EBBS: An Ensemble with Bi-Level Beam Search for Zero-Shot Machine Translation

Yuqiao Wen[*1] **Behzad Shayegh**[*1] **Chenyang Huang**[1] **Yanshuai Cao**[2] **Lili Mou**[1 3]

Abstract

The ability of zero-shot translation emerges when we train a multilingual model with certain translation directions; the model can then directly translate in unseen directions. Alternatively, zero-shot translation can be accomplished by pivoting through a third language (e.g., English). In our work, we observe that both direct and pivot translations are noisy and achieve less satisfactory performance. We propose EBBS, an ensemble method with a novel bi-level beam search algorithm, where each ensemble component explores its own prediction step by step at the lower level but all components are synchronized by a "soft voting" mechanism at the upper level. Results on two popular multilingual translation datasets show that EBBS consistently outperforms direct and pivot translations, as well as existing ensemble techniques. Further, we can distill the ensemble's knowledge back to the multilingual model to improve inference efficiency; profoundly, our EBBS-distilled model can even outperform EBBS as it learns from the ensemble knowledge.

1. Introduction

Machine translation is a widely applicable NLP task that aims to translate a text from a source language to a target language (Brown et al., 1990; Bahdanau et al., 2015). The Transformer architecture (Vaswani et al., 2017) and pretrained large language models (Radford et al., 2019; Lewis et al., 2020) have largely improved translation performance, especially in the supervised setting (Raffel et al., 2020), where a model can learn from large volumes of parallel corpora. However, machine translation remains challenging for low-resource languages, because there are not enough data for large neural networks to learn these languages (Radford et al., 2019; Muennighoff et al., 2023).

We specifically focus on multilingual translation in the *zero-shot* setting, where the system is required to translate between unseen language pairs. Since collecting parallel data and training individual models for every translation pair are prohibitively expensive, it is common to build a single multilingual system (Johnson et al., 2017; Fan et al., 2021) that can perform translation for all language pairs, most of which are zero-shot translation directions that do not involve a high-resource language (e.g., English). These models work by prepending a language-indicator token; the zero-shot translation ability emerges as the model generalizes from trained language pairs and is able to perform *direct translation* for unseen ones (Liu et al., 2021; Wicks & Duh, 2022). The main drawback of such multilingual models is that they are noisy in the zero-shot setting due to the lack of supervision, and as a result, they tend to generate low-quality translations (Zhang et al., 2020; Liu et al., 2021).

Alternatively, zero-shot translation can be performed by *pivoting* (Wu & Wang, 2007; 2009), where the model first translates the input into a high-resource language such as English, which is then translated to the target language. However, pivoting requires two translation steps, often leading to an accumulation of errors (Babych et al., 2007; Gu et al., 2019).

In this paper, we propose an ensemble approach that aggregates direct and pivot translations in order to build a stronger multilingual translation model from weak ones. Building an ensemble for text generation is nuanced as it involves a sequence of word predictions. Word-level ensembles aggregate predictions at each generation step, which is usually achieved by averaging the predicted probabilities (Sennrich et al., 2016a; Freitag et al., 2017; Shanbhogue et al., 2023). This may not be ideal for zero-shot translation as the predictions are too noisy, making the averaged probabilities overly

[*]Project partially done during Mitacs internship at Borealis AI. [1]Department of Computing Science & Alberta Machine Intelligence Institute (Amii), University of Alberta [2]Borealis AI [3]Canada CIFAR AI Chair. Correspondence to: Yuqiao Wen <yq.when@gmail.com>, Behzad Shayegh <the.shayegh@gmail.com>, Chengyang Huang <chenyangh@ualberta.ca>, Yanshuai Cao <yanshuai.cao@borealisai.com>, Lili Mou <doublepower.mou@gmail.com>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria. Copyright 2024 by the author(s).

smooth. On the other hand, minimum Bayes risk decoding (MBR, Bickel & Doksum, 2015) can be considered a sequence-level voting ensemble, but existing MBR methods are only able to *select* from weak and noisy candidates given by the direct and pivot translations.

To this end, we propose an ensemble decoding algorithm with bi-level beam search (EBBS). Our EBBS performs two levels of beam search at each generation step: at the lower level, beam search is applied individually to each ensemble component; at the upper level, the ensemble maintains a shared beam by voting and synchronizing the candidates (sub-sequences) in lower-level beams. Unlike word-level ensembles (Freitag et al., 2017; Shanbhogue et al., 2023), EBBS does not average the predicted distributions, encouraging individual predictors to explore their own preferences; unlike sequence-level MBR ensembles (Kobayashi, 2018; Eikema & Aziz, 2020), EBBS does not select from a candidate set, and thus is more flexible since votings are performed throughout the generation process.

We conducted experiments on IWSLT (Cettolo et al., 2017) and Europarl (Koehn, 2005), two popular multilingual datasets for zero-shot machine translation. Results show that EBBS can generate high-quality translations and outperform existing ensemble techniques. In addition, we used EBBS-generated data for distillation to further improve the multilingual model. The experiment shows that such a distilling process encourages the model to learn from high-quality translations produced by EBBS, allowing it to outperform EBBS with no inference overhead compared with direct translation.

2. Related Work

Machine translation. In NLP, machine translation is a long-standing task that aims to rewrite text from one language to another without changing the meaning. Traditional research in translation has been mainly centered on the supervised setting, utilizing manually crafted rules (Forcada et al., 2011; Dugast et al., 2007) and statistical methods (Brown et al., 1990; Koehn, 2009); more recently, neural machine translation systems have considerably improved the performance (Vaswani et al., 2017; Raffel et al., 2020). However, translation remains challenging for low-resource languages, where neural models do not have enough parallel data to train on.

Translation for low-resource languages largely relies on *zero-shot* techniques, where no parallel text is available for a particular translation direction. In general, zero-shot translation can be accomplished in a monolingual or multilingual setting. With monolingual data, the most common approach is to build language-specific autoencoders that share the same latent space of semantics; translation is then achieved by plugging in the decoder of the desired language (Lample et al., 2018a;b; Mohiuddin & Joty, 2020).

In this paper, we focus on the multilingual setting, where one model can translate between multiple languages (Dabre et al., 2020). Usually, parallel texts only exist for a high-resource language such as English, leaving translations between low-resource languages zero-shot (e.g., Italian to Dutch) (Johnson et al., 2017; Fan et al., 2021). In this setting, the most common approach is to train the multilingual model on English-centric data, and the zero-shot translation ability naturally emerges during the training process (Johnson et al., 2017; Scao et al., 2022).

A key challenge for multilingual models is task interference, where too many languages tend to degrade model performance (Zaremoodi et al., 2018; Wang et al., 2020). As a result, research in this direction has been alleviating such interference by developing various parameter-separation schemes (Baziotis et al., 2022; Chronopoulou et al., 2023) and using gradient-based methods to update language-specific parameters (Wang & Zhang, 2022; He et al., 2023). In our work, we use a standard Transformer model following (Johnson et al., 2017) and (Liu et al., 2021). Our proposed ensemble algorithm EBBS is compatible with the above approaches, as it is agnostic to model architectures.

Ensemble methods. In a model ensemble, multiple machine learning systems are integrated so as to form a stronger one (Dong et al., 2020; Yang et al., 2023). *Bagging*, a classic ensemble technique, works by training multiple models with different portions of data and combining their predictions through averaging or voting (Breiman, 1996; Bühlmann & Yu, 2002). Another popular ensemble approach is *boosting*, where different models are trained sequentially, with each subsequent model addressing the mistakes of the previous ones (Schapire, 2003; Hastie et al., 2009; Natekin & Knoll, 2013). Unfortunately, bagging and boosting are not compatible with our setting, because we build an ensemble with a single model. Alternatively, *stacking* combines the outputs by training a meta-model (Wolpert, 1992; Ganaie et al., 2022), but this does not apply to our zero-shot setting either because we do not have groundtruth signals to train the meta-model. Even though these ensemble techniques may be applied to supervised text generation (Freitag et al., 2017; Kobayashi, 2018; Hendy et al., 2021), they are still not ideal as they do not take advantage of the sequential nature of sentences.

Unlike previous work, our EBBS performs bi-level beam search, exploring different components' own predictions and synchronizing them by a "soft voting" mechanism at every step. Our approach is specifically suited to the sequence generation process.

3. Approach

In this section, we first explain our ensemble components in §3.1. In §3.2, we propose EBBS, a novel ensemble decoding algorithm. Finally, we describe in §3.3 knowledge distillation with EBBS-decoded outputs for efficiency considerations.

3.1. Ensemble Components

In this work, we focus on zero-shot multilingual machine translation, which requires a system to perform translations for multiple languages, where some translation directions are unseen.

Specifically, our multilingual model is an encoder–decoder Transformer with a byte pair encoding tokenizer (Sennrich et al., 2016b) shared among all languages. The encoder can capture the semantics of tokens in different languages, whereas the decoder translates the encoded text into the desired language based on a target-language indicator token (Johnson et al., 2017; Fan et al., 2021).

We follow the standard English-centric training (Johnson et al., 2017; Liu et al., 2021), where the multilingual model is trained using parallel data with English on one side (e.g., German-to-English and English-to-Romanian). As mentioned in §1, the zero-shot ability emerges during such training, and the model is able to perform direct translation between unseen language pairs (e.g., German-to-Romanian) (Dabre et al., 2020; Ranathunga et al., 2023). An alternative approach is pivot translation, where the multilingual model performs two translations using a high-resource language as a pivot (e.g., first translating German to English, and then English to Romanian).

However, both direct and pivot translations have major weaknesses: the quality of direct translation tends to be low due to the lack of parallel data, whereas pivot translation suffers from error accumulation as it requires two translation steps (Babych et al., 2007; Gu et al., 2019).

In this paper, we would like to build an ensemble of direct and pivot translations to boost translation quality, where each translation path results in an ensemble component. Commonly used ensemble methods such as averaging and voting may not work well for text generation. Voting, for example, chooses the most voted prediction, but in text generation, the components' votes often do not share anything in common, because there could be tens of thousands of tokens in the vocabulary. An averaging ensemble, on the other hand, averages the predicted distributions of all components, potentially leading to an overly smooth distribution. As reported in a pilot study in the appendix of (Fan et al., 2021), a multilingual averaging ensemble only achieves a small improvement of 0.2 BLEU points over direct translation. Overall, ensemble methods have not been widely explored for zero-shot multilingual translation in previous literature.

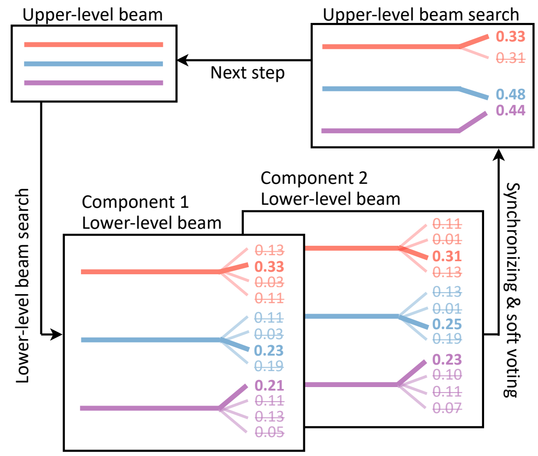

Figure 1. Illustration of our EBBS algorithm.

3.2. Our Proposed EBBS Algorithm

We propose an ensemble with bi-level beam search (EBBS), a novel decoding algorithm that enables different ensemble components to collaborate and vote on each other's partial generations with two levels of beam search.

At the lower level, each ensemble component performs beam search individually, exploring its own preferred regions of the sentence space. At the upper level, EBBS synchronizes the lower-level beam candidates through a voting mechanism, only keeping the most promising partial generations in a shared, upper-level beam. This allows the ensemble components to vote out spurious partial candidates and improve zero-shot translation performance.

Concretely, we assume there are K ensemble components $p_1, \cdots, p_K$, each predicting the probability of the next word given some prefix.

For the 0th decoding step, EBBS initializes the upper-level beam by $\overline{B}_0 = \langle \textsc{bos}, 1 \rangle$, suggesting that a sequence is forced to start with a special beginning-of-sequence token BOS with probability 1.

For step t, each ensemble component performs lower-level beam search individually, based on the prefixes in the last step's shared beam $\overline{B}_{t-1}$:

$$\underline{B}_{t,k} = \text{top-}Z\{ \langle \mathbf{y}_{1:t-1} \oplus \mathrm{y}, \ \ p \cdot p_k(\mathrm{y}|\mathbf{y}_{1:t-1}, \mathbf{x}) \rangle : \langle \mathbf{y}_{1:t-1}, p \rangle \in \overline{B}_{t-1}, \ \ \mathrm{y} \in V \} \tag{1}$$

for $k = 1, \cdots, K$. Here, top-Z selects Z-many sequences with the highest probabilities, $\oplus$ represents string concatenation, V is the vocabulary, and $p_k(\mathrm{y}|\mathbf{y}_{1:t-1}, \mathbf{x})$ is the kth ensemble component's predicted probability at step t given the prefix $\mathbf{y}_{1:t-1}$ and input $\mathbf{x}$.

At the upper level, EBBS synchronizes the lower-level individual beams $\underline{B}_{t,k}$, for $k = 1, \cdots, K$, into a shared, upper-level beam through a soft-voting mechanism, where the candidate set C_t is the union of the sequences in lower-level beams:

$$C_t = \bigcup_k \{\mathbf{y} : \langle \mathbf{y}, p\rangle \in \underline{B}_{t,k}\} \tag{2}$$

We evaluate each candidate in C_t and compute its overall vote as the sum of the probabilities.

$$\overline{B}_t = \text{top-}Z \left\{ \left\langle \mathbf{y}, \sum_{\substack{k:\ k=1,\cdots,K \\ \langle \mathbf{y}', p\rangle \in \underline{B}_{t,k}:\ \mathbf{y}'=\mathbf{y}}} p \right\rangle : \ \mathbf{y} \in C_t \right\} \tag{3}$$

In this way, the upper level synchronizes all ensemble components with the shared beam $\overline{B}_t$ for the next step of generation.

Intuitively, our voting scheme gives an ensemble component Z-many votes, each weighted by the predicted probability. The votes (probabilities) are then tallied (summed) for each candidate to form the upper-level beam. Our bi-level beam search terminates when we have Z-many terminated sequences in the shared beam, and returns the sequence with the highest score[1] as the ensemble output. We provide the detailed pseudocode for EBBS in Algorithm 1 and an illustration in Figure 1.

Discussion. As shown in Algorithm 2 (Appendix B), traditional beam search keeps a fixed-size beam of high-likelihood partial sequences. To build an ensemble with multiple predictors, it is tempting to directly average their probabilities $p(\mathbf{y}|\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} p_k(\mathbf{y}|\mathbf{x})$ as the score for beam search (Line 8, Algorithm 2), which has been experimented in previous work (Sennrich et al., 2016a; Shanbhogue et al., 2023).

However, our intuition suggests that such an approach may suffer from the *over-smoothing problem* (Wei et al., 2019; Wen et al., 2023b): when multiple translations (known as *modes*) are plausible given an input, the ensemble process will overly smooth out the modes by probability averaging.

By contrast, EBBS allows each ensemble component to explore its own mode (Lines 4–11, Algorithm 1). In Figure 1, for example, the top sequence yields two plausible next tokens, suggested by each component in the lower level; their probabilities are not smoothed out in our approach, unlike averaging ensembles. The upper level performs soft voting (Lines 12–19, Algorithm 1) so as to maintain tractable inference.

[1]For selecting the final output, we follow standard implementations and normalize the joint probability by length, i.e., taking the geometric mean of step-wise probabilities (Wolf et al., 2019; Ott et al., 2019). Otherwise, beam search algorithms are often biased towards short sequences (Meister et al., 2020).

3.3. EBBS-Based Distillation

Algorithm 1 Our EBBS Algorithm

Input: $\mathbf{x}$: input sentence; Z: beam size
K: number of scorers; $p_1, \cdots, p_K$: scorers

1 $H \leftarrow \emptyset$ ▷ candidate outputs
$\overline{B}_0 \leftarrow \{\langle \text{BOS}, 1\rangle\}$ ▷ upper-level beam
for $t = 1, 2, \cdots$ **do**
2 ▷ lower: individual beam search
for $\langle \mathbf{y}_{1:t-1}, p\rangle \in \overline{B}_{t-1}$ **do**
3 **for** $k = 1, \cdots, K$ **do**
4 $\underline{B}_{t,k} \leftarrow \emptyset$ ▷ lower-level beam
for $\mathrm{y} \in V$ **do**
5 $p' \leftarrow p_k(\mathrm{y}|\mathbf{y}_{1:t-1}, \mathbf{x})$
$\underline{B}_{t,k}.\,\mathrm{add}(\langle \mathbf{y}_{1:t-1} \oplus \mathrm{y}, p \cdot p'\rangle)$
6 $\underline{B}_{t,k} \leftarrow \underline{B}_{t,k}.\,\mathrm{top}(Z)$
7 ▷ upper: beam synchronization
$D \leftarrow$ empty dictionary
for $k = 1, \cdots, K$ **do**
8 **for** $\langle \mathbf{y}, p\rangle \in \underline{B}_{t,k}$ **do**
9 **if** $\mathbf{y} \in D$ **then**
10 $D[\mathbf{y}] \leftarrow p + D[\mathbf{y}]$
11 **else**
12 $D[\mathbf{y}] \leftarrow p$
13 $\overline{B}_t \leftarrow D.\,\mathrm{top}(Z)$
▷ check for termination
for $\langle \mathbf{y}, p\rangle \in \underline{B}_t$ **do**
14 **if** $\mathrm{y}_t = \text{EOS}$ **then**
15 $H.\,\mathrm{add}(\langle \mathbf{y}, p\rangle)$
if $|H| = Z$ **then**
16 **return** $H.\,\mathrm{top}(1)$

To improve inference efficiency, we perform knowledge distillation based on the outputs of our EBBS algorithm. In particular, we follow (Kim & Rush, 2016) and apply a sequence-level knowledge distillation loss, treating the output $\hat{\mathbf{y}}$ of our ensemble (serving as a *teacher*) as the pseudo-groundtruth for finetuning the multilingual translation model (serving as a *student*):

$$\mathcal{L}_{\text{KD}} = -\sum_{t=1}^{|\hat{\mathbf{y}}|} \log p(\hat{y}_t | \hat{\mathbf{y}}_{1:t-1}, \mathbf{x}) \tag{4}$$

Our distilling method is an ensemble-then-distill process. This differs from a straightforward practice of multi-teacher distillation, where the student learns from the union of teachers' outputs (Wu et al., 2021). The commonly applied cross-entropy loss is known to yield overly smooth distributions (Wen et al., 2023a;b), and the problem becomes more severe with multiple teachers, leading to less satisfactory performance of union distillation (Shayegh et al., 2024). On the contrary, our approach provides the student with a consolidated pseudo-groundtruth translation, causing less confusion during the distillation process especially when teachers disagree.

4. Experiments

4.1. Settings

We evaluated EBBS on two popular benchmark datasets for zero-shot machine translation: IWSLT (Cettolo et al., 2017), which contains 4 languages (with English) and 6 zero-shot directions; and Europarl v7 (Koehn, 2005), which contains 9 languages and 56 zero-shot directions.

We used BLEU scores (Papineni et al., 2002) (in particular, SacreBLEU (Post, 2018)) as our main evaluation metric,[2] which is one of the most widely used metrics for translation (Fan et al., 2021; Scao et al., 2022). For in-depth analyses, we further adopted other popular translation metrics, including the character-level n-gram F score (chrF2++) (Popović, 2017), the translation edit rate (TER) (Snover et al., 2006), and a more recent, neural network-based metric called COMET (Rei et al., 2020).

We replicated (Liu et al., 2021) and trained a multilingual translation system as our base model. Specifically, the neural architecture in (Liu et al., 2021) is a 5-layer encoder–decoder Transformer for IWSLT, but has 8 layers for Europarl to accommodate more training data and languages. Appendix C provides additional experimental details.

For EBBS, we used a beam size of five for both upper- and lower-level beams. In our experiment, we implemented standard beam search for comparison, where we also used a beam size of five, following the common practice (Meister et al., 2020). A comprehensive beam analysis can be found in Appendix D.

4.2. Competing Methods

We comprehensively compare our EBBS with direct/pivot translation and other ensemble methods.

Direct/pivot translation. For direct translation, we applied beam search on the multilingual model to translate in unseen directions. For pivot translation (Wu & Wang, 2007; 2009; Vamvas & Sennrich, 2022), we used English as the pivot because we have parallel data for translations both from and to English.

[2] We use BLEUn to denote the n-gram overlap and BLEU to denote the brevity-penalized geometric mean of BLEUn for $n = 1, \cdots, 4$. The exact evaluation scripts are available in our codebase (Footnote 1).

Word-level averaging ensemble. Averaging is one of the most widely used ensemble techniques in text generation (Sennrich et al., 2016a; Freitag et al., 2017; Shanbhogue et al., 2023). Essentially, the ensemble components' probabilities are first averaged before being fed to the standard beam search (Algorithm 2).

Word-level voting ensemble. The voting ensemble, common in classification tasks, picks the output class based on the number of votes from ensemble components (given by argmax). However, voting is not common in text generation, because argmax may select completely different words by the ensemble components due to the large vocabulary size, making voting ineffective. As a remedy, we pick the word by the highest probability when there is a tie for votes.

Sequence-level voting ensemble. Minimum Bayes risk (MBR) decoding is originally designed as a single-model decoding algorithm, where it selects a sequence from a set of beam search results based on similarity (Eikema & Aziz, 2020; Müller & Sennrich, 2021). Here, we use it as a sequence-level ensemble technique, where the candidates are the output sequences from different ensemble components. Let $C = \{\mathbf{y}_1, \cdots, \mathbf{y}_K\}$ be the set of candidate outputs given by K ensemble components. The best output is selected as

$$\mathbf{y}^* = \operatorname*{argmax}_{\mathbf{y}\in C} \sum_{\mathbf{y}'\in C\setminus\{\mathbf{y}\}} \text{BLEU}(\mathbf{y}, \mathbf{y}') \tag{5}$$

where $\text{BLEU}(\mathbf{h}, \mathbf{r})$ computes the BLEU score between a hypothesis $\mathbf{h}$ and a reference $\mathbf{r}$. In essence, MBR selects an output that resembles others most, using BLEU as the similarity metric.

4.3. Results and Analysis

Main results. Our experiment starts by a replication of the base multilingual model (Liu et al., 2021). As shown in Rows 1–2, Table 1, the results are generally close, indicating that our replication is successful and ready for ensemble research. Further, we tried English pivoting (Row 3), a common zero-shot translation method. In our experiments, we find that it does not outperform direct translation, as pivoting methods may suffer from the error accumulation problem due to two-step translation.

We then compare different ensemble techniques, including our proposed EBBS. We notice that IWSLT contains four languages (with English); thus we have two available pivoting directions (excluding source and target), which, along with direct translation, are our three ensemble components. For Europarl, it contains nine languages; for performance and efficiency concerns (to be shown in Figure 2), we also consider three translation paths as our ensemble components: direction translation, English pivoting, and a second

	#	Method	Average	it-nl	it-ro	nl-it	nl-ro	ro-it	ro-nl
IWSLT	1	Direct translation (Liu et al., 2021)†	17.7	18.5	17.8	17.9	15.5	19.6	16.8
	2	Direct translation (our replication)	17.29	17.46	17.48	18.23	14.63	19.65	16.26
	3	Pivoting (en)	16.19	17.49	15.09	16.79	13.05	18.34	16.37
	4	Word-level averaging ensemble	16.52	16.48	16.49	17.53	13.80	19.07	15.77
	5	Word-level voting ensemble	16.99	17.58	16.38	17.78	14.13	19.21	16.84
	6	Sequence-level voting ensemble (MBR)	16.66	16.54	16.51	17.72	13.64	19.58	15.98
	7	EBBS (ours)	18.24	19.52	17.09	19.06	14.58	20.75	18.45
	8	Direct w/ EBBS distillation (ours)	**18.92**	**19.86**	**18.80**	**19.73**	**15.39**	**21.23**	**18.48**

	#	Method	Average	da-de	da-es	da-fi	da-fr	da-it	da-nl	da-pt
Europarl	1	Direct translation (Liu et al., 2021)†	26.9	24.2	33.1	18.1	30.6	26.1	26.3	29.9
	2	Direct translation (our replication)	27.74	26.24	33.64	18.95	31.01	26.58	27.36	30.38
	3	Pivoting (en)	27.67	25.15	33.79	18.63	31.45	27.12	26.71	30.82
	4	Word-level averaging ensemble	27.76	26.21	33.64	18.88	31.09	26.71	27.40	30.37
	5	Word-level voting ensemble	27.75	25.93	33.90	18.47	31.29	27.08	26.74	30.84
	6	Sequence-level voting ensemble (MBR)	27.85	25.88	33.93	19.10	31.47	27.12	26.98	30.49
	7	EBBS (ours)	28.36	26.32	34.28	19.43	31.97	27.67	**27.78**	31.08
	8	Direct w/ EBBS distillation (ours)	**28.54**	**26.75**	**34.68**	**19.89**	**32.00**	**27.69**	27.61	**31.19**

Table 1. Main results of BLEU scores on IWSLT and Europarl. The best results are in **bold**; the second best results are underlined. † indicates cited results; others were obtained by our experimentation.

pivot.[3]

We study the common ensemble technique of word-level averaging (Row 4), which has been used in previous translation research (Freitag et al., 2017). As we can see, the averaging ensemble performs worse than direct translation on IWSLT, but is slightly better on Europarl. Our zero-shot results are different from (Freitag et al., 2017), which shows a word-level averaging ensemble of random seeds can improve performance in the supervised setting. This is because models trained with different random seeds exhibit similar behavior, and averaging their probabilities achieves a denoising effect. However, our ensemble components differ drastically in terms of their strengths and expertise due to the different translation paths (direct and pivot translations). Thus, word averaging fails to improve translation quality in our setting.

Alternatively, voting ensembles can also be applied, at either the word level or the sequence level. As seen, word-level voting is not effective, as it does not achieve significant improvements (Row 5). This is expected because the voted words (top predictions) by the ensemble components may not overlap due to the large vocabulary size. In such cases, the algorithm defaults to choosing the word with the highest probability, causing the ensemble to follow the most peaked distributions.

Sequence-level voting should also be done in a soft manner, and minimum Bayes risk (MBR) decoding can be thought of as using a Bayes risk to softly "vote" the candidate outputs. As seen from Row 6, such a method works relatively well on Europarl, achieving the second-highest performance across all ensemble methods; however, it works poorly on the IWSLT dataset. The main drawback of sequence-level voting is that it can only *select* one of the ensemble components' output. This may not work well when the individual ensemble components are weak, especially with the small IWSLT dataset. Such a selective sequence-level ensemble cannot integrate different expertise of its components during generation.

Dataset	Method	Avg. BLEU	Wins	Losses
IWSLT	Direct	17.29	2	4
	Ensemble	**18.24**	**4**	**2**
Europarl	Direct	27.85	4	52
	Ensemble	**28.44**	**52**	**4**
Overall	Direct	26.83	6	56
	Ensemble	**27.45**	**56**	**6**
p-value	3e-11			

Table 2. Pairwise comparison on all 66 zero-shot directions in both datasets. The p-value is given by a two-sided binomial test.

Unlike existing ensemble methods, our EBBS algorithm achieves higher performance in most directions on both datasets. Noticing that Europarl contains 56 zero-shot directions, we could only present in Table 1 the first seven directions based on the order provided by the dataset, due to the space limit. Table 2 further shows a pairwise comparison against direct translation (a strong baseline in our experiment) in all zero-shot directions. As seen, EBBS achieves higher performance in 56 out of 62 cases across two datasets, showing strong statistical evidence for its effectiveness, with a p-value of 3e-11 in a two-sided binomial test.

We also evaluate EBBS-based distillation (Row 8, Table 1). Again, since Europarl has 56 zero-shot directions, we follow the standard practice (Fan et al., 2021) and select a subset of directions, namely, Danish to other languages, to save computational cost. As seen in Row 8, EBBS-based distillation consistently achieves the highest performance

[3]We use the first available language in the order of Spanish, German, and French. For example, Spanish-to-German translation will have to use French as the pivot. These languages are chosen because they have the most content on the Internet according to the Web Technology Surveys (`https://w3techs.com/technologies/overview/content_language`).

Dataset	Method		BLEU$^{\uparrow}$	BLEU1$^{\uparrow}$	BLEU2$^{\uparrow}$	BLEU3$^{\uparrow}$	BLEU4$^{\uparrow}$	chrF2++$^{\uparrow}$	TER$^{\downarrow}$	COMET$^{\uparrow}$
IWSLT	EBBS		19.52	51.87	25.12	13.88	8.02	45.63	71.36	0.7341
	Direct Translation	No distillation	17.46	50.49	23.01	12.01	6.66	43.73	72.02	0.7088
		Direct distillation	18.10	50.37	23.53	12.63	7.17	44.48	72.86	0.7144
		Union distillation	17.80	49.21	23.01	12.51	7.10	44.93	75.92	0.7221
		EBBS distillation	**20.13**	**53.20**	**26.06**	**14.33**	**8.26**	**46.46**	**69.28**	**0.7428**
Europarl	EBBS		26.10	57.07	31.00	19.76	13.28	52.75	65.63	0.8340
	Direct Translation	No distillation	25.33	56.32	30.08	19.01	12.78	52.32	66.56	0.8276
		Direct distillation	25.44	56.54	30.28	19.13	12.79	52.61	66.34	0.8286
		Union distillation	25.53	56.58	30.34	19.18	12.91	52.63	66.27	0.8282
		EBBS distillation	**25.92**	**56.76**	**30.68**	**19.57**	**13.24**	**52.73**	**66.04**	**0.8307**

Table 3. Comparison of various distilling methods for Italian-to-Dutch translation. $^{\uparrow/\downarrow}$The higher/lower, the better.

in all directions (except for Danish-to-Dutch translation). This shows that an EBBS-distilled model can outperform EBBS, which is not surprising because learning can smooth out the noise of various heuristics (Deshmukh et al., 2021; Jolly et al., 2022), such as the ensemble algorithm in our scenario. Importantly, EBBS-based distillation achieves significantly higher translation quality with *no inference overhead* compared with direct translation.

Distillation analysis. We compare EBBS-based distillation with other distilling methods. Here, we only focus on Italian-to-Dutch[4] translation to save computational cost.

In particular, we consider two alternative distilling methods: direct and union distillation. Direct distillation finetunes the multilingual model with its own predictions based on direct translation. Union distillation, on the other hand, takes the union of the teachers' outputs (direct and pivot translations) for training, which is under a controlled experimental setup, because it uses exactly the same translation paths as our EBBS-based distillation.

As seen in Table 3, both direct and union distillation marginally improve the performance compared with no distillation. Intriguingly, learning from the union of multiple teachers is not necessarily better than learning from the best teacher (namely, direct translation). This is because multiple teachers may provide conflicting training signals and confuse the student model.

On the contrary, our EBBS-based distillation consistently outperforms direct and union distillation on both datasets. This shows that our ensemble-then-distill approach is able to consolidate the knowledge of multiple teachers to better train the student model.

Analysis of voting methods in EBBS. In our EBBS algorithm, the lower-level beams are synchronized into a shared upper-level beam by voting. Specifically, EBBS uses a mechanism of top-Z sum voting, where we add the ensemble components' probabilities for each appearance of a candidate in the lower-level beam, shown in Eqn. (3). Here, we analyze a few alternative voting methods for EBBS.

If EBBS adopts total-sum voting, it still uses lower-level beams to find candidates, but adds all components' probabilities together. This is equivalent to applying the common averaging ensemble to the top-Z candidates. However, it differs from our approach, because in total-sum voting, a component will vote even if the candidate does not appear in its own lower-level beam; the probability after voting in Eqn. (3) is substituted with $\frac{1}{K}\sum_k p_k(\mathbf{y}|\mathbf{x})$. As shown in Table 4, EBBS with total-sum voting performs worse than direct translation, suggesting the importance of ignoring the components whose lower-level beam does not contain the candidate. This is analogous to nucleus sampling (Holtzman et al., 2019), where the long tail of a distribution is mainly noise and should be ignored.

Other voting schemes that EBBS may use include 0/1 voting and max voting. The former selects the candidates that appear most in the lower-level beams, disregarding the probability values (unless for ties); the latter chooses the maximum probability across the lower-level beams, which gives preference to sequences through a maximization bias (Hasselt, 2010; van Hasselt et al., 2016). As seen, EBBS performs relatively well with both of these voting schemes, achieving a decent improvement over the baseline approach; however, their performance is worse than our top-Z sum voting.

Overall, the proposed bi-level beam search ensemble is effective with different voting schemes (except for the total-sum voting), and our top-Z sum voting works the best among these variants.

Analysis of ensemble components. In Table 5, we analyze the ensemble components to better understand our ensemble technique for zero-shot machine translation. As seen, direct translation is an effective approach, which is consistent with previous literature (Fan et al., 2021; Liu et al., 2021). English pivoting achieves higher performance for some metrics but lower for others; it is not conclusively better than direct translation, probably because of the error accumu-

[4]We could only afford one translation direction for this analysis, because we need to train different models for all competing distilling methods. This differs from Table 1, where we follow previous work and perform EBBS-based distillation for Danish to other languages. We chose Italian-to-Dutch translation here, because it is the first one in IWSLT, conveniently also available in Europarl.

Voting scheme	BLEU↑	BLEU1↑	BLEU2↑	BLEU3↑	BLEU4↑	chrF2++↑	TER↓	COMET↑
None (beam search)	25.33	56.32	30.08	19.01	12.78	52.32	66.56	0.8276
Total-sum	25.27	56.67	30.27	19.07	12.65	52.19	66.12	0.8311
Max	25.81	56.89	30.76	19.51	13.09	52.46	65.92	0.8300
0/1	25.84	56.99	30.78	19.49	13.05	52.61	65.80	0.8322
Top-Z sum (ours)	**26.10**	**57.07**	**31.00**	**19.76**	**13.28**	**52.75**	**65.63**	**0.8340**

Table 4. Comparison of different ensemble variants, using Italian-to-Dutch translation in the Europarl dataset as the testbed.

Method	BLEU↑	BLEU1↑	BLEU2↑	BLEU3↑	BLEU4↑	chrF2++↑	TER↓	COMET↑
Direct translation	25.33	56.32	30.08	19.01	12.78	52.32	66.56	0.8276
Pivoting (en)	25.08	56.76	30.29	19.06	12.66	51.92	66.24	0.8322
Pivoting (es)	24.40	55.38	29.08	18.22	12.09	51.71	67.91	0.8192
Pivoting (pt)	24.34	55.46	29.02	18.13	12.02	51.61	67.68	0.8191
Pivoting (fr)	24.20	55.41	29.02	18.00	11.84	51.61	67.84	0.8208
Pivoting (de)	23.65	55.33	28.69	17.67	11.54	50.70	67.89	0.8157
Pivoting (da)	23.12	54.81	27.96	17.12	11.12	50.36	69.00	0.8156
Pivoting (fi)	20.74	53.54	26.10	15.43	9.79	48.11	70.59	0.8051
Our EBBS	**26.10**	**57.07**	**31.00**	**19.76**	**13.28**	**52.75**	**65.63**	**0.8340**

Table 5. The performance of direct/pivot translation and our EBBS for Italian-to-Dutch translation on Europarl.

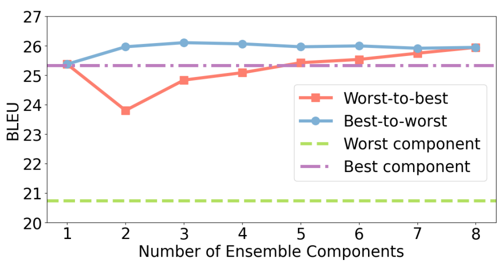

Figure 2. Analysis of the number of ensemble components for Italian-to-Dutch translation on Europarl.

lation problem. Pivoting through non-English languages degrades the performance to a large extent because lacking supervision along the pivoting path leads to two steps of zero-shot translation. EBBS, on the other hand, combines the strengths of individual components and consistently outperforms them in all metrics.

We further study how EBBS performs with different numbers of ensemble components. Specifically, we analyze two incremental ensemble settings: best-to-worst and worst-to-best. In both cases, we start with direct translation; then we incrementally add the next "best" or "worst" pivot translation according to Table 5.

Figure 2 shows the trends of incremental ensembles. If we add the best pivot directions, the performance peaks at three ensemble components; interestingly, the inclusion of weaker components does not affect EBBS much. On the other hand, adding the worst pivot translation at the beginning leads to an immediate drop of 1.6 BLEU points, which then largely recovers with the second pivot. This is reasonable because the worst pivot (Finnish) is 4.6 BLEU points lower than direct translation, and EBBS cannot decide on which of the two ensemble components to trust; despite this, the performance of EBBS is still much better than the average performance of the components. With a second pivot, there is a third "opinion" when the first two components "disagree." The performance continues to rise if more and stronger components are added. In fact, our ensemble even surpasses the baseline with 4 weakest pivot translations, each of which is at least 1 BLEU point lower than the baseline. This demonstrates that EBBS is flexible and works well with both strong and weak ensemble components.

Additional results. We show additional results in the appendices. D: Analysis of beam size, E: Analysis of inference efficiency, F: Entropy of distilled models, and G: Case study.

5. Conclusion and Future Work

In this work, we address ensemble-based zero-shot machine translation by directly translating and pivoting through different languages. We further design a novel bi-level beam search algorithm (called EBBS) for decoding. We evaluated EBBS on two popular zero-shot translation datasets, IWSLT and Europarl. Results show that EBBS outperforms existing ensemble techniques, and that the high-quality translations produced by EBBS can be used for distillation to improve translation efficiency (and sometimes also output quality).

Our EBBS is a general ensemble algorithm that can be potentially applied to various sequence generation tasks; however, we limited our scope to zero-shot machine translation in this paper due to the background of the project. In the future, we plan to explore EBBS for large language models (LLMs). Building an ensemble of LLMs encounters additional challenges because different LLMs tend to use different tokenizers. We may get inspirations from semi-CRF (Sarawagi & Cohen, 2004), where different tokenizations may be marginalized out during the generation process.

Limitations. See Appendix A.

References

Babych, B., Hartley, A., and Sharoff, S. Translating from under-resourced languages: Comparing direct transfer against pivot translation. In *Proceedings of Machine Translation Summit*, 2007. URL https://aclanthology.org/2007.mtsummit-papers.5.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1409.0473.

Baziotis, C., Artetxe, M., Cross, J., and Bhosale, S. Multilingual machine translation with hyper-adapters. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1170–1185, 2022. URL https://aclanthology.org/2022.emnlp-main.77.

Bickel, P. J. and Doksum, K. A. *Mathematical Statistics: Basic Ideas and Selected Topics*. CRC Press, 2015. URL https://bickel.stat.berkeley.edu/teaching.

Breiman, L. Bagging predictors. *Machine Learning*, 24:123–140, 1996. URL https://link.springer.com/article/10.1007/BF00058655.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990. URL https://aclanthology.org/J90-2002.

Bühlmann, P. and Yu, B. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002. URL https://www.jstor.org/stable/1558692?seq=1&cid=pdf-reference.

Cettolo, M., Federico, M., Bentivogli, L., Niehues, J., Stüker, S., Sudoh, K., Yoshino, K., and Federmann, C. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the International Conference on Spoken Language Translation*, pp. 2–14, 2017. URL https://aclanthology.org/2017.iwslt-1.1.

Chronopoulou, A., Stojanovski, D., and Fraser, A. Language-family adapters for low-resource multilingual neural machine translation. In *Proceedings of the the Workshop on Technologies for Machine Translation of Low-Resource Languages*, pp. 59–72, 2023. URL https://aclanthology.org/2023.loresmt-1.5.

Dabre, R. and Kurohashi, S. MMCR4NLP: Multilingual multiway corpora repository for natural language processing. *arXiv preprint arXiv:1710.01025*, 2017. URL http://arxiv.org/abs/1710.01025.

Dabre, R., Chu, C., and Kunchukuttan, A. A survey of multilingual neural machine translation. *ACM Computing Surveys*, 53(5), 2020. URL https://doi.org/10.1145/3406095.

Deshmukh, A. A., Zhang, Q., Li, M., Lin, J., and Mou, L. Unsupervised chunking as syntactic structure induction with a knowledge-transfer approach. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 3626–3634, 2021. URL https://aclanthology.org/2021.findings-emnlp.307.

Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258, 2020. URL https://link.springer.com/article/10.1007/s11704-019-8208-z.

Dugast, L., Senellart, J., and Koehn, P. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the Workshop on Statistical Machine Translation*, pp. 220–223, 2007. URL https://aclanthology.org/W07-0732.

Eikema, B. and Aziz, W. Is MAP decoding all you need? The inadequacy of the mode in neural machine translation. In *Proceedings of the International Conference on Computational Linguistics*, pp. 4506–4520, 2020. URL https://aclanthology.org/2020.coling-main.398.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Auli, M., and Joulin, A. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021. URL http://jmlr.org/papers/v22/20-1307.html.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127–144, 2011.

Freitag, M., Al-Onaizan, Y., and Sankaran, B. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*, 2017. URL http://arxiv.org/abs/1702.01802.

Ganaie, M., Hu, M., Malik, A., Tanveer, M., and Suganthan, P. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151,

2022. URL https://www.sciencedirect.com/science/article/pii/S095219762200269X.

Gu, J., Wang, Y., Cho, K., and Li, V. O. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1258–1268, 2019. URL https://aclanthology.org/P19-1121.

Hasselt, H. Double q-learning. In *Advances in Neural Information Processing Systems*, 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/091d584fced301b442654dd8c23b3fc9-Paper.pdf.

Hastie, T., Rosset, S., Zhu, J., and Zou, H. Multi-class AdaBoost. *Statistics and Its Interface*, 2(3):349–360, 2009. URL https://www.intlpress.com/site/pub/pages/journals/items/sii/content/vols/0002/0003/a008/.

He, D., Pham, M.-Q., Ha, T.-L., and Turchi, M. Gradient-based gradual pruning for language-specific multilingual neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 654–670, 2023. URL https://aclanthology.org/2023.emnlp-main.43.

Hendy, A., Gad, E. A., Abdelghaffar, M., ElMosalami, J. S., Afify, M., Tawfik, A. Y., and Hassan Awadalla, H. Ensembling of distilled models from multi-task teachers for constrained resource language pairs. In *Proceedings of the Conference on Machine Translation*, pp. 130–135, 2021. URL https://aclanthology.org/2021.wmt-1.8.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rygGQyrFvH.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. URL https://aclanthology.org/Q17-1024.

Jolly, S., Zhang, Z. X., Dengel, A., and Mou, L. Search and learn: Improving semantic coverage for data-to-text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10858–10866, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/21332.

Kim, Y. and Rush, A. M. Sequence-level knowledge distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. URL https://aclanthology.org/D16-1139.

Kobayashi, H. Frustratingly easy model ensemble for abstractive summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 4165–4176, 2018. URL https://aclanthology.org/D18-1449.

Koehn, P. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit*, pp. 79–86, 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.

Koehn, P. *Statistical Machine Translation*. Cambridge University Press, 2009. URL https://www.cambridge.org/core/books/statistical-machine-translation/94EADF9F680558E13BE759997553CDE5.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=rkYTTf-AZ.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. Phrase-based & neural unsupervised machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 5039–5049, 2018b. URL https://aclanthology.org/D18-1549.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020. URL https://aclanthology.org/2020.acl-main.703.

Liu, D., Niehues, J., Cross, J., Guzmán, F., and Li, X. Improving zero-shot translation by disentangling positional information. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pp. 1259–1273, 2021. URL https://aclanthology.org/2021.acl-long.101.

Meister, C., Cotterell, R., and Vieira, T. If beam search is the answer, what was the question? In *Proceedings of the Conference on Empirical Methods in Natural Language*

Processing, pp. 2173–2185, 2020. URL https://aclanthology.org/2020.emnlp-main.170.

Mohiuddin, T. and Joty, S. Unsupervised word translation with adversarial autoencoder. *Computational Linguistics*, 46(2):257–288, 2020. URL https://aclanthology.org/2020.cl-2.2.

Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. A. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, pp. 50358–50376, 2023. URL https://openreview.net/forum?id=j5BuTrEj35.

Müller, M. and Sennrich, R. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pp. 259–272, 2021. URL https://aclanthology.org/2021.acl-long.22.

Natekin, A. and Knoll, A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7:21, 2013. URL https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021/full.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 48–53, 2019. URL https://aclanthology.org/N19-4009.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. URL https://aclanthology.org/P02-1040.

Popović, M. chrF++: Words helping character n-grams. In *Proceedings of the Conference on Machine Translation*, pp. 612–618, 2017. URL https://aclanthology.org/W17-4770.

Post, M. A call for clarity in reporting BLEU scores. In *Proceedings of the Conference on Machine Translation*, pp. 186–191, 2018. URL https://aclanthology.org/W18-6319.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., and Kaur, R. Neural machine translation for low-resource languages: A survey. *ACM Computing Survey*, 55(11), 2023. URL https://doi.org/10.1145/3567592.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. COMET: A neural framework for MT evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2685–2702, 2020. URL https://aclanthology.org/2020.emnlp-main.213.

Sarawagi, S. and Cohen, W. W. Semi-Markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems*, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/eb06b9db06012a7a4179b8f3cb5384d3-Paper.pdf.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., et al. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. URL https://arxiv.org/abs/2211.05100.

Schapire, R. E. The boosting approach to machine learning: An overview. *Nonlinear Estimation and Classification*, pp. 149–171, 2003. URL https://link.springer.com/chapter/10.1007/978-0-387-21579-2_9.

Sennrich, R., Haddow, B., and Birch, A. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the Conference on Machine Translation*, pp. 371–376, 2016a. URL https://aclanthology.org/W16-2323.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, 2016b. URL https://aclanthology.org/P16-1162.

Shanbhogue, A. V. K., Xue, R., Saha, S., Zhang, D., and Ganesan, A. Improving low resource speech translation with data augmentation and ensemble strategies. In *Proceedings of the International Conference on Spoken Language Translation*, pp. 241–250, 2023. URL https://aclanthology.org/2023.iwslt-1.21.

Shayegh, B., Cao, Y., Zhu, X., Cheung, J. C., and Mou, L. Ensemble distillation for unsupervised constituency parsing. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RR8y0WKrFv.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. A study of translation edit rate with targeted human annotation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas: Technical Papers*, pp. 223–231, 2006. URL https://aclanthology.org/2006.amta-papers.25.

Vamvas, J. and Sennrich, R. NMTScore: A multilingual analysis of translation-based text similarity measures. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 198–213, 2022. URL https://aclanthology.org/2022.findings-emnlp.15.

van Hasselt, H., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2094–2100, 2016. URL https://ojs.aaai.org/index.php/AAAI/article/view/10295.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, Q. and Zhang, J. Parameter differentiation based multilingual neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11440–11448, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/21396.

Wang, Z., Lipton, Z. C., and Tsvetkov, Y. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 4438–4450, 2020. URL https://aclanthology.org/2020.emnlp-main.359.

Wei, B., Lu, S., Mou, L., Zhou, H., Poupart, P., Li, G., and Jin, Z. Why do neural dialog systems generate short and meaningless replies? A comparison between dialog and translation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 7290–7294, 2019. URL https://ieeexplore.ieee.org/document/8682634.

Wen, Y., Hao, Y., Cao, Y., and Mou, L. An equal-size hard EM algorithm for diverse dialogue generation. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=k5PEHHY4spM.

Wen, Y., Li, Z., Du, W., and Mou, L. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 10817–10834, 2023b. URL https://aclanthology.org/2023.acl-long.605.

Wicks, R. and Duh, K. The effects of language token prefixing for multilingual machine translation. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pp. 148–153, 2022. URL https://aclanthology.org/2022.aacl-short.19.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. URL https://arxiv.org/abs/1910.03771.

Wolpert, D. H. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. URL https://www.sciencedirect.com/science/article/pii/S0893608005800231.

Wu, C., Wu, F., and Huang, Y. One teacher is enough? Pre-trained language model distillation from multiple teachers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pp. 4408–4413, 2021. URL https://aclanthology.org/2021.findings-acl.387.

Wu, H. and Wang, H. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, pp. 856–863, 2007. URL https://aclanthology.org/P07-1108.

Wu, H. and Wang, H. Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP*, pp. 154–162, 2009. URL https://aclanthology.org/P09-1018.

Yang, Y., Lv, H., and Chen, N. A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review*, 56(6):5545–5589, 2023. URL https://link.springer.com/article/10.1007/s10462-022-10283-5.

Zaremoodi, P., Buntine, W., and Haffari, G. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 656–661, 2018. URL https://aclanthology.org/P18-2104.

Zhang, B., Williams, P., Titov, I., and Sennrich, R. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, 2020. URL https://aclanthology.org/2020.acl-main.148.

Algorithm 2 Beam Search

Input: $\mathbf{x}$: input sentence
Z: beam size
p: scorer

1 $H \leftarrow \emptyset$ ▷ candidate outputs
$B_0 \leftarrow \{\langle \textsc{bos}, 1 \rangle\}$ ▷ beam candidates
for $t = 1, 2, \cdots$ **do**
2 ▷ core of beam search
$B \leftarrow \emptyset$
for $\langle \mathbf{y}_{1:t-1}, p' \rangle \in B_{t-1}$ **do**
3 **for** $\mathrm{y} \in V$ **do**
4 $p' \leftarrow p' \cdot p(\mathrm{y} | \mathbf{y}_{1:t-1}, \mathbf{x})$
$B.\operatorname{add}(\langle \mathbf{y}_{1:t-1} \oplus \mathrm{y}, p' \rangle)$
5 $B_t \leftarrow B.\operatorname{top}(Z)$
▷ check for termination
for $\langle \mathbf{y}_{1:t}, p' \rangle \in B_t$ **do**
6 **if** $\mathrm{y}_t = \textsc{eos}$ **then**
7 $H.\operatorname{add}(\langle \mathbf{y}, p' \rangle)$
if $|H| = Z$ **then**
8 **return** $H.\operatorname{top}(1)$

A. Limitations

Our work features both algorithmic design and empirical effectiveness, but may also have limitations. A potential threat to validity is that we mainly used automatic metrics, including BLEU scores (in particular, SacreBLEU), chrF2++, TER, and COMET. We have not performed human evaluation, due to both practical and ethical concerns. Practically, it is difficult to find qualified annotators because our multilingual setting requires an annotator to know a large number of languages in addition to English. Ethically, it is inappropriate to ask annotators to study these languages before or during annotation, which is exhausting and may last for extended periods of time. That being said, our work uses the most standard metrics, including both overlapping-based and neural network-based ones. Results are generally consistent in all metrics on both datasets, with a strong statistical confidence level based on 62 translation directions. Moreover, we have provided detailed analyses and case studies to further illustrate how our EBBS works. Therefore, we deem our evaluation ethical, appropriate, and adequate.

B. Beam Search

We show the standard beam search in Algorithm 2 for a comparison with our proposed EBBS. In general, beam search takes a scorer p as the input and approximately finds the highest-scored sequence, by expanding its search tree with all the vocabulary (Lines 6–9) but only keeping the top-Z partial candidates (Line 10) at each generation step. Unlike EBBS, beam search is not specifically designed to work with multiple scorers, and we show in our main analysis that applying beam search with averaged probabilities of the ensemble components is not an ideal approach for ensemble decoding.

C. Details of Our Experiments

Dataset details. We evaluated our methods using IWSLT 2017 (Cettolo et al., 2017) and Europarl v7 (Koehn, 2005). Table 6 provides a summary of the languages.

The IWSLT 2017 translation dataset features multilingual data derived from TED talks. We followed previous work and used a standard split for zero-shot evaluation (Dabre & Kurohashi, 2017; Liu et al., 2021). In particular, IWSLT contains English-centric training data for Italian, Dutch, and Romanian, while evaluation is performed in six zero-shot directions. IWSLT is a relatively small dataset, which tests our method's ability to generalize from few languages.

Lower	Upper	BLEU↑	BLEU1↑	BLEU2↑	BLEU3↑	BLEU4↑	chrF2++↑	TER↓	COMET↑
1	1	24.98	56.75	30.08	18.84	12.46	51.97	66.13	0.8255
3	3	25.99	57.04	30.92	19.65	13.17	52.65	**65.60**	0.8326
5	5	26.10	**57.07**	**31.00**	19.76	13.28	52.75	65.63	0.8340
7	7	26.10	57.06	30.98	19.74	13.30	52.75	65.69	0.8346
9	9	**26.12**	57.02	30.99	**19.78**	**13.31**	**52.79**	65.76	**0.8352**
5	1	25.06	56.70	30.10	18.88	12.49	52.07	66.20	0.8264
5	2	25.63	56.90	30.56	19.29	12.87	52.48	65.92	0.8311
5	3	25.94	56.93	30.87	19.61	13.14	52.69	65.79	0.8330
5	4	26.03	57.02	30.95	19.69	13.21	52.74	65.75	0.8332
5	5	**26.10**	**57.07**	**31.00**	**19.76**	**13.28**	**52.75**	**65.63**	**0.8340**

Table 7. Beam size analysis for Italian-to-Dutch translation on the Europarl dataset.

Europarl is a multilingual dataset crawled from the proceedings of the European Parliament. We again followed previous work (Liu et al., 2021) and evaluated our methods with a standard split for the zero-shot setting, containing English-centric data for eight languages with a total of 56 zero-shot evaluation directions. We adopted their non-overlapping setting: in the original corpus, a sentence may be translated into multiple languages, and the non-overlapping setup chooses only one target translation for each input. This prevents potential data-leaking problems. Europarl contains more data and languages than IWSLT, which further tests our method's ability to generalize across multiple languages.

Code	Language	IWSLT	Europarl
da	Danish		✓
de	German		✓
en	English	✓	✓
es	Spanish		✓
fi	Finnish		✓
fr	French		✓
it	Italian	✓	✓
nl	Dutch	✓	✓
pt	Portugese		✓
ro	Romanian	✓	

Table 6. The languages in the IWSLT and Europarl datasets.

Implementation details. We directly adopted the neural architecture and hyperparameters in (Liu et al., 2021). In particular, we used 5- and 8-layer encoder–decoder models for IWSLT and Europarl, respectively. For both datasets, we had 512 hidden units and 8 attention heads. Our BLEU scores are based on SacreBLEU (Post, 2018) with the following specifications: `BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.5.1`.

In our presentation of beam search and the proposed EBBS, we describe the scorer as the multiplication of step-wise probabilities. In implementation, we used the sum of log-probabilities for numerical stability. Moreover, our EBBS is built on top of the popular fairseq framework (Ott et al., 2019), using their beam search implementation as the backbone. Consequently, we inherit standard beam search implementation techniques such as length normalization and max length constraints, which are not detailed in our pseudocode.

D. Analysis of Beam Size

We analyze the effect of different beam sizes on our EBBS algorithm. First, we study the setting where the lower- and upper-level beam sizes are matched. As seen in the top half of Table 7, the performance tends to increase with a larger beam size and eventually plateaus at around five, which is consistent with the practice of standard beam search (Meister et al., 2020).

Further, we analyze the setting where the upper- and lower-level beam sizes are not matched. Generally, the upper-level beam size should not exceed the lower-level beam size, because otherwise the upper-level beam may not be fully filled by the ensemble components. As shown in the bottom half of Table 7, EBBS performs better with larger upper-level beam sizes. This is understandable because a larger upper-level beam allows EBBS to explore more candidates in general.

Overall, our analysis shows that EBBS is robust and works well with a variety of beam sizes. Based on this experiment and efficiency considerations, we used a beam size of five for both upper- and lower-level beams in our main experiments.

E. Analysis of Inference Efficiency

We analyze the efficiency of our ensemble approach. As seen in Figure 3, the inference scales almost linearly, which is reasonable as we need to perform inference for all the components. The trend shows that it is computationally feasible to build an ensemble of even more components.

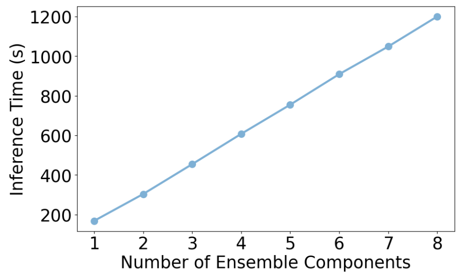

Figure 3. Inference time analysis on the test set of Italian-to-Dutch translation from the Europarl dataset. Experiments were conducted on an AMD EPYC 7313 CPU and an NVIDIA RTX A6000 GPU, with an inference batch size of 300 samples.

Dataset	Method		BLEU	Entropy
IWSLT	EBBS		19.52	-
	Direct translation	No distillation	17.46	2.46
		Direct distillation	18.10	1.62
		Union distillation	17.80	1.80
		EBBS distillation	**20.13**	1.70
Europarl	EBBS		**26.10**	-
	Direct translation	No distillation	25.33	2.06
		Direct distillation	25.44	1.44
		Union distillation	25.53	1.59
		EBBS distillation	25.92	1.51

Table 8. Entropy of various distillation techniques on Italian-to-Dutch translation.

Further, the analysis suggests that our EBBS-distilled model achieves a speedup of multiple times compared with EBBS, because after distillation the model is used by direct translation. This is a significant result, because our EBBS-based distillation not only speeds up the EBBS ensemble approach, but also retains (if not improving) the translation quality of EBBS (§4.3).

F. Entropy of Distilled Models

We would like to understand why EBBS-based distillation largely outperforms other methods, such as union distillation (§4.3). Our hypothesis is that cross-entropy distillation loss with diverse samples may lead to an overly smooth distribution, which in turn would affect the model performance (Wen et al., 2023b; Shayegh et al., 2024).

We show the average prediction entropy of our distilled models in Table 8. For some input $\mathbf{x}$ and generation step t, the prediction entropy is

$$H = -\sum_{\mathrm{y}\in V} p(\mathrm{y}|\hat{\mathbf{y}}_{1:t-1}, \mathbf{x}) \log p(\mathrm{y}|\hat{\mathbf{y}}_{1:t-1}, \mathbf{x})$$

A large entropy generally indicates that the model is less certain, producing a more uniform prediction, whereas a low entropy indicates that the model is confident, producing a more peaked distribution.

As seen in Table 8, the model without distillation yields the highest entropy, suggesting that it is uncertain about zero-shot translation probably due to a lack of training signals.

Union distillation trains the model from the union of ensemble components' outputs. It reduces prediction entropy compared with no distillation, but due to the nature of cross-entropy loss, it remains the highest among all distillation variants. Direct distillation is based on direct translation only, reinforcing the model's current belief and thus producing the lowest entropy. On the contrary, our EBBS-based distillation achieves a moderate entropy on both datasets.

It should be emphasized that the entropy analysis merely shed light on how different distillation methods behave, but the entropy itself does not indicate the quality of a model. We quote BLEU scores from Table 3, which has suggested that our EBBS-based distillation achieves similar or higher performance compared with EBBS, consistently outperforming other distillation methods.

G. Case Study

Table 9 shows examples of direct, pivot, and EBBS translations. As seen, pivot and direction translations are prone to low-quality output, but EBBS enables them to correct each other's mistakes. In the first example, say, our EBBS generally follows the sentence structure of direct translation, where the Italian word "divertimento" (*fun*) is mistranslated to the Dutch word "ontspanning" (*relaxation*), but our EBBS corrects it to "plezier" (*pleasure*), advocated by English pivoting and voted by all ensemble components.

IWSLT	
Input	ho sempre creduto che trasformare la paura in divertimento sia il dono della creatività. (*I have always believed that turning fear into fun is the gift of creativity.*)
Reference	Ik heb altijd geloofd dat het omzetten van angst in plezier de gift is van creativiteit. (*I have always believed that turning fear into joy is the gift of creativity.*)
Direct translation	Ik geloofde altijd dat het transformeren van angst in ontspanning de gift van creativiteit is (*I always believed that transforming anxiety into relaxation is the gift of creativity.*)
English-pivoting	Omdat ik altijd geloofde om angst in plezier te transformeren, is het geschenk van creativiteit. (*Because I always believed to transform fear into pleasure is the gift of creativity.*)
Romanian-pivoting	In feite hebben we altijd gedacht dat het transformeren van angst in divergentie de gift van creativiteit is. (*In fact, we have always thought that transforming fear into divergence is the gift of creativity*)
EBBS	Ik geloofde altijd dat het transformeren van angst in plezier de gift van creativiteit is. (*I always believed that transforming fear into pleasure is the gift of creativity.*)
Europarl	
Input	si poteva avvertire una forte tensione. (*a strong tension could be felt.*)
Reference	Er was veel spanning zichtbaar. (*There was a lot of tension visible.*)
Direct translation	Er was grote spanning te ontgaan. (*There was great tension to be escaped.*)
English-pivoting	Er zou veel spanningen kunnen zijn ontstaan. (*A lot of tensions could have arisen.*)
Spanish-pivoting	Mocht een sterke spanning kunnen worden aangekondigd. (*Should a strong tension can be announced.*)
EBBS	Er was veel spanning geweest. (*There had been a lot of tension.*)

Table 9. Case studies, where the source language is Italian and the target is Dutch. We provide English interpretations in (*italic*) for non-English text using Google Translate.

Acknowledgments

The research is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), a Mitacs Accelerate project, the Amii Fellow Program, the Canada CIFAR AI Chair Program, an Alberta Innovates Program, and the Digital Research Alliance of Canada (alliancecan.ca). In the title, we used the open-domain artwork "The Great Wave off Kanagawa" by Katsushika Hokusai.