**Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 11, No. 1, 2023, 95-102.**

Shahrood University of Technology

**Journal of Artificial Intelligence and Data Mining (JAIDM)**
Journal homepage: http://jad.shahroodut.ac.ir

**Research paper**

# Automatic Post-editing of Hierarchical Attention Networks for Improved Context-aware Neural Machine Translation

Mohammad Mehdi Jaziriyan and Foad Ghaderi*

*Human-Computer Interaction Lab., Faculty of Electrical and Computer Engineering Tarbiat Modares University, Tehran, Iran.*

**Article Info**

**Article History:**
*Received 28 July 2022*
*Revised 02 October 2022*
*Accepted 14 November 2022*

*DOI:10.22044/jadm.2022.12152.2367*

**Keywords:**
*Context-Aware Neural Machine Translation, Document-Level Neural Machine Translation, Neural Machine Translation*

**Corresponding author: fghaderi@modares.ac.ir (F. Ghaderi).*

## Abstract

Most of the existing neural machine translation (NMT) methods translate sentences without considering the context. It is shown that exploiting inter- and intra-sentential context can improve the NMT models and yield to better overall translation quality. However, providing document-level data is costly, so properly exploiting contextual data from monolingual corpora would help translation quality. In this work, we propose a new method for context-aware neural machine translation (CA-NMT) using a combination of hierarchical attention networks (HAN) and automatic post-editing (APE) techniques to fix discourse phenomena when there is lack of context. HAN is used when we have a few document-level data, and APE can be trained on vast monolingual document-level data to improve the results further. The experimental results show that combining HAN and APE can complement each other to mitigate contextual translation errors and further improve CA-NMT by achieving reasonable improvement over HAN (i.e. BLEU score of 22.91 on En-De news-commentary dataset).

## 1. Introduction

Machine translation (MT) is one of the most critical tasks in natural language processing (NLP). With the advent of modern deep neural networks, developing machine translation models has become faster in training and inference, and also become more accessible compared to statistical machine translation. Early neural machine translation architectures used recurrent neural networks (RNNs) like gated recurrent unit (GRU), and long short-term memory (LSTM) modules for translation [1, 2]. Attention mechanism was used later in machine translation, and achieved high quality translated texts, and aligned source and target words properly [3, 4].

Recently, transformer architectures have been used widely, and could improve the speed and quality of machine translation significantly. Transformers use dot product attention, which is faster than the additive attention mechanism. Moreover, it is possible to train the modules in parallel, and hence, increase the speed in different applications. The NLP researchers exploit efficient transformer-based architectures by using and extending this new architectural concept to develop new models or fine-tune the existing ones [1].

Some new language models like BERT [2], GPT [3], BART [4] are extensions of the transformer architecture.

These models are trained on massive unlabeled text corpora and then fine-tuned on downstream tasks in natural language processing applications. Interestingly, most of them achieved state-of-the-art results on multiple downstream tasks, e.g. machine translations, sentiment analysis, and natural language inference. Moreover, models like BART, mT5 [5], and mBART [6] are used to fine-tune data in parallel corpora in the language translation field. Despite their good performance, the downside of these models is that it is difficult to train them. They usually use millions of parameters, and hence, need computationally expensive resources.

Most of the previous works in neural machine translation (NMT) build their models at sentence-level, and do not consider context beyond one sentence. In contrast, context-aware neural machine translation (CA-NMT) addresses this issue by considering more than one sentence to resolve the problem of lack of appropriate context when translating [11, 12].

So far, different paradigms have been proposed for context-aware machine translation, e.g. concatenating consecutive sentences [7], using additional encoders [8], hierarchical attention networks [9], and cache memory-based methods [15, 16].

Despite the achievements, there is still a long way to reach human-level translation quality. More specifically, context-aware NMT is a complex problem, and the researchers need to carefully and efficiently model context and propose new approaches and architectures to mitigate translation problems.

Most of the context-aware architectures that are proposed for NMT use additional networks to model the context. Hierarchical attention networks show promising results in such models by considering previous sentences to extract context [12, 11, 14]. They can learn to find the most related previous sentences, and then attend to most related words in those sentences to use in translating the current sentence.

One of the main problems of context-aware methods is the lack of document-level bilingual corpora, and because of this, context-aware models can not properly utilize additional sentences. In order to alleviate this problem, the document-level monolingual text in the target-side language could help the model benefit from context, and learn the structures and relations between consecutive sentences. This monolingual text can be created using round-trip translation (RTT) [10], which translates the monolingual target-side text into source language, and then translates it back to the target language. With the help of the automatic post-editing (APE) technique, the challenge of the lack of bilingual corpus becomes less concerning. However, solely using APE could not solve some inconsistencies in translation process.

In this work, we employ hierarchical attention networks on a small document-level corpus and combine it with monolingual automatic post-editing trained on a large monolingual document level corpus to fix the remaining inconsistencies in the context-aware network in order to generate coherent and consistent translations.

Our contributions for this work are as follows:

- Proposing a new hybrid method for context-aware neural machine translation with the combination of HAN [9] and APE [11] networks.
- Generating new 9M synthetic monolingual target-side document in German language for training APE. This corpus consists of pairs of synthetic and original texts.

The rest of this paper is organized as what follows. In Section 2, we review the related works in NMT, CA-NMT, and APE. In Section 3, we propose our method. in Section 4, we express our experimental setup and discuss our experimental results. Finally, in Section 5, we conclude our work.

## 2. Related Works

In this section, we divide the related works into three parts. In the first part, we review the neural machine translation literature. Next, we get to know the methods proposed in context-aware neural machine translation. In the last part of this section, we discuss automatic post-editing methods.

### 2.1. Neural machine translation

In neural machine translation, encoder-decoder architectures try to model the conditional probability of the target sentence given the source sentence. Encoders try to encode source sentence $X$ in a fixed-length context vector $c$, and decoders use this vector to generate the translation $Y$; one token at a time [12]. One of the first methods in sequence-to-sequence neural machine translation was proposed by [13]. They proposed an encoder-decoder framework using LSTM network for end-to-end neural machine translation. Their proposed architecture achieved a BLEU score of 34.8 on WMT'14 En-Fr test set. They also suggested reversing the source side sequence could help the model learn long-term dependencies.

When using fixed-length context vectors to encode input data, long sequences cannot be encoded properly. To solve this problem, Bahdanau *et al.* [14] have proposed an attention-based mechanism that does not encode the whole sentence in a fixed-length vector; instead, all encoder hidden states participate in creating the context vector [12]. This model learns to optimize the parameters of each of these encoder hidden states by using the last hidden states of the decoder at timestamp $t$. The authors showed that their method could learn alignments between the source and target words.

Many architectures tried to fix the recurrent problem of RNNs. Gehring *et al.* [15] have proposed a model based on a convolutional neural network that uses CNN layers for encoding the representations. The next milestone for neural machine translation and natural language processing was the proposal of transformer architecture. Vaswani *et al.* have proposed a network that use a new self-attention layer without any recurrence and convolution [1]. They got the state-of-the-art results at the time on WMT14 En-De and En-Fr corpora with the BLEU score of 28.4 and 41.8, respectively.

Recently with the advent of pre-trained language models, many models were proposed to improve translation quality by using the hidden representation learned in an unsupervised manner [2]. Self-supervised learning is the task of learning bidirectional representations from the text itself, and this can be achieved by randomly masking a few words in an unlabeled text and making the model to learn these words. Thus the model trained with this objective can learn bidirectional context representations from the text itself.

BERT is on top of the idea of self-supervised learning and trained with masked language modeling (MLM) objective to learn masked word context representations, besides Next Sentence Prediction (NSP) objective to learn consecutive sentences. BERT [2] uses transformer encoder, so it is primarily used on natural language understanding tasks [28]. BART [4] is a sequence to sequence encoder-decoder model trained on 160 million unlabeled text corpora and used for sequence-to-sequence tasks like summarization and translation besides natural language understanding downstream tasks. BART is trained with the objective of MLM by removing and permuting words to help the model fix and predict the masked words. BART can also be used for translation tasks and get proper results. Another similar architecture, which is based on BART language model and training objective, is multilingual BART (mBART). This model was trained on multiple unlabeled texts, and achieved a considerable performance in translating multiple languages in sentence-level and document-level configurations. The downside of this model is that it has around ten times more parameters than the baseline transformer model.

In the recent years, transformer models formed the basis for NMT models, and pretrained sequence-to-sequence models further improved this task.

### 2.2. Context-aware NMT methods

One of the first approaches for incorporating context in a neural machine translation model was simply concatenating consecutive sentences [7]. This method downside is model cannot properly learn contextual information and using word attention weightings.

Other method to utilize contextual information is by using additional encoder or decoder. Wang et al. have proposed a new method that used the last three sentences as context [8] and to model this, they used an encoder to encode last three sentences. Their model fixed 29 out of 38 ambiguous problems and 24 out of 32 inconsistency errors in the test set and gained 2 score improvements over vanilla RNN.

Another approach to context-aware NMT was to use cache memory module, i.e. the words of the last sentence are considered as input into a cache memory. This can help to encode previous word representations for learning context but managing the cache memory was a difficult task [16].

In paper [17], the linguistic phenomena that arise in the absence of context in translation is examined, and the authors proposed additional decoder in order to help model learn these discourse phenomena to tackle this problem.

Some other research works in the category of additional encoders focus on using hierarchical attention networks (HANs) that work by learning two-layer abstraction to hierarchically learn and find relevant sentences and words in a document for translation [14, 19], which we utilize these methods in our work.

### 2.3. Automatic post-editing methods

Automatic Post-Editing (APE) can be helpful in multiple scenarios, e.g. improving machine translation, utilizing information not accessible by decoders, helping human post-editors to enhance their translation, and translation domain adaptation [18]. However, proper post-editing is a costly and time-consuming process. The authors of [19] proposed an artificial dataset called eSCAPE that has triplet of source sentences, target sentences, and their translations.

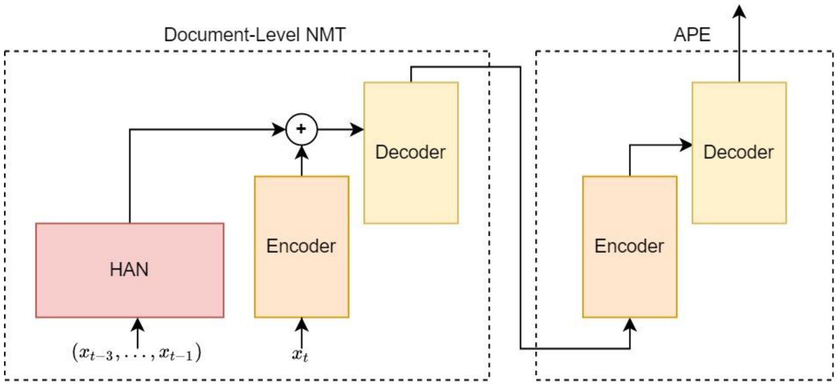

**Figure 1. Proposed architecture.**

Using only the target side sentences is a common approach for APE. Authors of [10], constructed synthetic data with round-trip translation (RTT), which means to translate from language *A* to language *B* and then vice versa to construct a synthetic corpus used in an encoder-decoder architecture for APE. In a similar research, authors of [11] used a document-level monolingual data in target language to exploit contextual information instead of single non-related sentence of previous work.

## 3. Proposed Model

As shown in Figure 1, we propose a new architecture combining the HAN and APE networks. The architecture of our proposed model consists of two parts. The first part is document-level neural machine translation model, which employs HAN architecture (Section 3.2). Our intuition for using HAN was that it could learn and use bilingual context appropriately. It is shown that this model can be useful in pronoun translation as well [9].

Using this network, we can encode current sentence using the last sentences as context. To use this model, we first train sentence-level NMT (Section 3.1). By employing and fine-tuning HAN-encoder, we then train a document-level neural machine translation that can get benefits of automatic post-editing.

In the second part of the model (Section 3.3), we employ automatic post-editing model trained on a massive target language data and can fix wrong translations, and also can improve lexical consistency and coherency.

For this part of the model, we used monolingual document level data and train the model following the approach proposed in [11]. We have monolingual data in target language, so we first translate target to source language with (trg → src) model then translate it back to the target language using (src → trg) model.

By utilizing and combining both context-aware model and APE, one can further improve performance of the neural machine translation and also improve multiple instances of discourse phenomena.

Details of the proposed method, including the sentence-level, document-level and automatic post-editing neural machine translation modules are presented in sequel.

### 3.1. Sentence-level neural machine translation

Models of neural machine translation are mostly based on the encoder-decoder architectures. In such networks, the encoder part receives an input sequence of $X = (x_1, x_2, ..., x_n)$, and generates an encoded representation. Using these representations, the decoder part generates target tokens $Y = (y_1, y_2, ..., y_n)$, one at a time in a left to right manner. The overall model tries to maximize conditional probability as follows:

$$P_\theta(Y \mid X) = \prod_{n=1}^{N} P_\theta(y_n \mid y_{<n}, X) \qquad (1)$$

where $P_\theta(Y \mid X)$ is conditional probability of target sentence *Y* given source sentence *X*.

Transformer architecture is similarly based on an encoder-decoder framework with stacked self-attention modules and fully connected layers in both encoder and decoder. The details of the transformer architecture are as follows:

**Self-attention module:** Attention module encodes each input word into three representations as *key*, *value* and *query*.

By evaluating the dot product of the *query* of a word with the *keys* of other words in the sequence, similarity of the words is measured [1].

**Transformer encoder:** The encoder of transformers consists of 6 stacked self-attention modules followed by fully connected layers. Positional embedding is also added to help the model learn the order of sentences. Encoder only attends to the input sequence itself, so it does not use any masked attentions. In the last layer, the sentence representation is encoded and the *key* and *value* representations are output to be used in the decoder module [1].

**Transformer decoder:** The decoder consists of two types of attention modules, i.e., self-attention on target sequences with the triangular mask to not attend to future target words, and attention module that attends to the whole source sentence representation using *key* and *value* pairs from the encoder. Similar to the encoder, the decoder also consists of 6 stacked layers [1].

### 3.2. Document-level neural machine translation

For document-level models, the NMT researchers try to model context as a conditional probability. The general formula is as follows:

$$P_\theta(Y \mid X) = \prod_{m=1}^{M} P_\theta(Y^m \mid X^m, D^{-m}) \qquad (2)$$

where $D^{-m}$ is document context consists of source and/or target sentences, and $Y^m$ and $X^m$ are the $m^{th}$ target and source sentences, respectively [20].

Among different methods for using context, in this paper, we use the hierarchical attention transformer network proposed by [9]. To this end, we use HAN-encoder, which employs HAN module in the encoder side of the model and has approved results on document-level translation. HAN works by using two abstraction of attention mechanism on sentence-level and word-level. It first finds related words at word-level, next finds the most related sentences between previous sentences.

### 3.3. Automatic post-editing

Post-editing model is similar to transformer; however, the inputs and outputs are in monolingual language. The APE models can be used to correct inconsistencies and improve coherency of translation models. For training an APE model, we need synthetic parallel monolingual data; therefore, we first build source to target translation model, and then use this model to translate our documents from De to En. Next, we build another target to source model to translate synthetic En data to De language and make our synthetic corpus. Therefore, for APE model, we used synthetic translations made from RTT as input, and the original monolingual data was used as the target of the model.

## 4. Experiments

### 4.1. Datasets

We used the En-De dataset provided in [20] as document-level bilingual data. This dataset consists of three corpora. Here, we use only the first two, i.e.

- **News-Commentary-2016 (NC-2016):** This part is extracted from the WMT news commentary v11 datasets; test sets consist of *test-news2015* and *test-news2016*.
- **TED-IWSLT-2017(IWSLT-2017):** This part of the dataset consists of transcripts of TED talks of English-German IWSLT-2017 dataset; *tst2016-2017* is used as test set.
- **Europarl7**

We used the same test sets as in [20]; details of the datasets are listed in Table 1.

**Table 1. Datasets statistics.**

| Data | Train | Validation | Test | All |
|---|---|---|---|---|
| News-Commentary 2016 | 220k | 2k | 3k | 225k |
| TED-IWSLT 2017 | 200k | 3k | 2.3k | 204k |
| Overall | 420k | 4k | 5.3k | 430k |

We used the document-level monolingual data provided by WMT 2019 competition to train our proposed architecture's automatic post-editing model. Multiple sources of crawled news were selected from documents published between 2010 and 2019, summing up to almost 9 million sentences. Furthermore, we added eSCAPE corpus [19], which consisted of 7 million WMT German parallel data in sentence-level data to expand our APE dataset.

### 4.2. Experimental setup

We used *Moses toolkit* to preprocess the data, i.e. we first normalized punctuations, and then tokenized the data. In the next step, we true-cased the input. For Byte pair encoding setup [21], we used *subword_nmt* with 32000 merges. We used *PyTorch* library with *Fairseq* framework for training our deep transformer models. All models were trained on a 1080Ti GPU with 11GB of memory. We finally report tokenized BLEU *multi-bleu.perl*, the detokenized BLEU (sacreBLEU) [22], METEOR score [23], and TER [24] to evaluate and compare the performance of the methods.

**Table 2. Proposed results.**

| Row | | NC-2016 | | | | | | | | IWSLT 2017 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sentence-level | | | | Document-level | | | | Sentence-level | | | | Document-level | | | |
| | | d.BLEU | t.BLEU | METEOR | TER | d.BLEU | t.BLEU | METEOR | TER | d.BLEU | t.BLEU | METEOR | TER | d.BLEU | t.BLEU | METEOR | TER |
| 1 | HAN | 22.08 | 22.01 | 41.44 | 70.2 | 22.82 | 22.8 | 42.11 | 67.95 | 23.97 | 23.95 | 42.96 | 67 | 24.42 | 24.38 | 44.04 | 67.89 |
| 2 | 1M | 21.13 | 21.17 | 38.97 | 70.09 | 21.61 | 21.73 | 39.97 | 68.19 | 22.32 | 22.34 | 40.64 | 68.17 | 22.79 | 22.87 | 41.8 | 69.19 |
| 3 | 3M | 21.2 | 21.25 | 39.19 | 70.06 | 21.81 | 21.85 | 40.08 | 67.98 | 22.6 | 22.62 | 41.11 | 67.83 | 23.16 | 23.19 | 42.37 | 68.62 |
| 4 | 7M | 21.33 | 21.37 | 39.57 | 70.04 | 21.99 | 22.05 | 40.24 | 67.81 | 23.1 | 23.14 | 41.59 | 67.74 | 23.63 | 23.69 | 42.76 | 68.36 |
| 5 | 9M | 21.62 | 21.67 | 39.81 | 70.24 | 22.38 | 22.43 | 40.48 | 68.11 | 23.4 | 23.46 | 41.89 | 67.36 | 23.81 | 23.86 | 43.03 | 68.25 |
| 6 | 16M | 21.95 | 22.03 | 41.35 | 70.14 | 22.61 | 22.64 | 41.96 | 67.81 | 23.47 | 23.51 | 42.67 | 67.27 | 24.02 | 24.06 | 43.76 | 68.2 |
| 7 | pairwise disjoint | 22.08 | 22.15 | 41.44 | 70.07 | 22.77 | 22.8 | 42.07 | 68.13 | 23.95 | 23.98 | 42.94 | 66.94 | 24.4 | 24.45 | 44.01 | 67.84 |
| 8 | paired sentences | **22.08** | **22.15** | **41.47** | **69.95** | **22.88** | **22.91** | **42.18** | **67.77** | **23.96** | **24** | **42.94** | **66.94** | **24.42** | **24.4** | **44.01** | **67.86** |

**Hyperparameters:** We used the same hyperparameters of transformer model as used in [1], i.e., 6 layers for each encoder and decoder, 8 attention heads with dimensionality of 512, and size of 2048 as the dimension of feed-forward layer. We selected the last three sentences as the context size in document-level modeling. In the training phase, we used initial learning rate of $7e^{-4}$ with one epoch warm up steps. Weight decay rate was set to 0.0001 and size of batch tokens was 8000 with 2 step gradient accumulation. Moreover, dropout and label smoothing rates were set to 0.3 and 0.1, respectively.

### 4.3. Experimental results

In order to investigate the best configuration for our proposed method, we designed and conducted three different experimental scenarios. The results of the proposed method in different scenarios and those of the HAN network are presented in Table 2. Details of different experimental scenarios are as follows:

- **Changing the number of document-level APE sentences:** We automatically generated about 9 million document-level German monolingual data for automatic post-editing model by using RTT. The data is produced using the German-English and English-German translation models described in the previous section. In our experiments, we conduct multiple experiments to assess the effect of increasing the number of sentences on evaluation scores from 1M to 16M, as described in rows 2 to 6 in Table 2.

  **Adding eSCAPE corpus:** We merged our 9 million synthetic data with the eSCAPE corpus [19] to generalize our data in multiple domains. In addition to the news category, the eSCAPE corpus contains a wide range of data in various domains such as subtitles, talks, and conversations, which can lead to generalization of our model. The corresponding addition of eSCAPE corpus mentioned in rows 6 to 8 in Table 2.
- **Concatenating two consecutive sentences (pairwise disjoint) and adding eSCAPE:** We pairwise concatenate the 9 million consecutive document-level sentences, i.e., (a, b, c, d) → (a, b), (c, d), where a, b, c, d indicate separate consecutive sentences. This way, the model can learn from context of consecutive sentences this approach used by [7].

  Following this approach in this experiment, our artificial data reached to 4.5 million sentences. We also combined this data with the eSCAPE corpus to obtain over 11 million sentences. The data was also integrated with the 200,000 IWSLT training corpus. The results of this setup are depicted in the 7th row of Table 2.
- **Concatenating each pair of sentences (paired sentences):** In this scenario, instead of using each sentence once, we used each sentence twice ((a, b, c, d) → (a, b), (b, c), (c, d)), so that the data eventually became 9 million, and each sentence appears twice, and then we merged them with the eSCAPE corpus. This yield to a dataset of around 16 million sentences. This configuration has 9 million consecutive sentences, in addition to eSCAPE and IWSLT corpus. The results are shown in the 8th row of Table 2.

It can be inferred from the trend of the results obtained in sentence-level experiments for both datasets (from 1 million to 16 million sentences) that the outcome of the algorithm is continuously improving.

To utilize the context-aware monolingual data, we continued the experiments under the defined scenarios. The second scenario, *pairwise disjoint*, results in better scores compared with those of the case of only using sentence-level data. The third scenario, *paired sentences*, results in substantial improvements in BLEU score. This improvement is more than those of last-mentioned methods. That means this approach can resolve inconsistencies in translation. As shown in Table 2, our best method achieves better results for the News Commentary-2016 dataset. This is expected, because our 9 million synthetic APE dataset consists of the German news text, which is considered as formal text. To mitigate deterioration effects of formal text on translation, and to add proper colloquial data, we expand this data with training set of IWSLT-2017 to further improve scores of informal texts. Because of limitation of our GPU memory and long text sentences of News-Commentary dataset, our results are a bit lower than the baseline method.

Our best configuration (APE + HAN) gets a BLEU score of 22.91 on the NC-2016 dataset and 24.45 on IWSLT-2017. To see the effect of our APE method on sentence-level data, we also report the results on the baseline system.

## 4. Conclusion

In this paper, we proposed a new method for context-aware NMT using hierarchical attention network and automatic post-editing model. The results confirm that our architecture can further improve the performance by increasing in BLEU score (22.91 on En-De News-Commentary) and by lowering translation error rate, and also get proper outputs for translation by fixing inconsistencies of both HAN and APE. This yields to improve the model and resolve multiple discourse phenomena in lack of context. Our experiments show that utilizing monolingual synthetic data in target-language properly can further improve the sentence-level APE model.

Therefore, it can be concluded that utilizing contextual information gained from synthetic monolingual APE dataset besides HAN can further improve context-aware neural machine translation models.

For future work, we aim to utilize other methods like pretrained language models in order to make the model take advantage of the hidden representations learned from millions of text corpora. We can also consider methods proposed for non-autoregressive APE. This method can also be used on other language pairs thus our proposed model is language agnostic.

We also plan to open source the 9 million sentence synthetic APE dataset which is trained on German language for future research and analysis.

Moreover, it is necessary to continue the experiments using more powerful computational resources to investigate about the upward trend of translation performance metrics in presence of larger post-editing datasets.

## References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., 2017, pp. 5998-6008.

[2] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, 2019.

[3] A. Radford and K. Narasimhan, "Improving Language Understanding by Generative Pre-training," 2018.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020.

[5] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," Online, 2021.

[6] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726-742, 2020.

[7] J. Tiedemann and Y. Scherrer, "Neural Machine Translation with Extended Context," in *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark, 2017.

[8] L. Wang, Z. Tu, A. Way, and Q. Liu, "Exploiting Cross-Sentence Context for Neural Machine Translation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2017.

[9] L. Miculicich, D. Ram, N. Pappas and J. Henderson, "Document-Level Neural Machine

Translation with Hierarchical Attention Networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018.

[10] M. Freitag, I. Caswell and S. Roy, "APE at Scale and Its Implications on MT Evaluation Biases," in *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Florence, Italy, 2019.

[11] E. Voita, R. Sennrich, and I. Titov, "Context-Aware Monolingual Repair for Neural Machine Translation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China, 2019.

[12] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., 2014, pp. 3104-3112.

[14] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations.*, 2015.

[15] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional Sequence to Sequence Learning," 2017.

[16] Z. Tu, Y. Liu, S. Shi, and T. Zhang, "Learning to Remember Translation History with a Continuous Cache," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 407-420, 2018.

[17] R. Chatterjee, M. Weller, M. Negri, and M. Turchi, "Exploring the Planet of the APEs: a Comparative Study of State-of-the-art Methods for MT Automatic Post-Editing," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China, 2015.

[18] M. Negri, M. Turchi, R. Chatterjee, and N. Bertoldi, "ESCAPE: a Large-scale Synthetic Corpus for Automatic Post-Editing," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018.

[19] S. Maruf, A. F. T. Martins, and G. Haffari, "Selective Attention for Context-aware Neural Machine Translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, 2019.

[20] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, 2016.

[21] M. Post, "A Call for Clarity in Reporting BLEU Scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium, 2018.

[22] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, 2005.

[23] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and R. Weischedel, "A Study of Translation Error Rate with Targeted Human Annotation," in *In Proceedings of the Association for Machine Transaltion in the Americas (AMTA 2006)*, 2006.

[24] E. Voita, R. Sennrich, and I. Titov, "When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019.

[25] E. Voita, P. Serdyukov, R. Sennrich, and I. Titov, "Context-Aware Neural Machine Translation Learns Anaphora Resolution," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 2018.

[26] S. Maruf and G. Haffari, "Document Context Neural Machine Translation with Memory Networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 2018.

[27] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," 2015.

[28] M. Kavehzadeh, M. M. Abdollah-Pour, S. Momtazi, "A Transformer-based Approach for Persian Text Chunking." *Journal of AI and Data Mining*, vol. 10, no. 3, 2022, 373-383.

# پس-ویرایش خودکار شبکه توجه سلسله مراتبی برای بهبود ترجمه ماشینی عصبی آگاه از متن

**محمد مهدی جزیرئیان و فواد قادری***

[۱] **دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران.**

ارسال ۲۰۲۲/۷/۲۸؛ بازنگری ۲۰۲۲/۱۰/۰۲؛ پذیرش ۲۰۲۲/۱۱/۱۴

## چکیده:

بیشتر روش‌های حال حاضر ترجمه ماشینی عصبی، جملات را بدون در نظر گرفتن زمینه متن ترجمه می‌کنند. نشان داده شده است بهره بردن از زمینه متن درون جمله‌ای و بین جمله‌ای به افزایش کیفیت ترجمه در مدل‌های ترجمه ماشینی کمک می‌کند. اگرچه فراهم آوردن دادگان سطح سند هزینه‌بر است، بهره‌گیری مناسب از دادگان تک زبانه می‌تواند در بهبود کیفیت ترجمه، کارایی خود را نشان دهد. در این پژوهش ما روشی نوین برای ترجمه ماشینی عصبی آگاه از متن ارائه کردیم که با تلفیق روش شبکه توجه سلسه مراتبی HAN و پس‌ویرایش خودکار APE تا مسئله پدیده‌های زبانی را که در نبود زمینه متن رخ می‌دهد را حل کنیم. شبکه HAN زمانی استفاده می‌شود که میزان کمی دادگان سطح سند داریم و APE می‌تواند با آموزش روی تعداد بسیار زیادی دادگان سطح سند تک زبانه در زبان مقصد آموزش داده شود تا بتواند دقت روش HAN را بهبود دهد. نتایج‌های آزمایش‌های ما نشان می‌دهد تلفیق روش‌های HAN و APE می‌توانند مکمل یکدیگر باشند و ترجمه ماشینی عصبی آگاه از متن را میزان بیشتری نسبت به HAN به صورت تنها بهبود دهند (به عبارت دیگر ما در این پژوهش به امتیاز بلو ۲۲٫۹۱ بر روی دادگان انگلیسی-آلمانی اخبار رسیدیم).

**کلمات کلیدی:** ترجمه ماشینی عصبی آگاه از متن، ترجمه ماشینی عصبی سطح سند، ترجمه ماشینی عصبی.