PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 108 JUNE 2017 27–36

Generating Alignments Using Target Foresight in Attention-Based Neural Machine Translation

Jan-Thorsten Peter, Arne Nix, Hermann Ney

Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Ahornstr. 55, 52056 Aachen, Germany

Abstract

Neural machine translation (NMT) has shown large improvements in recent years. The currently most successful approach in this area relies on the attention mechanism, which is often interpreted as an alignment, even though it is computed without explicit knowledge of the target word. This limitation is the most likely reason that the quality of attention-based alignments is inferior to the quality of traditional alignment methods. Guided alignment training has shown that alignments are still capable of improving translation quality. In this work, we propose an extension of the attention-based NMT model that introduces target information into the attention mechanism to produce high-quality alignments. In comparison to the conventional attention-based alignments, our model halves the AER with an absolute improvement of 19.1% AER. Compared to GIZA++ it shows an absolute improvement of 2.0% AER.

1. Introduction

The field of machine translation has seen a drastic shift in recent years since it has been demonstrated that end-to-end neural machine translation (NMT) models (Bahdanau et al., 2015) are able to outperform traditional phrase-based systems on numerous tasks. A key component of the approach introduced by Bahdanau et al. is the attention mechanism, which has been subject to a lot of research (Luong et al., 2015; Tu et al., 2016; Mi et al., 2016a; Sankaran et al., 2016; Feng et al., 2016; Cohn et al., 2016). The attention mechanism produces a distribution over the source sentence for every decoding step. This distribution is often interpreted as a soft alignment between the source and target sentence. It has been shown that incorporating alignment in-

Corresponding author: peter@cs.rwth-aachen.de
Cite as: Jan-Thorsten Peter, Arne Nix, Hermann Ney. Generating Alignments Using Target Foresight in Attention-Based Neural Machine Translation. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 27–36. doi: 10.1515/pralin-2017-0006.

formation during training as an additional objective function can improve the overall performance of the system (Chen et al., 2016). This indicates that the alignment problem is still relevant.

The relation between attention and alignments provides the motivation for this work, which aims at using the attention-based NMT approach to generate word alignments. However, the attention mechanism has a disadvantage compared to regular word alignment methods. While the word alignment is computed including the knowledge of the whole source and target sentence, the neural network knows only previously seen words on the target side. To remove this disadvantage, we extend the standard attention computation by introducing knowledge of the target word to which we want to align.

2. Related Work

Based on the NMT approach by Bahdanau et al. (2015) researchers have tried to improve the translation quality by modifying the attention mechanism. Most methods add various features to the attention computation (Tu et al., 2016; Mi et al., 2016a; Sankaran et al., 2016; Feng et al., 2016; Cohn et al., 2016), while others attempt to change the attention mechanism itself (Zhang et al., 2016). External alignments have been utilized to teach the network to mimic them by adding them to the objective function during training (Chen et al., 2016; Mi et al., 2016b).

Even though most of these approaches interpret the attention as a soft alignment, to the best of our knowledge, there have been only four publications that empirically measure the impact of their approach on the alignment quality (Tu et al., 2016; Mi et al., 2016a,b; Sankaran et al., 2016). These investigations use the Saer (Tu et al., 2016), Aer (Och and Ney, 2003) and F_1 metrics to measure the alignment quality. All authors noticed an improvement in alignment quality by applying their extensions to the attention mechanism, but as Mi et al. (2016b) report, there is still a significant qualitative difference to state-of-the-art alignments.

A method to create alignments using posterior regularization was presented by Ganchev et al. (2010) and Tamura et al. (2014) which used a special purpose recurrent neural network to create alignments.

3. Neural Machine Translation

The neural machine translation approach, as introduced by Bahdanau et al. (2015), is composed of three main components: The encoder, the attention mechanism, and the decoder (Figure 1). The encoder is a bidirectional recurrent neural network (RNN) which is applied to the input sentence f_1^J to produce the source representation h_1^J, where J is the sentence length. In each decoder step $i = 1, \ldots, I$ the encoder state for each source position $j = 1, \ldots, J$ is used to compute the attention energies $\tilde{\alpha}_{ij}$. For this a single hidden layer with weights W_a, U_a and an additional transformation vector

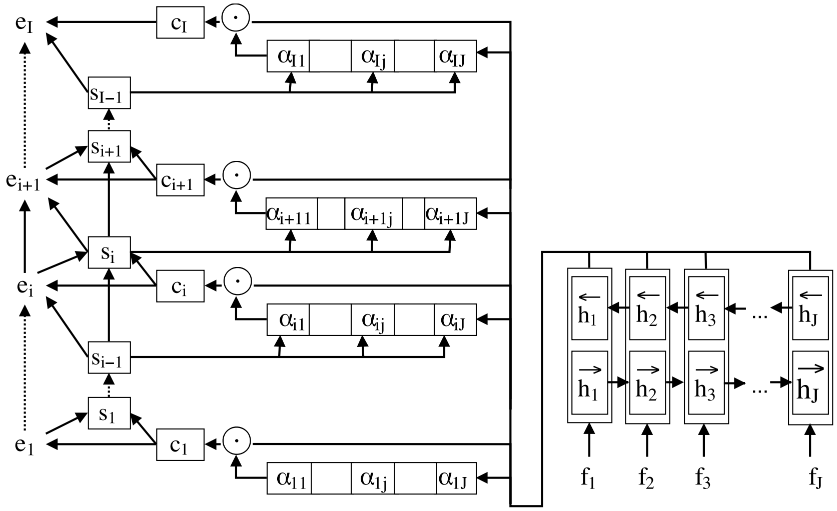


Figure 1. The unmodified attention-based NMT model (Bahdanau et al., 2015)

v_a is applied to the previous decoder state s_{i-1} and the relevant source representation h_j.

$$\tilde{\alpha}_{ij} := v_a^T \tanh(W_a s_{i-1} + U_a h_j) \tag{1}$$

The energies are converted into the attention weights α_{ij} by normalization with a softmax function over all $j = 1, \ldots, J$. These weights are used to compute the context vector c_i as a weighted sum of the encoder representations h_1^J.

This context vector c_i is handed over to the decoder which generates the output word e_i while taking the previously generated output e_{i-1}, the old decoder state s_{i-1} and the context vector c_i as inputs. At the end of each decoding step, the hidden decoder state s_i is updated w.r.t. the previous hidden state s_{i-1}, the context vector c_i and the generated output word e_i.

An extension to the standard training procedure for NMT models is introduced by guided alignment training (Chen et al., 2016; Mi et al., 2016b). This approach is designed to benefit from state-of-the-art alignments by defining an additional cost function that gives feedback explicitly to the components of the attention mechanism. This second loss function is computed for a set of N training samples as the cross-entropy between the soft alignment α_{ij} extracted from the attention mechanism and a given target alignment A_{ij}, provided by e.g.GIZA++ (Och and Ney, 2003):

$$\mathcal{L}_{\text{al}}(A, \alpha) := -\frac{1}{N} \sum_n \sum_{i=1}^{I^{(n)}} \sum_{j=1}^{J^{(n)}} A_{ij}^{(n)} \log \alpha_{ij}^{(n)} \tag{2}$$

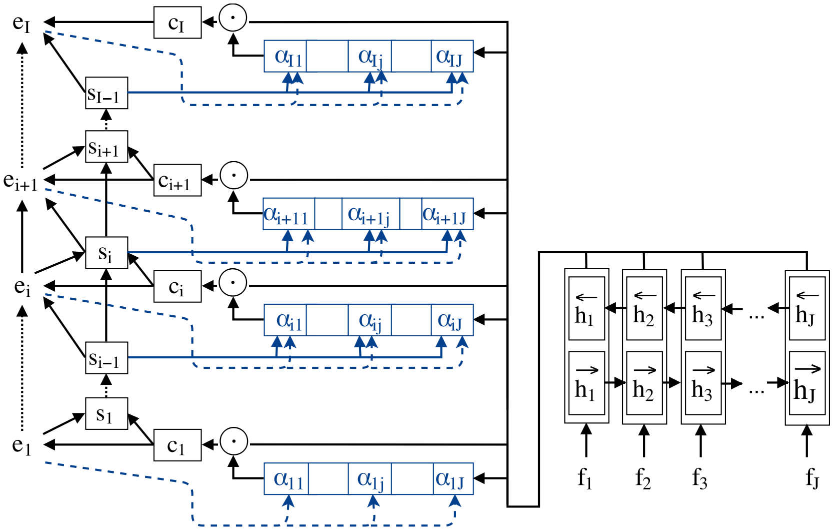


Figure 2. Attention-based NMT with target foresight, the dotted lines show how the current target word is feedback to the alignment computation.

To integrate this additional error measure into the traditional training process a new network loss function is defined as the weighted sum of the standard decoder cost function and the introduced alignment cost function.

4. Target Foresight

Since the introduction of the IBM models (Brown et al., 1993), alignments have always been important for statistical machine translation. And even though the attention mechanism (Bahdanau et al., 2015) does not explicitly generate an alignment, approaches like guided alignment training (Chen et al., 2016) and the analysis by Tu et al. (2016) indicate that the information encoded in the attention weights is related to an alignment from source to target side.

The aim of this work is to explore the alignment capabilities of the attention-based NMT model and to create alignments that are optimized for NMT. The latter is important since the attention mechanism does not assign weights to the source words, but to the encoder representation that is generated from these words. This representation may consequently encode information about neighboring words in the source sentence.

Nevertheless, we interpret the attention weights as a soft alignment for the remaining sections of this work and try to improve the alignment quality compared to the standard attention mechanism. We follow the example of traditional alignment methods and use the knowledge of the target reference sentence $\hat{e}_1^{\hat{I}}$ to improve the alignment quality of the attention. Therefore, we introduce the target word of the current decoding step $\hat{e}_i$ as additional input for the attention energy computation:

$$\tilde{\alpha}_{ij} = v_a^T \tanh(W_a s_{i-1} + U_a h_j + V_a \hat{e}_i). \quad (3)$$

We refer to this approach as *target foresight (TF)*, since the network is allowed to use the foresight of the target word $\hat{e}_i$ to determine the corresponding source position that should be aligned to $\hat{e}_i$. Figure 2 shows the additional connection added to the NMT model.

To further investigate the target foresight approach, we propose three different methods to be applied during training. First we add random noise to the value of $\tilde{\alpha}_{ij}$, which is supposed to prevent the encoding of target-word information in the attention weights. The second approach is to freeze the values of all weight matrices except for the attention parameters in the update steps of the training. The last approach is to train target foresight using guided alignment training (Chen et al., 2016; Mi et al., 2016b). This approach works by enforcing the network not to diverge too far from a given alignment. It allows however to chose a different alignment point if the improvement in the translation cost is large enough.

5. Experiments

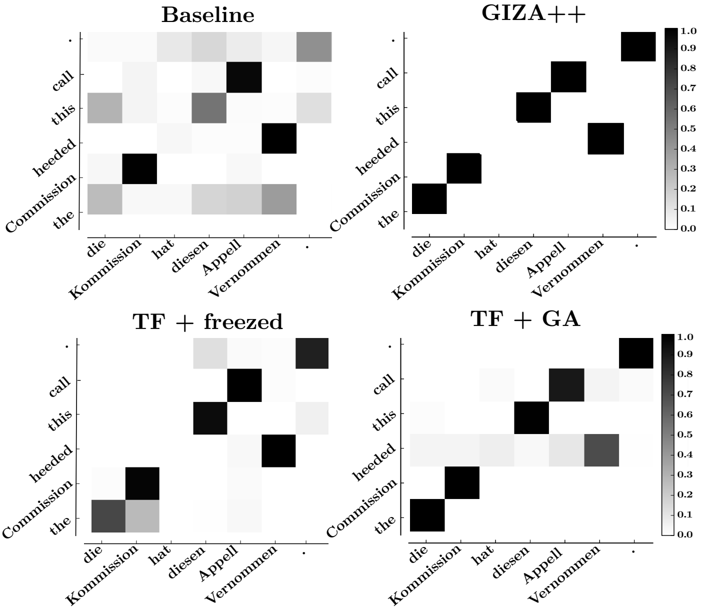
To evaluate the effectiveness of our approach we compare it to GIZA++ (Och and Ney, 2003), the BerkeleyAligner[1], `fast_align` (Dyer et al., 2013), and an unmodified attention-based model.

5.1. Setup

The translation models we use for all experiments in this work are based on the attention-based NMT approach by Bahdanau et al. (2015). We use a word-embedding size of 620 for the projection layer and a 30K shortlist of the most frequent words. The decoder and both directed RNNs of the bi-directional encoder are implemented as gated recurrent units. These RNNs as well as the attention layer have an internal dimension of 1000 nodes. For decoding, we use a beam-size of 12. Our implementation is based on the Blocks framework (Van Merriënboer et al., 2015) and the deep-learning library Theano (Bergstra et al., 2010).

To evaluate the alignment quality of our models, we use a set of 504 bilingual sentence pairs that were extracted from the Europarl (Koehn, 2005) German-to-English task and manually aligned by human annotators. We use this test set to evaluate the alignment quality on AER (Och and Ney, 2003) and SAER (Tu et al., 2016). To evaluate the soft alignment with AER, we convert it into a hard alignment by extracting the

[1] https://code.google.com/archive/p/berkeleyaligner



Baseline

GIZA++

TF + freezed

TF + GA

Figure 3. Attention weight matrices visualized in heat map form. Generated by the NMT Baseline, GIZA++, target foresight with freezed encoder and decoder parameters (TF + freezed) and target foresight with guided alignment training (TF + GA)

position with the largest alignment weight in both directions and merged them by applying Och's refined method (Och and Ney, 2003).

The network was trained on the Europarl corpus (Koehn, 2005) excluding the test set using AdaDelta (Zeiler, 2012) for learning rate adaption. Excluding the test data is done to evaluate the performance of the attention-based model on unseen data as it is the case when used for translation. It also shows that target foresight can easily be used to align unseen data without the need to retrain the model, while still outperforming traditional methods that have been trained including the test data. The training data consists of 1.2 million bilingual sentences of 32 and 34 million running words in German and English, respectively. The training is performed for 250K iterations with a batch-size of 40 and evaluated every 10K iterations. The development set of the IWSLT2013 German→English shared translation task[2] is used to select the best performing model which is then evaluated on the IWSLT2013 test as well as on the Europarl alignment test set.

[2] http://www.iwslt2013.org

Model	Alignment Test AER %	 SAER %
fast_align	27.9	33.0
GIZA++	21.0	26.8
BerkeleyAligner	20.5	26.4
Attention-Based	38.1	63.6
+ Guided alignment	29.8	38.0
+ Target foresight with fixed en-/decoder	33.9	55.6
+ Target foresight with guided alignment	**19.0**	34.9
+ converted to hard alignment	**19.0**	**24.6**

Table 1. Comparison of target foresight with the pure attention-based approach (with and without guided alignment) and other alignment methods.

5.2. Results

Table 1 shows that GIZA++ creates a far better alignment than `fast_align` and that the BerkeleyAligner creates an even slightly better result. In comparison the attention mechanism produces an AER of 17.6% worse than the BerkeleyAligner.

Interpreting the attention of the attention-based approach as an alignment results in 38.1% AER. If we train the network using guided alignment, we can reduce the AER to 29.8%.

Using the target foresight directly to create an alignment produces no usable results. The network does not learn any meaningful alignment, but uses the attention weights to encode the target word $\hat{e}_i$. It is in nearly all cases able to reproduce the target word on the output layer, even though $\hat{e}_i$ is only given to compute the alignment. Furthermore the computed alignment has no meaningful correlation with the correct alignment. To prevent this behavior, we try to make it harder to encode the target word into the attention weights, by applying noise to the alignment weights and the outputs of the corresponding network components. We also tried to initialize the encoder and decoder using the weights from our trained baseline network. We omitted these numbers since unfortunately none of these techniques gave usable results and used the following methods instead.

Fixing the encoder and decoder weights of our baseline network and training the attention layer for just additional 2000 iterations results in an improvement of 4.2% AER and 8.0% SAER.

Pairing the guided alignment training with the target foresight training yields an AER of 19.0%. This is an improvment of 10.8% compared to only using guided alignment. Compared to the BerkeleyAligner it improved by 1.5% and by 2.0% compared to GIZA++. Note the latter two still perform better considering the SAER score.

An expalaination for this behavior is that the design of SAER makes it easier for systems with hard-alignments to perform well than system using soft-alignments.

To elaborate this point: Even if the soft and the hard-alignment create the correct alignment, the soft-alignment would most likely receive a lower score since is very unlikely that it predicts the correct point with 100% certainty. Most alignment points are predicted correctly by our systems in this task. This allows the hard-alignments to produce a perfect score at most points. The soft-alignments gives these points also the highest probability, but distributes its probability mass more evenly and recives therefore a lower score than the hard-alignment.

To solve this we compute the Saer score also using the hard-alignment that we use to compute the Aer score. This gave us a corresponding Saer score that is 10.3% better than its soft equivalent. Using this comparison, the generated alignment outperforms all baseline methods on both evaluation metrics. We obtain an alignment which is superior to the baseline alignments and also to the standard guided alignment approach.

To verify that the obtained alignments can be used to improve the performance of an NMT, system we evaluate the guided alignment training on the IWSLT2013 task. We apply our NMT alignment model to produce a soft alignment for the Europarl training corpus and use it in guided alignment training. The resulting score of 18.8% Bleu was an improvement of 0.4% Bleu compared to a model trained using the GIZA++ alignment and 2.8% compared to the NMT baseline system. We also observe an improvement of 1.3% Aer.

6. Conclusion

This work shows that attention-based models are capable of generating alignments that improve the BerkeleyAligner alignments by 1.5% Aer. Using target foresight we are able to improve the Aer by 19.1% compared to the baseline attention mechanism and outperform the GIZA++ alignments by 2.0% Aer absolute and 9.5% relative using training with guided alignment. Additionally, we have shown that the new alignments can be used to improve the training of NMT models. The approach presented in this work shows also that it is possible to train one model and reuse it to align unseen data with a precision that outperforms the classical alignment methods.

Training the network to produce high quality alignments proves to be a hard task. The network seems to encode the knowledge of the target word in the attention weights and produces a non-usable alignment, but guided alignment training seems to counteract this effectively. In future work, we plan to find a way to achieve the same strong alignment without using guided alignment training.

Acknowledgements

The work reported in this paper results from two projects, SEQCLAS and QT21. SEQCLAS has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement n^{o} 694537. QT21 has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n^{o} 645452. The work reflects only the authors' views and neither the European Commission nor the European Research Council Executive Agency are responsible for any use that may be made of the information it contains.

Bibliography

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

Bergstra, James, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: A CPU and GPU math compiler in Python. In *Proc. 9th Python in Science Conf*, pages 1–7, 2010.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Comput. Linguist.*, 19(2):263–311, June 1993. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=972470.972474`.

Chen, Wenhu, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. Guided Alignment Training for Topic-Aware Neural Machine Translation. Austin, Texas, 2016. Association for Machine Translation in the Americas.

Cohn, Trevor, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. Incorporating structural alignment biases into an attentional neural translation model. *arXiv preprint arXiv:1601.01085*, 2016.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparametrization of IBM model 2. In *Proceedings of the NAACL 7th Workshop on Syntax, Semantics and Structure in Statistical Translation*, Atlanta, Georgia, USA, June 2013.

Feng, Shi, Shujie Liu, Mu Li, and Ming Zhou. Implicit Distortion and Fertility Models for Attention-based Encoder-Decoder NMT Model. *arXiv preprint arXiv:1601.03317*, 2016.

Ganchev, Kuzman, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior Regularization for Structured Latent Variable Models. *J. Mach. Learn. Res.*, 11:2001–2049, Aug. 2010. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1756006.1859918`.

Koehn, Philipp. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Em-*

pirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D15-1166`.

Mi, Haitao, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. A Coverage Embedding Model for Neural Machine Translation. *arXiv preprint arXiv:1605.03148*, 2016a.

Mi, Haitao, Zhiguo Wang, and Abe Ittycheriah. Supervised Attentions for Neural Machine Translation. *arXiv preprint arXiv:1608.00112*, 2016b.

Och, Franz Josef and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

Sankaran, Baskaran, Haitao Mi, Yaser Al-Onaizan, and Abe Ittycheriah. Temporal Attention Model for Neural Machine Translation. *arXiv preprint arXiv:1608.02927*, 2016.

Tamura, Akihiro, Taro Watanabe, and Eiichiro Sumita. Recurrent Neural Networks for Word Alignment Model. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P14-1138`.

Tu, Zhaopeng, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling Coverage for Neural Machine Translation. In *54th Annual Meeting of the Association for Computational Linguistics*, 2016.

Van Merriënboer, Bart, Dzmitry Bahdanau, Vincent Dumoulin, Dmitriy Serdyuk, David Warde-Farley, Jan Chorowski, and Yoshua Bengio. Blocks and fuel: Frameworks for deep learning. *arXiv preprint arXiv:1506.00619*, 2015.

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

Zhang, Biao, Deyi Xiong, and Jinsong Su. Recurrent Neural Machine Translation. *arXiv preprint arXiv:1607.08725*, 2016.

Address for correspondence:
Jan-Thorsten Peter
`peter@cs.rwth-aachen.de`
Human Language Technology and Pattern Recognition Group
RWTH Aachen University
Ahornstr. 55, 52056 Aachen, Germany