Two Heads are Better than One? Verification of Ensemble Effect in Neural Machine Translation

Chanjun Park[1], **Sungjin Park**[2], **Seolhwa Lee**[3], **Taesun Whang**[4], **Heuiseok Lim**[1†]
[1]Korea University, [2]NAVER Corp., [3]University of Copenhagen, [4]Wisenut Inc.
{bcj1210, whiteldark, limhseok}@korea.ac.kr
sungjin.park@navercorp.com
taesunwhang@wisenut.co.kr

Abstract

In the field of natural language processing, ensembles are broadly known to be effective in improving performance. This paper analyzes how ensemble of neural machine translation (NMT) models affect performance improvement by designing various experimental setups (i.e., intra-, inter-ensemble, and non-convergence ensemble). To an in-depth examination, we analyze each ensemble method with respect to several aspects such as different attention models and vocab strategies. Experimental results show that ensembling is not always resulting in performance increases and give noteworthy negative findings.

1 Introduction

Ensemble is a technique for obtaining accurate predictions by combining the predictions of several models. In neural machine translation (NMT), ensembles are most closely related to vocabulary (vocab). In particular, by aggregating the prediction results of multiple models, the ensemble averages the probability values over the vocab of the softmax layer (Garmash and Monz, 2016; Tan et al., 2020).

Most existing studies on ensembling for NMT focus on improving the performance of shared tasks. For example, in WMT's shared task, almost every participating team applied the ensemble technique to improve performance (Fonseca et al., 2019; Chatterjee et al., 2019; Specia et al., 2020). However, in most cases, only experimental results that improved performance by applying the ensemble technique are introduced; in-depth comparative analysis is rarely conducted (Wei et al., 2020; Park et al., 2020a; Lee et al., 2020). In this study, we attempt to investigate three main aspects regarding ensembles for machine translation.

First, we investigate the ensemble effect when using various vocab strategies and different attention models. For the vocab that plays the most important role in the machine translation ensemble, three different experimental conditions—independent vocab, share vocab, and share embedding—are applied to two different attention networks (Bahdanau et al., 2014; Vaswani et al., 2017).

Second, we investigate which among intra-ensemble and inter-ensemble is more effective for performance improvement. Notably, intra-ensemble is an ensemble of identical models, while inter-ensemble represents an ensemble between models that follow different network structures.

Third, we analyze the effect of the non-converging model on ensemble performance. Most existing studies create an ensemble using only those models that have been fitted. However, we perform in-depth comparative analysis experiments, raising the question of whether the non-converging model has only negative effects.

2 Ensemble Design

2.1 Ensemble in NMT

Ensemble prediction is a representative method for improving the translation performance of NMT systems. A commonly reported method involves aggregating predictions by training different models of the same architecture in parallel. Then, during decoding, we average the probabilities over the output layers of the target vocab at each time step.

In this study, we follow the above method for ensembles using the same model architecture (*i.e.,* intra-ensemble). Because the target vocabs are the same, ensembles of components with different model structures (*i.e.,* inter-ensemble) also follow the same method. We conduct experiments on intra- and inter-ensemble effects on LSTM-Attention (Bahdanau et al., 2014) and Transformer (Vaswani et al., 2017) networks, combined with various vocab strategies. A detailed description of the vocab strategies is provided in the next section.

Proceedings of the Second Workshop on Insights from Negative Results in NLP, pages 23–28

ISBN 978-1-954085-93-0

2.2 Vocab Strategies

Independent vocab means learning separate weights from each encoder and decoder without any connection or communication between the source and target languages. Most NMT research follows this methodology (Cho et al., 2014; Vaswani et al., 2017; Park et al., 2021b).

Share vocab means that the model uses a common vocab for a combination of the source and target languages (Lakew et al., 2018). That is, the encoder and decoder interact within the same vocab, and can refer to each other's vocabs, thus making the model more robust.

Share embedding goes a step beyond sharing the source–target vocabs, and shares the vocab embedding matrix of the encoder and decoder (Liu et al., 2019). It enables the sharing of vocab from various languages through one integrated embedding space. Consequently, it has been widely used in recent multilingual NMT (Aharoni et al., 2019).

2.3 Experimental Design

2.3.1 Design of Intra- and Inter-ensemble

Intra-ensemble is an ensemble of identical models. We use the LSTM-Attention and Transformer networks with three different weights for the combinations to average the probabilities of ensemble. Inter-ensemble represents an ensemble of models that follow different network structures. We experiment with different combinations of the two attention-based models and vocab strategies. In this experiment, we aim to suggest directions for creating a better ensemble technique by analyzing the effect of intra- and inter-ensemble combined with the vocab strategy and size of vocabs. Moreover, all experiments compare vocab size (*i.e.,* 32k and 64k) by considering performance difference with respect to vocab capacity.

2.3.2 Design of Non-convergence Ensemble

In general, ensembles comprise well-fitted models; however, we conduct experiments to examine how models with less convergence affect the ensemble. Non-converging models are trained using ¼ of the iterations needed for convergent models. Consequently, we can determine whether non-converging models will cause only negative effects on the ensemble.

Vocab size	Cases	Baseline	Intra-ensembles			
			$\{w_1,w_2\}$	$\{w_1,w_3\}$	$\{w_2,w_3\}$	$\{w_1,w_2,w_3\}$
32,000	LSTM$_{ind}$	24.51	24.44 ▼	24.45 ▼	24.40 ▼	**24.47** ▼
	LSTM$_{sv}$	21.36	21.34 ▼	21.33 ▼	**21.49** ▲	21.36 -
	LSTM$_{se}$	21.49	21.40 ▼	21.47 ▼	**21.50** ▲	21.41 ▼
	Transformer$_{ind}$	33.40	33.71 ▲	**34.13** ▲	33.82 ▲	33.92 ▲
	Transformer$_{sv}$	29.23	29.48 ▲	29.80 ▲	29.70 ▲	**29.88** ▲
	Transformer$_{se}$	29.54	29.89 ▲	29.92 ▲	29.96 ▲	**30.19** ▲
64,000	LSTM$_{ind}$	25.02	24.86 ▼	24.98 ▼	24.96 ▼	**25.03** ▲
	LSTM$_{sv}$	22.89	22.83 ▼	22.79 ▼	**22.92** ▲	22.79 ▼
	LSTM$_{se}$	22.94	22.92 ▼	22.91 ▼	22.95 ▲	**22.98** ▲
	Transformer$_{ind}$	32.45	33.75 ▲	33.82 ▲	33.91 ▲	**33.97** ▲
	Transformer$_{sv}$	30.37	30.60 ▲	30.80 ▲	**31.02** ▲	30.97 ▲
	Transformer$_{se}$	30.82	30.96 ▲	31.22 ▲	31.14 ▲	**31.28** ▲

Table 1: Performance of intra-ensembles (combinations of vocab sizes and attention networks). The baseline score is the average of the three models that have different weights. Note that the bold numbers indicate the best score in each case.

3 Experimental Settings and Results

3.1 Experimental Setup

In this study, we use the Korean–English parallel corpus released on AI Hub [1] as the training data (Park and Lim, 2020). Several studies (Park et al., 2020b, 2021a) have adopted this corpus for Korean language NMT research. The total amount of sentence pairs is 1.6M. We randomly extract 5k sentence pairs twice from the training data, and use these data for the validation and test sets.

We employ sentencepiece (Kudo and Richardson, 2018) for subword tokenization. The performance evaluation of all the translation results are proceeds with BLEU score by leveraging multi-bleu.perl script given by Moses.

3.2 Results

Our negative findings and their insights are illustrated by **NF** and **Insight**, respectively. The performance results of the baseline models (seen as recipes of an ensemble) are shown in Tables 1 to 4.

3.2.1 Comparison of Intra-ensemble Effect

We show the results of applying the vocab strategies to two different models, namely LSTM-Attention and Transformer with three different weights (*i.e.,* w_1, w_2, and w_3) for intra-ensemble in Table 1. Additionally, we compare the combinations of those weights to investigate the apparent intra-ensemble effect.

Table 1 shows the significant variation in ensemble effect, according to the vocab strategies. The Transformer and LSTM-Attention models exhibit the highest performance in the order of independent vocab (ind), share embedding (se), and share

[1] https://aihub.or.kr/aidata/87

vocab (sv) in both vocab sizes (32k and 64k, respectively).

NF1: *Although Lakew et al. (2018); Park et al. (2021a) found that share vocab (sv) is effective when subword tokenization is applied as a pre-tokenize step during training, it has a negative effect in model training.* However, we find that sharing the vocab improves performance; nevertheless, sharing the embedding space is more helpful. However, training with independent vocab strategy shows the highest performance without interference.

To an in-depth examination, we analyze the intra-ensemble performance with respect to four aspects: i) different attention models, ii) vocab strategy, iii) vocab size, and iv) the number of models in the ensemble.

i) Different attention models We investigate the influence of the different attention networks on an ensemble. Self-attention-based networks refine (▲) all vocab strategies; however, there are more cases without performance improvement than those with performance improvement using the Bahdanau attention-based networks. That is, **NF2:** *specifically, with the Bahdanau attention network, there is a case in which a negative result (▼) occurred in an ensemble*. This result is interpreted as a difference in the robustness (*i.e.,* with minimum performance degradation) and capacity (*i.e.,* parallelism) of the model, as the following interpretations show. The Bahdanau attention network is exposed to problems with long-term dependencies (Bengio et al., 1993), resulting in the weak processing of long-sequences and requiring more data than self-attention. Furthermore, the Bahdanau attention network is well-known for not being context-aware, leading to variance in model prediction (Gao et al., 2021). Thus, **Insight:** *it can be seen that there is a lack of capacity and robustness in the Bahdanau attention network. Owing to this, it can be inferred that this network has a negative influence on the ensemble effect.*

ii) Vocab strategy We observe that there is performance variation among the vocab strategies. Our finding is in line with the aforementioned result in terms of the ensemble effect being the same as the ordering in LSTM-Attention, which is *ind, se* and *sv*. This is reasonable because of the previous result; however, **NF3:** *mixing the vocab (i.e., sv) has a negative effect on the ensemble performance.*

Vocab size	Cases	Intra (Baseline)	Inter
32,000	LSTM$_{ind}$ + Transformer$_{ind}$	34.13	31.70 (-2.43)
	LSTM$_{sv}$ + Transformer$_{sv}$	29.88	27.46 (-2.42)
	LSTM$_{se}$ + Transformer$_{se}$	30.19	27.25 (-2.94)
64,000	LSTM$_{ind}$ + Transformer$_{ind}$	33.97	31.95 (-2.02)
	LSTM$_{sv}$ + Transformer$_{sv}$	31.02	28.98 (-2.04)
	LSTM$_{se}$ + Transformer$_{se}$	31.28	28.97 (-2.31)

Table 2: Performance of inter-ensembles (combinations of vocab sizes and attention networks). Here, the column "Intra" records the highest score among the two different models, according to each vocabulary strategy in Table 1.

iii) Vocab size As illustrated in Table 1, the performance of intra-ensemble models shows vast differences owing to vocab sizes. We confirm that a vocab size of 64k is more effective than that of 32k; consequently, we theorize that vocab size is closely related to the effect of ensemble. In the Transformer ensemble with independent vocab (*i.e.,* Transformer$_{ind}$), the BLEU score is improved by 0.73 in the baseline model at 32k; in contrast, the BLEU score is improved by 1.52 at 64k, which is an improvement of more than two times. In other words, **NF4:** *even a slight alteration of vocab size significantly affects the ensemble performance*, and we know that a broader capacity leads to better performance when conducting vocab prediction using softmax.

iv) Number of ensemble models We explore the number of ensembles, and further validate the performance using the model combinations. **NF5:** *Contrary to the expectation that the number and performance of the ensemble models would show a positive correlation, this was not the case.* As shown in Table 1, only six cases, i.e., 50% of the 12 cases, demonstrate a good score in the three models ($\{w_1, w_2, w_3\}$) of the ensemble. The remaining six cases demonstrate a good score in two models ($\{w_1, w_3\}$, $\{w_2, w_3\}$). This result proves the statement of NF5.

3.2.2 Intra-ensemble or Inter-ensemble?

Inter-ensemble is feasible if the same vocab is used across the two models. Therefore, an ensemble of Transformer and LSTM-Attention model with the corresponding vocab strategy can be created; a comparison of the performance results with intra-ensembles is presented in Table 1. The results for inter-ensembles are shown in Table 2.

This result shows that the baseline (*i.e.,* Intra) exhibits better performance than inter-ensembles. Notably, inter-ensembles show a negative effect.

Vocab size	Cases	Baseline		Intra-ensembles with non-convergence							Δ%
		Best Intra	w_{nc}	$\{w_{nc}, w_1\}$	$\{w_{nc}, w_2\}$	$\{w_{nc}, w_3\}$	$\{w_{nc}, w_1, w_2\}$	$\{w_{nc}, w_1, w_3\}$	$\{w_{nc}, w_2, w_3\}$	$\{w_{nc}, w_1, w_2, w_3\}$	
32,000	LSTM$_{ind}$	**24.47**	19.06	22.84	22.79	22.84	23.54	23.53	23.56	23.75	-4.93
	LSTM$_{sv}$	**21.49**	16.11	19.25	19.32	19.31	20.14	20.11	20.28	21.42	-7.05
	LSTM$_{se}$	**21.50**	17.20	19.94	20.13	20.12	20.71	20.65	20.77	20.92	-4.82
	Transformer$_{ind}$	**34.13**	31.87	33.37	33.87	33.70	33.83	34.03	34.12	34.04	-0.82
	Transformer$_{sv}$	**29.88**	27.81	28.94	28.94	29.38	29.53	29.42	29.51	29.59	-1.84
	Transformer$_{se}$	**30.19**	27.72	29.01	29.23	29.56	29.67	29.96	29.83	29.93	-1.96
64,000	LSTM$_{ind}$	**25.03**	19.54	23.57	23.74	23.64	24.55	24.53	24.37	24.56	-3.57
	LSTM$_{sv}$	**22.92**	18.87	21.76	21.76	21.77	22.35	22.36	22.39	22.54	-3.43
	LSTM$_{se}$	**22.98**	17.64	21.07	21.21	21.22	22.09	22.11	22.15	22.43	-5.33
	Transformer$_{ind}$	33.97	31.22	33.23	33.68	33.71	33.79	33.85	34.14 ▲	**34.29** ▲	-0.46
	Transformer$_{sv}$	31.02	28.73	29.90	30.50	30.64	30.52	30.80	**31.03** ▲	30.90	-1.31
	Transformer$_{se}$	**31.28**	28.41	30.16	30.48	30.62	30.79	31.05	31.05	31.18	-1.66

Table 3: Performance of combinations of intra-ensembles using non-convergence models (w_{nc}) with vocab sizes and attention networks. $\Delta\%$ represents the average relative rate (i.e., the difference) $\{w_{nc}, w_1\}$ to $\{w_{nc}, w_1, w_2, w_3\}$ over "Best Intra." Note that the bold numbers represent the best score in each case.

Vocab size	Cases	Baseline	Inter-Ensembles			Δ%
		Best Inter	C(LSTM) & NC(Transformer)	NC(LSTM) & C(Transformer)	NC(LSTM) & NC(Transformer)	
32,000	LSTM$_{ind}$ + Transformer$_{ind}$	**31.70**	30.40	29.74	28.09	-7.22
	LSTM$_{sv}$ + Transformer$_{sv}$	**27.46**	26.22	25.32	23.64	-8.74
	LSTM$_{se}$ + Transformer$_{se}$	**27.25**	26.12	25.66	24.30	-6.94
64,000	LSTM$_{ind}$ + Transformer$_{ind}$	**31.95**	30.87	30.26	28.96	-6.01
	LSTM$_{sv}$ + Transformer$_{sv}$	**28.98**	27.41	27.56	26.15	-6.70
	LSTM$_{se}$ + Transformer$_{se}$	**28.97**	27.21	27.16	24.99	-8.69

Table 4: Performance of combinations of inter-ensembles with non-convergence (NC) and convergence (C) conditions along with vocab sizes and attention networks. $\Delta\%$ represents the average relative rate (i.e., the differences), from first to third columns, of inter-ensembles over "Best Inter." Note that the bold numbers indicate the best score in each case.

That is, **NF6:** *inter-ensemble exhibits a negative effect on performance, resulting in performance degradation in all cases.* It seems that the heterogeneous model architecture from the two different models acted as a hindrance to performance improvement.

3.2.3 Does Non-convergence Ensemble Cause Negative Results?

In this section, we investigate the effect of non-convergence on intra- and inter-ensembles. We choose the model with the best score (intra- and inter-ensembles) from Table 1 and Table 2, respectively, as target models for comparison.

The performance results of intra- and inter-ensemble with non-convergence models are illustrated in Table 3 and Table 4, respectively.

Intra-ensemble In Table 3, intra-ensemble with a non-convergence model leads to negative results compared to the baseline model (*i.e.,* Best Intra) in LSTM-Attention. Using the Transformer model as a baseline generally lead to performance degradation; however, the decrease is relatively small. There are a few exceptions (▲) that show that non-converging models with Transformer sometimes perform better when ensembled together.

These results revealed that **NF7:** *the Transformer model is more robust than the LSTM-Attention model and stronger under adverse conditions.* Additionally, it is inferred that the underfitted model plays a role in noise injection, boosting performance. **Insight:** *This result is a meaningful in that even a non-convergence model, which many researchers neglect, can help improve performance.*

Inter-ensemble As detailed in Table 4, the performance decreased in all cases, and **NF8:** *non-converging model causes a highly negative result in inter-ensembles compared to intra-ensembles.* In conclusion, inter-ensemble provide negative results in all cases for the experiments conducted in this study.

4 Conclusion

Most researchers consider it common sense that ensembles are better; however, few studies have conducted any type of close verification. In this study, we perform various tests based on three experimental designs related to the ensemble technique, and demonstrate its negative aspects. Thus, we provide insights into the positives and negatives of ensembling for machine translation. In the future, we plan to conduct expanded experiments based on different language pairs.

Acknowledgments

This work was supported by Institute for Information communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques) and MSIT(Ministry of Science and ICT), Korea, under the ICT Creative Consilience program(IITP-2021-2020-0-01819) supervised by the IITP(Institute for Information & communications Technology Planning Evaluation)

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio, Paolo Frasconi, and Patrice Simard. 1993. The problem of learning long-term dependencies in recurrent networks. In *IEEE international conference on neural networks*, pages 1183–1188. IEEE.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the wmt 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Erick Fonseca, Lisa Yankovskaya, André FT Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10.

Peng Gao, Shijie Geng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. 2021. Scalable transformers for neural machine translation. *arXiv preprint arXiv:2106.02242*.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Surafel M Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. *arXiv preprint arXiv:1811.01137*.

Jihyung Lee, WonKee Lee, Jaehun Shin, Baikjin Jung, Young-Gil Kim, and Jong-Hyeok Lee. 2020. Postech-etri's submission to the wmt2020 ape shared task: Automatic post-editing with cross-lingual language model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 777–782.

Xuebo Liu, Derek F Wong, Yang Liu, Lidia S Chao, Tong Xiao, and Jingbo Zhu. 2019. Shared-private bilingual word embeddings for neural machine translation. *arXiv preprint arXiv:1906.03100*.

Chanjun Park, Sugyeong Eo, Hyeonseok Moon, and Heui-Seok Lim. 2021a. Should we find another model?: Improving neural machine translation performance with one-piece tokenization method without model modification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 97–104.

Chanjun Park, Chanhee Lee, Yeongwook Yang, and Heuiseok Lim. 2020a. Ancient korean neural machine translation. *IEEE Access*, 8:116617–116625.

Chanjun Park and Heuiseok Lim. 2020. A study on the performance improvement of machine translation using public korean-english parallel corpus. *Journal of Digital Convergence*, 18(6):271–277.

Chanjun Park, Kinam Park, Hyeonseok Moon, Sugyeong Eo, and Heuiseok Lim. 2021b. A study on performance improvement considering the balance between corpus in neural machine translation. *Journal of the Korea Convergence Society*, 12(5):23–29.

Chanjun Park, Yeongwook Yang, Kinam Park, and Heuiseok Lim. 2020b. Decoding strategies for improving low-resource machine translation. *Electronics*, 9(10):1562.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, et al. 2020. Findings of the wmt 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91.

Liang Tan, Lin Li, Yifeng Han, Dong Li, Kaixi Hu, Dong Zhou, and Peipei Wang. 2020. An empirical study on ensemble learning of multimodal machine

translation. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, pages 63–69. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Daimeng Wei, Hengchao Shang, Zhanglin Wu, Zhengzhe Yu, Liangyou Li, Jiaxin Guo, Minghan Wang, Hao Yang, Lizhi Lei, Ying Qin, et al. 2020. Hw-tsc's participation in the wmt 2020 news translation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 293–299.