Self-Induced Curriculum Learning in Self-Supervised Neural Machine Translation

Dana Ruiter
Saarland University
DFKI GmbH

Josef van Genabith
Saarland University
DFKI GmbH

Cristina España-Bonet
DFKI GmbH

druiter@lsv.uni-saarland.de
{josef.van_genabith,cristinae}@dfki.de

Abstract

Self-supervised neural machine translation (SSNMT) jointly learns to identify and select suitable training data from comparable (rather than parallel) corpora and to translate, in a way that the two tasks support each other in a virtuous circle. In this study, we provide an in-depth analysis of the sampling choices the SSNMT model makes during training. We show how, without it having been told to do so, the model self-selects samples of increasing (*i*) complexity and (*ii*) task-relevance in combination with (*iii*) performing a denoising curriculum. We observe that the dynamics of the mutual-supervision signals of both system internal representation types are vital for the extraction and translation performance. We show that in terms of the Gunning-Fog Readability index, SSNMT starts extracting and learning from Wikipedia data suitable for high school students and quickly moves towards content suitable for first year undergraduate students.

1 Introduction

Human learners, when faced with a new task, generally focus on simple examples before applying what they learned to more complex instances. This approach to learning based on sampling from a curriculum of increasing complexity has also been shown to be beneficial for machines and is referred to as *curriculum learning* (CL) (Bengio et al., 2009). Previous research on curriculum learning has focused on selecting the best distribution of data, i.e. order, difficulty and closeness to the final task, to train a system. In such a setting, data is externally prepared for the system to ease the learning task. In our work, we follow a complementary approach: we design a system that selects by itself the data to train on, and we analyse the selected distribution of data, order, difficulty and closeness to the final task, without imposing it beforehand. Our method resembles *self-paced learning* (SPL) (Kumar et al., 2010), in that it uses the emerging model hypothesis to select samples online that fit into its space as opposed to most curriculum learning approaches that rely on judgements by the target hypothesis, i.e. an external *teacher* (Hacohen and Weinshall, 2019) to design the curriculum.

We focus on machine translation (MT), in particular, self-supervised machine translation (SSNMT) (Ruiter et al., 2019), which exploits the internal representations of an emergent neural machine translation (NMT) system to select useful data for training, where each selection decision is dependent on the current state of the model. Self-supervised learning (Raina et al., 2007; Bengio et al., 2013) involves a primary task, for which labelled data is not available, and an auxiliary task that enables the primary task to be learned by exploiting supervisory signals within the data. In SSNMT, both tasks, data extraction and learning MT, enable and enhance each other. This and the mutual supervision of the two system internal representations lead to a self-induced curriculum, which is the subject of our investigation.

In Section 2 we describe related work on CL, focusing on MT. Section 3 introduces the main aspects of self-supervised neural machine translation. Here, we analyse the performance of both the primary and the auxiliary tasks. This is followed by a detailed study of the self-induced curriculum in Section 4 where we analyse the characteristics of the distribution of training data obtained in the auxiliary task of the system. We conclude and present ideas for further work in Section 5.

2 Related Work

Machine translation has experienced major improvements in translation quality due to the introduction of neural architectures (Cho et al., 2014;

Bahdanau et al., 2015; Vaswani et al., 2017). However, these rely on the availability of large amounts of parallel data. To overcome the need for labelled data, unsupervised neural machine translation (USNMT) (Lample et al., 2018a; Artetxe et al., 2018b; Yang et al., 2018) focuses on the exploitation of very large amounts of monolingual sentences by combining denoising autoencoders with back-translation and multilingual encoders. Further combining these with phrase tables from statistical machine translation leads to impressive results (Lample et al., 2018b; Artetxe et al., 2018a; Ren et al., 2019; Artetxe et al., 2019). USNMT can be combined with pre-trained language models (LMs) (Conneau and Lample, 2019; Song et al., 2019; Liu et al., 2020). Brown et al. (2020) train a very large LM on billions of monolingual sentences which allows them to perform NMT in a few-shot setting. Self-supervised NMT (SSNMT) (Ruiter et al., 2019) is an alternative approach focusing on *comparable*, rather than *parallel* data. The internal representations of an emergent NMT system are used to identify useful sentence pairs in comparable documents. Selection depends on the current state of the model, resembling a type of self-paced learning (Kumar et al., 2010).

Data selection in SSNMT is directly related to **curriculum learning**, the idea of presenting training samples in a *meaningful* order to benefit learning, e.g. in the form of faster convergence or improved performance (Bengio et al., 2009). Inspired by human learners, Elman (1993) argues that a neural network's optimization can be accelerated by providing samples in order of increasing complexity. While **sample difficulty** is an intuitive measure on which to base a learning schedule, curricula may focus on other metrics such as **task-relevance** or **noise**.

To date, **curriculum learning in NMT** has had a strong focus on the relevance of training samples to a given translation task, e.g. in domain adaptation. van der Wees et al. (2017) train on increasingly relevant samples while gradually excluding irrelevant ones. They observed an increase in BLEU over a static NMT baseline and a significant speed-up in training as the data size is incrementally reduced. Zhang et al. (2019) adapt an NMT model to a domain by introducing increasingly domain-distant (*difficult*) samples. This seemingly contradictory behavior of benefiting from both increasingly difficult (domain-distant) *and* easy (domain-relevant) samples has been analyzed by Weinshall et al. (2018), showing that the initial phases of training benefit from easy samples with respect to a hypothetical competent model (*target hypothesis*), while also being *boosted* (Freund and Schapire, 1996) by samples that are difficult with respect to the current state of the model (Hacohen and Weinshall, 2019). In Wang et al. (2019), both domain-relevance and denoising are combined into a single curriculum.

The denoising curriculum for NMT proposed by Wang et al. (2018) is related to our approach in that they also use *online data selection* to build the curriculum based on the current state of the model. However, the noise scores for the dataset at each training step depend on fine-tuning the model on a small selection of clean data, which comes with a high computational cost. To alleviate this cost, Kumar et al. (2019) use reinforcement learning on the pre-scored noisy corpus to jointly learn the denoising curriculum with NMT. In Section 3.2 we show that our model exploits its self-supervised nature to perform denoising by selecting parallel pairs with increasing accuracy, without the need of additional noise metrics.

Difficulty-based curricula for NMT that take into account sentence length and vocabulary frequency have been shown to improve translation quality when samples are presented in increasing complexity (Kocmi and Bojar, 2017). Platanios et al. (2019) link the introduction of difficult samples with the NMT models' *competence*. Other difficulty-orderings have been explored extensively in Zhang et al. (2018), showing that they, too, can speed-up training without a loss in translation performance.

SSNMT jointly learns to find and extract similar sentence pairs from comparable data and to translate. The extractions can be compared to those obtained by **parallel data mining** systems where strictly parallel sentences are expected. Beating early feature-based approaches, sentence representations obtained from NMT systems or tailored architectures are achieving a new state-of-the-art in parallel sentence extraction and filtering (España-Bonet et al., 2017; Grégoire and Langlais, 2018; Artetxe and Schwenk, 2019; Hangya and Fraser, 2019; Chaudhary et al., 2019). Using a highly multilingual sentence encoder, Schwenk et al. (2019) scored Wikipedia sentence pairs across various language combinations (*WikiMatrix*). Due to its multi-

lingual aspect and the close similarity with the raw Wikipedia data we use, we also use scored WikiMatrix data for one of the comparisons (Section 3.2).

3 Self-Supervised Neural Machine Translation (SSNMT)

SSNMT is a joint data selection and training framework for machine translation, introduced in Ruiter et al. (2019). SSNMT enables learning NMT from *comparable* rather than parallel data, where comparable data is a collection of multilingual topic-aligned documents.[1] Its basic architecture uses the semantic information encoded in the internal representations of a standard NMT system to determine at training time if an input sentence pair is *similar enough* or not, and therefore whether it should be used for training or not. Selection is made online, so, the more the semantic representations improve during training, the more truly parallel sentence pairs are selected. Because of this, the nature of the selected pairs naturally evolves during training, and this evolution is what we analyze as self-induced curriculum learning in Section 4.

SSNMT is based on a bidirectional NMT system $\{L1, L2\}\rightarrow\{L2, L1\}$ where the engine learns to translate simultaneously from a language $L1$ into another language $L2$ and vice-versa with a single encoder and a single decoder. This is important in the self-supervised architecture because it represents the two languages in the same semantic space. In principle, the input data to train the system is a monolingual corpus of sentences in $L1$ and a monolingual corpus of sentences in $L2$ and the system learns to find and select similar sentence pairs. In order to speed-up training, we use a comparable corpus such as Wikipedia, where we can safely assume that there are comparable (similar) and parallel sentence pairs in related documents D_{L1}, D_{L2}.

Given a document pair D_{L1}, D_{L2}, the SSNMT system encodes each sentence of each document into two fixed-length vectors C_w and C_h

$$C_w = \sum_t w_t, \qquad C_h = \sum_t h_t, \qquad (1)$$

where w_t is the word embedding and h_t the encoder output at time step t. For each of the *sentence representations* s, all combinations of sentences $s_{L1} \times s_{L2} \parallel s_{L1} \in D_{L1}$ and $s_{L2} \in D_{L2}$ are encoded and scored using the *margin-based* measure by Artetxe and Schwenk (2019) with $k = 4$.

What follows is a selection process, that identifies the top scoring s_{L2} for each s_{L1} and vice-versa. If a pair $\{s_{L1}, s_{L2}\}$ is top scoring for both language directions *and* for both sentence representations, it is accepted without involving any hyperparameter or threshold. This is the high precision, medium recall approach in Ruiter et al. (2019). Whenever enough pairs have been collected to create a batch, the system trains on it, updating its weights, improving both its translation and extraction ability to fill the next batch.

3.1 Translation Quality

Experimental Setup We use Wikipedia (WP) as a comparable corpus and download the English, French, German and Spanish dumps,[2] pre-process them and extract comparable articles per language pair using `WikiTailor`[3] (Barrón-Cedeño et al., 2015; España-Bonet et al., 2020). All articles are normalized, tokenized and truecased using standard `Moses` (Koehn et al., 2007) scripts. For each language pair, a shared byte-pair encoding (BPE) (Sennrich et al., 2016) of 100 k merge operations is applied. Following Johnson et al. (2017), a language tag is added to the beginning of each sequence.

The number of sentences, tokens and average article length is reported in Table 1. For validation we use *newstest2012* (NT12) and for testing *newstest2013* (NT13) for *en–es* and *newstest2014* (NT14) or *newstest2016* (NT16) for *en–*$\{fr, de\}$. The SSNMT implementation[4] builds on the transformer base (Vaswani et al., 2017) in `OpenNMT` (Klein et al., 2017). All systems are trained using a batch size of 50 sentences with maximum length of 50 tokens.

Monolingual embeddings trained using `word2vec` (Mikolov et al., 2013)[5] on the complete WP editions are projected into a common multilingual space via `vecmap`[6] (Artetxe et al., 2017) to attain bilingual embeddings between *en–*$\{fr$,*de*,*es*$\}$. These initialise the NMT word embeddings (C_w).

[1] Wikipedia is an example; the French article on *Paris* is different from the German one. They are not translations of each other, but they are on the same topic.

[2] Dumps were downloaded on January 2019 from `dumps.wikimedia.org/`

[3] `github.com/cristinae/WikiTailor`

[4] `github.com/ruitedk6/comparableNMT`

[5] `github.com/tmikolov/word2vec`

[6] `github.com/artetxem/vecmap`

	WP, L1			WP, L2			EP, L1		EP, L2	
L1–L2	# Sent.	# Tokens	Sent./Article	# Sent.	# Tokens	Sent./Article	# Sent.	# Tokens	# Sent.	# Tokens
en–fr	117 / 42	2693/1205	28	38/25	644/710	16	1+6	25+80	1+3	27+87
en–de	117 / 37	2693/987	29	51/30	1081/742	24	1+9	25+180	1+7	26+192
en–es	117 / 35	2693/937	32	27/20	691/572	17	1+7	24+84	1+4	26+91

Table 1: Millions of sentences and tokens for the corpora used. For Wikipedia (WP), we report the sizes for both the monolingual/comparable editions; for Europarl (EP), true+false splits (see Section 3.2).

	SSNMT						SotA	
	L1-to-L2			L2-to-L1			L1-to-L2	L2-to-L1
L1–L2	BLEU	TER	METEOR	BLEU	TER	METEOR	BLEU	BLEU
en–fr	29.5±.6	51.9±.6	46.4±.6	27.7±.6	53.4±.7	30.3±.4	45.6/25.1/37.5	–/24.2/34.9
en–de	15.2±.5	68.5±.7	30.3±.5	21.2±.6	62.8±.9	25.4±.4	37.9/17.2/28.3	–/21.0/35.2
en–es	28.6±.7	52.6±.7	47.8±.7	28.4±.7	54.1±.7	30.5±.4	–/–/–	–/–/–

Table 2: Automatic evaluation of SSNMT on NT14 (fr) NT16 (de) NT13 (es). Most right columns show the comparison with three SotA systems for supervised NMT (Edunov et al., 2018) / USNMT (Lample et al., 2018b) / pre-trained+LM USNMT (Song et al., 2019).

As a control experiment and purely in order to analyse the quality of the SSNMT data selection auxiliary task, we use the Europarl (EP) corpus (Koehn, 2005). The corpus is pre-processed in the same way as WP, and we create a synthetic comparable corpus from it as explained in Section 3.2. For these experiments, we use the same data for validation and testing as mentioned above.

Automatic Evaluation We use BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Lavie and Agarwal, 2007) to evaluate translation quality. For calculating BLEU, we use `multi-bleu.perl`, while TER and METEOR are calculated using the `scoring` package[7] which also provides confidence scores. SSNMT translation performance training on the en–$\{fr, de, es\}$ comparable Wikipedia data is reported in Table 2 together with a comparison to the current state-of-the-art (SotA) in supervised and (pre-trained) USNMT. SSNMT is on par with the current SotA in USNMT, outperforming it by 3–4 BLEU points in en–fr with lower performance on en–de (∼3 BLEU). Note that unsupervised systems such as Lample et al. (2018b) use more than $400\,M$ monolingual sentences for training while SSNMT uses an order of magnitude less by exploiting comparable corpora. However, once unsupervised NMT is combined with LM pre-training, it outperforms SSNMT (which does not use LM pre-training) by large margins, i.e. around 7 BLEU points for en–fr and 13 BLEU for en–de.

[7] kheafield.com/code/scoring.tar.gz

3.2 Data Extraction Quality

Experimental Setup To get an idea of the data extraction performance of an SSNMT system, we perform control experiments on synthetic comparable corpora, as there is no underlying ground truth to Wikipedia. For these purposes, we use the en–$\{fr,de,es\}$ versions of Europarl. After setting aside $1M$ parallel pairs as *true* samples to evaluate SSNMT data extraction performance, the target sides of all remaining source-target pairs in EP are scrambled to create non-parallel (*false*) source-target pairs. In order to keep the synthetic comparable corpora close to the statistics of the original comparable Wikipedias, we control the EP true:false (parallel:non-parallel) sentence pair ratio to mimic the ratios we observe in our extractions from WP. We assume that all WP sentences accepted by SSNMT are true (parallel) examples, and that the number of false examples (non-parallel) are the rejected ones. With this, we estimate base true:false ratios of 1:4 for en–$\{fr,es\}$ and 1:8 for en–de.[8] The false samples created from EP are oversampled in order to meet this ratio given that there are $1M$ true samples. Further, we calculate the average article length of the comparable WPs and split the synthetic comparable samples into pseudo-articles with this length. The statistics of

[8] In a manual evaluation annotating 10 randomly sampled WP articles for L1 and L2 in en–$\{fr,es,de\}$ each, the true:false ratios resulted 3:8 for en–fr, 1:4 for en–es and 1:8 for en–de which validate the assumption.

the synthetic pseudo-comparable EPs are reported in Table 1. We then train and evaluate the SSNMT system on the synthetic comparable data.

Automatic Evaluation The pairs SSNMT extracts from the pseudo-comparable EP articles at each epoch are compared to the $1M$ ground truth pairs to calculate *epoch-wise* extraction precision (P) and recall (R). Further, we also take the concatenation of all extracted sentences from the very beginning up to a certain epoch in training in order to report *accumulated* P and R. As we are interested in the final extraction decision based on the intersection of both representations C_w and C_h (*dual*), but also in the decisions of each single representation (C_w, C_h), we report the performance for all three representation combinations on EP_{enfr} in Figure 1. Similar curves are observed for EP_{ende} and EP_{enes}, which are considered in the discussion below.

At the beginning of training, the extraction **precision** of each representation itself is fairly low with P∈[0.45,0.66] for C_w and P∈[0.14,0.40] for C_h. The fact that C_w is initialized using pre-trained embeddings, while C_h is not, leads to the large difference in initial precision between the two. As both representations are combined via their intersections, the final decision of the model is high precision already at the beginning of training with values between 0.78–0.87. As training progresses and the internal representations are adapted to the task, the precision of C_h is greatly improved, leading to an overall high precision extraction which converges at 0.96–0.99. This development of extracting parallel pairs with increasing precision is in fact an instantiation of a **denoising curriculum** as described by Wang et al. (2018).

The **recall** of the model, being bounded by the performance of the weakest representation, is very low at the beginning of training (R∈ [0.03,0.04]) due to the lack of task knowledge in C_h. However, as training progresses and C_h improves, the accumulated extraction recall of the model rises to high values of 0.95–0.98. Interestingly, the epoch-wise recall is much lower than the accumulated, which provides evidence for the hypothesis that SSNMT models extracts different *relevant* samples at different points in training, such that it has identified most of the relevant samples at some point during training, but not at every epoch.

It should be stressed that the successful extraction of increasingly precise pairs in combination

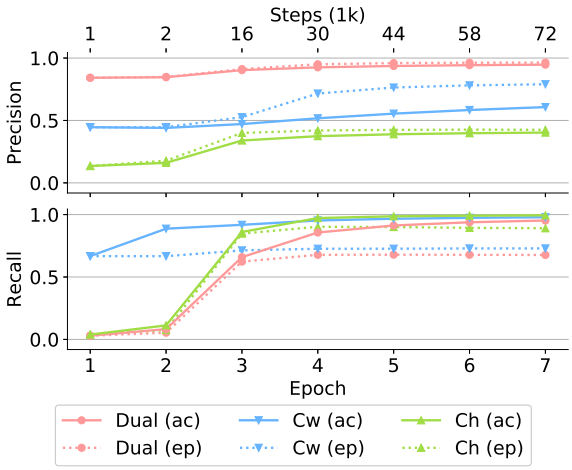

Figure 1: Accumulated (ac) and epoch-wise (ep) precision and recall on the en–fr EP-based synthetic comparable data.

with high recall is the result of the dynamics of both internal representations C_w and C_h. As C_h is less informative at the beginning of training, C_w guides the final decision at such early stages to ensure high precision; and as C_w is high in recall throughout training, C_h ensures a gentle growth in final recall by setting a good lower bound. The intersection of both ensures that errors committed by one can be caught by the other; effectively a mutual supervision between representations. The results in Figure 1 show that the SSNMT self-induced curriculum is able to identify parallel data in comparable data with high precision and recall.

Comparison with WikiMatrix Because of the close similarity with our WP data, we compare on the en–$\{fr, de, es\}$ corpora in WikiMatrix (Schwenk et al., 2019), which we pre-process as described in Section 3.1. As these data sets consist of preselected mined sentence pairs together with their similarity scores, a manual threshold θ needs to be set to extract sentence pairs for training supervised NMT. We run the extraction script using $\theta = 1.04$, which Schwenk et al. (2019) recommend as a *good choice for most language pairs*, and use the resulting data to train a supervised NMT system.

The results are summarized in the bottom two rows in Table 3. Confidence intervals ($p = 95\%$) are calculated using bootstrap resampling (Koehn, 2004). For en–fr, the supervised system trained on WikiMatrix outperforms SSNMT trained on WP by 3–4 BLEU points, while the opposite is the case for en–de, where SSNMT achieves 1–5

	#Pairs$_{enfr}$	$en2fr$	$fr2en$	#Pairs$_{ende}$	$en2de$	$de2en$	#Pairs$_{enes}$	$en2es$	$es2en$
NMT$_{init}$	2.14M	21.8±.6	21.1±.5	0.32M	3.4±.3	4.7±.3	2.51M	27.0±.7	25.0±.7
NMT$_{mid}$	3.14M	29.0±.6	26.6±.6	1.13M	11.2±.4	15.0±.6	3.96M	28.3±.7	26.1±.7
NMT$_{end}$	3.17M	28.8±.6	26.5±.6	1.18M	11.9±.5	15.3±.5	3.99M	28.3±.7	26.2±.7
NMT$_{all}$	5.38M	26.8±.7	25.2±.6	2.21M	11.6±.5	15.0±.6	5.41M	27.9±.6	25.9±.8
SSNMT	5.38M	29.5±.6	27.7±.6	2.21M	14.4±.6	18.1±.6	5.41M	28.6±.7	28.4±.7
WikiMatrix	2.76M	33.5±.6	30.1±.6	1.57M	13.2±.5	12.2±.5	3.38M	29.6±.7	26.9±.8

Table 3: BLEU scores of a supervised NMT system trained on the unique pairs collected by SSNMT in the first (NMT$_{init}$), intermediate (NMT$_{mid}$), final (NMT$_{end}$) and all (NMT$_{all}$) epochs of training tested on N13/N14.

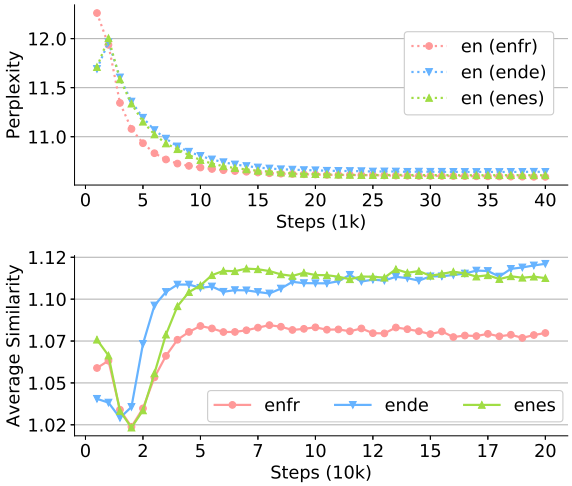

Figure 2: Perplexities on the English data extracted by SSNMT (top) and average similarity scores of the accepted pairs (bottom).

BLEU points more. For en–es, both approaches are not statistically significantly different. The variable performance of the two approaches may be due to the varying appropriateness of the extraction threshold θ in WikiMatrix. For each language and corpus, a new optimal threshold needs to be found; a problem that SSNMT avoids by its use of two representation types that complement each other during extraction without the need of a manually set threshold. The results show that SSNMT's self-induced extraction and training curriculum is able to deliver translation quality on a par with supervised NMT trained on externally preselected mined parallel data (WikiMatrix).

4 Self-Induced SSNMT Curricula

4.1 Order & Closeness to the MT Task

As a first indicator of the existence of a preferred choice in the order of the extracted sentence pairs, we compare the performance of SSNMT with different supervised NMT models trained on the WP data extracted by SSNMT at different points in training. We consider specific per-epoch data sets extracted in the first, intermediate and final epochs of training, as well as cumulative data of all unique sentence pairs extracted over all epochs. We then train four supervised NMT systems (NMT$_{init}$, NMT$_{mid}$, NMT$_{end}$, NMT$_{all}$) on these data sets. The difference in the **translation quality** using only the data selected at different epochs reflects the evolving closeness of the data to the final translation task: we expect data extracted in later epochs of the SSNMT training to include more sentences which are parallel, as demanded by a translation task, and therefore to achieve a higher translation quality.

For each language pair and system, the first four rows in Table 3 show the number of sentence pairs extracted for training and the BLEU score achieved. The evolving SSNMT training curriculum outperforms all supervised versions across all tested languages. Notably, performance is 1–3 BLEU points above the supervised system trained on all extracted data, despite the fact that the SSNMT system is able to extract only a small amount of data in its first epochs, compared to the fully supervised NMT$_{all}$, that, at every epoch, has access to all data that was ever extracted at any of the SSNMT epochs. This suggests that the SSNMT system is able to exclude previously accepted false positives in later epochs, while training supervised NMT on the complete data extracted by SSNMT leads to a recurring visitation at each epoch of the same erroneous samples. Similar to a **denoising curriculum**, the quality and quantity of the extracted data grows as training continues for all languages, as the concatenation of the data extracted across epochs (NMT$_{all}$) is always outperformed by the last and thus largest epoch (NMT$_{end}$), despite the data for NMT$_{all}$ being much larger in size.

An indicator of the **closeness of the curriculum to the final task** is the **similarity** between the selected sentence pairs during training. We

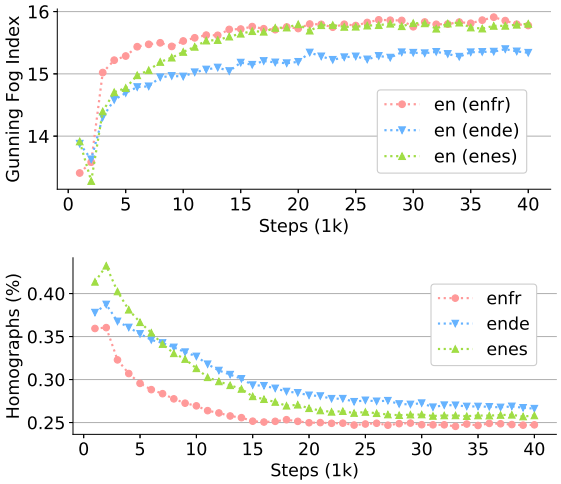

Figure 3: Gunning Fog Index (top) and percentage of homographs (bottom) of extracted English data seen during the first 40 k steps in training.

estimate similarity between pairs by their margin-based scores (Artetxe and Schwenk, 2019) during training. At the beginning of training, the average similarity between extracted pairs is low, but it quickly rises within the first 100 k training steps to values close to $margin$ 1.07 (en–fr) and $margin$ 1.12 (en–$\{de,es\}$). This evolution is depicted in Figure 2 (bottom). The increase in mean similarity of the accepted pairs provides empirical evidence for our hypothesis that internal representations of translations grow closer in the cross-lingual space, and the system is able to exploit this by extracting increasingly similar and accurate pairs.

4.2 Order & Complexity

Establishing the complexity of a sentence is a complex task by itself. Complexity can be estimated by the loss of an instance with respect to the gold or target. In our self-supervised approach, there is no target for the sentence extraction task, so we try to infer complexity by other means.

First, we study the behaviour of the average **perplexity** throughout training. Perplexities of the extracted data are estimated using a LM trained with `KenLM` (Heafield, 2011) on the monolingual WPs for the four languages in our study. We observe the same behaviour in the four cases illustrated by the English curves plotted in Figure 2 (top). Perplexity drops heavily within the first 10 k steps for all languages and models. This indicates that the data extracted in the first epoch includes more *outliers*, and the distribution of extracted sentences moves closer to the average observed in the monolingual WPs as training advances. The larger number of outliers at the beginning of training can be attributed to the larger number of homographs (bottom Figure 3) and short sentences at the beginning of training, leading to a skewed distribution of selected sentences.

The presence of **homographs** is vital for the self-supervised system in its initialization phase. At the beginning of training, only word embeddings, and therefore C_w, are initialized with pre-trained data, while C_h is randomly initialized. Thus, words that have the same index in the shared vocabulary, homographs, play an important role in identifying similar sentences using C_h, making up around 1/3 of all tokens observed in the first epoch. As training progresses, and both C_w and C_h are adapted to the training data, the prevalence of homographs drops and the extraction is now less dependent on a shared vocabulary. The importance of homographs for the initialization raises questions on how SSNMT performs on languages that do not share a script and it is left for future work.

Finally, we analyze the complexity of the sentences that an SSNMT system selects at different points of training by measuring their **readability**. For this, we apply a modified version of the **Gunning Fog Index** (GF) (Gunning, 1952), which is a measure predicting the years of schooling needed to understand a written text given the complexity of its sentences and vocabulary. It is defined as:

$$\text{GF} = 0.4 \left[\left(\frac{w}{s} \right) + 100 \left(\frac{c}{w} \right) \right] \qquad (2)$$

where w and s are the number of words and sentences in a text. c is the number of *complex words*, which are defined as words containing more than 2 syllables. The original formula excluded several linguistic phenomena from the complex word definition such as compound words, inflectional suffixes or familiar jargon; we do not apply all the language-dependent linguistic analysis.

Since our training data is based on Wikipedia articles, the diversity in the complexity of the sentences is limited to the range of complexities observed in Wikipedia. Figure 4 (right) shows the per-sentence GF distributions over the sentences found in the monolingual WPs. We plot the probability density function for the sentence-level GF Index for the four WP editions estimated via a kernel density estimation. Each distribution is made up of two overlapping distributions: one at the lower end of the sentence complexity scale containing short

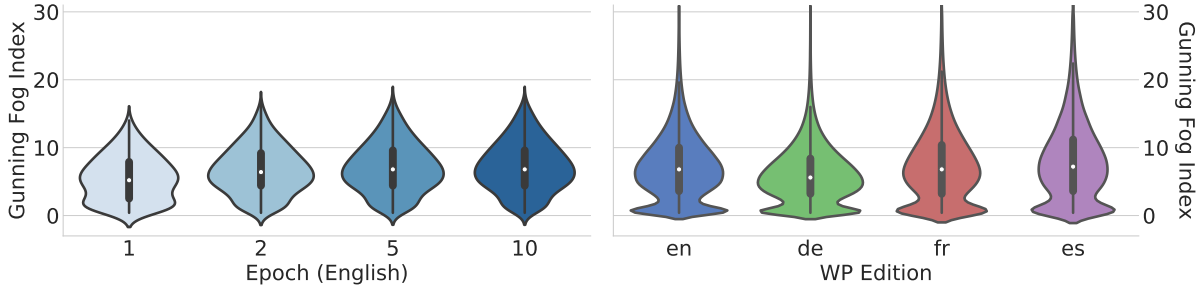

Figure 4: Kernel density estimated Gunning Fog distributions and box plots over extracted en (en–de) sentences at different points in training (left) and over the monolingual Wikipedias (right).

article titles and headers, and one with a higher average complexity and larger standard deviation containing content sentences.

To study the behaviour during training, we compare the Gunning Fog distributions of the English data extracted at the beginning, middle and end of training SSNMT$_{ende}$ with that of the original WP$_{en}$. In the extracted data, we observe that compared with WP the overlapping distributions are less pronounced and that there is no trail of highly complex sentences. This is due to (*i*) the pre-processing of the input data, which removes sentences containing less than 6 tokens, thus removing most WP titles and short sentences, and (*ii*) the length accepted in our batches, which is constrained to 50 tokens per sentence, removing highly complex strings. Apart from this, the distributions in the middle and the end of training come close to the underlying one, but we observe a large number of very simple sentences in the first epoch. This shows that the system extracts mostly simple content at the beginning of training, but soon moves towards complex sentences that were previously not yet identifiable as parallel.

A more detailed evolution is depicted in Figure 3 (top). We collect extracted sentences for each 1 k training steps and report their "text"-level GF scores.[9] Here we observe how the complexity of the sentences extracted rises strongly within the first 20 k steps of training. For English, most models start with text that is suitable for high school students (grade 10–11) and quickly turn to more complex sentences suited for first year undergraduate students (∼13 years of schooling); a **curriculum of growing complexity**. The GF mean of the full set of sentences in the English Wikipedia is ∼12, which corresponds to a high school senior. For all other languages, a similar trend of growing sentence complexity is observed.

[9]Note that GF is a text level score. In Figure 4 we show sentence level GF distributions, while in Figure 3 (top) we show GF scores for "texts" consisting of sentences extracted over a 1 k training step period.

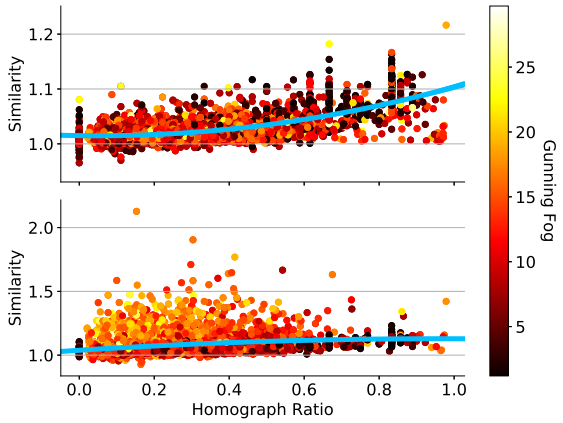

Figure 5: Margin-based similarity, homograph ratio and Gunning Fog index for the first 10 k extracted sentences in the first (top) and last (bottom) epoch of en–fr training. The solid blue line shows a second order polynomial regression between the homograph ratio and similarity.

4.3 Correlation Analysis

So far, the variables under study, similarity and complexity —GF and homograph ratio—, have been observed as a function of the training steps. In order to uncover the correlations between the variables themselves, we calculate the Pearson Correlation Coefficient (r) between them on the extracted pairs of the en–fr SSNMT model during its first and last epoch. As shown in the previous sections of the paper, most differences appear in the first epoch and the behaviour across languages is comparable.

At the beginning of training (Figure 5, top) there is a positive correlation ($r = 0.43$) between homograph ratio and similarity, naturally pointing to the importance of homographs for identifying

similar pairs at the beginning of training. This is supported by a weak negative correlation between GF and homograph ratio ($r = -0.28$), indicating that sentences with more homographs tend to be less complex. While there is no significant correlation between GF and similarity in the first epoch ($r = -0.07$), in the last epoch of training (Figure 5, bottom), we observe a moderate positive relationship indicating that more complex sentences tend to come with a higher similarity ($r = 0.30$). At this point, homographs become less important for the extraction and sentences without homographs are now also extracted in large numbers, indicated in terms of a weaker positive correlation between the homograph ratio and the similarity ($r = 0.25$). The relationship between the homograph ratio and the GF stays stable ($r = -0.27$), as can be expected since the two values are not dependent on the MT model's state (C_w and C_h), as opposed to the similarity score.

5 Summary and Conclusions

This paper explores self-supervised NMT systems which jointly learn the MT model and how to find its supervision signal in comparable data; i.e. how to identify and select similar sentences. This association makes the system naturally and internally evolve its own curriculum without it having been externally enforced. We observe that the dynamics of mutual-supervision of both system internal representations, C_w and C_h, is imperative to the high recall and precision parallel data extraction of SSNMT. Their combination for data selection over time instantiates a **denoising curriculum** in that the percentage of non-matching pairs, i.e. non-translations, decreases from 18% to 2%, with an especially fast descent at the beginning of training.

Even if the quality of extraction increases over time, lower-similarity sentence pairs used at the beginning of training are still relevant for the development of the translation engine. We analyze the translation quality of a supervised NMT system trained on the epoch-wise data extracted by SSNMT and observe a continuous increase in BLEU. Analogously, we also analyze the similarity scores of extracted sentences and observe that they also increase over time. As extracted pairs are increasingly similar, and precise, the extraction itself instantiates a secondary **curriculum of growing task-relevance**, where the task at hand is NMT learning with parallel sentences.

A tertiary **curriculum of increased sample complexity** is observed via an analysis of the extracted data's Gunning Fog indices. Here, the system starts with sentences suitable for initial high school students and quickly moves towards content suitable for first year undergraduate students: an overachiever indeed as the norm over the complete WP is end of high school level.

Lastly, by estimating the perplexity with an external LM trained on WP, we observe a steep decrease in perplexity at the beginning of training with fast convergence. This indicates that the extracted data quickly starts to resemble the underlying distribution of all WP data, with a larger amount of outliers at the beginning. These outliers can be accounted for by the importance of homographs at that point. This raises the question of how SSNMT will perform on really distant languages (less homographs) or when using smaller BPE sizes (more homographs), which is something that we will examine in our future work.

Acknowledgments

The project on which this paper is based was funded by the German Federal Ministry of Education and Research under the funding code 01IW17001 (Deeplee). Responsibility for the content of this publication is with the authors.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations, ICLR*.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3197–3203.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA.

Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba, and Lluís Màrquez. 2015. A Factory of Comparable Corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, New York, NY, USA. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey L. Elman. 1993. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71 – 99.

Cristina España-Bonet, Alberto Barrón-Cedeño, and Lluís Màrquez. 2020. Tailoring and Evaluating the Wikipedia for in-Domain Comparable Corpora Extraction. *arXiv preprint arXiv:2005.01177*.

Cristina España-Bonet, Adám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An empirical analysis of NMT-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350.

Yoav Freund and Robert E. Schapire. 1996. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, pages 148–156, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Francis Grégoire and Philippe Langlais. 2018. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1442–1453, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

Guy Hacohen and Daphna Weinshall. 2019. On the power of curriculum learning in training deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2535–2544, Long Beach, California, USA. PMLR.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy. Association for Computational Linguistics.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 379–386, Varna, Bulgaria. INCOMA Ltd.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

Gaurav Kumar, George Foster, Colin Cherry, and Maxim Krikun. 2019. Reinforcement learning based curriculum optimization for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2054–2061, Minneapolis, Minnesota. Association for Computational Linguistics.

M. P. Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1189–1197. Curran Associates, Inc.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations, ICLR*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. 2007. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning*, pages 759–766, New York, NY, USA. ACM.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. In *The 33rd AAAI Conference on Artificial Intelligence (AAAI 2019)*, pages 241–248.

Dana Ruiter, Cristina España-Bonet, and Josef van Genabith. 2019. Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,

pages 1828–1834, Florence, Italy. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, (ACL 2016), Volume 1: Long Papers*, pages 1715–1725, Berlin, Germany.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *In Proceedings of the Association for Machine Translation in the Americas (AMTA) 2006*, pages 223–231.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019. Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1282–1292, Florence, Italy. Association for Computational Linguistics.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation*, pages 133–143. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark. Association for Computational Linguistics.

Daphna Weinshall, Gad Cohen, and Dan Amir. 2018. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5238–5246, Stockholm, Sweden. PMLR.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55. Association for Computational Linguistics.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint arXiv:1811.00739*.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1903–1915, Minneapolis, Minnesota. Association for Computational Linguistics.