What does Attention in Neural Machine Translation Pay Attention to?

Hamidreza Ghader and **Christof Monz**
Informatics Institute, University of Amsterdam, The Netherlands
h.ghader, c.monz@uva.nl

Abstract

Attention in neural machine translation provides the possibility to encode relevant parts of the source sentence at each translation step. As a result, attention is considered to be an alignment model as well. However, there is no work that specifically studies attention and provides analysis of what is being learned by attention models. Thus, the question still remains that how attention is similar or different from the traditional alignment. In this paper, we provide detailed analysis of attention and compare it to traditional alignment. We answer the question of whether attention is only capable of modelling translational equivalent or it captures more information. We show that attention is different from alignment in some cases and is capturing useful information other than alignments.

1 Introduction

Neural machine translation (NMT) has gained a lot of attention recently due to its substantial improvements in machine translation quality achieving state-of-the-art performance for several languages (Luong et al., 2015b; Jean et al., 2015; Wu et al., 2016). The core architecture of neural machine translation models is based on the general encoder-decoder approach (Sutskever et al., 2014). Neural machine translation is an end-to-end approach that learns to encode source sentences into distributed representations and decode these representations into sentences in the target language. Among the different neural MT models, attentional NMT (Bahdanau et al., 2015; Luong et al., 2015a) has become popular due to its capability to use the most relevant parts of the source sentence at each translation step. This capability also makes the attentional model superior in translating longer sentences (Bahdanau et al., 2015; Luong et al., 2015a).

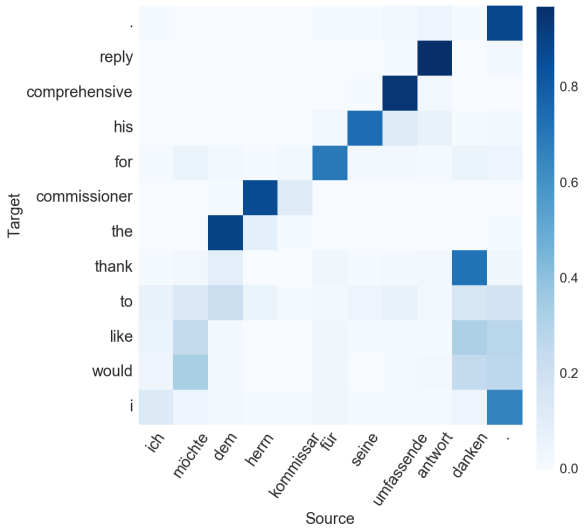


Figure 1: Visualization of the attention paid to the relevant parts of the source sentence for each generated word of a translation example. See how the attention is 'smeared out' over multiple source words in the case of "would" and "like".

Figure 1 shows an example of how attention uses the most relevant source words to generate a target word at each step of the translation. In this paper we focus on studying the relevance of the attended parts, especially cases where attention is 'smeared out' over multiple source words where their relevance is not entirely obvious, see, e.g., "would" and "like" in Figure 1. Here, we ask whether these are due to errors of the attention mechanism or are a desired behavior of the model.

Since the introduction of attention models in neural machine translation (Bahdanau et al., 2015) various modifications have been proposed (Luong et al., 2015a; Cohn et al., 2016; Liu et al., 2016). However, to the best of our knowledge

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 30–39,
Taipei, Taiwan, November 27 – December 1, 2017

there is no study that provides an analysis of what kind of phenomena is being captured by attention. There are some works that have looked to attention as being similar to traditional word alignment (Alkhouli et al., 2016; Cohn et al., 2016; Liu et al., 2016; Chen et al., 2016). Some of these approaches also experimented with training the attention model using traditional alignments (Alkhouli et al., 2016; Liu et al., 2016; Chen et al., 2016). Liu et al. (2016) have shown that attention could be seen as a reordering model as well as an alignment model.

In this paper, we focus on investigating the differences between attention and alignment and what is being captured by the attention mechanism in general. The questions that we are aiming to answer include: Is the attention model only capable of modelling alignment? And how similar is attention to alignment in different syntactic phenomena?

Our analysis shows that attention models traditional alignment in some cases more closely while it captures information beyond alignment in others. For instance, attention agrees with traditional alignments to a high degree in the case of nouns. However, it captures other information rather than only the translational equivalent in the case of verbs.

This paper makes the following contributions: 1) We provide a detailed comparison of attention in NMT and word alignment. 2) We show that while different attention mechanisms can lead to different degrees of compliance with respect to word alignments, global compliance is not always helpful for word prediction. 3) We show that attention follows different patterns depending on the type of the word being generated. 4) We demonstrate that attention does not always comply with alignment. We provide evidence showing that the difference between attention and alignment is due to attention model capability to attend the context words influencing the current word translation.

2 Related Work

Liu et al. (2016) investigate how training the attention model in a supervised manner can benefit machine translation quality. To this end they use traditional alignments obtained by running automatic alignment tools (GIZA++ (Och and Ney, 2003) and fast_align (Dyer et al., 2013)) on the training data and feed it as ground truth to the attention network. They report some improvements in translation quality arguing that the attention model has learned to better align source and target words. The approach of training attention using traditional alignments has also been proposed by others (Chen et al., 2016; Alkhouli et al., 2016). Chen et al. (2016) show that guided attention with traditional alignment helps in the domain of e-commerce data which includes lots of out of vocabulary (OOV) product names and placeholders, but not much in the other domains. Alkhouli et al. (2016) have separated the alignment model and translation model, reasoning that this avoids propagation of errors from one model to the other as well as providing more flexibility in the model types and training of the models. They use a feed-forward neural network as their alignment model that learns to model jumps in the source side using HMM/IBM alignments obtained by using GIZA++.

Shi et al. (2016) show that various kinds of syntactic information are being learned and encoded in the output hidden states of the encoder. The neural system for their experimental analysis is not an attentional model and they argue that attention does not have any impact for learning syntactic information. However, performing the same analysis for morphological information, Belinkov et al. (2017) show that attention has also some effect on the information that the encoder of neural machine translation system encodes in its output hidden states. As part of their analysis they show that a neural machine translation system that has an attention model can learn the POS tags of the source side more efficiently than a system without attention.

Recently, Koehn and Knowles (2017) carried out a brief analysis of how much attention and alignment match in different languages by measuring the probability mass that attention gives to alignments obtained from an automatic alignment tool. They also report differences based on the most attended words.

The mixed results reported by Chen et al. (2016); Alkhouli et al. (2016); Liu et al. (2016) on optimizing attention with respect to alignments motivates a more thorough analysis of attention models in NMT.

3 Attention Models

This section provides a short background on attention and discusses two most popular attention models which are also used in this paper. The first model is a non-recurrent attention model which is equivalent to the "global attention" method proposed by Luong et al. (2015a). The second attention model that we use in our investigation is an input-feeding model similar to the attention model first proposed by Bahdanau et al. (2015) and turned to a more general one and called *input-feeding* by Luong et al. (2015a). Below we describe the details of both models.

Both non-recurrent and input-feeding models compute a context vector c_i at each time step. Subsequently, they concatenate the context vector to the hidden state of decoder and pass it through a non-linearity before it is fed into the softmax output layer of the translation network.

$$\tilde{h}_t = tanh(W_c[c_t; h'_t]) \tag{1}$$

The difference of the two models lays in the way they compute the context vector. In the non-recurrent model, the hidden state of the decoder is compared to each hidden state of the encoder. Often, this comparison is realized as the dot product of vectors. Then the comparison result is fed to a softmax layer to compute the attention weight.

$$e_{t,i} = h_i^T h'_t \tag{2}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{|x|} \exp(e_{t,j})} \tag{3}$$

Here h'_t is the hidden state of the decoder at time t, h_i is ith hidden state of the encoder and $|x|$ is the length of the source sentence. Then the computed alignment weights are used to compute a weighted sum over the encoder hidden states which results in the context vector mentioned above:

$$c_i = \sum_{i=1}^{|x|} \alpha_{t,i} h_i \tag{4}$$

The input-feeding model changes the context vector computation in a way that at each step t the context vector is aware of the previously computed context c_{t-1}. To this end, the input-feeding model feeds back its own $\tilde{h}_{t-1}$ to the network and uses the resulting hidden state instead of the context-independent h'_t, to compare to the hidden states of the encoder. This is defined in the following equations:

$$h''_t = f(W[\tilde{h}_{t-1}; y_{t-1}]) \tag{5}$$

$$e_{t,i} = h_i^T h''_t \tag{6}$$

Here, f is the function that the stacked LSTM applies to the input, y_{t-1} is the last generated target word, and $\tilde{h}_{t-1}$ is the output of previous time step of the input-feeding network itself, meaning the output of Equation 1 in the case that context vector has been computed using $e_{t,i}$ from Equation 6.

	RWTH data
# of sentences	508
# of alignments	10534
% of sure alignments	91%
% of possible alignments	9%

Table 1: Statistics of manual alignments provided by RWTH German-English data.

4 Comparing Attention with Alignment

As mentioned above, it is a commonly held assumption that attention corresponds to word alignments. To verify this, we investigate whether higher consistency between attention and alignment leads to better translations.

4.1 Measuring Attention-Alignment Accuracy

In order to compare attentions of multiple systems as well as to measure the difference between attention and word alignment, we convert the hard word alignments into soft ones and use cross entropy between attention and soft alignment as a loss function. For this purpose, we use manual alignments provided by RWTH German-English dataset as the hard alignments. The statistics of the data are given in Table 1. We convert the hard alignments to soft alignments using Equation 7. For unaligned words, we first assume that they have been aligned to all the words in the source side and then do the conversion.

$$Al(x_i, y_t) = \begin{cases} \frac{1}{|A_{y_t}|} & \text{if } x_i \in A_{y_t} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Here A_{y_t} is the set of source words aligned to target word y_t and $|A_{y_t}|$ is the number of source words in the set.

After conversion of the hard alignments to soft ones, we compute the *attention loss* as follows:

$$L_{At}(y_t) = -\sum_{i=1}^{|x|} Al(x_i, y_t) \log(At(x_i, y_t)) \quad (8)$$

Here x is the source sentence and $Al(x_i, y_t)$ is the weight of the alignment link between source word x_i and the target word (see Equation 7). $At(x_i, y_t)$ is the attention weight $\alpha_{t,i}$ (see Equation 3) of the source word x_i, when generating the target word y_t .

In our analysis, we also look into the relation between translation quality and the quality of the attention with respect to the alignments. For measuring the quality of attention, we use the attention loss defined in Equation 8. As a measure of translation quality, we choose the loss between the output of our NMT system and the reference translation at each translation step, which we call *word prediction loss*. The word prediction loss for word y_t is logarithm of the probability given in Equation 9.

$$p_{nmt}(y_t \mid y_{<t}, x) = softmax(W_o \tilde{h}_t) \quad (9)$$

Here x is the source sentence, y_t is target word at time step t, $y_{<t}$ is the target history given by the reference translation and $\tilde{h}_t$ is given by Equation 1 for either non-recurrent or input-feeding attention models.

Spearman's rank correlation is used to compute the correlation between attention loss and word prediction loss:

$$\rho = \frac{\mathrm{Cov}(R_{L_{At}}, R_{L_{WP}})}{\sigma_{R_{L_{At}}} \sigma_{R_{L_{WP}}}} \quad (10)$$

where $R_{L_{At}}$ and $R_{L_{WP}}$ are the ranks of the attention losses and word prediction losses, respectively, Cov is the covariance between two input variables, and $\sigma_{R_{L_{At}}}$ and $\sigma_{R_{L_{WP}}}$ are the standard deviations of $R_{L_{At}}$ and $R_{L_{WP}}$.

If there is a close relationship between word prediction quality and consistency of attention versus alignment, then there should be high correlation between word prediction loss and attention loss. Figure 2 shows an example with different levels of consistency between attention and word alignments. For the target words "will" and "come" the attention is not focused on the manually aligned word but distributed between the aligned word and other words. The focus of this paper is examining cases where attention does not follow alignment, answering the questions whether those cases represent errors or desirable behavior of the attention model.

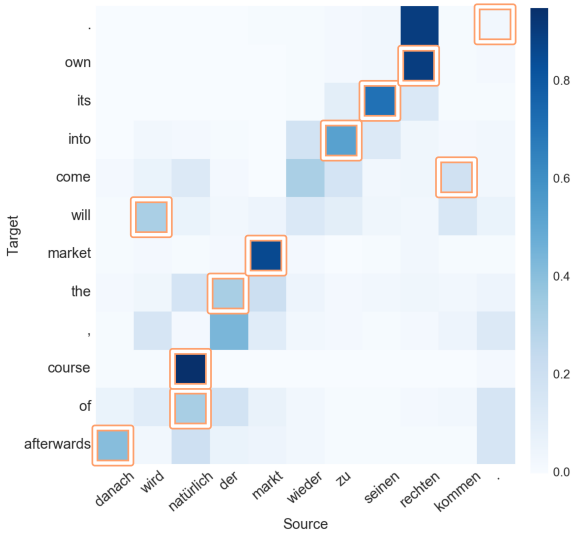


Figure 2: An example of inconsistent attention and alignment. The outlined cells show the manual alignments from the RWTH dataset (see Table 1). See how attention is deviated from alignment points in the case of "will" and "come".

4.2 Measuring Attention Concentration

As another informative variable in our analysis, we look into the attention concentration. While most word alignments only involve one or a few words, attention can be distributed more freely. We measure the concentration of attention by computing the entropy of the attention distribution:

$$E_{At}(y_t) = -\sum_{i=1}^{|x|} At(x_i, y_t) \log(At(x_i, y_t)) \quad (11)$$

5 Empirical Analysis of Attention Behaviour

We conduct our analysis using the two different attention models described in Section 3. Our first attention model is the global model without input-feeding as introduced by Luong et al. (2015a). The second model is the input-feeding model (Luong et al., 2015a), which uses recurrent attention. Our

System	test2014	test2015	test2016	RWTH
Non-recurrent	17.80	18.89	22.25	23.85
Input-feeding	19.93	21.41	25.83	27.18

Table 2: Performance of our experimental system in BLEU on different standard WMT test sets.

NMT system is a unidirectional encoder-decoder system as described in (Luong et al., 2015a), using 4 recurrent layers.

We trained the systems with dimension size of 1,000 and batch size of 80 for 20 epochs. The vocabulary for both source and target side is set to be the 30K most common words. The learning rate is set to be 1 and a maximum gradient norm of 5 has been used. We also use a dropout rate of 0.3 to avoid overfitting.

Data	# of Sent	Min Len	Max Len	Average Len
WMT15	4,240,727	1	100	24.7

Table 3: Statistics for the parallel corpus used to train our models. The length statistics are based on the source side.

5.1 Impact of Attention Mechanism

We train both of the systems on the WMT15 German-to-English training data, see Table 3 for some statistics. Table 2 shows the BLEU scores (Papineni et al., 2002) for both systems on different test sets.

Since we use POS tags and dependency roles in our analysis, both of which are based on words, we chose not to use BPE (Sennrich et al., 2016) which operates at the sub-word level.

	non-recurrent	input-feeding	GIZA++
AER	0.60	0.37	0.31

Table 4: Alignment error rate (AER) of the hard alignments produced from the output attentions of the systems with input-feeding and non-recurrent attention models. We use the most attended source word for each target word as the aligned word. The last column shows the AER for the alignment generated by GIZA++.

We report alignment error rate (AER) (Och and Ney, 2000), which is commonly used to measure alignment quality, in Table 4 to show the difference between attentions and human alignments provided by RWTH German-English dataset. To compute AER over attentions, we follow Luong et al. (2015a) to produce hard alignments from attentions by choosing the most attended source word for each target word. We also use GIZA++ (Och and Ney, 2003) to produce automatic alignments over the data set to allow for a comparison between automatically generated alignments and the attentions generated by our systems. GIZA++ is run in both directions and alignments are symmetrized using the grow-diag-final-and refined alignment heuristic.

	non-recurrent	input-feeding
Attention loss	0.46	0.25

Table 5: Average loss between attention generated by input-feeding and non-recurrent systems and the manual alignment over RWTH German-English data.

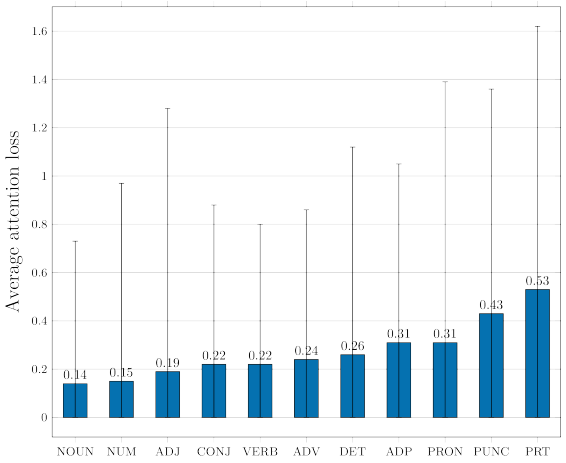
As shown in Table 4, the input-feeding system not only achieves a higher BLEU score, but also uses attentions that are closer to the human alignments.

Table 5 compares input-feeding and non-recurrent attention in terms of attention loss computed using Equation 8. Here the losses between the attention produced by each system and the human alignments is reported. As expected, the difference in attention losses are in line with AER.

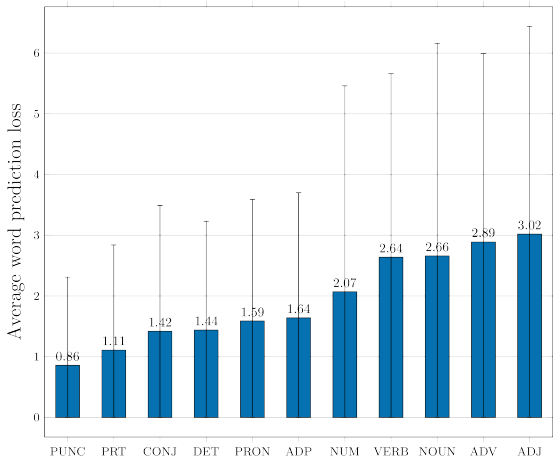
The difference between these comparisons is that AER only takes the most attended word into account while attention loss considers the entire attention distribution.

5.2 Alignment Quality Impact on Translation

Based on the results in Section 5.1, one might be inclined to conclude that the closer the attention is to the word alignments the better the translation. However, Chen et al. (2016); Liu et al. (2016); Alkhouli et al. (2016) report mixed results by optimizing their NMT system with respect to word prediction and alignment quality. These findings warrant a more fine-grained analysis of attention. To this end, we include POS tags in our analysis and study the patterns of attention based on POS tags of the target words. We choose POS tags be-



(a) Average attention loss based on the POS tags of the target side.



(b) Average word prediction loss based on the POS tags of the target side.

Figure 3: Average attention losses and word prediction losses from the input-feeding system.

Tag	Meaning	Example
ADJ	Adjective	large, latest
ADP	Adposition	in, on, of
ADV	Adverb	only, whenever
CONJ	Conjunction	and, or
DET	Determiner	the, a
NOUN	Noun	market, system
NUM	Numeral	2, two
PRT	Particle	's, off, up
PRON	Pronoun	she, they
PUNC	Punctuation	;, .
VERB	Verb	come, including

Table 6: List of the universal POS tags used in our analysis.

cause they exhibit some simple syntactic characteristics. We use the coarse grained universal POS tags (Petrov et al., 2012) given in Table 6.

To better understand how attention accuracy affects translation quality, we analyse the relationship between attention loss and word prediction loss for individual part-of-speech classes. Figure 3a shows how attention loss differs when generating different POS tags. One can see that attention loss varies substantially across different POS tags. In particular, we focus on the cases of NOUN and VERB which are the most frequent POS tags in the dataset. As shown, the attention of NOUN is the closest to alignments on average. But the average attention loss for VERB is almost two times larger than the loss for NOUN.

Considering this difference and the observations in Section 5.1, a natural follow-up would be to focus on getting the attention of verbs to be closer to alignments. However, Figure 3b shows that the average word prediction loss for verbs is actually smaller compared to the loss for nouns. In other words, although the attention for verbs is substantially more inconsistent with the word alignments than for nouns, the NMT system translates verbs more accurately than nouns on average.

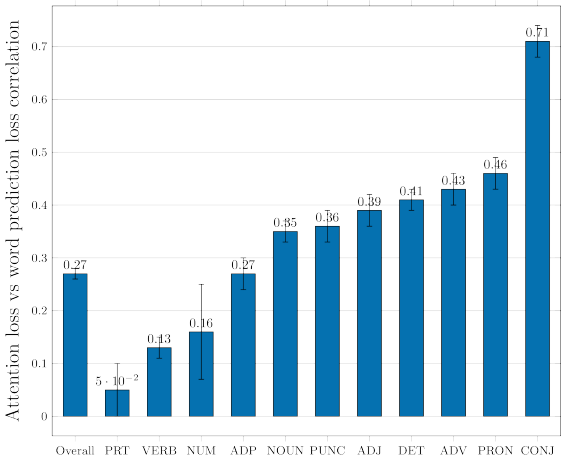
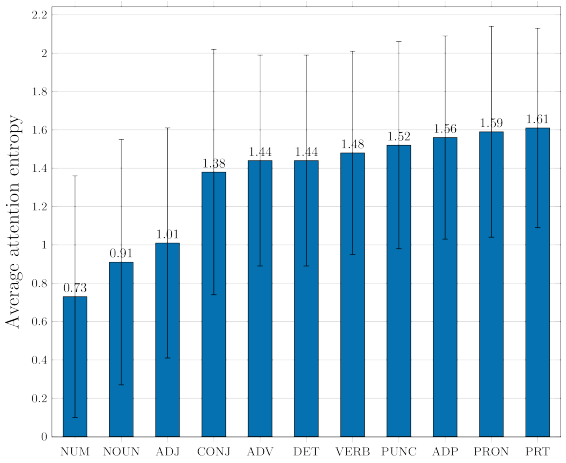


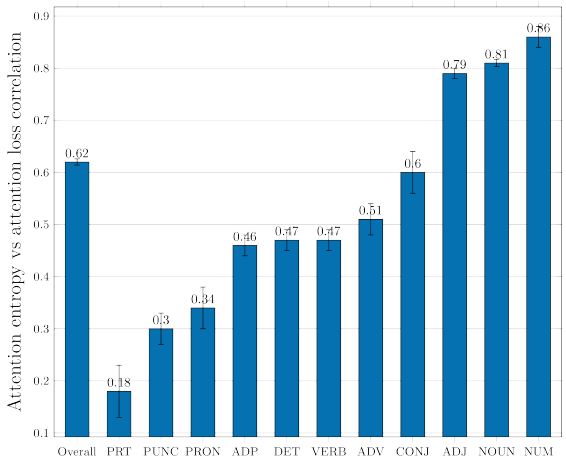
Figure 4: Correlation between word prediction loss and attention loss for the input-feeding model.

To formalize this relationship we compute Spearman's rank correlation between word prediction loss and attention loss, based on the POS tags of the target side, for the input-feeding model, see Figure 4.

The low correlation for verbs confirms that attention to other parts of source sentence rather than the aligned word is necessary for translating verbs and that attention does not necessarily have to follow alignments. However, the higher correla-



(a) Average attention entropy based on the POS tags.



(b) Correlation between attention entropy and attention loss.

Figure 5: Attention entropy and its correlation with attention loss for the input-feeding system.

tion for nouns means that consistency of attention with alignments is more desirable. This could, in a way, explain the mixed result reported for training attention using alignments (Chen et al., 2016; Liu et al., 2016; Alkhouli et al., 2016). Especially the results by Chen et al. (2016) in which large improvements are achieved for the e-commerce domain which contains many OOV product names and placeholders, but no or very weak improvements were achieved over common domains.

5.3 Attention Concentration

In word alignment, most target words are aligned to one source word. The average number of source words aligned to nouns and verbs is 1.1 and 1.2 respectively. To investigate to what extent this also holds for attention we measure the attention concentration by computing the entropy of the attention distribution, see Equation 11.

Figure 5a shows the average entropy of attention based on POS tags. As shown, nouns have one of the lowest entropies meaning that on average the attention for nouns tends to be concentrated. This also explains the closeness of the attention to alignments for nouns. In addition, the correlation between attention entropy and attention loss in case of nouns is high as shown in Figure 5b. This means that attention entropy can be used as a measure of closeness of attention to alignment in the case of nouns.

The higher attention entropy for verbs, in Figure 5a, shows that the attention is more distributed compared to nouns. The low correlation between attention entropy and word prediction loss (see Figure 6) shows that attention concentration is not required when translating into verbs. This also confirms that the correct translation of verbs requires the systems to pay attention to different parts of the source sentence.

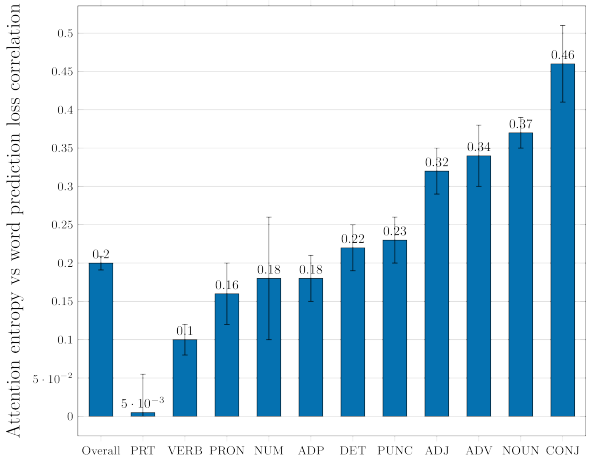


Figure 6: Correlation of attention entropy and word prediction loss for the input-feeding system.

Another interesting observation here is the low correlation for pronouns (PRON) and particles (PRT), see Figure 6. As can be seen in Figure 5a, these tags have more distributed attention comparing to nouns, for example. This could either mean that the attention model does not know where to focus or it deliberately pays attention to multiple, somehow relevant, places to be able to produce a better translation. The latter is supported by the relatively low word prediction losses, shown in the Figure 3b.

POS tag	roles(attention %)	description
NOUN	punc(16%)	Punctuations[1]
	pn(12%)	Prepositional complements
	attr(10%)	Attributive adjectives or numbers
	det(10%)	Determiners
VERB	adv(16%)	Adverbial functions including negation
	punc(14%)	Punctuations
	aux(9%)	Auxiliary verbs
	obj(9%)	Objects[2]
	subj(9%)	Subjects
CONJ	punc(28%)	Punctuations
	adv(11%)	Adverbial functions including negation
	conj(10%)	All members in a coordination[3]

Table 7: The most attended dependency roles with their received attention percentage from the attention probability mass paid to the words other than the alignment points. Here, we focus on the POS tags discussed earlier.

5.4 Attention Distribution

To further understand under which conditions attention is paid to words other than the aligned words, we study the distribution of attention over the source words. First, we measure how much attention is paid to the aligned words for each POS tag, on average. To this end, we compute the percentage of the probability mass that the attention model has assigned to aligned words for each POS tag, see Table 8.

POS tag	attention to alignment points %	attention to other words %
NUM	73	27
NOUN	68	32
ADJ	66	34
PUNC	55	45
ADV	50	50
CONJ	50	50
VERB	49	51
ADP	47	53
DET	45	55
PRON	45	55
PRT	36	64
Overall	54	46

Table 8: Distribution of attention probability mass (in %) over alignment points and the rest of the words for each POS tag.

One can notice that less than half of the attention is paid to alignment points for most of the POS tags. To examine how the rest of attention in each case has been distributed over the source sentence we measure the attention distribution over dependency roles in the source side. We first parse the source side of RWTH data using the ParZu parser (Sennrich et al., 2013). Then we compute how the attention probability mass given to the words other than the alignment points, is distributed over dependency roles. Table 7 gives the most attended roles for each POS tag. Here, we focus on POS tags discussed earlier. One can see that the most attended roles when translating to nouns include adjectives and determiners and in the case of translating to verbs, it includes auxiliary verbs, adverbs (including negation), subjects, and objects.

6 Conclusion

In this paper, we have studied attention in neural machine translation and provided an analysis of the relation between attention and word alignment. We have shown that attention agrees with traditional alignment to a certain extent. However, this differs substantially by attention mechanism and the type of the word being generated. We have shown that attention has different patterns based on the POS tag of the target word. The concentrated pattern of attention and the relatively high correlations for nouns show that training the attention with explicit alignment labels is useful for generating nouns. However, this is not the case for verbs, since the large portion of attention being paid to words other than alignment points, is already capturing other relevant information. Training attention with alignments in this

[1] Punctuations have the role "root" in the parse generated using ParZu. However, we use the pos tag to discriminate them from tokens having the role "root".

[2] Attention mass for all different objects are summed up.

[3] Includes all different types of conjunctions and conjoined elements.

case will force the attention model to forget these useful information. This explains the mixed results reported when guiding attention to comply with alignments (Chen et al., 2016; Liu et al., 2016; Alkhouli et al., 2016).

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers 639.022.213 and 612.001.218.

References

Tamer Alkhouli, Gabriel Bretschner, Jan-Thorsten Peter, Mohammed Hethnawi, Andreas Guta, and Hermann Ney. 2016. Alignment-based neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 54–65, Berlin, Germany. Association for Computational Linguistics.

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

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*.

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided alignment training for topic-aware neural machine translation. *AMTA 2016, Vol.*, page 121.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3093–3102, Osaka, Japan. The COLING 2016 Organizing Committee.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015a. Effective approaches to attention-based neural machine translation. pages 1412–1421.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015b. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China, October 1-8, 2000.*

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Slav Petrov, Dipanjan Das, and Ryan Mcdonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Language Resources and Evaluation Conference*.

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

Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting synergies between open resources for german dependency parsing, pos-tagging, and morphological analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 601–609, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NIPS)*, pages 3104–3112.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.