Improving Neural Machine Translation Using Noisy Parallel Data through Distillation

Praveen Dakwale
Informatics Institute
University of Amsterdam
p.dakwale@uva.nl

Christof Monz
Informatics Institute
University of Amsterdam
c.monz@uva.nl

Abstract

Due to the scarcity of parallel training data for many language pairs, quasi-parallel or comparable training data provides an important alternative resource for training machine translation systems for such language pairs. Since comparable corpora are not of as high quality as manually annotated parallel data, using them for training can have a negative effect on the translation performance of an NMT model. We propose distillation as a remedy to effectively leverage comparable data where the training of a student model on combined clean and comparable data is guided by a teacher model trained on the high-quality, clean data only. Our experiments for Arabic-English, Chinese-English, and German-English translation demonstrate that distillation yields significant improvements compared to off-the-shelf use of comparable data and performs comparable to state-of-the-art methods for noise filtering.

1 Introduction

Traditional machine translation systems are trained on parallel corpora consisting of sentences in the source language aligned to their translations in the target language. However, for many language pairs substantial amounts of high-quality parallel corpora are not available. On the other hand, for many languages, another useful resource known as comparable corpora can be obtained relatively easily in substantially larger amounts. Such comparable corpora can be created by crawling large monolingual data in the source and target languages from multilingual news portals such as Agence France-Presse (AFP), BBC news, Euronews etc. Source and target sentences in these monolingual corpora are then aligned by automatic document and sentence alignment techniques (Munteanu and Marcu, 2005). Such a bitext extracted from comparable data is usually not of the same quality as annotated parallel corpora. Recent research has shown that building models from low-quality data can have a degrading effect on the performance of recurrent NMT models (Khayrallah and Koehn, 2018). Therefore, there is a growing interest in filtering and sampling techniques to extract high-quality sentence pairs from such large noisy parallel texts.

Recently, the "Parallel corpus filtering" (Koehn et al., 2018) shared task was held at WMT-2018. This task aims at extracting high-quality sentence pairs from Paracrawl[1], which is a large noisy parallel corpus. Most of the participants in this task, used rule-based pre-filtering followed by a classifier-based scoring of sentence pairs (Barbu and Barbu Mititelu, 2018; Junczys-Dowmunt, 2018; Hangya and Fraser, 2018). A subset sampled with a fixed number of target tokens is then used to train recurrent NMT systems in order to evaluate the relative quality of the filtered bitexts. Some of the submissions show good translation performance for the German-English translation task by training on the filtered bitext only. In this paper, we propose a strategy to leverage additional low-quality bitexts without any filtering when used in conjunction with a high-quality parallel corpus. Motivated by the "knowledge distillation" frame-

[1] https://paracrawl.eu/

Arabic-English (ISI bitext)	
Src:	وقررت وزارة العدل الهولندية ابعاده رغم طلب الاردن تسليمه في اطار قضية تهريب مخدرات.
Trg:	The Dutch justice ministry decided to expel the Iraqi Kurd despite Amman's demand that he be handed over to Jordanian authorities.
Human:	The Dutch Justice Ministry decided to deport him, despite Jordan's request to hand him over as part of a drug smuggling case.
Chinese-English (ISI bitext)	
Src:	美国提出的报复清单是中国政府绝对不能接受的。
Trg:	And the Chinese side would certainly not accept the unreasonable demands put forward by the Americans concerning the protection of intellectual property rights.
Human:	The revenge list proposed by America will definitely not be accepted by Chinese government.
German-English (Paracrawl)	
Src:	Der Elektroden Schalter KARI EL22 dient zur Füllstandserfassung und -regelung von elektrisch leitfähigen Flüssigkeiten .
Tgt:	The KARI EL22 electrode switch is designed for the control of conductive liquids .
Human:	The electrode switch KARI EL22 is used for level detection and control of electrically conductive liquids.

Table 1: Noisy sentence pair example from ISI bitext (Arabic-English and Chinese-English) and Paracrawl (De-En). Fragments in red in either source or target side has no corresponding equivalent fragment on the respective aligned side.

work of Hinton et al. (2014), we propose "distillation" as a strategy to exploit comparable training data for training an NMT system. In our distillation strategy, we first train a teacher model on the clean parallel data, which then guides the training of a final student model trained on the combination of clean and noisy data. Our experimental results demonstrate that for Arabic-English and Chinese-English translation, distillation not only helps to successfully utilize noisy comparable corpora without any performance degradation, but it also outperforms one of the best performing filtering techniques reported in Koehn et al. (2018). In addition, we conduct similar experiments for German-English translation and observe that while simply adding noisy data to the training data pool degrades performance, our distillation approach still yields slight improvements over the baseline.

In Section 2, we discuss the relevant literature in NMT as well as in other deep learning based tasks which aim to utilize low quality training corpus. In Section 3, we provide a brief discussion of the type of noise in the comparable data, the architecture of the NMT model used in our experiments, and the knowledge distillation framework proposed by Hinton et al. (2014). In Section 4, we describe our strategy to use knowledge distillation for training with noisy data. We discuss our experimental settings including datasets and parameters in Section 5 and results in Section 6.

2 Related work

Khayrallah and Koehn (2018) reported that NMT models can suffer substantial degradation from adding noisy bitexts when compared to a baseline model trained on high-quality parallel text only. The "Parallel corpus filtering" (Koehn et al., 2018) task evaluated submissions based on NMT systems trained only on the bitext filtered from Paracrawl. However, given that many language pairs have at least some small amount of high-quality parallel corpora (which is also used by many of the participants to train a classifier for scoring the noisy data), it is important to investigate whether a bitext filtered using these proposed techniques results in any additional improvements in conjunction with the original high-quality data. Filtering techniques involve discarding a sentence pair with low confidence score. However, a sentence pair with a low score may still have fragments in the source and target sentences which can provide useful contexts. Our results show that for a recurrent NMT model, filtering the noisy bitext below a specific threshold using one of the best techniques submitted to the filtering task (known as "Dual conditional cross entropy filtering" (Junczys-Dowmunt, 2018)) yields only small improvements.

In the machine learning literature, various methods have been proposed for efficient learning with label noise. One of the recent methods is the bootstrapping (Reed et al., 2014) approach where improved labels for noisy or unlabeled data can be obtained by predictions of another classifier. For NMT, forward translations of the noisy bitext can be used as a variant of bootstrapping where the target side of the noisy bitext can be replaced by translations of the source sentence obtained by a model trained on the high-quality data. However, a better alternative for NMT would be to use back-translations (Sennrich et al., 2016), i.e., to replace the source side of the noisy bitext by translations of the target side obtained by a model trained in the reverse direction. Our experiments show that although backward translations of noisy data cause lower degradations than the original noisy data, they provide only moderate improvements. Moreover, cleansing the comparable data by back-translation is expensive as it requires the generation of pseudo source sentences using beam search decoding. Fine-tuning (Miceli Barone et al., 2017) is a well-known technique for domain adaptation for NMT but can also be used as a possible solution for training with noisy data where the idea is to first pre-train on noisy data and then continue training on high-quality data.

Our experiments show that when using noisy data for training NMT models, fine-tuning fails to provide any additional improvements. Moreover, bootstrapping based on filtering and back-translation, as explained above, show only small improvements over a model trained on high-quality data only. In order to overcome the dependence on filtering-based data selection or other data cleaning approaches and to leverage all available noisy data, in this paper, we propose knowledge distillation based training on combined clean and noisy sentence pairs.

It is very important to note that, as has been pointed out in Koehn et al. (2018), the aim of *"Parallel corpus filtering"* task proposed at WMT18 was not to select data relevant for a targeted domain, but to focus on the selection of high quality data that is relevant to all domains. Similarly, in this paper, we do not aim to propose a technique for domain adaptation for NMT but to propose a technique to leverage low quality or noisy training data for training high performing NMT models.

Although knowledge distillation has been used as a solution to other problems of NMT such as model compression (Kim and Rush, 2016), domain adaptation (Dakwale and Monz, 2017) or transfer learning for low-resource languages (Chen et al., 2017) and for leveraging noisy data for image recognition (Li et al., 2017), our approach is the first attempt to exploit distillation for training NMT systems with noisy data.

3 Background

3.1 Noise in the training corpora

Khayrallah and Koehn (2018) analyzed the Paracrawal corpus, identifying various types of noise in this corpus. They found that although there are some instances of incorrect language, untranslated sentences, and non-linguistic characters, the majority of noisy samples (around 41%) are misaligned sentences due to faulty document or sentence alignment. This results in alignments of incorrect source to target sentence fragments.

Similarly, a well-known noisy bitext commonly used for training machine translation systems for Arabic-English and Chinese-English is the ISI bitext created by automatically aligning sentences from monolingual corpora extracted from AFP and Xinhua, respectively (Munteanu and Marcu, 2005). This alignment method first searches for articles representing similar stories in two separate monolingual corpora for source and target languages using cross-lingual information retrieval with the help of a dictionary. Then parallel sentences are aligned by calculating word overlaps between each candidate sentence pair followed by a maximum entropy classifier. Since the bitexts are extracted from monolingual corpora for source and target languages, there is rarely any noise due to misspelling, wrong re-ordering or non-linguistic characters. The majority of noise in the resulting aligned bitext is due to limitations of the sentence alignment technique often resulting in sentence pairs which are partial translations of each other with additional fragments on either the source or target side.

Table 1 shows some examples of noisy sentence pairs for German-English (from the Paracrawl corpus) and Arabic-English (from the ISI bitext). The fragments marked red in the source sentence have no correspondence on the target side. We refer the reader to (Khayrallah and Koehn, 2018) and (Munteanu and Marcu, 2005) for a more detailed description of the types of noise in the respective

corpora.

3.2 Neural Machine Translation

We employ an NMT system based on Bahdanau et al. (2014). This is a simple encoder-decoder network where both the encoder and decoder are multilayer recurrent neural networks (we use LSTM's). Given an input sentence $[(x_1, x_2, ..., x_n)]$, the encoder converts it into a sequence of hidden state representations $[(h_1, h_2, ..., h_n)]$.

$$h_i = f_{encoder}(x_i, h_{i-1}) \tag{1}$$

Here, $f_{encoder}$ is an LSTM unit. The decoder is another multi-layer RNN which predicts a target sequence $y = (y_1, y_2, ...y_m)$. The probability of generation of a token y_i at position *'i'* on the target side is conditioned on the last target token y_{i-1}, the current hidden state of the decoder s_j, and the context vector c_j which is a conditional representation of the source sequence relevant to target position *'i'*. The probability of the sentence is computed as the product of the probabilities of all target tokens.

$$p(\mathbf{y}) = \prod_{j}^{m} p(y_j|y_1, ...y_{j-1}, \mathbf{x}) = \prod_{j}^{m} g(y_{j-1}, s_j, c_j) \tag{2}$$

g is a multi-layer feed-forward neural network with a nonlinear transformation. A softmax layer is applied on the output of the feedforward network *g*, which generates the probability of each word in the target vocabulary. Here, s_j is the hidden state representation corresponding to each token in the target sequence generated by the decoder RNN.

$$s_j = f_{dec}(s_{j-1}, y_{j-1}, c_j) \tag{3}$$

The context vector c_j is computed using an attention mechanism (Luong et al., 2015) as the weighted sum of the hidden states h_i of the encoder.

$$c_j = \sum_{i=1}^{n} \alpha_{ji} h_i \tag{4}$$

where α_{ji} are attention weights corresponding to each encoder hidden state output h_i calculated as follows :

$$\alpha_{ji} = \frac{exp(a(s_{j-1}, h_i))}{\sum_{k=1}^{n} exp(a(s_{j-1}, h_k))} \tag{5}$$

Activations $a(s, h)$ are calculated by using a scoring function such as dot product between the current decoder state s_{j-1} and each of the hidden states h_i of the encoder. The end-to-end network is trained by minimizing the negative log-likelihood over the training data. The log-likelihood loss is defined as

$$L_{NLL}(\theta) = -\sum_{j=1}^{n}\sum_{k=1}^{|V|}(y_j = k) * \log(p(y_j = k|x;\theta)) \tag{6}$$

Where y_j is the output distribution generated by the network at each time-step and *k* is the true class label, i.e., the reference target word at each time step selected from a fixed vocabulary *V*. The outer summation is the total loss computed as the sum over the complete target sequence.

3.3 Knowledge Distillation

Knowledge Distillation is a framework proposed in Hinton et al. (2014) for training compressed "student" networks by using supervision from a large teacher network. Assuming, we have a teacher network with large dimension size trained on a large amount of data, a smaller student network with much smaller dimension size can be trained to perform comparable or even better than the teacher by learning to mimic the output distributions of the teacher network on the same data. This is usually done by minimizing cross-entropy or KL-divergence loss between the two distributions. Formally, if we have a teacher network trained on the same data and with a learned distribution $q(y|x;\theta_T)$, the student network (model parameters represented by θ) can be trained by minimizing the following loss:

$$L_{KD}(\theta, \theta_T) = -\sum_{k=1}^{|V|} \mathrm{KL}\Big(q(y|x;\theta_T)\; p(y|x;\theta)\Big) \tag{7}$$

where θ_T is the parameter distribution of the teacher network. Commonly, this loss is interpolated with the log-likelihood loss which is calculated with regard to the target labels for the in-domain data:

$$L(\theta, \theta_T) = (1-\lambda)L_{NLL}(\theta) + \lambda L_{KD}(\theta, \theta_T) \tag{8}$$

In order to allow the student network to encode the similarities among the output classes, Hinton et al. (2014) suggests to generate a smoother distribution called 'soft-targets' by increasing the temperature of the softmax of both teacher and student network.

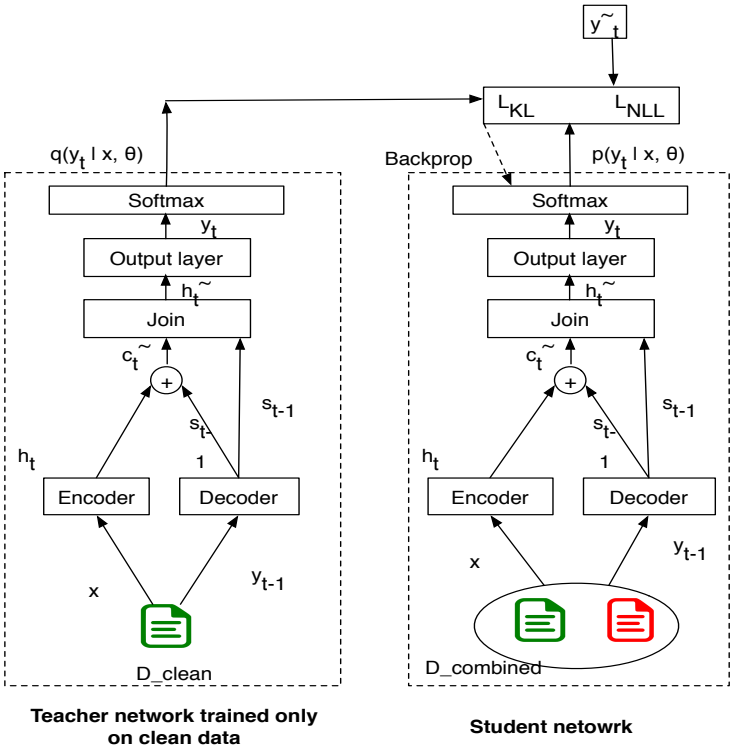

Figure 1: Distillation for noisy data. Both the teacher and student network have same architecture. Teacher network is trained only on the clean data, student network is trained for two losses : L_{NLL} wrt target labels and L_{KL} wrt to output distribution of teacher network

4 Knowledge distillation for noisy data

We discuss the main intuition and idea behind using knowledge distillation for noisy labels. A detailed analysis is given in Li et al. (2017). As shown in Figure 1, the idea is to first train the teacher model f on the clean data D_{clean} and then transfer the knowledge from the teacher to a student network which is trained on the entire dataset D by optimising the following loss:

$$L_D(y_i, f(x_i)) = \lambda l(y_i, f(x_i)) + (1-\lambda) l(s_i, f(x_i)) \quad (9)$$

where $s_i = f_{D_{clean}}(x_i)/\tau$ and τ is the temperature of the softmax. In equation 9, the student model is trained on the combination of two loss functions, the first term is the cross-entropy loss l between the prediction of the student model and the ground truth y_i, while the second term is the cross-entropy or KL-divergence between the output distributions of the student model and the teacher model. λ is a parameter to balance the weight between the two losses. Assuming the second loss to be cross-entropy, Equation 9 can be re-written as:

$$L_D(y_i, f(x_i)) = l(\lambda(y_i) + (1 - \lambda(s_i), f(x_i))) \quad (10)$$

Li et al. (2017) define $y_i^\lambda = \lambda y_i + (1 - \lambda)(s_i, f(x_i))$ as pseudo-label which is a combination of the given noisy label y_i and the prediction s_i from the teacher model. They provide an analysis based on the comparison between the risks involved in training directly on the noisy labels or training on the boot-strapped labels as compared to training on the pseudo label as defined above. They show that training on the pseudo label, for some values of λ defined through distillation, involves lower risks than direct training or bootstrapping. Therefore, a better model can be trained by driving the pseudo labels closer to the ground truth label. In case of comparable corpora training for NMT, instead of learning only from uncertain ground truth labels, the student model also benefits from the predictions of the teacher model while learning to imitate it.

5 Experiments

5.1 Comparisons

We compare our technique to standard scenarios of training on clean and noisy data. Further, we compare to the commonly used strategy of fine-tuning as well as back-translation which is an adapta-

	Clean		Noisy	
	Source	Size	Source	Size
Arabic-English	LDC	300k	ISI bitext	1.1m
Chinese-English	LDC	550k	ISI bitext	550k
German-English	WMT-17	5.1M	Para$_{rn}$	5.1M
German-English	WMT-17	5.1M	Filt$_{toks=100M}$	4.6M

Table 2: Datasets and statistics. Para$_{rn}$ = Randomly sampled subset of Paracrawl. Filt$_{toks=100M}$ = 100 million target token subsample submitted by (Junczys-Dowmunt, 2018)

tion of self-learning or bootstrapping methods. We carry out the following experimental comparisons:

- **Training on parallel data only:** The standard practice in NMT is to train on the high-quality parallel data only. This experiment is also the primary baseline for comparing the proposed method.

- **Training on comparable data only:** We conduct this experiment to demonstrate the substantial difference between the performance of the models trained on only noisy data or only on clean data.

- **Training on combined comparable and parallel data:** This experiment demonstrates the effect of adding comparable data to the baseline training data pool.

- **Fine-tuning:** The standard practice commonly used for domain adaptation. For noisy data, the idea is to first train the model on noisy data and then continue training on clean data.

- **Back-translation:** Back-translation has been proposed as a method to incorporate additional monolingual data for NMT (Sennrich et al., 2016). This is done by training an NMT system in reverse of the desired direction thus obtaining pseudo-source sentences for the additional monolingual target sentences. By applying back-translation, we discard the original source sentence in the comparable data and replace them with the pseudo-source sentences. The back-translated comparable data is then added to the clean parallel data.

- **Dual cross entropy filtering:** As discussed in the introduction, (Junczys-Dowmunt, 2018)) reported the best results for the Parallel Corpus Filtering task for WMT-18. They used the dual cross-entropy method in which sentence pairs in the noisy corpus are ranked based on forward and backward losses for each sentence pair with respect to NMT models trained on clean data in forward and reverse direction. We consider this filtering method as a competitive baseline for our approach.

Note that back-translation requires beam-search based decoding which is quite expensive for large amount of comparable data.

5.2 Datasets and Parameters

We conduct experiments for Arabic to English, Chinese to English, and German-English NMT. As a commonly used representative of comparable data, we consider all AFP sources from the ISI Arabic-English bitext (LDC2007T08) with a size of 1.1M sentence pairs and Xinhua news sources for the Chinese-English bitext (LDC2007T09) with a size of 550K sentence pairs. Both corpora are created by automatically aligning (Munteanu and Marcu, 2005) sentences from monolingual corpora. For Arabic-English, we compose the parallel data consisting of 325k sentence pairs from various LDC catalogues [2]

For Chinese-English, a parallel text of 550k parallel sentence pairs from LDC catalogues[3] is used. Note that for Arabic-English, the size of the comparable corpus is approximately 4 times that of the parallel data while for Chinese-English, the comparable corpora size is the same as that of the parallel corpus[4]. A byte pair encoding of size 20k is trained on the parallel data for the respective languages. NIST MT05 is used as dev set for both language pairs and MT08, MT09 as test set for Arabic-English and MT-06, MT-08 as test set for Chinese-English. Translation quality is measured in terms of case-sensitive 4-gram BLEU (Papineni et al., 2002). Approximate randomization (Noreen., 1989; Riezler and Maxwell, 2005) is used to detect statistically significant differences.

For German-English, we use high-quality data from the training corpus provided for WMT-17 (Bojar et al., 2017). For the noisy data, we randomly sample a bitext of equal size from the raw

[2]LDC2006E25, LDC2004T18, several Gale corpora, LDC2004T17, LDC2005E46 and LDC2004E13.

[3]LDC2003E14, LDC2005T10 and LDC2002E18.

[4]We are aware of the fact that much larger high-quality training data are available for Chinese-English, which result in a higher baseline. However, in order to simulate a scenario where the amount of clean data equals that of the comparable data, we downsample the size for our experiments.

	Arabic-English			Chinese-English		
	MT05	MT08	MT09	MT05	MT06	MT08
Parallel only	57.7	46.1	49.9	28.8	27.5	20.3
Comparable only	48.9 (-8.8)	**32.7** (-13.4)	**36.0**(-13.9)	11.3(-12.1)	**10.2**(-17.5)	**5.2** (-15.1)
Combined (Parallel + Comparable)	55.2 (-2.5)	44.2 (-1.9)	**47.9**(-2)	27.7(-1.1)	26.7 (-0.8)	**18.3** (-2)
Parallel +Comparable$_{bck}$	**60.4**(+2.7)	47.5 (+1.4)	**51.0**(+1.1)	**29.1**(+0.3)	27.2(-0.3)	19.8 (-0.5)
Fine-tuning	**56.1**(-1.6)	46.6 (+0.5)	**50.3**(+0.4)	**25.1** (-3.7)	**23.5** (-4)	**17.2** (-3.1)
Dual cross Entropy Filtering						
Parallel +Comparable$_{filt-25\%}$	**59.9** (+2.2)	**47.4** (+1.4)	**51.1** (+1.2)	**19.7**(-9.1)	20.9 (-6.6)	**16.8** (-3.5)
Parallel +Comparable$_{filt-50\%}$	59.2 (+1.5)	46.8 (+0.7)	**50.9**(+1)	**20.4** (-8.4)	**21.8**(-5.7)	**17.0** (-3.3)
Parallel+Comparable$_{filt-75\%}$	56.7 (-1)	**44.9** (-1.2)	49.1(-0.8)	**21.5** (-7.3)	**22.3** (-5.2)	**17.5**(-2.8)
Knowledge Distillation						
KD	**62.3** (+4.6)	**48.4** (+2.3)	**52.3**(+2.4)	**29.4** (+0.6)	**28.2** (+0.5)	**21.1** (+0.8)

Table 3: Performance of various training strategies for Arabic/Chinese-English. **Comparable**$_{bck}$ = Back-translated comparable corpora. **KD** =Knowledge distillation. Boldfaced = Significant differences at $p < 0.01$.

Paracrawl corpus ("very noisy" 1 billion English tokens) similar to Khayrallah and Koehn (2018). To be able to compare with the best filtering method, we also use a bitext of 100M target tokens submitted by Junczys-Dowmunt (2018) (available from the shared task website using a score file) which is filtered using their proposed "Dual cross entropy" score. A BPE of 32k is trained on the WMT-17 training data, newstest15 is used as dev set and newstest16 and newstest17 are used as test set. Table 2 summarizes clean and noisy training data for all language pairs.

We train an LSTM-based encoder-decoder model as described in Luong et al. (2015) using the Open-NMT-python toolkit (Klein et al., 2017), with both embeddings and hidden layers of size 1000. The maximum sentence length is restricted to 80 tokens. Parameters are optimized using Adam with an initial learning rate of 0.001, a decay rate of 0.5 (after every 10k steps), a dropout probability of 0.2 and label smoothing of 0.1. A fixed batch size of 64 is used. Model weights are initialized uniformly within [-0.02, 0.02]. We train for a maximum of 200k steps and select the model with best BLEU score on the development set for the final evaluation and decode with a beam size of 5.

6 Results

First, we compare the primary baseline with direct off-the-shelf use of noisy data without any filtering or noise reduction strategies. As can be seen in Table 3, for both Arabic-English and Chinese-English, the performance of an NMT system trained on comparable data only is substantially worse (up to -13.9 BLEU for Ar-En and -17.5 BLEU for Zh-En) as compared to clean data. Although for Arabic-English, the size of the noisy data is 4 times that of the clean data, while for Chinese-English, it is of equal size. Adding this noisy data to the clean data degrades translation performance (-2 BLEU for both Ar-En and Zh-En). The relative difference between the performance drop between the two language pairs can be attributed to the size of the comparable data.

Replacing the source side of the noisy data with back-translations sightly improves the BLEU score for Arabic-English (up to $+1.4$) but slightly degrades translation quality for Chinese-English (-0.3 BLEU compared to the baseline). Nevertheless, this is still an improvement over direct off-the-shelf addition of the original noisy bitext. This implies that back-translation replacement does provide some degree of data cleaning.

Fine-tuning for noisy data shows only slight improvements for Ar-En (up to $+0.5$ BLEU) and none for Zh-En (up to -4 BLEU drop). For both language pairs, we apply the dual cross-entropy filtering method of (Junczys-Dowmunt, 2018) by ranking sentence pairs in the comparable data according to the dual cross entropy and select subsamples from the top 50% and 75% of the full comparable bitext. Filtering at 50% shows significant ($+1$ BLEU) improvements for Arabic-English, whereas for Chinese-English this filtering results in performance even worse than adding all data, implying that cross entropy based filtering does not retain high-quality sentences from this

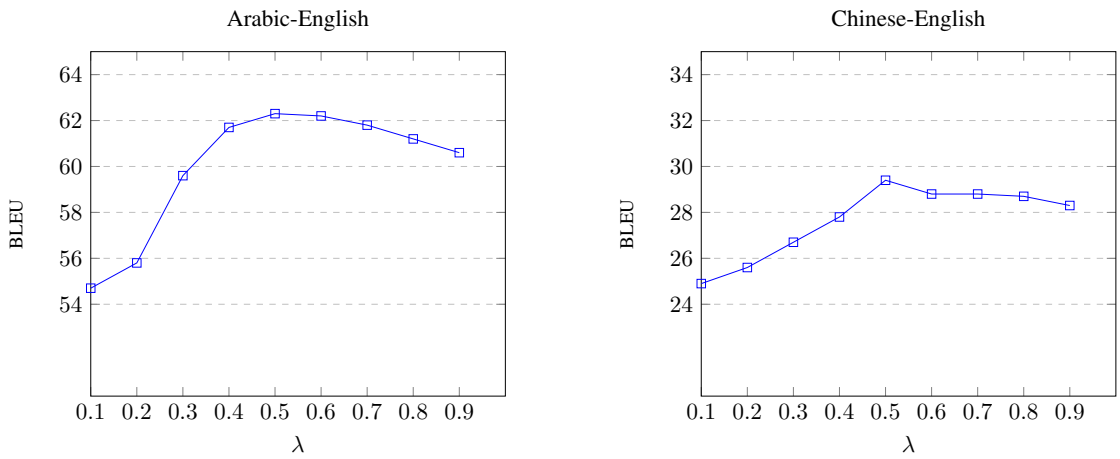

Figure 2: Variation in BLEU score for different values of λ for Arabic-English and Chinese-English

comparable bitext.

On the other hand, the proposed distillation strategy outperforms filtering as well as back-translation replacement for both language pairs. The improvements for Arabic-English are substantially higher ($+4.6$ BLEU for the dev set and $+2.4$ for the test set), while only a small improvement for Chinese-English is observed. Nevertheless, distillation provides significant improvements as compared to direct addition of the noisy data. The improvement with knowledge distillation shown in Table 3 correspond to the best improvements with respect to different values of λ. In Figure 2, we show the effect of varying values for λ (between 0.1 and 0.9) on the translation performance over the development set (MT05). For both the language pairs $\lambda = 0.5$ yields the best performance. As shown in Table 4, for German-English, there is a substantial difference (-16.4 BLEU) between the performance of a model trained on clean data only vs. one trained on randomly sampled Paracrawl data. Khayrallah and Koehn (2018) reported a degradation of up to -9 BLEU when combining clean and noisy data. However, we observe only a 1 BLEU drop for the same setting. Nevertheless, directly adding noisy data seems to provide no additional improvements. Similarly, fine-tuning on the clean data does not show any improvements. On the other hand, applying the proposed distillation over this combined bitext shows slight improvement of 0.3 BLEU over the clean data baseline.

For a comparison with "Dual cross entropy filtering", we use the filtered bitext submitted by Junczys-Dowmunt (2018) and add it to the training data, which also degrades BLEU by -1. Again, applying distillation over this filtered bitext combined with the clean data set shows an improvement of 0.9 BLEU over the clean-data baseline. As shown in Figure 3, we evaluate the performance variation for different values of λ using the 'Randomly sampled (100M target tokens) paracrawl' against the newstest'15 development set. Similar to the other two language pairs, we observe that the best BLEU score is achieved for $\lambda = 0.5$.

7 Conclusion

In this paper, we explored the effectiveness of using comparable training data for neural machine translation. Our experiments show that depending on the size of the noisy data, the performance of an NMT model can suffer significant degradations. Further, we show that noisy cleaning methods such as filtering and back-translation of noisy data show only slight improvements over the baseline. Moreover, fine-tuning fails to show any significant improvements when used for noisy data.

To overcome these problems, we proposed distillation as a remedy to efficiently leverage noisy data for NMT where we train a primary NMT model on the combined training data with knowledge distillation from the teacher network trained on the clean data only. Our experiments show that distillation can help to successfully utilize low-quality comparable data resulting in significant improvements as compared to training directly on the noisy data.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers 639.022.213 and

	test15	test16	test17
WMT (Parallel only)	25.2	30.0	26.0
Randomly sampled Paracrawl			
Para$_{rn}$	**14.6** (-10.6)	**10.2** (-19.8)	9.6 (-16.4)
WMT (Parallel) + Para$_{rn}$	**24.1**(-1.1)	**29.0** (-1)	25.0 (-1)
Fine-tuning (Para$_{rn}$)	**21.8** (-3.4)	**24.4** (-5.6)	**21.1** (-4.9)
Knowledge distillation (WMT + Para$_{rn}$)	**25.6** (+0.4)	**30.3** (+0.3)	**26.3** (+0.3)
Parcrawl filtered with dual cross entropy			
Filt$_{toks=100M}$ only	**24.0** (-1.2)	**28.8** (-1.2)	**24.6** (-1.4)
WMT + Filt$_{toks=100M}$	**24.1** (-1.1)	28.7 (-0.3)	**25.0** (-1)
Fine-tuning (Filt$_{toks=100M}$)	**23.9** (-1.3)	29.1 (-0.9)	25.1(-0.9)
Knowledge distillation (WMT + Filt$_{toks=100M}$)	**26.1**(+1.1)	**31.3** (+0.3)	**26.9**(+0.9)

Table 4: German-English results. **WMT** = Only clean Data, **Para**$_{rn}$ = Randomly sampled 5.1 million sentence pairs from Paracrawl. **Filt**$_{toks=100M}$ = 100 million target tokens filtered (Junczys-Dowmunt, 2018)

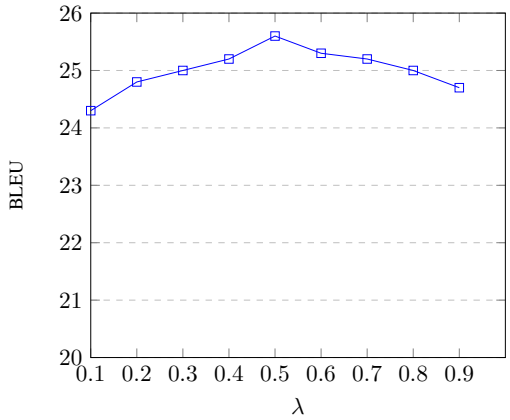

Figure 3: Variation in BLEU score for different values of λ for German-English when trained with randomly sampled paracrawl data

612.001.218.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Barbu, Eduard and Verginica Barbu Mititelu. 2018. A hybrid pipeline of rules and machine learning to filter web-crawled parallel corpora. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 880–884, Belgium, Brussels, October. Association for Computational Linguistics.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

Chen, Yun, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935. Association for Computational Linguistics.

Dakwale, Praveen and Christof Monz. 2017. Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data. In *Proceedings of the 16th Machine Translation Summit*.

Hangya, Viktor and Alexander Fraser. 2018. An unsupervised system for parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 895–900, Belgium, Brussels, October. Association for Computational Linguistics.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. In *NIPS 2014 Deep Learning Workshop*.

Junczys-Dowmunt, Marcin. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *WMT*.

Khayrallah, Huda and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Melbourne, Australia. Association for Computational Linguistics.

Kim, Yoon and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the*

2016 Conference on Empirical Methods in Natural Language Processing, pages 1317–1327, Austin, Texas, November. Association for Computational Linguistics.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Koehn, Philipp, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the wmt 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 739–752, Belgium, Brussels, October. Association for Computational Linguistics.

Li, Yuncheng, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Jia Li. 2017. Learning from noisy labels with distillation. *CoRR*, abs/1703.02391.

Luong, Thang, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal. Association for Computational Linguistics.

Miceli Barone, Antonio Valerio, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1489–1494. Association for Computational Linguistics.

Munteanu, Dragos Stefan and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, pages 477–504, December.

Noreen., Eric W. 1989. *Computer Intensive Methods for Testing Hypotheses. An Introduction*. Wiley-Interscience.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Reed, Scott E., Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping. *CoRR*, abs/1412.6596.

Riezler, Stefan and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96. Association for Computational Linguistics.