UNSUPERVISED PRETRAINING FOR SEQUENCE TO SEQUENCE LEARNING

Prajit Ramachandran*
University of Illinois at Urbana-Champaign
prajitram@gmail.com

Peter J. Liu, Quoc V. Le
Google Brain
{peterjliu,qvl}@google.com

ABSTRACT

Sequence to sequence models are successful tools for supervised sequence learning tasks, such as machine translation. Despite their success, these models still require much labeled data and it is unclear how to improve them using unlabeled data, which is much less expensive to obtain. In this paper, we present simple changes that lead to a significant improvement in the accuracy of seq2seq models when the labeled set is small. Our method intializes the encoder and decoder of the seq2seq model with the trained weights of two language models, and then all weights are jointly fine-tuned with labeled data. An additional language modeling loss can be used to regularize the model during fine-tuning. We apply this method to low-resource tasks in machine translation and abstractive summarization and find that it significantly improves the subsequent supervised models. Our main finding is that the pretraining accelerates training and improves generalization of seq2seq models, achieving state-of-the-art results on the WMT English→German task. Our model obtains an improvement of 1.3 BLEU from the previous best models on both WMT'14 and WMT'15 English→German. Our ablation study shows that pretraining helps seq2seq models in different ways depending on the nature of the task: translation benefits from the improved generalization whereas summarization benefits from the improved optimization.

1 INTRODUCTION

Sequence to sequence (*seq2seq*) models (Sutskever et al., 2014; Cho et al., 2014; Kalchbrenner & Blunsom, 2013) are extremely effective on a variety of tasks that require a mapping between a variable-length input sequence to a variable-length output sequence (Sutskever et al., 2014; Vinyals et al., 2015b;a; Bahdanau et al., 2015; Chan et al., 2015; Bahdanau et al., 2016; Vinyals & Le, 2015; Shang et al., 2015; Nallapati et al., 2016; Wu et al., 2016). The main weakness of sequence to sequence models, and deep networks in general, lies in the fact that they can easily overfit when the amount of supervised training data is small.

While it has been suggested that unlabeled data can be used to address this weakness, there is little evidence of a simple and general method achieving a significant improvement on important baselines. In this work, we propose a simple and effective technique for using unsupervised pretraining to improve seq2seq models. Our proposal is to initialize both encoder and decoder networks with pretrained language models. More specifically, to train a seq2seq model mapping from a source domain to a target domain, we first train two language models. One language model is trained on an unlabeled corpus of the source domain and the second is trained on an unlabeled corpus of the target domain. The source-side language model is used to initialize the encoder and the target-side language model is used to initialize the decoder. Finally, the entire seq2seq model is fine-tuned with the labeled corpus.

We benchmark this method on machine translation for English→German and abstractive summarization in a low-resource setting on CNN and Daily Mail articles. Our ablation study shows that among many other possible choices of using a language model in seq2seq with attention, the above proposal works best. Our study also shows that, for translation, the main gains come from the

*Work done as an intern on Google Brain.

improved generalization due to the pretrained features, whereas for summarization the gains come from the improved optimization due to pretraining the encoder which has been unrolled for hundreds of timesteps. On both tasks, our proposed method always improves generalization on the test sets. Our main result is that a seq2seq model, with pretraining, exceeds the strongest possible baseline in both neural machine translation and phrase-based machine translation. Our model obtains an improvement of 1.3 BLEU from the previous best models on both WMT'14 and WMT'15 English→German. On abstractive summarization, our method achieves competitive results to the strongest baselines.

2 Unsupervised Pretraining for Sequence to Sequence Learning

In the following section, we will describe our basic unsupervised pretraining procedure for sequence to sequence learning and how to modify sequence to sequence learning to effectively make use of the pretrained weights. We then show several extensions to improve the basic model.

2.1 Basic Procedure

The basic procedure of our approach is to pretrain both the encoder and decoder networks in the sequence to sequence framework with language models, which can be trained on large amounts of unlabeled text data. This can be seen in Figure 1, where the parameters in the shaded boxes are pretrained. In the following we will describe the method in detail by using machine translation as an example application, but the method can be applied to all sequence to sequence learning tasks.

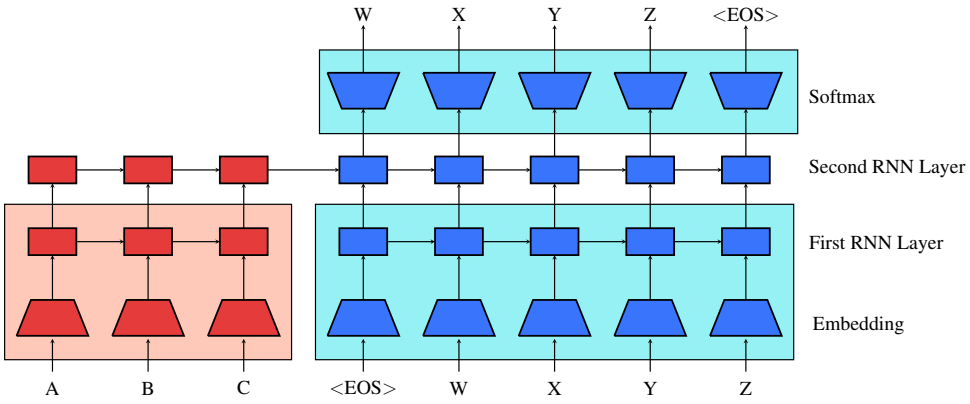

Figure 1: Pretrained sequence to sequence model. The red parameters are the encoder and the blue parameters are the decoder. All parameters in a shaded box are pretrained, either from the source-side (light red) or target-side (light blue) language model. Otherwise, they are randomly initialized.

First, two monolingual datasets are collected, one for the source side language $\mathcal{D}_{src}$, and one for the target side language $\mathcal{D}_{tgt}$. A language model (LM) is trained on each dataset independently, giving an LM trained on the source side corpus L_{src}, and an LM trained on the target side corpus L_{tgt}. L_{src} and L_{tgt} can be different sizes. To simplify our explanation we assume that the LMs have one LSTM layer (Hochreiter & Schmidhuber, 1997), though this technique works regardless of the number of layers.

After two language models are trained, a multi-layer seq2seq model M is constructed. The embedding and first LSTM layer of the encoder are initialized with the corresponding trained weights of L_{src}. Likewise, the embedding, first LSTM layer, and softmax of the decoder are initialized with the corresponding trained weights of L_{tgt}. We call the encoder and decoder parameters that are pretrained θ_{enc} and θ_{dec} respectively. All other LSTM layers are initialized randomly. There is no connection between the first LSTM layer of the encoder and the first LSTM layer of the decoder. In this sense, the first LSTM layer serves as a feature extractor. Finally, during the training of M, all the parameters of the seq2seq model θ_M are fine-tuned on the labeled dataset D_{label}.

Another approach would be to pretrain the entire LSTM, instead of just the embeddings, first LSTM layer, and softmax. However, this requires pretraining a new model for every architecture change, which can be very costly given that a large language model can take on the order of a week or weeks to train (Jozefowicz et al., 2016). Instead, we find that training just two LMs and using our initialization scheme is sufficient to demonstrate significant gains while being very flexible.

2.2 IMPROVING THE MODEL

We also employ three additional methods to further improve the model above. The three methods are: a) Monolingual language modeling losses, b) Residual connections and c) Attention over multiple layers. The three methods are shown in Figure 2 and will be discussed below.

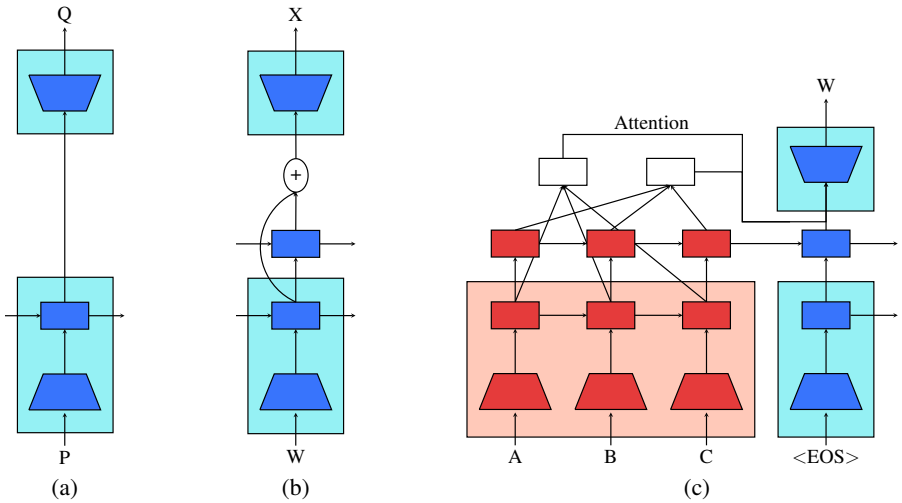

Figure 2: Improvements to the baseline model. (a) monolingual language modeling loss, (b) residual connection, and (c) attention over multiple layers.

Monolingual language modeling losses: After the seq2seq model M is initialized with the two LMs, it is fine-tuned with a labeled dataset. However, there is no requirement that the pretrained parameters stay close to their original pretrained values. That is, the seq2seq model can forget the learned information from language modeling and overfit on the labeled set. To combat this problem, we incorporate additional monolingual language modeling losses to optimize M:

$$\mathcal{L} = \mathcal{L}_{label}(\mathcal{D}_{label}; \theta_M) + \alpha \mathcal{L}_{src}(\mathcal{D}_{src}; \theta_{src}) + \beta \mathcal{L}_{tgt}(\mathcal{D}_{tgt}; \theta_{tgt})$$

The additional losses $\mathcal{L}_{src}$ and $\mathcal{L}_{tgt}$ regularize the seq2seq model by forcing it to correctly model both the normal seq2seq task and the original monolingual language modeling tasks. Note that $\mathcal{L}_{src}$ only involves the pretrained encoder parameters θ_{src} and $\mathcal{L}_{tgt}$ only involves the pretrained decoder parameters θ_{tgt}. This means that the parameters that were pretrained must be useful for both the original language modeling task and the new seq2seq task. In practice, we implemented this joint unsupervised and supervised objective by a round-robin training scheme. First, θ_{enc} and θ_{dec} are optimized with respect to $\mathcal{L}_{src}$ and $\mathcal{L}_{tgt}$ and then all parameters θ_M are updated with respect to $\mathcal{L}_{label}$. In this paper, we used $\alpha = \beta = 1$. We found this regularization gave a large improvement in final test performance, and is an important component of our method.

This is also the motivation for why there is no connection between the first LSTM layers of the encoder and decoder. Without a connection, the first LSTM layers just serve as feature extractors. If the additional language modeling costs are added, the first LSTM layers can focus purely on deriving powerful features for both the language modeling and seq2seq task, without needing to spend capacity for communication between the encoder and decoder.

Residual connections: As described, the input vector to the decoder softmax layer is a random vector because the high level (non-first) layers of the LSTM are randomly initialized. This slows

down training and introduces random gradients to the pretrained parameters, reducing the effectiveness of pretraining. One method to avoid this problem is to freeze all pretrained weights at the start of training and only train the randomly initialized weights. Afterwards, all parameters can be fine-tuned together. We used a simpler solution by introducing a residual connection (He et al., 2016) from the output of the first LSTM layer directly to the input of the softmax. So,

$$h_{\text{softmax}} = h_1 + \text{LSTM}_N(\ldots(\text{LSTM}_2(h_1))\ldots)$$

where LSTM_i represents the i^{th} LSTM layer. This residual connection makes the initialized decoder equivalent to L_{tgt} with random noise added to the input of the softmax. We found this formulation improved performance and decreased the time to convergence. In the experiments we show below, we used a gated residual connection with a bias initialized to a large positive as in Gulcehre et al. (2015), but later experimentation showed that a regular residual connection works just as well.

Attention over multiple layers: In all our models, we used an attention mechanism (Bahdanau et al., 2015). We experimented with attending over both the top level LSTM states, as is done normally, and the first layer LSTM states of the encoder in order to get low and high level features. Given a query vector q_t from the decoder, encoder states from the first layer $h_1^1, \ldots, h_T^1$, and encoder states from the last layer $h_1^L, \ldots, h_T^L$, we compute the attention context vector c_t as follows:

$$\alpha_i = \frac{\exp(q_t \cdot h_i^N)}{\sum_{j=1}^{T} \exp(q_t \cdot h_j^N)} \qquad c_t^1 = \sum_{i=1}^{T} \alpha_i h_i^1 \qquad c_t^N = \sum_{i=1}^{T} \alpha_i h_i^N \qquad c_t = [c_t^1; c_t^N]$$

Note that attention weights α_i are only computed once using the top level encoder states. Furthermore, since our pretraining scheme makes very few assumptions about the model structure, we also experimented with passing the attention vector c_t as input into the next timestep (Luong et al., 2015b). We do not pass c into the first LSTM layer, because the size of the first LSTM layer is predetermined by L_{tgt} which only takes in the embedding as input. Instead, c is passed as input to the second LSTM layer by concatenating it with the output of the first LSTM layer.

We used all three improvements in our experiments. However, in general we noticed that the benefits of the attention modifications are minor in comparison with the benefits of the additional language modeling objectives and residual connections.

3 Related Work

Pretraining was once an essential technique for training deep networks (Hinton et al., 2006; Bengio et al., 2007; Dahl et al., 2010; Erhan et al., 2010). However, better optimization techniques and big labeled datasets rendered pretraining unnecessary for training large, deep networks (Krizhevsky et al., 2012; Nair & Hinton, 2010; Russakovsky et al., 2015; He et al., 2016). Some works in computer vision even found that pretraining can hurt when a lot of labeled data is available (Paine et al., 2014). In our work, we find that pretraining helps sequence to sequence models.

Dai & Le (2015) was amongst the rare studies which showed the benefits of pretraining in a semi-supervised learning setting. They found that pretraining LSTMs gave the state of the art performance for document classification and that increasing the amount of unlabeled data gave better performance for low resource settings. Zoph et al. (2016) showed that initializing a low resource seq2seq model with a seq2seq model trained on a different, higher resource language pair gave large performance gains. Luong et al. (2015a) found that multi-task learning with an autoencoding objective can improve translation performance. Lacoste et al. (2016) found it necessary to pretrain a character level seq2seq model due to the long unroll length. Zhang & Zong (2016) found it useful to add an additional task of sentence reordering of source-side monolingual data for neural machine translation. Firat et al. (2016) introduced fine-tuning techniques that gave promising results for zero-resource neural machine translation.

Sennrich et al. (2015b) introduced a data augmentation technique called backtranslation for neural machine translation. In order to increase the size of the labeled dataset, a target→source model is trained with the labeled dataset and used to translate monolingual target data into the source domain. The new synthetic pairs are added to the labeled dataset. Backtranslation yielded state of the art performance. However this technique does not have a natural analogue to other domains that are not

invertible, like summarization. We note that their technique is complementary to ours, and may lead to additional gains in machine translation. Cheng et al. (2016) proposed using monolingual data for neural machine translation by jointly training a source→target model, target→source model, source autoencoder, and target autoencoder. Their method demonstrates strong improvements. However, it is sensitive to how the monolingual corpora are picked, has to do an expensive top-k search every step for the autoencoding objective, and shows no improvement when using both target and source monolingual data over using just target monolingual data.

Gulcehre et al. (2015) is the work closest to ours. They combine an LM with an already trained seq2seq model by fine-tuning additional deep output layers. However, their method produces small improvements over the supervised baseline ($\leq$ 0.5 BLEU). We suspect that their method does not produce significant gains because (i) the models are trained independently of each other and are not fine-tuned (ii) the LM is combined with the seq2seq model after the last layer, wasting the benefit of the low level LM features, and (iii) only using the LM on the decoder side. Venugopalan et al. (2016) addressed (i) but still experienced minor improvements. Using pretrained GloVe (Pennington et al., 2014) vectors for embeddings was more important for their model.

4 EXPERIMENTS

In the following section, we apply our approach to two important tasks in seq2seq learning: machine translation and abstractive summarization. On each task, we compare against the previous best systems. We also perform ablation experiments to understand the behavior of each component of our method.

4.1 MACHINE TRANSLATION

Dataset and Evaluation: For machine translation, we evaluate our method on the WMT English→German task (Bojar et al., 2015). We used the WMT 14 training dataset, which is slightly smaller than the WMT 15 dataset. Because the dataset has some noisy examples, we used a language detection system to filter the training examples. Sentences pairs where the either source was not English or the target was not German were thrown away. This resulted in around 4 million training examples. Following Sennrich et al. (2015b), we use subword units (Sennrich et al., 2015a) with 89500 merge operations, giving a vocabulary size around 90000. The validation set is the concatenated newstest2012 and newstest2013, and our test sets are newstest2014 and newstest2015. Evaluation on the validation set was with case-sensitive BLEU (Papineni et al., 2002) on tokenized text using `multi-bleu.perl`. Evaluation on the test sets was with case-sensitive BLEU on detokenized text using `mteval-v13a.pl`. The monolingual training datasets are the News Crawl English and German corpora, each of which has more than a billion tokens.

Experimental settings: The language models were trained in the same fashion as (Jozefowicz et al., 2016) We used a 1 layer 4096 dimensional LSTM with the hidden state projected down to 1024 units (Sak et al., 2014) and trained for one week on 32 Tesla K40 GPUs. Our seq2seq model was a 3 layer model, where the second and third layers each have 1000 hidden units. The monolingual objectives, residual connection, and the modified attention were all used. We used the Adam optimizer (Kingma & Ba, 2015) and train with asynchronous SGD on 16 GPUs for speed. We used a learning rate of 5e-5 which is multiplied by 0.8 every 50K steps after an initial 400K steps, gradient clipping with norm 5.0 (Pascanu et al., 2013), and dropout of 0.2 on non-recurrent connections (Zaremba et al., 2014). We used early stopping on validation set perplexity. A beam size of 10 was used for decoding. Our ensemble is constructed with the 5 best performing models on the validation set, which are trained with different hyperparameters.

Results: Table 1 shows the results of our method in comparison with other baselines. Our method achieves a new state of the art for single model performance on both newstest2014 and newstest2015. In fact, our best single model outperforms the previous state of the art ensemble of 4 models. Our ensemble of 5 models matches or exceeds the previous best ensemble of 12 models.

Ablation study: In order to better understand the effects of pretraining, we conducted an ablation study by modifying the pretraining scheme. Figure 3 shows the drop in validation BLEU of various

System	*ensemble?*	*BLEU* *newstest2014*	*newstest2015*
Supervised Neural MT (Jean et al., 2015)	single	-	22.4
Backtranslation (Sennrich et al., 2015b)	single	22.7	25.7
Backtranslation (Sennrich et al., 2015b)	ensemble 4	23.8	26.5
Backtranslation (Sennrich et al., 2015b)	ensemble 12	**24.7**	27.6
Ours pretrained	single	**24.0**	**27.0**
Ours pretrained	ensemble 5	**24.7**	**28.1**

Table 1: English→German performance on WMT test sets.

ablations compared with the full model. The *full model* uses LMs trained with monolingual data to initialize the encoder and decoder, in addition to the language modeling objective. In the following, we interpret the findings of the study. Note that some findings are specific to the translation task.

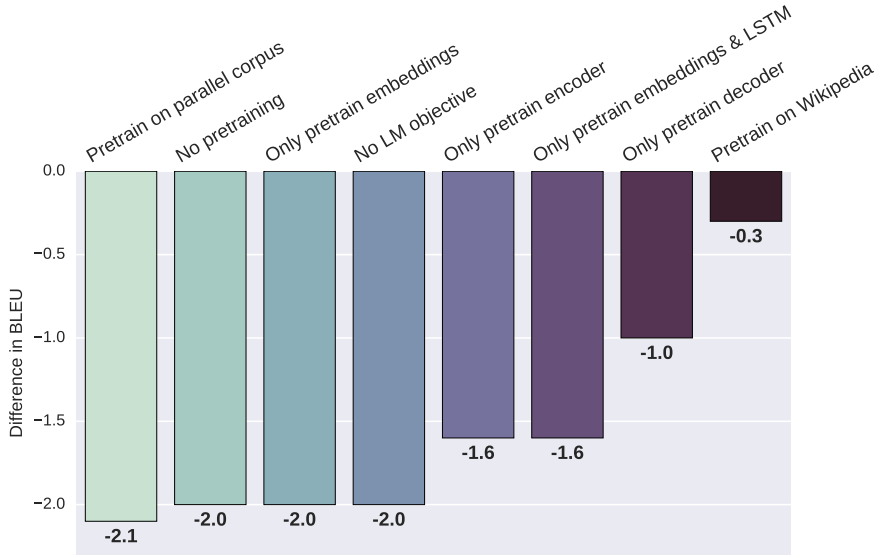

Figure 3: English→German ablation study measuring the difference in validation BLEU between various ablations and the full model. More negative is worse. The full model uses LMs trained with monolingual data to initialize the encoder and decoder, plus the language modeling objective.

Pretraining the decoder is better than pretraining the encoder: Only pretraining the encoder leads to a 1.6 BLEU point drop while only pretraining the decoder leads to a 1.0 BLEU point drop. This suggests that it is more important to pretrain the decoder than the encoder for translation. One possible explanation is that the decoder has a 'tougher job' than the encoder because it has to both understand the semantic meaning of the source sentence and learn to generate the target sentence. In contrast, the encoder only needs to capture the semantic meaning of the source side. Pretraining the decoder is especially helpful for generation because it is initialized as a language model that can already generate in the target language.

Pretrain as much as possible because the benefits compound: Naïvely, one might expect the benefits of pretraining to be additive. That is, the gains of pretraining the entire model is sum of the gains of only pretraining the encoder and only pretraining the decoder. Concretely, given the drops of no pretraining at all (-2.0) and only pretraining the encoder (-1.6), the additive estimate of the drop of only pretraining the decoder side is $-2.0 - (-1.6) = -0.4$. However the actual drop is -1.0 which is a much larger drop than the additive estimate. This suggests that the benefits of pretraining various components are not independent but actually compound on top of each other.

Pretraining the softmax is important: Pretraining only the embeddings and first LSTM layer gives a large drop of 1.6 BLEU points. This suggests that pretraining the decoder softmax is important. Since the gradients from the non-pretrained softmax will be random at the start of training, the pretrained embeddings and LSTM weights will be overwritten with noise which will neutralize the benefits of pretraining. The performance of this ablation may be improved if the pretrained weights were frozen for some iterations before fine-tuning the entire model jointly. The size of the softmax may also influence the importance of pretraining the softmax, but we did not explore this.

The language modeling objective is a strong regularizer: The drop in BLEU points of pretraining the entire model and not using the LM objective is as bad as using the LM objective without pretraining. This suggests that the LM objective acts as a strong regularizer that prevents the weights from drifting too far away from their pretrained state. This ensures that the high quality features that are extracted by the original language model continue to be used. Furthermore, the results imply that the LM objective improves performance regardless of pretraining or not. Future work can use this method to improve their performance even if no pretraining is done.

Pretraining on a lot of unlabeled data is essential for learning to extract powerful features: Surprisingly, if the model is initialized with LMs that are trained on the source part and target part of the *parallel* corpus, the drop in performance is as large as not pretraining at all. This suggests that the primary benefits of pretraining for the translation task come not from the better parameter initialization scheme, but from having learned to extract powerful features. Because the parallel corpus is an order of magnitude smaller than monolingual corpora, the LMs are not able to learn powerful features. Furthermore, the experiments show that the domain of the unlabeled corpora is not as important as the size. For this, we trained two LMs on the English and German Wikipedia. Wikipedia is a different domain than the parallel corpus, which consists mainly of translations from the European Parliment and online news articles (Bojar et al., 2015). The number of tokens in the Wikipedia corpora is comparable to that of the News Crawl monolingual corpora. Initializing with the Wikipedia LMs leads to only a minor drop in performance. This implies that pretraining is effective as long as enough unlabeled data is available.

Very low resource experiments: We also benchmarked the pretraining technique on a very low resource machine translation setup. We first took a random subset of 100K examples from the entire English→German parallel corpus, which is 2.5% of the entire corpus. We trained a randomly initialized baseline and a pretrained model on the subset. For the pretrained model, fixing the embeddings and softmax to their pretrained weights gave the best performance. This is an important advantage of pretraining because the embedding and softmax usually make up the majority of parameters in the model. Freezing them can help prevent severe overfitting. Table 2 shows that the pretraining technique gives a significant improvement of 5.3 BLEU points over the baseline.

System	*BLEU*
Baseline supervised	4.3
Pretrained	9.6

Table 2: Validation set results for training on a 100K example subset of the English→German parallel corpus. The pretrained model freezes the embedding and softmax to their pretrained value.

We were also interested in tracking the performance of pretraining as a function of dataset size. For this, we trained a a model with and without pretraining on random subsets of the English→German corpus. Both models use the additional LM objective. The results are summarized in Figure 4. When a 100% of the labeled data is used, the gap between the pretrained and no pretrain model is 2.0 BLEU points. However, that gap grows when less data is available. When trained on 20% of the labeled data, the gap becomes 3.8 BLEU points. This demonstrates that the pretrained models degrade less as the labeled dataset becomes smaller.

4.2 Abstractive Summarization

Dataset and Evaluation: For a low-resource abstractive summarization task, we use the CNN/Daily Mail corpus from (Hermann et al., 2015). Following Nallapati et al. (2016), we modify the data collection scripts to restore the bullet point summaries. The task is to predict the bullet

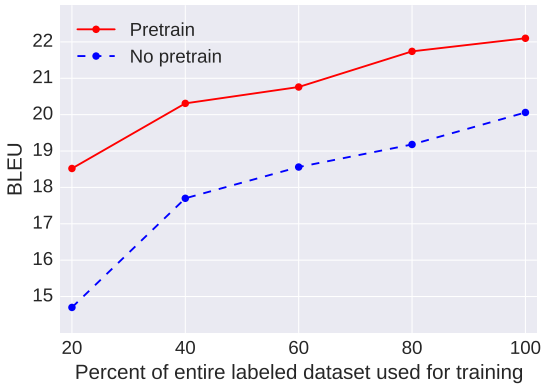

Figure 4: Validation performance of pretraining vs. no pretraining when trained on a subset of the entire labeled dataset for English→German translation.

point summaries from a news article. The dataset has fewer than 300K document-summary pairs. To compare against Nallapati et al. (2016), we used the anonymized corpus. However, for our ablation study, we used the non-anonymized corpus.[1] We evaluate our system using full length ROUGE (Lin, 2004). For the anonymized corpus in particular, we considered each highlight as a separate sentence following Nallapati et al. (2016). In this setting, we used the English Gigaword corpus (Napoles et al., 2012) as our larger, unlabeled "monolingual" corpus, although all data used in this task is in English.

Experimental settings: We use subword units (Sennrich et al., 2015a) with 31500 merges, resulting in a vocabulary size of about 32000. We use up to the first 600 tokens of the document and predict the entire summary. Only one language model is trained and it is used to initialize both the encoder and decoder, since the source and target languages are the same. However, the encoder and decoder are not tied. The LM is a one-layer LSTM of size 1024 trained in a similar fashion to Jozefowicz et al. (2016). For the seq2seq model, we use the same settings as the machine translation experiments. The only differences are that we use a 2 layer model with the second layer having 1024 hidden units, and that the learning rate is multiplied by 0.8 every 30K steps after an initial 100K steps.

Results: Table 3 summarizes our results on the anonymized version of the corpus. Our pretrained model is only able to match the previous baseline seq2seq of Nallapati et al. (2016). However, our model is a unidirectional LSTM while they use a bidirectional LSTM. They also use a longer context of 800 tokens, whereas we used a context of 600 tokens due to GPU memory issues. Furthermore, they use pretrained word2vec (Mikolov et al., 2013) vectors to initialize their word embeddings. As we show in our ablation study, just pretraining the embeddings itself gives a large improvement.

System	*ROUGE-1*	*ROUGE-2*	*ROUGE-L*
Seq2seq + pretrained embeddings (Nallapati et al., 2016)	32.49	11.84	29.47
+ temporal attention (Nallapati et al., 2016)	**35.46**	**13.30**	**32.65**
Ours pretrained	32.56	11.89	29.44

Table 3: Results on the anonymized CNN/Daily Mail dataset.

Ablation study: We performed an ablation study similar to the one performed on the machine translation model. The results are reported in Figure 5. Here we report the drops on ROUGE-1, ROUGE-2, and ROUGE-L on the non-anonymized validation set.

[1]We encourage future researchers to use the non-anonymized version because it is a more realistic summarization setting with a larger vocabulary. Our numbers on the non-anonymized test set are 35.56 ROUGE-1, 14.60 ROUGE-2, and 25.08 ROUGE-L. We did not consider highlights as separate sentences.

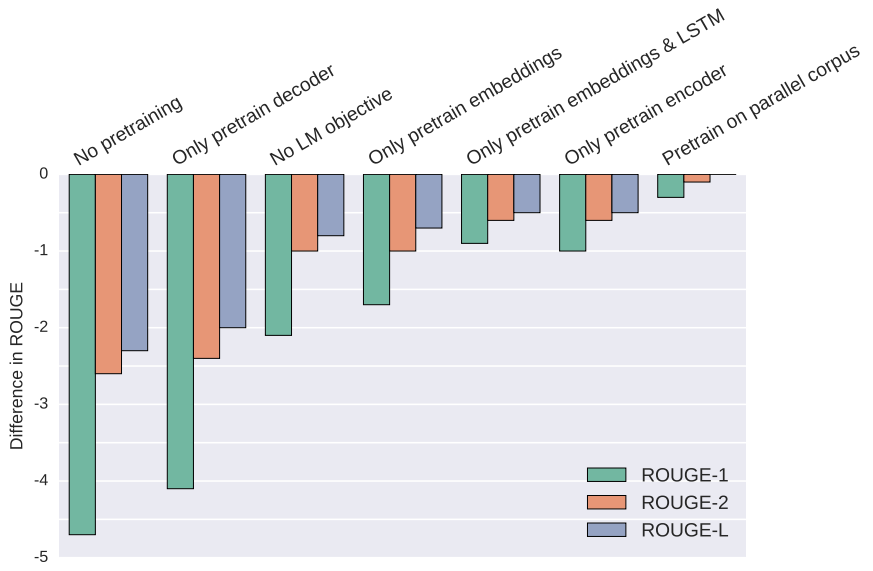

Figure 5: Summarization ablation study measuring the difference in validation ROUGE between various ablations and the full model. More negative is worse. The full model uses LMs trained with unlabeled data to initialize the encoder and decoder, plus the language modeling objective.

Pretraining improves optimization: There is a large drop in performance without any pretraining at all. However, in contrast with the machine translation model, it is more beneficial to only pretrain the encoder than only the decoder of the summarization model. One interpretation is that because the encoder is unrolled for hundreds of timesteps, the pretrained encoder weights enable the gradient to flow much further back in time than randomly initialized weights. This also explains why pretraining on the parallel corpus is as good as pretraining on the unlabeled corpus, which was not observed in the translation model. The pretrained weights improve optimization, which enables a better model to be trained. Interestingly, only pretraining the embeddings gives a significant improvement over no pretraining at all. Randomly initialized encoder embeddings may be difficult to learn because the long unrolling of the encoder leads to vanishing gradient (Pascanu et al., 2013). Pretraining the embeddings eases the burden of optimization and gives the model a good foundation to learn with.

The language modeling objective is a strong regularizer: A model without the LM objective has a nontrivial drop in ROUGE scores. Since there are only a few hundred thousand examples in the dataset, the model can easily overfit. The LM objective combats this by forcing the encoder to maintain its ability to generate English. However, unlike the translation task, pretraining is significantly more important than the LM objective. This is probably because the LM objective helps primarily with regularization while pretraining helps with optimization, which is more important in the long document summarization setting.

Human evaluation: As ROUGE may not be able to capture the quality of summarization, we also performed a small qualitative study to understand the human impression of the summaries produced by different models. We took 200 random documents and compared the performance of a pretrained and non-pretrained system. The document, gold summary, and the two system outputs were presented to a human evaluator who was asked to rate each system output on a scale of 1-5 with 5 being the best score. The system outputs were presented in random order and the evaluator did not know the identity of either output. The evaluator noted if there were repetitive phrases or sentences in either system outputs. Unwanted repetition was also noticed by Nallapati et al. (2016).

Table 4 and 5 show the results of the study. In both cases, the pretrained system outperforms the system without pretraining in a statistically significant manner. The better optimization enabled by pretraining improves the generated summaries and decreases unwanted repetition in the output.

$NP > P$	$NP = P$	$NP < P$
29	88	83

Table 4: The count of how often the no pretrain system (*NP*) achieves a higher, equal, and lower score than the pretrained system (*P*) in the side-by-side study where the human evaluator gave each system a score from 1-5. The sign statistical test gives a p-value of < 0.0001 for rejecting the null hypothesis that there is no difference in the score obtained by either system.

		No pretrain	
		No repeats	Repeats
Pretrain	No repeats	67	65
	Repeats	24	44

Table 5: The count of how often the pretrain and no pretrain systems contain repeated phrases or sentences in their outputs in the side-by-side study. McNemar's test gives a p-value of < 0.0001 for rejecting the null hypothesis that the two systems repeat the same proportion of times. The pretrained system clearly repeats less than the system without pretraining.

5 Conclusion

We showed that pretraining sequence to sequence models is a simple but effective technique that can aid in both generalization and optimization. Our scheme involves pretraining two language models in the source and target domain, and initializing the embeddings, first LSTM layers, and softmax of a sequence to sequence model with the weights of the language models. Using an extra language modeling objective during fine-tuning helps regularize the model. In our experiments, we showed that pretraining can improve results in tasks with hundreds of thousands or millions of examples. A key advantage of this technique is that it is flexible and can be applied to a large variety of tasks, including conversation modeling for chatbots or question answering. We hope that future work uses pretraining as a reliable way to improve performance on sequence to sequence tasks.

Acknowledgments

We thank George Dahl, Andrew Dai, Laurent Dinh, Stephan Gouws, Geoffrey Hinton, Rafal Jozefowicz, Pooya Khorrami, Phillip Louis, Ramesh Nallapati, Arvind Neelakantan, Xin Pan, Abi See, Rico Sennrich, Luke Vilnis, Yuan Yu and the Google Brain team for their help with the project.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Yoshua Bengio, et al. End-to-end attention-based large vocabulary speech recognition. In *ICASSP*, 2016.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *NIPS*, 2007.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 2015.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Semi-supervised learning for neural machine translation. *arXiv preprint arXiv:1606.04596*, 2016.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

George Dahl, Marc'Aurelio Ranzato, Abdel rahman Mohamed, and Geoffrey E. Hinton. Phone recognition with the mean-covariance restricted boltzmann machine. In *NIPS*. 2010.

Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. In *NIPS*. 2015.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman-Vural, and Kyunghyun Cho. Zero-resource translation with multi-lingual neural machine translation. *arXiv preprint arXiv:1606.04164*, 2016.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*. 2015.

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. Montreal neural machine translation systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 2015.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *EMNLP*, 2013.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Alexandre Lacoste, Andrew Fandrianto, Daniel Hewlett, David Berthelot, Illia Polosukhin, Jay Han, Llion Jones, and Matthew Kelcey. WikiReading: A novel large-scale language understanding task over wikipedia. In *ACL*, 2016.

Chin-Yew Lin. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. In *ICLR*, 2015a.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015b.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*. 2013.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. Sequence-to-sequence RNNs for text summarization. *arXiv preprint arXiv:1602.06023*, 2016.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX '12, pp. 95–100, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2391200.2391218`.

Tom Le Paine, Pooya Khorrami, Wei Han, and Thomas S. Huang. An analysis of unsupervised pre-training in light of recent advances. *arXiv preprint arXiv:1412.6597*, 2014.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *ACL*, 2002.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *ICML*, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 2015.

Hasim Sak, Andrew W. Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH*, 2014.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015a.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015b.

Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*. 2014.

Subhashini Venugopalan, Lisa Anne Hendricks, Raymond Mooney, and Kate Saenko. Improving LSTM-based video description with linguistic knowledge mined from text. *arXiv preprint arXiv:1604.01729*, 2016.

Oriol Vinyals and Quoc V. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey E. Hinton. Grammar as a foreign language. In *NIPS*. 2015a.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015b.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

Jiajun Zhang and Chengqing Zong. Exploiting source-side monolingual data in neural machine translation. In *EMNLP*, 2016.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *EMNLP*, 2016.

APPENDIX

SELECTED SUMMARIZATION OUTPUTS

Source Document
(cnn) like phone booths and typewriters , record stores are a vanishing breed – another victim of the digital age . camelot music . virgin megastores . wherehouse music . tower records . all of them gone . corporate america has largely abandoned brick - and - mortar music retailing to a scattering of independent stores , many of them in scruffy urban neighborhoods . and that s not necessarily a bad thing . yes , it s harder in the spotify era to find a place to go buy physical music . but many of the remaining record stores are succeeding – even thriving – by catering to a passionate core of customers and collectors . on saturday , hundreds of music retailers will hold events to commemorate record store day , an annual celebration of , well , your neighborhood record store . many stores will host live performances , drawings , book signings , special sales of rare or autographed vinyl and other happenings . some will even serve beer . to their diehard customers , these places are more than mere stores : they are cultural institutions that celebrate music history (the entire duran duran oeuvre , all in one place !) , display artifacts (aretha franklin on vinyl !) , and nurture the local music scene (hey , here s a cd by your brother s metal band !) . they also employ knowledgeable clerks who will be happy to debate the relative merits of blood on the tracks and blonde on blonde . or maybe , like jack black in high fidelity , just mock your lousy taste in music . so if you re a music geek , drop by . but you might think twice before asking if they stock i just called to say i love you .
Ground Truth summary
saturday is record store day , celebrated at music stores around the world . many stores will host live performances , drawings and special sales of rare vinyl .
No pretrain
corporate america has largely abandoned brick - brick - mortar music . many of the remaining record stores are succeeding – even thriving – by catering to a passionate core of customers .
Pretrained
hundreds of music retailers will hold events to commemorate record store day . many stores will host live performances , drawings , book signings , special sales of rare or autographed vinyl .

Table 6: The pretrained model outputs a highly informative summary, while the no pretrain model outputs irrelevant details.

Source Document
(cnn) hey , look what i did . that small boast on social media can trigger a whirlwind that spins into real - life grief , as a texas veterinarian found out after shooting a cat . dr. kristen lindsey allegedly shot an arrow into the back of an orange tabby s head and posted a proud photo this week on facebook of herself smiling , as she dangled its limp body by the arrow s shaft . lindsey added a comment , cnn affiliate kbtx reported . my first bow kill , lol . the only good feral tomcat is one with an arrow through it s head ! vet of the year award ... gladly accepted . callers rang the phones hot at washington county s animal clinic , where lindsey worked , to vent their outrage . web traffic crashed its website . high price of public shaming on the internet then an animal rescuer said that lindsey s prey was probably not a feral cat but the pet of an elderly couple , who called him tiger . he had gone missing on wednesday , the same day that lindsey posted the photo of the slain cat . cnn has not been able to confirm the claim . as the firestorm grew , lindsey wrote in the comments underneath her post : no i did not lose my job . lol . psshh . like someone would get rid of me . i m awesome ! that prediction was wrong . the clinic fired lindsey , covered her name on its marquee with duct tape , and publicly distanced itself from her actions . our goal now is to go on and try to fix our black eye and hope that people are reasonable and understand that those actions do nt anyway portray what we re for here at washington animal clinic , said dr. bruce buenger . we put our heart and soul into this place . the clinic told wbtx that lindsey was not available for comment . cnn is reaching out to her . she removed her controversial post then eventually shut down her facebook page . callers also complained to the brenham police department and washington county animal control , as her facebook post went viral . the sheriff s office in austin county , where the cat was apparently shot , is investigating , and lindsey could face charges . its dispatchers were overloaded with calls , the sheriff posted on facebook . we are asking you to please take it easy on our dispatchers . as soon as the investigation is complete , we will post the relevant information here on this page , the post read . animal rights activists are pushing for charges . animal cruelty must be taken seriously , and the guilty parties should be punished to the fullest extent of the law , said cat advocacy activist becky robinson . her organization , alley cat allies , is offering a $ 7,500 reward for evidence leading to the arrest and conviction of the person who shot the cat . but others stood up for lindsey . she s amazing . she s caring , said customer shannon stoddard . she s a good vet , so maybe her bad choice of posting something on facebook was not good . but i do nt think she should be judged for it . she dropped off balloons at the animal clinic for lindsey with a thank you note . cnn s jeremy grisham contributed to this report .
Ground Truth summary
dr. kristen lindsey has since removed the post of her holding the dead cat by an arrow . her employer fired her ; the sheriff s office is investigating . activist offers $ 7,500 reward .
No pretrain
dr. kristen lindsey allegedly shot an arrow into the back of an orange orange tabby s head . it s the only good good tomcat is one with an arrow through it s head ! vet vet of the year award .
Pretrained
lindsey lindsey , a texas veterinarian , shot an arrow into the back of an orange tabby s head . she posted a photo of herself smiling , as she dangled its limp body by the arrow s shaft . lindsey could face charges , the sheriff s department says .

Table 7: The pretrained model outputs a highly relevant summary but makes a mistake on the feline executioner's name. The no pretrain model degenerates into irrelevant details and repeats itself.

Source Document
eugenie bouchard s run of poor form continued as the top seed was beaten 6 - 3 , 6 - 1 by american lauren davis in the second round at the family circle cup in charleston on wednesday . davis , 21 , had lost her only career meeting with bouchard , but was in control this time against the world no 7 . davis won nine of the final 11 games of the match and broke bouchard s serve twice in the final set to pull off the upset . eugenie bouchard fires down a serve during her second - round match at the family circle cup bouchard shows her frustrations during her straight - sets defeat by lauren davis on wednesday i ve never beaten here before , so i came out knowing i had nothing to lose , said davis , ranked 66th in the world . bouchard was a semi-finalist last year but had struggled in the lead - up to the charlston green - clay event , losing three of her last five matches to lower - ranked opponents . davis used precise groundstrokes to keep bouchard on her heels throughout the second set . davis broke bouchard s serve to love to take a 3 - 1 lead as the 21 - year - old canadian had a double fault , then sailed a forehand long to fall behind . two games later , davis hit a backhand to send back bouchard s 102mph serve for a winner down the left sideline and a 5 - 1 lead . davis hits a forehand on her way to an impressive win against the canadian top seed bouchard has struggled recently , this time slumping to her fourth defeat in six matches the match ended on bouchard s forehand into the net , davis waving to the cheering crowd . bouchard said : as soon as it was over , definitely a bit of anger , but also kind of this confusion , slash , like quest to find what s wrong , like this kind of searching feeling that , ok , like i know something s off . i know something s not right . so i want to find it . i was definitely a little bit slow today , overpowered . usually , i m the one dominating . so it was definitely , just , i do nt know , just not good .
Ground Truth summary
eugenie bouchard suffered her fourth defeat in six matches . the canadian top seed lost to lauren davis at the family circle cup . world no 66 davis won 6 - 3 , 6 - 1 in the second round in charleston . davis won nine of the final 11 games of the match to seal victory . click here for all the latest news from charleston .
No pretrain
bouchard beat american lauren davis 6 - 3 , 6 - 1 in the second round of the family circle cup in charleston on wednesday . bouchard had lost her only career meeting with bouchard but was in control this time against the world no 7 . bouchard won nine of the final 11 games of the match and broke bouchard s serve twice in the final set to pull off the upset .
Pretrained
eugenie bouchard was beaten 6 - 3 , 6 - 1 by american lauren davis in the second round . davis had lost her only career meeting with bouchard , but was in control this time against the world no 7 . davis hit a backhand to send back bouchard s 102mph serve for a winner down the left sideline .

Table 8: Both models output a relevant summary, but the no pretrain model uses the same name to refer to both players.

Source Document
(cnn) mike rowe is coming to a river near you . sometimes , you hear about a person who makes you feel good about humanity , but bad about yourself , rowe says . on thursday s episode of somebody s got ta do it , rowe meets up with chad pregracke , the founder of living lands & waters , who does just that . pregracke wants to clean up the nation s rivers one piece of detritus at a time . his quota ? always more . read mike rowe s facebook post on how to break our litter habit . since he founded the nonprofit in 1998 at the ripe age of 23 , pregracke and more than 87,000 volunteers have collected 8.4 million pounds of trash from u.s. waterways . those efforts helped him earn the 2013 cnn hero of the year award , along with numerous other honors . wherever you are , no matter if there s a stream , a creek , a lake , whatever , that needs to be cleaned up , you can do it . just organize it and do it , he told cnn s anderson cooper after his win . pregracke also gives rowe a tour of the 150 - foot , solar - powered barge that the living lands & waters staff calls home during lengthy cleanups . the part - home , part - office , part - dumpster has seven bedrooms , two bathrooms , a classroom and a kitchen – and just happens to be made from a recycled strip club . according to the organization s latest annual report , pregracke has made it his mission in 2015 to remove 500,000 more pounds of trash . if you d like to help achieve this goal , visit his website to learn how to help : livinglandsandwaters.org / get - involved / .
Ground Truth summary
chad pregracke was the 2013 cnn hero of the year . mike rowe visited pregracke for an episode of somebody s got ta do it .
No pretrain
rowe meets up with chad pregracke , founder of living lands & waters . pregracke and more than 87,000 volunteers collected 8.4 million pounds of trash from u.s. waterways .
Pretrained
rowe is the founder of living lands & waters , who does just that . pregracke also gives rowe a tour of the 150 - foot barge that the living lands & waters gets .

Table 9: A failure case. The pretrained model outputs irrelevant details while the no pretrain model successfully summarizes the document.

SELECTED ENGLISH→GERMAN OUTPUTS

Source
Mayor Bloomberg told reporters that, because of that court order, the city had suspended the reopening of the public space and protesters were informed, however, that local laws do not allow them to re-install with camping shops and sleeping bags.
Ground Truth
Bürgermeister Bloomberg stellt vor der Presse klar , das aufgrund dieser richterlichen Anordnung die erneute Öffnung des Platzes für den Publikumsverkehr und die Demonstranten aufgehoben worden sei . Die Demonstranten wies er darauf hin , dass die Stadtgesetze ihnen nicht erlaubten , sich erneut mit Zelten und Schlafsäcken an diesem Ort einzurichten .
No pretrain
Der Bürgermeister Bloomberg sagte den Reportern , dass die Stadt aufgrund dieser Gerichtsentscheidung die Wiedereröffnung des öffentlichen Raumes und die Information der Demonstranten ausgesetzt habe , dass die lokalen Gesetze ihnen nicht erlauben , mit den Campingplätzen und Schlafsäcken neu zu installieren .
Pretrained
Bürgermeister Bloomberg erklärte gegenüber Journalisten , dass die Stadt aufgrund dieser Gerichtsentscheidung die Wiedereröffnung des öffentlichen Raums ausgesetzt habe und dass die Demonstranten darüber informiert wurden , dass die örtlichen Gesetze es ihnen nicht erlauben würden , sich mit Campingplätzen und Schlafsälen neu zu installieren .

Table 10: The no pretrain model makes a complete mistranslation when outputting ”und die Information der Demonstranten ausgesetzt habe”. That translates to ”the reopening of the public space and the information [noun] of the protesters were suspended”, instead of informing the protesters. Furthermore, it wrongly separated the two sentences, so the first sentence has extra words and the second sentence is left without a subject. The pretrained model does not make any of these mistakes. However, both models make a vocabulary mistake of ”zu installieren”, which is typically only used to refer to installing software. A human evaluator fluent in both German and English said that the pretrained version was better.

Source
The low February temperatures, not only did they cause losses of millions for the agricultural sector, but they limited the possibilities of the state economy to grow, causing a contraction of the economic activity in general of 3.6 percent in the first half of the year, mainly supported by the historic fall of 31.16 per cent in agriculture, which affected the dynamics of other economic sectors.
Ground Truth
Die niedrigen Temperaturen im Februar verursachten nicht nur Verluste in Millionenhöhe in der Landwirtschaft , sondern steckten darüber hinaus dem Wachstum der Staatswirtschaft enge Grenzen und verursachten im ersten Vierteljahr einen allgemeinen Rückgang der Wirtschaftstätigkeit um 3,6 Prozent Dieser geht hauptsächlich auf den historischen Abbau der landwirtschaftlichen Entwicklung um 31,16 Prozent zurück , der sich bremsend auf weitere Wirtschaftssektoren auswirkte .
No pretrain
Die niedrigen Temperaturen im Februar führten nicht nur zu Verlusten für die Landwirtschaft , sondern sie beschränkten die Möglichkeiten der staatlichen Wirtschaft , wachsen zu wachsen , wodurch die Wirtschaftstätigkeit insgesamt von 3,6 Prozent in der ersten Hälfte des Jahres , hauptsächlich durch den historischen Rückgang von 31.16 % in der Landwirtschaft , beeinflusst wurde , was die Dynamik anderer Wirtschaftssektoren betraf .
Pretrained
Die niedrigen Temperaturen im Februar führten nicht nur zu Verlusten von Millionen für den Agrarsektor , sondern beschränkten die Möglichkeiten der Staatswirtschaft , zu wachsen , was zu einer Schrumpfung der Wirtschaftstätigkeit im Allgemeinen von 3,6 Prozent in der ersten Hälfte des Jahres führte , hauptsächlich durch den historischen Einbruch von 316 Prozent in der Landwirtschaft , der die Dynamik anderer Wirtschaftsbereiche beeinflusst hatte .

Table 11: The human evaluator noted that the pretrained version is better, as it correctly captures the meaning and sentence structure of the middle. The no pretrain model does not misses translating the word ”million”, repeats itself in ”wachsen zu wachsen”, and puts the verb ”beeinflusst wurde” is an unnatural position. However, the pretrained model makes a mistake in the percentage (316% instead of 31.16%).

Source
To facilitate the inception of the Second World War, they allowed bankers and politicians to create a latent conflict situation by saddling Germany with huge war reparations, thereby making a radicalist example of the impoverished masses, it remained only to introduce a sufficiently convincing culprit and a leader with a simple solution, while also creating a multi-racial Czechoslovakia with a strong German minority to play, and indeed did, the role of a fifth colony, once the war had been ignited.
Ground Truth
Um den Zweiten Weltkrieg einfacher entfachen zu können , ließen die Banker durch die Politik eine latente Konfliktsituation schaffen , indem sie Deutschland mit gigantischen Kriegsreparationen belegten ; dadurch schufen sie die Voraussetzung verarmter Massen , so dass sie den Deutschen nur noch einen ausreichend starken Führer unterjubeln mussten , der die Schuldigen benannte und einfache Lösungen anbot ; ein weiterer Faktor war die Schaffung des Vielvölkerstaates Tschechoslowakei mit einer starken deutschen Minderheit , die die Funktion einer fünften Kolonne einnehmen sollte und auch einnahm , um den Kriegsbrand zu entfachen .
No pretrain
Um die Gründung des Zweiten Weltkriegs zu erleichtern , ermöglichte es den Bankern und Politikern , eine latente Konfliktlage zu schaffen , indem sie Deutschland mit enormen Reparationsforderungen konfrontierte , wodurch ein radikalislamistisches Beispiel der verarmten Massen entstand , es blieb nur , einen ausreichend aussagekräftigen Schuldigen und einen Führer mit einer einfachen Lösung zu etablieren , während gleichzeitig eine multi-ethnische Tschechoslowakei mit einer starken deutschen Minderheit zu spielen war und tatsächlich die Rolle einer fünften Kolonie war .
Pretrained
Um die Einführung des Zweiten Weltkrieges zu erleichtern , ließen sie Banker und Politiker eine latente Konfliktlage schaffen , indem sie Deutschland mit riesigen Reparationszahlungen belieferten , wodurch ein radikalislamistisches Beispiel der verarmten Massen entstand , es blieb nur , einen ausreichend überzeugenden Schuldigen und einen Führer mit einer einfachen Lösung zu präsentieren , während gleichzeitig eine multiethnische Tschechoslowakei mit einer starken deutschen Minderheit geschaffen wurde , um zu spielen , und tatsächlich , die Rolle einer fünften Kolonie , sobald der Krieg entfacht worden war .

Table 12: An example where the English source is poorly worded. Both models output poor translations, but the evaluator noted that the pretrained version is still better than the no pretrain version. Interestingly, both models mistranslate ”radical” as ”radikalislamistisches”, which means ”radical Islam”, which is probably a bias in the training data.

Source
The total vote count will also be done if at the end of the ordinary calculation is established that the difference between the winner and the candidate placed on second position is equal to or less than one percentage point, as long as there is a request of the representative of the political party whose candidate came on the second position, case in which there will be excluded the electoral boxes that have been considered during the partial recount.
Ground Truth
Die Stimmenauszählung kann auch in ihrer Gesamtheit erfolgen , wenn nach Abschluss der ordentlichen Berechnung festgestellt wird , dass der Unterschied zwischen dem mutmaßlichen Gewinner und dem Kandidaten auf dem zweiten Platz gleich oder geringer als ein Prozent ist , vorausgesetzt es liegt ein ausdrücklicher Antrag von einem Vertreter der Partei , deren Kandidat Zweiter geworden ist , vor . In diesem Fall würden die Wahlpakete , die einer teilweisen Auszählung ausgesetzt wurden , ausgeschlossen .
No pretrain
Die gesamte Stimmenanzahl wird auch dann erreicht , wenn am Ende der ordentlichen Berechnung festgestellt wird , dass der Unterschied zwischen dem Sieger und dem Kandidaten , der auf der zweiten Position liegt , gleich oder weniger als einen Prozentpunkt beträgt , vorausgesetzt , dass der Vertreter der Partei , deren Kandidat auf der zweiten Position ist , der Fall ist , in dem die Wahlunterlagen , die während der teilweisen Rückzählung berücksichtigt wurden , ausgeschlossen werden .
Pretrained
Die Gesamtzahl der Stimmzettel wird auch dann durchgeführt , wenn am Ende der ordentlichen Berechnung festgestellt wird , dass der Unterschied zwischen dem Gewinner und dem auf den zweiten Platz platzierten Kandidaten gleich oder weniger als einen Prozentpunkt beträgt , solange es einen Antrag des Vertreters der politischen Partei gibt , dessen Kandidat auf die zweite Position kam , in dem es die Wahlzettel ausklammert , die während der Teilzählung berücksichtigt wurden .

Table 13: Another example where the English source is poorly worded. Both models get the structure right, but have a variety of problematic translations. Both models miss the meaning of ”total vote count”. They both also translate ”electoral boxes” poorly - the no pretrain model calls it ”electoral paperwork” while the pretrained model calls it ”ballots”. These failures may be because of the poorly worded English source. The human evaluator found them both equally poor.