Data-efficient Neural Text Compression with Interactive Learning

Avinesh P.V.S and **Christian M. Meyer**
Research Training Group AIPHES and UKP Lab
Computer Science Department, Technische Universität Darmstadt
www.aiphes.tu-darmstadt.de, www.ukp.tu-darmstadt.de

Abstract

Neural sequence-to-sequence models have been successfully applied to text compression. However, these models were trained on huge automatically induced parallel corpora, which are only available for a few domains and tasks. In this paper, we propose a novel interactive setup to neural text compression that enables transferring a model to new domains and compression tasks with minimal human supervision. This is achieved by employing active learning, which intelligently samples from a large pool of unlabeled data. Using this setup, we can successfully adapt a model trained on small data of 40k samples for a headline generation task to a general text compression dataset at an acceptable compression quality with just 500 sampled instances annotated by a human.

1 Introduction

Text compression is the task of condensing one or multiple sentences into a shorter text of a given length preserving the most important information. In natural language generation applications, such as summarization, text compression is a major step to condense the extracted important content of the source documents. But text compression can also be applied in a wide range of related applications, including the generation of headlines (Filippova et al., 2015), captions (Wubben et al., 2016), subtitles (Vandegehinste and Pan, 2004; Luotolahti and Ginter, 2015), and the compression of text for small screens (Corston-Oliver, 2001).

Neural *sequence-to-sequence* (Seq2Seq) models have shown remarkable success in many areas of natural language processing and specifically in natural language generation tasks, including text compression (Rush et al., 2015; Filippova et al., 2015; Yu et al., 2018; Kamigaito et al., 2018). Despite their success, Seq2Seq models have a major drawback, as they require huge parallel corpora with pairs of source and compressed text to be able to learn the parameters for the model. So far, the size of the training data has been proportional to the increase in the model's performance (Koehn et al., 2003; Suresh, 2010), which is a major hurdle if only limited annotation capacities are available to manually produce a corpus. That is why existing research employs large-scale automatically extracted compression pairs, such as the first sentence and the presumably shorter headline of a news article. However, such easy-to-extract source data is only available for a few tasks, domains, and genres and the corresponding models do not generalize well from the task of headline generation to other text compression tasks.

In this paper, we propose an *interactive setup* to neural text compression, which learns to compress based on user feedback acquired during training time. For the first time, we apply *active learning* (AL) methods to neural text compression, which greatly reduces the amount of the required training data and thus yields a much more data-efficient training and annotation workflow. In our experiments, we find that this approach enables the successful transfer of a model trained on headline generation data to a general text compression task with a minimum of parallel training instances.

The objective of AL is to efficiently select unlabeled instances that a user should annotate to advance the training. A key component of AL is the choice of the *sampling strategy*, which curates the samples in order to maximize the model's performance with a minimum amount of user interaction. Many AL sampling strategies have proven effective for human-supervised natural language processing tasks other than compression (Hahn et al., 2012; Peris and Casacuberta, 2018; Liu et al., 2018).

In our work, we exploit the application of uncertainty-based sampling using attention disper-

Proceedings of NAACL-HLT 2019, pages 2543–2554
Minneapolis, Minnesota, June 2 - June 7, 2019.

sion and structural similarity for choosing samples to be annotated for our interactive Seq2Seq text compression model. We employ the AL strategies for (a) learning a model with a minimum data, and (b) adapting a pretrained model with few user inputs to a new domain.

In the remaining paper, we first discuss related work and introduce the state-of-the-art Seq2Seq architecture for the neural text compression task. Then, we propose our novel interactive compression approach and demonstrate how batch mode AL can be integrated with neural Seq2Seq models for text compression. In section 4, we introduce our experimental setup, and in section 5, we evaluate our AL strategies and show that our approach successfully enables (a) learning the Seq2Seq model with a minimum of data, (b) transfer of a pretrained headline generation model to a new compression task and dataset with minimal user interaction. To encourage further research and enable reproducing our results, we publish our code as open-source software.[1]

2 Related Work

In this section, we discuss related work to our research concerning: (1) neural text compression models, (2) existing text compression corpora and (3) active learning for neural models.

Neural text compression. Neural text compression models can be broadly classified into two categories: (a) deletion-based extractive models and (b) abstractive models. The goal of the deletion-based models is to delete unimportant words from a source text to generate a shorter version of the text. In contrast, abstractive models generate a shorter text by inserting, reordering, reformulating, or deleting words of the source text.

Previously, deletion-based extractive methods explored various modeling approaches, including the noisy-channel model (Knight and Marcu, 2002; Turner and Charniak, 2005), integer linear programming (Clarke and Lapata, 2007), variational autoencoders (Miao and Blunsom, 2016), and Seq2Seq models (Filippova et al., 2015). Similarly, recent abstractive models have seen tree-to-tree transduction models (Cohn and Lapata, 2013) and variations of Seq2Seq models, such as attention (Rush et al., 2015), attentive long short-term memory (LSTM) models (Wubben et al., 2016) and operation networks where the Seq2Seq model decoder is replaced with a deletion decoder and a copy-generate decoder (Yu et al., 2018).

Filippova et al. (2015) show that Seq2Seq models without any linguistic features have the ability to delete unimportant information. Kamigaito et al. (2018) incorporate higher-order dependency features into a Seq2Seq model and report promising results. Rush et al. (2015) propose an attention-based Seq2Seq model for generating headlines. Chopra et al. (2016) further improve this task with recurrent neural networks. Although Seq2Seq models show state-of-the-art results on different compression datasets, there is yet no work which investigates whether large training corpora are needed to train neural compression models and if there are efficient ways to train and adapt them to other datasets with few annotations.

Text compression corpora. Early publicly available text compression datasets are manually curated but small (Cohn and Lapata, 2008; Clarke and Lapata, 2006, 2008). These datasets are typically used by unsupervised approaches as they are 200 times smaller in size compared to the annotated data used for training state-of-the-art supervised approaches. Filippova and Altun (2013) introduce an extractive compression dataset of 250k headline and first sentence compression pairs based on Google News, which they use for training a supervised compression method. Similarly, Rush et al. (2015) create another large abstractive dataset of 4 million headline and first sentence compression pairs from news articles extracted from the Annotated Gigaword corpus (Napoles et al., 2012). Although these datasets are large, they predominantly address headline generation for news.

Creating such large corpora manually for a new task or domain is hard. Toutanova et al. (2016) pioneered the manual creation of a multi-reference compression dataset MSR-OANC with 6k sentence–short paragraph pairs from business letters, newswire, journals, and technical documents sampled from the Open American National Corpus[2]. They provide five crowd-sourced rewrites for a fixed compression ratio and also acquire quality judgments. This dataset covers multiple genres compared to the large automatically collected compression datasets, and Toutanova et al. (2016) show that neural Seq2Seq

[1] https://github.com/UKPLab/NAACL2019-interactiveCompression

[2] https://www.anc.org/data/oanc

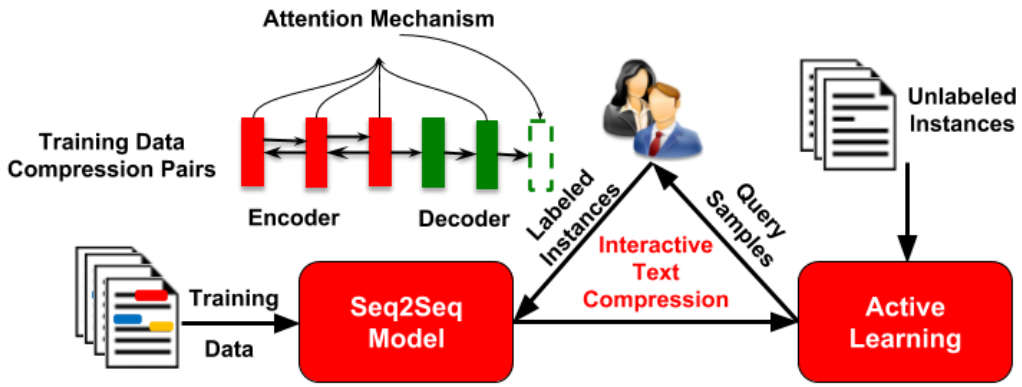

Figure 1: Pipeline of our interactive text compression model. The pipeline is divided into three main components: (1) Neural Seq2Seq text compression model, (2) active learning, and (3) interactive text compression

models trained on headline generation datasets fail to achieve state-of-the-art results as compared to an ILP-based unsupervised method. In our work, we go beyond that and investigate strategies to easily adapt pretrained models to such small datasets employing minimal user input.

Active learning for neural models. AL has been successfully applied to various natural language processing tasks, including corpus annotation (Hahn et al., 2012; Yan et al., 2011), domain adaptation (Chan and Ng, 2007), personalized summarization (P. V. S. and Meyer, 2017), machine translation (Haffari and Sarkar, 2009), language generation (Mairesse et al., 2010), and many more. Only recently, it has been applied to neural models: Wang et al. (2017a) propose an AL approach for a black box semantic role labelling (SRL) model where the AL framework is an add-on to the neural SRL models. Peris and Casacuberta (2018) use AL in neural machine translation. They propose quality estimation sampling, coverage sampling, and attention distraction sampling strategies to query data for interactive machine translation. Liu et al. (2018) additionally propose an AL simulation trained on a high-resource language pair to transfer their model to low-resource language pairs. In another line of research, Sener and Savarese (2018) discuss a core-set AL approach as a batch sampling method for neural image classification based on convolutional neural networks. Although AL techniques have been widely used in natural language processing, to our knowledge, there is yet no work on the use of AL for neural text compression. We fill this gap by putting the human in the loop to learn effectively from a minimal amount of interactive feedback and for the first time, we explore this data-efficient AL-based approach to adapt a model to a new compression dataset.

3 Approach

To address this research problem, we first describe the neural Seq2Seq text compression models we use. Then, we introduce our active learning strategies to select the training samples interactively for in-domain training as well as for domain adaptation, and we describe a novel interactive neural text compression setup. Figure 1 illustrates the main components of our system.

3.1 Neural Seq2Seq Text Compression

In this work, we employ state-of-the-art Seq2Seq models with attention (Seq2Seq-gen) (Rush et al., 2015) and pointer-generated networks with coverage (Pointer-gen) (See et al., 2017) as our base models, which we use for our AL-based interactive text compression setup.

Both Seq2Seq models are built upon the encoder-decoder framework by Sutskever et al. (2014). The encoder encodes the input sequence $\mathrm{x} = (x_1, x_2.., x_n)$ represented by an embedding matrix into a continuous space using a bidirectional LSTM network and outputs a sequence of hidden states. The decoder is a conditional bidirectional LSTM network with attention distribution (Luong et al., 2015)

$$a_i^j = \frac{\exp(e_i^j)}{\sum_{k=1}^{n} \exp(e_k^j)} \quad (1)$$

where e_i^j is computed at each generation step j with the encoder states h_i^{enc} and the decoder states h_j^{dec}:

$$e_i^j = q \cdot \tanh(W_h^{\text{enc}} h_i^{\text{enc}} + W_h^{\text{dec}} h_j^{\text{dec}} + b_{\text{att}}) \quad (2)$$

where q , W_h^{enc}, W_h^{dec} and b_{att} are learnable parameters. The attention distribution a_i^j is used to compute the weighted sum of the encoder hidden states, also known as the context vector

$$c_j^* = \sum_i^n a_i^j h_i^{\text{enc}} \tag{3}$$

To obtain the vocabulary distribution P_j^{vocab} at generation step j, we concatenate the fixed context vector with the decoder state h_j^{dec} and pass it through two linear layers:

$$P_j^{\text{vocab}} = \text{softmax}(W_v(W_v'[h_j^{\text{dec}}; c_j^*] + b_v') + b_v) \tag{4}$$

where W_v, W_v', b_v and b_v' are learnable parameters. P_j^{vocab} is a probability distribution over all words in the vocabulary V. Based on the vocabulary distribution, the model generates the target sequence $\text{y} = y_1, y_2, \ldots, y_m$, $m \leq n$ with

$$y_j = \text{argmax}_w P_j^{\text{vocab}}(w),\ w \in V \tag{5}$$

for each generation step j.

Finally during training, we define the loss function for generation step j as the negative log likelihood of the target word y_j and the overall loss function for the target word sequence as $\mathcal{L}$:

$$\mathcal{L} = \frac{1}{m} \sum_{j=0}^{m} -\log P_j^{\text{vocab}}(y_j) \tag{6}$$

Another state-of-the-art approach we use for our experiments is the pointer-generator networks (Pointer-gen) proposed by See et al. (2017). This model uses a pointer-generator network that determines a probability function to generate the words from the vocabulary V or copy the words from the source text by sampling from the attention distribution a_i^j as shown in Eq. 8. The model achieves this by calculating an additional generation probability p_{gen} for generation step j, which is calculated from the context vector c_j^*, the decoder state h_j^{dec}, and the current input to the decoder $x_j^{'}$:

$$p_{\text{gen}} = \sigma(W_c^T c_j^* + W_{h^{\text{dec}}}^T h_j^{\text{dec}} + W_{x'}^T x_j^{'} + b_{\text{gen}}) \tag{7}$$

$$P_j(w) = p_{\text{gen}} P_j^{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i=0}^{n} a_i^j \tag{8}$$

where vectors W_c, $W_{h^{\text{dec}}}$, $W_{x'}$, b_{gen} are learnable parameters, n is the number of words in the source text and σ is the sigmoid function.

The model also uses an extra feature of coverage to keep track of words generated by the model to discourage repetition. In the coverage model, a coverage vector is calculated which is the sum of the attention distribution across all the previous decoding steps and it is passed on as an extra input to the attention mechanism:

$$c_i^j = \sum_{k=0}^{j-1} a_i^k \tag{9}$$

$$e_i^j = q \cdot \tanh(W_h^{\text{enc}} h_i^{\text{enc}} + W_h^{\text{dec}} h_j^{\text{dec}} + W_c c_i^j + b_{\text{att}}) \tag{10}$$

where W_c is an additional learnable parameter.

3.2 Active Learning

Toutanova et al. (2016) show that Seq2Seq models, which perform well on large news headline generation datasets, fail to achieve good performance on their MSR-OANC multi-genre compression dataset. A major issue with training Seq2Seq models is the lack of domain-specific data and the expensive process to create parallel compression pairs. It is therefore indispensable to minimize the cost of data annotation. Thus, AL comes into play whose key element is to find a strategy for selecting samples the user should annotate which yield a more efficient training process. For text compression, we suggest AL strategies to maximize the model's coverage and the diversity of the samples. To this end, we build upon work in uncertainty sampling by (Peris and Casacuberta, 2018; Wang et al., 2017b) and propose a new strategy to predict the sample diversity at a structural level.

Coverage constraint sampling (Coverage-AL). An important factor on which text compression models are evaluated is the coverage (Marsi et al., 2010). Coverage can be defined as the text compression models being able to learn the deletion or generation rules from the training samples and apply them on an input source text. Wu et al. (2016) first proposed the idea of using attention weights to calculate coverage penalty for active learning based machine translation systems. The attention weights were further extended by Peris and Casacuberta (2018) to estimate an attention

dispersion based uncertainty score for a sentence. The idea of attention dispersion is that if the neural Seq2Seq compression model is uncertain then the attention weights will be dispersed across the source text while generating the target words. The samples with higher dispersion will have their attention weights uniformly distributed across the source sentences. Thus, the goal is to find the samples with high uncertainty based on attention dispersion. As we want to define to the extent to which the attention distribution differs from a normal distribution we propose to use a skewness score. The skewness score calculates the attention dispersion while decoding a target word y_j.

$$\text{skewness}(y_j) = \frac{\frac{1}{n}\sum_{i=1}^{n}(a_i^j - \frac{1}{n})^3}{(\frac{1}{n}\sum_{i=1}^{n}(a_i^j - \frac{1}{n})^2)^{3/2}} \quad (11)$$

a_i^j is the attention weight assigned by the attention layer to the i-th source word when decoding the j-th target word and $\frac{1}{n}$ is the mean of the attention weights of the target word y_j.

The skewness for a normal distribution is zero, and since we are interested in the skewness of samples with heavy tails, we take the negative of the skewness averaged across all target words to obtain the uncertainty coverage score C_{score}.

$$C_{\text{score}}(\text{x}, \text{y}) = \frac{\sum_{j=1}^{m} -\text{skewness}(y_j)}{m} \quad (12)$$

where m is the number of target words.

Diversity constraint sampling (Diversity-AL). Diversity sampling methods have been used in information retrieval (Xu et al., 2007) and image classification (Wang et al., 2017b). The core idea is that samples that are highly similar to each other typically yield little new information and thus low performance. Similarly, to increase the diversity of the samples in neural text compression, we propose a novel scoring metric to measure the diversity of multiple source texts at a structural level. Our intuition is that integrating part-of-speech, dependency and named entity information is useful for text compression, e.g., to learn which named entities are important and how to compress a wide range of phrase types and syntactically complex sentences. Thus, we consider part of speech tags, dependency trees, and named entity embeddings and calculate the structural similarity of the source text with regard to the target text. We use a multi-task convolutional neural network similar to Søgaard and Goldberg (2016) trained on OntoNotes and Common Crawl to learn the structural embeddings consisting of tag, dependency and named entity embeddings. The diversity score D_{score} is calculated using the cosine distance between the average of the structural embeddings of the words in the source sentence and the average of the structural embeddings of the words in the target compression as in Eq. 13:

$$D_{\text{score}}(\text{x}, \text{y}) = \frac{\text{E}_{\text{struc}}(\text{x}) \cdot \text{E}_{\text{struc}}(\text{y})}{||\text{E}_{\text{struc}}(\text{x})|| \cdot ||\text{E}_{\text{struc}}(\text{y})||} \quad (13)$$

where $\text{E}_{\text{struc}}(\cdot)$ is the average structural embedding of a text.

These AL sampling strategies are applied interactively while training to make better use of the data by selecting the most uncertain instances. Additionally, both strategies can be applied for domain adaptation by actively querying user annotations for a domain-specific dataset in an interactive text compression setup, which we describe next.

3.3 Interactive Text Compression

In this subsection, we introduce our interactive text compression setup. Our goal is to select the batch of samples for training efficiently with minimal samples and to become able to transfer the models to new datasets for different domains and genres with few labeled data.

We consider an initial collection of parallel instances $D = \{(\text{x}_i, \text{y}_i) \mid 1 \leq i \leq N\}$ consisting of pairs of input text x_i and their corresponding compression y_i. Additionally, we consider unlabeled instances $D' = \{\text{x}_i \mid i > N\}$, for which we only know the uncompressed source texts. Our goal is to sample sets of unlabeled instances $S_t \subset D'$ which should be annotated by a user in each time step t. The interactive compression model can only see the labeled pairs from the initial dataset D in the beginning, but then incrementally learns from the user annotations.

Algorithm 1 provides an overview of our interactive compression setup. The inputs are the labeled compression pairs D and the unlabeled source texts D'. D is used to initially train the neural text compression model M. In line 5, we start the interactive feedback loop iterating over $t = 0, \ldots, T$. We first sample a set of unlabeled source texts S_t (line 6) by using our AL strategies introduced in section 3.2 and then loop over each of the unlabeled samples to be annotated

or supervised by the human in line 10. As the user feedback in the current time step of sample S_t, we obtain the compressions Y_t of the sampled source texts S_t from the user and use them for online training of the model M. After T iterations or if there are no samples left for querying (i.e., $S_t = \emptyset$), we stop the iteration and return the updated Seq2Seq model M.

Algorithm 1 Interactive Text Compression

```
1:  procedure InteractiveCompression()
2:      input: Text Compression Pairs D,
3:             Unlabeled Text D'
4:      M ← learnSeq2Seq(D)
5:      for t = 0, ..., T do
6:          S_t ← getSample(D')
7:          if S_t = ∅ then
8:              return M
9:          else
10:             Y_t ← queryUser(S_t)
11:             M ← update(M, S_t, Y_t)
12:             D' ← D' − S_t
13:         end if
14:     end for
15: end procedure
```

4 Experimental Setup

4.1 Data

For our experiments, we use the large Google News text compression corpus[3] by Filippova and Altun (2013), which contains 250k automatically extracted the deletion-based compressions from aligned headlines and first sentences of news articles. Recent studies on text compression have extensively used this dataset (e.g., Zhao et al., 2018; Kamigaito et al., 2018). We carry out in-domain active learning experiments on the Google News compression corpus.

To evaluate our interactive setup, we adapt the trained models to the MSR-OANC text compression corpus by Toutanova et al. (2016), which contains 6k crowdsourced multi-genre compressions from the Open American National Corpus. This corpus is well-suited to evaluate our interactive setup, since it is sourced from mixture of newswire, letters, journals, and non-fiction genres, in contrast to the Google News corpus covering only newswire.

Dataset	# Train	# Dev	# Test
Google News	195,000	5,000	10,000
MSR-OANC	5,000	448	785

Table 1: Statistics of the compression datasets

For evaluating the compressions against the reference compressions, we use a Python wrapper[4] of the ROUGE metric (Lin, 2004) with the parameters suggested by Owczarzak et al. (2012) yielding high correlation with human judgments (i.e., with stemming and without stopword removal).[5]

4.2 Preprocessing and Parameters

To preprocess the datasets, we perform tokenization. We obtain the structural embeddings for a sentence using spaCy[6] embeddings learned using a multi-task convolutional neural network.

To evaluate and assess the effectiveness of our active learning-based sampling approaches, we set up our interactive text compression approach for the two state-of-the-art Seq2Seq models consisting of a generative model (Seq2Seq-gen) and a generate-and-copy model (Pointer-gen) as described in Section 3.1. For the neural Seq2Seq text compression experiments, we set the beam size and batch size to 10 and 30 respectively. We use the Adam optimizer (Kingma and Ba, 2015) for the gradient-based optimization. Finally, the parameters for the neural network parameters like weights and biases are randomly initialized.

In order to asses the effectiveness of AL for neural text compression we extend the OpenNMT[7] implementations with our interactive framework following Algorithm 1. The sampling strategy selects instances to be annotated interactively by the user in batches. Next, the neural text compression model is incrementally updated with the selected samples.

Due to the presence of a human in the loop, it typically demands real user feedback, but the cost of collecting sufficient data for various settings of our models is prohibitive. Thus in our experiments, the users were simulated by using the com-

[3] https://github.com/google-research-datasets/sentence-compression

[4] https://github.com/pltrdy/files2rouge

[5] -n 2 -c 95 -r 1000 -a -m

[6] https://spacy.io/

[7] https://github.com/OpenNMT/OpenNMT-py

Methods	UB			Random			Coverage-AL			Diversity-AL		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
Seq2Seq-gen	59.94	52.08	59.78	61.60	50.03	61.37	**62.89**	**51.38**	**62.56**	62.54	50.19	62.13
Pointer-gen	79.26	71.77	79.08	71.61	61.15	71.28	**78.11**	**70.50**	**77.89**	77.45	70.30	77.38

Table 2: ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-L (RL) achieved by the state-of-the-art models using our sampling strategies evaluated on the Google compression test set. Bold marks best AL strategy.

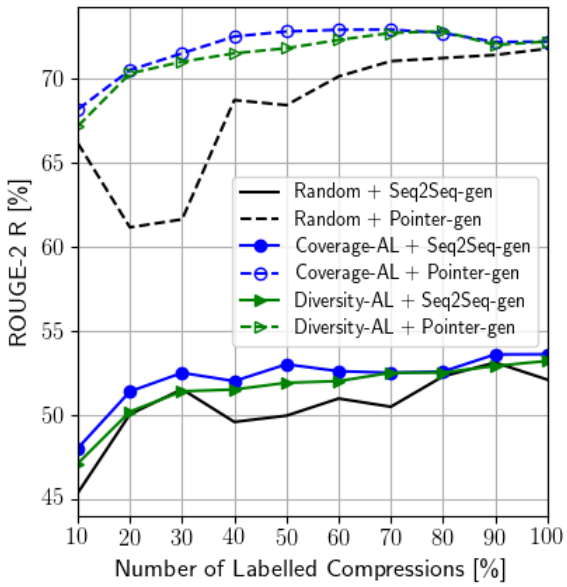

Figure 2: Analysis of the active learning approaches combined with state-of-the-art Seq2Seq compression models on Google compression dataset while varying the training sizes.

pression pairs from our corpus as the sentences annotated by the user.

5 Results and Analysis

Our experiments address two main research questions for in-domain training and domain adaptation of neural text compression:

- Which active learning strategies are useful in text compression to select training samples such that higher performance can be achieved with a minimum of labeled instances?
- Which instances are to be annotated interactively by the user such that the model adapts quickly to a new dataset?

In-domain Active Learning. For in-domain active learning experiments, we choose the Google News text compression training corpus and sample for corpus sizes between 10% and 100% in ten percent point steps. As a baseline, we use a random sampling strategy to test the state-of-the-art Seq2Seq neural text compression models. Figure 2 suggests that our coverage-based sampling (Coverage-AL) and diversity-based sampling (Diversity-AL) strategies outperform the random sampling strategy throughout all training sizes. A key observation is that our sampling strategies are behind the upper bound by just 0.5% ROUGE-2 when only 20% of the training data is used. Table 2 illustrates the results of our sampling strategies when 20% of the data is used for training. All the results are in comparison to the upper bound (UB) receiving 100% of the training data.

Coverage-AL performs better than the Diversity-AL for both the Seq2Seq-gen and Pointer-gen models. However, they are still not effective in the Seq2Seq-gen model where random sampling performs on par with the active learning sampling approaches. We believe this is due to the Seq2Seq-gen model's inability to copy from the source text in the sampled set as a consequence of active learning in the batch setting. Whereas for Pointer-gen model, we observed that both Coverage-AL and Diversity-AL strategies of adding new samples for training had a greater impact when the model has not adapted. We attribute the effectiveness of the Coverage-AL strategy over Diversity-AL to the exploitation of the model uncertainty, as the Diversity-AL only uses the similarity based on the samples, but misses to integrate the model uncertainty.

Table 3 presents an example sentence compression pair from the Google News dataset and the generated compressions of both neural Seq2Seq models when using one of the three sampling strategies. The example shows that detailed descriptions like the names of the ships "JING GANGSHA" and "HENG SHUI" are dropped by all models. In particular, the Seq2Seq-gen model has the problem of generating words not present in the original text (e.g.,

Source text:	Two Chinese war ships , " JING GANGSHA " and " HENG SHUI " arrived at the port of Trincomalee on 13 th January 2014 on a good will visit .
Reference:	Two Chinese war ships , arrived at the port of Trincomalee will visit .
Seq2Seq-gen	
+ Random:	Two Chinese war ships , arrived at the port of toddlers on 13 th January 2014 .
+ Coverage-AL:	Two Chinese war ships , arrived at the port of Trincomalee on a good will visit .
+ Diversity-AL:	Two Chinese war ships arrived at the port of Scottsbluff on 13 th .
Pointer-gen	
+ Random:	Two Chinese war ships , arrived at the port of Trincomalee on 13 th January 2014 .
+ Coverage-AL:	Two Chinese war ships arrived at the port of Trincomalee will visit .
+ Diversity-AL:	Two Chinese war ships , arrived at the port of Trincomalee .

Table 3: In-domain active learning example sentence and compressions for Google News compression dataset when using 20% of labelled compressions with Random, Coverage-AL, Diversity-AL sampling strategies

Methods	MSR-OANC ID			Random			Coverage-AL			Diversity-AL		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
Seq2Seq-gen	30.05	10.42	26.87	33.51	13.60	30.26	**35.10**	**15.00**	**32.78**	34.85	14.92	32.41
Pointer-gen	35.24	16.57	32.56	38.19	21.87	37.94	**39.59**	**24.87**	**37.02**	39.42	24.70	36.86

Table 4: ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-L (RL) achieved by the state-of-the-art models using our sampling strategies when interactively retrained using 10% of the MSR-OANC training set. The results are in comparison to the models trained on in-domain training set (MSR-OANC ID). Bold marks best AL strategy.

"toddlers", "Scottsbluff"). In contrast, the Pointer-gen model's ability to copy from the original text restrains the model from generating irrelevant words. Although Diverysity-AL based models recognized the phrasal constructs crucial for the sentence meaning, Coverage-AL generated the closest compression to the reference.

Active learning for domain adaptation. To test our interactive Seq2Seq model using active learning strategies for the domain adaptation scenario, we train the model on the Google News compression corpus and test it on the multi-genre MSR-OANC compression dataset. Additionally, for domain adaptation, the neural Seq2Seq model is updated incrementally using our interactive compression Algorithm 1. The sampling strategies select the instances to be interactively annotated by the user. As the cost of interactive experimentation with real users, we use simulated feedback from the labeled sentence compressions from the MSR-OANC training data. The two sampling strategies used for in-domain active learning are used for interactive compression with the state-of-the-art Seq2Seq models. Table 4 illustrates the results of the interactive text compression model when applied to the MSR-OANC text compression dataset. One interesting observation is the fact that our sampling strategies at 10% of the training data ($\approx$ 500 samples) perform better than models trained on in-domain training data (MSR-OANC ID) with 5k training instances by +8.3% and +8.2% ROUGE-2.

Figure 3 shows the results for the various sample sizes of the 5k training instances. The results show a similar trend as the active learning for the interactive data-selection scenario. The Coverage-AL and Diversity-AL strategies do not show significant differences from each other. However, the two active learning strategies achieve on average +2.5% ROUGE-2 better results than the random sampling. The results demonstrate that the use of relevant training samples is useful for transferring the models to new domains and genres.

Table 5 shows an example from the MSR-OANC compression dataset. The example illustrates similar compression properties as seen in the in-domain settings. In particular, the two models learned to drop appositions, optional modifiers, detailed clauses, etc. Additionally, we also observed that the difficult cases where those where there is little to be removed, but due to higher compression ratios during the training, the models removed more than required. This confirms the

Source text:	Given the urgency of the situation in Alaska , Defenders needs your immediate assistance to help save Alaska 's wolves from same - day airborne land - and - shoot slaughter .
Reference:	Given the urgency of the situation in Alaska , Defenders needs your immediate assistance saving Alaska 's wolves from slaughter .
Seq2Seq-gen	
+ Random:	Immediate assistance to save Alaska's tundra .
+ Coverage-AL:	Sometimes needs your assistance to help save Alaska 's wolves .
+ Diversity-AL:	The situation in Alaska, help save Alaska 's tundra .
Pointer-gen	
+ Random:	Immediate assistance to help save Alaska s wolves .
+ Coverage-AL:	The urgency of the situation in Alaska , Defenders needs your immediate assistance .
+ Diversity-AL:	Defenders needs your assistance to help save Alaska 's wolves .

Table 5: Domain adaptation example from the MSR-OANC dataset when trained on a 20% of labelled compressions with Random, Coverage-AL, and Diversity-AL sampling strategies

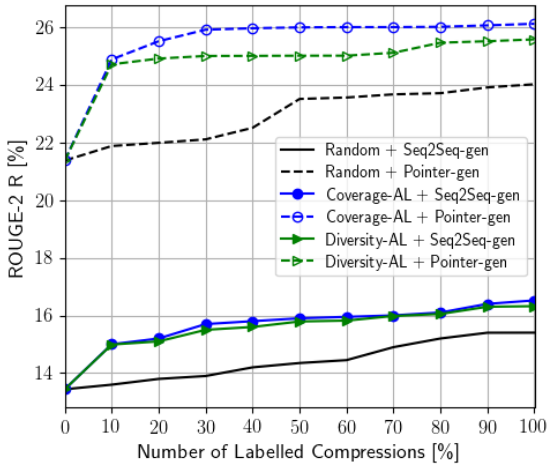

Figure 3: Analysis of the active learning for domain adaptation on the MSR-OANC dataset while varying the training data.

cause for lower ROUGE scores compared to the Google News corpus.

6 Conclusion

We propose a novel neural text compression approach using a neural Seq2Seq method with an interactive setup that aims at (a) learning an in-domain model with a minimum of data and (b) adapting a pretrained model with few user inputs to a new domain or genre. In this paper, we investigate two uncertainty-based active learning strategies with (a) a coverage constraint using attention dispersion and (b) a diversity constraint using structural similarity to make better use of the user in the loop for preparing training data pairs. The active learning based data selection methodology samples the data such that the most uncertain samples are available for training first. Experimental results show that the selected samples achieve comparable performance to the state-of-the-art systems, but trained on 80% less in-domain training data. Active learning with an interactive text compression model helps in transferring models trained on a large parallel corpus for a headline generation task to a general compression dataset with just 500 sampled instances. Additionally, the same in-domain active learning based data selection shows a notable performance improvement in an online interactive domain adaptation setup. Our experiments demonstrate that instead of more training data, relevant training data is essential for training Seq2Seq models in both in-domain training as well as domain adaptation.

In future work, we plan to explore several lines of work. First, we intend to investigate further applications of our interactive setup, e.g., in movie subtitle compression or television closed captions where there is no sufficient training data to build neural models. On a more general level, the interactive setup and the active learning strategies presented can also be used for other natural language processing tasks, such as question answering, to transfer a model to a new domain or genre.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) under grant No. GRK 1994/1. We also acknowledge the useful suggestions of the anonymous reviewers.

References

Yee Seng Chan and Hwee Tou Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), June 23–30, 2007, Prague, Czech Republic*, pages 49–56.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT), June 12–17, 2016, San Diego CA, USA*, pages 93–98.

James Clarke and Mirella Lapata. 2006. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL), July 17–21, 2006, Sydney, Australia*, pages 377–384.

James Clarke and Mirella Lapata. 2007. Modelling compression with discourse constraints. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL), June 28–30, 2007, Prague, Czech Republic*, pages 1–11.

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*, pages 137–144.

Trevor Cohn and Mirella Lapata. 2013. An abstractive approach to sentence compression. *ACM Transactions on Intelligent Systems and Technology*, 4(3):41:1–41:35.

Simon Corston-Oliver. 2001. Text compaction for display on very small screens. In *Proceedings of the NAACL Workshop on Automatic Summarization, June 3, 2001, Pitsburgh, PA, USA*, pages 89–98.

Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP) September 17–21, 2015, Lisbon, Portugal*, pages 360–368.

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP) October 18–21, 2013, Seattle, WA, USA*, pages 1481–1491.

Gholamreza Haffari and Anoop Sarkar. 2009. Active learning for multilingual statistical machine translation. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL/IJCNLP), August 2–7, 2009, Singapore*, pages 181–189.

Udo Hahn, Elena Beisswanger, Ekaterina Buyko, and Erik Faessler. 2012. Active learning-based corpus annotation - the pathojen experience. In *American Medical Informatics Association Annual Symposium (AMIA), November 3–7, 2012, Chicago, IL, USA*.

Hidetaka Kamigaito, Katsuhiko Hayashi, Tsutomu Hirao, and Masaaki Nagata. 2018. Higher-order syntactic attention network for longer sentence compression. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL/HLT), June 1–6, 2018, New Orleans, LA, USA*, pages 1716–1726.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR), May 7–9, 2015, San Diego, CA, USA*.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL), May 27–June 1, 2003, Edmonton, Canada*, pages 48–54.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the ACL Workshop "Text Summarization Branches Out", July 25-26, 2004, Barcelona, Spain*, pages 74–81.

Ming Liu, Wray L. Buntine, and Gholamreza Haffari. 2018. Learning to actively learn neural machine translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL), October 31–November 1, 2018, Brussels, Belgium*, pages 334–344.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), September 17–21, 2015, Lisbon, Portugal, pages 1412–1421.

Juhani Luotolahti and Filip Ginter. 2015. Sentence compression for automatic subtitling. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA), May 11–13, 2015, Vilnius, Lithuania*, pages 135–143.

François Mairesse, Milica Gašić, Filip Jurčiček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve J. Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), July 11–16, 2010, Uppsala, Sweden*, pages 1552–1561.

Erwin Marsi, Emiel Krahmer, Iris Hendrickx, and Walter Daelemans. 2010. On the limits of sentence compression by deletion. In *Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation*, volume 5790 of *Lecture Notes in Artificial Intelligence*, pages 45–66.

Yishu Miao and Phil Blunsom. 2016. Language as a latent variable: Discrete generative models for sentence compression. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), November 1–4, 2016, Austin, TX, USA*, pages 319–328.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX), June 7–8, 2012, Montréal, Canada*, pages 95–100.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of the NAACL-HLT Workshop on Evaluation Metrics and System Comparison for Automatic Summarization, June 3-8, 2012, Montréal, Canada*, pages 1–9.

Avinesh P. V. S. and Christian M. Meyer. 2017. Joint optimization of user-desired content in multi-document summaries by learning from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), July 30–August 4, Vancouver, Canada*, pages 1353–1363.

Álvaro Peris and Francisco Casacuberta. 2018. Active learning for interactive neural machine translation of data streams. In *Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL), October 31–November 1, 2018, Brussels, Belgium*, pages 151–160.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), September 17–21, 2015, Lisbon, Portugal*, pages 379–389.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), July 30–August 4, Vancouver, Canada*, pages 1073–1083.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *Proceedings of the 6th International Conference on Learning Representations (ICLR), May 6–9, New Orleans, LA, USA*.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), August 7–12, 2016, Berlin, Germany*.

Bipin Suresh. 2010. Inclusion of large input corpora in statistical machine translation. Technical report, Stanford University.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montréal, Canada*, pages 3104–3112.

Kristina Toutanova, Chris Brockett, Ke M. Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), November 1-4, 2016, Austin, TX, USA*, pages 340–350.

Jenine Turner and Eugene Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), June 25-30, 2005, Ann Arbor, MI, USA*, pages 290–297.

Vincent Vandegehinste and Yi Pan. 2004. Sentence compression for automated subtitling: A hybrid approach. In *Proceedings of the ACL Workshop "Text Summarization Branches Out", July 25-26, 2004, Barcelona, Spain*, pages 89–95.

Chenguang Wang, Laura Chiticariu, and Yunyao Li. 2017a. Active learning for black-box semantic role labeling with neural factors. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), August 19–25, 2017, Melbourne, Australia*, pages 2908–2914.

Gaoang Wang, Jenq-Neng Hwang, Craig S. Rose, and Farron Wallace. 2017b. Uncertainty sampling based active learning with diversity constraint by sparse selection. In *19th IEEE International Workshop*

on Multimedia Signal Processing (MMSP), October 16–18, 2017, Luton, UK, pages 1–6.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*, 1609.08144.

Sander Wubben, Emiel Krahmer, Antal van den Bosch, and Suzan Verberne. 2016. Abstractive compression of captions with attentive recurrent neural networks. In *Proceedings of the Ninth International Natural Language Generation Conference (INLG), September 5–8, 2016, Edinburgh, UK*, pages 41–50.

Zuobing Xu, Ram Akella, and Yi Zhang. 2007. Incorporating diversity and density in active learning for relevance feedback. In *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2–5, 2007, Proceedings*, pages 246–257.

Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G. Dy. 2011. Active learning from crowds. In *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML), June 28–July 2, 2011, Bellevue, WA, USA*, pages 1161–1168.

Naitong Yu, Jie Zhang, Minlie Huang, and Xiaoyan Zhu. 2018. An operation network for abstractive sentence compression. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING), August 20-26, 2018, Santa Fe, NM, USA*, pages 1065–1076.

Yang Zhao, Zhiyuan Luo, and Akiko Aizawa. 2018. A language model based evaluator for sentence compression. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), July 15–20, 2018, Melbourne, Australia*, pages 170–175.