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

Dana Ruiter
Saarland University
DFKI GmbH
druiter@lsv.uni-saarland.de

Josef van Genabith
Saarland University
DFKI GmbH
{josef.van_genabith,cristinae}@dfki.de

Cristina España-Bonet
DFKI GmbH

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. 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 \( \{L_1, L_2\} \rightarrow \{L_2, L_1\} \) where the engine learns to translate simultaneously from a language \( L_1 \) into another language \( L_2 \) 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 \( L_1 \) and a monolingual corpus of sentences in \( L_2 \) 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 \( D_{L_1}, D_{L_2} \).

Given a document pair \( D_{L_1}, D_{L_2} \), the SSNMT system encodes each sentence of each document into two fixed-length vectors \( C_w \) and \( C_h \)

\[
C_w = \sum_t w_t, \quad C_h = \sum_t h_t, \quad (1)
\]

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 \( s_{L_1} \times s_{L_2} \parallel s_{L_1} \in D_{L_1} \) and \( s_{L_2} \in D_{L_2} \) 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 \( s_{L_2} \) for each \( s_{L_1} \) and vice-versa. If a pair \( \{s_{L_1}, s_{L_2}\} \) 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, 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 news-test2012 (NT12) and for testing news-test2013 (NT13) for \( en-es \) and news-test2014 (NT14) or news-test2016 (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 \( C_t \).

---

\(^1\)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/cristine/WikiTailor

\(^4\)github.com/ruitedk6/comparableNMT

\(^5\)github.com/tmikolov/word2vec

\(^6\)github.com/artetxem/vecmap
The corpus is pre-processed in TEOR (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.

### 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\(\times\)\(10^6\) 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 \(en-de\).

### 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 parallel pairs as \(1+:1\)\(\times\)\(4\) for \(en-fr\), \(1+:1\)\(\times\)\(8\) for \(en-de\), and \(1+:1\)\(\times\)\(8\) for \(en-es\). The false samples created from 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.

#### 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).

<table>
<thead>
<tr>
<th>L1-L2</th>
<th>BLEU</th>
<th>TER</th>
<th>METEOR</th>
<th>BLEU</th>
<th>TER</th>
<th>METEOR</th>
<th>BLEU</th>
<th>TER</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(en-fr)</td>
<td>29.5±.6</td>
<td>51.9±.6</td>
<td>46.4±.6</td>
<td>27.7±.6</td>
<td>53.4±.7</td>
<td>30.3±.4</td>
<td>45.6/25.1/37.5</td>
<td>-/24.2/34.9</td>
<td></td>
</tr>
<tr>
<td>(en-de)</td>
<td>15.2±.5</td>
<td>68.5±.7</td>
<td>30.3±.5</td>
<td>21.2±.6</td>
<td>62.8±.9</td>
<td>25.4±.4</td>
<td>37.9/17.2/28.3</td>
<td>-/21.0/35.2</td>
<td></td>
</tr>
<tr>
<td>(en-es)</td>
<td>28.6±.7</td>
<td>52.6±.7</td>
<td>47.8±.7</td>
<td>28.4±.7</td>
<td>54.1±.7</td>
<td>30.5±.4</td>
<td>-/-/-/-</td>
<td>-/-/-/-</td>
<td></td>
</tr>
</tbody>
</table>

\(^7\)kheafield.com/code/scoring.tar.gz

\(^8\)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\textsubscript{en-fr} in Figure 1. Similar curves are observed for EP\textsubscript{ende} and EP\textsubscript{enes}, which are considered in the discussion below.

At the beginning of training, the extraction precision of each representation itself is fairly low with $P \in [0.45,0.66]$ for $C_w$ and $P \in [0.14,0.40]$ 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 \in [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 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 \textit{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 $\theta$ 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 \textit{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 \textit{en–de}, where SSNMT achieves 1–5
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\textsubscript{init}, NMT\textsubscript{mid}, NMT\textsubscript{end}, NMT\textsubscript{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\textsubscript{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\textsubscript{all}) is always outperformed by the last and thus largest epoch (NMT\textsubscript{end}), despite the data for NMT\textsubscript{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

<table>
<thead>
<tr>
<th>#Pairs\textsubscript{enfr}</th>
<th>en2fr</th>
<th>fr2en</th>
<th>#Pairs\textsubscript{ende}</th>
<th>en2de</th>
<th>de2en</th>
<th>#Pairs\textsubscript{enes}</th>
<th>en2es</th>
<th>es2en</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT\textsubscript{init}</td>
<td>2.14M</td>
<td>21.8±.6</td>
<td>21.1±.5</td>
<td>0.32M</td>
<td>3.4±.3</td>
<td>4.7±.3</td>
<td>2.51M</td>
<td>27.0±.7</td>
</tr>
<tr>
<td>NMT\textsubscript{mid}</td>
<td>3.14M</td>
<td>29.2±.6</td>
<td>26.6±.6</td>
<td>1.13M</td>
<td>11.2±.4</td>
<td>15.0±.6</td>
<td>3.96M</td>
<td>28.3±.7</td>
</tr>
<tr>
<td>NMT\textsubscript{end}</td>
<td>3.17M</td>
<td>28.8±.6</td>
<td>26.5±.6</td>
<td>1.18M</td>
<td>11.9±.5</td>
<td>15.3±.5</td>
<td>3.99M</td>
<td>28.3±.7</td>
</tr>
<tr>
<td>NMT\textsubscript{all}</td>
<td>5.38M</td>
<td>26.8±.7</td>
<td>25.2±.6</td>
<td>2.21M</td>
<td>11.6±.5</td>
<td>15.0±.6</td>
<td>5.41M</td>
<td>27.9±.6</td>
</tr>
<tr>
<td>SSNMT</td>
<td>5.38M</td>
<td>29.5±.6</td>
<td>27.7±.6</td>
<td>2.21M</td>
<td>14.4±.6</td>
<td>18.1±.6</td>
<td>5.41M</td>
<td>28.6±.7</td>
</tr>
<tr>
<td>WikiMatrix</td>
<td>2.76M</td>
<td>33.5±.6</td>
<td>30.1±.6</td>
<td>1.57M</td>
<td>13.2±.5</td>
<td>12.2±.5</td>
<td>3.38M</td>
<td>29.6±.7</td>
</tr>
</tbody>
</table>

Table 3: BLEU scores of a supervised NMT system trained on the unique pairs collected by SSNMT in the first (NMT\textsubscript{init}), intermediate (NMT\textsubscript{mid}), final (NMT\textsubscript{end}) and all (NMT\textsubscript{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).
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 10k 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 SS-NMT 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:

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

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 3: Gunning Fog Index (top) and percentage of homographs (bottom) of extracted English data seen during the first 40k steps in training.](image-url)
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$_{en-de}$ 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 1k training steps and report their “text”-level GF scores. Here we observe how the complexity of the sentences extracted rises strongly within the first 20k 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.

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
Analogously, we also analyze the similarity scores which jointly learn the MT model and how to find (This paper explores self-supervised NMT systems). The relationship between the homograph ratio and development of the translation engine. We analyze 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


