Debugging Neural Machine Translations

Matīss Rikters

Tilde
Vienības gatve 75A, Riga, Latvia, LV-1004
{matiss.rikters}@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. The purpose of the tool is to help researchers and developers find weak and faulty example translations that their NMT systems produce without the need for reference translations. Our tool also includes an option to directly compare translation outputs from two different NMT engines or experiments. In addition, we present a demo website of our tool with examples of good and bad translations: http://attention.lielakeda.lv.

Keywords: Neural machine translation, Visualization tool, Attention mechanism

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) [2]. 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 of the existing tools for inspecting results of statistical phrase-based approaches are either not compatible or serve little purpose in dealing with neural network generated output.

In this paper, 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 visualizing 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.

Lupeikiene A., Matulevičius R., Vasilecas O. (eds.):
Baltic DB&IS 2018 Joint Proceedings of the Conference Forum and Doctoral Consortium.

The structure of this paper is as follows: Section 1.1 summarizes related work on tools for inspecting translation outputs and alignments; Section 2 introduces some concepts of the baseline tool - how it scores translations and displays the visualizations in different environments, as well as outlines the improvements made to make it more useful for debugging machine translation output. In section 3 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. Finally, we conclude in Section 4 and introduce plans for directions of future work and research in the area.

1.1 Related Work

The foundation of our tool is based on the paper of Rikters et al. [13], who introduce visualization of NMT attention and use attention-based scoring of NMT as described by Rikters and Fishel [14]. While in general it can be useful to quickly find sentences with “scrambled" attention alignments, it does have several flaws like considering completely untranslated sentences as good. This consistently misleads users when sorting data sets by confidence and looking for the highest scoring examples. Another shortcoming is the ability to only visualize a translation from one system at a time, making it slightly tricky to directly compare how multiple systems handle the same inputs.

In contrast, both iBLEU [9] — a web-based tool for visualizing BLEU [10] scores — and MT-ComparEval [8] — which builds upon iBLEU by adding supplementary visualizations, scores and metrics — can easily work with multiple MT outputs and even a set of human references. A downside for these tools is that the reference translation set is always mandatory and can't be left out. While it is useful to verify how the system performs in a controlled environment (when the expected result - reference translations - is known beforehand), more often than not the strangest abnormalities appear when using arbitrary data.

NMT frameworks like Nematus [15], Neural Monkey [4] or OpenNMT [7] have some forms of visualization, but they mainly handle representation of the translation process instead of the translation results. For instance, OpenNMT has a separate repository for visualization tools[1] that can generate visualizations of embeddings or beam search. Neural Monkey utilizes the built-in visualizations of TensorFlow [1] that can show the compute graph and multiple types of histograms from the training progress.

2 Visualization Tool

The basis of our visualization tool is described in full detail in the baseline paper [13]. 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

[1] VisTools - https://github.com/OpenNMT/VisTools

repository[2] 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[3] [6], we have also added out-of-the-box support for working with attention alignments from OpenNMT and Sockeye[4] [5] frameworks.

2.1 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) penalizes attention deficiency and excessive attention per input token.

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

- Absentmindedness Penalties ($\text{AP}_{\text{out}},\ _{\text{in}}$) penalize output tokens that pay attention to too many input tokens, or input tokens that produce too many output tokens.

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

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

- Overlap Penalty (OP) penalizes 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).

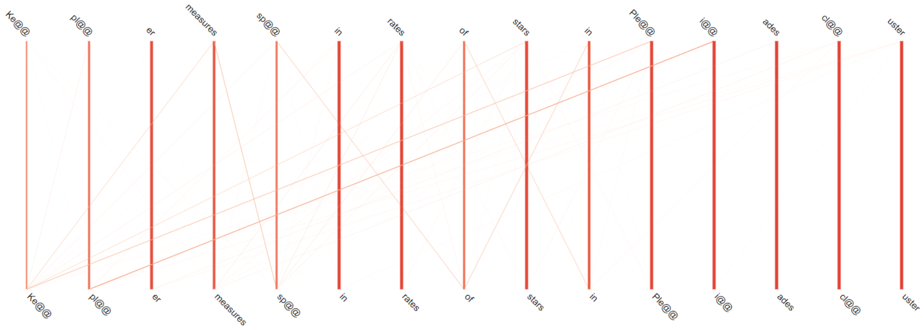
$$\text{OP} = (0.8 + (L_t * 0.01)) * (3 - ((1 - S) * 5)) * (0.7 + S) * tan(S) \tag{4}$$

- 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) .

[2] NMT Attention Alignment Visualizations: https://github.com/M4t1ss/SoftAlignments

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

[4] Sockeye: https://github.com/awslabs/sockeye



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 1. 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%;

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

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 subsequence.

Since the baseline confidence score considered only the attention alignments when calculating the final value, examples like shown in Figure 1 received particularly high values due to consistent one-to-one attention alignments. The updated score takes care of this problem by penalizing hypothesis sentence that is overly similar to the input source.

2.2 Web Interface

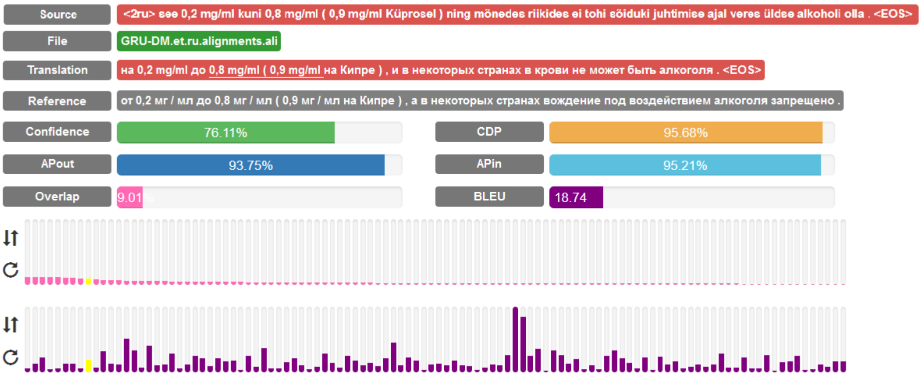
The web interface is the primary point of interaction with the tool. Aside from browsing visualizations, ordering data sets by confidence scores and exporting visualizations as images, that are all clarified in the baseline paper, we introduce several significant changes to the system. The first one 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 2.3);

- the ability to provide reference translations, if available, to display next to the hypothesis and calculate BLEU scores (Section 2.4);
- the ability to directly compare translations and alignments from two different NMT systems (Section 2.5).

2.3 Overlap

As mentioned in Section 2.1, 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 2. The system also underlines the longest matching substring between the source and translation in cases where the overlap is high enough (over 10%). An example is shown in Figure 2, where the overlap ratio is 20.19%.



Source: <2ru> 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 . <EOS>

File: GRU-DM.et.ru.alignments.ali

Translation: 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 . <EOS>

Reference: от 0,2 мг / мл до 0,8 мг / мл (0,9 мг / мл на Кипре) , а в некоторых странах вождение под воздействием алкоголя запрещено .

Confidence: 76.11%

CDP: 95.68%

APout: 93.75%

APin: 95.21%

Overlap: 9.01

BLEU: 18.74

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

2.4 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 2). Unlike overlap, the BLEU scores do not influence the overall confidence scores.

2.5 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, OpenNMT, Marian and Soceye) or even an arbitrary input file, as long as it's formatted according to the specification provided in the readme [5].

Figure 3 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 color-coded to match the colors 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.

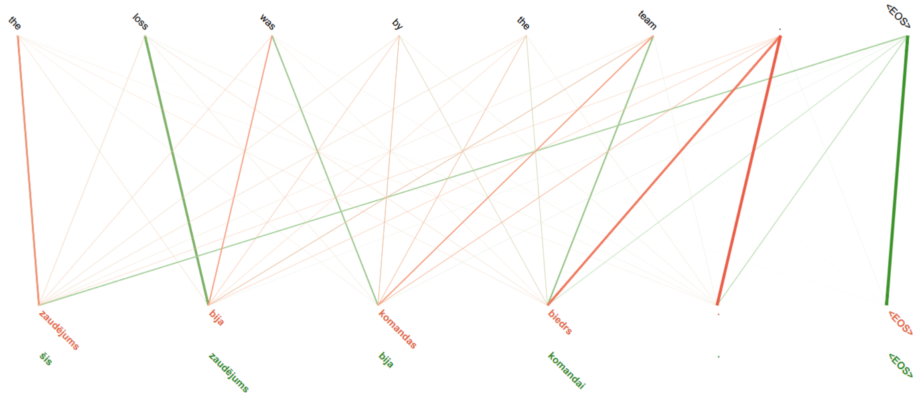
Figure 4 illustrates another example with strong attention alignments and a high overlap ratio (94.03%) between source and translated sentences from one system compared to a weak, but at least better translation from another system. The final confidence score for the second translation is strongly influenced 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.

3 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, AP_{in} or 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.

[5] 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.

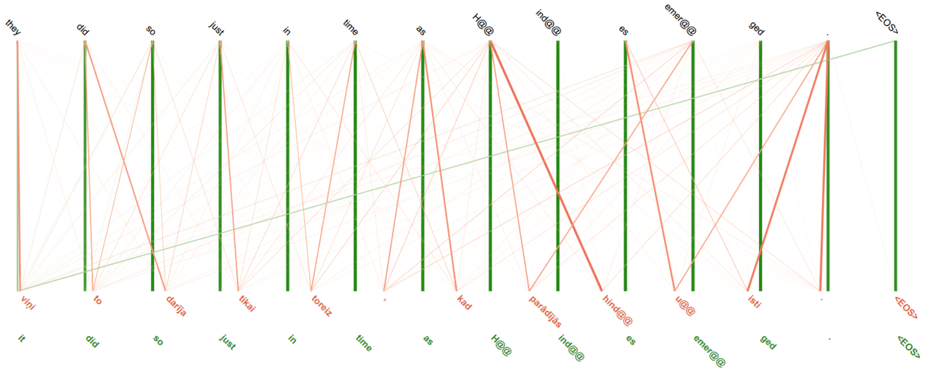
Figure 3. A direct comparison of attention alignments for translating the same sentence with two different NMT systems.

- 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 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 should help.
- Sentences with a low BLEU score, but normal or even high confidence, CDP, AP$_{in}$ and AP$_{out}$

  The BLEU metric has its flaws and one of them is comparing each hypothesis to only one reference, while it is often possible to translate the same sentence in several different ways. 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 evaluations.

A separate recommendation specifically for comparing two translations is to look at the attention alignment lines and try to find ones with source tokens having strong alignments to different hypothesis tokens, while maintaining relatively similar confidence scores. Such translations are often synonyms.



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 4. 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%.

4 Conclusion

In this paper, we described our conversion of a visualization tool into an instrument for debugging output form neural machine translation systems by improving the attention alignment scoring and confidence estimation of the baseline. The tool is intended to help researchers better understand how their systems perform by enabling to quickly locate better and worse translations in a 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 interface, providing users with an explanation for the result of a specific translation.

In a future version of the system we may include other reference-based MT scoring metrics for more variety of scoring and sorting. Some examples of metrics may include chrF [12] or TER [16]. 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 3.

In addition to the reference-based metrics, there still are some referenceless approaches yet to be utilised. For instance, borrowing ideas from parallel corpora filtering [11] 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.

Another ongoing challenge is to find a way of better representing attention alignments generated by multi-layer neural networks. While in recurrent neural network NMT systems this is rarely a problem, more modern approaches like convolution neural networks [3] and transformer neural networks [17] require training of deeper models to achieve competitive quality translation results. This, however, results in each layer paying attention only to a subset of the input sentence. Even when all attentions are summed up, the result looks like every source token is connected to every hypothesis token as can be seen in Figure 5.

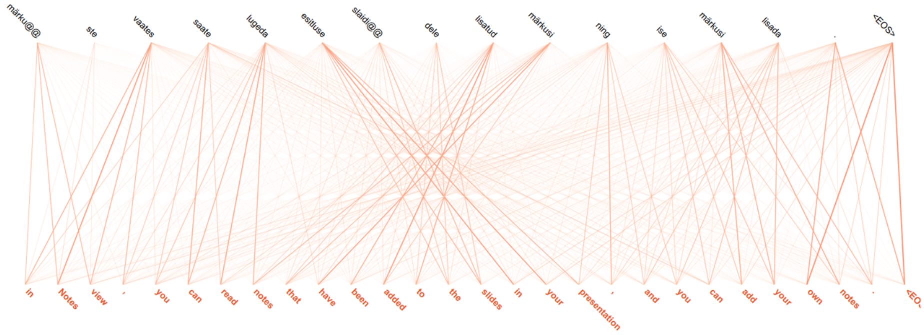


Figure 5. An example of attention alignments from a 15-layer encoder and 15-layer decoder convolutional neural machine translation system trained with FairSeq.

5 Acknowledgments

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

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473 (2014), http://arxiv.org/abs/1409.0473
3. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. arXiv preprint arXiv:1705.03122 (2017)
4. Helcl, J., Libovický, J.: Neural Monkey: An open-source tool for sequence learning. The Prague Bulletin of Mathematical Linguistics pp. 5–17 (2017). https://doi.org/10.1515/pralin-2017-0001
5. Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., Post, M.: Sockeye: A Toolkit for Neural Machine Translation. ArXiv e-prints (Dec 2017), https://arxiv.org/abs/1712.05690

6. 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., Birch, A.: Marian: Fast neural machine translation in c++. arXiv preprint arXiv:1804.00344 (2018), https://arxiv.org/abs/1804.00344
7. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M.: OpenNMT: Open-Source Toolkit for Neural Machine Translation. ArXiv e-prints (2017)
8. Klejch, O., Avramidis, E., Burchardt, A., Popel, M.: Mt-compareval: Graphical evaluation interface for machine translation development. The Prague Bulletin of Mathematical Linguistics 104(1), 63–74 (2015)
9. Madnani, N.: ibleu: Interactively debugging and scoring statistical machine translation systems. In: Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on. pp. 213–214. IEEE (2011)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. . . . of the 40Th Annual Meeting on . . . pp. 311–318 (2002). https://doi.org/10.3115/1073083.1073135, http://dl.acm.org/citation.cfm?id=1073135
11. Pinnis, M., Krišlauks, R., Miks, T., Deksne, D., Šics, V.: Tilde's machine translation systems for wmt 2017. In: Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers. pp. 374–381. Association for Computational Linguistics, Copenhagen, Denmark (September 2017), http://www.aclweb.org/anthology/W17-4737
12. Popović, M.: chrf: character n-gram f-score for automatic mt evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395 (2015)
13. Rikters, M., Fishel, M., Bojar, O.: Visualizing neural machine translation attention and confidence. The Prague Bulletin of Mathematical Linguistics 109(1), 39–50 (2017)
14. Rikters, M., Fishel, M.: Confidence through attention. In: Proceedings of The 16th Machine Translation Summit (2017)
15. Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Barone, A.V.M., Mokry, J., et al.: Nematus: a toolkit for neural machine translation. EACL 2017 p. 65 (2017)
16. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of association for machine translation in the Americas. vol. 200. Citeseer (2006)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR abs/1706.03762 (2017), http://arxiv.org/abs/1706.03762