Baltic J. Modern Computing, Vol. 6 (2018), No. 4, pp. 403–417
http://dx.doi.org/10.22364/bjmc.2018.6.4.06

Debugging Translations of Transformer-based Neural Machine Translation Systems

Matīss RIKTERS, Mārcis PINNIS

Tilde,
Vienības gatve 75A, Riga, Latvia
matiss.rikters@tilde.lv, marcis.pinnis@tilde.lv

Abstract In this paper, we describe a tool for debugging the output and attention weights of neural machine translation (NMT) systems and for improved estimations of confidence about the output based on the attention. We dive deeper into ways for it to handle output from transformer-based NMT models. Its purpose is to help researchers and developers find weak and faulty translations that their NMT systems produce without the need for reference translations. We present a demonstration website of our tool with examples of good and bad translations: http://attention.lielakeda.lv.

Keywords: Neural Machine Translation, Attention Mechanism, Transformer Models

1. Introduction

As one of the primary use-cases for the modern computer, automated translation of texts from one language into another or machine translation (MT) has evolved vastly since its early days in the 1950s. There have been several large paradigm shifts that have greatly impacted the field of MT - rule-based MT (RBMT), statistical MT (SMT) and neural network MT (NMT) (Bahdanau et al., 2014). With each paradigm shift detailed understanding of how the system produces its final translation has changed from fully clear in the case of RBMT to slightly less, but often still predictable in SMT, to often completely unpredictable in NMT. Many current tools for inspecting results of statistical phrase-based approaches are either not compatible or serve little purpose in dealing with neural network generated output.

To address the lack of tools for inspection and analysis of NMT translations, we propose a tool for browsing, inspecting and comparing translations specifically designed for NMT output. The tool uses the attention weights that correspond to specific token pairs, which are generated during the decoding process, by turning them into one of several visual representations that can help humans better understand how the output

translations were produced. Aside from just visualising attention alignments, the tool also uses them to estimate the confidence in translation, which allows to distinguish acceptable outputs from completely unreliable ones. For this, no reference translations are required.

The structure of this paper is as follows: Section 2 summarises related work on tools for inspecting translation outputs and alignments; Section 3 introduces the key concepts of the baseline tool - how it scores translations and displays the visualisations in different environments, as well as outlines the improvements made to make it more useful for debugging machine translation output. In section 4, we give an overview of how to make the most use of our tool in finding odd translations, what to look for when comparing them and possible causes of errors. Section 5 talks about the challenges introduced by multi-layer models like transformers and section 6 - about how to deal with them. Finally, we conclude the paper in Section 7 and introduce plans for future work in the area.

2. Related Work

The foundation of our tool is based on the paper of Rikters et al. (2017), who introduce visualisation of NMT attention and use attention-based scoring of NMT as described by Rikters and Fishel (2017). While in general it can be useful to quickly find sentences with "scrambled" attention alignments, it gets more challenging when having to deal with output from multi-layer neural networks. This tends to mislead users when sorting data sets by confidence and looking for the highest scoring examples.

2.1. Attention Averaging

The general intuition is that transformer models do learn to pay more attention to specific source sentence tokens while generating translation tokens just like attentional RNNs. Since each attention head in each layer shows different results, it becomes non-trivial to decipher which one or several matrices, if any, has learnt the alignment representation. Averaging attention probabilities over all attention heads in all layers provides a solution to obtain a single attention matrix for a translated sentence.

We trained transformer and RNN NMT models using data from the highest-ranking English-Latvian system in WMT 2017 (Pinnis et al., 2017) and used both systems to translate formatting-rich documents. To compare the quality of attention averaging to the established RNN attention alignments. we performed a small-scale human evaluation on the formatting transfer between source and translated documents. The human evaluation showed that the averaged transformer alignments are just as acceptable as RNN alignments.

2.2. Guided Alignments

Chen et al. (2016) claim that translation of unknown out of vocabulary (OOV) words is linked to soft alignment dispersion and may be the source of some translation errors. To improve the alignments and the output translations, the authors propose to use the IBM model 4 Viterbi alignments as additional input data during training. They experiment with adding alignments produced by *GIZA++* to RNN-based NMT systems. The authors report improvements in alignment distributions as well as overall translation quality.

Liu et al. (2016) also attempt to improve attention alignments produced by RNN-based NMT systems. In addition to *GIZA++* alignments, they experiment with *fast_align* and add several heuristics. The authors report that the significantly faster *fast_align* generates slightly lower quality alignments and they improve NMT output quality and soft alignments just as well.

We did a similar experiment as with the averaging by training an NMT system with guided alignments (*fast_align*) and translating formatted documents to perform human evaluation. The evaluation showed that the model with guided alignments is able to transfer document formatting slightly better than the averaged alignments and RNN attention alignments.

3. Visualisation Tool

The basis of our visualisation tool is described in full detail in the baseline paper (Rikters et al., 2017). It requires source and translated sentences along with the corresponding attention alignments from NMT systems as input files and can provide a visual overview in a command line environment (Linux Terminal or Windows Powershell) or a web browser of any modern device. It is published in a GitHub repository[1] and open-sourced with the MIT License. In the further subsections of the paper, we will outline only core components and focus more on highlighting improvements and differences.

In addition to Nematus, Neural Monkey and Marian[2] (Junczys-Dowmunt et al., 2018), we have also added out-of-the-box support for working with attention alignments from OpenNMT and Sockeye[3] (Hieber et al., 2017) frameworks.

3.1. Confidence Scores

This section outlines how the confidence scores are calculated and outlines what is how the final score differs from the baseline.

The four main metrics that we use for scoring translations are:

– **Coverage Deviation Penalty** (CDP) penalises attention deficiency and excessive attention per input token.

$$\mathrm{CDP} = -\frac{1}{L_s}\sum_j \log\left(1 + \left(1 - \sum_i \alpha_{ji}\right)^2\right) \tag{1}$$

– **Absentmindedness Penalties** ($\mathrm{AP}_{\mathrm{out},\ \mathrm{in}}$) penalise output tokens that pay attention to too many input tokens, or input tokens that produce too many output tokens.

$$\mathrm{AP}_{out} = -\frac{1}{L_s}\sum_i \sum_j \alpha_{ji} \cdot \log \alpha_{ji} \tag{2}$$

$$\mathrm{AP}_{in} = -\frac{1}{L_s}\sum_j \sum_i \alpha_{ij} \cdot \log \alpha_{ij} \tag{3}$$

[1] NMT Attention Alignment Visualisations: `https://github.com/M4t1ss/SoftAlignments`

[2] Marian: `https://github.com/marian-nmt/marian`

[3] Sockeye: `https://github.com/awslabs/sockeye`

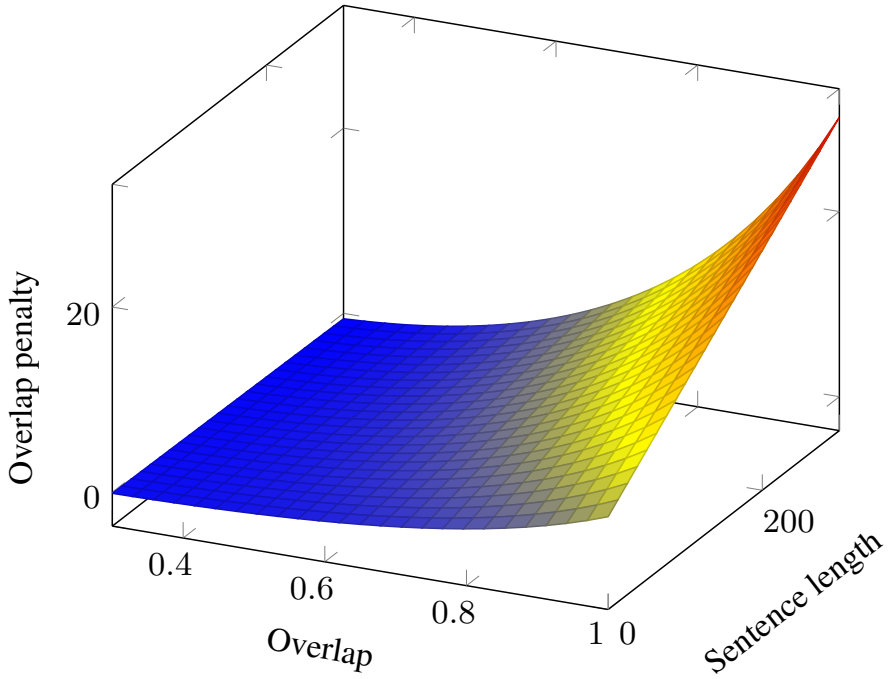


Figure 1. A plot of the overlap penalty.

– **Overlap Penalty** (OP) penalises translations that copy large fractions from source sentences. A stronger penalty is allocated to longer sentences that copy large amounts from the source while shorter ones get more tolerance (e.g., the three-word English sentence "Thanks Barack Obama." can be perfectly translated into "Paldies Barack Obama." although 2/3 of words in the translation are the same in the source). A plot of how the penalty increases in relation to the source-translation overlap and source sentence length is shown in Figure 1.

$$\text{OP} = (0.8 + (L_t * 0.01)) * (3 - ((1 - S) * 5)) * (0.7 + S) * tan(S) \quad (4)$$

– **Confidence** is the sum of the three main metrics – CDP, AP_{in} and AP_{out} and the similarity penalty, when the similarity between input and output sentences is high (similarity > 0.3) .

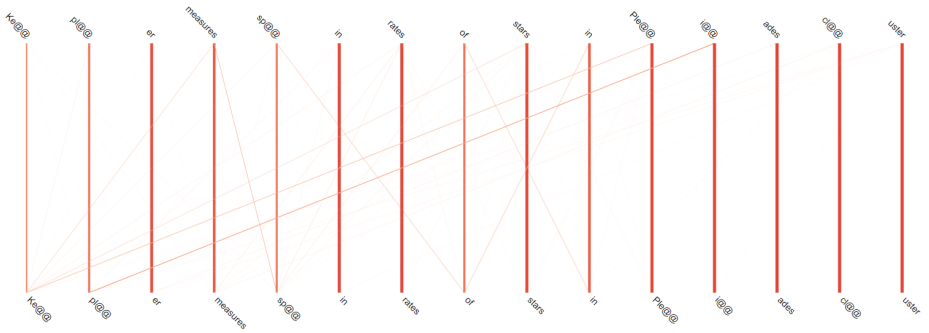
$$confidence = \begin{cases} CDP + AP_{out} + AP_{in}, & \text{if } similarity < 0.3 \\ CDP + AP_{out} + AP_{in} - OP, & \text{otherwise} \end{cases} \quad (5)$$

In all of the metrics L_s is the length of the source sentence; L_t - length of the target sentence; S - similarity between the source sentence and the translation on the scale of 0 - 1; α_{ji} - the attention weight between source token *i* and translation token *j*.

Changes have been introduced to the final confidence score by first calculating the similarity ratio between input and output sentences and then adding a further penalty

only if the similarity is high enough. The similarity is calculated by finding the longest contiguous matching sub-sequence.

Since the baseline confidence score considered only the attention alignments when calculating the final value, examples like shown in Figure 2 received particularly high values due to consistent one-to-one attention alignments. The updated score takes care of this problem by penalising hypothesis sentence that is overly similar to the input source.



Source: Kepler measures spin rates of stars in Pleiades cluster
Hypothesis: Kepler measures spin rates of stars in Pleiades cluster
Reference: Keplers izmēra zvaigžņu griešanās ātrumu Plejādes zvaigznājā.

Figure 2. An example of a translated sentence that exhibits a verbatim rendition of the input. CDP: 100.0%; AP_{out}: 98.84%; AP_{in}: 98.85%; Baseline Confidence: **95.44%**; Updated Confidence: **25.02%**;

3.2. Web Interface

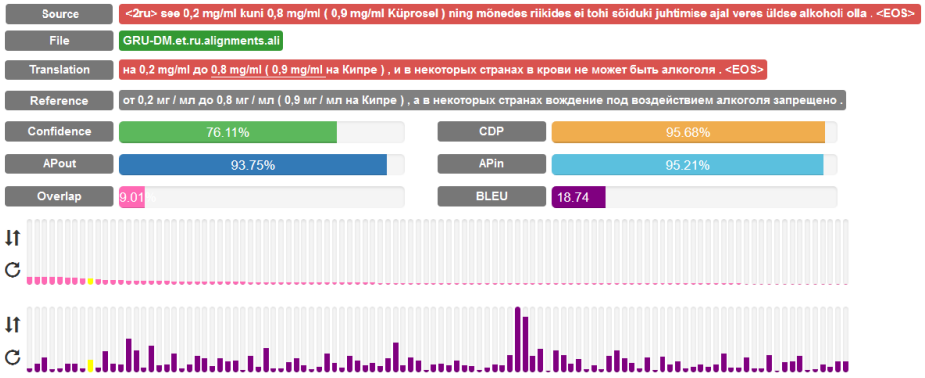
The web interface is the primary point of interaction with the tool. Aside from browsing visualisations, ordering data sets by confidence scores and exporting visualisations as images, that are all clarified in the baseline paper, we introduce several significant changes to the system. The first one is a technical update on how data is served — loading is performed asynchronously in the background and thereby eliminating long wait times to view the proceeding sentences in a large data set. The three major additions are:

- the addition of source-translation overlap percentage alongside the four base scores (Section 3.3);
- the ability to provide reference translations, if available, to display next to the hypothesis and calculate BLEU scores (Section 3.4);
- the ability to directly compare translations and alignments from two different NMT systems (Section 3.5).

3.3. Overlap

As mentioned in Section 3.1, the updated confidence score considers hypotheses translations that are long and have a significant overlap with the source sentence as a worse

translations, while tolerating considerable overlap for shorter sentences. In addition to contributing to the final confidence score, the overlap ratio has been added as an individual score for sorting, navigating and comparing sentences from a data set as shown in Figure 3. The system also underlines the longest matching sub-string between the source and translation in cases where the overlap is high enough (over 10%). An example is shown in Figure 3, where the overlap ratio is 20.19%.



Source: see 0,2 mg/ml kuni 0,8 mg/ml (0,9 mg/ml Küprosel) ning mõnedes riikides ei tohi sõiduki juhtimise ajal veres üldse alkoholi olla.

Hypothesis: на 0,2 mg/ml до 0,8 mg/ml (0,9 mg/ml на Кипре) , и в некоторых странах в крови не может быть алкоголя.

Match: 0,8 mg/ml (0,9 mg/ml

Figure 3. An example translation from Estonian into Russian, showing useful features for debugging translation outcomes - underlining of the longest matching sub-string between the source and translated sentences; sorting translations by overlap (pink bars) or BLEU score (purple bars); reference translation (grey background).

3.4. References and BLEU

We believe that simply displaying the reference next to the hypothesis is helpful more often than not. Having provided references also allows to calculate BLEU scores for the translations, providing yet another dimension for sorting (Figure 3). Unlike overlap, the BLEU scores do not influence the overall confidence scores.

Both overlap and BLEU score calculation and output has also been added to the terminal interface of the tool (Figure 4).

3.5. Comparing Translations

The final major addition to the tool is the option to directly compare two translations of the same source sentence. To perform the comparison, all source sentences for both input data sets must match, but the target sentences may differ in output token order as well as count. Comparisons may be performed between translations obtained from any two of the five currently supported NMT frameworks (Nematus, Neural Monkey,

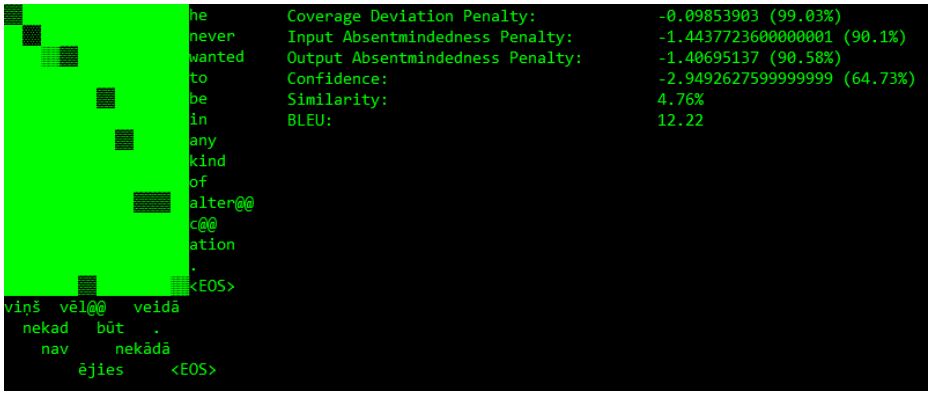


Figure 4. An example of the updated terminal interface output.

OpenNMT, Marian and Sockeye) or even an arbitrary input file, as long as it's formatted according to the specification provided in the readme [4].

Figure 5 shows an example comparison of a sentence translated by two different NMT systems. On the top row is the source text and the bottom rows represent output from each individual NMT system colour-coded to match the colours of the alignment lines. The second hypothesis (in green) exhibits stronger and more reliable output alignments to the content words while the first shows strong alignments coming from the stop sign. In this example neither hypothesis matches the reference, but since it is only two words long for a source sentence of triple the length, it can hint to an oversimplified translation by the translator (assuming English was the original) and does not mean that both hypotheses are completely wrong. In fact, the second hypothesis is a fairly decent representation of the source sentence.

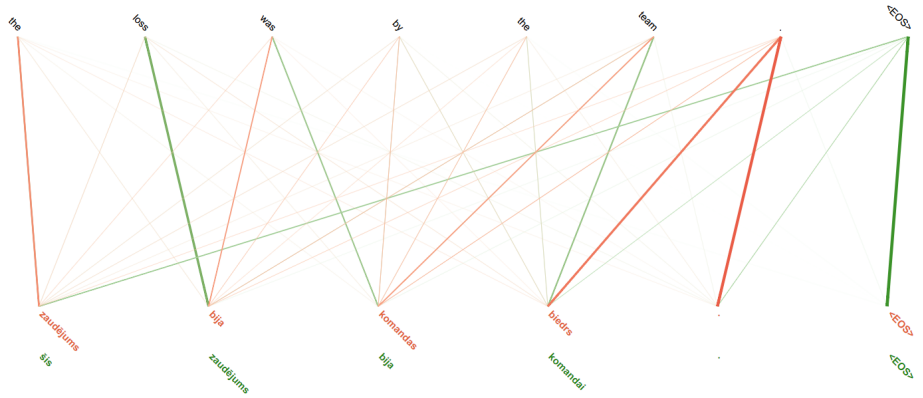
Figure 6 illustrates another example with strong attention alignments and a high overlap ratio (94.03%) between source and translated sentences from one system compared to a weak, but at least better translation from another system. The final confidence score for the second translation is strongly influenced by the high overlap, even though the sentence is not particularly long. In similar conditions, the confidence score of the second hypothesis calculated by the baseline system would be very close to 100% due to its complete disregard for the actual words of the source and hypothesis sentences.

4. Recipes for Debugging

In this section, we summarise several tips and tricks that may come in handy when using the tool to look for faulty translations of various kinds. Here we also list common causes associated with the problems. Some peculiarities to pay attention to may include:

– **Short sentences with a low confidence, CDP, $\mathbf{AP}_{in}$ or $\mathbf{AP}_{out}$**
All of the metrics do not necessarily need to be low, but translations that exhibit at least one of them to be under 30% are often worth looking into.

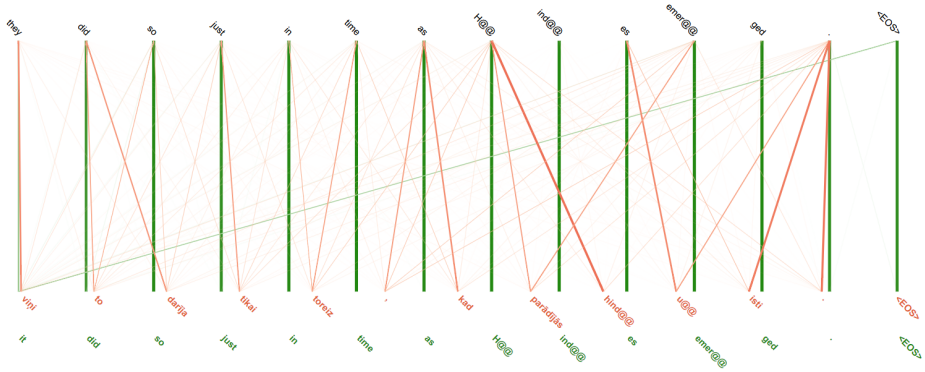
[4] Using other input formats - `https://github.com/M4t1ss/SoftAlignments\#how-to-get-alignment-files-from-nmt-systems`



Source: the loss was by the team.
Hypothesis 1: zaudējums bija komandas biedrs.
Hypothesis 2: šis zaudējums bija komandai.
Reference: zaudē komanda.

Figure 5. A direct comparison of attention alignments for translating the same sentence with two different NMT systems.

– **Long sentences with a high overlap**
As stated before, for short, several words long sentences it may be completely normal to have an overlap of 50% or more, but if it occurs in sentences that are 10 or more words long, it may indicate that the system has only partially translated the



Source: they did so just in time as Hindes emerged.
Hypothesis 1: viņi to darīja tikai toreiz , kad parādījās hinduisti.
Hypothesis 2: it did so just in time as Hindes emerged.
Reference: viņiem tas izdevās pēdējā brīdī.

Figure 6. A comparison of lower and higher scoring hypotheses from two different NMT systems. Scores for Hypothesis 1 (orange): Confidence **53.1%**; Overlap **0.9%**. Scores for Hypothesis 2 (green): Confidence **28.63%**; Overlap **94.03%**.

source or not translated anything at all. When completely untranslated sentences are found, it is worth checking the training data for any source-target sentence pairs that are equal. Removing them from the training data will reduce cases where words or even full sentences are left untranslated.

– **Sentences with a low BLEU score, but normal or even high confidence, CDP, $\mathbf{AP}_{in}$ and $\mathbf{AP}_{out}$**
The BLEU metric has its flaws and one of them is comparing each hypothesis to only one reference, although many sentences can actually have more than one correct translation. In cases when the only low-scoring metric output by the tool is the BLEU score, it is often that the translation is perfectly good, but just different from the reference. Such sentences are often useful examples to show that lower BLEU scores of neural MT systems do not necessarily represent lower quality translations and are cheaper to find than performing full manual human evaluation.

5. Dealing with Output from Multi-layer Models

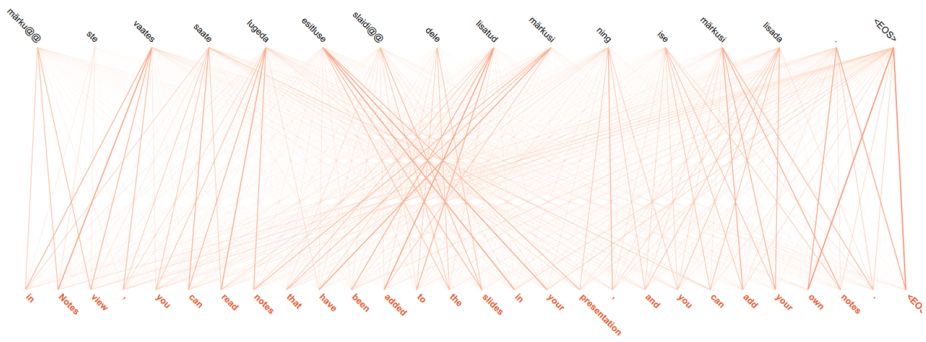


Figure 7. An example of attention alignments from a 15-layer encoder and 15-layer decoder convolutional neural machine translation system trained with FairSeq.

An ongoing challenge is to find a way of how to better acquire attention alignments generated by multi-layer neural networks. While in recurrent neural network NMT systems this is rarely a problem, more modern approaches like convolutional neural networks (Gehring et al., 2017) and transformer neural networks (Vaswani et al., 2017) require training of deeper models to achieve translation results of competitive quality. This, however, results in uncertainties of how to interpret attention weights, including whether they encompass reliable alignment information. Even when all attention matrices are summed up, the result looks like every source token is connected to every hypothesis token as can be seen in Figure 7.

Out of all modern NMT approaches that are built as deep multi-layer neural networks, the transformer-based NMT systems currently achieve state-of-the-art translation quality results for most language pairs (as shown by the results of the WMT shared task for news translation (Bojar et al., 2018)). Therefore, we chose to investigate how

they work and how the attention information can be made useful for debugging output translations.

5.1. Transformer Models

Vaswani et al. (2017) proposed a novel neural network architecture, the Transformer, which relies only on the attention mechanism to draw global dependencies between input and output. It has an encoder-decoder structure using multiple stacked self-attention and point-wise, fully connected layers for both the encoder and decoder. One of the big advantages of training self-attentional models is that they are highly parallelizable, as they do not employ the recurrent connections of recurrent neural networks (RNNs).

A typical transformer model would consist of six layers of which each would consist of eight attention heads. This means that there are 48 source-to-target attention matrices that the neural network can use for translation purposes.

Aside from visualising and interpreting NMT output, attention alignments are also used to get hard word alignments in order to correctly translate structured documents and reconstruct the structure after translating (Pinnis et al., 2018b). To achieve similar results with transformer-based NMT models, several approaches have been explored, such as learning guided alignments [5], averaging attention matrices [6] and using *fast_align* (Dyer et al., 2013) to generate alignments after the translation has been produced (e.g., Pinnis et al. (2018a)). The latter approach is of no use for interpreting NMT output as it uses a separate model and only attempts to guess what the alignments are after the result has been produced. The other two are worth looking into.

6. Experiments and Results

We used the previously mentioned averaged transformer and guided alignment transformer models to determine, which approach is better suited for our debugging tool to quickly identify faulty and suspicious translations. Both models were trained using data from the Tilde's unconstrained submission to the WMT 2017 shared task on news translation (Pinnis et al., 2017). The averaged transformer model was trained using the Sockeye NMT toolkit on the fully processed (including factorisation and morphology-driven word splitting) dataset of the Tilde's unconstrained English-Latvian submission (46.04 million sentence pairs in total). The guided alignment model was trained using the Marian toolkit on the same dataset, but without factorisation and without morphology driven word splitting (only byte-pair encoding was performed). Both models were trained until convergence and reached about the same quality on news domain (WMT17 development set) and general domain data (ACCURAT development corpus (Skadiņa et al., 2012)).

In order to determine the usefulness, we aimed to answer two main questions - 1) do the resulting attention alignments represent actual relations between the source sentence and the output translation, and 2) do the confidence scores produce similar results using these alignments. Regarding the first question, it is important to understand whether the alignments from transformer models for high-quality translations actually represent word-by-word source-translation alignments and/or relevant phrases. As for

[5] Make guided alignment work with transformer - `https://git.io/fx5zy`

[6] Transformer Attention Probabilities - https://github.com/awslabs/sockeye/pull/504

the second question - even if the first one is not fully confirmed, the alignments may still be useful for finding translation errors. Therefore, if the resulting transformer attention alignments help in producing distinct and sortable confidence scores, they will be considered useful.

Table 1. NMT system quality on news (NewsTest 2017) and general (ACCURAT) data.

	BLEU	
	News	General
Attention Averaging	21.96	39.15
Guided Alignments	22.05	39.20

6.1. Attention Averaging

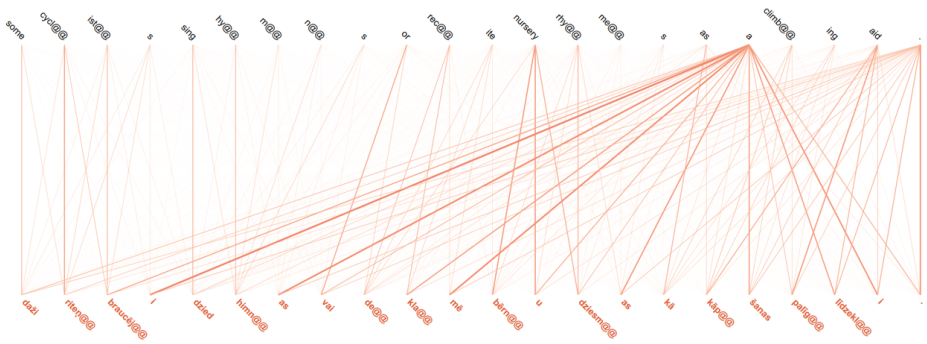
Figure 8 exhibits rather dispersed attention alignments for an acceptable translation. A significant amount of the attention is focused on the stop mark in the end of the source sentence and even more on the word “a”, which clearly should not be connected to so many output tokens. Such a distribution of attention alignments for RNN-based models would indicate that the model had problems translating some or most input tokens and an unsuccessful translation had been produced. During manual inspection, we noticed that most results exhibit similar outputs by having an excessive amount of attention focused to one or two source tokens, while the translations themselves were good. This indicates that the answer to the first question is negative.

To answer the second question, we sorted the test set of 2000 sentences by each of the confidence scores and looked for low-scoring and relatively short sentences. All scores exhibited a large number of false-positives, mainly due to dispersed attention alignments. Such behaviour means that the attention-based scores that are computed from transformer models with attention averaging cannot aid in finding poorly translated sentences.

6.2. Guided Alignments

The top part of Figure 9 shows how the same sentence is translated with the system that was trained using guided alignments. In this example, the translation is noticeably worse, but the alignment lines in the visualisation are much stronger and less scattered. The computed confidence score of 48.65% seems fairly adequate, as the sentence-level BLEU score is also quite low - 12.69. This leads to believe that the learned alignments do a better job in representing relations between source and translated tokens, answering positively to the first question. The example shows that the attention is mainly dispersed in places where words are split in subword units (ending with ‘@@’).

To see how attention alignments change after joining subword units and the respective attentions into full words, we summed attention weights over source subword units and averaged attention weights over target subword units. The soft attention alignments acquired with this method for the same sentence can be seen in the lower part of Figure 9. As expected, the alignments became stronger and it also improved the confidence



Source: Some cyclists sing hymns or recite nursery rhymes as a climbing aid.
Hypothesis: Daži riteņbraucēji dzied himnas vai deklamē bērnu dziesmas kā kāpšanas palīglīdzekli.
Reference: Daži riteņbraucēji dzied himnas vai skaita bērnudārza dzejoļus, lai vieglāk tiktu kalnā.

Figure 8. Attention alignment example of a translation from English into Latvian with a transformer-based NMT model and attention averaging.

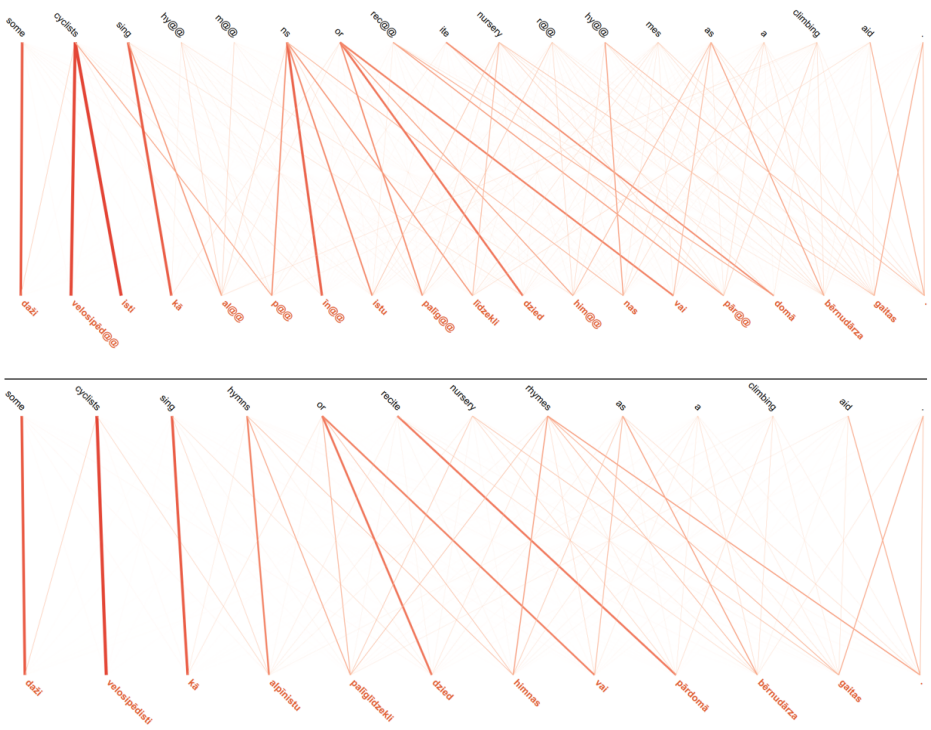
scores. This allowed to better single out several of the very-worst translations of the set when sorting it by AP_{out} and CDP.

7. Conclusion

In this paper, we described how our visual NMT debugging tool handles output from multi-layer neural networks, such as the recent and very popular Transformer models. We explored two scenarios of preparing attention alignments from transformer-based NMT to be compatible with our tool. We found that the guided alignment training strategy yields the best results for quickly locating better and worse translations in arbitrary test sets. Compared to other similar tools, ours relies on the confidence scores and does not require reference translations to facilitate this easier navigation, but it only benefits with additional features that are enabled when the references are provided. This allows to integrate it, for example, in an NMT system with a web-based interface, providing users with an explanation for the result of a specific translation.

In a future version of the system, we plan to include other reference-based MT scoring metrics for more variety of scoring and sorting. Some examples of metrics may include chrF (Popović, 2015) or TER (Snover et al., 2006). Another idea for future work would be to list and order specific best, worst or interesting examples of translations. This could be done by considering the recipes from Section 4.

In addition to the reference-based metrics, there exist other reference-less approaches yet to be utilised. For instance, borrowing ideas from parallel corpora filtering (Pinnis et al., 2017), such as 1) source-hypothesis sentence length difference; 2) language identification for the hypothesis; 3) digit mismatch between the source and hypothesis; 4) foreign or corrupt symbol checking for the hypothesis.



Source:	Some cyclists sing hymns or recite nursery rhymes as a climbing aid.
Hypothesis:	Daži velosipēdisti kā alpīnisma palīglīdzekli dzied himnas vai pārdomā bērnudārza gaitas.
Reference:	Daži riteņbraucēji dzied himnas vai skaita bērnudārza dzejoļus, lai vieglāk tiktu kalnā.

Figure 9. Attention alignment example of a translation from English into Latvian with a transformer-based NMT model trained using guided alignments. Attentions displayed on the subword level (top) and word level (bottom). Confidence: 48.65%; CDP: 85.78%; AP_{out}: 84.09%; AP_{in}: 88.78% (top), Confidence: 66.73%; CDP: 99.96%; AP_{out}: 90.10%; AP_{in}: 90.92% (bottom).

8. Acknowledgements

The research has been supported by the European Regional Development Fund within the research project “Neural Network Modelling for Inflected Natural Languages” No. 1.1.1.1/16/A/215.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (wmt18).

In *Proceedings of the Third Conference on Machine Translation*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Chen, W., Matusov, E., Khadivi, S., and Peter, J.-T. (2016). Guided alignment training for topic-aware neural machine translation. *AMTA 2016, Vol.*, page 121.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 644–648, Atlanta, USA.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2017). Sockeye: A Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Liu, L., Utiyama, M., Finch, A., and Sumita, E. (2016). Neural machine translation with supervised attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3093–3102.

Pinnis, M., Krišlauks, R., Miks, T., Deksne, D., and Šics, V. (2017). Tilde's machine translation systems for wmt 2017. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 374–381, Copenhagen, Denmark. Association for Computational Linguistics.

Pinnis, M., Rikters, M., and Krišlauks, R. (2018a). Tilde's Machine Translation Systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation (WMT 2018), Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Pinnis, M., Skadiņš, R., Šics, V., and Miks, T. (2018b). Integration of neural machine translation systems for formatting-rich document translation. In *International Conference on Applications of Natural Language to Information Systems*, pages 494–497. Springer.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Rikters, M. and Fishel, M. (2017). Confidence through attention. In *Proceedings of The 16th Machine Translation Summit*.

Rikters, M., Fishel, M., and Bojar, O. (2017). Visualizing neural machine translation attention and confidence. *The Prague Bulletin of Mathematical Linguistics*, 109(1):39–50.

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiș, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M. L., and Pinnis, M. (2012). Collecting and using comparable corpora for statistical machine translation. In Calzolari, N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 438–445, Istanbul, Turkey. European Language Resources Association (ELRA).

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200/6. Citeseer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Received November 8, 2018 , accepted December 13, 2018