Collective Wisdom: Improving Low-resource Neural Machine Translation using Adaptive Knowledge Distillation

Fahimeh Saleh
Monash University
first.last@monash.edu

Wray Buntine
Monash University
first.last@monash.edu

Gholamreza Haffari
Monash University
first.last@monash.edu

Abstract

Scarcity of parallel sentence-pairs poses a significant hurdle for training high-quality Neural Machine Translation (NMT) models in bilingually low-resource scenarios. A standard approach is transfer learning, which involves taking a model trained on a high-resource language-pair and fine-tuning it on the data of the low-resource MT condition of interest. However, it is not clear generally which high-resource language-pair offers the best transfer learning for the target MT setting. Furthermore, different transferred models may have complementary semantic and/or syntactic strengths, hence using only one model may be sub-optimal. In this paper, we tackle this problem using knowledge distillation, where we propose to distill the knowledge of *ensemble of teacher* models to a single *student* model. As the quality of these teacher models varies, we propose an effective adaptive knowledge distillation approach to dynamically adjust the contribution of the teacher models during the distillation process. Experiments on transferring from a collection of six language pairs from IWSLT to five low-resource language-pairs from TED Talks demonstrate the effectiveness of our approach, achieving up to +0.9 BLEU score improvement compared to strong baselines.

1 Introduction

Neural models have been revolutionising machine translation (MT), and have achieved state-of-the-art for many high-resource language pairs (Chen et al., 2018; Stahlberg, 2019; Maruf et al., 2019). However, the scarcity of bilingual parallel corpora is still a major challenge for training high-quality NMT models (Koehn and Knowles, 2017). Transfer learning by fine-tuning, from a model trained for a high-resource language-pair, is a standard approach to tackle the scarcity of the data in the target low-resource language-pair (Dabre et al., 2017; Kocmi and Bojar, 2018; Saleh et al., 2019; Kim et al., 2019). However, this is a one-to-one approach, which is not able to exploit models trained for multiple high-resource language-pairs for the target language-pair of interest. Furthermore, models transferred from different high-resource language-pairs may have complementary syntactic and/or semantic strengths, hence using a single model may be sub-optimal.

Another appealing approach is multilingual NMT, whereby a single NMT model is trained by combining data from multiple high-resource and low-resource language-pairs (Johnson et al., 2017; Ha et al., 2016; Neubig and Hu, 2018). However, the performance of a multilingual NMT model is highly dependent on the types of languages used to train the model. Indeed, if languages are from very distant language families, they lead to negative transfer, causing low translation quality in the multilingual system compared to the counterparts trained on the individual language-pairs (Tan et al., 2019a; Oncevay et al., 2020). To address this problem, (Tan et al., 2019b) has proposed a knowledge distillation approach to effectively train a multilingual model, by selectively distilling the knowledge from individual teacher models to the multilingual student model. However, still all the language pairs are trained in a single model with a blind contribution during training.

Algorithm 1:

```
Input    : $\mathcal{D}_{LR} := \{(\boldsymbol{x}_1, \boldsymbol{y}_1), .., (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, low resource
           dataset, Individual models $\{\theta^l\}_{l=1}^{L}$ for $L$ language pairs,
           Total training epochs: $N$
Output   : $\theta_{LR}$: low-resource model
Randomly initialize low-resource model $\theta_{LR}$ ;
$n = 0$ ;
while $n < N$ do
    $D_{LR} = random\_permute(\mathcal{D}_{LR})$ ;
    $\boldsymbol{b}_1, .., \boldsymbol{b}_M = create\_minibatches(\mathcal{D}_{LR})$ ;
    $m = 1$ ;
    while $m \leq M$ do
        // compute contribution weights;
        for $l \in L$ do
            $\Delta_l = -ppl(\theta^l(b_m))$ ;
        $\boldsymbol{\alpha} = \mathrm{softmax}(\Delta_1, .., \Delta_L)$ ;
        // compute the gradient ;
        $\boldsymbol{g} = \nabla_{\theta_{LR}} \mathcal{L}_{ALL}^{adaptive}(\boldsymbol{b}_m, \theta_{LR}, \{\theta^l\}_1^L, \boldsymbol{\alpha})$ ;
        // updates the parameters using the optimiser ADAM ;
        $\theta_{LR} = \mathrm{update\_param}(\theta_{LR}, \boldsymbol{g})$ ;
        $m = m + 1$ ;
    $n = n + 1$ ;
```

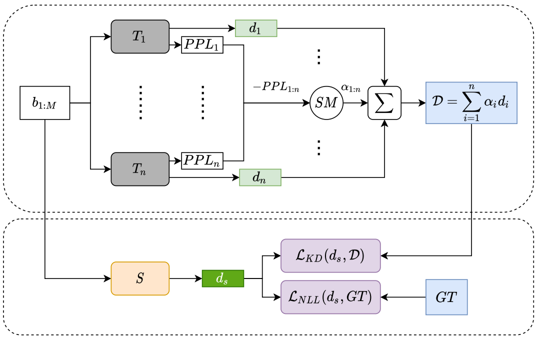
T_1
d_1
PPL_1
$b_{1:M}$
$-PPL_{1:n}$
SM
$\alpha_{1:n}$
$\sum$
$\mathcal{D} = \sum_{i=1}^{n} \alpha_i d_i$
T_n
PPL_n
d_n
S
d_s
$\mathcal{L}_{KD}(d_s, \mathcal{D})$
$\mathcal{L}_{NLL}(d_s, GT)$
GT

Figure 1: Adaptive Knowledge Distillation. **(Top)** Teachers' contribution weight calculation. $T_{1:n}$ and $d_{1:n}$ denote the freezed teacher models and their corresponding probability distributions respectively. **(Bottom)** Training the student with adaptive knowledge distillation. S, SM, and GT denote the student model, softmax function, and ground-truth respectively.

In this paper, we propose a many-to-one transfer learning approach which can effectively transfer models from multiple high-resource language-pairs to a target low-resource language-pair of interest. As the fine-tuned models from different high-resource language pairs can have complementary syntactic and/or semantic strengths in the target language-pair, our idea is to distill their knowledge into a single student model to make the best use of these teacher models. We further propose an effective adaptive knowledge distillation (AKD) approach to dynamically adjust the contribution of the teacher models during the distillation process, enabling making the best use of teachers in the ensemble. Each teacher model provides dense supervision to the student via dark knowledge (Hinton et al., 2015) using a mechanism similar to label smoothing (Szegedy et al., 2016; Müller et al., 2019), where the amount of smoothing is regulated by the teacher. In our AKD approach, the label smoothing coming from different teachers is combined and regulated, based on the loss incurred by the teacher models during the distillation process.

Experiments on transferring from a collection of six language pairs from IWSLT to five low-resource language-pairs from TED Talks demonstrate the effectiveness of our approach, achieving up to +0.9 BLEU score improvements compared to strong baselines.

2 Adaptive Knowledge Distillation

We address the problem of low-resource NMT, assuming that we have access to models for high resource languages, and data for low resource model. Our approach relies on two main steps, (i) Transferring from high-resource to low-resource language-pairs by fine tuning the high-resource models using the small amount of bilingual data, and (ii) Adaptive distillation of knowledge from the teacher models to the student model.

More specifically, given a training dataset for a low-resource language-pair, $\mathcal{D}_{LR} := \{(\boldsymbol{x}_1, \boldsymbol{y}_1), .., (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ and multiple individual high resource NMT models $\{\theta^l\}_{l=1}^{L}$ fine-tuned on $\mathcal{D}_{LR}$ (teachers), we are interested in training a single NMT model (student) by adaptively distilling knowledge from all teachers based on their effectiveness to improve the accuracy of the student. Knowledge distillation (KD) is a process of improving the performance of a simple *student* model by using a distribution over soft labels obtained from an expert *teacher* model instead of hard ground-truth labels (Hinton et al., 2015). The training objective to distill the knowledge from a single teacher to the student involves,

$$-\sum_{\boldsymbol{x},\boldsymbol{y} \in \mathcal{D}_{LR}} \sum_{t=1}^{|\boldsymbol{y}|} \sum_{v \in V} Q(v|\boldsymbol{y}_{<t}, \boldsymbol{x}, \theta^l) \log P(v|\boldsymbol{y}_{<t}, \boldsymbol{x}, \theta_{LR}) \tag{1}$$

where θ^l and θ_{LR} are the parameters of the teacher and student models, respectively. $P(. \mid .)$ is the conditional probability with the student model and $Q(. \mid .)$ denotes the output distribution of the teacher model. According to Equation 1, knowledge distillation provides dense training signal as *each* word in the vocabulary (V) contributes to the training objective, regulated by a weight coming from the teacher.

This is in contrast to the negative log-likelihood training objective, which only provides supervision signal based on the correct target words according to the bilingual training data,

$$\mathcal{L}_{NLL}(\mathcal{D}_{LR}, \theta_{LR}) := -\sum_{\boldsymbol{x},\boldsymbol{y} \in \mathcal{D}_{LR}} \sum_{t=1}^{|\boldsymbol{y}|} \log P(y_t | \boldsymbol{y}_{<t}, \boldsymbol{x}, \theta_{LR}). \tag{2}$$

Given a collection of teacher models $\{\theta_l\}_{l=1}^{L}$, we pose the following training objective,

$$\begin{aligned} &\mathcal{L}_{KD}^{adaptive}(\mathcal{D}_{LR}, \theta_{LR}, \{\theta^l\}_1^L, \boldsymbol{\alpha}) := \\ -\sum_{\boldsymbol{x},\boldsymbol{y} \in \mathcal{D}_{LR}} \sum_{l=1}^{L} \alpha_l \sum_{t=1}^{|\boldsymbol{y}|} &\sum_{v \in V} Q(v | \boldsymbol{y}_{<t}, \boldsymbol{x}, \theta^l) \log P(v | \boldsymbol{y}_{<t}, \boldsymbol{x}, \theta_{LR}) \end{aligned} \tag{3}$$

where α_l regulates the contribution of the l-th teacher. We dynamically adjust the contribution weights over the course of the distillation process, in order to effectively address the knowledge gap of the student during the training process. This is achieved based on the rewards (negative perplexity) attained by the teachers on the data, where these values are passed through a softmax transformation to turn into a distribution. To stabilize these contribution weights over the course of the training process, we smooth them using a running geometric average. The student model is trained end-to-end with a weighted combination of losses coming from the ensemble of teachers and the data,

$$\mathcal{L}_{ALL}^{adaptive}(\mathcal{D}_{LR}, \theta_{LR}, \{\theta^l\}_1^L, \boldsymbol{\alpha}) := \quad \lambda_1 \mathcal{L}_{NLL}(\mathcal{D}_{LR}, \theta_{LR}) + \lambda_2 \mathcal{L}_{KD}^{adaptive}(\mathcal{D}_{LR}, \theta_{LR}, \{\theta^l\}_1^L, \boldsymbol{\alpha}) \tag{4}$$

where $\lambda_1 = 0.5$ and λ_2 is started from 0.5 and gradually increased to 3 following the annealing function of (Bowman et al., 2015) in our experiments. Our approach is summarized in Algorithm 1 and Figure 1.

3 Experiments

3.1 Settings

Data. We conduct our experiments on the European languages of IWSLT and TED datasets. The language pairs with more than 100K training data are considered as high-resource and the ones less than 15k are assumed as low-resource. The high-resource models are trained on IWSLT2014 (ru,de,it,pl,nl,es-en). IWSLT 2014 MT task data (sl-en) (Cettolo et al., 2014), and TED talk data (gl/et/nb/eu-en) (Qi et al., 2018) are used as low-resource languages. Detail about the preprocessing step and the statistics of data and language codes based on ISO 639-1 standard[1] are listed in Section 1.1 of Appendix A.

Training configuration. Individual low-resource and high-resource NMT models are trained on the low-resource data. The first trained from scratch and the later by finetuning with the vanilla transformer architecture. For multilingual NMT, we train a single model with all high-resource and the up-sampled of low-resource language pairs by using a decoder language embedding layer to identify the type of language during the inference step. Multilingual selective knowledge distillation (Tan et al., 2019b) is trained with all language pairs while matching the outputs of each low-resource model simultaneously through knowledge distillation. For training our approach, we fine-tune the high-resource models with low-resource languages and treat them as teachers. When training on the low-resource language, we load teacher models into memory and train a single low-resource model (student) from scratch while using the weighted average of teachers' probabilities based on their contribution weight. In order to make clear how different teachers contribute during training the student, we illustrate contribution weights of all teachers for first 30 iterations of different mini-batches during the training in Figure 2.

Model configuration. All models are trained with Transformer architecture (Vaswani et al., 2017), with the model hidden size of 256, feed-forward hidden size of 1024, and 2 layers, implemented in Fairseq framework (Ott et al., 2019). We use the Adam optimizer (Kingma and Ba, 2015) and an inverse square root schedule with warm-up (maximum LR 0.0005). We apply dropout and label smoothing with a rate of 0.3 and 0.1 respectively. The source and target embeddings are shared and tied with the last layer. We train with half-precision floats on one V100 GPU, with at most 4028 tokens per batch.

[1] http://www.loc.gov/standards/iso639-2/php/English_list.php

MT Task	Individual	Transfer Learning from (x→en) model						Multi-Lingual		Multi-Teacher
x→en	student	ru	de	it	es	pl	nl	Uniform	Selec. KD	Adap. KD
sl	10.58	10.36	14.09	13.29	16.89	17.63	16.67	15.97	16.17	**18.35**
nb	26.38	32.24	32.77	31.90	30.04	30.66	32.86	30.06	31.08	**33.72**
gl	13.87	11.88	17.66	21.90	**27.49**	16.67	17.05	25.27	25.08	24.50
eu	6.50	9.54	10.68	9.92	11.00	10.50	10.02	10.11	11.03	**11.38**
et	10.15	12.18	14.85	14.93	15.53	14.25	13.66	14.91	15.15	**16.20**

Table 1: BLEU scores of the translation tasks from five languages into English. Selective KD is based on (Tan et al., 2019b).

Contribution weight setting	gl-en	nb-en
Adaptive contribution	24.50	33.72
Equal contribution	19.10	32.60

Table 2: Effect of different contribution settings.

Contribution temperature	eu-en	sl-en
with adaptive temp	11.38	18.35
without temp	10.52	18.05

Table 3: Effect of adaptive temperature.

3.2 Results

In Table 1, we compare our approach with individual NMT models, transferred models from high-resource language pairs, multilingual NMT, and multilingual selective knowledge distillation (Tan et al., 2019b). We selected the best models according to the SacreBLEU[2] score on the validation set. In our experiments, bold numbers indicate the best results and underlined numbers show the second best ones. Transfer learning results are inline with the language family relationships (Littell et al., 2017). The high-resource languages which are linguistically close to the low-resource languages have the most impact on low-resource model's improvement. Likewise, the contribution weights of different teachers are consistent with the performance of the teachers as hypothesized (See results in Table 1 and Figure 2). According to Table 1, the multilingual models (with and without knowledge distillation) are less accurate than at least one of the transferred models from high-resource languages[3]. This suggests a weak link may exist between the impact of each high-resource language and its contribution during the training multilingually. Adaptive knowledge distillation compensates this blind collaboration between teachers by weighting the teachers' contributions particularly for the cases where majority of teachers and student are linguistically close such as "nb-en". The qualitative examples are presented in Section 1.4 of Appendix A. It is worth noting that, we empirically observed when there is more diversity in teachers (e.g, in case of "gl-en" in Table 1), adaptive KD underperforms compared to the best teacher and we hypothesise this happens because there is an empirically dominant teacher ("es"). This observation suggests that a prior effort for choosing the proper teacher languages (e.g., based on the language family information) will directly impact the performance of the low-resource NMT model.

4 Analysis

4.1 Contribution Weight Analysis

To analyse the effect of teachers' contribution weights, we compare two different contribution settings: *(i) Adaptive contribution:* which assigns the contribution weights to all the teachers based on their performance per mini-batch as explained in Section 2. *(ii) Equal contribution:* which gives all the teachers the same contribution weights. According to Table 2, the equal contribution setting is not as effective as the adaptive contribution especially for the languages with more inconsistent teachers (based on BLEU score) e.g., "gl-en".

4.2 Contribution Temperature Scaling

Through the experiments, we observed that when most of the teachers do not agree (in terms of perplexity), a constant temperature is not an ideal option. An alternative is to adaptively change the value of the temperature given the *agreement* among the teachers determined based on the distance between the

[2] SacreBLEU signature: BLEU+case.mixed+numrefs.1+ smooth.exp+tok.none+version.1.3.1

[3] Except for the Basque language which is extremely low-resource and is linguistically as distant to all the languages in the multilingual setting.

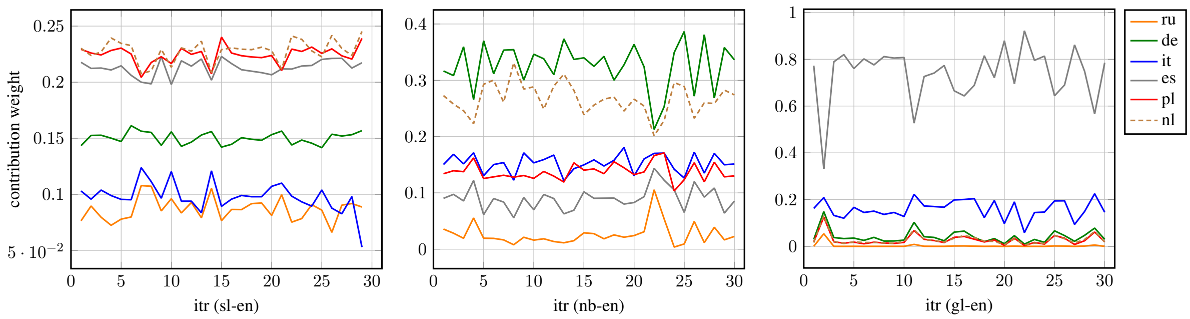

Figure 2: Teachers' contribution weights during the training of low-resource NMT models for "sl-en", "gl-en", and "nb-en" language pairs, first 30 iterations for different mini-batches.

maximum and minimum perplexity between teachers which can be formulated as:

$$\tau = \frac{1 - (\max(S) - \min(S))}{N} \tag{5}$$

where S is the output of the softmax operation on the negative perplexity of all N teachers and $(\max(\mathrm{S}) - \min(\mathrm{S}))$ is inversely proportional to the extent of the agreement between teachers. Such temperature scaling encourages the contribution of better teachers in case of the existence of a disagreement, while it allows similar contributions when all teachers agree on a mini-batch. Table 3 shows the effect of adaptive temperature for two languages.

5 Conclusion

In this paper, we present an adaptive knowledge distillation approach to improve NMT for low-resource languages. We address the inefficiency of the original transfer learning and multilingual learning by making wiser use of all high-resource languages and models in an effective collaborative learning manner. Our approach shows its effectiveness in translation of low-resource languages especially when there are complementary knowledge in multiple high-resource languages from the same linguistic family and it is not explicitly clear which language has more impact in every mini-batch of low-resource training data. Experiments on the translation of five extremely low-resource languages to English show improvements compared to the strong baselines.

Acknowledgements

This work was supported by the Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE) `(www.massive.org.au).`

References

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In *Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam*, volume 57.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *Institute for Anthropomatics and Robotics*, 2(17/03):17.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *ACL 2017*, page 28.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2019. A survey on document-level machine translation: Methods and evaluation. *arXiv preprint arXiv:1912.08494*.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 4696–4705.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. *arXiv preprint arXiv:2004.14923*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. FAIRSEQ: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana, June. ACL.

Fahimeh Saleh, Alexandre Bérard, Ioan Calapodescu, and Laurent Besacier. 2019. Naver Labs Europe's systems for the document-level generation and translation task at WNGT 2019. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 273–279.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. ACL.

Felix Stahlberg. 2019. Neural machine translation: A review. *arXiv preprint arXiv:1912.02047*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.

Xu Tan, Jiale Chen, Di He, Yingce Xia, QIN Tao, and Tie-Yan Liu. 2019a. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 962–972.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019b. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*.

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.

Appendix A:

Experiments and Analysis

Data Preprocessing

We use the European languages of IWSLT[4] and TED[5] datasets in our experiments as listed in Table 4. We filter the parallel corpus with `langid.py` (Lui and Baldwin, 2012) and remove sentences with a length ratio greater than 1.5. All the sentences are first tokenized with the Moses tokenizer[6] and then segmented with BPE segmentation (Sennrich et al., 2016) with a learned BPE model by 32k merge operations on all languages. We keep the output vocabulary of the teacher and student models the same to make the knowledge distillation feasible.

High-resource Languages						
Language name	Russian	German	Italian	Spanish	Polish	Dutch
Code	ru	de	it	es	pl	nl
size (#sent(k))	153\6.9\5.5	160\7.2\6.7	167\7.5\5.5	169\7.6\5.5	128\5.8\5.4	153\6.9\5.3
Low-resource Languages						
Language name	Basque	Galician	Norwegian	Slovenian	Estonian	
Code	eu	gl	nb	sl	et	
size (#sent(k))	3.3\0.3\0.3	8.4\0.6\1	14\0.8\0.8	14.5\1.4\0.6	7.7\0.7\1	

Table 4: Language names and statistics for bilingual resources (Language→English), (train\dev\test)

Translation Examples

Table 5 showcases the generated English translations by the individual student, all the teachers, and student trained through adaptive knowledge distillation from Norwegian language. This example shows that while there is a diversity between different teachers' translations e.g., for the verb of *"provoke"*, the student is impacted by the agreement of the majority of teachers. Moreover, this example shows that our adaptive KD model captures the best of all teachers resulting in a higher quality translation.

Ref	And great creativity is needed to do what it does so well : to provoke us to think differently with dramatic creative statements .
Individual	kepler **great** mission mission to do it as **well** : to grow us to think **with dramatic** creativity .
Teacher (ru-en)	and the first **creativity** needed **to do what it does** : to promote us to think about the **dramatic** creativity .
Teacher (de-en)	now , the future **creativity** needs to do it as it does : to **provoke** us to **think differently with dramatic** creative expression .
Teacher (it-en)	now , the future **creativity** is needed **to do what it does** so **well** : to provocate us to **think differently** about **dramatic** reactive .
Teacher (es-en)	the future of **creativity** to do that as it's doing so good : to provocate us to **think differently** about **dramatic** creativity .
Teacher (pl-en)	the future of **creativity to do what it does** so good : to promise others **with dramatic** creativity .
Teacher (nl-en)	now , the frequent **creativity** is to make it that it makes so good : to **provoke** us with **dramatic** creative .
Proposed Adapt. KD	now , they need **great creativity to do what it does** so **well** : **provoke** us to **think differently with dramatic** creativity.

Table 5: The generated outputs from the individual student, all teachers, and student trained with multi-teachers (Proposed Adapt. KD) for "nb-en" MT task. Some of the correct keyword translations are indicated with green color while hallucinations are represented by red. The bold-green shows the best of the teachers' output which is also captured with the student.

[4] https://wit3.fbk.eu/

[5] https://github.com/neulab/word-embeddings-for-nmt

[6] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl