Self-Attentive Residual Decoder for Neural Machine Translation

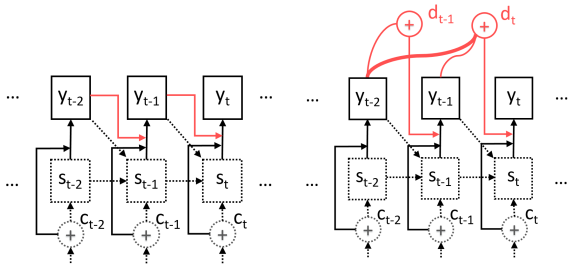
Lesly Miculicich Werlen[*,†], **Nikolaos Pappas**[*], **Dhananjay Ram**[*,†],
Andrei Popescu-Belis[‡]
[*]Idiap Research Institute, Switzerland,
[†]École polytechnique fédérale de Lausanne (EPFL), Switzerland,
[‡]HEIG-VD / HES-SO, Switzerland
{lmiculicich, npappas, dram}@idiap.ch
andrei.popescu-belis@heig-vd.ch

Abstract

Neural sequence-to-sequence networks with attention have achieved remarkable performance for machine translation. One of the reasons for their effectiveness is their ability to capture relevant source-side contextual information at each time-step prediction through an attention mechanism. However, the target-side context is solely based on the sequence model which, in practice, is prone to a recency bias and lacks the ability to capture effectively non-sequential dependencies among words. To address this limitation, we propose a target-side-attentive residual recurrent network for decoding, where attention over previous words contributes directly to the prediction of the next word. The residual learning facilitates the flow of information from the distant past and is able to emphasize any of the previously translated words, hence it gains access to a wider context. The proposed model outperforms a neural MT baseline as well as a memory and self-attention network on three language pairs. The analysis of the attention learned by the decoder confirms that it emphasizes a wider context, and that it captures syntactic-like structures.

1 Introduction

Neural machine translation (NMT) has recently become the state-of-the-art approach to machine translation (Bojar et al., 2016). Several architectures have been proposed for this task (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014; Gehring et al., 2017; Vaswani et al., 2017), but the attention-based NMT model designed by Bahdanau et al. (2015) is still considered the de-facto baseline. This architecture is composed of two recurrent neural networks (RNNs), an encoder and a decoder, and an attention mechanism between them for modeling a soft word-alignment. First, the model encodes the complete source sentence, and then decodes one word at a time. The decoder has access to all the context on the source side through the attention mechanism. However, on the target side, the contextual information is represented only through a fixed-length vector, namely the hidden state of the decoder. As observed by Bahdanau et al. (2015), this creates a bottleneck which hinders the ability of the sequential model to learn longer-term information effectively.



(a) Baseline NMT decoder (b) Self-attentive residual dec.

Figure 1: Comparison between the decoder of the baseline NMT and the proposed decoder with self-attentive residual connections.

As pointed out by Cheng et al. (2016), sequential models present two main problems for natural language processing. First, the memory of the encoder is shared across multiple words and is prone to bias towards the recent past. Second, such models do not fully capture the structural composition of language. To address these limitations, several recent models have been proposed, namely memory networks (Cheng et al., 2016; Tran et al., 2016; Wang et al., 2016) and self-attention networks (Daniluk et al., 2016; Liu and Lapata, 2018). We experimented with these methods, applying them to NMT: *memory RNN* (Cheng et al., 2016) and *self-attentive RNN* (Daniluk et al., 2016). How-

ever, we observed no significant gains in performance over the baseline architecture.

In this paper, we propose a self-attentive residual recurrent decoder, presented in Figure 1b, which, if unfolded over time, represents a densely-connected residual network. The self-attentive residual connections focus selectively on previously translated words and propagate useful information to the output of the decoder, within an attention-based NMT architecture. The attention paid to the previously predicted words is analogous to a read-only memory operation, and enables the learning of syntactic-like structures which are useful for the translation task.

Our evaluation on three language pairs shows that the proposed model improves over several baselines, with only a small increase in computational overhead. In contrast, other similar approaches have lower scores but a higher computational overhead. The contributions of this paper can be summarized as follows:

- We propose and compare several options for using self-attentive residual learning within a standard decoder, which facilitates the flow of contextual information on the target side.
- We demonstrate consistent improvements over a standard baseline, and two advanced variants, which make use of memory and self-attention on three language pairs (English-to-Chinese, Spanish-to-English, and English-to-German).
- We perform an ablation study and analyze the learned attention function, providing additional insights on its actual contributions.

2 Related Work

Several studies have been proposed to enhance sequential models by capturing longer contexts. Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is the most commonly used recurrent neural network (RNN), because its internal memory allows to retain information from a more distant past than a vanilla RNN. Several studies attempt to increase the memory capacity of LSTMs by using memory networks (Weston et al., 2015; Sukhbaatar et al., 2015). For instance, Cheng et al. (2016) incorporate different memory cells for each previous output representation, which are later accessed by an attention mechanism. Tran et al. (2016) include a memory block to access recent input words in a selective manner. Both methods show improvements on language modeling. For NMT, Wang et al. (2016) presented a decoder enhanced with an external shared memory. Memory networks extend the capacity of the network and have the potential to read, write, and forget information. Our method, which attends over previously predicted words, can be seen as a read-only memory, which is simpler but computationally more efficient because it does not require additional memory space.

Other studies aim to improve the modeling of source-side contextual information, for example through a context-aware encoder using self-attention (Zhang et al., 2017), or a recurrent attention NMT (Yang et al., 2017) that is aware of previously attended words on the source-side in order to better predict which words will be attended in future. Additionally, variational NMT (Zhang et al., 2016a) introduces a latent variable to model the underlying semantics of source sentences. In contrast to these studies, we focus instead on the contextual information *on the target side*.

The application of self-attention mechanisms to RNNs have been previously studied, and in general, they seem to capture syntactic dependencies among distant words (Liu and Lapata, 2018; Soltani and Jiang, 2016; Lee et al., 2017; Lin et al., 2017). Daniluk et al. (2016) explore different approaches to self-attention for language modeling, leading to improvements over a baseline LSTM and over memory-augmented methods. However, the methods do not fully utilize a longer context. The main difference with our approach is that we apply attention on the output embeddings rather than the hidden states. Thus, the connections are independent of the recurrent layer representations, which is beneficial to NMT, as we show below.

Our model relies on residual connections, which have been shown to improve the learning process of deep neural networks by addressing the vanishing gradient problem (He et al., 2016). These connections create a direct path from previous layers, helping the transmission of information. Recently, several architectures using residual connections with LSTMs have been proposed for sequence prediction (Zhang et al., 2016b; Kim et al., 2017; Zilly et al., 2017; Wang and Tian, 2016). To our knowledge, our study is the first one to use self-attentive residual connections within residual RNNs for NMT. In parallel to our study, a similar method was recently proposed for sentiment analysis (Wang, 2017).

3 Background: Neural Machine Translation

Neural machine translation aims to compute the conditional distribution of emitting a sentence in a target language given a sentence in a source language, denoted by $p_\Theta(y|x)$, where Θ is the set of parameters of the neural model, and $y = \{y_1, ..., y_n\}$ and $x = \{x_1, ..., x_m\}$ are respectively the representations of source and target sentences as sequences of words. The parameters Θ are learned by training a sequence-to-sequence neural model on a corpus of parallel sentences. In particular, the learning objective is to maximize the following conditional log-likelihood:

$$\max_{\Theta} \frac{1}{N} \sum_{n=1}^{N} \log(p_\Theta(y|x)) \quad (1)$$

The models typically use gated recurrent units (GRUs) (Cho et al., 2014) or LSTMs (Hochreiter and Schmidhuber, 1997). Their architecture has three main components: an encoder, a decoder, and an attention mechanism.

The goal of the encoder is to build meaningful representations of the source sentences. It consists of a bidirectional RNN which includes contextual information from past and future words into the vector representation h_i of a particular word vector x_i, formally defined as follows:

$$h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}] \quad (2)$$

Here, $\overrightarrow{h_i} = f(x_i, h_{i-1})$ and $\overleftarrow{h_i} = f(x_i, h_{i+1})$ are the hidden states of the forward and backward passes of the bidirectional RNN respectively, and f is a non-linear function.

The decoder (see Figure 1a) is in essence a recurrent language model. At each time step, it predicts a target word y_t conditioned over the previous words and the information from the encoder using the following posterior probability:

$$p(y_t|y_1, ..., y_{t-1}, c_t) \approx g(s_t, y_{t-1}, c_t) \quad (3)$$

where g is a non-linear multilayer function. The hidden state of the decoder s_t is defined as:

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \quad (4)$$

and depends on a *context vector* c_t that is computed by the attention mechanism.

The attention mechanism allows the decoder to select which parts of the source sentence are more useful to predict the next output word. This goal is achieved by considering a weighted sum over all hidden states of the encoder as follows:

$$c_t = \sum_{i=1}^{m} \alpha_i^t h_i \quad (5)$$

where α_i^t is a weight calculated using a normalized exponential function a, also known as *alignment function*, which computes how good is the match between the input at position $i \in \{1, ..., n\}$ and the output at position t:

$$\alpha_i^t = softmax(e_i^t) \quad (6)$$

$$e_i^t = a(s_{t-1}, h_i) \quad (7)$$

Different types of alignment functions have been used for NMT, as investigated by Luong et al. (2015). Here, we use the one originally defined by Bahdanau et al. (2015).

4 Self-Attentive Residual Decoder

The decoder of the attention-based NMT model uses a skip connection from the previously predicted word to the output classifier in order to enhance the performance of translation. As we can see in Eq. (3), the probability of a particular word is calculated by a function g which takes as input the hidden state of the recurrent layer s_t, the representation of the previously predicted word y_{t-1}, and the context vector c_t. Within g, these quantities are typically summed up after going through simple linear transformations, hence the addition of y_{t-1} is indeed a skip connection as in residual networks (He et al., 2016). In theory, s_t should be sufficient for predicting the next word given that it is dependent on the other two local-context components according to Eq. (4). However, the y_{t-1} quantity makes the model emphasize the last predicted word for generating the next word. How can we make the model consider a broader context?

To answer this question, we propose to include into the decoder's formula skip connections not only from the previous time step y_{t-1}, but from all previous time steps from y_0 to y_{t-1}. This defines a residual recurrent network which, unfolded over time, can be seen as a densely connected residual network. These connections are applied to all previously predicted words, and reinforce the memory of the recurrent layer towards what has been translated so far. At each time step, the model

decides which of the previously predicted words should be emphasized to predict the next one. In order to deal with the dynamic length of this new input, we use a target-side summary vector d_t that can be interpreted as the representation of the decoded sentence until the time t in the word embedding space. We therefore modify Eq. (3) replacing y_{t-1} with d_t:

$$p(y_t|y_1, ..., y_{t-1}, c_t) \approx g(s_t, d_t, c_t) \quad (8)$$

The replacement of y_{t-1} with d_t means that the number of parameters added to the model is dependent only on the calculation of d_t. Figure 1b illustrates the change made to the decoder. We define two methods for summarizing the context into d_t, which are described in the following sections.

4.1 Mean Residual Connections

One simple way to aggregate information from multiple word embeddings is by averaging them. This average can be seen as the sentence representation until time t. We hypothesize that this representation is more informative than using only the embedding of the previous word. Formally:

$$d_t^{avg} = \frac{1}{t-1} \sum_{i=1}^{t-1} y_i \quad (9)$$

4.2 Self-Attentive Residual Connections

Averaging is a simple and cheap way to aggregate information from multiple words, but may not be sufficient for all kinds of dependencies. Instead, we propose a dynamic way to aggregate information in each sentence, such that different words have different importance according to their relation with the prediction of the next word. We propose to use a shared self-attention mechanism to obtain a summary representation of the translation, i.e. a *weighted average representation* of the words translated from y_0 to y_{t-1}. This mechanism aims to model, in part, important non-sequential dependencies among words, and serves as a complementary memory to the recurrent layer.

$$d_t^{cavg} = \sum_{i=1}^{t-1} \alpha_i^t y_i \quad (10)$$

$$\alpha_i^t = softmax(e_i^t) \quad (11)$$

The weights of the attention model are computed by a scoring function e_i^t that predicts how important each previous word (y_0, ..., or y_{t-1}) is for the current prediction y_t.

We experiment with two different scoring functions, as follows:

$$e_i^t = v^\intercal tanh(W_y y_i + W_s s_t) \ \textit{(content+scope)} \quad (12)$$

$$\text{or } e_i^t = v^\intercal tanh(W_y y_i) \ \textit{(content)} \quad (13)$$

where $v \in \mathbb{R}^e$, $W_y \in \mathbb{R}^{e \times e}$, and $W_s \in \mathbb{R}^{e \times d}$ are weight matrices, e and d are the dimensions of the embeddings and hidden states respectively. Firstly, we study the scoring function noted *content+scope*, as proposed by Bahdanau et al. (2015) for NMT. Secondly, we explore a scoring function noted as *content*, which is calculated based only on the previous hidden states of the decoder, as proposed by Pappas and Popescu-Belis (2017). In contrast to the first attention function, which makes use of the hidden vector s_t, the second one is based only on the previous word representations, therefore, it is independent of the current prediction representation. However, the normalization of this function still depends on t.

5 Other Self-Attentive Networks

To compare our approach with similar studies, we adapted two representative self-attentive networks for application to NMT.

5.1 Memory RNN

The *Memory RNN* decoder is based on the proposal by Cheng et al. (2016) to modify an LSTM layer to include a memory with different cells for each previous output representation. Thus at each time step, the hidden layer can select past information dynamically from the memory. To adapt it to our framework, we modify Eq. (4) as:

$$s_t = f(\tilde{s}_t, y_{t-1}, c_t) \quad (14)$$

$$\text{where} \quad \tilde{s}_t = \sum_{i=1}^{t-1} \alpha_i^t s_i \quad (15)$$

$$\alpha_i^t = softmax(e_i^t) \quad (16)$$

$$e_i^t = a(h_i, y_{t-1}, \tilde{s}_{t-1}) \quad (17)$$

5.2 Self-Attentive RNN

The *Self-Attentive RNN* is the simplest one proposed by Daniluk et al. (2016), and incorporates a summary vector from past predictions calculated with an attention mechanism. Here, the attention is applied over previous hidden states. This decoder is formulated as follows:

$$p(y_t|y_1, ..., y_{t-1}, c_t) \approx g(s_t, y_{t-1}, c_t, \tilde{s}_t) \quad (18)$$

$$\text{where} \quad \tilde{s}_t = \sum_{i=1}^{t-1} \alpha_i^t s_i \tag{19}$$

$$\alpha_i^t = softmax(e_i^t) \tag{20}$$

$$e_i^t = a(s_i, s_t) \tag{21}$$

Additional details of the formulations in Sections 3, 4, and 5 are described in the Appendix A.

6 Experimental Settings

6.1 Datasets

To evaluate the proposed MT models in different conditions, we select three language pairs with increasing amounts of training data: English-Chinese (0.5M sentence pairs), Spanish-English (2.1M), and English-German (4.5M).

For English-to-Chinese, we use a subset of the UN parallel corpus (Rafalovitch and Dale, 2009)[1], with 0.5M sentence pairs for training, 2K for development, and 2K for testing. For training Spanish-to-English MT, we use a subset of WMT 2013 (Bojar et al., 2013), corresponding to Europarl v7 and News Commentary v11 with ca. 2.1M sentence pairs. Newstest2012 and Newstest2013 were used for development and testing respectively. Finally, we use the complete English-to-German set from WMT 2016 (Bojar et al., 2016) with a total of ca. 4.5M sentence pairs. The development set is Newstest2013, and the testing set is Newstest2014. Additionally, we include as testing sets Newstest2015 and Newstest2016, for comparison with the state of the art. We report translation quality using (a) BLEU over *tokenized* and *truecased* texts, and (b) NIST BLEU over *detokenized* and *detruecased* texts[2].

6.2 Model Configuration

We use the implementation of the attention-based NMT baseline provided in `dl4mt-tutorial`[3] developed in Python using Theano (Team et al., 2016). The system implements an attention-based NMT model, described above, using one layer of GRUs (Cho et al., 2014). The vocabulary size is 25K for English-to-Chinese NMT, and 50K for Spanish-to-English and English-German. We use the byte pair encoding (BPE) strategy for out-of-vocabulary words (Sennrich et al., 2016b). For all cases, the maximum sentence length of the training samples is 50, the dimension of the word embeddings is 500, and the dimension of the hidden layers is 1,024. We use dropout with a probability of 0.5 after each layer. The parameters of the models are initialized randomly from a standard normal distribution scaled to a factor of 0.01. The loss function is optimized using Adadelta (Zeiler, 2012) with $\epsilon = 10^{-6}$ and $\rho = 0.95$ as in the original paper. The systems were trained in 7–12 days for each model on a Tesla K40 GPU at the speed of about 1,000 words/sec.

[1] `http://www.uncorpora.org/`

[2] Scrips from Moses toolkit (Koehn et al., 2007): BLEU *multi-bleu*, NIST BLEU *mteval-v13a.pl*, *tokenizer.perl*, *truecase.perl*.

[3] `https://github.com/nyu-dl/dl4mt-tutorial`

Models	|Θ|	BLEU En–Zh	BLEU Es–En
SMT baseline	–	21.6	25.2
NMT baseline	108.7M	22.6	25.4
+ Memory RNN	109.7M	22.5	25.5
+ Self-attentive RNN	110.2M	22.0	25.1
+ Mean residual connections	108.7M	23.6	25.7
+ Self-attentive residual connections	108.9M	**24.0**	**26.3**

Table 1: BLEU score (multi-bleu) on *tokenized* text. The highest score per dataset is marked in bold. The self-attentive residual connections make use of the *content* attention function. $|\Theta|$ indicates the number of parameters per model.

7 Analysis of the Results

Table 1 shows the BLEU scores and the number of parameters used by the different NMT models. Along with the NMT baseline, we included a statistical machine translation (SMT) model based on Moses (Koehn et al., 2007) with the same training/tuning/test data as the NMT. The performance of *memory RNN* is similar to the baseline and, as confirmed later, its focus of attention is mainly on the prediction at $t-1$. The *self-attentive RNN* method is inferior to the baseline, which can be attributed to the overhead on the hidden vectors that have to learn the recurrent representations and the attention simultaneously. The proposed models outperform the baseline, and the best scores are obtained by the NMT model with *self-attentive residual connections*. Despite their simplicity, the *mean residual connections* already improve the translation, without increasing the number of parameters.

Tables 2 and 3 show further experiments with the proposed methods on various English-German test sets, compared to several previous systems. Table 2 shows BLEU values calculated by *multi-*

Models	BLEU NT14	NT15
NMT (unk. word repl.) (Luong et al., 2015)	20.9	–
Context-aware NMT (Zhang et al., 2017)	22.57	–
Recurrent attention NMT (Yang et al., 2017)	22.1	25.0
Variational NMT (Zhang et al., 2016a)	–	25.49
NMT baseline	22.3	24.8
+ Memory RNN	22.6	24.9
+ Self-attentive RNN	22.0	24.3
+ Mean residual connections	22.9	24.9
+ Self-attentive residual connections	**23.2**	**25.5**

Table 2: BLEU score (multi-bleu) on *tokenized* text for English-to-German on *Newstest (NT) 2014, and 2015*. The highest score per dataset is marked in bold. The self-attentive residual connections makes use of the *content* attention function.

Models	BLEU (NIST) NT14	NT15	NT16
Winning WMT	20.1	24.4	**34.2**
NMT (BPE) (Sennrich et al., 2016b)	–	22.8	–
Syntax NMT (Nadejde et al., 2017)	–	–	29.0
NMT Baseline	21.0	24.4	28.8
+ Mean residual connections*	21.4	24.7	29.6
+ Self-attentive residual connections**	**21.7**	**25.0**	29.7

Table 3: NIST BLEU scores on *detokenized* and *detruecased* text for English-to-German on *Newstest (NT) 2014, 2015, 2016*. Significance test: * $p < 0.05$, ** $p < 0.01$. The Winning WMT systems are listed in the text below.

bleu, and includes the NMT system proposed by Luong et al. (2015) which replaces unknown predicted words with the most strongly aligned word on the source sentence. Also, the table includes other systems described in Section 2. Additionally, Table 3 shows values calculated by the NIST BLEU scorer, as well as results reported by the "Winning WMT" systems for each test set respectively: UEDIN-SYNTAX (Williams et al., 2014), UEDIN-SYNTAX (Williams et al., 2015), and UEDIN-NMT (Sennrich et al., 2016a). Also, we include the results reported by Sennrich et al. (2016b) for a baseline encoder-decoder NMT with BPE for unknown words similar to our configuration, and finally the system proposed by Nadejde et al. (2017), an explicit syntax-aware NMT that introduces combinatory categorial grammar (CCG) supertags on the target side by predicting words and tags alternately. The comparison with this work is relevant for the analysis described later in Section 8.2. The results confirm that the

Attention function	BLEU En-Zh	Es-En
Content+Scope	23.1	25.6
Content	**24.0**	**26.3**

Table 4: BLEU scores for two scoring variants of the attention function of the proposed decoder.

self-attentive residual connections improve significantly the translations. To evaluate the significance of the improvements against the NMT baseline, we performed a one-tailed paired t-test.

7.1 Impact of the Attention Function

We now examine the two scoring functions that can be used for the *self-attentive residual connections* model presented in Eq. (12), considering English-to-Chinese and Spanish-to-English. The BLEU scores are presented in Table 4: the best option is the *content* matching function, which depends only on the word embeddings. The *content+scope* function, which depends additionally on the hidden representation of the current prediction is better than the baseline but scores lower than *content*. The idea that the importance of the context depends on the current prediction is appealing, because it can be interpreted as learning internal dependencies among words. However, the experimental results show that it does not necessarily lead to the best translation. On the contrary, the *content* attention function may be extracting representations of the whole sentence which are easier to learn and generalize.

7.2 Performance According to Human Evaluation

Manual evaluation on samples of 50 sentences for each language pair helped to corroborate the conclusions obtained from the BLEU scores, and to provide a qualitative understanding of the improvements brought by our model. For each language, we employed one evaluator who was a native speaker of the target language and had good knowledge of the source language. The evaluators ranked three translations of the same source sentence – one from each of our models: *baseline*, *mean residual connections*, and *self-attentive residual connections* – according to their translation quality. The three translations were presented in a random order, so that the system that had generated them could not be identified. To integrate the judgments, we proceed in pairs, and count the

	Ranking (%)								
System	**En–Zh**			**Es–En**			**En–De**		
	$>$	$=$	$<$	$>$	$=$	$<$	$>$	$=$	$<$
Mean vs. Baseline	26	56	18	20	64	16	28	58	24
Self-attentive vs. Baseline	28	60	12	28	56	16	32	54	14
Self-attentive vs. Mean	24	62	14	28	58	14	32	56	12

Table 5: Human evaluation of sentence-level translation quality on three language pairs. We compare the models in pairs, indicating the percentages of sentences that were ranked higher ($>$), equal to ($=$), or lower ($<$) for the first system with respect to the second one. The values correspond to percentages (%).

Systems	d	**Perplexity**
LSTM (Daniluk et al., 2016)	300	85.2
LSTM + Attention (Daniluk et al., 2016)	296	82.0
LSTM + 4-gram (Daniluk et al., 2016)	968	75.9
LSTM + Mean residual connections	296	80.2
LSTM + Self-attentive residual connections	296	80.4

Table 6: Evaluation of the proposed methods on language modeling. The number of parameter for all models is 47M.

number of times each system was ranked higher, equal to, or lower than another competing system. The results shown in Table 5 indicate that the *self-attentive residual connections* model outperforms the one with *mean residual connections*, and both outperform the baseline, for all three language pairs. The rankings are thus identical to those obtained using BLEU in Tables 1 and 3.

7.3 Performance on Language Modeling

To examine whether language modeling (LM) can benefit from the proposed method, we incorporate the residual connections into a neural LM. We use the same setting as Daniluk et al. (2016) for a corpus of Wikipedia articles (22.5M words), and we compare with two methods proposed in the same paper, namely attention LSTM and 4-gram LSTM. As shown in Table 6, the proposed models outperform the LSTM baseline as well as the self-attention model, but not the 4-gram LSTM. Experiments using 4-gram LSTM for NMT showed poor performance (13.9 BLEU points for English-Chinese) which can be attributed to the difference between the LM and NMT tasks. Both tasks predict one word at a time conditioned over previous words, however, in NMT the previous target-word-inputs are not given, they have to be generated by the decoder. Thus, the output could be conditioned over previous erroneous predictions affecting in higher proportion the 4-gram LSTM

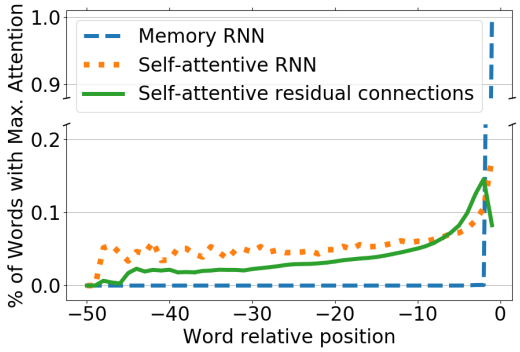


Figure 2: Percentage of words that received maximum attention at a given relative position, ranging from -1 to -50 (maximum length).

model. This shows that even if a model improves language modeling, it does not necessarily improve machine translation.

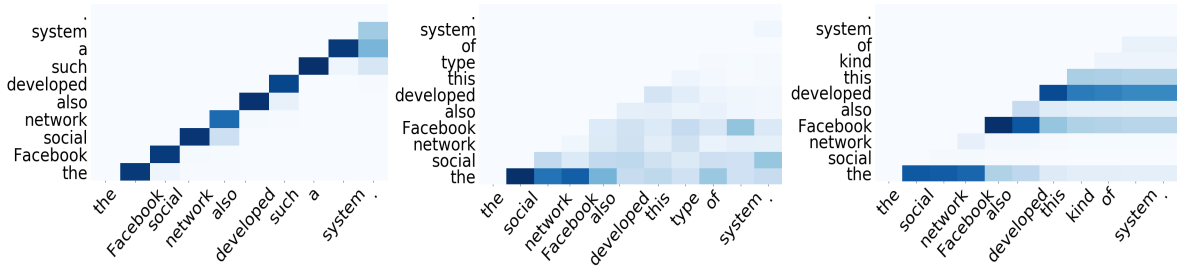
8 Qualitative Analysis

8.1 Distribution of Attention

Figure 2 shows a comparison of the distribution of attention of the different self-attentive models described in this paper, on Spanish-to-English NMT (the other two language pairs exhibit similar distributions). The values correspond to the number of words which received maximal attention for each relative position (x-axis). We selected, at each prediction, the preceding word with maximal weight, and counted its relative position. We normalized the count by the number of previous words at the time of each prediction.

We observe that the *memory RNN* almost always selects the immediately previous word ($t-1$) and ignores the rest of the context. On the contrary, the other two models distribute attention more evenly among all previous words. In particular, the *self-attentive RNN* uses a longer context than the *self-attentive residual connections* but, as the performance on BLEU score shows, this fact does not necessarily mean better translation.

Figure 3 shows the attention to previous words generated by each model for one sentence translated from Spanish to English. The matrices present the target-side attention weights, with the vertical axis indicating the previous words, and the color shades at each position (cell) representing the attention weights. The weights of the *memory RNN* are concentrated on the diagonal, indicating that the attention is generally located on the previous word, which makes the model al-



(a) Memory RNN (b) Self-attentive RNN (c) Self-attentive residual connections

Figure 3: Matrix of distribution of the attention weights to previous words. The vertical axis represents the previous words. A darker shade indicates a higher attention weight.

Algorithm 1 Binary Parse Tree

Require: **A** matrix of attention of size $N \times N$
Require: **s** sentence as list of words of size N
function SPLIT($tree$, **A**, **s**)
  $n \leftarrow length(s)$
  $i \leftarrow 0$
  while $max(\mathbf{A}[:][i]) = 0$ or $i < n$ **do**
    $i \leftarrow i + 1$
  end while
  $tree.addChild(\mathbf{s}[0:i])$
  if $i < n$ **then**
    $subtree \leftarrow newTree()$
    SPLIT($subtree$, $\mathbf{A}[i:n][i:n]$, $\mathbf{s}[i:n]$))
    $tree.addChild(subtree)$
  end if
end function
$tree \leftarrow newTree()$; SPLIT($tree$, **A**, **s**)

most equivalent to the baseline. The weights of the *self-attentive RNN* show that attention is more distributed towards the distant past, and they vary for each word because the attention function depends on the current prediction. This model tries to find dependencies among words, although complex relations seem difficult to learn. On the contrary, the proposed *self-attentive residual connections* model strongly focuses on particular words, and we present a wider analysis of it in the following section.

8.2 Structures Learned by the Model

When visualizing the matrix of attention weights generated by our model (Figure 3c), we observed the formation of sub-phrases which are grouped depending on their attention to previous words. To build the sub-phrases in a deterministic fashion, we implemented Algorithm 1, which iteratively splits the sentence into two sub-phrases every time the focus of attention changes to a new word, from left-to-right. The results are binary tree structures containing the sub-phrases, exemplified in Figure 4.

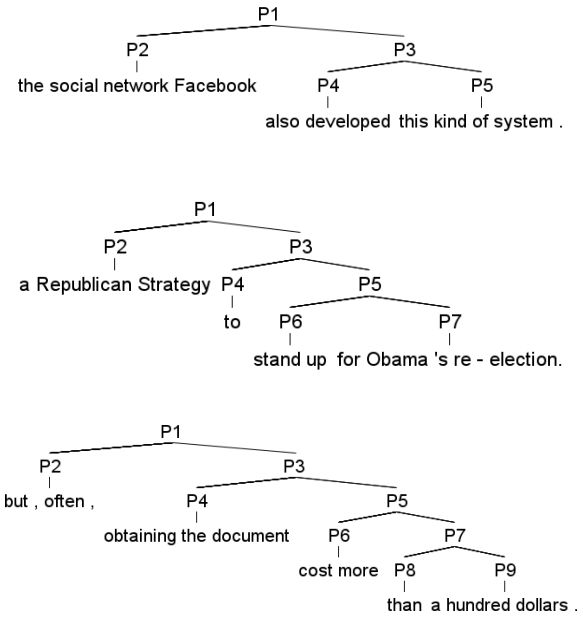


Figure 4: Examples of hypothesized syntactic structures obtained with Algorithm 1.

We formally evaluate the syntactic properties of the binary tree structures by comparing them with the results of an automatic constituent parser (Manning et al., 2014), using the ParsEval approach (Black et al., 1991), i.e. by counting the precision and recall of constituents, excluding single words. Our models reaches a precision of 0.56, which is better than the precision of 0.45 obtained by a trivial right-branched tree model[4]. Note that these structures were neither optimized for parsing nor learned using part-of-speech tagging as most parsers do. Our interpretation of the results is that they are "syntactic-like" structures. However, given the simplicity of the model, they could also be viewed as more limited structures, similar

[4]A model constructed by dividing iteratively one word and the rest of the sentence, from left-to-right.

Better than baseline	
S:	Estudiantes y profesores se están tomando a la ligera la fecha.
R:	Students and teachers are taking the date lightly.
B:	Students and teachers are *being taken lightly to the date*.
O:	Students and teachers are **taking the date lightly**.
S:	No porque compartiera su ideología, sino porque para él los Derechos Humanos son indivisibles.
R:	Not because he shared their world view, but because for him, human rights are indivisible.
B:	Not because *I* share his ideology, but because *he is indivisible by human rights*.
O:	Not because **he** shared his ideology, but because **for him human rights are indivisible**.
Worse than baseline	
S:	El gobierno intenta que no se construyan tantas casas pequeñas.
R:	The Government is trying not to build so many small houses.
B:	The government is trying *not to build so many small houses*.
O:	The government is trying **to ensure that so many small houses are not built**.
S:	Otras personas pueden tener niños .
R:	Other people can have children.
B:	*Other people can* have children.
O:	**Others may** have children.

Table 7: Examples from Spanish to English.

to sentence chunks.

8.3 Translation Examples

Table 7 shows examples of translations produced with the baseline and the *self-attentive residual connections* model. The first part shows examples for which the proposed model reached a higher BLEU score than the baseline. Here, the structure of the sentences, or at least the word order, are improved. The second part contains examples where the baseline achieved better BLEU score than our model. In the first example, the structure of the sentence is different but the content and quality are similar, while in the second one lexical choices differ from the reference.

9 Conclusion

We presented a novel decoder which uses self-attentive residual connections to previously translated words in order to enrich the target-side contextual information in NMT. To cope with the variable lengths of previous predictions, we proposed two methods for context summarization: *mean residual connections* and *self-attentive residual connections*. Additionally, we showed how similar previous proposals, designed for language modeling, can be adapted to NMT. We evaluated the methods over three language pairs: Chinese-to-English, Spanish-to-English, and English-to-German. In each case, we improved the BLEU score compared to the NMT baseline and two variants with memory-augmented decoders. A manual evaluation over a small set of sentences for each language pair confirmed the improvement. Finally, a qualitative analysis showed that the proposed model distributes weights throughout an entire sentence, and learns structures resembling syntactic ones.

As future work, we plan to enrich the present attention mechanism with the *key-value-prediction* technique (Daniluk et al., 2016; Miller et al., 2016) which was shown to be useful for language modeling. Moreover, we will incorporate relative positional information to the attention function. To encourage further research in *self-attentive residual connections* for NMT an other similar tasks, our code is made publicly available[5].

This work is part of the project *Towards Document-Level Neural Machine Translation* (Miculicich Werlen, 2017).

Acknowledgments

We are grateful for support to the European Union under the Horizon 2020 SUMMA project (grant n. 688139, see www.summa-project.eu). We would also like to thank James Henderson for his valuable feedback and suggestions.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, San Diego, USA.

Ezra W. Black, Steven Abney, Daniel P. Flickenger, Claudia Gdaniec, Ralph Grishman, Philip Harrison, Donald Hindle, Robert J. P. Ingria, Frederick Jelinek, Judith L. Klavans, Mark Y. Liberman, Mitchell P. Marcus, Salim Roukos, Beatrice Santorini, and Tomek Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove*, California, USA.

[5] https://github.com/idiap/Attentive_Residual_Connections_NMT

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Michał Daniluk, Tim Rocktäschel, Johannes Welbl, and Sebastian Riedel. 2016. Frustratingly short attention spans in neural language modeling. In *Proceedings of the International Conference on Learning Representations*, San Juan, Puerto Rico.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia. PMLR.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. 2017. Residual lstm: Design of a deep recurrent architecture for distant speech recognition. *arXiv preprint arXiv:1701.03360*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2017. Recurrent additive networks. *arXiv preprint arXiv:1705.07393*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the International Conference on Learning Representations*, Toulon, France.

Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

Thang Luong, Hieu Pham, and D. Christopher Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Lesly Miculicich Werlen. 2017. Towards document-level neural machine translation. Technical report, Idiap.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.

Maria Nadejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn,

and Alexandra Birch. 2017. Predicting target language CCG supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 68–79, Copenhagen, Denmark. Association for Computational Linguistics.

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research*, 58:591–626.

Alexandre Rafalovitch and Robert Dale. 2009. United Nations General Assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit*, volume 12, pages 292–299.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the ACL (Vol. 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rohollah Soltani and Hui Jiang. 2016. Higher order recurrent neural networks. *arXiv preprint arXiv:1605.00064*.

Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

The Theano Development Team, Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, et al. 2016. Theano: A python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*.

Ke Tran, Arianna Bisazza, and Christof Monz. 2016. Recurrent memory networks for language modeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 321–331, San Diego, California. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Cheng Wang. 2017. Rra: Recurrent residual attention for sequence learning. *arXiv preprint arXiv:1709.03714*.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2016. Memory-enhanced decoder for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 278–286, Austin, Texas. Association for Computational Linguistics.

Yiren Wang and Fei Tian. 2016. Recurrent residual learning for sequence classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 938–943, Austin, Texas. Association for Computational Linguistics.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *Proceedings of the International Conference on Learning Representations*, San Diego, USA.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, Eva Hasler, and Philipp Koehn. 2014. Edinburgh's Syntax-Based Systems at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 207–214, Baltimore, Maryland, USA. Association for Computational Linguistics.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, and Philipp Koehn. 2015. Edinburgh's Syntax-Based Systems at WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 199–209, Lisbon, Portugal. Association for Computational Linguistics.

Zichao Yang, Zhiting Hu, Yuntian Deng, Chris Dyer, and Alex Smola. 2017. Neural machine translation with recurrent attention modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 383–387, Valencia, Spain. Association for Computational Linguistics.

Matthew D Zeiler. 2012. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Biao Zhang, Deyi Xiong, Jinsong Su, and Hong Duan. 2017. A context-aware recurrent encoder for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2424–2432.

Biao Zhang, Deyi Xiong, jinsong su, Hong Duan, and Min Zhang. 2016a. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas. Association for Computational Linguistics.

Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James Glass. 2016b. Highway long short-term memory RNNs for distant speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5755–5759.

Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutník, and Jürgen Schmidhuber. 2017. Recurrent highway networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4189–4198, International Convention Centre, Sydney, Australia. PMLR.