A Text Sentiment Classification Method Enhanced by Bi-GRU and Attention Mechanism

Dongdong Li(✉), Xiaohou Shi, and Meiling Dai

China Telecom Corporation Research Institute Beijing, Beijing, China
{lidd4,daiml1}@chinatelecom.com

Abstract. Text sentiment analysis is a natural language processing technique designed to identify the emotional tendencies expressed in text. In recent years, this field has garnered significant attention and is widely used in practical applications. For example, sentiment analysis is employed for brand reputation management on social media, public opinion monitoring, and risk control in fields such as finance, medicine, and politics. Sentiment analysis is also utilized in tasks such as personalized recommendation and natural language generation. Despite the numerous methods and techniques proposed and applied in text sentiment analysis research, challenges and problems persist. During the sentiment classification process, text data exhibits problems such as uncertainty and semantic diversity, noise, and errors, leading to low accuracy and efficiency of sentiment analysis models. To enhance sentiment analysis accuracy and efficiency, this paper proposes an improved text sentiment classification method based on Bi-GRU and self-attention mechanism. The attention mechanism is initially fused with the update gate of the Bi-GRU gating unit to obtain important feature information in the text content. Subsequently, the Bi-GRU is followed by a self-attention mechanism to perform secondary screening on the text features, and the softmax function is applied to text vectors for sentiment classification, significantly enhancing the accuracy of sentiment classification. The proposed method is tested on the public dataset Yelp Dataset Challenge, and the experimental results indicate a considerable improvement in the accuracy of sentiment classification.

Keywords: Sentiment classification · Attention mechanism · Bi-GRU

1 Introduction

In recent years, significant progress has been made in sentiment analysis research. Deep learning methods have been widely applied in sentiment classification, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms. For instance, Yoon Kim proposed a sentence-level sentiment classification approach based on CNNs [1]. This method transforms the text into a matrix and applies multiple convolution kernels to extract the feature representation. Pengfei Liu introduced a multi-task learning-based RNN method for text and sentiment classification [2], which optimizes multiple related tasks to enhance the model's ability to learn shared data characteristics. Zichao Yang proposed a hierarchical attention mechanism-based method for

Y. Zhang et al. (Eds.): CENet 2023, LNEE 1125, pp. 176–186, 2024.
https://doi.org/10.1007/978-981-99-9239-3_18

text classification [3]. This method utilizes a two-level attention mechanism to learn text representation and better capture semantic information, thus improving sentiment classification performance. Richard Socher et al. proposed a semi-supervised recurrent autoencoder method for sentiment classification [4], which uses a recursive autoencoder to learn structural information and feature representation and semi-supervised learning to utilize unlabeled data. This approach has achieved promising results on sentiment classification tasks. Moreover, Bo Pang and Lillian Lee proposed a sentiment analysis method based on subjectivity summary [5], which separates the text into subjective and objective parts and summarizes the subjective part using the minimum cut algorithm for sentiment analysis. This method provides a significant idea for future sentiment analysis research.

This paper presents a novel Bi-GRU network model integrated with an attention update gate. The proposed model employs attention scores to regulate the update gate, thereby enhancing its performance. Moreover, the model is optimized and combined with a self-attention mechanism to boost the accuracy of sentiment classification. To mine information based on the similarity between words rather than their order, a self-attention mechanism is added after the Bi-GRU model. This approach avoids information loss for longer sentences during sentiment classification, and yields promising classification results.

1.1 Bi-GRU (Bidirectional Gated Recurrent Unit)

The development of Long Short-Term Memory (LSTM) [6] networks has led to the emergence of numerous network variants, including the widely adopted Gated Recurrent Unit (GRU) [7] network. GRUs have demonstrated comparable performance to LSTMs in addressing issues such as vanishing and exploding gradients, as well as capturing long-term dependencies.

Compared to LSTMs, the GRU network utilizes only two gate structures. The first gate combines the forget and input gates from LSTMs into a single update gate, denoted as z_t, which helps maintain a balance between input and forget operations. The second gate, referred to as the reset gate rt, regulates the level of dependence on previous state information, with lower values indicating a reduced level of dependence.

The network structure diagram of GRU is illustrated in Fig. 1.

The calculation process of the recurrent unit in the GRU network can be outlined as follows: at time t, the input vector x_t and the hidden layer state h_{t-1} from the previous time step *t-1* are taken as input. The reset gate and the update gate outputs, z_t and r_t, respectively, are computed using Eqs. (3)–(4). The candidate state, $h_t{}'$, is then updated using Eq. (5), and the hidden layer state, h_t, is updated using Eq. (6).

$$\mathrm{r}_t = \sigma(W_{rx}x_t + W_{rh}h_{t-1} + b_r) \tag{1}$$

$$\mathrm{z}_t = \sigma(W_{zx}x_t + W_{zh}h_{t-1} + b_z) \tag{2}$$

$$h'_t = \tanh(W_{hx}x_t + W_{hr}r_t h_{t-1} + b_h) \tag{3}$$

$$h_t = (1 - z_t)h_{t-1} + \mathrm{z}_t h'_t \tag{4}$$

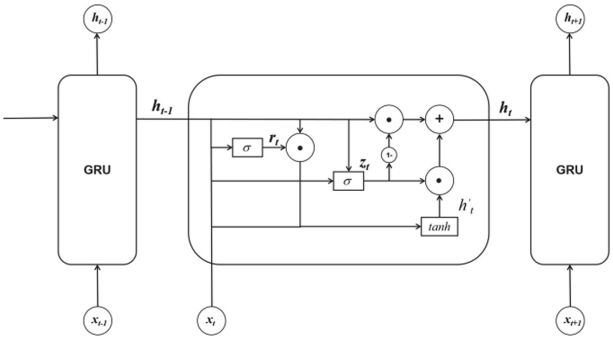

Fig. 1. GRU network structure diagram

In the formula, W_{rx}, W_{rh}, W_{zx}, W_{zh}, W_{hx}, and W_{hr} represent the weight matrix of the update gate, reset gate, and hidden layer, respectively, and b_r, b_z, and b_h are bias vectors.

The Bi-directional Gated Recurrent Unit (Bi-GRU) network comprises two GRU layers—the forward and the reverse layer. The forward propagation GRU calculates the sequence information of the current time step, while the backward propagation GRU reads the same sequence in reverse, introducing the reverse sequence information. The output layer of the network is interconnected with both layers of the GRU, with all neurons in the output layer incorporating both forward and reverse information during the network training process. Figure 3 depicts the specific architecture of the Bi-GRU network (Fig. 2).

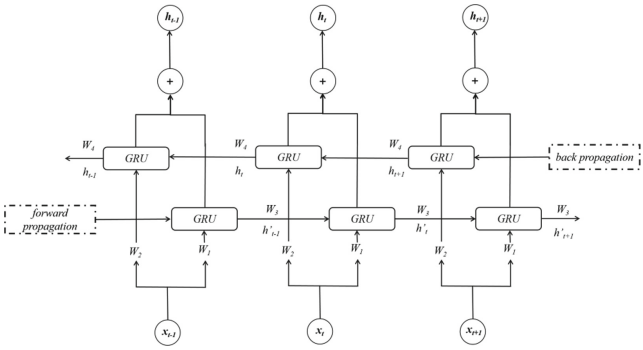

Fig. 2. Bi-GRU network structure diagram

The bidirectional recurrent network, as illustrated in Fig. 3, is constructed by combining two unidirectional recurrent networks. These networks share the same input and operate in opposing directions, with information flow proceeding in both directions.

Additionally, the two networks exhibit structural symmetry, independently performing computations using Eqs. (7) and (8), updating their states, generating outputs, and subsequently connecting the outputs in both directions according to Eq. (9).

$$h'_t = f(W_1 x_t + W_3 h'_t + b'_t) \tag{5}$$

$$h_t = f(W_2 x_t + W_4 h_{t-1} + b_t) \tag{6}$$

$$H_t = h'_t \oplus h_t \tag{7}$$

The formula presented includes the following variables: $h_t^{'}$, h_t, x_t, and H_t. These variables respectively represent the state of the hidden layer for forward propagation, the state of the hidden layer for backpropagation, the input value of the input neuron, and the output value of the hidden layer state at the given moment. Additionally, $h^{'}_{t-1}$ and h_{t+1} represent the state of the forward propagation hidden layer at time $t-1$ and the back propagation hidden layer at time $t+1$, respectively. The activation function of the hidden layer is represented by f, while the vector splicing operation is denoted by the symbol $\oplus$. Furthermore, the variables $b_t^{'}$ and b_t respectively represent the bias vectors of the forward propagation hidden layer and the back propagation hidden layer. Finally, W_1, W_2, W_3, and W_4 correspond to the weight matrix of different components.

2 Bi-GRU Based on Attention Mechanism

In this study, we propose a novel emotion classification model, Bi-GRU', which combines the BiGRU model and the Attention mechanism. The overall architecture of the model is illustrated in the Fig. 3, which can be divided into three main parts: text preprocessing, vectorization, and classifier. The first part, text preprocessing, involves preparing the input text data for further processing, including steps such as tokenization and stemming. In the second part, vectorization, the preprocessed text is transformed into numerical vectors, which can be effectively processed by the model. Finally, the classifier utilizes the Bi-GRU' architecture to classify the emotion expressed in the input text.

In the text preprocessing stage, the first step involves cleaning the text data by removing stop words and line breaks, unifying the case of English letters in the English data set, and serializing the data. Next, the processed data is vectorized using word2vector to convert the text data into a vector. Finally, the word vector is fed into the classifier for processing. In this stage, the BiGRU' model and the forced forward attention mechanism are used to learn the data and extract important features. The BiGRU' model filters the input information through the update gate and the reset gate, and extracts important features from longer input sequences. The attention mechanism is used to weight the key information in the input text, and assign different weight information to the words in the text to learn which words are more important, so that the model can better capture emotional information in classification tasks.

The GRU model uses the update gate to determine the influence degree of the output of the previous hidden layer on the output of the current hidden layer. However, traditional

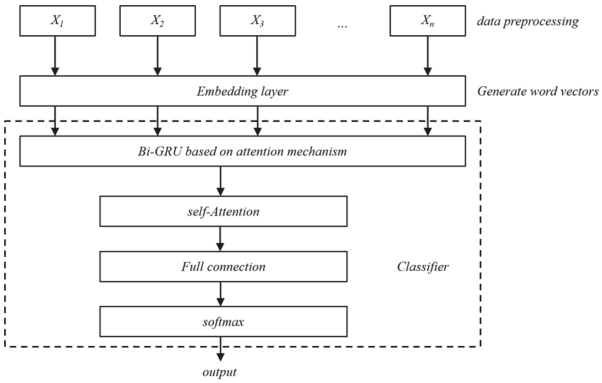

Fig. 3. Sentiment classification model architecture diagram

update gates mainly rely on historical information and newly received information, and may not effectively extract important information in longer input sequences. To address this issue, the attention mechanism is used to selectively focus on relevant input elements and improve the overall performance of the model. This paper proposes a sentiment classification model that combines the attention mechanism with Bi-GRU. To improve the ability of GRU to extract important feature information from text, an attention mechanism is added to the update gate of GRU. The attention score of the GRU update gate is calculated using the following formula:

$$\mathrm{u}_i = \tanh(W_w \mathrm{x}_i + b_w) \tag{8}$$

$$\mathrm{a}_i = soft\max(u_i) \tag{9}$$

In the above formula, W_w and b_w are the weight coefficients and offsets of the feature vectors, x_i is the currently input feature vector, and a_i is the attention score, which acts on the update gate in the GRU structure. Figure 5 is a structure diagram of the improved GRU model based on the attention mechanism (Fig. 4).

The calculation formula is as follows:

$$\mathrm{z}_t = \sigma(w_r \cdot [h_{t-1}, x_t]) \tag{10}$$

$$z_t' = a_t * z_t \tag{11}$$

$$\mathrm{r}_t = \sigma(w_r \cdot [h_{t-1}, x_t]) \tag{12}$$

$$\tilde{h}_t = \tanh(w_{\underset{h}{\sim}} \cdot [r_t * h_{t-1}, x_t]) \tag{13}$$

$$\mathrm{h}_t' = (1 - z_t\prime) * h_{t-1} + z_t * \tilde{h}_t \tag{14}$$

In the above formula, x_t represents the word vector of the t-th word segment, r_t is the reset gate, z_t is the original update gate of GRU, z_t' is the update gate with the

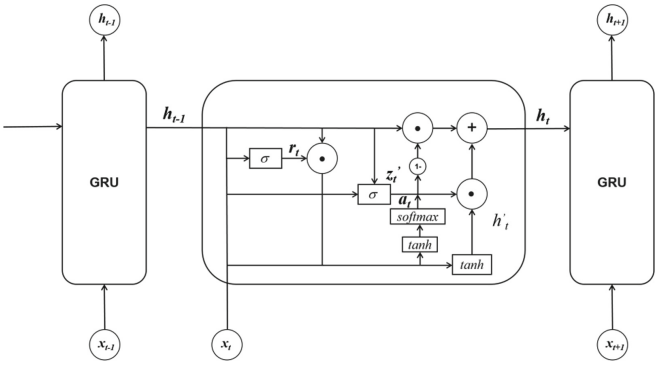

Fig. 4. Improved GRU structure diagram

attention mechanism added, h_t' and h_{t-1} are the hidden GRU state, a_t is the attention score in Formula (11), and σ is the sigmoid activation function. The enhanced update gate not only relies on the historical information of the previous moment and the newly received input information, but also on the attention score of the current information. The attention score reflects the importance of information and its impact on the current state. Information with higher attention scores are assigned larger weights, resulting in higher values for the update gate and are retained for further processing. Conversely, information with lower attention scores are assigned smaller weights, resulting in smaller values for the update gate and thus discarded. This mechanism improves the ability of the GRU model to extract essential information in the text and enhances its feature extraction ability (Fig. 5).

In this paper, the BiGRU' model is obtained by bidirectionalizing the GRU' model with the attention mechanism added to the update gate. BiGRU' is similar to the BiGRU model structure, as shown in Fig. 6.

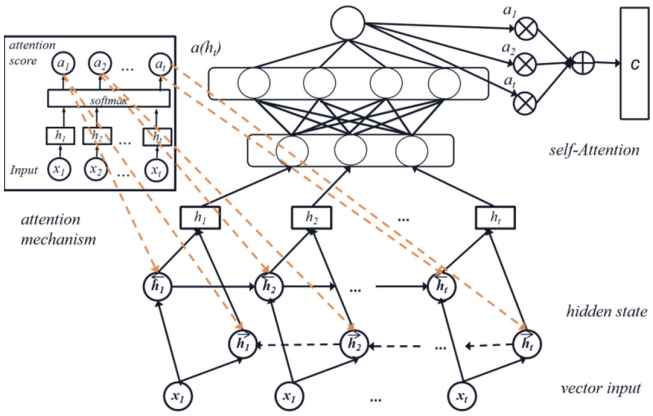

Fig. 5. BiGRU' network structure diagram

Given the word vector is x_t, t ∈ [1, L], L is the length of the text, W_e is the weight matrix of BiGRU', then the word vectorization of the text is expressed as:

$$x_t = W_e w_t, t \in [1, L] \tag{15}$$

The calculation formula of BiGRU' is:

$$\overrightarrow{h_t} = AGRU(x_t), t \in [1, L] \tag{16}$$

$$\overleftarrow{h_t} = AGRU(x_t), t \in [1, L] \tag{17}$$

In this method, $\overrightarrow{h}$ represents the hidden state of the word during forward propagation, and $\overleftarrow{h}$ represents the hidden state of the word during backpropagation. Concatenating $\overrightarrow{h}$ and $\overleftarrow{h}$, $h_i = [\overrightarrow{h}, \overleftarrow{h}]$, can obtain the bidirectional semantic information of the word vector.

This study also uses the self-attention mechanism after improving Bi-GRU to further integrate the important feature information of the text. The self-attention mechanism learns the hidden state weight at each moment t and extracts the feature information of the text by calculating the similarity between words. This mechanism does not depend on the order of words and retains important feature information. The specific calculation formula of self-attention is as follows:

$$\mathrm{e}_t = u_{att}^T \tanh(W_{att} h_t + b_{att}) \tag{18}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{n} \exp(e_k)} \tag{19}$$

$$c = \sum_{t=1}^{n} \alpha_t h_t \tag{20}$$

In the formula, ${u^T}_{att}$, W_{att}, and b_{att} are related parameter matrices of self-attention. h_t represents the hidden state at time t, and α_t represents the attention weight of the state hidden state at time t. The final weighted vector representation c of the text can be obtained by weighting and summing the hidden state ht through Formula (22). Finally, c is passed through the softmax function to obtain the sentiment classification result.

3 Experiment and Analysis

3.1 Data Preprocessing

This paper employs the Twitter Tweet Comments Dataset to train and evaluate sentiment analysis algorithms. The Twitter Sentiment Analysis Dataset comprises tweets from the highly representative social media platform Twitter, which provides a rich source of freely expressed emotions and thoughts. The dataset consists of over 15,000 tweets

labeled with positive, negative, or neutral sentiment. The first step in the pre-processing of the dataset involves the removal or replacement of irrelevant information, noise, and unnecessary characters in the raw data with appropriate symbols. For instance, URLs, punctuation marks, numbers, and special characters in tweets are removed. Word2vector is then utilized to initialize the word embedding information of the comment text. Subsequently, the dataset is randomly partitioned into a training set and a test set in an 8: 2 ratio, which are employed for model training and performance evaluation, respectively. The training set comprises 12,000 instances, and the test set comprises 3000 instances.

3.2 Evaluation Index

This paper employs accuracy, recall, and F_1 as evaluation metrics, with the respective calculation formulas presented below:

$$accuracy = \frac{TP + TF}{TP + TF + FP + FN} \tag{21}$$

$$recall = \frac{TP}{TP + FN} \tag{22}$$

$$precision = \frac{TP}{TP + FP} \tag{23}$$

$$\mathrm{F1} = \frac{2 * recall * precision}{precison + recall} \tag{24}$$

Among these metrics, TP represents the number of samples that were correctly classified as positive samples, while TF represents the number of samples that were correctly classified as negative samples. In contrast, FP represents the number of samples that were actually negative but were misclassified, and FN represents the number of samples that were actually positive but were misclassified.

In this experiment, we compared the performance of the BiAGRU' model, which incorporates an attention mechanism, with the following four models: the word2vector-GRU model, which combines word2vector word vectors with a GRU text classifier, the Bi-GRU model, the Bi-LSTM model, and the GRU-Attention model. Notably, the GRU-Attention model only integrates the self-attention mechanism discussed in Chap. 2 after the Bi-GRU layer and does not apply attention to the update gate of the GRU layer.

The parameter configurations for the experimental model Bi-GRU' are presented in Table 1

3.3 Analysis of the Experimental Results of the Data Set

Table 2 presents the analysis outcomes of each model on the Twitter dataset, while the classification results for each category are displayed in Fig. 7.

As depicted in Fig. 7, the proposed implementation model exhibits superior performance in terms of accuracy, recall, and F1 score when compared with other models. Additionally, the analysis of Table 2 and Fig. 7 highlights that both BiGRU and BiLSTM models outperform the word2vector-GRU model, indicating that the bidirectional

Table 1. Model parameter setting table.

Parameter name	Parameter value
Word vector dimension	100
Learning rate	0.02
Loss function	Cross-entropy loss
Batch size	150
Dropout	0.1
Optimizer	SGD

Table 2. Sentiment analysis results of each model.

Models	Accuracy	Recall	F1
word2vector-GRU	0.846	0.839	0.841
Bi-GRU	0.873	0.871	0.874
Bi-LSTM	0.868	0.867	0.869
Bi-GRU-attention	0.883	0.884	0.886
Bi-GRU'	0.902	0.901	0.905

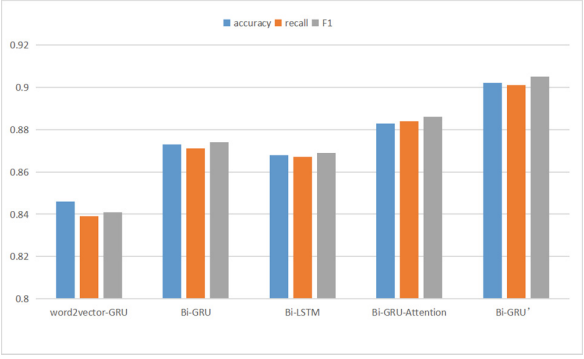

Fig. 6. Comparison of various model results

model performs better than the unidirectional model. Furthermore, the Bi-GRU' model, which utilizes a double-layer attention mechanism, surpasses the Bi-GRU-Attention model, despite both models using BiGRU and attention to extract text information. The difference in performance arises from the number of layers of attention, which suggests the effectiveness of the attention mechanism added to the update gate in Bi-GRU'. Additionally, while the performance of Bi-GRU and Bi-LSTM models is similar, the

double-layer attention mechanism incorporated in the Bi-GRU' model yields significantly better results than both Bi-GRU and Bi-LSTM models. The findings demonstrate that the proposed model, which builds on BiGRU, achieves outstanding performance in all aspects.

To further confirm the model's effectiveness, this study conducted additional experiments on the Amazon product review dataset, which contains millions of product reviews and ratings across different categories (e.g., books, electronics, household items, etc.). A test set of 76,537 items was randomly selected, and models including word2vector-GRU, Bi-GRU, Bi-LSTM, Bi-GRU-Attention, and Bi-GRU' were tested on the Amazon product review text in the test set. The experimental results are presented in Fig. 7.

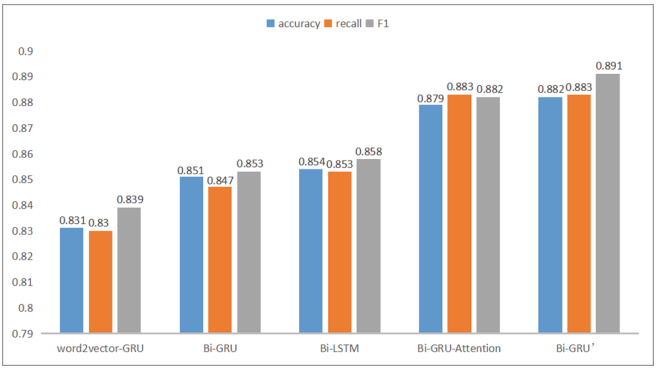

Fig. 7. Comparison of various model results

The results depicted in Fig. 8 show that the models with the added attention layer outperform the models without it in various performance indicators, confirming the effectiveness of the attention layer. Although the BiGRU-Attention model has a slightly higher recall rate and F1 value, its accuracy rate is lower. Overall, the performance of the Bi-GRU' model is better. The experimental results indicate that the Bi-GRU' model proposed in this paper, with the addition of a two-layer attention mechanism, can better capture the context and has a superior overall performance in sentiment analysis.

4 Conclusion

Traditional sentiment analysis models often overlook the context and the influence of crucial words on sentiment analysis. Most models rely on stacked neural network models and attention mechanisms. To overcome these limitations, this paper proposes a BiGRU network model with an attention update gate that utilizes the attention score to regulate the update gate. The model is optimized and combined with a forward attention mechanism to enhance its accuracy. Experimental results demonstrate the efficacy of the proposed model. In future studies, we aim to explore the integration of different attention mechanisms and GRU, optimize the loss function, and evaluate the model's effectiveness in various domains.

References

1. Yoon, K.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)
2. Liu, P., Qiu, X., Huang, X.: Recurrent neural network for text classification with multi-task learning. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI) (2016)
3. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (2016)
4. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., YNg, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2013)
5. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL) (2004)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8) (1997)
7. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)