, e.g., unable to generate coherent and compact summary in arbitrary length or shorter than one sentence.

In contrast, generative summarization builds semantic representation of a document and creates a summary with sentences not necessarily presenting in the original document explicitly. When the generated summary is required to be a single compact sentence, we name the summarization task as *headline generation* (Dorr et al., 2003). Most previous works heavily rely on modeling latent linguistic structures of input document, via syntactic parsing and semantic parsing, which always bring inevitable errors and degrade summarization quality.

Recent years have witnessed great success of deep neural models for various natural language processing tasks (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Ranzato et al., 2015) including text summarization. Taking neural headline generation (NHG) for example, it learns to build a large neural network, which takes a document as input and directly outputs a compact sentence as headline of the document. Compared with conventional generative methods, NHG exhibits the following advantages: (1) NHG is fully data-driven, requiring no linguistic information. (2) NHG is completely end-to-end, which does not explicitly model latent linguistic structures, and thus prevents error propagation. Moreover, the attention mechanism (Bahdanau et al., 2015) is introduced in NHG, which learns a soft alignment over input document to generate more accurate headline (Rush et al., 2015).

Nevertheless, NHG still confronts a significant problem: current models are mostly optimized at the word level instead of sentence level, which prevents them from capturing various aspects of summarization quality. In fact, it is essentially desirable to incorporate the implicit sentence-wise information contained in the evaluation criteria, e.g. ROUGE, into NHG model.

To address this issue, we propose to apply the *minimum Bayes risk* technique in tuning NHG

---

Corresponding author: Z. Liu (liuzy@tsinghua.edu.cn)

model with respect to evaluation metrics. Specifically, we utilize *minimum risk training* (MRT), which aims at minimizing a sentence-wise loss function over the training data. To the best of our knowledge, although MRT has been widely used in many NLP tasks such as statistical machine translation (Och, 2003; Smith and Eisner, 2006; Gao et al., 2014; Shen et al., 2015), it has not been well considered in the research of text summarization.

We conduct experiments on three real-world datasets in English and Chinese respectively. Experiment results show that, NHG with MRT can significantly and consistently improve the summarization performance as compared to NHG with MLE, and other baseline systems. Moreover, we explore the influence of employing different evaluation metrics and find the superiority of our model stable in MRT.

## 2 Background

In this section, we formally define the problem of neural headline generation and introduce the notations used in our model. Denote the input document  $\mathbf{x}$  as a sequence of words  $\{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ , where each word  $\mathbf{x}_i$  comes from a fixed vocabulary  $V$ . Headline generator aims to take  $\mathbf{x}$  as input, and generates a short headline  $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$  with length  $N < M$ , such that the conditional probability of  $\mathbf{y}$  given  $\mathbf{x}$  is maximized. The log conditional probability can be further formalized as:

$$\log \Pr(\mathbf{y}|\mathbf{x}; \theta) = \sum_{j=1}^N \log \Pr(\mathbf{y}_j|\mathbf{x}, \mathbf{y}_{<j}; \theta), \quad (1)$$

where  $\mathbf{y}_{<j} = \{\mathbf{y}_1, \dots, \mathbf{y}_{j-1}\}$ ,  $\theta$  indicates model parameters. That is, the  $j$ -th word  $\mathbf{y}_j$  in headline is generated according to all  $\mathbf{y}_{<j}$  generated in past and the input document  $\mathbf{x}$ . In NHG, we adopt an encoder-decoder framework to parameterize  $\Pr(\mathbf{y}_j|\mathbf{x}, \mathbf{y}_{<j}; \theta)$ , as shown in Fig. 1.

The encoder of the model encodes the input document  $\mathbf{x}$  into low-dimensional vectors  $(\mathbf{h}_1, \dots, \mathbf{h}_i, \dots, \mathbf{h}_M)$  using bi-directional recurrent neural network with GRU units, where  $\mathbf{h}_i$  is the concatenation of forward and backward states corresponding to word  $x_i$ . Then, the decoder sequentially generates headline words based on these vectors and decoder hidden states, using a uni-directional GRU recurrent neural network

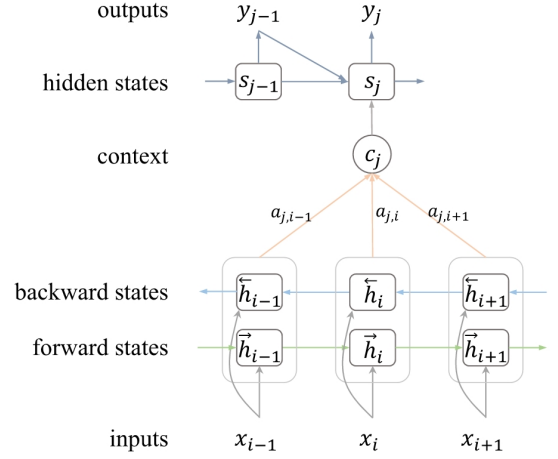


Figure 1: The framework of NHG.

with attention, i.e.,

$$\Pr(\mathbf{y}_j|\mathbf{x}, \mathbf{y}_{<j}; \theta) = \Pr(\mathbf{y}_j|\mathbf{c}_j, \mathbf{s}_j, \mathbf{y}_{<j}, \theta), \quad (2)$$

where  $\mathbf{c}_j$  stands for the context for generating the  $j$ -th headline word and is calculated utilizing the attention mechanism.  $\mathbf{s}_j$  is the  $j$ -th hidden state of the decoder, and  $\theta$  denotes a set of model parameters. Please refer to (Bahdanau et al., 2015; Sutskever et al., 2014) for more details.

## 3 Minimum Risk Training for NHG

Given a dataset  $D$  with large-scale document-headline pairs  $\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{x}^{(T)}, \mathbf{y}^{(T)})\}$ , we propose to use minimum risk training to optimize model parameters instead of the conventional maximum likelihood estimation. We employ the famous ROUGE evaluation metrics to compose the expected loss function. In this section, we introduce our basic idea in detail.

### 3.1 Minimum Risk Training

In the traditional training strategy, the optimized NHG model parameters are estimated by maximizing the log likelihood of generated headlines over training set  $D$ :

$$\mathcal{L}_{\text{MLE}}(\theta) = \sum_{(\mathbf{x}, \mathbf{y}) \in D} \log \Pr(\mathbf{y}|\mathbf{x}; \theta). \quad (3)$$

According to Eq. (1), the training procedure is fundamentally maximizing the probability of each word in headline step by step, which will inevitably lose global information. Moreover,  $\mathbf{y}_{<j}$  are authentic words from reference headline in the

training phase, while in the training phase they are predicted by model. It will lead to error propagation and inaccurate headline generation.

In order to tackle these problems, we propose to use minimum risk training (MRT) strategy. Given a document  $\mathbf{x}$ , we define  $\mathcal{Y}(\mathbf{x}; \theta)$  as the set of all possible headlines generated with parameters  $\theta$ . Regarding  $\mathbf{y}'$  as the reference headline of  $\mathbf{x}$ , we denote  $\Delta(\mathbf{y}', \mathbf{y})$  as the distance between  $\mathbf{y}$  and a generated headline  $\mathbf{y}'$ . MRT defines the objective loss function as follows:

$$\mathcal{L}_{\text{MRT}}(\theta) = \sum_{(\mathbf{x}, \mathbf{y}) \in D} \mathbb{E}_{\mathcal{Y}(\mathbf{x}; \theta)} \Delta(\mathbf{y}', \mathbf{y}). \quad (4)$$

Here  $\mathbb{E}_{\mathcal{Y}(\mathbf{x}; \theta)}$  indicates the expectation over all possible headlines. Thus the objective function of MRT can be further formalized as:

$$\mathcal{L}_{\text{MRT}}(\theta) = \sum_{(\mathbf{x}, \mathbf{y}) \in D} \sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{x}; \theta)} \Pr(\mathbf{y}' | \mathbf{x}; \theta) \Delta(\mathbf{y}', \mathbf{y}). \quad (5)$$

In this way, MRT manages to minimize the expected loss by perceiving the distance as a measure of assessing the overall risk. Nevertheless, it is usually time-consuming and inefficient to enumerate all possible instances. For simplicity, we draw a subset of samples  $\mathcal{S}(\mathbf{x}; \theta) \subset \mathcal{Y}(\mathbf{x}; \theta)$  from the current probability distribution of generated headlines. The loss function can be approximated as:

$$\mathcal{L}_{\text{MRT}}(\theta) = \sum_{(\mathbf{x}, \mathbf{y}) \in D} \sum_{\mathbf{y}' \in \mathcal{S}(\mathbf{x}; \theta)} \frac{\Pr(\mathbf{y}' | \mathbf{x}; \theta)^\epsilon}{\sum_{\mathbf{y}^* \in \mathcal{S}(\mathbf{x}; \theta)} \Pr(\mathbf{y}^* | \mathbf{x}; \theta)^\epsilon} \Delta(\mathbf{y}', \mathbf{y}),$$

where  $\epsilon$  is a hyper-parameter that controls the smoothness of the objective function (Och, 2003). A proper  $\epsilon$  value can significantly enhance the effectiveness of MRT. In the experiment, we set  $\epsilon$  to  $5 \times 10^{-3}$ .

### 3.2 ROUGE

MRT exploits the distance between two sentences to compose the loss function, which enables us to directly optimize NHG model with respect to a specific evaluation metric of the task. As we know, the most widely used evaluation metric for document summarization is ROUGE (Lin, 2004). The basic idea of ROUGE is to count the number of overlapping units between computer-generated

summaries and the reference summaries, such as overlapped n-grams, word sequences, and word pairs. It is the most common evaluation metric in Document Understanding Conference (DUC), a large-scale summarization evaluation sponsored by NIST (Lin, 2004). When training English models, we adopt negative recall value of ROUGE-1,2 and L to compose  $\Delta(\mathbf{y}', \mathbf{y})$ . For Chinese models, we utilize negative F1 value of ROUGE-1,2 and L to compose  $\Delta(\mathbf{y}', \mathbf{y})$ .

## 4 Experiments

In this paper, we evaluate our methodology on both English and Chinese headline generation tasks. We first introduce the datasets and evaluation metrics used in the experiment. Then we demonstrate that our model performs the best compared with state-of-the-art baseline systems. We also analyze the influence of different parameters in detail to gain more insights.

### 4.1 Datasets and Evaluation Metrics

**English Datasets.** In the experiment, we utilize the English Gigaword Fifth Edition (Parker et al., 2011) corpus, containing 10 million news articles<sup>1</sup> with corresponding headlines. We follow the experimental settings in (Rush et al., 2015) to collect 4 million article-headline pairs as the training set.

We use the dataset from DUC-2004 Task-1 as our test set. It consists of 500 news articles, each of which is paired with four human-generated reference headlines.<sup>2</sup> We also take Gigaword test set<sup>3</sup> to evaluate our models. We use the DUC-2003 evaluation dataset of size 624 as development set to tune the hyper-parameters in MRT.

(6) **Chinese Dataset.** We conduct experiments on the Chinese LCSTS dataset (Hu et al., 2015), consisting of article-headline pairs extracted from Sina Weibo<sup>4</sup>. LCSTS are composed of three parts, containing 2.4 million, 10,666 and 1,106 article-headline pairs respectively. Those pairs in Part-II

<sup>1</sup>To avoid noises in articles and headlines that may influence the performance, we filter out headlines with by-lines, extraneous editing marks and question marks. For English dataset, we preprocess the corpus with tokenization and lower-casing.

<sup>2</sup>The dataset can be downloaded from <http://duc.nist.gov/> with agreements.

<sup>3</sup>This dataset is provided by (Rush et al., 2015).

<sup>4</sup>The website of Sina Weibo is <http://weibo.com/>. A typical news article posted in Weibo is limited to 140 Chinese characters, and the corresponding headline is usually set in a pair of square brackets at the beginning of the news article.

and Part-III are labeled with relatedness scores by human annotation that indicate how relevant an article and its headline are<sup>5</sup>. In the experiment, we take Part-I of LCSTS as training set, Part-II as development set and Part-III as test set. We only reserve those pairs with scores no less than 3. It is worth mentioning that, we take Chinese characters as inputs of NHG instead of words in order to prevent the influence of Chinese word segmentation errors. In addition, we replace all digits with # for both English and Chinese corpus.

**Evaluation metrics.** In the experiment, we use ROUGE (Lin, 2004), as introduced in Section 3.2, to evaluate the performance of headline generation.

Following (Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016), for DUC2003 and DUC2004, we report recall scores of ROUGE-1( $R1_R$ ), ROUGE-2( $R2_R$ ) and ROUGE-L( $RL_R$ ) with official 75 bytes ceiling limit. And following (Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016), for Gigaword test set, we report full-length F1 scores of ROUGE-1( $R1_{F1}$ ), ROUGE-2( $R2_{F1}$ ) and ROUGE-L( $RL_{F1}$ ). Since a shorter summary tends to get lower recall score, when testing on DUC datasets, we set the minimum length of a generated headline as 10. Note that we report 75 bytes capped recall scores only. In this case summaries that longer than 75 bytes obtain no bonus on recall scores. Due to full-length F1 makes the evaluation result unbiased to summary length, we set no limitation to headline length when testing on Gigaword test set.

For Chinese, we report full-length F1 scores ( $R1_{F1}$ ,  $R2_{F1}$  and  $RL_{F1}$ ) following previous works (Hu et al., 2015; Gu et al., 2016). We set no length limitation on Chinese experiments either.

## 4.2 Baseline Systems

### 4.2.1 English Baseline systems

**TOPIARY** (Zajic et al., 2004) is the winner system of DUC2004 Task-1. It utilizes linguistic-based sentence compression method and unsupervised topic detection at the same time.

**MOSES+** (Rush et al., 2015) generates headlines based on MOSES, a widely-used phrase-based machine translation system (Koehn et al., 2007). It also enlarges the phrase table and uses

MERT to improve the quality of generated headlines.

**ABS** and **ABS+** (Rush et al., 2015) are both attention-based neural models that generate short summary for given sentence. The difference is that ABS+ extracts additional n-gram features at word level to revise the output of ABS model.

**RAS-Elman** and **RAS-LSTM** (Chopra et al., 2016) both utilize convolutional encoders that take input words and word position information into account. They also make use of attention-based decoders. The difference is that, RAS-Elman selects Elman RNN (Elman, 1990) as decoder, while RAS-LSTM selects long short term memory architecture (Hochreiter and Schmidhuber, 1997).

**BWL**, namely big-words-lvt2k-lsent (Nallapati et al., 2016) implements a trick that restricts the vocabulary size at the decoder end, by means of constructing the vocabulary of documents in each mini-batch respectively (Jean et al., 2015).

All the English baseline systems listed above except TOPIARY utilize Gigaword dataset for training, as described in Section 4.1.

### 4.2.2 Chinese Baseline systems

**RNN-context** (Hu et al., 2015) is a simple character based encoder-decoder architecture that takes the concatenation of all hidden states at the encoder end as the input of decoder end.

**COPYNET** (Gu et al., 2016) incorporates copying mechanism into sequence-to-sequence framework, which replicates certain segments from the input sentence into the output sentence.

## 4.3 Implementation Details

In MLE, the word embeddings are randomly initialized and then updated during training. In MRT, we initialize model parameters using the optimized parameters learned from NHG with MLE. For English models, we set the word embedding dimension to 620, the hidden unit size to 1,000 and the vocabulary size to 30,000. The corresponding values for Chinese models are 400, 500 and 3,500 respectively. In particular, the size of subset  $\mathcal{S}(\mathbf{x}; \theta)$  in Eq.(6) has a great impact on the performance. When the size is too small, the sampling will not be sufficient. When the size is too large, the learning time will grow correspondingly. In this paper, we set the size to 100 to achieve a trade-off between effectiveness and efficiency. These samples are drawn from the probability distribution of generated headlines by the up-to-date

<sup>5</sup>Each pair in Part-II is labeled by only one annotator, and in Part-III is by three annotators.

		evaluation metric		
criterion	loss	R1	R2	RL
MLE	N/A	23.70	7.85	21.20
MRT	−R1	<b>28.81</b>	9.58	<b>25.31</b>
	−R2	26.94	9.56	24.01
	−RL	28.19	<b>9.64</b>	25.02

Table 1: Effects of distance measures on the English validation set. −R1, −R2 and −RL represent the opposite value of ROUGE-1, ROUGE-2 and ROUGE-L respectively.

		evaluation metric		
criterion	loss	R1	R2	RL
MLE	N/A	24.61	8.52	22.00
MRT	−R1	<b>29.84</b>	10.24	<b>26.33</b>
	−R2	27.97	10.18	24.99
	−RL	29.18	<b>10.44</b>	25.88

Table 2: Effects of using different distance measures on the English test set.

NHG model<sup>6</sup>. We use AdaDelta (Zeiler, 2012) to adapt learning rates in stochastic gradient descent for both MLE and MRT. We utilize no dropout or regularization, but we take gradient clipping during training and the training is early stopped based on DUC2003 data. All our models are trained on GeForce GTX TITAN GPU. For NHG+MLE on the English dataset, it takes about 2.5 hours for each 10,000 iterations, For NHG+MRT, it takes about 3.75 hours. During testing, we use beam search of size 10 (Chopra et al., 2016) to generate headlines.

#### 4.4 Choices of Model Setup

In the training process of NHG model, there are several significant factors that greatly influence the performance, such as the choice of distance measure in loss function and the treatment of unknown words. To determine the most appropriate choices of model setup, we investigate the effects of these factors on the development set respectively.

##### 4.4.1 Effect of Distance Measure

As described in Section 3.2, the distance  $\Delta(\mathbf{y}', \mathbf{y})$  in the loss function is computed by the negative value of ROUGE. We investigate the effect of utilizing various distance measures in MRT. Table 1 shows the experiment results on English development set using different evaluation metrics. We find that all NHG+MRT models consistently outperform NHG+MLE, which indicates

<sup>6</sup>An alternative subset building strategy is to choose top- $k$  headlines. Considering the efficiency and parallel architecture of GPUs, we opt sampling.

	$R1_R$	$R2_R$	$RL_R$
Original	28.08	9.19	25.00
Ignore	28.81	9.58	25.31
Copy	<b>29.68</b>	<b>9.98</b>	<b>25.94</b>
Mapping	29.62	9.94	25.91

Table 3: Effect of using different UNK Post Processing methods on English development set.

	$R1_R$	$R2_R$	$RL_R$
Input-only	27.17	8.98	23.96
Extended-input	28.08	9.19	24.50
Full-vocab	<b>29.68</b>	<b>9.98</b>	<b>25.94</b>

Table 4: Effect of using different restrictions of output vocabulary on English development set.

that the MRT technique is robust when loss function varies. −R1<sub>R</sub> statistically brings significant improvement for all evaluation metrics over MLE, one possible reason is that R1 score correlates well with human judgement (Lin and Och, 2004). Hence we decide to utilize −R1<sub>R</sub> as the default semantic measure in the experiments (e.g., in Section 4.5). In addition, Table 2 shows that this argument is still valid on DUC2004.

##### 4.4.2 Effect of UNK Post Processing

In the training procedure of NHG model, a common experiment setup is to keep a fixed size of vocabulary on both input and output side. These vocabularies are usually shortlists that only contain most frequent words, so that the out of vocabulary words are usually mapped to a special token “UNK”.

There are three typical post-processing methods to deal with UNK tokens. A simplest way is to ignore them, and we denote it as **Ignore**. Another way (Jean et al., 2015) is to copy words from original input directly, and we denote it as **Copy**. The third way is to replace the unknown words according to a dictionary built upon the whole training set, and we denote it as **Mapping**. We conduct experiments on the English development set to investigate the performance of these methods. The fixed vocabulary size in NHG model is set to 30,000. Experiment results shown in table 3 indicate that the “Copy” method performs the best among three methods and generally improves the original model. Hence, we decide to utilize it as the default post processing method in our experiments on the test set.

#### 4.5 Evaluation Results

Table 5 shows the evaluation results of headline generation on different English test sets.

System	Training	Model Architecture	DUC-2004			Gigaword		
			$R1_R$	$R2_R$	$RL_R$	$R1_{F1}$	$R2_{F1}$	$RL_{F1}$
Non-neural systems								
TOPIARY	—	Linguistic-based	25.12	6.46	20.12	-	-	-
MOSES+		Phrase-based	26.50	8.13	22.85	-	-	-
Neural systems								
ABS	MLE	Attention-based enc + NNLM	26.55	7.06	22.05	29.55	11.32	26.42
ABS+		ABS + Extractive tuning	28.18	8.49	23.81	29.76	11.88	26.96
RAS-Elman		CNN enc + Elman-RNN dec	28.97	8.26	24.06	33.78	15.97	31.15
RAS-LSTM		CNN enc + LSTM dec	27.41	7.69	23.06	32.55	14.70	30.03
BWL		G-RNN enc + G-RNN dec + trick	28.35	9.46	24.59	33.17	16.02	30.98
this work	MLE	G-RNN enc + G-RNN dec	24.92	8.60	22.25	32.67	15.23	30.56
	MRT	G-RNN enc + G-RNN dec	<b>30.41</b>	<b>10.87</b>	<b>26.79</b>	<b>36.54</b>	<b>16.59</b>	<b>33.44</b>

Table 5: Comparison with baseline systems on DUC-2004 and Gigaword English test sets. G-RNN stands for Gated Recurrent Neural Networks, enc and dec are shorts for encoder and decoder respectively.

System	Training	Model	LCSTS		
			$R1_{F1}$	$R2_{F1}$	$RL_{F1}$
RNN-context	MLE	G-RNN enc + G-RNN dec + minimum length	29.9	17.4	27.2
COPYNET		G-RNN enc + G-RNN dec + Copy mechanism	35.0	22.3	32.0
<i>this work</i>	MLE	G-RNN enc + G-RNN dec	34.9	23.3	32.7
	MRT	G-RNN enc + G-RNN dec	<b>38.2</b>	<b>25.2</b>	<b>35.4</b>

Table 6: Comparison with baseline systems on Chinese test set. Note that the RNN-context has the same model architecture as ours. But they set a minimum length limit when decoding.

The baseline systems are introduced in Section 4.2. These results indicate that NHG model with MLE achieves comparable performance to existing headline generation systems. Moreover, replacing MLE with MRT significantly and consistently improves the performance of NHG model, and outperforms the state-of-the-art systems on both test set.

Similar results can be observed from the experiment results on Chinese headline generation task as well, as shown in Table 6<sup>7</sup>. NHG with MRT improves the ROUGE scores up to over 3 points compared with baseline systems. We also notice that MLE model is already better than (Hu et al., 2015) and comparable to (Gu et al., 2016). This indicates that a character based model indeed performs good on Chinese summary task. Moreover, when evaluating with F1 scores, longer summaries would be penalized and get lower scores.

Figure 2 shows the R1 scores of headlines generated by NHG+MLE and NHG+MRT on the English dataset with respect to input lengths. As we can see, NHG+MRT consistently improves over NHG+MLE for all lengths.

To reduce the computation complexity when training NHG model, a possible approach is to restrict the size of vocabulary for generated headlines. There are three typical methods to deal with

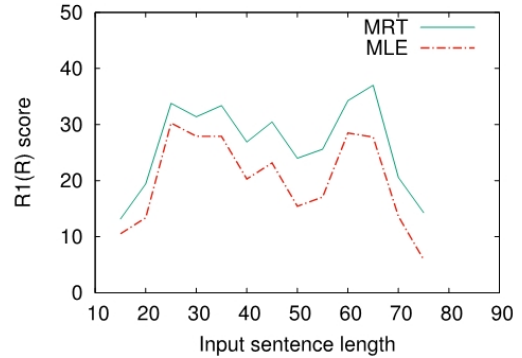


Figure 2: Recall scores of ROUGE-1 on DUC-2004 test set over various input lengths.

the size. The first one is to restrict the output words of headline within the input sentence, denoted as **Input-only**. The second one is to construct an extended vocabulary that includes similar words with those appear in the input sentence, denoted as **Extended-input**. The third one is to use full vocabulary, denoted as **Full-vocab**. Table 4 illustrates the experiment results of using different restrictions on the English development set. The extended vocabulary is constructed by collecting 100 nearest neighbors for each input word, according to pre-trained Google-News word vectors (Mikolov et al., 2013). We observe that the “Input-only” and “Extended-input” achieve comparable performance while using hundreds of times less vocabulary. It indicates that it is feasible to utilize these tricks to train NHG model much more

<sup>7</sup>The MRT result reported here is obtained by taking the negative F1 score of ROUGE-1 as loss function. Several related experiment results are not given due to the length limit.

<b>Article (1):</b>	Jose Saramago became the first writer in Portuguese to win the Nobel prize for literature on Thursday , his personal delight was seconded by a burst of public elation in his homeland .
<b>Reference:</b>	Jose Saramago becomes first writer in Portuguese to win Nobel prize for literature
<b>NHG+MLE:</b>	Portuguese becomes Portuguese president to win the nobel prize for literature
<b>NHG+MRT:</b>	Jose Saramago is the first writer in the Portuguese language to win Nobel
<b>Article (2):</b>	A slain Russian lawmaker was honored Tuesday as a martyr to democratic ideals in a stately funeral service in which anger mingled freely with tears .
<b>Reference:</b>	Russian lawmaker buried beside greats; mourned as martyr; killers unknown.
<b>NHG+MLE:</b>	Slain Russian lawmaker remembered as martyr to democracy ( Moscow )
<b>NHG+MRT:</b>	Slain Russian lawmaker honored as martyr in stately funeral service
<b>Article (3):</b>	Voting mainly on party lines on a question that has become a touchstone in the debate over development and preservation of wilderness , the Senate on Thursday approved a gravel road through remote wildlife habitat in Alaska .
<b>Reference:</b>	Senate approves 30-mile road in Alaskan wilderness; precedent? veto likely.
<b>NHG+MLE:</b>	US senate passes law allowing road drilling in Alaska , Alaska
<b>NHG+MRT:</b>	Senate passes gravel road through Alaska wildlife habitat in Alaska

Table 7: Examples of original articles, reference headlines and generated outputs by different training strategy on DUC-2004 test set.

efficiently.

#### 4.6 Case Study

We present several examples for comparison as shown in Table 7. We can observe that: (1) NHG with MRT is generally capable of capturing the core information of an article. For example, the main subject in Article 1 is “Jose Saramago”. NHG+MRT can successfully find the correct topic and generate a headline about it, but NHG+MLE failed. (2) When both systems capture the same topic, NHG+MRT can generate more informative headline. For Article 2, NHG+MLE generates “remembered as” when NHG+MRT generates “honored as”. Considering the context, “honored as” would be more appropriate. (3) NHG+MLE usually suffer from generating duplicated words or phrases in headlines. As shown in Article 3, NHG+MLE repeats the phrase “Alaska” several times which leads to a semantically incomplete headline. NHG+MRT seems to be able to overcome this problem, benefitting from directly optimizing sentence-level ROUGE.

### 5 Related Work

Headline generation is a well-defined task standardized in DUC-2003 and DUC-2004. Various approaches have been proposed for headline generation: rule-based, statistical-based and neural-based.

The rule-based models create a headline for a news article using handcrafted and linguistically motivated rules to guide the choice of a potential headline. Hedge Trimmer (Dorr et al., 2003) is a representative example of this approach which creates a headline by removing constituents from

the parse tree of the first sentence until it reaches a specific length limit. Statistical-based methods make use of large scale training data to learn correlations between words in headlines and articles (Banko et al., 2000). The best system on DUC-2004, TOPIARY (Zajic et al., 2004) combines both linguistic and statistical information to generate headlines. There is also method make use of knowledge bases to generate better headlines. With the advances of deep neural networks, there are growing works that design neural networks for headline generation. (Rush et al., 2015) proposes an attention-based model to generate headlines. (Filippova et al., 2015) proposes a recurrent neural network with long short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) for headline generation. (Gu et al., 2016) introduces copying mechanism into encoder-decoder architecture inspired by the Pointer Networks (Vinyals et al., 2015).

In this work, we propose the NHG model realized by a bidirectional recurrent neural network with gated recurrent units. We also propose to apply minimum risk training (MRT) to optimize parameters of NHG model. MRT has been widely used in machine translation (Och, 2003; Smith and Eisner, 2006; Gao et al., 2014; Shen et al., 2015), but less been explored in document summarization. To the best of our knowledge, this work is the first attempt to utilize MRT in neural headline generation.

### 6 Conclusion and Future Work

In this paper, we build an end-to-end neural headline generation model, which does not require heavy linguistic analysis and is fully data-

driven. We apply minimum risk training for model optimization, which effectively incorporates sentence-wise information by taking various evaluation metrics into consideration. Evaluation result shows that NHG with MRT achieves significant and consistent improvements on both English and Chinese datasets, as compared to state-of-the-art baseline systems including NHG with MLE. There are still many open problems to be explored as future work: (1) Besides article-headline pairs, there are also rich plain text data not considered in NHG training. We will investigate the probability of integrating these plain texts to enhance NHG for semi-supervised learning. (2) We will investigate the hybrid approach of incorporating NHG with other successful headline generation approaches like sentence compression models.

## References

- [Bahdanau et al.2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.
- [Banko et al.2000] Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of ACL*, pages 318–325.
- [Cho et al.2014] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of EMNLP*, pages 1724–1734.
- [Chopra et al.2016] Sumit Chopra, Michael Auli, Alexander M. Rush, and SEAS Harvard. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of NAACL*.
- [Dorr et al.2003] Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of HLT-NAACL*, pages 1–8.
- [Elman1990] Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.
- [Filippova et al.2015] Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with lstms. In *Proceedings of EMNLP*, pages 360–368.
- [Gao et al.2014] Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. 2014. Learning continuous phrase representations for translation modeling. In *Proceedings of ACL*, pages 699–709.
- [Gu et al.2016] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of ACL*.
- [Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, pages 1735–1780.
- [Hu et al.2015] Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. In *Proceedings of EMNLP*, pages 1967–1972.
- [Jean et al.2015] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of ACL-IJCNLP*, pages 1–10.
- [Koehn et al.2007] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*, pages 177–180.
- [Lin and Och2004] Chin-Yew Lin and FJ Och. 2004. Looking for a few good metrics: Rouge and its evaluation. *NTCIR Workshop*.
- [Lin2004] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL*, pages 74–81.
- [Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, pages 3111–3119.
- [Nallapati et al.2016] Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization. *CoRR*.
- [Och2003] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167.
- [Parker et al.2011] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition, june.
- [Ranzato et al.2015] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.
- [Rush et al.2015] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of EMNLP*, pages 379–389.
- [Shen et al.2015] Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2015. Minimum risk training for neural machine translation. *arXiv preprint arXiv:1512.02433*.



- [Smith and Eisner2006] David A Smith and Jason Eisner. 2006. Minimum risk annealing for training log-linear models. In *Proceedings of COLING/ACL*, pages 787–794.
- [Sutskever et al.2014] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*, pages 3104–3112.
- [Vinyals et al.2015] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems* 28, pages 2674–2682. Curran Associates, Inc.
- [Zajic et al.2004] David Zajic, Bonnie Dorr, and Richard Schwartz. 2004. Bbn/umd at duc-2004: Topiary. In *Proceedings of HLT-NAACL*, pages 112–119.
- [Zeiler2012] Matthew D. Zeiler. 2012. Adadelata: An adaptive learning rate method. *Computer Science*.