Attention With Sparsity Regularization for Neural Machine Translation and Summarization

Jiajun Zhang, *Member, IEEE*, Yang Zhao, Haoran Li, and Chengqing Zong, *Senior Member, IEEE*

***Abstract*—The attention mechanism has become the *de facto* standard component in neural sequence to sequence tasks, such as machine translation and abstractive summarization. It dynamically determines which parts in the input sentence should be focused on when generating each word in the output sequence. Ideally, only few relevant input words should be attended to at each decoding time step and the attention weight distribution should be sparse and sharp. However, previous methods have no good mechanism to control this attention weight distribution. In this paper, we propose a sparse attention model in which a sparsity regularization term is designed to augment the objective function. We explore two kinds of regularizations: L_∞-norm regularization and minimum entropy regularization, both of which aim to sharpen the attention weight distribution. Extensive experiments on both neural machine translation and abstractive summarization demonstrate that our proposed sparse attention model can substantially outperform the strong baselines. And the detailed analyses reveal that the final attention distribution indeed becomes sparse and sharp.**

***Index Terms*—Sequence to sequence learning, attention mechanism, sparsity regularization, machine translation, summarization.**

I. Introduction

ATTENTION mechanism is introduced in [1] for neural machine translation and quickly becomes the *de facto* component in neural sequence prediction tasks, such as machine translation, document summarization and image caption. The attention model leads to the state-of-the-art performance in these tasks [2]–[8], no matter which kind of network architecture is employed, such as recurrence [4], convolution [9] and purely attention [10].

In this article, we concentrate ourselves on neural machine translation and abstractive summarization, in which the attention mechanism guides the sequential prediction process by selectively focusing on specific words in the input sequence when predicting each output word. This procedure calculates a weight distribution over all the input words at each decoding time step and the higher the attention weights the more contribution of the corresponding words. Since only few relevant input words are responsible to generate a specific output word, the attention distribution should be sparse and sharp in most cases. However, the conventional training algorithm has no good mechanism to control the attention distribution.

The existing work on improving the attention model mainly attempt at incorporating the prior knowledge to influence the attention weights. For instance, [2], [11] assume that the generation of each output word only attends to a local continuous window of the input sequence and propose a local attention model. Inspired by statistical machine translation [12] in which a coverage vector is maintained to record whether or not a source word has been translated already, the coverage model is introduced into neural machine translation and abstractive summarization [7], [13], [14]. Similarly, the fertility mechanism [15], [16] is proposed to model the desired number of output words for each source word. [17], [18] suppose that the attention weight distribution should be consistent with the automatically learned word alignments and then present the supervised attention models accordingly.

Note that, these previous methods do not provide effective constraints for the attention distribution at each specific decoding step. Take Chinese-to-English neural machine translation in Fig. 1 for example. When predicting the target word after *is*, the attention distribution in NMT focuses uniformly on three source words while the correct one should attend most to the word *xia4du2(poison)*. The uniform distribution makes the decoder difficult to determine which source word should be translated at this decoding step. As a result, the wrong target word UNK is produced.

Ideally, the attention weight distribution should be sparse and sharp. The sparsity constraint requires that most of the input words have the attention weights close to 0.0. The sharpness constraint requires that the attention weight of the most relevant input words should be as big as possible. Some recent studies show that sparsity can be achieved by refining the softmax function which calculates the attention weights [19]–[21]. This kind of approaches employ a soft-thresholding operator that sets the weights under a threshold to zeros. As for sharpness, [22], [23] demonstrate that slightly modifying the softmax function with low temperature can achieve this goal. Note that these methods

Manuscript received June 10, 2018; revised November 12, 2018; accepted November 19, 2018. Date of publication November 28, 2018; date of current version December 14, 2018. This work was supported in part by the Natural Science Foundation of China under Grant 61673380 and in part by the Beijing Advanced Innovation for Language Resources of Beijing Language and Culture University. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ani Nenkova. (*Corresponding author: Jiajun Zhang*.)

J. Zhang, Y. Zhao, and H. Li are with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jjzhang@nlpr.ia.ac.cn; yang.zhao@nlpr.ia.ac.cn; haoran.li@nlpr.ia.ac.cn).

C. Zong is with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China, with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China (e-mail: cqzong@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TASLP.2018.2883740

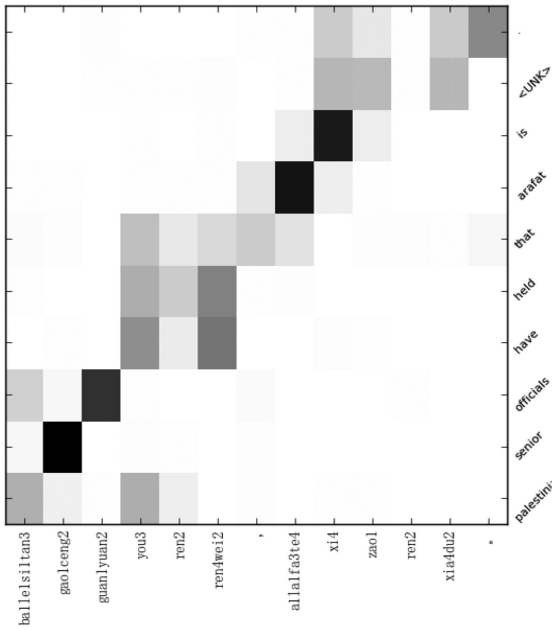


Fig. 1. A sequence to sequence translation example in which the attention distribution does not focus on the most related word *xia4du2 (poison)* when generating the word after *is* and *UNK* is produced rather than the correct translation *poisoned*.

utilize hard constraints. The sparsity and sharpness of the attention weight distribution are not directly optimized according to an objective function.

Instead of refining the softmax function, we propose in this article a novel attention model in which a sparsity regularization term is introduced to control the attention weight distribution. We investigate two kinds of regularizations. One is the L_∞-**norm regularization** that dynamically maximizes the biggest attention weight and enforces other attention weights to approach 0.0. The other is the **minimum entropy regularization** which makes the attention weight distribution as sharp as possible by minimizing the entropy of the distribution. The regularizer influences only the objective function and does not introduce any additional network parameters. Therefore, the sparse model can be applied in any attention-based sequence to sequence learning framework.

We evaluate the proposed methods on two recognized sequence prediction tasks, namely neural machine translation and abstractive summarization. The translation experiments on Chinese-English and German-English demonstrate that both of the proposed sparsity regularization terms can significantly improve the translation performance over the strong baseline. Furthermore, our proposed method can also substantially boost the word alignment quality even compared to the coverage model [13]. The abstractive summarization experiments on CNN/Daily Mail dataset show that our minimum entropy regularizer can significantly outperform the state-of-the-art methods.

II. Attention-Based Sequence to Sequence Learning

Our attention model with sparsity constraints can be employed in any attention-based sequence to sequence learning framework. Without loss of generality, we use the framework similar to Google's Neural Machine Translation (GNMT) [4], in which both of the encoder and decoder apply the stacked LSTM (Long-Short Term Memory [24]) layers. The network structure is shown in Fig. 2.

Given the input text[1] $X = (x_1, x_2, \cdots, x_{T_x})$, the encoder transforms X into a sequence of abstract context representations $C = (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_{T_x})$ whose size is the same as the length of the input text. Then, from the context vectors C the decoder generates the output sequence[2] $Y = (y_1, y_2, \cdots, y_{T_y})$ one word each time by maximizing the probability of $p(y_i|y_{<i}, C)$. Note that x_j (y_i) is the j_{th} (i_{th}) word (or token) in the input (output) sequence and we utilize $\mathbf{x}_i$ $(\mathbf{y}_j)$ to denote the embedding of x_j (y_i). Next, we briefly review the encoder introducing how to obtain C and the decoder addressing how to calculate $p(y_i|y_{<i}, C)$.

The **encoder** employs m stacked LSTM layers to learn the context vectors $C = (\mathbf{h}_1^m, \mathbf{h}_2^m, \cdots, \mathbf{h}_{T_x}^m)$. In the k-th layer ($k > 1$), $\mathbf{h}_j^k$ is calculated as follows:

$$\mathbf{h}_j^k = \text{LSTM}\left(\mathbf{h}_{j-1}^k, \mathbf{h}_j^{k-1}\right) \tag{1}$$

For notation simplification, we omit here the cell memory state from the previous time step when calculating the LSTM unit.

In the first layer ($k = 1$), $\mathbf{h}_j^1$ is obtained through a bidirectional LSTM:

$$\overrightarrow{\mathbf{h}_j^1} = \text{LSTM}\left(\overrightarrow{\mathbf{h}_{j-1}^1}, \mathbf{x}_j\right) \tag{2}$$

$$\overleftarrow{\mathbf{h}_j^1} = \text{LSTM}\left(\overleftarrow{\mathbf{h}_{j+1}^1}, \mathbf{x}_j\right) \tag{3}$$

Given $\overrightarrow{\mathbf{h}_j^1}$ and $\overleftarrow{\mathbf{h}_j^1}$, $\mathbf{h}_j^1$ is calculated with a feed-forward neural network[3] $\mathbf{h}_j^1 = \tanh(W_h^l \cdot \overrightarrow{\mathbf{h}_j^1} + W_h^r \cdot \overleftarrow{\mathbf{h}_j^1} + b_h)$.

The **decoder** computes the conditional probability $p(y_i|y_{<i}, C)$ with the help of **attention** mechanism [1] that leverages different input context $\mathbf{c}_i$ at different decoding time step:

$$p(y_i|y_{<i}, C) = p(y_i|y_{<i}, \mathbf{c}_i) = \text{softmax}(W\hat{\mathbf{z}}_i)) \tag{4}$$

where $\hat{\mathbf{z}}_i$ is the attention output:

$$\hat{\mathbf{z}}_i = \tanh\left(W_c[\mathbf{z}_i^l; \mathbf{c}_i]\right) \tag{5}$$

in which $\mathbf{z}_i^l$ is the top hidden state of the decoder network and $\mathbf{z}_i^k$ in the k-th layer is computed using the following formula:

$$\mathbf{z}_i^k = \text{LSTM}\left(\mathbf{z}_{i-1}^k, \mathbf{z}_i^{k-1}\right) \tag{6}$$

If $k = 1$, $\mathbf{z}_i^1$ will be calculated by combining $\hat{\mathbf{z}}_{i-1}$ as feed input [2]:

$$\mathbf{z}_i^1 = \text{LSTM}\left(\mathbf{z}_{i-1}^1, \mathbf{y}_{i-1}, \hat{\mathbf{z}}_{i-1}\right) \tag{7}$$

The dynamic context vector $\mathbf{c}_i$ is the weighted sum of the source-side context vectors and is calculated by the attention

[1] In machine translation, the input text is a source sentence. And in abstractive summarization, the input text is a document (or a paragraph).

[2] The output sequence is the target language sentence in machine translation and is the condensed short summary in abstractive summarization.

[3] Normally, $\overrightarrow{\mathbf{h}_j^k}$ and $\overleftarrow{\mathbf{h}_j^k}$ are concatenated to form $\mathbf{h}_j^k$. In this work, we use feed-forward neural network instead since it leads to better translation performance.

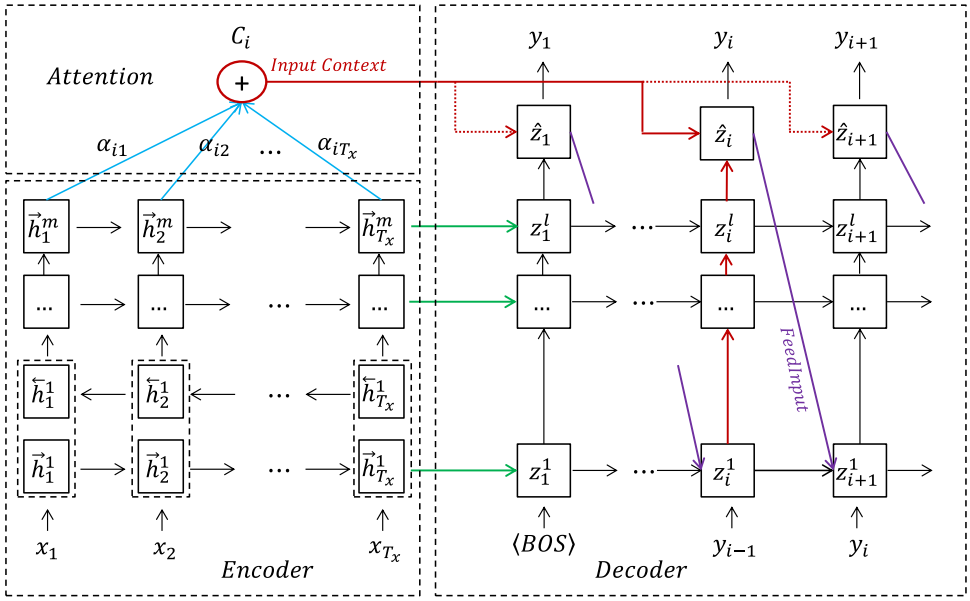


Fig. 2. It is the architecture of the sequence to sequence learning framework in which stacked LSTMs are employed for both encoder and decoder. The attention mechanism is the bridge between the input and output.

model just as illustrated in the left part of Fig. 2:

$$\mathbf{c}_i = \sum_{j=1}^{T_x} \alpha_{ij} \mathbf{h}_j^m \tag{8}$$

where α_{ij} is a normalized item calculated as follows:

$$e_{ij} = \mathbf{v}_a^\intercal \tanh\left(W_a \mathbf{z}_i^l + U_a \mathbf{h}_j^m\right) \tag{9}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'} \exp(e_{ij'})} \tag{10}$$

In which $\mathbf{v}_a$ is a parameter vector. The greater the value of the variable α_{ij}, the more contribution of the j-th input word to the generation of the i-th output word.

Given the ***training data*** $\mathcal{D} = \{(X^{(n)}, Y^{(n)})\}_{n=1}^N$ (source-target sentence pairs in machine translation and document-summary pairs in abstractive summarization), all the parameters θ are optimized to maximize the following conditional log-likelihood:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{T_y} \log p\left(y_i^{(n)} | y_{<i}^{(n)}, X^{(n)}, \theta\right) \tag{11}$$

III. Attention With Sparsity Regularization

The standard attention model introduced in the previous section (Eq. 10) has no direct constraints on the weight distribution $\alpha_i = (\alpha_{i1}, \cdots, \alpha_{ij}, \cdots, \alpha_{iT_x})$. The previous work [13], [15] and our empirical study find that the objective function (Eq. 11) cannot guarantee reasonable weight distribution in which only very few relevant weights are positive and others are close to 0. Especially, the attention weight distributions may be uniform as illustrated in Fig. 3(a). Obviously, this kind of attention weight distribution is not so reasonable and should be constrained during training.

Intuitively, the attention weight distribution α_i for each output word should be both sparse and sharp. Accordingly, we attempt to fully investigate the regularization method that controls the distribution of the attention vector α_i in the objective function (Eq. 11):

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{T_y} \left(\log p(y_i^{(n)} | y_{<i}^{(n)}, X^{(n)}, \theta) + \lambda R(\alpha_i)\right) \tag{12}$$

in which $R(\alpha_i)$ is the regularization term that controls the attention weight distribution at each decoding time step. λ is the hyper-parameter which balances the log-likelihood and the regularization term.

In the optimization theory, L_0-norm and L_1-norm are the most frequently used optimization methods to control the sparsity of the parameters which are usually coefficients of variables. L_0-norm measures the number of non-zero elements in a vector and minimizing L_0-norm will guarantee the sparsity of the parameter vector. Since L_0-norm is computationally difficult to use due to its non-convex property, L_1-norm is usually employed to approximate L_0-norm.

However, we will show here that L_1-norm is not suitable in our scenario. For the attention weight distribution α_i, L_1-norm is:

$$||\alpha_i||_1 = \sum_{j=1}^{T_x} |\alpha_{ij}| \tag{13}$$

Since α_i is a normalized vector (see Eq. 10), the above L_1-norm $||\alpha_i||_1$ will always be constant 1.0. It is easy to see that L_1-norm cannot adequately constrain the attention weight distribution. Accordingly, we explore in this work other two regularization methods: maximum L_∞-norm regularizer and minimum entropy regularizer.

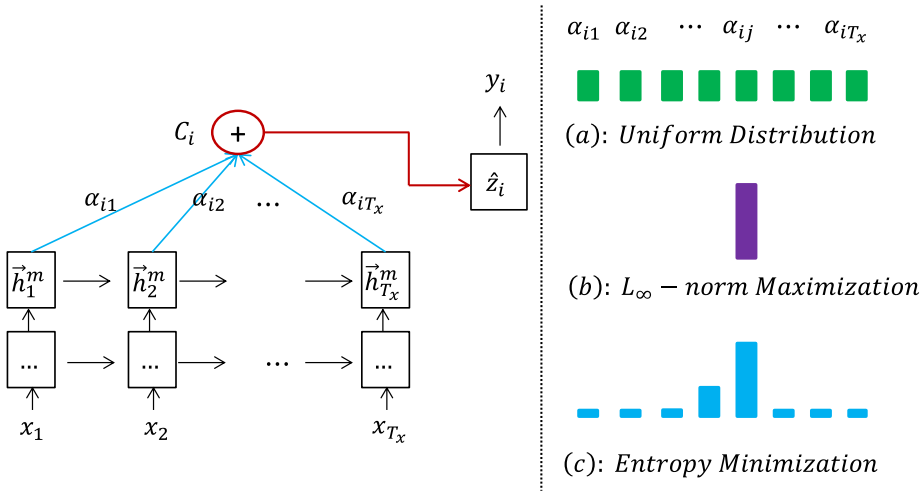


Fig. 3. The attention weight distribution, in which (a) shows a special case of uniform distribution in standard attention model; (b) gives the distribution constrained with a L_∞-norm regularization; and (c) illustrates the distribution obtained by minimizing its entropy.

A. Maximum L_∞-Norm Regularizer

L_∞-norm returns the maximum element of a value vector and it is not commonly used in the optimization problems. In the image processing community, several researchers [25], [26] employ **minimum L_∞-norm** methods to solve the geometric reconstruction problems, in which the value vector is the geometric distance between the measured image and the reprojected structure and motion. Minimizing the maximum value of the distance vector leads to well-reconstructed structures.

In our scenario, the value vector α_i is a probability distribution and sum of the vector is a constant 1.0. Under the summation constraint, the minimum L_∞-norm method will enforce the attention weight distribution to be as uniform as possible. Therefore, we instead utilize the **maximum L_∞-norm** method and design the regularization term in Eq. 12 as:

$$R(\alpha_i) = L_\infty(\alpha_i) = ||\alpha_i||_\infty = \max(\alpha_i) \tag{14}$$

Provided α_i sum to one, maximizing the biggest value of the weight distribution will make the maximum element approach 1.0 and others approximate 0.0. In attention-based sequence to sequence learning, this optimization method will make the generation process focus on the most relevant input word at each decoding step. Fig. 3(b) illustrates an attention weight distribution optimized by the maximum L_∞-norm regularizer.

This optimization method assumes that the generation of each output word depends only on one input word. However, this assumption does not hold in many cases, in which two or more input words contribute to one output word generation. For example, the Chinese multi-word *suo3you3 de0 ren2* is translated into only one English word *everyone*. Thus, when generating the target word *everyone*, the attention model should focus on three continuous words *suo3you3 de0 ren2*. Such many-to-one mappings are common in machine translation. The maximum L_∞-norm regularizer is more like a hard constraint and is unable to handle many-to-one cases, and it is more reasonable to design a soft constraint for the attention weight distribution. Accordingly, we introduce the minimum entropy approach.

B. Minimum Entropy Regularizer

In statistical learning community, **maximum entropy** principle is widely used to design a model which satisfies the prior constraints but makes uniform assumption for unknown events [27]. In contrast, **minimum entropy** is not often used. [28] and [29] apply minimum entropy as a regularizer to encode priors. For instance, [29] designs a novel semi-supervised learning framework in which unlabeled data is modeled as a prior distribution and the minimum entropy regularizer is employed to meet this prior.

In the attention-based sequence to sequence model, we argue that a good attention weight distribution should be sparse and sharp. Accordingly, the prior distribution has a low entropy and we therefore design a minimum entropy regularizer to encode this prior as follows.

$$Ent(\alpha_i) = -\sum_{j=1}^{T_x} \alpha_{ij} \log(\alpha_{ij})$$

$$R(\alpha_i) = -Ent(\alpha_i) = \sum_{j=1}^{T_x} \alpha_{ij} \log(\alpha_{ij}) \tag{15}$$

Since the overall objective is to maximize Eq. 12, the regularization term $R(\alpha_i)$ should be the negative entropy $-Ent(\alpha_i)$.

By minimizing the entropy of the attention weight distribution α_i, we expect that the mass of the attention weights are assigned to the few relevant source words. Fig. 3(c) demonstrates a specific attention weight distribution optimized by the minimum entropy regularizer.

C. Parameter Optimization for Regularizers

Gradient decent algorithm is usually applied during the back-propagation process to optimize the network parameters according to the objective function. Obviously, the objective function should be differentiable everywhere.

The difference between our new objective function Eq. 12 and the baseline Eq. 11 only lies in the regularization term $R(\alpha_i)$.

Since the entropy function (Eq. 15) is differentiable convex, it is trivial to learn the network parameters for the minimum entropy regularizer. However, L_∞ (Eq. 14) is nondifferentiable and requires some approximation techniques to calculate the gradients.

Following [30], we employ the **subgradient method** to deal with this pointwise maximum function. Specifically, we use *weak subgradient calculus* which can compute a subgradient for nondifferentiable function $f(x)$ as long as we can evaluate $f(x)$.

Given $L_\infty(\alpha_i) = \max\{\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{iT_x}\}$, we define $I(\alpha_i) = \{j|\alpha_{ij} = L_\infty(\alpha_i)\}$. The weak subgradient calculus utilizes the subgradient of any α_{ik} $(k \in I(\alpha_i))$ to approximate $\partial L_\infty(\alpha_i)$. If the norm is obtained by just one index k, the weak subgradient can be calculated as follows:

$$\frac{\partial L_\infty(\alpha_i)}{\partial \alpha_{ij}} = \text{sign}(\alpha_{ik})\delta_{kj} \tag{16}$$

Where δ_{kj} is the Kronecker delta function and $\delta_{kj} = 1$ if $k = j$, otherwise $\delta_{kj} = 0$.

IV. Comparison to Other Approaches

A. Softmax With Temperature

Modifying the softmax function with a temperature parameter can sharpen the weight distribution to some extent [22]:

$$\alpha_{ij} = \frac{\exp(e_{ij}/T)}{\sum_k \exp(e_{ik}/T)} \tag{17}$$

Where $T > 0$ is a temperature parameter and it will be the standard softmax function if $T = 1$. When we raise the temperature T, the result distribution of α_i will be softer. Conversely, the attention weight distribution will be sharper when T gets smaller. Fig. 4 shows the distribution for different temperatures. [23] further proposes **Gumbel-Softmax** as follows:

$$\alpha_{ij} = \frac{\exp(\log(e_{ij}) + g_j)/T)}{\sum_k \exp(\log(e_{ik}) + g_k)/T)} \tag{18}$$

where g_j is drawn from Gumbel(0,1) and readers can refer to [23] for more details.

However, these methods do not best fit our scenario. In previous studies, the input dimension of the softmax function (usually the neuron number of the final layer) is fixed. However, in our scenario the input dimension of the softmax function equals to the input sentence length which varies sentence by sentence. A fixed temperature may not be suitable for all the sentences with different lengths.

B. Sparsemax

Sparsity of the weight distribution can also be achieved by refining the softmax function. As an alternative to softmax, [19] presents **Sparsemax** transformation which can project the attention scores $\mathbf{e}_i = \{e_{i1}, e_{i2}, \cdots, e_{iT_x}\}$ onto the probability simplex $\alpha_\mathbf{i}$:

$$sparsemax(\mathbf{e}_i) := \underset{\alpha_i \in \Delta^{T_x}}{\text{argmin}} \, ||\alpha_i - \mathbf{e}_i||^2 \tag{19}$$

where $\alpha_i \in \Delta^{T_x}$ is a $(T_x - 1)$-dimensional probability simplex and therefore $\Delta^{T_x} := \{\alpha_i \in \mathcal{R}^{T_x} \,|\, \sum_j \alpha_{ij} = 1, \alpha_{ij} \geq 0\}$.

The weight distribution α_i can be obtained by:

$$sparsemax_j(\mathbf{e}_i) = \max\{0, \mathbf{e}_{ij} - \tau(\mathbf{e}_i)\} \tag{20}$$

in which $\tau(\mathbf{e}_i)$ is a soft threshold that is calculated according to the scores $\mathbf{e}_i$. Please see [19] for detailed information. [20], [21] generalize sparsemax and applied the strategy into neural machine translation. However, they report that these methods are just comparable to the softmax baseline in translation quality.

We can see that the core idea behind these approaches is to make the attention sparse through setting the weights below a threshold to zeros. It is more like hard constraint and the attention weights are not directly optimized according to an objective function.

V. Incorporating the Coverage Model

Our sparsity regularization shapes the attention weight distribution and chooses the most relevant input words at each decoding step. However, it has no mechanism to control the relationship between weight distributions at different decoding steps. As we emphasized before that the proposed regularization method can be adopted in any attention-based sequence to sequence learning framework, we can employ the attention mechanism enhanced with sophisticated techniques, such as the coverage model [7], [13], [14], as our baseline. In this work, besides the conventional attention-based model, we also apply our regularization methods to [13]'s coverage-based attention model which is just a slight modification of Eq. 9:

$$e_{ij} = \mathbf{v}_a^\intercal \tanh\left(W_a z_i^l + U_a \mathbf{h}_j^m + \mathcal{C}_{i-1,j}\right) \tag{21}$$

Where the new item $\mathcal{C}_{i-1,j}$ denotes the coverage degree of the input word x_j before decoding time step i and can be updated dynamically using a neural network:

$$\mathcal{C}_{i,j} = f\left(\mathcal{C}_{i-1,j}, \alpha_{ij}, \mathbf{z}_i^l, \mathbf{h}_j^m\right) \tag{22}$$

In which $f(\cdot)$ is a nonlinear activation function. It is easy to see that our proposed regularization method is orthogonal to the coverage model. In the experiments, we will combine these two techniques together to figure out the effectiveness.

VI. Experimental Setup

In this section we describe the experimental settings for machine translation and abstractive summarization.

A. Machine Translation

1) Dataset: We evaluate the proposed regularized attention model on both Chinese-to-English and German-to-English translation tasks. For the Chinese-to-English task, the training data includes about 2.1 M sentence pairs extracted from LDC corpora.[4] For validation, we choose NIST 2003 (MT03) dataset. For testing, we use NIST 2004-2006 (MT04-06), and NIST 2008 (MT08) datasets.

[4]LDC2000T50, LDC2002L27, LDC2002T01, LDC2002E18, LDC2003E07, LDC2003E14, LDC2003T17, LDC2004T07.

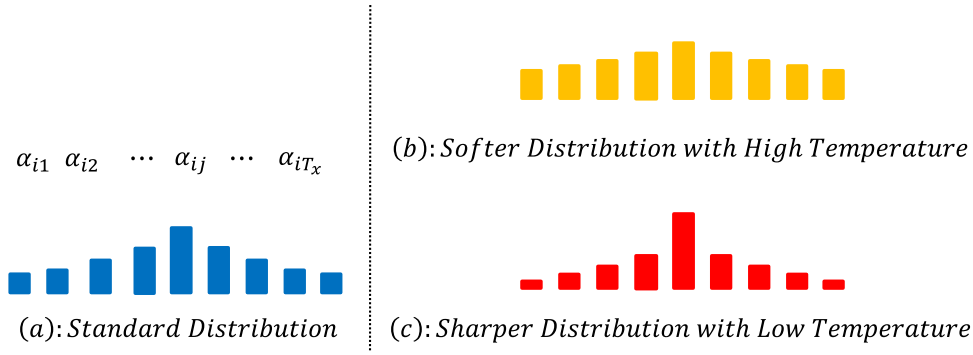


Fig. 4. Different attention weight distributions, in which (a) shows a weight distribution in standard attention model, (b) gives the softer weight distribution using high temperature in softmax, and (c) illustrates the sharper weight distribution with low temperature.

For the German-to-English task, we utilize the same subset of the WMT 2014 training corpus employed by [2], [31], [32]. It contains 4.5 M sentence pairs. The concatenation of news-test 2012 and news-test 2013 is used as the validation set. The news-test 2014 is employed as the test set.

2) Training and Evaluation Details: The baseline attention-based NMT and all the proposed methods are implemented by reusing Zoph_RNN[5] toolkit which is written in C++/CUDA and provides efficient training across multiple GPUs. The encoder includes two stacked LSTM layers and the first layer employs the bidirectional LSTMs. The decoder also contains two stacked LSTM layers followed by the softmax layer.

For Chinese-English, we keep the most frequent 50 K words as vocabulary. For Germen-English, we employ BPE [33] with 50 K merge operations. The dimension of word embedding and the size of hidden layers are all set to 1000.

Each NMT model is trained on Titan X using stochastic gradient decent (SGD) algorithm. We use a mini batch size of $B = 128$. We also apply dropout on each layer to avoid overfitting, and the dropout rate is set to 0.2. At test time, we employ beam search with beam size $b = 12$. We use case-insensitive BLEU score [34] to evaluate the final translation quality. Significance test is performed using the pairwise re-sampling approach [35].

3) Translation Systems: To test the effectiveness of our regularized attention models in machine translation, we investigate and compare multiple systems.

NMT: It is the baseline NMT system whose architecture is the same as that of Google's NMT system [4] (as shown in Fig. 2). Our system is implemented by extending Zoph_RNN with bidirectional encoding and global attention.

NMT+Coverage: It is the NMT system augmented with a coverage model for the attention [13] as shown in Eq. 21.

NMT+Temp: It is the system that utilizes **softmax with temperature parameter** in the attention model [22].

NMT+Gumbel-Softmax: It is the system that utilizes **Gumbel-Softmax** to calculate attention weights [23].

NMT+Sparsemax: It is the system that utilizes **Sparsetmax** for attention weight calculation [19].

[5]https://github.com/isi-nlp/Zoph_RNN

NMT+L_∞-norm: It is our proposed NMT system that employs **maximum L_∞-norm** regularization in the sparse attention model.

NMT+Entropy: It is our proposed NMT system which applies **minimum entropy** regularization.

NMT+Cov+Ent: It is a combination of two methods, namely **NMT+Coverage** and **NMT+Entropy**: This system explores the potential of the attention mechanism that combines both of the coverage model and the minimum entropy regularization model.

In regularization methods, we set balance item $\lambda = 0.01$ in Eq. 12 for all the experiments. We investigate the influence of different λ settings in the experiment analysis part.

B. Abstractive Summarization

1) Dataset: We perform abstractive summarization on the CNN/Daily Mail dataset [36]–[38], which consists of online news articles paired with short summaries. There are respectively 781 and 56 tokens on average in each article and the summary. Using the scripts provided by [37], the dataset is split into three parts: 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs.

2) Training and Evaluation Details: The baseline attention-based abstractive summarization model is shown in Fig. 2. The encoder is a single-layer bidirectional LSTM and the decoder is a single-layer unidirectional LSTM. The encoder and the decoder are connected with a global attention layer.

For fair comparison, we employ the same settings as done in [7]. All the models use a vocabulary of most frequent 50 K words for both input and output. We set the dimension of word embedding and hidden states to 128 and 256 respectively.

Our summarization models are trained on Titan X using Adagrad [39] (the same as [7]) with a learning rate 0.15. The batch size is set to 16. At test time, we adopt beam search with beam size $b = 4$ to produce summaries. All the model results are evaluated with ROUGE metric [40].

3) Summarization Systems: We investigate multiple abstractive summarization systems (ABS) to evaluate the effectiveness of our proposed attention regularization (minimum entropy regularizer) method.

ABS: It is the baseline abstractive summarization system using the global attention model.

TABLE I
TRANSLATION RESULTS (BLEU SCORE) FOR DIFFERENT TRANSLATION METHODS IN CHINESE-ENGLISH TASK. **ALL** DENOTES CONCATENATION OF ALL THE TEST SETS. $(0.8, 0.6, 0.4, 0.2, 0.1)$ DENOTE THE TEMPERATURES USED IN THE SOFTMAX FUNCTION DURING ATTENTION WEIGHT CALCULATION. Δ COLUMN PRESENTS THE IMPROVEMENT OF EACH METHOD COMPARED TO THE BASELINE NMT ON **ALL** TEST SET. * AND ** INDICATE THAT THE GAINS ARE STATISTICALLY SIGNIFICANT AT THE LEVEL $p < 0.05$ AND $p < 0.01$ RESPECTIVELY. + DENOTES THAT *NMT+Cov+Ent* PERFORMS SIGNIFICANTLY BETTER THAN *NMT+Coverage* AT THE LEVEL $p < 0.05$. THE COLUMN **ENTROPY** DENOTES AVERAGE ENTROPY PER TOKEN OF THE ATTENTION WEIGHT DISTRIBUTION ON **ALL** TEST SET

Method	MT03 (Dev)	MT04	MT05	MT06	MT08	ALL	Δ	Entropy
NMT	40.77	43.02	40.49	40.59	30.68	39.24	-	1.566
NMT+Coverage	41.23	43.35	40.98	40.94	30.83	39.65	**+0.41**	1.499
NMT+Temp-0.8	40.74	42.77	40.86	40.81	30.70	39.33	+0.09	1.202
NMT+Temp-0.6	40.39	42.36	40.50	40.31	30.67	39.11	-0.13	0.831
NMT+Temp-0.4	39.38	41.62	39.81	39.69	30.21	38.34	-0.90	0.591
NMT+Temp-0.2	38.18	40.52	38.52	38.44	29.21	37.16	-2.08	0.331
NMT+Temp-0.1	37.42	39.96	37.78	37.93	28.70	36.55	-2.69	0.208
NMT+Gumbel-Softmax-0.8	40.54	42.21	39.84	39.61	30.74	38.72	-0.52	1.230
NMT+Sparsemax	41.07	43.08	40.65	41.02	31.01	39.52	+0.28	0.810
NMT+L_∞-Norm	41.17	43.65	40.89	40.82	31.11	39.66	**+0.42**	1.223
NMT+Entropy	41.33	43.80	**41.61**	41.08	31.28	39.93	**+0.69***	0.915
NMT+Cov+Ent	**41.42**	**43.89**	41.37	**41.67**	**31.64**	**40.32**	**+1.08**$^{**+}$	0.903

ABS+Pointer: Based on ABS, the pointer-generator system proposed in [7] uses a pointer network that can directly copy words from input to output. This model is capable to handle OOV words and can generate better summaries.

ABS+Pointer+Cov: The pointer-generator system further introduces the coverage mechanism into the ABS model [7] to overcome under and over generation problems.

ABS+Pointer+Ent: Based on ABS+Pointer, we further employ our **minimum entropy** regularizer to control the attention distribution.

ABS+Pointer+Cov+Ent: It is the new system which combines the proposed **minimum entropy** regularizer with copy and coverage mechanisms.

TABLE II
TRANSLATION RESULTS (BLEU SCORE) FOR DIFFERENT TRANSLATION METHODS IN GERMAN-ENGLISH TASK

Method	tst12-13	tst14	Δ
NMT	25.37	26.26	-
NMT+Coverage	25.84	26.69	**+0.43**
NMT+Entropy	26.38	27.13	**+0.87***
NMT+Cov+Ent	**26.62**	**27.45**	**+1.19**$^{**+}$

TABLE III
EVALUATION RESULTS OF THE WORD ALIGNMENT QUALITY. THE LOWER THE SCORE, THE BETTER THE ALIGNMENT QUALITY

Method	AER	SAER
NMT	45.57	61.43
NMT+Coverage	43.68	59.38
NMT+Entropy	43.46	58.94
NMT+Cov+Ent	**43.37**	**58.46**

VII. RESULTS AND ANALYSIS

A. Machine Translation

1) Translation Quality: Table I reports the translation performance of different systems on the Chinese-English task. The results are split into seven parts. As shown in the first part of the table, the baseline is very strong and it obtains more than 40.0 BLEU score on most of the data sets. When augmenting the attention model with coverage information (the second part in Table I), the system **NMT+Coverage** can reach a moderate improvement of 0.41 BLEU points, suggesting that the coverage knowledge is beneficial to the attention-based NMT, which is in line with the findings of [13], [14].

The third part in Table I gives the results of the method **NMT+Temp**. When applying the temperature parameter, only smaller values (less than 1.0) can lead to sharper distributions. Thus, we try five different settings (from 0.8 to 0.1). We find that the temperature strategy cannot produce better translations. The performance is at most on par with the baseline if we use the temperature parameter 0.8. However, the translation quality decreases dramatically when the temperature parameter is smaller than 0.4. The **Gumbel-Softmax** method (the fourth part in Table I) cannot lead to better performance neither. It indicates that the temperature strategy is not a good method to control the attention weight distribution in neural machine translation. In contrast, the **Sparsemax** method can obtain a marginal improvement of average 0.28 BLEU points.

The last two parts in Table I demonstrate that the proposed regularization methods are much beneficial to the attention-based NMT. Specifically, the maximum L_∞-norm regularization (**NMT+L_∞-Norm**) achieves comparable results with the coverage model **NMT+Coverage**, while the minimum entropy regularization (**NMT+Entropy**) can obtain significant gains of 0.69 BLEU points over the baseline. The improvement can be up to 1.08 BLEU points if we combine coverage information with sparsity regularization constraints. Since NMT+Entropy outperforms NMT+L_∞-Norm, we will employ **minimum entropy** regularization in the remaining experiments.

Similar phenomena can be observed from the German-English translation results as shown in Table II. The finding is that **NMT+Entropy** outperforms **NMT+Coverage** and

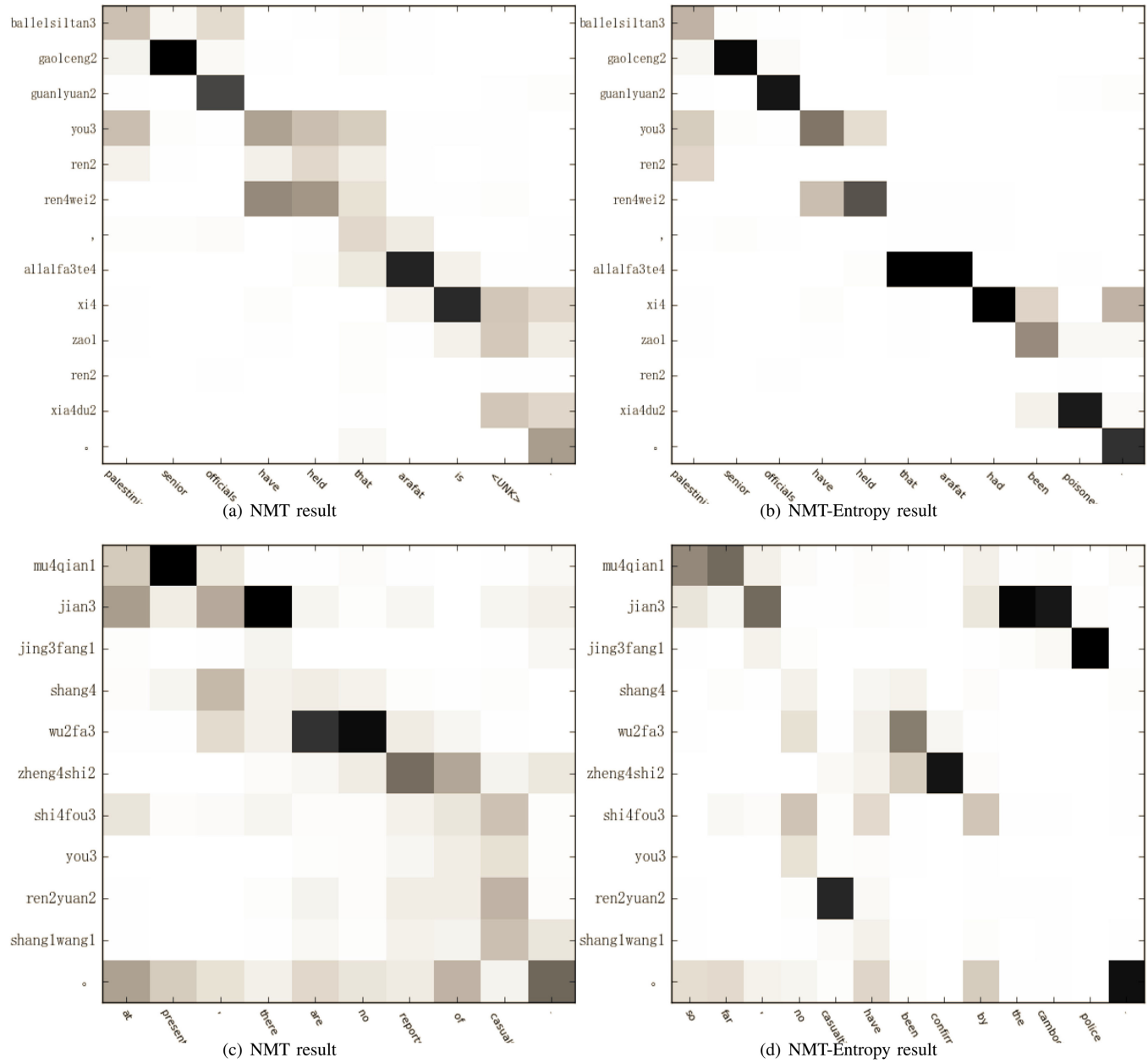


Fig. 5. Alignment examples: our regularization method leads to much more sparse and sharp attention weights, and accordingly obtains better translation results. (b) Correctly generates the word *poisoned* compared to (a). (d) obtains the correct translation *local residents* although there are still errors.

NMT+Cov+Ent performs best. Considering that only one reference is available, the improvements are very promising. Both of Table I and Table II strongly suggest that our sparsity regularization method and the coverage model are complementary to each other. It is quite reasonable because the sparsity regularization controls the current attention weight distribution while the coverage model makes use of previous attentions.

2) Word Alignment Quality: Better attention weight distribution may lead to more accurate word alignments. To figure out whether it is the case, we carry out a word alignment experiment on the evaluation dataset including 900 manually word aligned Chinese-English sentence pairs [41]. Following [13], we employ alignment error rate (AER) [42] and soft alignment error rate (SAER) to evaluate the alignment quality.

$$SAER = 1 - \frac{|M_A \times M_S| + |M_A \times M_P|}{|M_A| + |M_S|} \tag{23}$$

in which M denotes alignment matrix. S and P are respectively sure and possible links in the reference alignment and each link is assigned probability 1.0 (others are 0.0) in matrices M_S and M_P. A is the candidate alignment which can be obtained by different NMT systems and each element $M_A(i, j)$ of M_A is the attention weight between the i-th target word and the j-th source word.

Table III reports the word alignment performance for different NMT systems. It is easy to see that the trends are similar

TABLE IV
TRANSLATION RESULTS FOR DIFFERENT λS IN EQ. 12 ON THE CHINESE-ENGLISH TASK

λ	**MT03 (Dev)**	**MT04**	**MT05**	**MT06**	**MT08**	**ALL**
$\lambda = 0.005$	40.91	43.60	41.36	40.89	30.92	39.78
$\lambda = 0.01$	**41.33**	**43.80**	**41.61**	**41.08**	**31.28**	**39.93**
$\lambda = 0.02$	40.87	43.53	41.25	40.96	30.84	39.70
$\lambda = 0.04$	40.63	43.06	41.07	40.72	30.57	39.38
$\lambda = 0.08$	40.35	42.65	40.56	40.50	30.41	39.05

to that in translation results. Both of the coverage model and the sparsity regularization method can substantially reduce the alignment errors under two metrics. Furthermore, our proposed minimum entropy regularization method **NMT+Entropy** performs better than the coverage model **NMT+Coverage**. These results indicate that our sparsity regularization method is very effective to obtain gooyd word alignments.

To better illustrate the effectiveness of our regularization method, Fig. 5 gives two translation examples with attention weight distribution. Comparing the right figures with the left ones, we can easily find that the proposed minimum entropy regularization method **NMT+Entropy** generates much better translations. The attention weight distribution is more sparse and sharp. For instance, Fig. 5(a) ignores the translation *poisoned* after *is* due to uniform attention weight distribution on multiple source words. However, our method **NMT+Entropy** (Fig. 5(b)) successfully generates the correct translation *poisoned* after *had been* by sharpening the weight distribution. To figure out the overall attention weight distribution, we also calculate and provide the entropy per token on the **ALL** test set (last column in Table I). The results show that our entropy method can get lower entropy while significantly boosting the translation quality.

3) Effect of Balance Factor λ: In order to investigate the influence of λ in Eq. 12, we explore different settings from 0.005 to 0.08, and run the experiments on Chinese-English task using the system **NMT+Entropy**.

Table IV lists the comparison results. The numbers clearly show that $\lambda = 0.01$ performs best on all the test sets. $\lambda = \{0.005, 0.02\}$ can reach comparable performance while larger λs ($\lambda = \{0.04, 0.08\}$) much decrease the translation quality. It suggests that λ is best to be set around 0.01.

B. Abstractive Summarization

1) Summarization Quality: Table V reports overall performance of different abstractive summarization methods. To have a thorough comparison, we also list the results obtained with the state-of-the-art extractive summarization method **Lead-3** [7]. The overall statistical results in Table V show that the extractive method **Lead-3** is very strong and the abstractive models can only perform on par with **Lead-3**. It indicates that there is much room for the abstractive methods to improve. Next, we just compare our method to other abstractive ones and investigate the difference within the same paradigm.

We can see from the first two parts that compared to the baseline attention-based sequence model **ABS**, adding the copy mechanism with pointer network (**ABS+Pointer**) can remarkably improve the summarization quality. The third part in Table V describes the effectiveness of our regularized attention model at the basis of the pointer network. Similar to neural machine translation, we also investigate different λs in **minimum entropy** regularizer. The figures in the table demonstrate that our proposed regularized attention model can substantially boost the summarization quality no matter which value is used for λ. Specifically, $\lambda = 0.04$ performs best and it can obtain an improvement of more than 1.0 ROUGE-L score (34.66 vs. 33.42) compared to the reported scores of [7].

The fourth part **ABS+Pointer+Cov** shows that the coverage mechanism is dramatically useful in abstractive summarization task. The last part in Table V reports the results obtained by combining the pointer network, the coverage mechanism and our proposed **minimum entropy** regularizer. We can observe that the summarization quality can be further improved by introducing our regularized attention model (especially when using $\lambda = 0.02$).

2) Some Examples: To better understand the models, we further investigate some specific examples which are listed in Fig. 6 and Fig. 7. In these figures, the original article and its reference summary are shown in the upper half. In the bottom half, we compare our attention regularization model **ABS+Pointer+Cov+Ent** to the state-of-the-art baseline **ABS+Pointer+Cov**. The words in blue indicate that they match the reference summary. Intuitively, more blue words lead to better summaries.

From the two examples, we can see that the system with our sparsity regularization method produces much better summaries which contain more overlapping contents with the references. It suggests that the proposed sparse attention model could focus on the most important information of the input article by sharpening the attention weight distribution.

VIII. RELATED WORK

Attention model becomes a de facto standard for neural machine translation [1], [4], [9], [10] and abstractive summarization [5], [37], [43]–[45]. The standard attention mechanism calculates the contribution of each input word to the generation of an output word. Many approaches are proposed in recently two years to improve the attention model [2], [13], [17], [19]–[21].

[13] and [14] design a coverage model to tackle the problem that the history attention weight distribution is ignored when predicting a target word in machine translation. The attention weights are accumulated in their methods and employed as a feature to encourage the new target word to focus on untranslated source words. [15] and [16] follow the similar idea by utilizing the fertility concept. [7] introduces the coverage mechanism into the abstractive summarization task and better summaries can be produced. These approaches make use of the history attention information, but do not control the attention distribution at the current decoding step.

[2] argues that we should not attend to all the source words at each decoding step. They dynamically compute the position of the central word that should be focused on and then just employ a fixed window as the local context. Instead of using fixed

TABLE V
SUMMARIZATION QUALITY USING DIFFERENT NEURAL SEQUENCE TO SEQUENCE MODELS. WE ALSO LIST THE RESULTS OBTAINED WITH THE STATE-OF-THE-ART EXTRACTIVE SUMMARIZATION METHOD *Lead-3*. REGARDING **ABS+POINTER** AND **ABS+POINTER+ENT**, WE REPORT BOTH SCORES OF OUR REIMPLEMENTATION AND THE ONES FROM [7]. ALL THE ROUGE SCORES ARE REPORTED WITH A 95% CONFIDENCE INTERVAL OF AT MOST ± 0.25, WHICH IS GENERATED BY THE OFFICIAL ROUGE EVALUATION SCRIPT

Method	ROUGE-1	ROUGE-2	ROUGE-L
ABS	31.33	11.81	28.83
ABS+Pointer (reimplementation)	35.83	15.34	32.62
ABS+Pointer ([7])	36.44	15.66	33.42
ABS+Pointer+Ent ($\lambda = 0.005$)	37.15	16.01	33.96
ABS+Pointer+Ent ($\lambda = 0.01$)	37.03	15.95	33.80
ABS+Pointer+Ent ($\lambda = 0.02$)	37.29	16.30	34.13
ABS+Pointer+Ent ($\lambda = 0.04$)	38.00	16.47	34.66
ABS+Pointer+Ent ($\lambda = 0.08$)	37.77	16.46	34.57
ABS+Pointer+Cov (reimplementation)	39.26	17.26	36.15
ABS+Pointer+Cov ([7])	39.53	17.28	36.38
Lead-3 ([7])	40.34	17.70	36.57
ABS+Pointer+Cov+Ent ($\lambda = 0.02$)	**39.96**	**17.49**	**36.71**
ABS+Pointer+Cov+Ent ($\lambda = 0.04$)	39.37	17.21	36.14
ABS+Pointer+Cov+Ent ($\lambda = 0.08$)	39.60	17.27	36.30

Article (truncated): microsoft has a new pilot program to hire people with autism for full-time positions within the company . the announcement was made on the company 's blog by microsoft corporate vp of worldwide operations , mary ellen smith , who has a teenage son , shawn , with autism . the positions will be based at its redmond campus in washington and the program will be run in partnership with specialisterne , a danish nonprofit that helps train people with autism for careers in it . smith wrote on the blog : ' it 's simple , microsoft is stronger when we expand opportunity and we have a diverse workforce that represents our customers . ' the announcement was made on the company 's blog by microsoft corporate vp of worldwide operations , mary ellen smith , who has a teenage son , shawn , with autism . ' the day my son was diagnosed i distinctly remember the final words of two doctors i overheard as my husband and i left the medical center , `` i think they understand . " ' i also remember how we walked to the car , pulled onto the road , drove 15 miles to our home , and entered the house . all in silence . we did not know what to say . ' but we do now . what we learned over the last 15 years was to find our voice . to model what a unique advantage a young man like my son shawn can offer . ' to think of where he is now , a college freshman and part-time employee , and where he has the potential to go , makes my husband and i so proud .

Reference Summary:
microsoft corporate vp of worldwide operations , mary ellen smith , who has a teenage son shawn with autism , made the announcement .
program will be run in partnership with specialisterne , a danish nonprofit that helps train people with autism for careers in it .
full-time positions based at its redmond campus , washington .

ABS+Poiner+Cov:
microsoft has a new pilot program to hire people with autism for full-time positions .
the announcement was made on the company 's blog by microsoft corporate vp of worldwide operations .
mary ellen smith wrote on the blog : ' it 's simple , microsoft is stronger when we expand opportunity and we have a diverse workforce that represents our customers ' .

ABS+Pointer+Cov+Ent:
microsoft has a new pilot program to hire people with autism for full-time positions within the company .
the announcement was made on the company 's blog by microsoft corporate vp of worldwide operations , mary ellen smith , who has a teenage son , shawn , with autism .
the positions will be based at its redmond campus in washington and the program will be run in partnership with specialisterne , a danish nonprofit that helps train people with autism for careers in it .

Fig. 6. Abstractive summarization example about *microsoft announcement*.

Article (truncated): scientists say they have found 50 galaxies that may contain intelligent alien races . these galaxies were found to be emitting ‘ unusually high ’ levels of radiation - possibly indicating ‘ the presence of a highly advanced civilisation . ’ within these galaxies , the researchers said it was possible that an alien race could be harnessing the power of the stars - emitting huge amounts of noticeable heat in the process . pennsylvania scientists have found evidence that we may not be alone . they found 50 galaxies emitting unusually high levels of radiation . this could be because aliens are harnessing the power of stars . however , further research is needed to confirm that is the case . shown is arp 220 , one promising galaxy candidate . the study was led by dr jason wright from the center for exoplanets and habitable worlds at penn state university in pennsylvania . the glimpsing heat from alien technologies survey (g-hat) , published in the astrophysical journal , details how the team thought they might see the infrared signature of a race far more advanced than ours . a proposed method for harnessing the power of an entire star is known as a dyson sphere . first proposed by theoretical physicist freeman dyson in 1960 , this would be a swarm of satellites that surrounds a star . they could be an enclosed shell , or spacecraft spread out to gather its energy - known as a dyson swarm . if such structures do exist , they would emit huge amounts of noticeable infrared radiation back on earth . but as of yet , such a structure has not been detected .

Reference Summary:
pennsylvania scientists find evidence that we might not be alone .
they found 50 galaxies emitting unusually high levels of radiation .
this could be because aliens are harnessing the power of entire stars .
however , further research is needed to confirm that is the case .

ABS+Poiner+Cov:
scientists say they have found 50 galaxies that may contain intelligent alien races .
they found 50 galaxies emitting unusually high levels of radiation .
within these galaxies , the researchers said it was possible that an alien race could be harnessing the power of the stars .

ABS+Pointer+Cov+Ent:
pennsylvania scientists have found 50 galaxies that may contain intelligent alien races .
they found 50 galaxies emitting unusually high levels of radiation .
this could be because aliens are harnessing the power of stars .
however , further research is needed to confirm that is the case .

Fig. 7. Abstractive summarization example about *intelligent galaxies*.

windows, [11] proposes a method which dynamically determines the window size. On one hand, it is difficult for this kind of local attention models to locate the central word. On the other hand, the attended source words of interest may be discontinuous and cannot be covered by a fixed window. In contrast, our regularization method learns a sparse and sharp attention weight distribution and encourages the decoder to focus on the most related source words in an implicitly way.

There are some studies addressing the sparsity and sharpness of the attention weight distribution by modifying the softmax function. [19]–[21] mainly employ a soft-thresholding operator which sets the weights under a threshold to zeros, leading

to sparse distribution. [22], [23] demonstrate that slightly refining the softmax function with low temperature can result in sharper weight distributions. Note that these approaches apply hard constraints. The sparsity and sharpness of the attention weight distribution are not directly optimized according to a well-defined objective function.

IX. Conclusion and Future Work

This paper has presented a novel attention model with sparsity regularization for seq2seq models. Maximum L_∞-norm and minimum entropy regularization methods are proposed to sharpen the attention weight distribution and impels the decoder to focus on the most relevant input words. Our experiments on both machine translation and abstractive summarization tasks demonstrate that the new attention mechanism is highly effective, achieving significant improvements. The results and analysis suggest that our regularization methods provide a flexible and effective way to learn reasonable attention weight distribution. In the future, we will apply our regularization methods to other tasks such as reading comprehension and image caption.

Acknowledgment

The authors would like to thank the reviewers for their valuable comments and suggestions.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Representations*, 2015. [Online]. Available: https://arxiv.org/pdf/1409.0473.pdf

[2] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[3] M. Junczys-Dowmunt, T. Dwojak, and H. Hoang, "Is neural machine translation ready for deployment? A case study on 30 translation directions," in *Proc. 13th Int. Workshop Spoken Lang. Process*. 2016. [Online]. Available: http://workshop2016.iwslt.org/downloads/IWSLT_2016_paper_4.pdf

[4] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, arXiv:1609.08144.

[5] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 379–389.

[6] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, 93–98.

[7] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. Assoc. Comput. Linguistics*, 2017, pp. 1073–1083.

[8] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 2048–2057.

[9] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. Int. Conf. Mach. Learning*, 2017, pp. 1243–1252.

[10] A. Vawani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[11] R. Shu and H. Nakayama, "An empirical study of adequate vision span for attention-based neural machine translation," in *Proc. 1st Workshop Neural Mach. Transl.*, 2017, pp. 1–10.

[12] P. Koehn *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proc. 45th Annu. Meeting ACL Interact. Poster Demonstration Sessions*, 2007, pp. 177–180.

[13] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Modeling coverage in neural machine translation," in *Proc. Assoc. Comput. Linguistics*, 2016, pp. 76–85.

[14] H. Mi, B. Sankaran, Z. Wang, and A. Ittycheriah, "Coverage embedding model for neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 955–960.

[15] T. Cohn, C. D. V. Hoang, E. Vymolova, K. Yao, C. Dyer, and G. Haffari, "Incorporating structural alignment biases into an attentional neural translation model," in *Proc. North Amer. Ch. Assoc. Comput. Linguistics*, 2016, pp. 876–885.

[16] S. Feng, S. Liu, M. Li, M. Zhou, and K. Q. Zhu, "Improving attention modeling with implicit distortion and fertility for machine translation," in *Proc. 26th Int. Conf. Comput. Linguistics*, 2016, pp. 3082–3092.

[17] H. Mi, Z. Wang, and A. Ittycheriah, "Supervised attentions for neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 2283–2288.

[18] L. Liu, M. Utiyama, A. Finch, and E. Sumita, "Neural machine translation with supervised attention," in *Proc. 26th Int. Conf. Comput. Linguistics*, 2016, pp. 3093–3102.

[19] A. Martins and R. Astudillo, "From Softmax to Sparsemax: A sparse model of attention and multi-label classification," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1614–1623.

[20] V. Niculae and M. Blondel, "A regularized framework for sparse and structured neural attention," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 3340–3350.

[21] C. Malaviya, P. Ferreira, and A. F. T. Martins, "Sparse and constrained attention for neural machine translation," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 370–376.

[22] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS 2014 Deep Learn. Workshop*, 2015. [Online]. Available: https://arxiv.org/pdf/1503.02531.pdf

[23] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *Proc. Int. Conf. Learn. Representations*, 2017. [Online]. Available: https://arxiv.org/pdf/1611.01144.pdf

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] K. Sim and R. Hartley, "Removing outliers using the l-infty-norm," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 485–494.

[26] F. Kahl and R. Hartley, "Multiple-view geometry under the l-infty-norm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1603–1617, Sep. 2008.

[27] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[28] M. Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction," *Neural Comput.*, vol. 11, no. 5, pp. 1155–1182, 1999.

[29] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Neural Inf. Process. Syst.*, 2005, pp. 529–536.

[30] L. Vandenberghe, "Subgradient methods," Dept. Elect. Comput. Eng., Univ. California, Los Angeles, CA, USA, Tech. Rep. EE236C, 2013.

[31] S. Shen *et al.*, "Minimum risk training for neural machine translation," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1683–1692.

[32] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, "Deep recurrent models with fast-forward connections for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 4, no. 1, pp. 371–383, 2016.

[33] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1715–1725.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistic*, 2002, pp. 311–318.

[35] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 388–395.

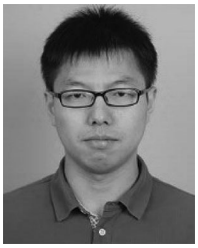
[36] K. M. Hermann *et al.*, "Teaching machines to read and comprehend," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 1693–1701.

[37] R. Nallapati, B. Zhou, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn.*, 2016, pp. 280–290.

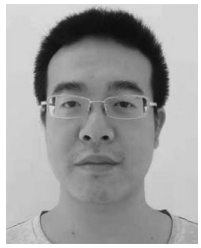
[38] R. Nallapati, F. Zhai, and B. Zhou, "SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents," in *Proc. Assoc. Adv. Artif. Intell.*, 2017, pp. 3075–3081.

[39] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. Jul., pp. 2121–2159, 2011.
[40] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, 2004, pp. 74–81.
[41] Y. Liu and M. Sun, "Contrastive unsupervised word alignment with non-local features." in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2295–2301.
[42] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.
[43] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1631–1640.
[44] P. Nema, M. Khapra, A. Laha, and B. Ravindran, "Diversity driven attention model for query-based abstractive summarization," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1063–1072.
[45] Q. Zhou, N. Yang, F. Wei, and M. Zhou, "Selective encoding for abstractive sentence summarization," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1095–1104.

Jiajun Zhang received the Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences. His research interests include machine translation, multilingual natural language processing, and deep learning.



Yang Zhao received the Bachelor's and Master's degree from Beijing Jiaotong University, Beijing, China. He is currently working toward the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include natural language processing and machine translation.



Haoran Li received the B.S. degree from Shandong University, Jinan, China, in 2013, and is currently working toward the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include natural language processing, summarization, and multimedia.

Chengqing Zong received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 1998. He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include natural language processing, machine translation, and sentiment analysis. He is a Member of the International Committee on Computational Linguistics and the President of Asian Federation of Natural Language Processing. He is an Associate Editor for the *ACM Transactions on Asian and Low-Resource Language Information Processing* and an Editorial Board Member of the IEEE INTELLIGENT SYSTEMS, the journal *Machine Translation*, and the *Journal of Computer Science and Technology*. He served ACL-IJCNLP 2015 as the PC Co-Chair, IJCAI 2017, IJCAI-ECAI 2018, and AAAI 2019 as the Area Chair, and IJCNLP 2017 as the General Chair.