ICIC Express Letters
Part B: Applications
Volume **12**, Number **1**, January **2021**

 ISSN 2185-2766
pp. 91–98

IMPROVING NEURAL MACHINE TRANSLATION WITH POS TAGS

Long Hong Buu Nguyen[1,*], Hung Duong Minh[2], Dien Dinh[1] and Thanh Le Manh[3]

[1]Faculty of Information Technology
University of Science
Vietnam National University, Ho Chi Minh City
227 Nguyen Van Cu Street, Ward 4, District 5, Ho Chi Minh City 700000, Vietnam
[*]Corresponding author: long.hb.nguyen@gmail.com; ddien@fit.hcmus.edu.vn

[2]Office of Finance and Facilities
University of Foreign Languages
Hue University
57 Nguyen Khoa Chiem Street, Hue City 53000, Vietnam
dmhung@hueuni.edu.vn

[3]Computer Science Department
University of Sciences
Hue University
77 Nguyen Hue Street, Hue City 53000, Vietnam
lmthanh1953@yahoo.com

Received June 2020; accepted September 2020

Abstract. *Neural Machine Translation (NMT) has certain capability to implicitly learn semantic and syntactic aspects of the language. However, recently there have been also attempts to add linguistic annotation into neural translation models, and steps towards more linguistically motivated models. This paper presents an extension of NMT model to incorporate additional Part-of-Speech (POS) tags into the attention mechanism effectively to yield further improvements. The context vector produced by source annotations and target hidden state is used for target POS tagging. Then, we improve word prediction by simultaneously utilizing the context vector from attention layer and the predicted target POS tags. Evaluating on translating between English and Vietnamese in two directions with a low resource setting in the domain of TED talks, we obtain promising results in BLEU scores over baseline methods.*
Keywords: Part-of-speech, Recurrent neural network, Sequence to sequence model, Neural machine translation, Attention model

1. **Introduction.** Neural Machine Translation (NMT) [1, 2] is a new model in Machine Translation (MT) powered by most recent advances in sequence to sequence learning frameworks [3, 4]. NMT has made great progresses and drawn much attention in recent years.

In practical applications, the most basic form of NMT is the encoder-decoder framework where an encoder encodes a source sequence into a fixed-size vector representation, and then a decoder generates the target sequence sequentially via neural networks. The attention layer comes between the encoder and the decoder and helps the decoder to pick only the encoded inputs that are important for each step of the decoding process and resolve the problems with the fixed-size vector. Currently, the attention-encoder-decoder framework has become a subject of great interest to academics and industry.

Attention mechanism plays an important role in NMT. However, the conventional attention module is only conducted on the representation of the surface words of the source

DOI: 10.24507/icicelb.12.01.91

sentence, which may not be enough to model complex alignments between a target word and source words. Numerous experiments have established that more complex attention mechanisms [5, 6, 7, 8, 9, 10] or external syntactic information [11, 12, 13, 14, 15, 16, 17] can leverage the performance of NMT. These works have indicated that the POS tags which are used as additional syntactic information are of great benefits to NMT models, potentially reducing language ambiguity and alleviating data sparseness. Unfortunately, these methods only used the POS tags to enhance word representation or post editing the translation results.

Our goal is to utilize bilingual POS tags to model better attention mechanism. Taking advantage of the existence of parallel corpora, we use our POS taggers to assign a correct POS tag for each word in the corpora. Since POS tagging is a simpler task than word prediction and the number of POS tags is much less than that of words, the POS tagging has achieved very good performance. In our model, NMT and bilingual POS tags are jointly modeled via multi-task learning. This implementation fully exploits bilingual POS tags in semantic learning and attention modeling and thus leads to better performance.

This paper presents how we applied bilingual POS tags to attention-encoder-decoder NMT framework. First, source POS tags are combined with words to provide an effective word representation. Second, correct POS tag is generated for the predicted target word beforehand. Then, attention results are refined with the guidance of predicted bilingual POS tags. Finally, the refined attention is used to predict the target word.

This method shows several advantages. First, it may be of great benefits to NMT models, potentially reducing data sparseness and semantic ambiguity problems. Second, the POS tag information can be complementary to the textual input by providing a higher level of information abstraction so the input word representation is better encoded.

The remainder of the paper is organized as follows. We introduce related work (Section 2), then present our method (Section 3), report the results (Section 4) and finally conclude and propose some elements for future work (Section 5).

2. **Related Work.**

Recent advances in deep learning research facilitate innovative ideas in machine translation. Though promising, NMT still lacks the ability of modeling deeper semantic and syntactic aspects of the language. In this work, NMT and bilingual POS tags are jointly modeled via multi-task learning. These studies are related to our work.

2.1. **Syntax-directed attention.**

The attentional encoder-decoder framework pioneered by Bahdanau et al. [2] is the core, opening a new trend in neural machine translation. Cohn et al. [8] incorporated structural alignment biases inspired from conventional statistical alignment models (e.g., IBM models 1, 2) to encourage more linguistic structures in the attention layer. Tu et al. [18] further proposed a so-called coverage vector to trace the attention history for flexibly adjusting future attentions. Recently, many efforts have been initiated on exploiting source or target-side syntax information to improve the performance of NMT. Eriguchi et al. [16] proposed a tree-to-sequence NMT model that introduces a tree-based encoder and adapts its attention model to consider both sequential and phrase hidden units. Later, Chen et al. [12] extended the tree-based encoder to a bidirectional one with tree-coverage attention mechanism. More directly, Chen et al. also extended the local attention model with syntax-distance constraint. Chen et al. [13] extended the local attention with syntax-distance constraint, which focuses on syntactically related source words with the predicted target word to learning a more effective context vector for predicting translation.

2.2. **Multi-task learning.**

Multi-task learning has attracted attention to improving NMT in recent work. The initial approach for multi-task learning for neural networks was presented in [19]. The authors used convolutional and feed forward networks for several tasks such as semantic parsing and POS tagging. This idea was extended to sequence

to sequence models in [10]. Using additional word factors like POS-tags has shown to be beneficial in NMT [17] even the POS-tags was only used to enhance input word representation. Zaremoodi et al. [20, 21] have explored the use of syntactic parsing, semantic parsing, and Named Entity Recognition (NER) to improve the performance of NMT in low-resource scenarios. Niehues and Cho [22] have made use of POS tagging and NER tasks to improve NMT. In this word, they used the same encoder to encode the POS and named entity tags as what was done with input words. This approach is really the same as [17] but it just used another strategy.

3. **Improvement of NMT with POS Tags.** Additional word factors like POS-tags has shown to be beneficial in NMT. The previous work [17] did deal with the topic similar to that of this paper, but the important difference is that they only used the POS-tags to enhance input word representation. In this work, we investigate the feasibility of factored model idea [23] not only into input word representation but also into attentional neural translation model [2]. We aim to find how the neural model can benefit from incorporating the additional POS tagging factor at a deeper layer of NMT, especially the attention layer.

3.1. **Neural machine translation.** We introduce the background of the encoder-decoder [4, 24] framework.

Given a set of sentence pairs $D = \{(\mathbf{x}, \mathbf{y})\}$, the encoder f_{enc} with parameters θ_{enc} maps an input sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ to a sequence of continuous representations $h^{enc} = \left(h_1^{enc}, h_2^{enc}, \ldots, h_n^{enc}\right)$ whose size varies concerning the source sentence length. The decoder f_{dec} with θ_{dec} generates an output sequence $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ by computing $P(y_t|y_{<t})$ as follows:

$$P(y_t|y_{<t}) = softmax(f_{dec}(h_{dec}, c_t)) \tag{1}$$

where h_{dec} is a sequence of continuous representations for the decoder and c_t is the context vector which can be calculated as follows:

$$c_t = \sum_{i=1}^{n} a_{t,i} h_i^{enc} \tag{2}$$

where $a_{t,i}$ is attention weight:

$$a_{t,i} = softmax(e_{t,i}) = \frac{\exp e_{t,i}}{\sum_{j=1}^{n} \exp e_{t,j}} \tag{3}$$

where $e_{t,i}$ is a similarity score between the source and target representations. The parameters of calculating cross-attention weight $a_{t,i}$ are denoted as θ_{attn}. The $e_{t,i}$ can be calculated [2]:

$$e_{t,i} = V_\alpha^T \tanh\left(W_1 s_{t-1}^{dec} + W_2 h_i^{enc}\right) \tag{4}$$

where W_1, W_2 are learned parameters of the attention layer.

After that, the target hidden state s_t is updated:

$$s_t = f_{enc}(s_{t-1}, y_{t-1}, c_t) \tag{5}$$

The encoder and decoder are trained to maximize the conditional probability of target sequence given a source sequence:

$$\mathcal{L}_t(\mathcal{D}; \boldsymbol{\theta}) = \sum_{d=1}^{|D|} \sum_{t=1}^{M} \log P(y_t|y_{<t}, x; \theta_{enc}, \theta_{dec}, \theta_{attn}) \tag{6}$$

where M is target sentence length.

Both the encoder and decoder can be implemented by the different basic neural models structures, such as Recurrent Neural Network (RNN) [4, 24, 25], Convolutional Neural Network (CNN) [26], and self-attention [27].

3.2. **The proposed method.** In our proposed model, NMT and bilingual POS tagging are jointly modeled via multi-task learning, where the predicted POS tags are utilized to improve attention model.

Given a set of sentence pairs $D = \{(\mathbf{x}, \mathbf{px}, \mathbf{y}, \mathbf{py})\}$, **px** and **py** are pre-annotated POS tag sequences of **x** and **y**, respectively. To encode the source-side information, at each input step, word vector and POS vector are concatenated to establish a common vector. Then, these common vectors are inputted into the forward RNN layer and the backward RNN layer to represent h^{enc}.

As illustrated in Figure 1, our neural encoder is similar to the encoder of standard NMT model which is built upon bi-directional RNNs. We extend our neural decoder to update the context vectors with POS tag information to yield better target representations.

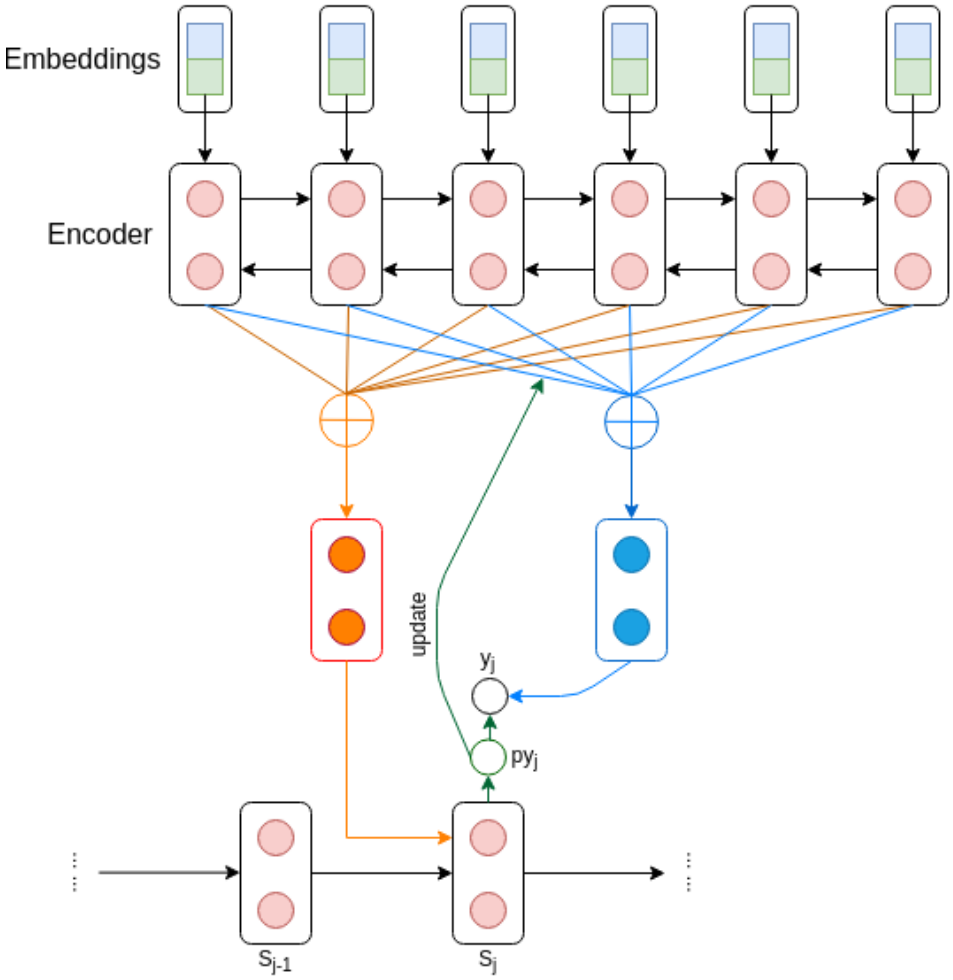

FIGURE 1. The overall architecture of the proposed model

The s_t is now used to predict the target POS tag at the tth timestep (instead of using to predict the target word). We apply a single layer neural network with an activation A (e.g., *sigmoid*, *tanh*, and *ReLu*) and a softmax classifier to obtain a POS tag probability distribution:

$$p_t = A(W_p s_t + b_p) \tag{7}$$

$$dp_t = softmax(W_{dp} p_t + b_{dp}) \tag{8}$$

where W_p, b_p and W_{dp}, b_{dp} are the weight matrices and biases, respectively. We also represent the POS tag embeddings based on dp_t for further use in updating the normal attention.

In addition to encoder state h_i^{enc} and target hidden state s_{j-1}, the target POS tag embedding is also used to update the attention. The context vector c'_t has similar equation as c_t but we concatenate the target hidden state s_{j-1} with target POS embeddings and p_t. Hence, the $e'_{t,i}$ is now calculated as:

$$e'_{t,i} = V_\alpha^T \tanh\left(W_3\left[s_{t-1}^{dec}; p_t; E_p dp_t\right] + W_4 h_i^{enc}\right) \quad (9)$$

where W_3, W_4 are learned parameters of the POS-updated attention layer.

4. Experiments.

4.1. Experimental settings.

We provide the information about the datasets, NMT configurations as well as the evaluation metrics.

4.1.1. *Datasets.*

We use two datasets: one for POS tagging and one for NMT.

We trained our English and Vietnamese POS taggers on English Penn Treebank[1] and Vietnamese Treebank corpora[2] respectively using the CRF toolkit[3]. We used both original POS tagset from both corpora as well as universal POS tagset [28]. Some statistics of POS datasets can be found in Table 1. We give result examples from our POS taggers in Table 2.

Table 1. Statistics of English and Vietnamese POS corpora

POS dataset	# tokens	# types	# sents	avg length
English	1,128,999	47,453	44,287	25.49
Vietnamese	1,364,450	43,389	50,000	27.30

Table 2. Example of POS taggers

Sentence	I	planted	a	food	forest	in	front	of	my	house	.
Original tags	NNS	VBD	DT	NN	NN	IN	NN	IN	PRP$	NN	.
Universal tags	PRON	VERB	DET	NOUN	NOUN	ADP	NOUN	ADP	PRON	NOUN	.

Sentence	Tôi	đã	trồng	môt	rừng	thưc	phẩm	ở	trước	nhà	tôi	.
Original tags	Pp	R	Vv	Nq	Nn	Nn	Nn	Cm	Nn	Nn	Pp	PU
Universal tags	PRON	ADV	VERB	NOUN	NOUN	NOUN	NOUN	CONJ	NOUN	NOUN	PRON	.

We conducted our NMT experiments on TED Talks which is a machine translation part of the IWSLT 2015 [29] and translate between English (en) ↔ Vietnamese (vi). Each TED talk is considered to be a document. For training, we used about 117K parallel sentences, and used tst2012 for tuning model parameters (phrase-based SMT) and early stopping (NMT). We evaluated on the official test sets tst2013 and tst2015. All details of NMT data statistics can be found in Table 3.

Table 3. Statistics of the English-Vietnamese datasets from IWSLT'15 MT track

MT dataset	# tokens		#types		# sents	avg length		# docs
	en	vi	en	vi		en	vi	
train	2,435,771	2,867,788	44,573	21,611	117,055	20.81	24.5	1,192
dev (tst2012)	27,988	34,298	3,518	2,170	1,553	18.02	22.08	14
test (tst2013)	26,729	33,683	3,676	2,332	1,268	21.08	26.56	18
test (tst2015)	20,850	26,235	3,127	2,059	1,080	19.31	24.29	12

[1]https://catalog.ldc.upenn.edu/LDC2015T13

[2]http://www.clc.hcmus.edu.vn

[3]https://taku910.github.io/crfpp/

4.1.2. *Set-up and configurations.* We compare these systems in our experiments:

- **SMT:** a famous phrase-based SMT which is implemented in Moses toolkit [30] with its standard configuration.
- **NMT:** for the NMT-related models, we implemented our model using pytorch deep learning library. We used a Long-Short Term Memory (LSTM) recurrent structure [31] for both source and target RNN sequences. The word embedding size is 250. The hidden layer dimension is 500. In the training phase, we used the default Adam optimizer [32] with a fixed learning rate of 0.0001. The batch size was 80 and the number of epochs was 10.
- **NMT + POS tags:** our proposed method to integrate the POS tags to the attention layer. The POS tagset for English is the Penn Treebank tagset and the POS tagset for Vietnamese is the Vietnamese Treebank tagset. The English-Vietnamese parallel corpus is pos-tagged by our pretrained taggers. The POS tag embedding size is 75. Other parameters are the same as the NMT model.
- **NMT + UPOS tags:** the original POS tagset is replaced with the universal POS tagset. We map the original POS tagset to the universal POS tagset before training our POS tagger. The number of the universal POS tags is less than the number of the original POS tags so the performance of the universal POS tagger is better than the performance of the original POS tagger. Other parameters are the same as the NMT with original POS tags model.

4.1.3. *Evaluation metrics.* We measure the end translation quality with case-insensitive BLEU [33]. We also apply bootstrapping re-sampling [30] to measuring the statistical significance ($p < 0.05$) of BLEU score differences between translation outputs of the proposed models compared to the baselines.

4.2. **Results and analysis.** We report our experimental results based on BLEU scores on English-Vietnamese translation task to further study the effectiveness of our model. Table 4 shows that the attentional model with our extensions is noticeably better than the vanilla NMT and SMT in terms of BLEU scores. The use of POS tags helps the decoder to have more opportunities to choose correct target words. Because the English POS tagset and Vietnamese POS tagset are different, the decoder has to implicitly learn how to map these different tags. On the other hand, the universal POS tags are the same for both languages and also have less number of tags than the original tags. This gives the decoder a better chance to learn the mapping and predict target tags more correctly. As proved by the results in Table 4, the NMT with universal POS tags (NMT + UPOS tags) gives the best BLEU scores in both testsets (i.e., tst2013 and tst2015) and both directions (i.e., $en \rightarrow vi$ and $vi \rightarrow en$).

Table 4. Experiments on translation systems

Methods	**en → vi**		**vi → en**	
	tst2013	**tst2015**	**tst2013**	**tst2015**
SMT	20.63	19.21	18.73	16.05
NMT	24.49	23.23	20.99	17.91
NMT + POS tags	25.17	23.47	21.85	19.11
NMT + UPOS tags	25.55	23.81	24.39	20.41

We show example outputs of the baseline and our methods in Table 5. The POS tags are not outputted but are extracted from the attention layer for illustration. This is not a long sentence so the POS tag predictions are correct. With the help of POS tags, the outputs are able to generate sentences more correctly. Especially, when the POS tag

Table 5. Example of translation systems

Source	I planted a food forest in front of my house .
Reference	Tôi đã trồng môt rừng thưc phẩm ở trước nhà tôi .
NMT	Tôi đã trồng môt rừng ở trước nhà .
NMT + POS tags	Tôi/Pp đã/R trồng/Vv môt/Nq rừng/Nn ở/Cm phía/Nn trước/Nn **nhà/Nn tôi/Pp** ./PU
NMT + UPOS tags	Tôi/PRON đã/ADV trồng/VERB môt/NOUN rừng/NOUN **thưc/NOUN phẩm/NOUN** trước/NOUN **nhà/NOUN tôi/PRON** ./.

mapping is 1-1 (i.e., the universal POS tag case), the output is better than the one with original POS tags.

Experimental results of deep models indicate that it is beneficial for NMT to explicitly incorporate linguistic knowledge by designing effective architecture, though the NMT with deep layers is able to learn linguistic knowledge to some extent. Our method can also be adapted to ConvS2S or Tranformer models because they also have the same attention mechanism. However, we consider this adaptation as our future work.

5. **Conclusion and Future Work.** We have proposed a novel attentional encoder-decoder for translation capable of integrating POS tags into neural machine translation. Experiments on English-Vietnamese corpora showed our model significantly outperforms sentence-level NMTs and achieved state-of-the-art performance on the datasets, which proved the effectiveness of our approach.

As our future work, we aim to explore whether the attentional neural translation model can benefit from other linguistic factors such as chunking tags, and named entity tags. To the best of our knowledge, this study can be considered as the first work towards fully-factored neural translation model for English-Vietnamese.

Acknowledgment. The authors would like to thank Computational Linguistics Center (University of Sciences, HCMC-VNU) for providing language resources.

REFERENCES

[1] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz and J. Makhoul, Fast and robust neural network joint models for statistical machine translation, *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, USA, pp.1370-1380, 2014.

[2] D. Bahdanau, K. Cho and Y. Bengio, Neural machine translation by jointly learning to align and translate, *Proc. of the ICLR*, San Diego, CA, USA, 2015.

[3] A. Graves, Generating sequences with recurrent neural networks, *arXiv e-Prints*, 2013.

[4] I. Sutskever, O. Vinyals and Q. V. Le, Sequence to sequence learning with neural networks, *Advances in Neural Information Processing Systems*, vol.27, pp.3104-3112, 2014.

[5] J. Su, J. Zeng, D. Xiong, Y. Liu, M. Wang and J. Xie, A hierarchy-to-sequence attentional neural machine translation model, *IEEE/ACM Trans. Audio Speech Lang. Process*, pp.623-632, 2018.

[6] J. Zhang, M. Wang, Q. Liu and J. Zhou, Incorporating word reordering knowledge into attention-based neural machine translation, *Proc. of the ACL*, 2017.

[7] S. Feng, S. Liu, N. Yang, M. Li, M. Zhou and K. Q. Zhu, Improving attention modeling with implicit distortion and fertility for machine translation, *Proc. of the COLING*, 2016.

[8] T. Cohn, C. D. V. Hoang, E. Vymolova, K. Yao, C. Dyer and G. Haffari, Incorporating structural alignment biases into an attentional neural translation model, *Proc. of the NAACL*, 2016.

[9] Y. Cheng, S. Shen, Z. He, W. He, H. Wu, M. Sun and Y. Liu, Agreement-based joint training for bidirectional attention-based neural machine translation, *Proc. of the IJCAI*, 2016.

[10] M. T. Luong, H. Pham and C. D. Manning, Effective approaches to attention-based neural machine translation, *Proc. of the EMNLP*, 2015.

[11] H. Chen, S. Huang, D. Chiang and J. Chen, Improved neural machine translation with a syntax-aware encoder and decoder, *Proc. of the ACL*, 2017.

[12] K. Chen, R. Wang, M. Utiyama, L. Liu, A. Tamura, E. Sumita and T. Zhao, Neural machine translation with source dependency representation, *Proc. of the EMNLP*, 2017.

[13] K. Chen, R. Wang, M. Utiyama, E. Sumita and T. Zhao, Syntax-directed attention for neural machine translation, *Proc. of the AAAI*, 2018.

[14] J. Li, D. Xiong, Z. Tu, M. Zhu, M. Zhang and G. Zhou, Modeling source syntax for neural machine translation, *Proc. of the ACL*, 2017.
[15] S. Wu, D. Zhang, N. Yang, M. Li and M. Zhou, Sequence-to-dependency neural machine translation, *Proc. of the ACL*, 2017.
[16] A. Eriguchi, K. Hashimoto and Y. Tsuruoka, Tree-to-sequence attentional neural machine translation, *Proc. of the ACL*, 2016.
[17] R. Sennrich and A. Birch, Linguistic input features improve neural machine translation, *Proc. of the CMT*, 2016.
[18] Z. Tu, Z. Lu, Y. Liu, X. Liu and H. Li, Modeling coverage for neural machine translation, *Proc. of the ACL*, 2016.
[19] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, Natural language processing (almost) from scratch, *The Journal of Machine Learning Research*, pp.2493-2537, 2011.
[20] P. Zaremoodi and G. Haffari, Neural machine translation for bilingually scarce scenarios: A deep multi-task learning approach, *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.
[21] P. Zaremoodi, W. Buntine and G. Haffari, Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation, *Proc. of the ACL*, 2018.
[22] J. Niehues and E. Cho, Exploiting linguistic resources for neural machine translation using multi-task learning, *Proc. of the 2nd Conference on Machine Translation*, 2017.
[23] P. Koehn and H. Hoang, Factored translation models, *Proc. of the 2007 Joint Conference on EMNLP-CoNLL*, Prague, Czech Republic, pp.868-876, 2007.
[24] K. Cho, B. van Merrienboer, D. Bahdanau and Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, *arXiv Preprint*, arXiv:1409.1259, 2014.
[25] S. Rikukawa, H. Mori and T. Harada, Recurrent neural network based stock price prediction using multiple stock brands, *International Journal of Innovative Computing, Information and Control*, vol.16, no.3, pp.1093-1099, 2020.
[26] J. Gehring, M. Auli, D. Grangier, D. Yarats and Y. N. Dauphin, Convolutional sequence to sequence learning, *Proc. of the ACL*, 2017.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems*, pp.5998-6008, 2017.
[28] S. Petrov, D. Das and R. McDonald, A universal part-of-speech tagset, *Proc. of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012.
[29] M. Cettolo, J. Niehues, S. Stuker, L. Bentivogli, R. Cattoni and M. Federico, The IWSLT 2015 evaluation campaign, *Proc. of the 12th International Workshop on Spoken Language Translation*, 2015.
[30] P. Koehn, Statistical significance tests for machine translation evaluation, *Proc. of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004.
[31] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation*, pp.1735-1780, 1997.
[32] D. P. Kingma and J. L. Ba, Adam: A method for stochastic optimization, *ICLR*, 2015.
[33] K. Papineni, S. Roukos, T. Ward and W. Zhu, BLEU: A method for automatic evaluation of machine translation, *Proc. of the ACL*, PA, USA, 2002.