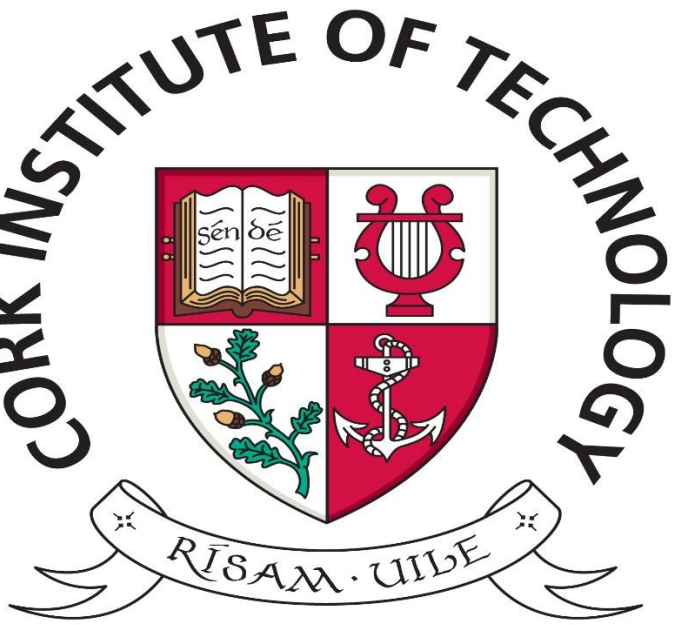


Actively Fake it Until you Make it in Neural Machine Translation

by

Frank Kelly

in the

Faculty of Engineering and Science

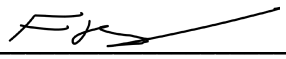
Department of Computer Science

May 2021

Declaration of Authorship

I, Frank Kelly, declare that this research proposal titled, `Actively Fake it Until You Make it in Neural Machine Translation" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a master's degree at Cork Institute of Technology.
- Where any part of this research Proposal has previously been submitted for a degree or any other qualification at Cork Institute of Technology or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. Except for such quotations, this project report is entirely my own work.
- I have acknowledged all main sources of help.
- Where the research proposal is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- I understand that my project documentation may be stored in the library at CIT, and may be referenced by others in the future.

Signed: 

Date: 9th May 2020

CORK INSTITUTE OF TECHNOLOGY

Abstract

Faculty of Engineering and Science

Department of Computer Science

Master of Science

by Frank Kelly

Model extraction attacks can create functionally equivalent models if given query access to deep neural networks (DNNs).

State-of-the-art results in machine translation are currently achieved by training DNNs on vast quantities of parallel corpora. Gathering and annotating large quantities of high-quality data requires substantial time and financing. Active learning model extraction attacks bypass this hurdle by focusing on extracting this valuable information from deployed DNN models.

Active learning model extraction approaches generate synthetic training datasets by having a victim model label an adversary's input data. Such model extraction attacks enable an adversary to obtain a functionally equivalent model. Active learning model extraction is problematic for the monetization of DNN APIs as it undermines the valuable intellectual property of deployed models.

This study aims to evaluate how effective the active learning forward translation process can be for generating synthetic parallel corpora for training Neural Machine Translation (NMT) Models. For experiments the active learning model extraction process is used to generate a target language pair from source data which is a part of an existing parallel corpus. A NMT model is trained on the synthetic parallel corpora and another identical NMT model is trained on the original parallel corpora.

Experiments showed using synthetic parallel corpora generated from active learning via random sample selection can be used to train functioning NMT models.

Active learning model extraction has been applied to various NLP tasks. However, to the best of my knowledge, this will be the first evaluation of active learning model extraction attack applied to a Neural Machine Translation model.

Acknowledgements

I would like to thank my supervisor Dr Mohammed Hasanuzzaman for his support throughout this master project.

I would like to add a huge thanks to my course coordinator Dr Ted Scully. Dr Ted Scully was amazingly supportive over the last two years, whereby no query I had was ever left unanswered.

Contents

Research Context and Contribution to the Research Field

What is Active Learning

Active learning is a simple unsupervised data augmentation technique used in machine learning (ML). (Settles, 2009). In active learning, the learner submits a query to an oracle (e.g. another ML model or human). This oracle processes the input data and returns an annotated output to the learner. (e.g. learner inputs image to oracle and oracle return classification label for image). This newly annotated data is then used by the learner for training purposes to improve its own performance.

Active learning is particularly useful in domains where a huge quantity of data is available but is costly to annotate. Annotation costs can be significant due to the sheer quantity of data required or high level of domain knowledge required to interpret data.

State of the art neural machine translation models have improved significantly in recent years (Vaswani et al., 2017). Overall neural machine translation is not at the same fluency of professional human annotators. However some state of the neural machine translation networks have outperformed human annotators in certain translation task (Ng et al., 2019; Popel et al., 2020). Due to these recent advancement active learning is has become an effective data augmentation technique when applied to machine translation data (Liu, Buntine, & Haffari, 2018)

Active learning is typically used non maliciously as a data augmentation technique to increase available training data to further improve model performance. Orekondy and colleagues (2019) used active learning from the malicious perspective of stealing an models' functionality by extracting valuable information from the victim model. This extracted information was then used to train a substitute model of approximately equivalent functionality.

Introduction to Model Extraction

A variety of extraction attacks have been developed which focus on various aspects of a model, such as decision boundary inference (Papernot, McDaniel et al. 2017), equation solving of parameter values (Tramèr, Zhang et al. 2016) and architecture extraction (Hu, Liang, et al. 2020). More general approaches approximate models' functionality by using active learning (Orekondy, Schiele et al. 2019).

Model extraction is a field of research within adversarial machine learning (Chakraborty, Alam et al. 2018). Adversarial threat models fall into two categories, white-box attacks (WBA) and black-box attacks BBA, Carlini, Athalye et al. 2019). During a WBA, the adversary has complete knowledge of a model's features (i.e., architecture, weights, and models trained dataset). In contrast, BBAs operate without knowledge about the model and only use past inputs to analyze weak points of the model. WBAs are more effective than BBA. However, for an attacker to gain the extensive knowledge needed for a WBA, severe security breaches need to occur. Thus WBA are far less likely.

Papernot and colleagues (2017) performed BBAs against MetaMind, an online deep learning API, resulting in 84.24% of adversarial examples (Szegedy, Zaremba et al. 2013) being misclassified. These results showed BBAs could be as effective as WBAs in a real-world setting.

A WBA's effectiveness within a BBA setting was achieved by extracting a substitute of the victim model. A substitute model was extracted by implementing a statistical querying method known as Jacobian Based Augmentation. For Jacobian Based Augmentation, the attacker queried the ML API with synthetic inputs chosen via a heuristic to produce outputs that fluctuate about the target model's decision boundary. With their resulting outputs, these inputs are used to build a substitute model that approximates the original victim model's decision boundary. Synthesis of inputs to decipher the decision boundary leads to efficient extraction of models . To determine a model's outputs for the entire input domain would require an infinite number of queries. Concentrating the search space to the decision boundary reduces the number of queries, making the extraction of complex deep neural

network (DNN) models feasible. With this substitute model, adversarial examples were created using white-box methods such as the fast gradient sign method (Goodfellow, Shlens et al. 2014).
A summary of (Papernot, McDaniel et al. 2017) approach is as follows:

1. **Initial data collection:** Initially collect input data that is relevant to the model domain (i.e. if image classifies digits then collect images of digits, if model classify birds then input birds)

2. **Architecture selection**: Use domain knowledge to select appropriate DNN model architecture (If image classifier use CNN architecture)

Substitute model training steps: Approach iteratively trains more accurate substitute model by repeating the following:

3. Labelling: Adversary Inputs data to a victim model which labels the data (i.e. active learning)

4. Training: Substitute model trains on the training data previously labeled within step 1

5. Augmentation: Jacobian-based data augmentation is used on the initial substitute training dataset to produce a more extensive substitute training set. The newly generated training points better represent the victim models' decision boundary.

6. Repeats steps 3 to 5 Figure 1 below illustrates a summary of the approach taken

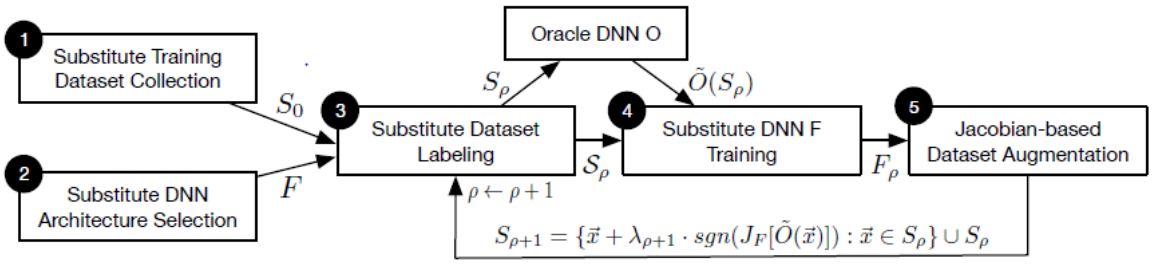


Figure 1 Approach taken to approximate decision boundary of victim model taken fromPapernot et al. (2017)

Papernot and colleagues (2017) trained on the benchmark MNIST dataset for six epochs. The achieved accuracy of 81.2% is considered low by state-of-the-art standards, however adversary examples crafted on the substitute model transfer to the victim model with a success rate of 84.24%. As the substitute model's decision boundary approximates the victim model, adversarial examples misclassified by the substitute model are also miss classified by the victim model.

This transfer rate of adversarial examples shows it is possible to compromise the victim model's output integrity by leveraging the extracted model to craft adversarial examples.

Active Learning model extraction applied to image classification.

Orekondy and collegues (2019) used a relatively simple form of active learning model extraction attack that effectively extracts the functionality of victim DNN models used for image classification.

The attacker used a pre-trained model to construct substitute models and had victim model produce labels with confidence scores from the victim models prediction outputs.

For the attack it was assumed an adversary lacks knowledge of victim models data and architecture. This Simplistic approach which can be broken into two steps

1. Adversary Inputs images to a victim model which labels the input data (i.e active learning)
2. Train a substitution model with the labelled input images from step1

Illustration of approach is shown below where:

A = Adversary, V= Victim, P(X)= model input data, Dv= Data annotated by expert

F = model, y= output from model, B = Queries,

π= Strategy to select images (i.e random selection of images or heuristic used to select images)

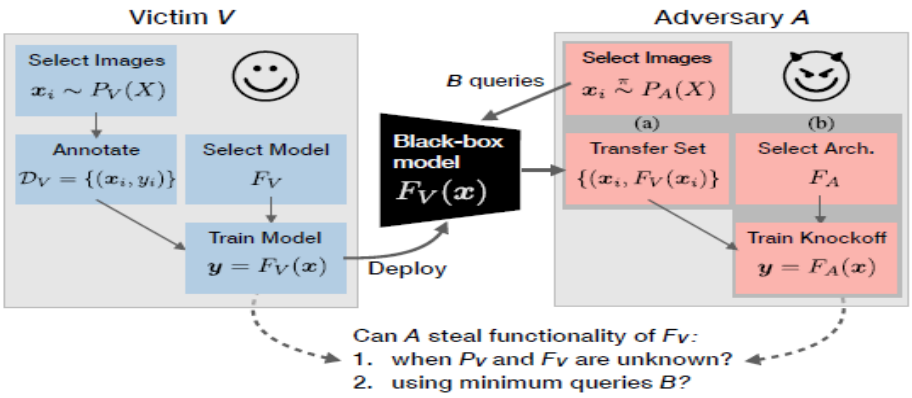


Figure 2 Illustrates adversary "A" extracting the functionality of victim "V". (Orekondy, Schiele, & Fritz, 2019)

Interestingly when this simple approach was applied to a real-world image classification API using benchmark image datasets CelebA (220k images) and OpenImages-Faces (98k images), substitute models achieved an accuracy of 0.76 to 0.82.

This result was achieved while using random selection strategies for adversarial input images. These results may be further enhanced by using a superior selection for adversarial input methods than random chance. Satisfactory models can be extracted using $30 worth of queries (Orekondy, Schiele et al. 2019). This attack's simplicity suggests that it would generalize well to the extraction of DNNs applied to other domains. Orekondy and collegues(2019) active learning attack is applied to applied to the image classification domain. This projects work differs from the previous work of Orekondy and collegues(2019) as the active learning attack is being applied to data with significantly different properties.

Active Learning model extraction applied to NLP Questioning and Answering

Active learning model extraction attacks have been implemented against DNN models applied to NLP tasks such as question and answering (Krishna, Tomar et al. 2019). Similarly, to the previous work of Orekondy and collegues (2019), the adversary uses pre-trained models as a base substitute model (Krishna, Tomar at al. 2019).

Then the adversary perfomes the following tasks:

1. Adversary Inputs images to a victim model which labels the input data (i.e active learning)
2. Train a substitution model with the labelled input images from step1

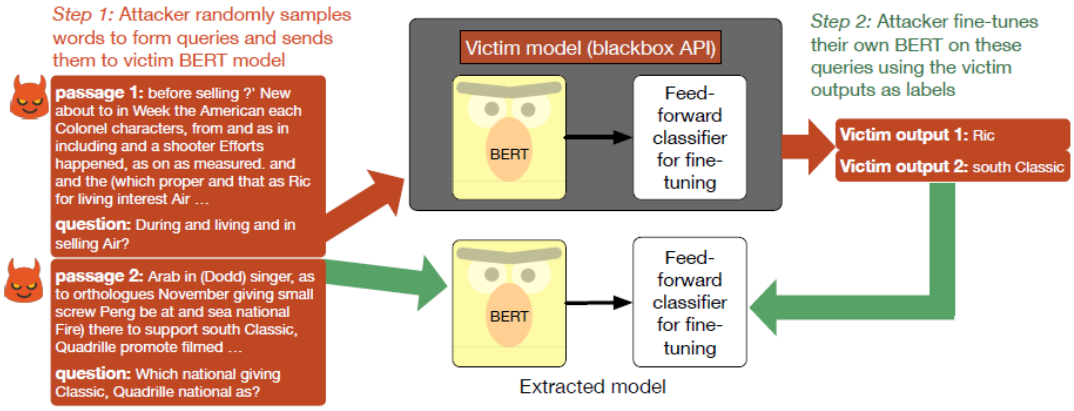


Figure 3 Illustration of active learning model extraction applied to pre trained BERT model used for questioning and answering (Krishna, Tomar, Parikh, Papernot, & Iyyer, 2019)

Model extraction results are stated below (Krishna et al., 2019).

Task	# Queries	Cost	Model	Accuracy	Agreement
SST2	67349	$62.35	VICTIM	93.1%	-
			RANDOM	90.1%	92.8%
			WIKI	91.4%	94.9%
			WIKI-ARGMAX	91.3%	94.2%
MNLI	392702	$387.82*	VICTIM	85.8%	-
			RANDOM	76.3%	80.4%
			WIKI	77.8%	82.2%
			WIKI-ARGMAX	77.1%	80.9%
SQuAD 1.1	87599	$115.01*	VICTIM	90.6 F1, 83.9 EM	-
			RANDOM	79.1 F1, 68.5 EM	78.1 F1, 66.3 EM
			WIKI	86.1 F1, 77.1 EM	86.6 F1, 77.6 EM
BoolQ	9427	$5.42*	VICTIM	76.1%	-
			WIKI	66.8%	72.5%
			WIKI-ARGMAX	66.0%	73.0%
	471350	$516.05*	WIKI (50x data)	72.7%	84.7%

Figure 4: A comparison of victim model against substitute model trained on random non-sensical sentences and substitute model trained on sentences from wiki corpus. (Krishna et al., 2019)

Two query generators were used to create input data to be labelled by the victim model. The random generator created random sentences using words from a wiki vocabulary. These random sentences were non-sensical. The wiki generator selected actual sentences from the corpus WikiText-103.

Randomly generated input data consistently produced extraction models of lower accuracy. However, the surprising accuracy of models generated from non-sensical data was still relatively close to the victim model. Results on the agreement between models suggest that the approach produced extracted models of significant fidelity. However, within the paper, it stated that agreement between models is lower on held-out data sets. On the Squad dataset(Rajpurkar, Zhang, Lopyrev, & Liang,

2016), extracted models using wiki agreement scores dropped to 59 F1. Low fidelity models were extracted using this method (Krishna, Tomar et al.,2019).

Active learning model extraction have been applied to a BERT based DNN used for NLP task of questioning and answering (Krishna et al., 2019). This project differentiates from previous work of Krishna and collegeus by performing a model extraction on a transformer based NMT model. There are similarities between machine translation and question and answering NLP tasks. However, machine translation is considered an NLP problem with unique properties. For instance, due to the unique nature of machine translation data, specific metrics designed for evaluating machine translation metrics as required to evaluate MT model outputs. One such metric is the BLEU score (Papineni, Roukos, Ward, & Zhu, 2002).

Closely related work in machine translation domain

A comprehensive study on active learning applied to neural machine translation was performed (Zhao, Zhang, Zhou, & Zhang, 2020) with significantly similarities to the work stated within this project. Similar to this project Zhao and colleagues (2020) used active learning to translate corpora to generate extra data for training. This extra generated corpus was combined with non-synthetic corpora to train a transformer model. Zhao and colleagues (2020) showed the effectiveness of active learning as they achieved a higher BLUE score with their combination of original data and synthetic data compared to a transformer model trained solely on the original dataset. This projects variation in experimental design evaluates how actually effective active learning model is when applied NMT as opposed to the compounded effect of active learned data combined with previous parallel corpora.

This project differs from the previous work of Zhao and colleagues (2020) as a synthetic dataset is solely used to train a NMT model and compares this to an NMT model trained on the original dataset. By using a purely synthetic dataset for training a NMT and comparing a NMT trained on an original dataset it gives an evaluation on how effective active learning can be at extracting information from a source model NMT.

Edunov and colleagues (2018) performed an analysis very similar to work performed within this project whereby performance of NMTs trained solely on synthetic data generated by training models was assessed. For their experiments they used use back translation (Sennrich, Haddow, & Birch, 2015a) to generate synthetic data. Back translation takes a monolingual corpus as the target language of the parallel training corpora and generates the synthetic source language via an alternative neural network. Back translating is more effective than froward translating used in active learning from the perspective of generating a higher BLEU scores. BLUE scores are calculated based on the target language sentences in comparison to the translations reference sentence. Therefore, by inputting a high-quality target sentence to synthetically generate the source sentence, the neural network output it more likely to output a higher target sentence in relation to the reference sentence. This projects work differentiate from Edunov and colleagues (2018) work as it evaluates the performance of synthetic data solely created in active learning model extraction attack model (i.e forward translate source sentence into target sentence).

Niu and colleagues (2018) evaluated the performance of synthetic data created from bi-directional translating. This approach incorporated synthetic data generated from source data (i.e forward translation) and target data (back translation). The generated synthetic data is combined with original data during evaluation. Similarly to other previously discussed related work the compounded effect of the synthetic data with the original data was measured (Niu et al., 2018). This projects variation in

experimental design evaluate how actually effective synthetic data is when applied to training NMT as opposed to the compounded effect of synthetic data combined with original parallel corpora.

Summary of Contribution

This project will leverage of the previous work (Krishna, Tomar et al. 2019;Orekondy, Schiele et al. 2019; Papernot, McDaniel et al. 2017) with regards to active learning model extraction and the threat it poses to the monetization of DNN APIs. Little work has been done to extract DNNs applied to the NLP domain (but see Krishna, Tomar at al. 2019; Pal, Gupta at al. 2020). Previous work examined functional extraction of DNNs concerning NLP tasks such as text classification and question answering (Krishna, Tomar at al. 2019; Pal, Gupta at al. 2020). However, model extraction attacks have not been applied to neural machine translation models.

The novel aspect of this project is the proposed extraction attacks are focused on NMT models. This focus is significant as nature of machine translation data is significantly different from all other model extraction attacks reviewed during the literature review.

Use of NMTs to generate synthetic parallel corpora were all done from the perspective of improving the performance of NMT using a data augmentation technique(Zhao et al., 2020), (Niu et al., 2018) and (Edunov et al., 2018).

Actions performed as part of this project are similar to the above-mentioned work, but this project looks at the application of a data augmentation technique to machine translation from the perspective of its ability to steal potential valuable intellectual property. The potential for theft of intellectual property is the primary reason the synthetic dataset is not combined with the original dataset when training NMT models for evaluation. The original dataset combined with the synthetic dataset has been shown to improve the absolute performance score of NMT models in the stated previous related work. However, by evaluating a NMT model trained solely on a synthetic dataset with a NMT model trained solely on its associated original data we can analyse how effective active learning is at extracting the victim models information/functionality.

As DNNs achieve the state-of-the-art result in machine translation tasks, it is essential to independently assess potential vulnerabilities in an adversarial setting so that effective defences and countermeasures to adversarial attacks can be created.
To the best of my knowledge, active learning extraction attacks applied to neural machine translation has not been previously performed.

Research Aim

The project aim is to evaluate how effective a synthetic parallel corpus created via active learning model extraction is at training a substitute neural machine translation model.

Research Objectives

- Compare difference in performance between NMT models trained on synthetic data and its associated original data.
- Analyse performance obtained from training on synthetic datasets of various sample sizes.
- Analyse NMT model performance with various translation evaluation metrics.
- Perform statistical significance test on evaluation metric results.

Experiment Design and Methodology

The literature review showed that active learning is a viable method for the extraction of DNN models but is understudied in relation to machine translations. The following sections outline the considerations that have been made for the realisation of stated project objectives.

Neural Machine Translation Model selection for Training

Review of team submissions to the conference of machine translation 2020 (WMT 2020) translation tasks (Bawden et al., 2020; Farajian, Lopes, Martins, Maruf, & Haffari, 2020) showed a prevalence for using transformer based NMT models. In a black box attack setting it is reasonable to assume that an attacker would use a transformer based NMT due to their prevalence in modern machine translation literature. Hence transformer based NMT are selected as the base architectures for the substitute models to be trained with experiments.

Transformer model

Previous to the deployment of transformer models (Vaswani et al., 2017), NMT models commonly used recurrent neural networks and long short-term memory to achieve state of the art results. Transformer models differentiate themselves from these previous models as there constructed using attention mechanisms.

When implementing attention mechanisms its assumed that the model input queries, token keys, and their values are all vectors. From a high-level, attention mechanisms map an input query and a set of token key-value pairs to an output. For each keys value, the weight assigned is calculated as a compatibility between the given query and the corresponding key.

The output is then calculated as a weighted sum of the compatibility values for the input query and the associated key.

Transformer models rely entirely on self-attention to generate outputs.

Scaled Dot Product Attention

The transformer model used for this experiment uses a form of attention known as Scaled Dot-Product Attention

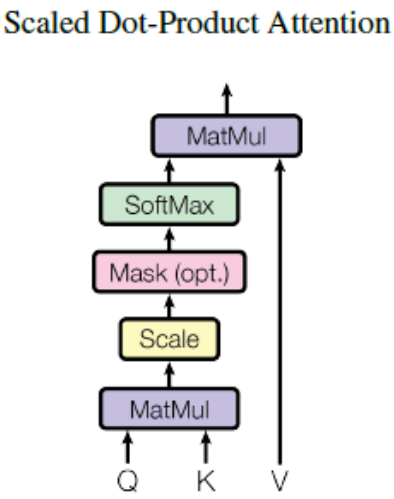


Figure 5 Scaled Dot Procut Attention source : (Vaswani et al., 2017)

Where:

- Q = queries packed into matrix,
- K = keys packed into matrix,
- V = values packed into matrix,
- d_k = dimensions of queries and keys,
- d_v = dimensions of values

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Figure 6 Scaled DoT Product Formulae ref (Vaswani et al., 2017)

A scaling factor is represented within this equation by $\frac{1}{\sqrt{d_k}}$. If queries of large dimensions are used, the calculated dot product will grow excessive in magnitude. As a result, the SoftMax function will consist of extremely small gradients. By applying $\frac{1}{\sqrt{d_k}}$ to the dot product it counter acts this effect by reducing the calculated dot product, as d_k increases in size.

Multi-Head Attention

For transformer models multi-head attention incorporates “Scaled Dot-Product” into multiple heads which execute in parallel. Multi-head attention is computational advantageous as it enables the model to process information jointly for several subspace representations at various positions.

Multi headed attention use different learned linear projections to produce projected versions of queries, keys and values to the dimensions of queries(d_k), dimensions of keys (d_k) and dimensions of values (d_v).

Figure 7 below illustrates the linear projection of queries, keys and values. Where h represents the number of attention heads operating in parallel. Each attention head operation takes as an input the linear projected versions of the query, key and value and produces an output of d_v-dimensional values. The outputs from all heads are combined via concatenation and finally projected once again.

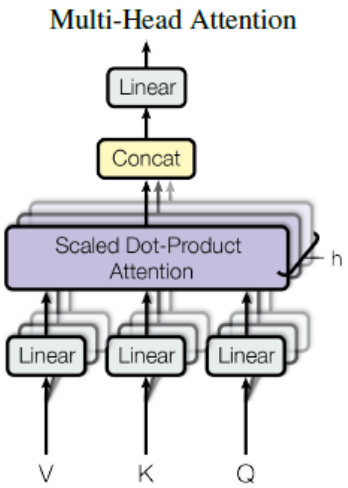


Figure 7 Multi-Head Attention contains multiple attention layers running simultaneously. Source (Vaswani et al., 2017)

Where:

- Q = queries packed into matrix,
- K = keys packed into matrix,
- V = values packed into matrix,
- d_k = dimensions of queries and keys,
- d_v = dimensions of values
- h = number of heads
- W^Q = projection matrix of queries ($W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$)
- W^K = projection matrix of keys ($W_i^K \in \mathbb{R}^{d_{model} \times d_k}$)
- W^V = projection matrix of values ($W_i^V \in \mathbb{R}^{d_{model} \times d_v}$)
- W^O = projection matrix of heads concatenated output ($W_i^O \in \mathbb{R}^{hd_v \times d_{model}}$)

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Figure 8 Multi-Head Attention Formulae (Vaswani et al., 2017)

Transformer Model Architecture

Transformer's architectures consist of an encoder and decoder stack. Encoder self-attention layers allow every position in the encoder interact with all positions of the previous layer of the encoder. Decoder self-attentions layers allow a position in the decoder interact with all previous positions in the decoder. However, the decoder differs from the encoder in that each position within the decoder is prevented from interacting with subsequent located positions within the decoder. Each encoder stack outputs memory key-value pairs which are inputted to the decoder layer multi head attention component. Each decoder outputs queries which are fed as inputs to the next encoder layer. This enables all positions of a decoder to interact with all positions of the input sequences.

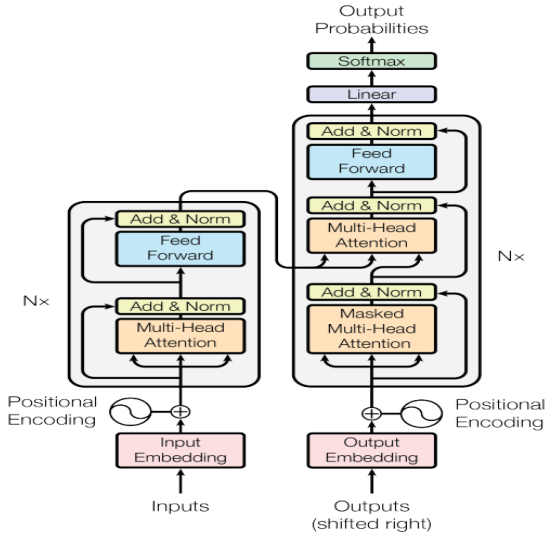


Figure 8 Transformer Model Architecture source: (Vaswani et al., 2017)

Encoder Structure

The encoder consists of 6 identical layers. An example of one of the layers is shown on the left-hand side of Figure 8 above. Each layer of the encoder consists of two components. First component is a multi-headed self-attention mechanism, and the second component is standard fully connected feed forward neural network. The output of each component and the residual input to each component is then layer normalised i.e LayerNorm(x +componentFunc(x)).

Decoder Structure

An example of one of the layers is shown on the right-hand side of Figure 8 above. The decoder is similar in structure to that of the encoder. Where the decoder has 6 identical layers. Also similarly the output of each component and the residual input to each component is then layer normalised.

The decoder differentiates from the encoder by inserting one extra component prior to the feed forward neural network component. The extra component takes the output from the encoder stack as input and performs multi head attention. The layers of the decoder also differ from that of the encoder as the first self-attention component is masked to prevent positions from appearing at later positions.

Decoder predictions for a position rely solely on the known outputs prior to that position. Ensuring only predictions from prior positions are considered is achieved by offsetting the output embedding by one position and the masking of the initial multi head attention component.

Victim Source Neural Machine Translation Model.

Experiment 1 Victim model

Training state-of-the-art NMT is computationally expensive. A transformer model trained on the WMT 2014 English to-German translation task with 8 NVIDIA P100 GPUS for 3.5 days achieved a BLEU score of 28.4 (Vaswani et al., 2017).

Fortunately, the machine translation open-source community OPEN-NMT makes available a transformer model based on the work of (Vaswani et al., 2017) at the following link:

https://opennmt.net/Models-tf/

OPEN-NMT publicly available transformer model is pre-trained on the WMT2014 English-German dataset. Reported BLEU score of OPEN-NMT pretrained model state below:

Test set	NIST BLEU
newstest2014	26.9
newstest2017	28

Figure 9: Reported BLUE scores achieved by publicly available ONMT model. Source: https://opennmt.net/Models-tf/

This freely available pre-trained NMT model is to be used to syntheses a dataset for active learning by translating sentences from English to German for experiment 1.

OEPN-NMT default specified configuration was used for source model implementation.

Experiment 2 to Experiment 7 Victim Model

For experiments 2 to 7 Facebooks FAIR's transformer model WMT2019 News translation task submission (Ng et al., 2019) was used to generate the synthetic data. A pretrained version of the Facebook-FAIR WMT 19 models is publicly available in the following code repository:

Ave.	Ave. z	System
90.3	0.347	Facebook-FAIR
93.0	0.311	Microsoft-WMT19-sent-doc
92.6	0.296	Microsoft-WMT19-doc-level
90.3	0.240	HUMAN
87.6	0.214	MSRA-MADL
88.7	0.213	UCAM
89.6	0.208	NEU
87.5	0.189	MLLP-UPV
87.5	0.130	eTranslation

Figure 10: Results of WMT2019 English to German Document News Translation Task: source : (Ng et al., 2019)

https://github.com/pytorch/fairseq/tree/master/examples/wmt19

Facebook submission placed first in the English to German document translation task. Facebook-FAIR even out ranked a human annotator at document translation. Though this model was used for document level translation task but Facebook-FAIR model it generates translation at the sentence level thus making it a perfect candidate for this project's experiments.

Facebook-FAIRs default specified configuration was used for source model implementation.

Processing Data

In theory it is possible for a neural machine translation model to perform end to end translation. However, to obtain acceptable levels of performance it is important to perform pre and post processing.

To perform pre-processing of data for these experiments an open-source unsupervised subword tokenizer and detokenizer designed for Neural-based text processing known as SentencePiece (Kudo & Richardson, 2018) is used.

SentencePiece uses segmentation algorithms:

- unigram language model (Kudo, 2018)
- byte-pair-encoding (BPE) (Sennrich, Haddow, & Birch, 2015b)

When inputting training data to a SentencePiece model it defines a fixed vocabulary size. Specifying a fixed size vocabulary is advantageous as it reduces the size of the vectorized input to the NMT model. This is turn reduces the quantity of memory consumed when training the model. However, a

disadvantage to specifying a fixed vocabulary is that it can lead to translation inaccuracy when handling corpora with unseen words.

SentecePiece library has four primary functions:

1. Normalizer (spm_encode)
2. Trainer (spm_train)
3. Encoder (spm_encode)
4. Decoder (spm_decode)

Normalization

Normalization is the process of taking semantically equivalent characters and then converting them into a standard form. Normalisation is an important process as it reduces the noise data set where three semantically identical words are condensed down to one. Common approaches to normalisation within NLP task are the use of stemming and lemmatization. However, the SentencePiece library implements NFKC normalisation which concentrates on converting similar Unicode characters to their standard form.

In the sample shown below (Fig. 11) the UCS4 sequence (e.g [45 304 300]) is converted into [1E14] (hex).

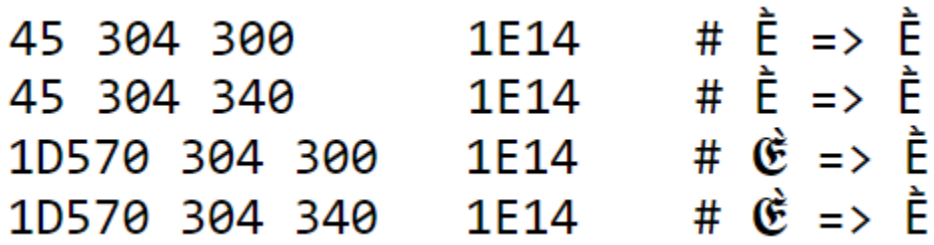


Figure 11: Example of NFKC normalization where similar Unicode's are converted to a canonical version.

Tokenisation

Tokenisation is the process of dividing a sequence of text into sub elements known as tokens. A simple form of tokenisation is the splitting of sentence where white spacing occurs (Fig. 12). However, the formation of tokens may also occur based on characters or special characters. Tokenisation is an important process as it allows for the vectorisation of inputs.

```
Frank's a student, This is his example of tokenisation.
['Frank', "'s", 'a', 'student', ',', 'This', 'is', 'his', 'example', 'of', 'tokenisation', '.']
```

Figure 12: Basic Example of tokenised sentence

When tokenising with SentencePiece its possible to use segmentation algorithms: unigram language model (Kudo, 2018), byte pair-encoding (BPE) (Sennrich et al., 2015b) , word, char. For this experiment unigram was chosen as the sub word segmentation algorithm for tokenisation.

Subword Tokenization

Subword tokenisation breaks the elements into smaller pieces. The goal of subword tokenisation is to obtain the most frequent and diverse sub words. By working with sub words the overall size of the vocabulary can be reduced

For instance, in the case where a rare word exists instead of having a single dedicated token for the rare word, the rare word can be recreated by combing sub word token.

Table 1: Example sub word token being recombined to encode alternative words

Words	**Sub word tokens**	**Vocabulary ID sequence**
sufficient	suff , i , cient	112 200 149
efficient	eff , i , cient	113 200 149
deficient	def , i , cient	99 200 149
ancient	an , cient	97 149

When applying subword segmentation it is possible to separate a single word into multiple combinations of segmentation.

Table 2: Example of Multiple Possible Sub Word Segmentations

Word	Segmented Combination 1	Segmented Combination 2	Segmented Combination 3
sufficient	suff , i , cient	su , ffici, ent	s, u, f, f, l, c, i, e,n , t

The presence of multiple alternative segmentations create a form of noise which increases robustness and accuracy when training neural translation models.

Lossless Tokenisation

One problem with subword tokenisation methods is that information regarding white spacing is not retained. This is problem is most relevant to Asian and Arabic languages, however with regards to European languages, the German language is famous for its large compound words which combine words by removing spacing (e.g das Bezirksschornsteinfegermeister→ the district chimney sweep).

To mitigate this problem SentencePiece stores the information used in encoding the normalised text in the encoded output. With this meta data it is possible to decode an encoded output as an inverse operation of the encoder (Fig. 13). The process of storing meta data within the encoded output for the purpose of inverting the encoding is termed as lossless tokenization.

$$\text{Decode}(\text{Encode}(\text{Normalize}(text))) = \text{Normalize}(text).$$

Figure 13 Sudo code for lossless tokenisation (Kudo & Richardson, 2018)

White spacing information is stored within the encoded output by appending a underscore to where white spacing occurs (Fig. 14).

```
Frank's a student, This is his example subword loss tokenisation.
_Fran k ' s _a _student , _This _is _his _example _sub w or d _loss _to ke ni s ation .
```

Figure 14: Example of subword loss tokenised sentence

This is useful for sub word tokenisation as it states whether segmentation occurs within a word or appears at the start of the sentence.

Unigram language model

The unigram language model is a sub word segmentation method which choose word segments, sub words and characters probabilistically (Kudo, 2018).

A unigram language model assumes every sub word occurs independently. Thus, a sequence of sub words probability of occurring is calculated as the product of each individual sub words probability of occurrence $p(x_i)$

X= (x1, …, X_M) ➔ where X is a sequence of sub words

$p(x_i)$ →probability of sub word instance

P(X) = $\prod_{i=1}^{M} p(x_i)$ → Product of sub words instances within sequence provides probability of Sub word sequence

$\forall_i x_i \in V, \sum_{x \in V} p(x) = 1$ →Where *V is a predetermined vocabulary,* summation of all sub words probabilities within the predetermined vocabulary = 1

x^*=arg max $P(x)$ → Where x^* is the most probable segmentation for the input sentence X.

When a vocabulary V is provided sub word occurrence probabilities $p(x_i)$ are estimated via the Expectation–maximization (EM) algorithm that maximizes the following likelihood ι assuming that $p(x_i)$ are hidden variables.

$$\iota = \sum_{s=1}^{|D|} \log\left(P\left(X^{(s)}\right)\right) = \sum_{s=1}^{|D|} \log\left(\sum_{x \epsilon S(X^{(s)})} P(x)\right)$$

In a practical setting the full set of vocabulary *V* is unknown. To find the joint optimization of a vocabulary set and their occurrence probabilities the following iterative algorithm is used:

1. Create a reasonably large size seed vocabulary from most frequent substrings and all characters present within training corpora.
2. Repeat the following sub step until vocabulary *v* reaches a desired vocabulary size.
    a. Fix the set of vocabulary V and optimize $p(x)$ with the enhance maximisation algorithm.
    b. Compute the loss_i for each subword x_i where $loss_i$ represents how likely the likelihood ι is reduced when subword x_i is removed from the current vocabulary.
    c. Sort the symbols by $loss_i$ and keep top $\eta\%$ of subwords. Note that subwords consisting of a single character are always kept avoiding out of vocabulary.

Mapping of Subword Text Tokens to Numerical IDs

Encoders are used for the pre-processing task of mapping vocabulary tokens to a numerical ID. These numerical IDs are a vector representation of the tokens.

Decoders are used for post processing of converting numerical IDs back to human readable text.

Mapping of tokens to IDs is a crucial step for using neural machine translations models. Mappings allow text sentences to be converted to sequences of IDs. These sequences of IDs are inputted to the NMT model which outputs processed translation in the form of numerical IDs. These numerical ID translations are then converted back into human readable text format by using a decoder.

Special tokens

When creating a vocabulary numerical IDs are reserved for special tokens used for marking the beginning of sentences (<s>), ending of sentences (</s>).

To allow NMT to handle out of vocabulary words a token representing unknown words are used (<unk>). During pre-processing of sentences, if a piece of text does not match any tokens within the vocabulary the unseen text is represented with the unknown tokens (<unk>) numerical ID. This prevents out of vocabulary errors from occurring and allows the NMT to translate the remainder of the sentence.

NMT required the dimension of all inputs to be the same. This is problematic from the perspective that length of sentences vary greatly and would. To circumvent this problem padding tokens (<pad>) are used to buffer sentences to ensure all inputs are dimensionally identical. Inversely, there is often outliers sentences of excessive length. It is good practice if the length of a sentence exceeds the input dimensions, truncation of the sentence is used during pre-processing.

If the truncation or omission of long sentences is not used it will lead to large memory requirement during processing of NMT which would be consumed by non-value adding padding information within sentences.

Training

Training Data and Batching

Pre-processing removed sentences which contained more than 200 and less than 2 input features. Language sentences pairs differing in ratio of size from 1:1.5 are also removed during pre-processing.

For experiments 1 to5, 100,000 sentences were randomly selected from the associated experiment corpus to form the original training dataset.

To evaluate effectiveness of active learning at differing scales of datasets sizes, experiments 6 used 3.5 million and experiment 7 used 11.5 million randomly selected sentences from the associated experiment corpus to form the original training dataset.

The source text of the original dataset (i.e English sentences) were then translated by the source model associated with the experiment to create the target text for the synthetic data set.

The synthetic data set consists of samples that have the exact same source text as the original dataset. The only difference between original data and synthetic dataset is that synthetic data target text is generated from the source NMT.

Dataset selection

Due to the availability of a native German speaker, who is also fluent in English to German all datasets used within this project are English to German parallel corpora. Any reference to datasets within this project refers to English to German parallel corpora.

Experiment 1

Experiment 1 used 100,000 samples from the WMT2017 news translation task training data. For the initial experiment, a dataset which is known to generate respectable BLEU scores on the WMT test sets was chosen.

See link: `http://data.statmt.org/wmt17/translation-task/training-parallel-nc-v12.tgz`

WMT 2017 training data consisted of the following publicly available corpora (Table 3).

File	Size
Europarl v7	628MB
Common Crawl corpus	876MB
News Commentary v12	162MB
Rapid corpus of EU press releases	156 MB

Table 3 : WMT 2017 training data corpora

Experiment 2

Paracrawl is a publicly available large dataset constructed by a web crawling website for parallel corpora (Banón et al., 2020). Paracrawl English to German dataset consisted in total of approximately 37M sentence pairs. For experiment 3 a random sample of 100,000 sentences was chosen from this massive dataset. Para crawl dataset set located at following link:

https://s3.amazonaws.com/web-language-models/paracrawl/release5.1/en-de.txt.gz

A web crawled dataset was selected for experiment 2 to examine the effect a noisy dataset would have when applying active learning.

Experiment 3 and 4

Experiment 3 and 4 use corpora created from web crawling. The dataset CCAligned (El-Kishky, Chaudhary, Guzman, & Koehn, 2019) was taken from the OPUS project (Tiedemann, 2016). The CCAligned is made of parallel or comparable web-documents. CCAligned English to German dataset consisted in total of 121.7M sentence. For experiment 3 and 4 a random sample of 100,000 sentences was chosen from CCAligned dataset. A web crawled dataset was selected for experiment 3 and 4 to examine the effect a noisy dataset would have when applying active learning.

CCaligned Dataset set is located at following link:

https://opus.nlpl.eu/download.php?f=CCAligned/v1/moses/de-en.txt.zip de-en.txt.zip

Experiment 5, 6 and 7

The UFAL dataset is a collection medical related parallel corpus. The UFAL medical dataset was recommended for use in the WMT 2020biomedical shared task (Bawden et al., 2020).

To gain access to UFAL dataset it is required to register for access at the link below:

https://ufal.mff.cuni.cz/ufal_medical_corpus

A domain specific corpus was used to evaluate how effective active learning can be when applied to a specific domain different to that of the pre-trained source model.

Test set selection

To assess the performance of models 5 test sets are used.

Withheld Corpus Data Testset

An additional sample of 20,000 test samples are taken when the training data samples are gathered from the associated experiments dataset. These 20,000 samples are to be representative of the data used to train the model but are not used during training. This withheld data test set will indicate how the model handles unseen data which representative of the training domain.

Withheld corpus datasets are to only consist of original data and will not contain any synthetic data. Synthetic data is forbidden for use in test sets as it may influence bias in results for synthetic data which would not otherwise be present.

WMT 2017 -2020 news Test sets

The choice of test datasets has a huge impact generated BLEU scores. It is meaningless to state BLEU score without stating the test set which was used to generate such scores. To make results comparable to the works peers within the MT community, fours WMT test set from 2017 to 2020 were used to

evaluate the performance of all experiment models. WMT test sets were sourced for experiments using the following publicly available code repository:

https://github.com/mjpost/sacrebleu

Justification for the use of multiple test datasets

It's known for performance of DNNs to vary greatly across different test sets. This occurs due to variation in the relevance and quality of training data in relation to the given test set data. Natural language is a particularly varied form of data due to wide range of semantics and structure found within a language. Machine translation data is even more varied again due to differences in nature between the source and target language pairs. Even relatively to other DNN applications, machine translation data is exceptional complex. Due to the data's complex nature, translation models typically require a huge quantity of data to generalise well to test sets. Thus, it expected that a machine translation models performance would vary widely across different datasets. To confirm that any difference in model performance was not due to chance, the same multiple test sets are applied experiments.

Hardware and Schedule

Models are trained on one google cloud platform compute engine VM instance. The VM instance has a NVIDIA TESLA V100 GPU and a n1-standard-8 machine type (i.e 8 vCPUs, Memory 30GB). The base transformer model executed 0.84 steps per second using the hyperparameters described throughout this report. Within each experiment each base model is trained for 12 hours. A checkpoint of the model is saved every A check point of the trained model is saved every 5000 steps. Final evaluation is performed on model checkpoints which have been trained for 50,000 steps.

Basis of Evaluation

Currently the gold standard in evaluation of machine translation is the use of human translators. Due to the sheer size of data associated with state-of-the-art translation models, using a human translator is a timely and costly process. To reduce the cost and time required to evaluate development translation systems, automatic evaluation metrics are used.

The following sections contains an assessment of the currently widely used automatic evaluation metrics and the Human evaluation methods used for assessing quality if machine translation outputs.

Automatic Machine Translations Metric

Development of machine translation automatic evaluation metrics is still an active area of research. Due to the unusual property of machine translation tasks where different outputs may be correct, specialized evaluation metrics are required.

Intuitively a good automatic evaluation metric for translations should consistently produce a meaningful score on translation quality, which can be used to score and rank translation systems accordingly.

There are currently metrics available for automatically evaluating machine translations, but they are not without faults. Prevalent automatic metrics compare a machine translation with that of a human translator and calculate translation quality based upon this comparison (Mathur, Baldwin, & Cohn, 2020).

Bilingual Evaluation Understudy (BLEU) metric (Papineni et al., 2002)

As of writing the dominant metric used for automatically evaluating machine translation model is the BLEU score. BLEU has been widely adopted to the NLP task of machine translation due to the following reasons (Post, 2018).

- Relatively little computation power required
- Applies to various languages pairs
- Somewhat correlates with human judgement

BLUE compares n-grams of machine translations with a human reference translation. Where more n-gram matches are counted, the machine translation is of greater quality. BLEU uses a modified unigram precision which is based on the standard precision measure.

A standard precision formula applied to unigram machine translation is as follows:

Where:

m = Tokenised machine translated sentence

h = Tokenised reference human translated sentence

t_p = Count of correctly translated tokens (i.e machine translated word is in *h*)

f_p= Count of Incorrectly translated tokens (i.e machine translated word is not in *h*)

$$Precision = \frac{t_p}{t_p + f_p}.$$

$$t_p = \sum_{x \in [m]} count(x \cap [h])$$

$$t_p + f_p = \sum_{x \in [m]} count(x)$$

$$Unigram\ Prescision = \frac{\sum_{x \in [m\,]} count(x \cap [h])}{\sum_{x \in [m]} count(x)}$$

However, the following examples shows how the standard precision score is not a suitable metric for evaluating machine translation.

Example of standard precision metric applied to machine translation

Reference Translation: Frank is a student in the student union

Machine Translation: student student student student student student

When standard precision is applied the following high precision, score is obtained for the unsuitable translation:

$$Unigram\ Prescision\ = \frac{1+1+1+1+1+1}{6} = \frac{6}{6}$$

The standard precision score does not account for scenarios where a token (e.g student) is present within the machine translation more than the reference translation. The modified precision metric implemented BLEU by corrects this problem by Count Clipping.

Where:

r_f = Frequency of machine translated token within *referenced translation.*

$$r_f = \sum_{y \in [h \cap m]} count(y)$$

$$Count_{clip} = \min\left(t_p, r_f\right)$$

$$Modified\ Unigram\ Prescision = \frac{\min\left(t_p, r_f\right)}{\sum_{x \in [m]} count(x)}$$

Example of modified unigram precision metric applied to machine translation.

Reference Translation: Frank is a student in the student union

Machine Translation: student student student student student student

When modified precision is applied the following low precision, score is obtained for the unsuitable translation:

$$Unigram\ Prescision = \frac{\min(6, 2)}{6} = \frac{2}{6}$$

A machine translation stating the same words as the human translation is considered as adequacy. To gain assessment of fluency, n-grams matches are used. Where greater n-gram length matches represent greater fluency. When assessing fluency the previously stated modified unigram precision metric is used but instead of counting unigrams, the matching token length is specified by the size of the n-gram.

Where:

m = Machine Translation tokens

h = Reference Translation tokens

t_p = Count of correctly translated n-gram

f_p= Count of Incorrectly translated n-gram

r_f = Frequency of machine translated n-grams within reference translation.

$$t_p = \sum_{x_{(n-1)} \in [m\ tokens]} count(x_{(n-1)} \cap [h\ tokens])$$

$$r_f = \sum_{y_{(n-1)} \in [h \cap m]} count\left(y_{(n-1)}\right)$$

$$Modified\ n-gram\ Prescision\ = \frac{\min\left(t_p, r_f\right)}{\sum_{x_{(n-1)}\in[m\ tokens]]} count(x_{(n-1)})}$$

Example of modified bi-gram precision metric applied to machine translation.

Reference Translation: Frank is a student in the student union.

Machine Translation: a student student student student student

$$Bigram\ Prescision\ p_2 = \frac{1}{5}$$

Evaluation of sentences will vary greatly from sentence to sentence. To account for this variation BLUE combines an entire test corpus when calculating modified n-gram precision.

Where:

M = Entire collection of all machine translations

m = Single tokenised machine translated sentence

H = Entire collection of all reference translations

h = Single tokenised reference translation sentence

t_p = Count of correctly translated n-gram within m

r_f = Frequency of machine translated n-gram within h.

$$p_n = \frac{\sum_{m\in[M]}\sum_{x_{(n-1)}\in[m]}\min\left(t_p, r_f\right)}{\sum_{m\in[M]}\ \sum_{x_{(n-1)}\in[m]} count(x_{(n-1)})}$$

The n-gram precision scores are combined via an average logarithm with uniform weighing for each of the n-grams.

$$Combined\ n\ gram\ precision\ score\ = \sum_{n=1}^{N} w_n \log p_n$$

Brevity penalty

Generated translation is required to be appropriate in length (i.e not to long and not to short). The modified precision evaluation metric penalises sentences which are excessive in length as it counts the correctly translated words. If the translated sentence is excessive in length, the translation is likely to have a greater number of false positives thus penalising excessive length. In contrast a short translation relative to a reference translation is mor likely to produce a high modified precision score even if the translation is of poor quality

Example of too short a translation with high precision score

Reference Translation: Is a

Machine Translation: Frank is a student at Munster Technological University Studying AI

The example translation a above would produce a high precision score despite it lacking significant information.

Thus, to penalise sentences of inappropriate length a brevity score is introduced.

Brevity penalty is calculated as follows:

- Sum the length of each reference sentences over the entire corpus.
- Sum the length of each translation over the entire corpus.
- Dived summed translation length by summed reference sentence length

If a translation exactly matches its reference sentence in length it produces a brevity score of 1.

Problems with BLEU

The major limitation of BLEU is that it does not consider sentence structure. Sentence structure is not evaluated by BLUE as location of n-grams words within sentences. Some fluency of sentence is assessed as n-gram size increases but is still considered a poor metric with regards grammatical assessment of sentences. The next major limitation of BLEU is that it does not consider the context and meaning or words. The inability for BLEU to properly assess context and meaning occurs as BLEU is based on n-gram matching where it blindly matches tokens. BLEU is a unstable metric when assessing MT model outputs at the sentence level and should be only used as a basis for evaluation when assessing aggregated MT outputs at the full system level (Papineni et al., 2002).

Reporting of BLUE scores can be problematic as different implementations can produce varying slightly varying results. Majority of open-source BLUE score implementations are originally based upon the http://www.statmt.org/moses/ repository. Variance occurs due to altering settings associated with smoothing, length penalty and maximum n-gram length. In order to reduce variance in results sacreBLUE was proposed in order to bring better consistencies evaluation via the BLEU metric(Post, 2018). To create consistent BLEU metric sacreBLUE performs internal tokenisation and pre-processing of the test references.

To create comparable BLEU scores the publicly available code repository https://github.com/mjpost/sacrebleu (Post, 2018) was used to supply WMT2017 to WMT2020 test datasets for the evaluation of outputs NMTs

Sentence Level BLEU score

Due to the large quantity of translated test sentences, sacreBLEU sentence level scoring functionality was used to help identify translations of interest. Applying the BLEU scoring metric at the sentence level is widely considered unreliable and should not be used as a basis for evaluation. However, it has been used as part of this project as a useful tool for marking interesting translation for further human inspection.

Metric for Evaluation of Translation with Explicit Ordering (METEOR) Metric

METEOR is an automatic metric which attempts to correct some the problems associated with BLEU such as a lack of recall and the lack of explicit word-matching between machine translation outputs and its associated reference translation (Banerjee & Lavie, 2005)

METEOR evaluates MT based on matching of unigrams between MT model translated output and an original reference translation (Banerjee & Lavie, 2005). These matched unigrams are used to calculate a precision and recall. A Fmean is calculated using a weighted harmonic mean with a lighter weighting assigned to precision then recall. See formula for calculating weighted harmonic F-score (Fig. 15).

$$Fmean = \frac{10PR}{R + 9P}$$

Figure 15:Weighted Harmonic mean F-Score. Source : (Banerjee & Lavie, 2005)

METOER also calculates a penalty score based on alignment of tokens between MT output translation and its associated reference translations (Fig. 16).

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched} \right)$$

Figure 16: METOER alignment penalty Source : (Banerjee & Lavie, 2005)

A chunk is a consecutive matching of tokens between MT output and reference translation. When a token no longer matches in the sentences sequence of tokens it is the end of the chunk and the next chuck begins (i.e. a form max n-gram matching). The fewer the number of chunks relative to the number of matching unigrams results in a smaller penalty score.

The final Meteor score is a combination of the calculated F-mean and the alignment penalty (Fig. 17).

$$Score = Fmean * (1 - Penalty)$$

Figure 17: METEOR score formulae: (Banerjee & Lavie, 2005)

The above Meteor score formula (Fig. 17) calculates the METEOR F-score for a single sentence. When calculating a MT systems overall performance on a corpus, the cumulative precision, cumulative recall, and cumulative penalty over the entire corpus. These aggregated values are then applied to the same F-mean score formula and subsequent METEOR score formulae.

Problems with METOER

Like BLEU, METEOR is a n-gram based evaluation metrics, so shares many of the same problems associated with BLEU.

Word and character n-gram F-scores (chrF++) metric (Popović, 2015)

Character n-gram F-score is a simple but effective automatic evaluation metric. They have been empirically shown to show high level correlation with human ranking assessments (Popović, 2015). To improve correlation with direct human assessment short word n-gram F-scores (i.e 2 word n-gram) are used in addition to character n-gram scores (Popović, 2017). N-gram F-scores metric is relatively simple to imply as it is language independent and robust against variance in some pre-processing steps such as tokenisation.

Formulae for generating n-gram based F scores is as follows:

Where:

ngramP= percentage of character/word n-grams in system translated sentences

ngramF= percentage of character/word n-grams in reference translation and system translated sentences

β = Coefficient for assigning β times more importance to recall than to precision. If β =1 then importance between recall and precision is equal, if β =2 then importance between recall is twice that of precision

$$ngram\beta = (1 + \beta^2)\frac{ngramP \cdot ngramR}{\beta^2 \cdot ngram + ngramR}$$

For evaluation of translation using n-gram-F scores the following publicly available repository was used: https://github.com/m-popovic/chrF.

Problems with chrF++

Like BLEU, chrF++ is a n-gram based evaluation metrics, so shares many of the same problems associated with BLEU.

Translation Error Rate (TER) Metric

TER is a simple automatic MT evaluation metric which sums the total number of modifications to a model's translation are required to match the reference translations.(Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006). Insertion, Deletion, movements, and switches are all considered as modifications which are tallied by the TER metric.

Problems with TER

TER is not a n-gram based evaluation metric but in essence stills share many of the same problems as it relies on matching tokens.

Modern Alternative Automatic Evaluation Metrics.

Automatic Machine Translation evaluation metrics is an active area of research due to faults in the currently predominantly used metrics. WMT2020 Metric shared task had 27 different metric submissions (Mathur, Wei, Freitag, Ma, & Bojar, 2020). Majority of submission similarly to BLEU rely on evaluating machine translation outputs against their associated reference translation. Four of the submission did significantly differ by submitting metrics which evaluated machine translation outputs by evaluating them to the original source translation. WMT2020 reported progress on source-based evaluation metrics but reference-based translation methods are still superior in performance.

A common theme among many of the competitor's submission is the use of contextual word embedding and predictor-estimator models. Results of the competition indicated that metrics based on contextual embedding such as YiSi (Lo, 2020) currently produce the best results. Unlike n-gram based methods, contextual word embedding approach can estimate the contextual meaning of translation.

Justification of Automatic Evaluation Metric Selected for Experiments

For all experiments BLEU is the primary metric for evaluating experimental results. The choice of BLEU as the primary valuation metric is due to current widespread use by the machine translation academic community. By reporting BLEU scores and tests sets the BLUE score was generated on it allows for comparisons to made with the works of others.

TER, METEOR and chrF++ were used for some experiments to confirm if magnitude of difference in model performance varied across metrics due the apparent strength and weaknesses of each metric. TER, METEOR and chrF++ are not as widely discussed as BLEU scores. However, TER, METEOR and chrF++ still do feature analysis of model's performance as they are computation inexpensive and are simple to consistently to apply experiments thus effort to reward is often justifiable for their implementation. Currently machine translation models are not widely assessed via contextual word

embedding based metrics. Due to this lack of use, no contextual word embedding methods as part of this project.

Human Evaluation Methods

Due to flaws in current automatic evaluation metrics for machine translation, human evaluation is still considered the gold standard when assessing the quality of translations. However even human annotators have faults as is evident when annotators have low level of agreement (Graham, Baldwin, & Mathur, 2015) Human evaluation primarily assesses sentences based on adequacy (i.e does translate sentence accuracy represent true meaning) and fluency. Disagreements on assessments even occur where the same person revisits sentences they previously assessed.(Callison-Burch, Fordyce, Koehn, Monz, & Schroeder, 2007). Effort have been made in the domain of MT to error in human evaluation by standardising approaches. WMT shared task have previously used the following methods:

Relative Ranking

For relative ranking assessment human judges rank different translation in order of perceived quality. (Callison-Burch et al., 2008) This approach dictates which translation is superior with denoting any information with regards to actual quality of translations. Relative ranking was the official gold standard evaluation metric from WMT shared task from 2008 to 2016.

Direct Assessment

Direct human assessment (DA) consist of assigning an absolute quality scores to each translated sentence based solely in comparison to the translation reference sentence (Bojar et al., 2016). To deal with conflicting human judgements the scores can be averaged too together.
Direct assessment has been the official gold standard for evaluation metrics since WMT 2017shared task.

Human UCCA-Based Evaluation of Machine Translation (HUME)

Hume is a semantic based human evaluation methods (Birch, Abend, Bojar, & Haddow, 2016). Unlike the relative ranking or direct assessment HUME provides insight into what translation errors are. Hume has been used as part of the WMT 2020 shared task.

Justification for Relative Ranking as Chosen Human Evaluation Method for Experiments

For experiments the outputs of test sets are evaluated by a native German speaker judge. The judge performs a relative ranking assessment whereby model A or model B were ranked based on the quality of the outputted German translation in comparison to the associated reference German sentence. The human evaluation is a blind test whereby the human judge had no information regarding the model used to generate translations being assessed. Ties where the judge cannot score the translation differently, are counted as NA. The result used to deem which model performed better consist only of the scores of where a model's translation was deemed better. As ties are indicative of the difference in performance, they were still tallied for analysis purpose.

Due to human resource constraints relative ranking was the chosen human evaluation metric. For a human judge to assign a consistent numerical absolute scoring they would require a guide and training. HUME would require significant training time and would take considerably longer to execute as it is a much more fine-grained approach.

The relative ranking method is coarser in comparison to the other human evaluation methods and does not produce information as descriptive as the alternative methods. However due to the ease at which it can be applied it was the chosen human evaluation metric for this project experiment due to

human resource constraints. Human evaluation is costly and time consuming. Due to resource constraints human evaluation is unfortunately not used as extensively as desired for this project to evaluate machine translation outputs.

Statistical Significance Tests in Machine Translation

To validate empirical results statistical multiple statistical significance test are applied. Statistical significance test plays a crucial role in validation of empirical results as it estimates the probability that results were not obtained coincidentally.

The focus of this project is to evaluate the performance of a neural translation model (M_t) trained on a language paired dataset created by humans O_d vs being trained on a language paired synthetic dataset created by a pre-trained NMT S_d . To evaluate performance evaluation metric E is used to quantify performance. E represents evaluation metrics BLEU and chrF

$\delta(M_t) = E(O_d, M_t) - E(S_d, M_t)$

Performance difference $\delta(M_t)$ between trained M_t is used as the test statistic for hypothesis testing.

Paired Difference Test

A paired difference test checks if there is a significant difference between paired sample data. The paired difference test performs a univariate analysis on a sample pair to verify the null hypothesis H_0 that the pair of two variables were taken from the same population distribution (Fisher, 1937) (i.e statistically both trained models have the same performance.)

$H_o: \delta(M_t) < 0$

The paired difference test creates a p-value which represents the probability the null hypothesis cannot be rejected. The lower a p-value, the less likely the sample pair is drawn from the same population distribution. A p-value less than 0.05 signifies the difference between the sample pair is statistically significant and the null hypothesis can be rejected. Paired difference test only provides confirmation that one model consistently performs differently than the other, however it does not evaluate the magnitude of the differences. To evaluate the magnitude in difference the mean and standard deviations of the metric scores were examined.

Wilcoxon signed rank test.(Wilcoxon, 1945, 1992)

As the experiment process varied for each experiment (i.e dataset set used, sample size, alterations in pre-processing) it cannot be assumed that calculated scores are normally distributed. The assumption that data samples do not come from a normally distributed population was verified by implementing a Shapiro-Wilk test (Shapiro & Wilk, 1965).

Due to data distribution assumptions a nonparametric version of paired sample t-test known as Wilcoxon signed-rank test was chosen to determine the statistical significance of experimental results. The null hypothesis for Wilcoxon signed-rank test is that the differences between paired sample values follow the symmetrical distribution around zero (Wilcoxon, 1945, 1992)

The procedure for a Wilcoxon signed rank test is as follows:
Where :

N = Sample size
Xo,i = Original sample data value
Xs,i = Synthetic sample data value

1. Calculate absolute difference between paired samples $|\mathrm{Xo,i} - \mathrm{Xs,i}|$

2. Exclude pairs where difference between sample equals zero
3. Rank the absolute differences in order from smallest (ranked 1) to greatest absolute (ranked N_r). if ranks are tied the average is assigned to those ranks (i.e rank 1 and 2 both become 1.5)
4. Each rank is assigned a sign (i.e – / +) according to the sample pair difference.
5. Calculate the sum of the signed ranks

$$W = \sum_{i=1}^{N_r} [sgn(X_{0,i} - X_{s,i})]$$

6. Use summed signed ranks as a test statistic to confirm null hypothesis H_0 should be rejected.

$$reject\ H_0\ if\ |W| > W_{critical,N_r}$$

A two-tail hypothesis was chosen to decrease the power of the test to obtain a conservative estimation if the population distribution of experimental results between the models was statistically significantly different or if any variations was due to random variation.

Multiple Optimizer Runs to account for parameter estimation noise

When training NMT models it possible to get variations in final model performance due to noise in parameter estimations. If model is trained and tested only once the measured performance may be on the higher or lower end of performance distribution. To account for this variation in performance and get better hypothesis testing it is advisable to perform multiple optimizers runs on machine translation models (Clark, Dyer, Lavie, & Smith, 2011). Whereby an optimizer runs are defined as the exact model training configuration executed multiple times.
For experiment 1 and experiment 7, three optimizer runs were performed creating 3 models trained on the same synthetic data and 3 models trained on the same original data. Ideally execution of more optimizer runs would have been preferable but was not performed due to hardware resource constraints.

Paired Bootstrap Resampling

The bootstrap method for statistical significance testing (Tibshirani & Efron, 1993) is often used with the BLEU evaluation metric to evaluate machine translation systems (Riezler and Maxwell III 2005). The bootstrap methods randomly take data points from the sample data to create many virtual sample datasets. These virtual datasets are treated as proxy data points generated by the original experiment.

To create the virtual datasets any data point within the original sample data can be placed within the virtual dataset more than once. Resampling is repeated until the virtual data set has the same sample size as the original dataset. This process is performed multiple times to create many virtual datasets with varying combination of datapoints from the original sample data (Fig.18). Each of these virtual datasets represents a plausible combination of values which may have been obtained during the actual experiment. Each of virtual dataset has its statistics calculated. The combination of sample statistics across the virtual datasets is then used for hypothesis testing to evaluate if results were statistically significant.

The primary assumption of the bootstrap methods is that the provided sample data is representative of the population distribution. Thus, when applying the boot strap method for statistical significance tests it was only applied to a single test dataset within the experiment. As the virtual data is assumed to be representative of the population distribution the additional data points converge on the correct population distribution and improve confidence intervals.

Set $c = 0$
Compute actual statistic of score differences $|S_{\mathrm{X}} - S_{\mathrm{Y}}|$ on test data
Calculate sample mean $\tau_B = \frac{1}{B}\sum_{b=0}^{B} |S_{\mathrm{X_b}} - S_{\mathrm{Y_b}}|$ over bootstrap samples $b = 0, \ldots, B$
For bootstrap samples $b = 0, \ldots, B$
    Sample with replacement from variable tuples for systems X and Y for test sentences
    Compute pseudo-statistic $|S_{\mathrm{X_b}} - S_{\mathrm{Y_b}}|$ on bootstrap data
    If $|S_{\mathrm{X_b}} - S_{\mathrm{Y_b}}| - \tau_B\ (+\tau) \geq |S_{\mathrm{X}} - S_{\mathrm{Y}}|$
        $c++$
$p = (c+1)/(B+1)$
Reject null hypothesis if p is less than or equal to specified rejection level.

Figure 18: Pseudo code of Bootstrap for Statistical Significance Test: Source: (Riezler and Maxwell III 2005)

Approximate randomization (AR) exchanges.

Fisher-Pitman's permutation test the null hypothesis on statistical distribution of data by calculating the test statistics on all possible translation of the test dataset. To evaluate different permutations of test set, translated sentences are exchanged with corresponding translation sentences between the two NMT systems under evaluation. If there is no significant difference between sentences, then shuffling of examples between test sets should not significantly alter the calculated metric score (Noreen, 1989).

The concept of this test is simple however the computation cost of generating all permutation is O^n Due to the computation cost of performing the Pitman's permutation test in it is more practical to use a variation of the test known as approximate randomization. Approximate randomization (AR) tests only randomly shuffles a fix quantity of sentences between systems instead of all possible permutations (Fig. 19). Calculated BLEU scores vary significantly between different test sets thus only hypothesis sentences for a reference sentence are only swapped with hypothesis sentences associated with the same reference sentences within the test set. Empirical evidence suggest that sampling based methods such as approximate randomization results in fewer type 1 errors than sampling free methods such as the Wilcoxon signed-rank test: (Riezler & Maxwell III, 2005)

Set $c = 0$
Compute actual statistic of score differences $|S_{\mathrm{X}} - S_{\mathrm{Y}}|$ on test data
For random shuffles $r = 0, \ldots, R$
    For sentences in test set
        Shuffle variable tuples between system X and Y with probability 0.5
    Compute pseudo-statistic $|S_{\mathrm{X_r}} - S_{\mathrm{Y_r}}|$ on shuffled data
    If $|S_{\mathrm{X_r}} - S_{\mathrm{Y_r}}| \geq |S_{\mathrm{X}} - S_{\mathrm{Y}}|$
        $c++$
$p = (c+1)/(R+1)$
Reject null hypothesis if p is less than or equal to specified rejection level.

Figure 19 Pseudo code of Approximate Randomization Exchange Statistical Significance Test: Source: (Riezler & Maxwell III, 2005)

Statistical Analysis Implementation Used for Experimental results

Experiment 1 and Experiment 7 Statistical Analysis

Three optimizer run have been performed for experiment 1 and experiment 7. For theses experiment the publicly available code repository https://github.com/jhclark/multeval was used(Clark et al., 2011). MultEval takes translations generated from NMT model created from multiple optimizers runs and calculates the scores of 3 popular machine translation metrics. To obtain descriptive statistics such as standard deviation and mean for each test set experiment 1 performed bootstrap resampling across the three optimizer runs. To obtain p-values for the 3 calculated metrics scores stratified approximate randomization across is performed the three optimizer runs for each test set. MultEval produces an output with statistical coefficients (Fig. 20).

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	12.1 (0.1/0.2/-)	19.0 (0.0/0.1/-)	72.2 (0.1/0.1/-)	96.9 (0.1/0.5/-)
system 1	17.2 (0.1/0.1/0.0001)	22.0 (0.1/0.0/0.0001)	66.4 (0.1/0.0/0.0001)	96.4 (0.1/0.3/0.0001)

Figure 20: Example of table produced from using MultEval

n = number of optimizers runs

s_sel = average variance over all optimizer runs due the given test set. A large s_sel indicates high variance within the test set and that a larger sample size may be required to obtain reliable results. s_sel is calculated using bootstrap resampling

s_opt= variance due to optimizer instability. Calculated from the collective metric score over all optimizer runs.

p=p-value calculated via random approximation.

The values located under the three metric scores (i.e BLEU, METEOR, TER) are the averaged metric score results across all optimizers runs for the given system. Due to the assumptions made by approximate randomization the statistical significance test should be assessed for the individual experiment and cannot be aggerated as across experiments with difference in data distribution.

Statistical analysis across all experiment

The nonparametric Wilcoxon signed rank test is used to assess if a statistically significant difference between NMT models trained on synthetic data and original data across all experiments exists. The Wilcoxon signed rank test is applied to BLEU scores results across all experimented to determine if statistical difference exists. The Wilcoxon signed rank test is applied to chrF+ scores result across all experimented to determine if a statistical difference also exists under the alternative evaluation metric.

Implemented Experiment Methodology

In total 7 experiments were executed. A summary of variations between the experiments is shown the table below:

Experiment	Corpus	Framework	Source Model Supplied by	Source Model BLUE Score on *	Optimizer Runs	Encoding	Training Data Size	Min-Max Sentence Length tokens	Updates during Training	Aprox Training Time (hrs)
1	WMT2017	Tensorflow	ONMT-TF	28	3	Sentece Piece	100,000	2 to 200	50k	17
2	Paracrawl	Pytorch	Fairseq	30.9	1	BPE	100,000	2 to 200	22K	2.5
3	OPUS	Pytorch	Fairseq	30.9	1	BPE	100,000	2 to 200	20k	2
4	OPUS	Pytorch	Fairseq	30.9	1	BPE	100,000	2 to 80	20k	2
5	UFAL	Pytorch	Fairseq	30.9	1	BPE	100,000	2 to 80	20k	2
6	UFAL	Pytorch	Fairseq	30.9	1	BPE	3,500,000	2 to 80	50k	4.5
7	UFAL	Pytorch	Fairseq	30.9	3	BPE	11,500,000	2 to 80	100k	12

*Stated BLEU score was achieved on WMT2018 New test set.

Except for variations in configuration stated in the above table, all 7 experiments follow the exact same process pipeline.

The active learning model extraction attack experiment pipeline is shown below (Fig. 21). Steps to active learning model extraction attack experiment pipeline are as follows:

1. Take the original source language sentences from parallel corpus and create a monolingual corpus.
2. Input monolingual original Source Language corpus into victim source model for translation.
3. Combine the original source sentences with its associated synthetic target sentences to create synthetic parallel corpora.
4. Perform identical pre-processing of synthetic parallel corpora and original parallel corpora.
5. Train NMT models of identical configuration on synthetic parallel corpora and original parallel corpora
6. Post training both NMT models performance are evaluated by translating the same withheld corpus data unseen by the NMT model
7. Post training both NMT models performance is then further evaluated by translating the same WMT test sets

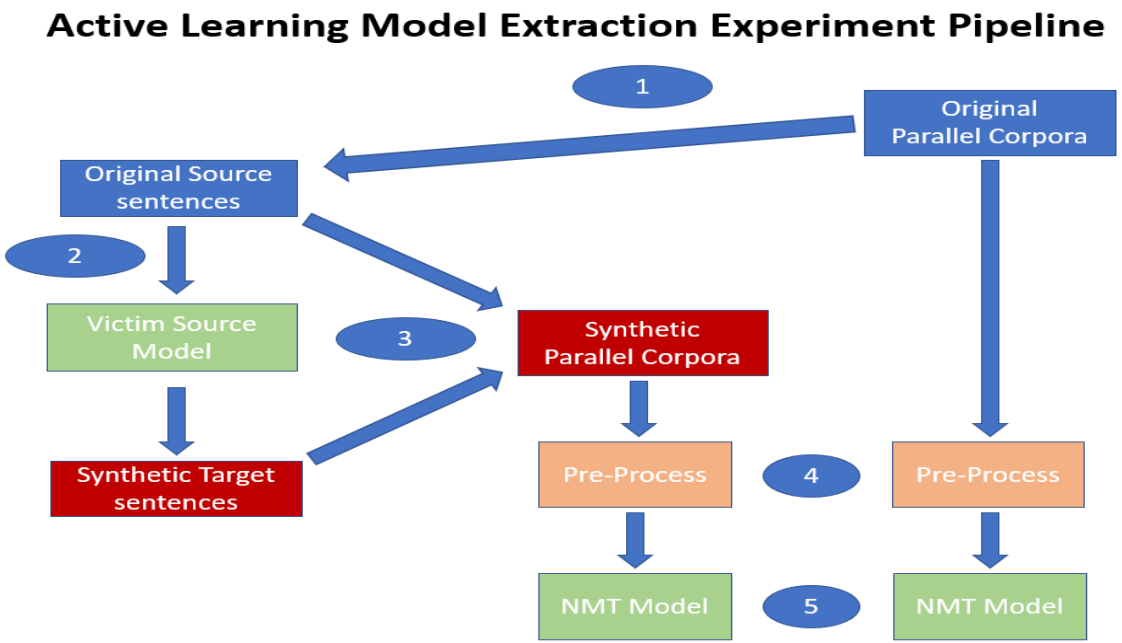


Figure 21: Process flow of Active Learning Model Extraction Experiment Pipeline

Note: Open-TF and Fairseq NMT model inference APIs can take raw inputs. These raw inputs are then pre-processing and post processing as required by the victim model. Thus, no pre-processing and post processing of victim data model was required.

Implementation and Results

Introduction

The previous chapter experiment methodology was designed to evaluate how effective active learning model extraction is at generating synthetic data for training neural machine translation models. This chapter present the results of NMT models trained on the original language paired dataset and the synthetically generated language pair dataset.

Evaluation of Results

Statistical significance tests

The results of the Wilcoxon Signed Rank Test Tailed Hypothesis Test on all combined test scores indicate that there is a difference in the distribution of scores generated by the NMT trained on the original dataset and the NMT trained on synthetic dataset. The means scores for the original data sets were lower than that of the synthetic data. Original data BLEU scores and character + word F scores had a higher standard deviation than that of the synthetic data (Table 4).

Trained Model Data	BLUE Score Mean	BLEU Score Standard Deviation	c6+w2-F2 Score Mean	c6+w2-F2 Score Standard Deviation
Original Data	10.63	10.39	34.17	12.88
Synthetic Data	14.48	9.4	41.49	8.48

Table 4:Desrciptive Statistics from analysis of all experimental results

Due to the surprising results observed as part of the initial statistical significance test (synthetic data scored better than original data), it was decided to perform another more conservative analysis. When evaluating scores produced by the various experiments, it was noted that UFAL data produced majority of the worst performing NMT models trained on original data.

For the 2nd Wilcoxon Signed Rank Test Tailed Hypothesis Test the UFAL data experimental results were treated as outliers and excluded (i.e experiment 5 to experiment 7 were excluded). UFAL data experiments were treated as outliers due to the extremely poor performance of the NMT model trained original data.

The 2nd Wilcoxon Signed Rank Test Tailed Hypothesis Test combine the scores of Experiments 1 to Experiment 4. The results of the 2nd Wilcoxon Signed Rank Test Tailed Hypothesis Test scores indicate that there is a difference in the distribution of scores generated by the NMT trained on the original dataset and the NMT trained on synthetic dataset. The means scores for the original data sets were lower than that of the synthetic data. Original data BLEU scores and character + word F scores had a lower standard deviation than that of the synthetic data (Table 5).

Trained Model Data	BLUE Score Mean	BLEU Score Standard Deviation	c6+w2-F2 Score Mean	c6+w2-F2 Score Standard Deviation
Original Data	14.05	10.25	41.18	8.07
Synthetic Data	16.74	11.09	44.12	8.57

Table 5: Descriptive Statistics from analysis of all experiments 1 to experiment 4

The more conservative analysis NMT models trained on synthetic data metric scores were still higher than that of NMT models trained on the original data. however Inversely to the first statistical

significance test the standard deviation of NMT models trained on synthetic data was higher than NMT models trained on original data.

For experiment 1 and experiment 7 the statistical significance test which used approximate randomization found that NMT models trained on synthetic parallel corpora performed statistically significantly different to the NMT model trained on original parallel corpora. The values located under the three metric scores (i.e BLEU, METEOR, TER) are the averaged metric score results across all optimizers runs for the given system. Higher scores on BLEU and METEOR, and the lower score on TER indicate that NMT models trained on synthetic data performed better than the baseline NMT models trained on original data.

Experiment 1 WMT17 Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	16.5 (0.3/0.0/-)	22.0 (0.1/0.1/-)	66.0 (0.3/0.1/-)	97.6 (0.3/0.2/-)
system 1	21.5 (0.3/0.2/0.0001)	25.0 (0.2/0.1/0.0001)	60.3 (0.4/0.1/0.0001)	98.9 (0.3/0.4/0.0001)

Note: For the full set of statistical significance test results associated with this project refer to appendix sections:

“Statistical Significance Test Applied to Combined Experimental Results”

“Experiment 1 Statistical Significance Test”

“Experiment 7 Statistical Significance Test”

Comparison of WMT Test Set Results Across All Experiments

A comparison of test set BLEU scores and chrF++ scores (Fig. 26) showed that best performing models were trained on synthetic parallel corpora during experiment1. It noted that the chrF++ evaluation metric reported less a smaller ratio of difference in performance between NMT trained on synthetic data vs the NMT trained on original data (Fig. 27).

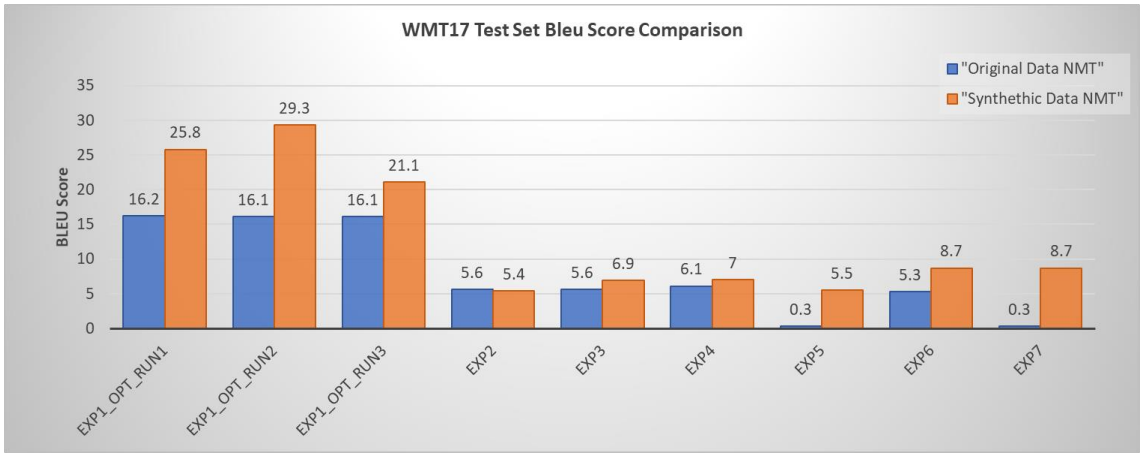


Figure 22 WMT2017 Test Set BLEU Score Comparison

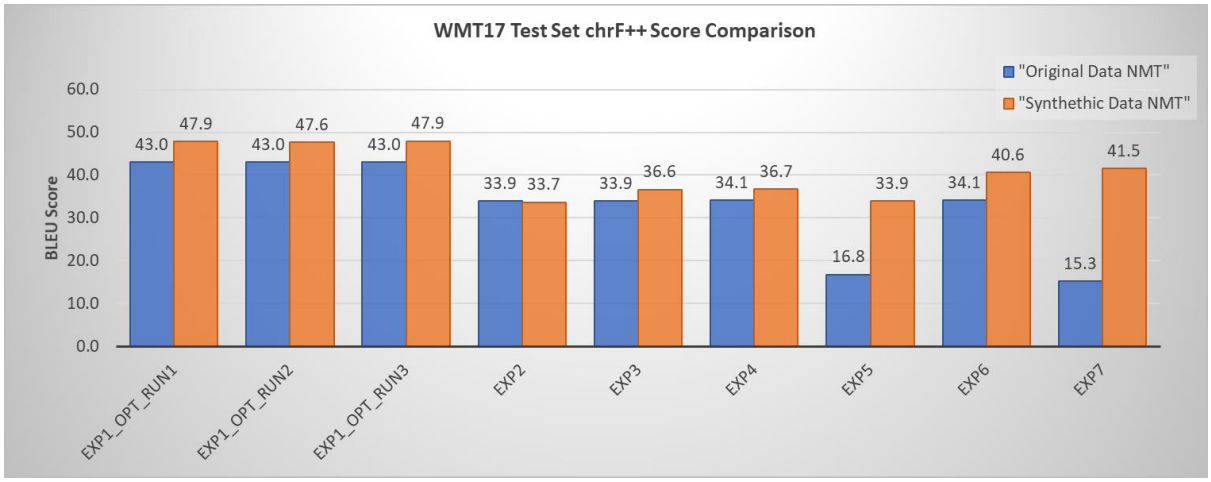
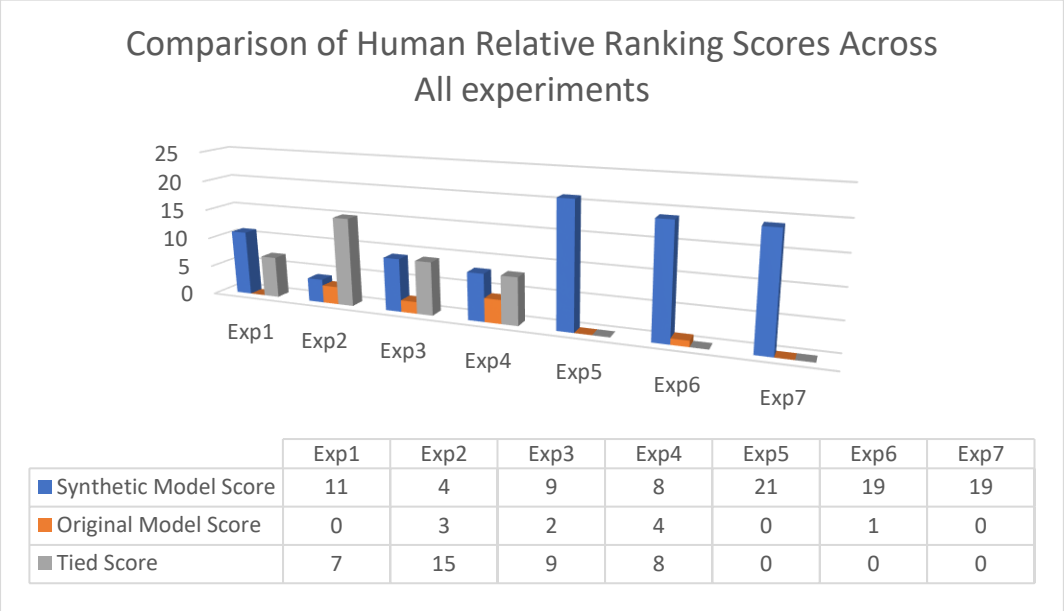


Figure 23: WMT2017 Test Set chrF++ Score Comparison

For the full set of comparisons made across all experiment refer to appendix section "Comparison of Test Results Across Experiments "

Human Evaluation Results

The relative ranking assessment performed by a native German speaker found the NMT model trained on synthetic data consistently performed better than the NMT trained on original data (Fig. 24). The human judge noted many ties and stated that the magnitude of difference performance of both models seemed consistently marginal for experiments 1 to experiment 4. For experiment 5 to experiment 7 the difference in performance was noted to be much more significant.



	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7
Synthetic Model Score	11	4	9	8	21	19	19
Original Model Score	0	3	2	4	0	1	0
Tied Score	7	15	9	8	0	0	0

Figure 24: Comparison of human evaluation results across all experiments

Note: To view the relative ranking sentence test sets and the human judges scoring of each sentence refer to appendix sections "Experiment # Human Evaluation Test Results"

Discussion

High BLEU scores during training

The unexpectedly high BLEU scores achieved during training for both original and synthetically trained is a sign that the model is overfitting. This overfitting is evident when test sets from non-withheld corpus data are used to evaluate model performance and a significant drop in BLEU score results are observed for most tests. See appendix sections “Experiment # NMT Training Charts”.

Synthetic Trained NMT performed better than NMT trained on original data.

Statistical significance tests on BLEU and chrF++ scores indicate that NMT models trained on synthetic data performed better than NMT models trained on the original data. See appendix for “Statistical Significance Test Applied to Combined Experimental Results”. On initial observation of these results, it was surprising to find that synthetically trained models performed better.

Transparency issue with neural networks

Lack of transparency of neural networks makes it difficult to interpret the NMT internally themselves as to why one model performed better than the other. Due to the notorious difficulty in deciphering the internal workings of neural networks, it was decided to concentrate the analysis on all data associated with experiment to determine why the NMT trained on synthetic data performed better the NMT trained on original data. This data analysis encompasses input data used during experiment training phase, inference inputs and inference outputs during experiment testing phase.

Visual inspection of Training Data

To analyse the training phase of NMTs a human visual inspection was performed on the inputs (see appendix section: “Samples of Sentences Used to Train Models”). The human visual inspection of the training data observed that translations annotated by the source models are of comparable quality to that of the original reference translation.

Human Evaluation of NMT Inference of Test Sets

During experimentation, a relative ranking assessment of the inference data was evaluated by a native German speaker judge (see appendix sections “Experiment # Human Evaluation Test Results”). The aggregated results of the experiments deemed the synthetically trained model superior to that of the original model. For all experiment which did not use the UFAL data a high number of ties were tallied. While the NMT trained on synthetic data did consistently perform better than the NMT trained on original data the high number of ties would indicate that the magnitude of difference in performance was not large (see appendix section: “Comparison of Human Evaluation Relative Ranking scores across experiments”).

Sentence length analysis

To assess why the NMT trained on synthetic data obtained a better BLEU score during inference, a comparison of generated sentence length was performed on both NMT outputs against that of the reference sentence. The basis of this analysis is that BLEU is a n-gram based metric. With BLEU a n-gram score is generated based on the translated sentences containing tokens the same tokens as the reference sentence and if translated sentences are too long, they are penalised by the brevity penalty. Thus, translated sentences which match the same length of reference sentence are likely to score higher on BLEU scores as they will not have missed a chance to match a token if not too short or penalised by brevity if too long. However, sentences are often not the exact same length when translated between languages.

Therefore, if two NMT models generated translation of the same quality (i.e same Fluency and meaning) but one NMT was more likely to translate sentences of the same length as the reference sentence it is more likely to score higher on a n-gram based metric.

Bias in n-gram based metrics due to matching sentence Length

The analysis of the test set translations found that the synthetic data trained NMT outputted more sentences the exact same length as the associated reference sentences on tests. To assess if there is bias in the synthetic training data to generate sentences of the same length, the length of the inputted source English sentences was compared against the length of the target German sentence. The analysis found that the input synthetic German sentence length matched the length of the source English sentence more often than it occurred for the original sentence length. Logically the length of the reference sentence is related to the source sentence, but it is not dependant on it. Thus, the correlation observed where the synthetic data sentence length matched the source length may be coincidental with regards to the synthetic model being more likely to output translated sentences matching in length to the reference sentence.

However, if the correlation between matching sentence length between source and target training data resulted in NMT models being probabilistically more likely to output translations the same length as the reference translation then this would offer an explanation as to why the synthetically trained models had bias to perform better on the BLEU, METEOR and chrF++ n-gram metrics.

If length bias exists, then TER based metrics may be also subject to bias as deletion and insertion edit actions are related to difference in sentence length.

Possible alternative evaluation metrics which may not be subject to such bias length could be automatic evaluation metrics based on contextual word embedding. Such automatics metrics attempt scores translations based on closeness of meaning between translated words. Such a metric may not be as subject boas to due to sentence length.

Potential Bias in Human Evaluation Test Results Due to Matching Sentence Length

The relative ranking human evaluation metric used a reference translation when comparing NMT output translations. As humans are notoriously subjected to bias, it is plausible that if two translations were of the same actual quality, the model which matched the reference translation may have been regarded as slightly superior due to the higher similarity in appearance. This bias may be contributed to the perceived better performance of the NMT trained on synthetic data within the human evaluation experiments.

Poor quality pairings when original parallel corpus was gathered.

When performing the human evaluation on results some obscure reference sentence which did not seem to fit with the source sentence were observed. A plausible cause for why the synthetic data trained models perform slightly better than that of the original data models is due to the poor pairing sentences for within the parallel corpus. The source model may have translated enough sentences that were a better representation of the source text than that matched during the initial creation of the parallel corpora. Unsupervised parallel corpora gathered by methods such as web crawling are particularly vulnerable to mismatch in pairing. Due to the sheer quantity of parallel sentences required to train state of the art NMT models, web crawled corpora are commonly used. For humans to annotate millions of high-quality sentence translations across various languages for all domains of knowledge which may use translation is impractical hence the use of unsupervised methods to generate corpora.

Automated generation and preparation of high quality parallel corpora is a difficult problem and still an active area of research (Koehn et al., 2020). Unsupervised generation of parallel corpora rely on text alignment algorithms and filtering of noisy unsuitable sentences. Due to the wide variety of website crawled for large datasets and the variability in translation quality of websites, web crawling cab gather large quantity of noisy non sensical sentences. Use of too much noisy data is known to decreased performance in NMT models (Joulin, Grave, Bojanowski, & Mikolov, 2016).

The noisy nature of web crawled data may have been the contributing factor in providing training data which resulted NMT outputs which did match the length of the reference sentence as often as the NMT model trained synthetic data.

Quality of samples generated by source model

Recent state of the art models NMT output sentences are approaching similar quality to that human translations (Popel et al., 2020). The success of the active learning method applied for this project experiments are entirely reliant on performance of the source model used to translate sentences for the synthetic data training set. Active learning has been applied extensively to statistical machine translation (Haffari, Roy, & Sarkar, 2009). With the recent staggering improvement in machine translation due to adoption of neural network models and the availability of hardware sources for applying such methods, active learning has become an even more viable method for augmenting parallel corpora than previous research has previously indicated. The translation outputs of the publicly available ONMT model for German to English are of good adequacy but are not at the same adequacy and fluency as a Human translator.

For experiment 2 to 7 the quality of the Facebooks FAIRs model placed higher than a human annotator on the WMT2019 document level news translation task. The adequacy of this model is exceptional, but the fluency of human translations is still considered better than current state of the art NMT solutions (Popel et al., 2020).

As NMT are not at the same level of fluency as human translation a great level of scrutiny was applied when assessing experimental results when it was observed that the NMT model trained on synthetic data performed better than the NMT trained on the original data.

Effect of dataset size

Experiment 5 to 7 used the same UFAL medical corpus but size of the corpus increased in size from experiment 5 to 7. Where:

- Experiment 5 sample size = 100K
- Experiment 6 sample size = 3.5M
- Experiment 7 sample size = 11.5M

By only altering the sample size it was used to examine the effect dataset size has on the NMT trained on synthetic data and the NMT trained on the original data. NMT models trained on the original data converged at significantly lower BLEU score of 10 on training validation data for experiment 5 and 7. Experiment 6 had a BLUE of 40 during training and had not fully converged when stopped after training for 50K updates.

Experiment 5 and 7 BLUE score did not score higher than 0.5. The withheld datasets did perform better but results were still poor with BLEU scores 13.6 and 6.2.

These results were surprising to see for when datasets increase in size a model is expected to generalize better and score higher on the unseen test data. Increase in performance with increasing size is typically seen when the extra training sample were relevant in relation to the unseen test data.

The NMT trained on synthetic data produced low BLUE scores but interestingly was still significantly better than NMT models trained on original data. In future work an analysis will be valuable to examine the variance of the synthetic data compared to the original data to confirm if the original data has a higher variance in tokens.

NMTs trained on synthetic data marginally improved on performance between sample size 100K and 3.5M however approximately the same score on WMT test sets was achieved on experiment with data size 3.5M and 11.5M. Experiment 5 and 7 abysmal performance when training on the original data may be due to overly noisy training data.

The UFAL medical corpus a has 84M/94M German/English tokens associated with medical domain and 716M/817M German/English tokens associated with more general domains. For experiment 5 to 7 a specified sample size of random sentences was taken from the UFAL corpus. This approach may have led to too a great a variety of tokens within the training data and not allow the model to train successfully due to the noisiness of the training data. As random samples were taken for all experiments its possible experiment 6 trained on a less noisy dataset compared relative to experiment 5 and 7.

The objective of evaluating the effectiveness of active learning model extraction on large scale data was not properly realised during experiments 5 to 7. The NMT trained on the original data was not representative of a well performing NMT that would be worthwhile for an adversary to attack. However, as the BLEU score and chrF++ score increased for the synthetic data between experiment 5 and 6, it seems to indicate that active learning model extraction could scale fine as size of the synthetic dataset size increases.

Effect of Noisy datasets

For all experiments word frequency threshold of 15 was used when performing byte pair encoding. This essentially means that a combination of characters must appear at least 15 times for it to be encoded. As a more varied dataset was used and sample sizes were increased, word frequency threshold should have been increased.

Teams within WMT translation shared task often implement extensive filtering of data sets. Noisy datasets can reduce the performance of models. Facebook FAIRs WMT 2019 submission used language identification filtering (Joulin et al., 2016) to remove sentence which not match the source and target languages of the parallel corpora. This method identified mis-match pairs and sentence which contain non sensical tokens thus reducing the noise within the training data. Noisy original data with poor pairings between the source and target language are a likely cause for the NMT trained on original data to performed poorly. Experiments 5 to 7 showed the viability of active learning as a data augmentation for machine translation as the synthetic data significantly outperformed the original data constructed from web crawled data.

Difference Between State-of-the-Art results

Hardware resources available for training models as part of this project were significantly less than that of the previous state of the art implementations. For instance, the publicly available source model used to annotate the synthetic data was trained on 128 Volta GPUs and dataset size of 27.5 million sentences. For all experiments in this project 1 V100 GPU was used. The v100 is a respectable GPU but

availability of 1 GPU was a significant disadvantage compared state of art distributed training implementations (Ott, Edunov, Grangier, & Auli, 2018).

For experiment 1 models were trained for 16 hours on one V100 GPU with 100,000 random training samples taken from the WMT2017 news translation task. Experiment1 optimizer run one synthetic data NMT achieved a respectable BLEU score of 29.3 on the WMT2017 news test set. To put this BLEU score into perspective the LMUs NMT submission in the WMT2017 news translation task achieved the best human evaluation score and had a BLEU score of 27.1 on the same WMT2017 news test set (Huck, Braune, & Fraser, 2017). LMUs NMT submissions implemented NMT systems used gated recurrent units, thus its architecture is significantly different to the transformer model used by this project. Hence these project models were able to achieve a higher BLEU score despite the small training dataset.

Transformer based models such as team Tohoku-AIP-NTT submissions to WMT2020 news translation shared task scored a BLEU score of 37.5 on the WMT 2020 news translation test set (Kiyono, Ito, Konno, Morishita, & Suzuki, 2020). This projects highest BLEU score on the WMT 2020 news translation test set was achieved in experiments 1's optimizer run1 ssynthetic data NMT with a BLEU score of 20.9. The WMT shared task competitor submission trained their model on approximately 44M sentence pairs. As this projects experiment 1's model was only trained on a sample size of 100,000 sentences the score o BLEU f 20.9 by the transformer-based model could still be considered impressive.

According to the tensor board training charts, experiment 1's model had still not converged upon time of stopping. If experiment 1 models training ran for longer this model and a larger sample size from the good quality data was used, the experiments models may have achieved a BLEU score closer to previous state of the art implementations of transformer-based models. However, as the projects primary aim was to evaluate the effectiveness of active learning applied to training of NMT models, all models in experiment 1 were prematurely stopped at 50,000 iterations, which was approximately 16 hours. As this experiment is measuring the difference between model performance as opposed to an absolute value this approach was acceptable with regards to this project's objectives.

It would have been preferable to obtain the best possible performance for each model. However, to perform an effective analysis many runs are required to be performed to gather enough evidence to prove results are statistically significant. Due to resource constraints the model was stopped premature so that a higher quantity of different runs could be performed with the limited available resources.

For experiment 2 to 7 the choice of datasets differed from datasets used by previous state-of-the-art implementations. Between experiments 2 to 7 the highest BLEU score achieved was 12.8. The choice of datasets is the most probable cause of poor BLEU score results, relative to experiment 1. The training charts of experiment 2 to 4 show that both synthetic data and original data models have converged on training data with BLEU scores greater than 45 on the validation data. When experiment 2, 3 and 4 models are tested on the withheld training data a range of BLEU scores ranging from 33 to 45.6 are achieved. When experiment 2, 3 and 4 models are tested on WMT2020 test set a range of BLEU scores ranging from 5.2 to 6 are achieved. High scores on withheld test data but poor scores on the WMT test sets indicate that experiment 2 to 7 training data samples were not representative of the WMT test sets and thus a poor choice for the WMT test sets. Resource constraints are constant in practice thus it is still worthwhile performing an analysis of on the difference between model performance which are trained in sub optimal state of the art conditions.

Effect of sentence length.

Both NMT models quality of translation diminished as source English sentence complexity (i.e length) increased. Sentence level analysis showed a correlation between longer sentence length and lower sentence level BLEU scores. As sentence level BLEU score are an unreliable metric, a human evaluation confirmed this phenomenon.

Difference between evaluation metrics

According to statistical significance tests both chrF++ and BLEU metric indicate NMT models trained on synthetic data performed better than the original models. Interestingly the chrF++ metric describes the magnitude of difference in performance between the synthetic and original data as less than that reported by the BLEU metric (Table 6).

Statistical Analysis	**BLEU Score Original / Synthetic mean**	**chrF++ Score Original / Synthetic mean**	**BLEU Score Ratio of Original to Synthetic mean**	**chrF++ Score Ratio of Original to Synthetic mean**
All scores	10.63 / 14.48	34.17 / 41.49	0.73	0.82
All scores except UFAL	14.05 / 16.74	41.18 / 44.12	0.84	0.93

Table 6:Difference in results ratio between chrF++ and BLEU metric

As these metrics were applied to the exact same NMT outputs, the different ratio in result illustrates the merit in using multiple metrics to evaluate the results. Currently it is common for the MT community to publish results only stating observed BLEU scores. The choice of consistently publishing BLEU scores makes sense from the perspective of making it possible to compare results across publications. However, the MT community should choose multiple metrics to be consistently applied when publishing results. Despite both BLEU and chrF++ being n-gram based evaluation metrics they produced different ratio of performance. A bigger discrepancy between results is even more likely to be observed between metrics based on different principles of analysis such CharcteTER (Wang, Peter, Rosendahl, & Ney, 2016) a TER based. When the computational cost of applying TER and n-gram based methods is compared with the computational cost of training the NMT being evaluated, applying additional metrics is computationally trivial.

Due to flaws in current evaluation metrics the use of an ensemble of metrics to evaluate results, may capture faults or benefits which may not otherwise be observed. Neither n-gram nor TER take into account context when evaluating translations. They only consider quantity of correctly/incorrectly translated words against reference sentences. However, many synonyms can be used to form equivalent translations, but are penalised by n-gram and TER metrics.

Word embedding metrics such as YiSi-1 (Lo, 2020) are more recent metrics which aim to capture the contextual significance of translation. Word embedding metrics currently achieve the highest correlation with human evaluation and thus deemed the highest performing evaluation metrics (Mathur, Wei, et al., 2020). It would be more challenging to apply learnt word embedding evaluation models consistently across different experimental results than n-gram and TER methods. However, an attempt to apply such methods as part of an ensemble of metrics would allow MT evaluation metrics to progress without comprising the current standard in results reporting.

Hazards of unsupervised learning

It is a known problem that NMT may produce unfortune mistranslation which a human would never produce. An example of this appeared during human evaluation of results where it was noted the sentence “hollande in Mali” was translated to “Holocaust in Mali” (Fig. 25).

Source_Sentence	Reference_Sentence	Original Hypothesis	Synthethic Hypothesis
hollande in mali	hollande in mali	hollande in mali	Holocaust in Mali

Figure 25: Inappropriate translation by synthetic model. Source: Experiment 4 UFAL corpus withheld corpus test set

According to a google search “hollande in Mali” refers to the previous French president Francois Hollandes visit to Mali (Fig 26).

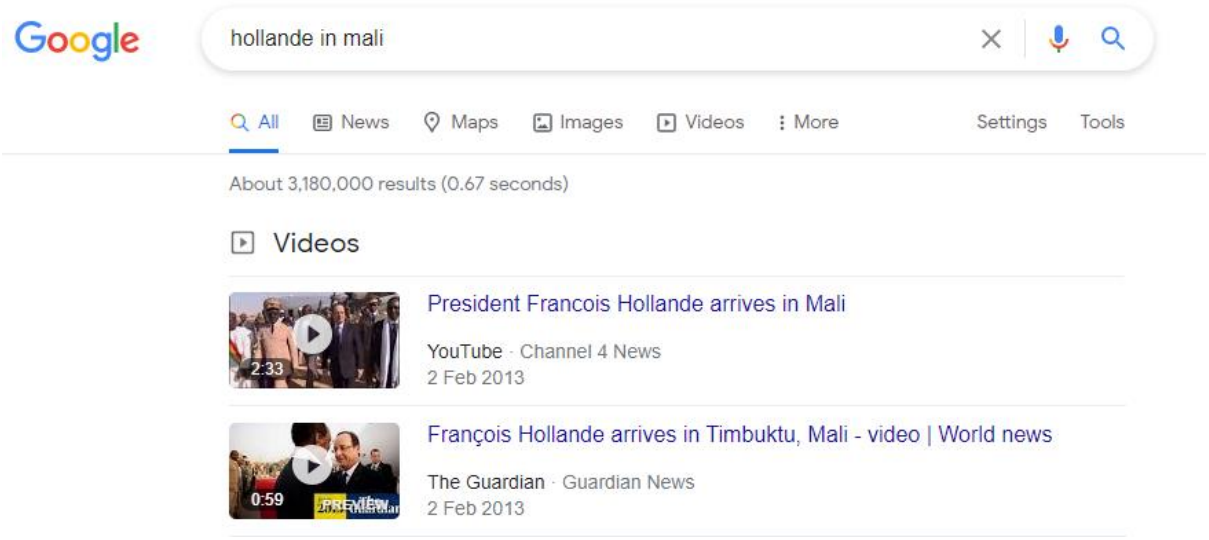


Figure 26:Google search result of “hollande in Mali"

The human evaluation test also noted inappropriate translation were produced by the originally trained model.

Source_Sentence	Reference_Sentence	Original_Hypothesis	Synthethic_Hypothesis	Description of Original_Hypothesis Mistake
The officers shot Clark seven times as he approached them.	Die Polizisten feuerten sieben Schüsse auf Clark ab, als er sich ihnen näherte.	er hat die sieben polizisten angegriffen.	die offiziere erschossen clark siebenmal, als er sich ihnen näherte.	Mistakingly reversed attackers and the attacked
When she asked United Healthcare why she was receiving a mountain of letters, she was told it was a coding issue, she said.	Als sie United Healthcare fragte, warum sie einen Berg von Briefen erhalten habe, wurde ihr gesagt, dass es sich um ein Codierungsproblem handele, sagte sie.	sie hat mich gefragt, ob es eine gute idee ist, sie zu töten.	als sie die gemeinsame gesundheitsversorgung fragte, warum sie einen berg briefe erhielt, wurde ihr gesagt, es sei ein codierungsthema, sagte sie.	Mistakingly translated as "she asked me if killing her was a good idea"

Figure 27: Inappropriate translations by original model. Source: Experiment 7 WMT20 test set

It would most certainly be a PR disaster if a model available to the public translated a president’s visit to a country as a holocaust for that country. Such an examples show a real danger in applying unsupervised learning methods such as active learning when training NMT models. When applying unsupervised learning methods such as active learning to NMT development, it is imperative that precautions are taken to prevent such events from occurring. Rudimentary methods such as rule-based filtering can remove such obvious mistranslation but fall short when handling contextually appropriate uses of words such as “holocaust “. This is a challenging problem, with many active academic shared tasks dedicated to problem of hate speech detection (Mandl, Modha, Kumar M, & Chakravarthi, 2020; Mubarak, Darwish, Magdy, Elsayed, & Al-Khalifa, 2020). Approach such text

classification ML models which leverage multilingual word embedding are actively used by social media companies such as Facebook to help classify content which is in breach of community policies. During review of datasets, it was noted that parallel corpuses created from web crawling contained much offensive content. As these models were not being released to the public, they were fine to still use, however unsupervised learning methods for NMT trained on such corpuses are likely to produce offensive outputs due to these offensive inputs.

Results in comparison to related work.

Niu and colleagues (2018) work used bidirectional translation to augment synthetic parallel corpora for training a NMT model. In contrast to this projects experimental results, forward directional translations (i.e source to target) had no significant increase in model performance. The stated likely cause for this was poor selection of source data (Niu et al., 2018). Niu and colleagues (2018) work combined synthetic and original data. If the extra synthetic data only added representation already covered by the existing corpora it would be expected to see such negligible improvement in performance. In contrast to this projects work, no original data is used by the NMT trained on synthetic data thus the impact of the synthetic data makes a dramatic impact on the performance of the model. Niu and colleagues (2018) trained models based on recurrent neural network architecture. This project used a transformer model which has been noted to perform better than recurrent neural networks in machine translation tasks. This may be why better performance by synthetic data was observed within this project work.

Niu and colleagues (2018) reported issues with data selection. Similarly, experiment 5 to 7 of this project was impacted also by poor data selection. Interesting though this poor selection of data had a significantly more negative effect on NMT trained solely on the original parallel corpora.

Edunov and colleagues (2018) used back translation to augment synthetic data for training NMT models. Similarly, to this project English to German NMT models were trained and tested on the WMT 2017 test set. NMT models scored 32.6 (Fig. 28, Edunov et al., 2018) and this project work achieved 29.3 on the WMT 2017 test set during experiment 1 optimizer run2.

	news2013	news2014	news2015	news2016	news2017	Average
bitext	27.84	30.88	31.82	34.98	29.46	31.00
+ beam	27.82	32.33	32.20	35.43	31.11	31.78
+ greedy	27.67	32.55	32.57	35.74	31.25	31.96
+ top10	28.25	33.94	34.00	36.45	32.08	32.94
+ sampling	28.81	34.46	34.87	37.08	32.35	33.51
+ beam+noise	29.28	33.53	33.79	37.89	32.66	33.43

Figure 28: Results achieved by (Edunov et al., 2018) on various test sets of WMT English-German when adding 24M synthetic sentence pairs obtained by various generation methods to a 5.2M sentence-pair bitext. (Edunov et al., 2018)

Edunov and colleagues (2018) used 5.2M original parallel sentences and combined it with 24M synthetic parallel sentences. In contrast, this project used a sample size of 100K synthetic sentences to achieve a BLUE score which was only 3 lowers. Comparing experiment 1 results (Edunov et al., 2018) confirms that NMT model trained on synthetic data within this project genuinely performed well. However, the closeness in scores despite the masive disprenency in sampel size could be due largely due to fortunate sample data selection relative to the tests sets used during evaluation.

Sample Selection

The active learning model extraction attack implemented during this project is simple to employ. For the project, all samples sentences were chosen randomly from a large dataset. More advanced implementations use heuristic methods for finding more appropriate sample sentences for training. If an adversary is restricted to obtain a high functioning model within a specified query budget it would be prudent to employ some form of heuristic to find high performing source sentences as opposed to the random selection used. During visual inspection of sentences, a correlation between long complex source sentences and poor target translations was observed. A simple heuristic would be to search sentences based on length so that these more challenging sentences would not appear within training corpora.

Total token entropy is a heuristic method originally applied to active learning with sequence models (Settles & Craven, 2008). Within uncertainty, sampling entropy can be used as a measure of informativeness as non-sparse populations distribution have a propensity for low entropy.

Zha and colleagues (2020) applied Total Token Entropy as a selection metric for sample training sentences and found sample selected via this method resulted in better NMT model performance than that compared to NMT models trained on randomly selected samples.

Experiment 2 to experiment 4 according to the training chart produced respectable test scores on withheld test data however performed poorly on WMT test sets. Due to the small sample size of 100K, both NMT models did not generalise well to the test sets. If a better sample selection method was used a higher BLEU score may have been achieved on the unseen test sets.

Conclusion

The aim of this project work is to evaluate how effective a synthetic parallel corpus created via active learning model extraction attack are at training a substitute neural machine translation model. Analysis of experimental results confirmed that using solely a synthetic dataset generated from active learning with random sample selection can be used to train functioning NMT models.

A transformer NMT trained solely on synthetic data from active learning achieved a respectable BLEU score of 29.3 on the WMT2017 news test set. To put this BLEU score into perspective the LMUs recurrent neural network NMT submission in the WMT2017 news translation task achieved the best human evaluation score and had a BLEU score of 27.1 on the same WMT2017 news test set. Surprisingly, the statistical analysis test of experimental results indicated that NMT models trained solely on synthetic data performed better than NMTs trained on the original data.

As NMT models have not surpassed human quality translation the experimental result was surprising. The analysis of the large datasets used during training was found to contain web crawled parallel corpora that infrequently had poor translation pairing. Noise within the original training data allowed the NMT model trained on synthetically generated data to better match reference translation.

The objective of evaluation of the effectiveness of active learning model extraction on large scale data and was not properly realised during experiments due to poor sample sentence selection. While using the larger datasets the NMT trained on the original data was not representative of a well-performing NMT that would be worthwhile for an adversary to attack. Thus, a comparison of the difference in performance between original and synthetic data would not be conducive to the assessment of the effectiveness of model extraction attacks.

However, as the NMT model trained on synthetic data performed better than web crawled data, it showed that active learning is an effective unsupervised method of data augmentation. High variability of both synthetic and original NMT model's performance is observed across the experiment. High variability of performance appeared to be caused by the random sample selection method. The greatest challenge in applying active learning model extraction attacks is the selection of high-quality sample sentences for translation.

Comprehensive analyses of results from experiments conclude that simple active learning model extraction attacks are a genuine threat to the monetization of NMT models.

Code used as part of this experiment is publicly available at the following link: https://github.com/Frankkelly1990/Actively_Fake_It_Until_You_Make_it_in_NMT.git

Future work

Apply query selection technique to improve quality of synthetic Dataset

Future experiments using active learning model extraction attacks should be done using a heuristic method for the selection of sample sentences. The sample selection heuristic should be used to generate multiple datasets for training multiple NMT models. The use of multiple generated datasets should be performed to assess if high performing NMT models can more consistently be produced than using datasets generated from randomly selected sentences.

Compare Synthetic Data Against Non-Noisy Translated Dataset

The work of this project showed that synthetically generated datasets performed better than the original data. Some of the original data in question were gathered via web crawling. While large datasets used by state-of-the-art NMT models contain web crawled data, it can negatively affect NMT performance due to noisy data. Results from this project provide a reason to suspect that the web crawl samples may have altered the difference in performance between the NMT model trained on synthetic data and the NMT model trained on original data.

To rule out this possibility, an experiment should be performed which use a fully human-annotated parallel corpus. Alternatively, a parallel corpus that is known to be high quality could be used as it would likely still be an improvement on some of the parallel corpora used for this project.

An adversary is more likely to perform a model extraction attack on a model which is using high-quality data that is hard to obtain. Thus, an evaluation of active learning model extraction attack applied using datasets that did not to non-web crawled data would be more prudent.

Combine targeted web crawling with active learning to generate data sets.

Results of this project indicate that parallel corpora generated from using an NMT model to translate web crawled data created better models than parallel corpora generated from using web crawling and text alignment algorithms.

An interesting experiment would be to perform web crawling within specified knowledge domains website and generate parallel corpora from the web pages using a high-performance NMT model to generate the target language pair. The quality of generated corpora should then be evaluated on an NMT model.

The aim of this evaluation is to determine if a dataset obtained via such an unsupervised method could contribute to the performance of state-of-the-art models.

Perform more extensive human evaluation of results.

This project had access to one German native speaker for performing the human evaluation. Due to the limits in human resources, the human evaluation was not as extensive as desired. In future work, a larger sample size for the human evaluation would be an improvement on the work presented within this project.

Use of Hume to perform semantic analysis of experiment so that comparisons could be made to confirm if NMT trained on synthetic data generated the same error as the NMT trained on original data.

Appendix

Experiment 1 – Results

Experiment 1 NMT Training Charts

Figures below illustrate experiment 1 validation BLEU score which calculated during training of the model.

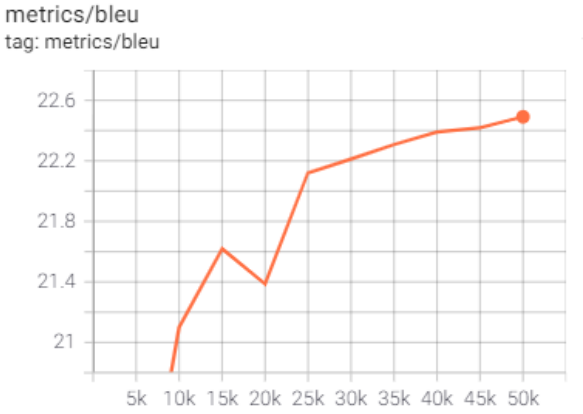


Figure 30 Experiment 1 Synthetic Dataset NMT validation BLEU score during training optimizer run 1

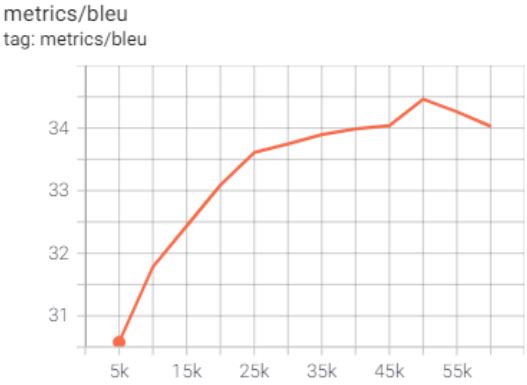


Figure 29 Experiment 1 Original Dataset NMT validation BLEU score during training

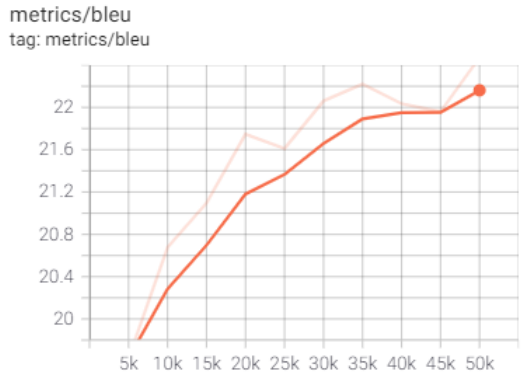


Figure 32 Experiment 1 Synthetic Dataset NMT validation BLEU score during training optimizer run 2

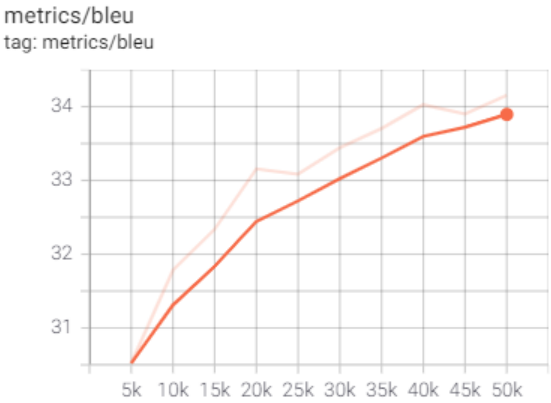


Figure 31 Experiment 1 Original Dataset NMT validation BLEU score during training optimizer run 2

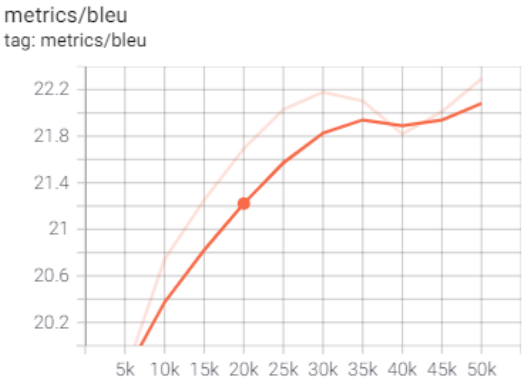


Figure 34: Experiment 1 Synthetic Dataset NMT validation BLEU score during training optimizer run 3

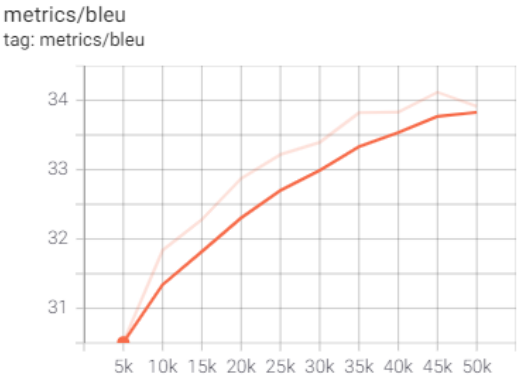


Figure 33: Experiment 1 Original Dataset NMT validation BLEU score during training optimizer run 3

Experiment 1 BLEU Score Test Results

Table below illustrates experiment 1 BLEU scores which are calculated on translation of sentences from test sets.

Experiment	Corpus	Optimizer Run	Test Data Set	Original Data BLEU Score	Synthetic Data BLEU Score	Difference in score (Original data score - Synthetic data Score)
experiment1	ONMT	optimizer_run1	Withheld Corpus Data	5.71	8.05	-2.34
experiment1	ONMT	optimizer_run2	Withheld Corpus Data	5.78	21.2	-15.42
experiment1	ONMT	optimizer_run3	Withheld Corpus Data	5.69	29.5	-23.81
experiment1	ONMT	optimizer_run1	wmt17	16.2	25.8	-9.6
experiment1	ONMT	optimizer_run1	wmt18	21.4	14.5	6.9
experiment1	ONMT	optimizer_run1	wmt19	17.9	8.04	9.86
experiment1	ONMT	optimizer_run1	wmt20	13.5	20.9	-7.4
experiment1	ONMT	optimizer_run2	wmt17	16.1	29.3	-13.2
experiment1	ONMT	optimizer_run2	wmt18	21.4	25.9	-4.5
experiment1	ONMT	optimizer_run2	wmt19	18.1	14.8	3.3
experiment1	ONMT	optimizer_run2	wmt20	13.5	8.06	5.44
experiment1	ONMT	optimizer_run3	wmt17	16.1	21.1	-5
experiment1	ONMT	optimizer_run3	wmt18	21.2	29.1	-7.9
experiment1	ONMT	optimizer_run3	wmt19	18.2	25.7	-7.5
experiment1	ONMT	optimizer_run3	wmt20	13.6	14.3	-0.7

Table 7: Experiment 1 BLEU Score results from Test Data Set Translations

Average difference between original data BLEU score and synthetic data BLEU Score: -4.79

Standard deviation between original data BLEU score and synthetic data BLEU Score: 8.7

Experiment 1 Character and word F Score results

Table below illustrates experiment 1 character and word F Score results which are calculated on translation of sentences from test sets.

Experiment	Corpus	Optimizer Run	Test Data Type	Original Data c6+w2-F2 Score	Synthetic Data c6+w2-F2 Score	Difference in score (Original data score - Synthetic data Score)
experiment1	ONMT	optimizer_run1	wmt17	43.0406	47.8923	-4.8517
experiment1	ONMT	optimizer_run1	wmt18	48.0218	54.6563	-6.6345
experiment1	ONMT	optimizer_run1	wmt19	44.8533	51.7829	-6.9296
experiment1	ONMT	optimizer_run1	wmt20	41.4373	39.4376	1.9997
experiment1	ONMT	optimizer_run1	Withheld Corpus Data	36.6609	40.4791	-3.8182
experiment1	ONMT	optimizer_run2	wmt17	42.9916	47.6496	-4.658
experiment1	ONMT	optimizer_run2	wmt18	48.3396	54.532	-6.1924
experiment1	ONMT	optimizer_run2	wmt19	45.0607	51.6609	-6.6002
experiment1	ONMT	optimizer_run2	wmt20	41.5094	39.4927	2.0167
experiment1	ONMT	optimizer_run2	Withheld Corpus Data	36.77	40.408	-3.638
experiment1	ONMT	optimizer_run3	wmt17	43.0374	47.9223	-4.8849
experiment1	ONMT	optimizer_run3	wmt18	48.0097	54.4877	-6.478
experiment1	ONMT	optimizer_run3	wmt19	45.2402	51.4649	-6.2247
experiment1	ONMT	optimizer_run3	wmt20	41.2061	38.8661	2.34
experiment1	ONMT	optimizer_run3	Withheld Corpus Data	36.7098	40.4067	-3.6969

Table 8: Experiment 1 character and word F Score results from test data set translations

Average difference between original data and synthetic data F scores: -3.88

Standard deviation between original data and synthetic data F scores: 3.19

Experiment 1 Statistical Significance Test

Baseline represents model trained on original data.

System 1 represents model trained on synthetic data.

Withheld Corpus Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	12.1 (0.1/0.2/-)	19.0 (0.0/0.1/-)	72.2 (0.1/0.1/-)	96.9 (0.1/0.5/-)
system 1	17.2 (0.1/0.1/0.0001)	22.0 (0.1/0.0/0.0001)	66.4 (0.1/0.0/0.0001)	96.4 (0.1/0.3/0.0001)

WMT17 Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	16.5 (0.3/0.0/-)	22.0 (0.1/0.1/-)	66.0 (0.3/0.1/-)	97.6 (0.3/0.2/-)
system 1	21.5 (0.3/0.2/0.0001)	25.0 (0.2/0.1/0.0001)	60.3 (0.4/0.1/0.0001)	98.9 (0.3/0.4/0.0001)

WMT18 Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	21.8 (0.3/0.1/-)	25.3 (0.1/0.1/-)	58.4 (0.3/0.3/-)	96.5 (0.3/0.1/-)
system 1	29.7 (0.3/0.2/0.0001)	29.4 (0.2/0.1/0.0001)	50.3 (0.3/0.1/0.0001)	97.4 (0.3/0.3/0.0001)

WMT19 Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	18.9 (0.3/0.1/-)	23.4 (0.2/0.2/-)	61.0 (0.4/0.3/-)	91.7 (0.4/0.3/-)
system 1	26.8 (0.4/0.1/0.0001)	27.6 (0.2/0.1/0.0001)	53.2 (0.4/0.1/0.0001)	94.0 (0.4/0.5/0.0001)

WMT20 Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	14.5 (0.3/0.1/-)	20.9 (0.2/0.0/-)	65.7 (0.4/0.1/-)	86.1 (0.5/0.5/-)
system 1	15.8 (0.4/0.3/0.0001)	20.4 (0.3/0.2/0.0001)	66.7 (0.5/0.3/0.0001)	72.4 (1.0/0.8/0.0001)

Experiment 1 Human Evaluation Test Results

Results of relative ranking human evaluation performed by native German speaker on experiment1 NMT model outputs of WMT17 test set.

Model A = NMT trained on Synthetic data.

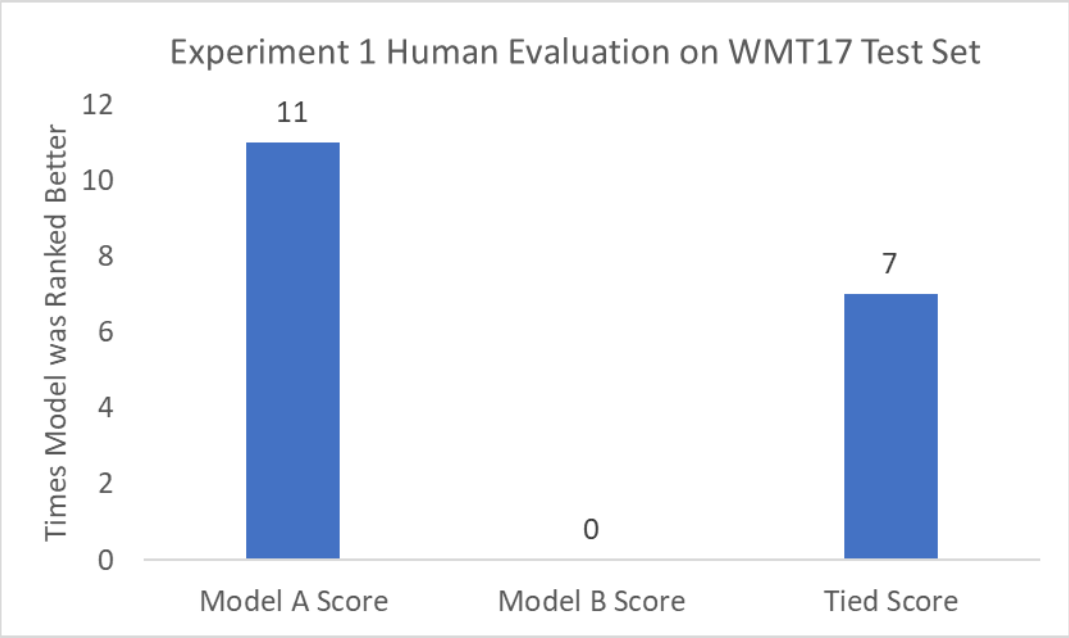
Model B = NMT trained on Original data.

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
but they are not sufficient .	aber sie reichen nicht aus .	aber sie reichen nicht aus.	sie sind aber nicht ausreichend.	same
but will it work ?	aber wird es funktionieren ?	aber wird es funktionieren?	aber wird es arbeiten?	a
a lot has changed since 2005 .	seit 2005 hat sich viel verändert .	Seit 2005 hat sich viel verändert.	Vieles hat sich seit 2005 verändert.	same
paris - who would have thought it ?	paris - wer hätte das gedacht ?	Paris - wer hätte das gedacht?	paris - wer hätte das gedacht?	same
one could now imagine much more clearly what might happen if a nuclear bomb exploded .	man konnte sich jetzt viel deutlicher vorstellen , was passieren könnte , wenn eine atombombe explodierte .	Man konnte sich jetzt viel deutlicher vorstellen, was passieren könnte, wenn eine Atombombe explodierte.	Man kann sich jetzt viel klarer vorstellen, was möglicherweise ein Atombomben explodierte.	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
but at least we know that we have asked the right questions .	aber zumindest wissen wir , dass wir die richtigen fragen gestellt haben .	aber zumindest wissen wir, dass wir die richtigen Fragen gestellt haben.	Aber zumindest wissen wir, dass wir die richtigen Fragen stellen.	a
we cannot prevent that .	dies können wir nicht verhindern .	Das können wir nicht verhindern.	wir können nicht verhindern.	a
that scheme is both absurd and dangerous .	diese strategie ist sowohl absurd als auch gefährlich .	Das ist sowohl absurd als auch gefährlich.	Das Vorhaben ist sowohl absurd als auch gefährlich.	same
but i remember the many children in my neighborhood who were not vaccinated .	aber ich erinnere mich an die vielen kinder in meiner nachbarschaft , die nicht geimpft wurden .	aber ich erinnere mich an die vielen Kinder in meiner Nachbarschaft, die nicht geimpft wurden.	aber ich erinnere mich an die vielen Kinder in meiner Nachbarschaft, die nicht gewirtschaftet wurden.	a
for there is such a thing as a lawless legality .	denn es gibt so etwas wie eine gesetzlose legalität .	denn es gibt so etwas wie eine rechtmäßige Legalität.	Denn so etwas gibt es eine Sache als gesetzwichssigkeit.	a
but what about longer-term risks ?	aber was ist mit längerfristigen risiken ?	aber was ist mit längerfristigen Risiken?	aber was ist mit längerfristigen Risiken?	same
nobody in israel will accept this .	niemand in israel wird das hinnehmen .	niemand in israel wird das akzeptieren.	niemand wird diese annehmen.	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
hosni mubarak has been sacrificed to save the military regime .	hosni mubarak wurde geopfert , um das militärregime zu retten .	hosni mubarak wurde geopfert, um das militärische Regime zu retten.	Hacksni mubarak wurde geopfert, um das militärische Regime zu retten.	a
the facts speak for themselves .	die tatsachen sprechen für sich .	Die Fakten sprechen für sich.	die Fakten sprechen für sich selbst.	same
putin is in no hurry , but he clearly knows what he wants .	putin hat keine eile , aber er weiß ganz klar , was er will .	Beeilen Sie sich in keine Eile, aber er weiß klar, was er will.	setzend, ist nicht eilig, aber er weiß deutlich, was er will.	same
frankly , all of us already know what needs to be done .	offen gesagt , wissen wir alle bereits , was getan werden muss .	offen gesagt, alle von uns wissen schon, was getan werden muss.	Seltsamerweise wissen wir alle bereits, was ein getan werden muss.	a
vaccines have always had a special meaning for me .	impfungen hatten für mich immer eine besondere bedeutung .	Impfstoffe haben für mich immer eine besondere Bedeutung.	Impfstoffe haben immer einen besonderen Sinn für mich.	a
this is true regardless of why women migrate .	das gilt unabhängig davon , warum frauen auswandern .	Das gilt unabhängig davon, warum Frauen wandern.	Hier geht es unabhängig davon, ob sich Frauen wandern.	a

Model A Total Score	Model B Total Score	Total ties
11	0	7



Human evaluation of experiment 1 deemed NMT trained on synthetically trained data have superior outputs on the WMT17 test set. Due to the significant quantity of ties, the judge considered the NMT trained synthetic data to only be slightly superior.

Experiment 2 – Results

Experiment 2 NMT Training Charts

Figures below illustrate experiment 2 validation BLEU score which calculated during training of the model.

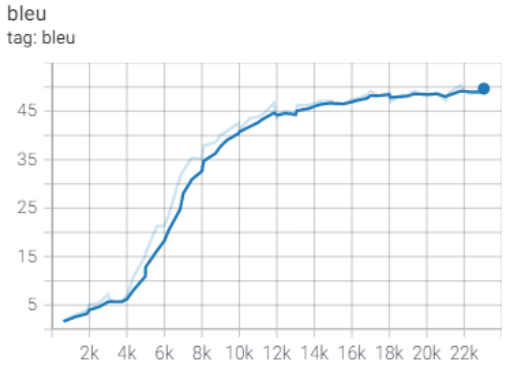


Figure 36 Experiment 2 Synthetic Dataset NMT validation BLEU score during training Single run.

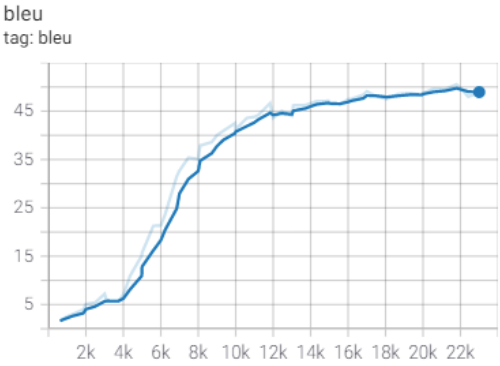


Figure 35 Experiment 2 Original Dataset NMT
validation BLEU score during training Single run

Experiment 2 BLEU Score Test Results

Experiment	Corpus	Optimizer Run	Test Data Set	Original Data BLEU Score	Synthetic Data BLEU Score	Difference in score (Original data score - Synthetic data Score)
experiment2	paracrawl	single_run	Withheld Corpus Data	38.43	38.72	-0.29
experiment2	paracrawl	single_run	wmt17	5.6	5.4	0.2
experiment2	paracrawl	single_run	wmt18	7.9	7.9	0
experiment2	paracrawl	single_run	wmt19	7.2	7.1	0.1
experiment2	paracrawl	single_run	wmt20	5.2	5.1	0.1

Table 9: Experiment 2 BLEU Score results from Test Data Set Translations

Average difference between original data and synthetic data BLEU scores: 0.022

Standard deviation between original data and synthetic data BLEU scores: 0.168333

Experiment 2 Character and word F Score results

Experiment	Corpus	Optimizer Run	Test Data Type	Original Data c6+w2-F2 Score	Synthetic Data c6+w2-F2 Score	Difference in score (Original data score - Synthetic data Score)
experiment2	paracrawl	single_run	wmt17	33.9448	33.6656	0.2792
experiment2	paracrawl	single_run	wmt18	37.4339	37.2972	0.1367
experiment2	paracrawl	single_run	wmt19	35.4057	35.0694	0.3363
experiment2	paracrawl	single_run	wmt20	33.1857	32.8424	0.3433
experiment2	paracrawl	single_run	Withheld Corpus Data	56.7052	56.7091	-0.0039

Table 10: Experiment 2 character and word F Score results from Test Data Set Translations

Average difference between original data and synthetic data F scores: 0.21832

Standard deviation between original data and synthetic data F scores: 0.133633

Experiment 2 Human Evaluation Test Results

Results of relative ranking human evaluation performed by native German speaker on experiment2 NMT model outputs of WMT18 test set.

Model A = NMT trained on Synthetic data.

Model B = NMT trained on Original data.

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
And no-one knows what is coming next.	Und niemand weiß, was als nächstes kommt.	und niemand weiß, was als nächstes kommt.	und niemand weiß, was als nächstes erscheint.	same
Maybe the two are connected?"	Vielleicht sind die beiden miteinander verbunden?"	vielleicht sind die beiden miteinander verbunden? "	vielleicht sind die beiden miteinander verbunden? "	same
“I don’t know exactly what they are going to say; it is odd for me," says Mordillo.	"Ich weiss nicht genau, was sie sagen werden, es ist seltsam für mich", sagt Mordillo.	"ich weiß nicht genau, was sie sagen werden, es ist für mich", sagt mordillo.	"ich weiß nicht genau, was sie sagen werden, es ist für mich", sagt mordillo.	same
I feel very good."	Ich fühle mich sehr gut".	ich fühle mich sehr gut ".	ich fühle mich sehr gut ".	same
It just shows how far we've come.	Es zeigt nur, wie weit wir gekommen sind.	es zeigt gerade, wie weit wir gekommen sind.	es zeigt gerade, wie weit wir gekommen sind.	same
Whenever they're ready ... ."	Wann immer sie bereit sind ... "	immer wenn sie bereit sind... ".	wann immer sie bereit sind... ".	b

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
But it's not much more than an idea.	Aber es ist nicht viel mehr als eine Idee.	aber es ist nicht viel mehr als eine idee.	aber es ist nicht viel mehr als eine idee.	same
And it especially matters when it's a guy like Hogan.	Und es ist besonders wichtig, wenn es ein Typ wie Hogan ist.	und es ist besonders wichtig, wenn es ein kerl ist wie hogan.	und es ist vor allem wichtig, wenn es ein kerl ist.	a
He likes the country for which he has worked for the last 50 years.	Er mag das Land, für das er in den letzten 50 Jahren gearbeitet hat.	er mag das land, für das er in den letzten 50 jahren arbeitete.	er mag das land, für das er in den letzten 50 jahren arbeitete.	same
"Once a country owns a club, everything is possible.	"Sobald ein Land einen Verein besitzt, ist alles möglich.	"einmal ein land besitzt, ist alles möglich.	"einmal ein land besitzt, ist alles möglich.	same
"Before, we would have already known what happened," said Villarreal, 46, nicknamed "El Flaco" for his slender build.	"Früher hätten wir schon gewusst, was passiert ist", sagte Villarreal, 46, mit dem Spitznamen "El Flaco" wegen seiner schlanken Körperstatur.	"vorher hätten wir schon gewusst, was passiert ist", sagt villarreal, 46, mit dem namen "el flaco" für seine schlanken gebaut.	"vorher hätten wir bekannt gegeben, was geschehen war", sagt villarreal, 46, mit dem namen "el flaco" für seine schlanken bauherr aufgebaut.	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
That is all absolutely terrible and I am glad that it is finally being looked into.	Das ist alles ganz schrecklich und ich bin froh, dass es endlich aufgearbeitet wird.	das ist absolut schrecklich und ich bin froh, dass es endlich in die augen geraten ist.	das ist absolut schrecklich bedauerlich, und ich bin froh, dass es endlich in die augen geraten ist.	same
Now I look back and think: 'Girl, be proud of yourself!'".	Nun schau ich zurück und denke mir: 'Mädchen, sei stolz auf dich!'."	jetzt komme ich zurück und denke: "mädchen, sei stolz auf dich!"	jetzt schaue ich zurück und denke: "mädchen, ich bin stolz auf dich!"	a

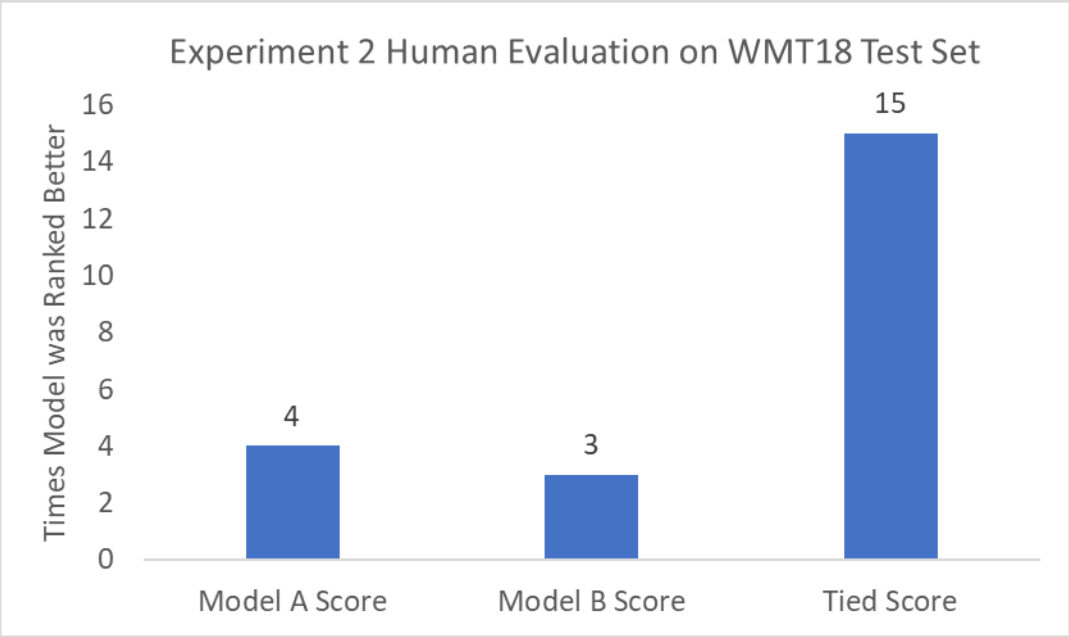
But it matters.	Aber es ist wichtig.	aber es ist wichtig.	aber es aber wichtig aber wichtig aber wichtig aber aber aber wichtig aber aber aber aber wichtig aber aber aber aber wichtig aber aber aber aber aber aber wichtig aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber wichtig aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber	a

			aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber aber	
This shouldn't be so hard.	Das sollte nicht so schwer sein.	das sollte doch nicht so schwer sein.	dies sollte nicht so sehr schwierig sein.	same

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
"I'm afraid," he says.	"Ich habe Angst", sagt er.	"ich fürchte", sagt er.	"ich fürchte mich", sagt er.	b
"The steelwork is quite complicated," he explains.	"Das Stahlwerk ist ziemlich kompliziert", erklärt er.	"die steelwork ist sehr kompliziert", erklärt er.	"die steelwerk ist sehr kompliziert", erklärt er.	same
"I find it so beautiful to look at the handwriting of someone who has been dead for 300 years," he says wistfully.	"Ich finde es so schön, auf die Handschrift von jemandem zu schauen, der seit 300 Jahren tot ist", sagt er wehmütig.	"ich finde es so schön, das handschriftliche schreiben von jemand, der seit 300 jahren tot ist", sagt er wistly.	"ich finde es so schön, als ich das handschreiben von jemandem, der seit 300 jahren tot ist", sagt er wistend.	same
"I moved up for Broner [at 140 pounds], but I'm not that big.	"Ich bin für Broner [auf 140 Pfund] aufgestiegen, aber ich bin nicht so groß.	"ich bewegte mich für broner [at], aber ich bin nicht so groß.	ich zog für broner [bei 140 pfund], aber ich bin nicht so groß.	b
"I know Sikorsky very well," the President said, "I have three of them."	"Ich kenne Sikorsky sehr gut", sagte der Präsident, "ich habe drei davon."	"ich kenne sikorsky sehr gut", sagte der präsident, "ich habe drei von ihnen".	"ich kenne sikorsky sehr gut", sagte der präsident, "ich habe drei".	same
Coe said he was also "pleased" that Russia accepted the criteria for its reintroduction.	Coe sagte, er sei auch "erfreut" darüber, dass Russland die Kriterien für seine Wiedereinführung akzeptiere.	coe sagte, er sei auch "erfreut", dass russland die kriterien für seine wiedereinführung akzeptiert.	coe sagte, er sei auch "erfreut", dass russland die kriterien für seine wiedereinführung akzeptiert.	same

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
Her statue speaks back to them the words of hope and welcome they need to hear.	Ihre Statue spricht zu ihnen die Worte der Hoffnung und des Willkommens, die sie hören müssen.	ihre statue spricht zu ihnen die worte der hoffnung und begrüßung, die sie hören müssen.	ihre statue redet sich mit den worten der hoffnung und willkommen, die sie hören müssen.	same

Model A Score	Model B Score	Tied Score
4	3	15



Human evaluation of experiment 2 deemed that NMT model trained on synthetically trained data more often produced superior quality translation than that of the original model. Due to the significant quantity of ties, the judge considered the NMT trained synthetic data to only be slightly superior.

Experiment 3 – Results

Experiment 3 NMT Training Charts

Figures below illustrate experiment 3 validation BLEU score which calculated during training of the model.

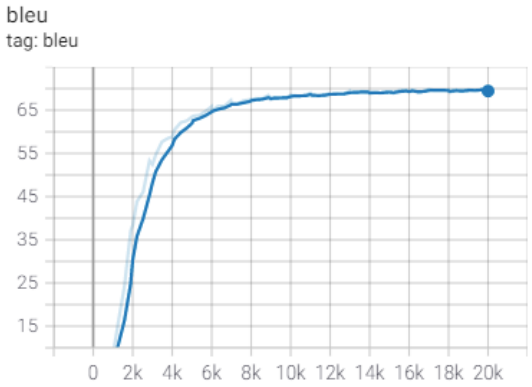


Figure 37 Experiment 3 Synthetic Dataset NMT validation BLEU score during training single run

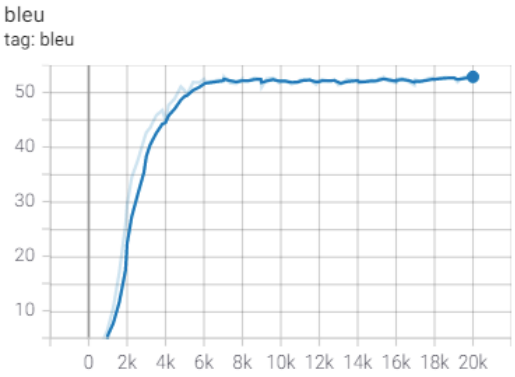


Figure 19 Experiment 3 Original Dataset NMT
validation BLEU score during training single run

Experiment 3 BLEU Score Test Results

Experiment	Corpus	Optimizer Run	Test Data Set	Original Data BLEU Score	Synthetic Data BLEU Score	Difference in score (Original data score - Synthetic data Score)
experiment3	opus	single_run	Withheld Corpus Data	38.43	33.31	5.12
experiment3	opus	single_run	wmt17	5.6	6.9	-1.3
experiment3	opus	single_run	wmt18	7.9	9.8	-1.9
experiment3	opus	single_run	wmt19	7.2	8.6	-1.4
experiment3	opus	single_run	wmt20	5.2	5.9	-0.7

Table 11: Experiment 3 BLEU Score results from Test Data Set Translations

Average difference between original data and synthetic data BLEU scores: -0.036

Standard deviation between original data and synthetic data BLEU scores: 2.606067

Experiment 3 Character and word F Score results

Experiment	Corpus	Optimizer Run	Test Data Type	Original Data c6+w2-F2 Score	Synthethic Data c6+w2-F2 Score	Difference in score (Original data score - Synthethic data Score)
experiment3	opus	single_run	wmt17	33.9448	36.6112	-2.6664
experiment3	opus	single_run	wmt18	37.4339	41.0231	-3.5892
experiment3	opus	single_run	wmt19	35.4057	38.5282	-3.1225
experiment3	opus	single_run	wmt20	33.1857	35.1337	-1.948
experiment3	opus	single_run	Withheld Corpus Data	56.7052	56.1585	0.5467

Table 12:Experiment 3 character and word F Score results from Test Data Set Translations

Average difference between original data and synthetic data F scores: -2.15588

Standard deviation between original data and synthetic data F scores: -1.45578

Experiment 3 Human Evaluation Test Results

Results of relative ranking human evaluation performed by native German speaker on experiment3 NMT model outputs of WMT19 test set.

Model A = NMT trained on Synthetic data.

Model B = NMT trained on Original data.

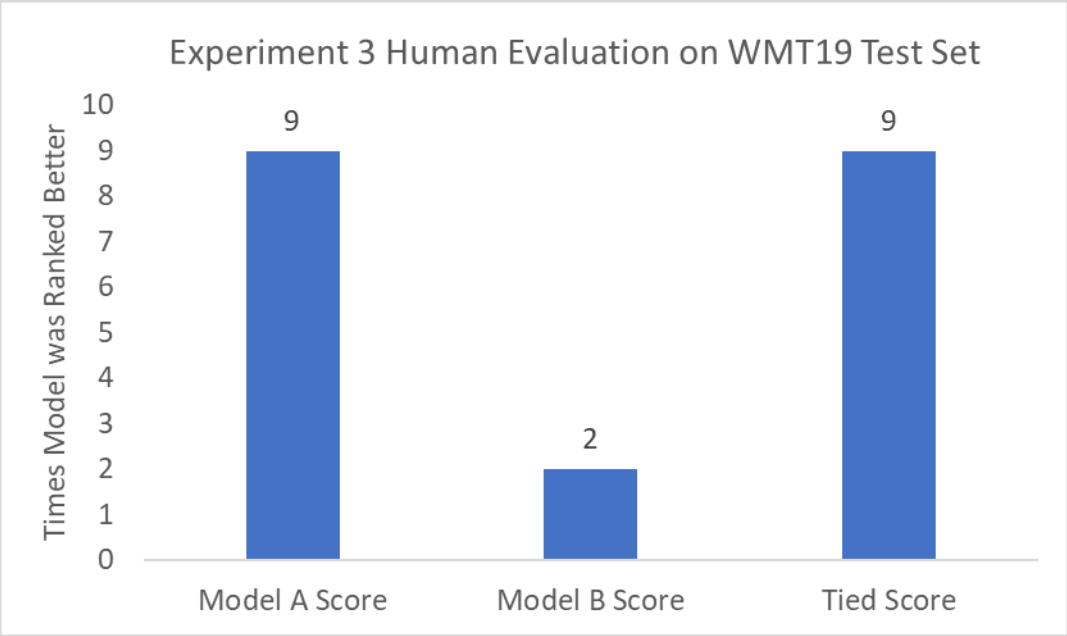
Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
"I think that means the messages were not delivered," he said.	„Ich denke, das bedeutet, dass die Botschaften nicht übermittelt wurden", sagte er.	"ich denke, das bedeutet, dass die nachrichten nicht geliefert wurden", sagte er.	"ich denke, dass die nachrichten nicht ausgeliefert wurden", sagte er.	a
He was speechless when he discovered what happened.	Er war sprachlos, als er herausfand, was passiert ist.	er war sprachlos, als er entdeckt hat, was passiert ist.	er wurde speechnisch, als er entdeckte, was passiert.	a
I knew how much he loved her."	Ich wusste, wie sehr er sie geliebt hatte.“	ich wusste, wie sehr er sie liebte.	ich wusste, wie viel er sie liebte ".	a
"My wife had sent her something that she didn't mean to say.	"Meine Frau hatte ihr etwas geschickt, was sie nicht sagen wollte.	"meine frau hatte ihr etwas geschickt, was sie nicht zu sagen hatte".	"meine frau hatte ihr etwas, das sie nicht gesagt hatte.	same
"I think this is a genus idea.	Ich denke, das ist eine großartige Idee.	ich denke, das ist eine genus-idee.	"ich denke, das ist eine genius idee.	same
'We know it won't change students' behaviour instantly.	Wir wissen, dass es das Verhalten der Studenten nicht sofort ändern wird.	"wir wissen, dass es das verhalten der studenten sofort ändern wird.	wir wissen, dass es nicht sofort das verhalten der schüler verändern wird.	b

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
Giles noted that there are more than 135 shark species in the area, but most are not considered dangerous.	Giles stellte fest, dass es mehr als 135 Haiarten in der Gegend gibt, aber die meisten gelten nicht als gefährlich.	vergoldet, dass es mehr als 135 shark-arten in der umgebung gibt, aber die meisten werden nicht als gefährlich angesehen.	giles bemerkte, dass es mehr als 135 shark auf dem gebiet gibt, aber die meisten werden nicht gefährlich.	same
I am so happy, so happy to get the cup back.	Ich bin so glücklich, so glücklich, den Cup zurück zu bekommen.	ich bin so glücklich, also glücklich, den pokal zurück zu bekommen.	ich bin so glücklich, also glücklich, um die tasse kaffe zu bekommen.	a
Now we have this great relationship."	Jetzt haben wir diese großartige Beziehung“.	jetzt haben wir diese großartige beziehung.	jetzt haben wir diese große beziehung ".	same
Who will care for the dog?	Wer kümmert sich um den Hund?	wer kümmert sich um den hund?	wer für den hund sorgen?	a
I think that's really one of the big questions in our time - how do we change that?	Ich denke, das ist wirklich eine der großen Fragen in unserer Zeit – wie können wir das ändern?	ich denke, das ist wirklich eine der großen fragen in unserer zeit - wie ändern wir das?	ich denke, das ist eine der großen fragen in unserer zeit - wie verwandeln wir das?	a
"The ruling is an opportunity to see that we need to overcome the past."	„Das Ergebnis ist eine Chance, zu sehen, dass wir die Vergangenheit überwinden müssen.“	"die ruling ist eine gelegenheit, zu sehen, dass wir die vergangenheit überwinden müssen".	"die herrschende ist eine chance zu sehen, dass wir die vergangenheit überwinden müssen".	same
"He was quite a quiet man, and he wasn't a boastful person," she said.	"Er war ein ziemlich ruhiger Mann, und er war keine bostache Person", sagte sie.	"er war recht leiser mann, und er war keine gotse person", sagte sie.	"er war ein sehr leiser mann, und er war kein böser mensch", sagte sie.	same

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
"We've been tested over our games this far, and we're still unbeaten, so I have to be happy," he said.	"Wir wurden durch unsere Spiele so weit getestet, und wir sind immer noch ungeschlagen, also muss ich glücklich sein", sagte er.	"wir wurden bisher über unsere spiele getestet, und wir sind immer noch unerreichbar, also musste ich mich freuen", sagte er.	"wir haben unsere spiele bisher getestet, und wir sind noch ungeschlagen, also habe ich glücklich", sagt er.	a
Maybe she didn't, but that's the worst body language I've ever seen."	Vielleicht hat sie es nicht getan, doch das war die schlechteste Körpersprache, die ich je gesehen habe."	vielleicht hat sie das nicht getan, aber das ist die schlimmste körpersprache, die ich je gesehen habe.	vielleicht hast du es nicht getan, aber das ist die schlimmste körpersprache, die ich je gesehen habe ".	a
However, instead of sending it to her husband, she sent it to Ms. Maurice, twice.	Anstatt ihn jedoch an ihren Mann zu schicken, schickte sie ihn zweimal an Frau Maurice.	anstatt ihn an ihren mann zu schicken, schickte sie ihn jedoch an frau maurice, zweimal.	anstatt sie an ihren ehemann zu senden, schickten sie sie sie an frau maurice, zwei mal.	same
They were probably better in the first half and we came out in the second half and were the better side.	Sie waren wahrscheinlich besser in der ersten Hälfte und wir mobilisierten uns in der zweiten Hälfte und waren die bessere Seite.	sie waren wahrscheinlich besser in der ersten hälfte und wir kamen aus der zweiten hälfte und waren die bessere seite.	sie waren wahrscheinlich besser als die hälfte und wir kamen in der zweiten hälfte und waren auf der bessere seite.	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
"But because Paddington was so real to him, it was almost like if you have a child who achieves something: you're proud of them even though it's not your doing really.	„Aber weil Paddington so echt für ihn war, war es fast so, als ob man ein Kind hat, das etwas erreicht: Du bist stolz auf das Kind, obwohl es nicht wirklich dein Tun ist.	"aber weil paddington so real war wie er war, war es fast so, als ob man ein kind hat, das etwas erreicht hat: man ist stolz darauf, auch wenn es nicht das wirklich macht.	"aber weil paddington so echt ged war, war es schon fast so, als ob du ein kind hast, der etwas erreicht: du bist stolz darauf, auch wenn es nicht wirklich ist.	same
Nickel mining is also important in the province, but is mostly concentrated in Morowali, on the opposite coast of Sulawesi.	Nickelbergbau ist auch in der Provinz wichtig, wird aber hauptsächlich in Morowali betrieben, an der gegenüberliegenden Küste von Sulawesi.	nickel mining ist auch in der provinz wichtig, wird aber hauptsächlich auf morowali konzentrieren, an der gegenüberliegenden küste des sulawesi.	der bergbau ist auch in der provinz liegt ebenfalls wichtig, ist aber meist in morowali, auf der gegenüberliegenden küste von sulawesi.	same
"No-one knew," she said.	„Keiner wusste etwas“, sagte sie.	"no-one wusste", sagte sie.	"niemand wusste", sagte sie.	b

Model A Score	Model B Score	Tied Score
9	2	9



Human evaluation of experiment 3 deemed NMT trained on synthetically trained data have superior outputs on the WMT19 test set. Due to the significant quantity of ties, the judge considered the NMT trained synthetic data to only be slightly superior.

Experiment 4 – Results

Experiment 4 NMT Training Charts

Figures below illustrate experiment 4 validation BLEU score which calculated during training of the model.

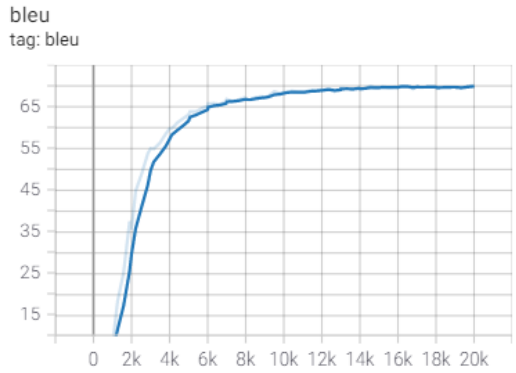


Figure 39 Experiment 4 Synthetic Dataset NMT validation BLEU score during training single run

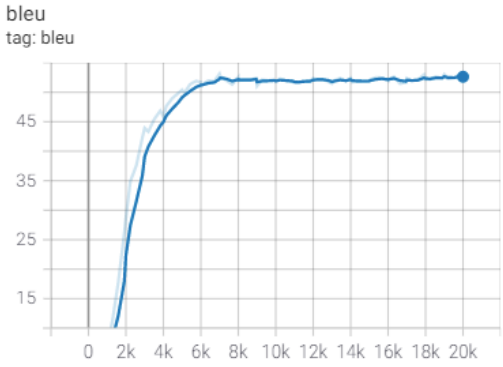


Figure 38 Experiment 4 Original Dataset NMT validation BLEU score during training single run

Experiment 4 BLEU Score Test Results

Experiment	Corpus	Optimizer Run	Test Data Set	Original Data BLEU Score	Synthethic Data BLEU Score	Difference in score (Original data score - Synthethic data Score)
experiment4	opus	single_run	Withheld Corpus Data	42.79	45.55	-2.76
experiment4	opus	single_run	wmt17	6.1	7	-0.9
experiment4	opus	single_run	wmt18	8.1	10	-1.9
experiment4	opus	single_run	wmt19	6.9	8.7	-1.8
experiment4	opus	single_run	wmt20	4.7	6	-1.3

Table 13 Experiment 4 BLEU Score results from Test Data Set Translations

Average difference between original data and synthetic data BLEU scores: -1.732

Standard deviation between original data and synthetic data BLEU scores: 0.627452

Experiment 4 Character and word F Score results

Experiment	Corpus	Optimizer Run	Test Data Type	Original Data c6+w2-F2 Score	Synthethic Data c6+w2-F2 Score	Difference in score (Original data score - Synthethic data Score)
experiment4	opus	single_run	wmt17	34.1115	36.7279	-2.6164
experiment4	opus	single_run	wmt18	37.4826	41.1643	-3.6817
experiment4	opus	single_run	wmt19	35.1679	38.2734	-3.1055
experiment4	opus	single_run	wmt20	32.0013	35.3034	-3.3021
experiment4	opus	single_run	Withheld Corpus Data	63.3767	65.8185	-2.4418

Table 14 Experiment 4 character and word F Score results from Test Data Set Translations

Average difference between original data and synthetic data F scores: -3.0295

Standard deviation between original data and synthetic data F scores: 0.451994

Experiment 4 Human Evaluation Test Results

Results of relative ranking human evaluation performed by native German speaker on experiment4 NMT model outputs of withheld corpus test set.

Model A = NMT trained on Synthetic data.

Model B = NMT trained on Original data.

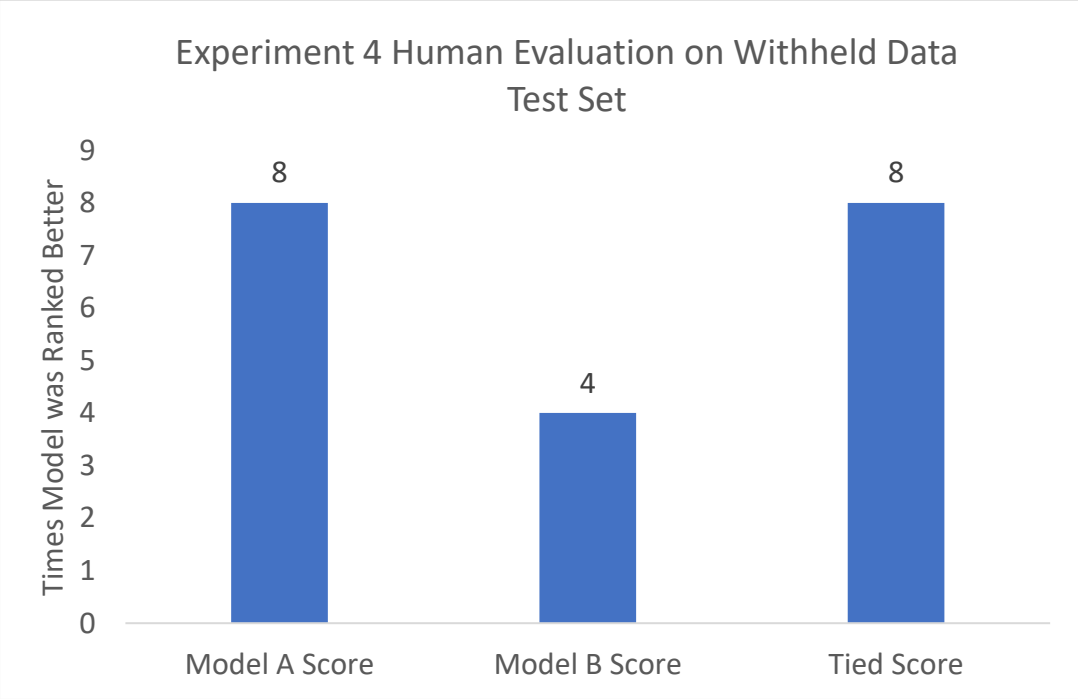
Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
please read the next paragraph before you buy cialis online!	lesen sie den nächsten absatz, bevor sie cialis online kaufen!	bitte lesen sie den nächsten absatz, bevor sie cialis online kaufen!	bitte lesen sie den folgenden abschnitt, bevor sie cialis online kaufen können!	a
welcome to our virtual tuning-store!	herzlich willkommen in unserem virtuellen tuning-store!	willkommen in unserem virtuellen tuning-store!	willkommen in unserem virtuellen tuning-geschäft!	b
because your eyes deserve the best	weil ihre augen das beste verdienen.	weil ihre augen das beste verdienen	weil ihre augen verdient das beste	a
only the best for you.	einfach nur das beste für sie.	nur das beste für sie.	nur das beste für sie.	same
or contact us..	oder kontaktieren sie uns...	oder kontaktieren sie uns..	oder kontaktieren sie uns..	same
please fill out these fields	bitte füllen sie diese felder aus:	bitte füllen sie diese felder aus	bitte füllen sie die folgenden felder aus	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
more than 12 people please email	mehr als 12 personen bitte e-mail an	mehr als 12 personen bitte e-mail	mehr als 12 personen bitte per e-mail	same
10 other products from the same category:	10 andere produkte der gleichen kategorie:	andere produkte der gleichen kategorie:	10 andere artikel der gleichen kategorie:	b
contact us for any further information	bitte kontaktieren sie uns für weitere informationen	kontaktieren sie uns für weitere informationen	kontaktieren sie uns für weitere informationen	same
dogs welcome (may to october)	hunde willkommen (mai bis oktober)	willkommen (mai bis oktober)	hunde willkommen (mai bis oktober)	b
welcome to cep paper processing,	herzlich willkommen bei cep paper processing,	willkommen bei cep paper processing,	willkommen bei cep paper processing,	same
working together in 3 simple steps ™	die zusammenarbeit in 3 einfachen schritten ™	zusammenarbeit in 3 einfachen schritten ™	gemeinsam in 3 einfachen schritten ™	a
so many years have you been looking and here you are,	so viele jahre haben sie gesucht, und hier sind sie,	so viele jahre suchen sie und hier sind sie,	so viele jahre haben sie gesucht und hier sind sie,	b
we are very happy about this great start in 2016 and give a huge thanks to all customers and visitors for their interest.	wir freuen uns über diesen guten start in 2016 und danken allen kunden und besuchern für ihr interesse.	wir freuen uns sehr über diesen großartigen start in 2016 und geben allen kunden und besuchern für	wir freuen uns sehr über diesen tollen start im jahr 2016 und geben ein großer dank aller kunden	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
		ihr interesse einen großen dank.	und besucher für ihr interesse.	
our team members all have responsibility into the heart, with good teamwork and social practice ability.	unsere teammitglieder alle haben verantwortung in das herz, mit guter teamwork und sozialpraxisfähigkeit.	unsere teammitglieder tragen alle verantwortung in das herz, mit guter teamarbeit und sozialen übungsfähigkeit.	unser team mitglieder alle haben verantwortung in das herz, mit guter teamwork und social practice.	a
muuuz: via why he chose you to be designer of this exhibition?	muuuz: via warum wählte er sie designer dieser ausstellung zu sein?	muuuz: warum hat er sie entschieden, designer dieser ausstellung zu sein?	muuuuz: warum wählte er dich bei der ausstellung dieser ausstellung?	a
you can find detailed information on homepage.	sie finden detaillierte informationen auf der homepage.	detaillierte informationen finden sie auf der homepage.	ausführliche informationen finden sie auf der startseite.	a
i have taken note of the data protection conditions.	die datenschutzbestimmungen habe ich zur kenntnis genommen.	ich habe die datenschutzbestimmungen zur kenntnis genommen.	ich habe die datenschutzerklärung zur kenntnis genommen.	same
there were many additional benefits discovered that include:	es gab viele zusätzliche vorteile entdeckt, enthalten:	es gab viele zusätzliche vorteile, die entdeckt wurden:	es wurden viele weitere vorteile gefunden, die hier eingeschlossen wurden:	same

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
country bulgaria croatia czech republic finland france germany italy netherlands poland romania slovenia spain sweden united kingdom model acunex ® acunex ® vario acunex ® variomax document type eds elfu	land bulgarien deutschland finnland frankreich großbritannien italien kroatien niederlande polen rumänien schweden slowenien spanien tschechische republik modell acunex ® acunex ® vario acunex ® variomax dokumenttyp eds elfu	land bulgarien dänemark tschechische republik finnland frankreich deutschland italien niederlande polen rumänien slowenien spanien schweden großbritannien modell acunex ® acunex ® vario acunex ® variomax dokumentstyp eds elfu	land bulgarien kroatien kroatien tschechische republik finnland frankreich italien niederlande polen slowenien slowenien slowenien slowenien spanien vereinigtes königreich modell acunex ® acunex ® vario acunex ® variomax document typ ds elfu	same

Model A Score	Model B Score	Tied Score
8	4	8



Human evaluation of experiment 4 deemed NMT trained on synthetically trained data have superior outputs on the withheld corpus test set. Due to the significant quantity of ties, the judge considered the NMT trained synthetic data to only be slightly superior.

Experiment 5 – Results

Experiment 5 NMT Training Charts

Figures below illustrate experiment 5 validation BLEU score which calculated during training of the model.

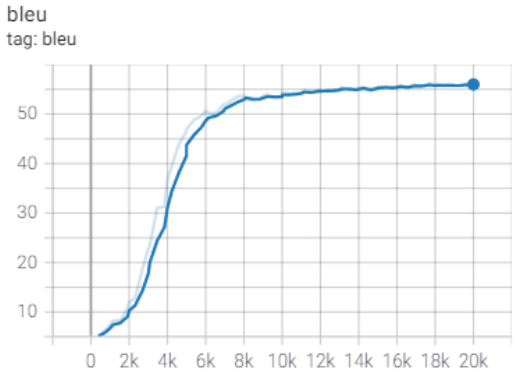


Figure 41 Experiment 5 Synthetic Dataset NMT validation BLEU score during training Single run.

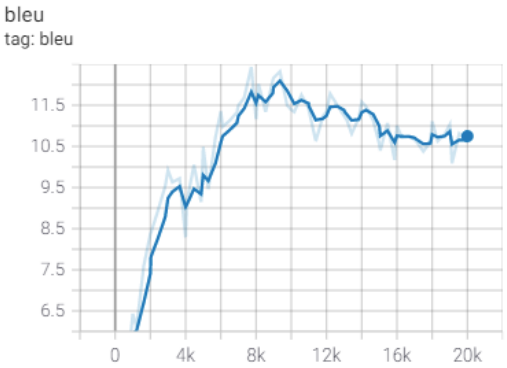


Figure 40 Experiment 5 Original Dataset NMT
validation BLEU score during training Single run.

Experiment 5 BLEU Score Test Results

Experiment	Corpus	Optimizer Run	Test Data Set	Original Data BLEU Score	Synthethic Data BLEU Score	Difference in score (Original data score - Synthethic data Score)
experiment5	UFAL	single_run	Withheld Corpus Data	13.57	12.79	0.78
experiment5	UFAL	single_run	wmt17	0.3	5.5	-5.2
experiment5	UFAL	single_run	wmt18	0.5	7.7	-7.2
experiment5	UFAL	single_run	wmt19	0.6	7.5	-6.9
experiment5	UFAL	single_run	wmt20	0.5	5.8	-5.3

Table 15 Experiment 5 BLEU Score results from Test Data Set Translations

Average difference between original data and synthetic data BLEU scores: -4.764

Standard deviation between original data and synthetic data BLEU scores: 2.888249

Experiment 5 Character and word F Score results

Experiment	Corpus	Optimizer Run	Test Data Type	Original Data c6+w2-F2 Score	Synthethic Data c6+w2-F2 Score	Difference in score (Original data score - Synthethic data Score)
experiment5	UFAL	single_run	wmt17	16.817	33.9292	-17.1122
experiment5	UFAL	single_run	wmt18	17.0783	37.8745	-20.7962
experiment5	UFAL	single_run	wmt19	17.0425	35.7811	-18.7386
experiment5	UFAL	single_run	wmt20	17.0055	33.7456	-16.7401
experiment5	UFAL	single_run	Withheld Corpus Data	29.3332	26.5839	2.7493

Table 16 Experiment 5 character and word F Score results from Test Data Set Translations

Average difference between original data and synthetic data F scores: -14.1

Standard deviation between original data and synthetic data F scores: 8.56

Experiment 5 Human Evaluation Test Results

Results of relative ranking human evaluation performed by native German speaker on experiment5 NMT model outputs of WMT18 test set.

Model A = NMT trained on Synthetic data.

Model B = NMT trained on Original data.

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
And no-one knows what is coming next.	Und niemand weiß, was als nächstes kommt.	und niemand weiß, was als nächstes kommt.	der zweite punkt lautet?	a
We can be glad that we have long overcome this time.	Wir können froh sein, dass wir diese Zeit längst überwunden haben.	wir können froh sein, dass wir diese zeit längst überwunden haben.	schön, dass wir lange zeit haben.	a
Menge: Yes, it is very variable.	Menge: Ja, es ist sehr variabel.	menge: ja, es ist sehr variabel.	ja, das ist sehr angenehm.	a
Menge: It is a nice, broad spectrum that I see here.	Menge: Es ist ein schönes, breites Spektrum, das ich hier sehe.	menge: es ist ein schönes, breites spektrum, das ich hier sehe.	das ist eine neue, lebhafte, gut.	a
That is all absolutely terrible and I am glad that it is finally being looked into.	Das ist alles ganz schrecklich und ich bin froh, dass es endlich aufgearbeitet wird.	das ist alles absolut schrecklich und ich bin froh, dass es endlich aussehen lässt.	ich bin mir ziemlich sicher, dass dies eine falsche rolle ist.	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
"Everything has been announced, everything will be implemented.	"Alles wurde angekündigt, alles wird umgesetzt.	revier quot; alles wurde angekündigt, alles wird umgesetzt.	topographisch ist es nicht möglich, alle unsere kolo in europa einzutreten.	a
According to Human Rights Watch, five young protesters were killed and many more injured.	Laut Human Rights Watch wurden fünf junge Demonstranten getötet und viele weitere verletzt.	gemäß den menschenrechten wurden fünf junge proteste getötet und viele weitere verletzt.	fünf jahre nach einem der ansprüche 1 bis 5, wobei der schutz der menschenwürde und der achtung vor allem der menschenrechte und der menschenwürde getan werden.	a
She needs to know it and see it and be raised in it.	Sie muss es wissen und es sehen und darin aufgezogen werden.	sie muss es wissen und es sehen.	sie soll es beinhalten, und das muss sie wissen.	a
Others have tried and failed.	Andere haben es versucht und sind gescheitert.	andere haben es versucht und gescheitert.	und die anderen haben es nicht getan.	a
Guglielmi said it is unclear if the incident also occurred on July 27.	Guglielmi sagte, es sei unklar, ob der Vorfall auch am 27. Juli stattgefunden habe.	guglielmi sagte, es sei unklar, wenn der vorfall auch am 27. juli fand.	elizabem ist auf der anderen seite sehr leicht.	a

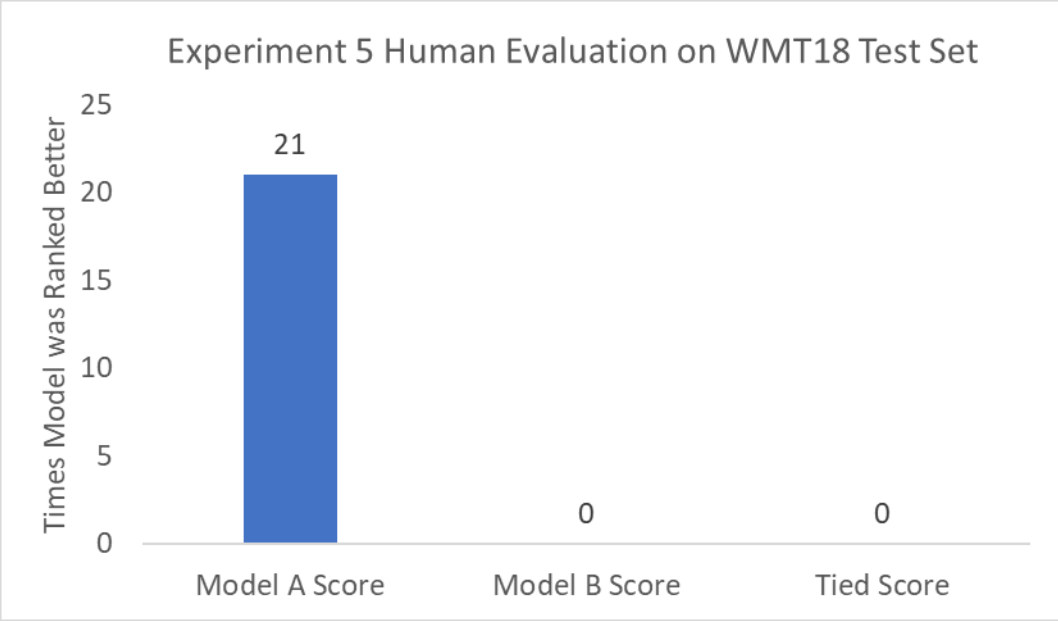
Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
And because in the crisis talks afterwards, he pretty much did everything wrong that you can do wrong.	Und weil er in der Krisenkommunikation danach so ziemlich alles falsch machte, was man falsch machen kann.	und weil er in den krisengesprächen danach alles falsch gemacht hat, was man falsch machen kann.	vielleicht hast du das gefühl, dass du in ordnung gebracht hast.	a
But it matters.	Aber es ist wichtig.	aber es ist wichtig.	aber ich habe es gewusst.	a
Polygamy is common here.	Polygamie ist hier üblich.	polyspielzeug ist hier üblich.	da ist polymoris.	a
If so, he will be the first.	Wenn ja, wird er der Erste sein.	wenn ja, wird er der erste sein.	der erste, der das sein mag.	a
Miller was not engaging in literary criticism - he was making it clear that these people are not welcome in Trump's USA.	Miller engagierte sich nicht in literarischer Kritik - er machte deutlich, dass diese Leute in Trumps USA nicht willkommen sind.	miller kam in analphabetischer kritik nicht an - er machte deutlich, dass diese menschen in trumpf nicht willkommen sind.	unsere konservativen englischen kollegen, die im bericht von frau wetten in new york sind, gibt es keine elemente, die in diesen fällen öffentlich sind.	a
"This documentary is not somebody speaking for us or speaking to us, it's us speaking," Davis said.	"Dieser Dokumentarfilm ist nicht jemand, der für uns spricht oder mit uns	lecquot; dieser dokumentationston ist nicht jemand, der für uns spricht oder mit uns spricht, er hat uns	da ist erstens die koloniale grundrechude, denn dieser mann hat keine grenzen mehr.	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
	spricht. Wir reden", sagte Davis.	siedenties zitronen zu sprechen; davis sagte.		
The heavily armoured dinosaur used red and white camouflage to hide from predators, and employed a shielding technique known as counter-shading, which is also used by many modern-day animals.	Der stark gepanzerte Dinosaurier verwendete eine rote und weiße Tarnung, um sich vor Raubtieren zu verstecken, und verwendete eine Abschirmungstechnik, die als Gegenbeschattung bekannt ist, die auch von vielen modernen Tieren verwendet wird.	die stark gepanzerte dinosaur nutzte rotes und weißes tarning, um sich vor predaten zu verstecken, und eine abschirmtechnik, die als gegenrasierklinge bekannt ist, die auch von vielen modernen tag verwendet wird.	der erste schritt bestand darin, eine entschiedene & quot; abgrenzung des nutzens & quot; zu verfolgen, und die missbildungen von brutalität wird vom typ zacken als extremer und als ganzes angesehen.	a
I felt really attacked, frightened and ashamed.	Ich fühlte mich wirklich angegriffen, verängstigt und beschämt.	ich fühlte mich wirklich angegriffen, reißt und schämt mich.	ich habe mich selbst geschätzt, herr ratspräsident.	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
Nevertheless, I do not at all conceal that he ran a strict regime there.	Trotzdem verschweige ich ja keineswegs, dass er dort ein strenges Regiment führte.	nichtsdestotrotz schätze ich überhaupt nicht, dass er dort ein strenges regime stieß.	ich habe ihn zwar stark beklagt, aber es gibt keine garantie, dass ich das recht habe.	a
The travel warning is also a response to a new Missouri law that would make it more difficult to sue a business for housing or employment discrimination.	Die Reisewarnung ist auch eine Antwort auf ein neues Missouri-Gesetz, das es schwieriger machen würde, ein Unternehmen wegen Wohnungs- oder Beschäftigungs-Diskriminierung zu verklagen.	die fahrtswarnung ist auch eine antwort auf ein neues missouritätsgesetz, das es schwieriger machen würde, ein unternehmen für wohn- oder beschäftigungsdiskriminierung gleichzusetzen.	in der tat nahmliche zirkus wird erwähnt, dass es erforderlich ist, einen neuen rechtsrahmen für den eintritt ins arbeitsleben zu schaffen.	a
He had an eye for the bigger picture while many others were busied with themselves.	Er hatte den Blick fürs Ganze, während viele andere mit sich selbst beschäftigt waren.	er hatte ein auge für das größere bild, während viele andere mit sich selbst hatten.	ein bißchen kurz davor hatten wir noch immer das angebot von	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
			menschen, die ihn in den griff bekommen haben.	

Model A Score	Model B Score	Tied Score
21	0	0



Human evaluation of experiment 5 deemed NMTs trained on synthetically trained data have superior outputs on the WMT18 test set.

Experiment 6 – Results

Experiment 6 NMT Training Charts

Figures below illustrate experiment 6 validation BLEU score which calculated during training of the model.

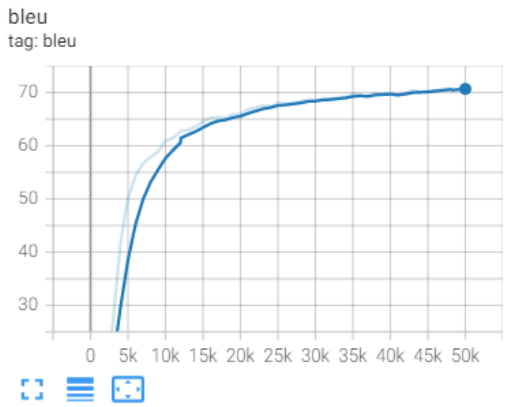


Figure 43 Experiment 6 Synthetic Dataset NMT validation BLEU score during training Single run.

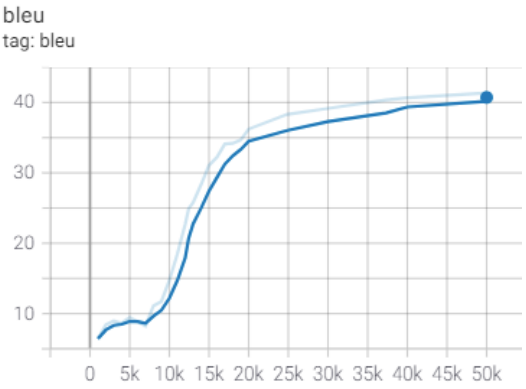


Figure 42 Experiment 6 Original Dataset NMT validation BLEU score during training Single run

Experiment 6 BLEU Score Test Results

Experiment	Corpus	Optimizer Run	Test Data Set	Original Data BLEU Score	Synthethic Data BLEU Score	Difference in score (Original data score - Synthethic data Score)
experiment6	UFAL	single_run	Withheld Corpus Data	35.52	21.24	14.28
experiment6	UFAL	single_run	wmt17	5.3	8.7	-3.4
experiment6	UFAL	single_run	wmt18	8.3	12.8	-4.5
experiment6	UFAL	single_run	wmt19	7.3	12.3	-5
experiment6	UFAL	single_run	wmt20	5.3	9.2	-3.9

Table 17 Experiment 6 BLEU Score results from Test Data Set Translations

Average difference between original data and synthetic data BLEU scores: -0.504

Standard deviation between original data and synthetic data BLEU scores: 7.41

Experiment 6 Character and word F Score results

Experiment	Corpus	Optimizer Run	Test Data Type	Original Data c6+w2-F2 Score	Synthetic Data c6+w2-F2 Score	Difference in score (Original data score - Synthethic data Score)
experiment6	UFAL	single_run	wmt17	34.1046	40.6254	-6.5208
experiment6	UFAL	single_run	wmt18	38.728	46.502	-7.774
experiment6	UFAL	single_run	wmt19	36.5335	44.5654	-8.0319
experiment6	UFAL	single_run	wmt20	33.6155	41.2286	-7.6131
experiment6	UFAL	single_run	Withheld Corpus Data	55.6192	38.8378	16.7814

Table 18:Experiment 6 character and word F Score results from Test Data Set Translations

Average difference between original data and synthetic data F scores: -2.63

Standard deviation between original data and synthetic data F scores: 9.72

Experiment 6 Human Evaluation Test Results

Results of relative ranking human evaluation performed by native German speaker on experiment 6 NMT model outputs of WMT20 test set.

Model A = NMT trained on Synthetic data.

Model B = NMT trained on Original data.

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
She said the duchess pushed them to plough ahead with their new ideas and spoke about how they have to keep pushing and going forward, knowing they are doing it for others who will follow.	Sie sagte, die Herzogin habe sie dazu gedrängt, ihre neuen Ideen zu verwirklichen, und sprach darüber, wie sie als Frauen ständig nach vorne streben müssen, im Bewusstsein, dass sie es für andere tun, die folgen werden.	sie sagte, die herzogin dränge sie, ihre neuen ideen voranzutreiben und sprach darüber, wie sie weiter voranschreiten und voranschreiten müssen, da sie wissen, dass sie es für andere tun, die folgen werden.	die herzogin hat gesagt, dass die herzogin sie <unk> hat, ihre neuen ideen voranzutreiben und darüber zu sprechen, wie sie weiter vorankommen und weitergehen müssen, wissen, dass sie es für andere tun werden, die ihr folgen werden.	A
The officers shot Clark seven times as he approached them.	Die Polizisten feuerten sieben Schüsse auf Clark ab, als er sich ihnen näherte.	die offiziere schossen siebenmal auf, als er sich ihnen näherte.	die offiziere haben clark sieben mal <unk>, als er sie angegriffen hat.	A

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
The average IPO return in 2019 is now about 9%, down from more than 30% at the end of June and more than 18% about two weeks ago.	Die durchschnittliche Emissionsrendite im Jahr 2019 liegt jetzt bei etwa 9 %, gegenüber mehr als 30 % Ende Juni und mehr als 18 % vor etwa zwei Wochen.	die durchschnittliche ipo-rendite im jahr 2019 beträgt heute etwa 9%, von mehr als 30% ende juni und mehr als 18% vor etwa zwei wochen.	im jahr 2019 beträgt die <unk> von ipo im jahr 2019 etwa 9%, von mehr als 30% ende juni und von mehr als 18% vor zwei wochen.	a
But, heartbroken at the thought of losing him also, they changed their minds and said they wanted him to be forgiven.	Doch aus Verzweiflung bei dem Gedanken, auch ihn zu verlieren, änderten sie ihre Meinung und sagten, sie wollten, dass ihm vergeben wird.	aber herzzerreißend über den gedanken, ihn zu verlieren, änderten sie ihre meinung und sagten, dass sie wollten, dass er vergeben wird.	aber herzzerbrochen an dem gedanken, ihn auch zu verlieren, sie haben ihre gedanken geändert und sagten, er wolle, dass er zu vergeben ist.	a
"I can assure you, this one is going to get solved," he said.	„Ich kann ihnen versichern, dass dieser Fall gelöst wird“, sagte er.	ich kann ihnen versichern, dass dieses hier gelöst wird, sagte er.	"ich kann ihnen versichern", sagte er.	a
"He is innocent. She was my daughter and he is my son," she said.	„Er ist unschuldig. Sie war meine Tochter und er ist mein Sohn“, sagte sie.	er ist unschuldig. sie war meine tochter und er ist mein sohn, sagte sie.	sie war meine tochter und er ist mein sohn ", sagte sie.	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
"While we already have among the strictest regimes on those products, if the evidence shows that further action is needed... we will not hesitate to take it."	„Zwar verfügen wir bereits über eine der strengsten Regelungen für diese Produkte, doch wenn die Fakten zeigen, dass weitere Maßnahmen erforderlich sind ... werden wir nicht zögern, diese einzuleiten.“	während wir bereits zu den engeren regelungen für diese produkte gehören, wenn die beweise zeigen, dass weitere maßnahmen erforderlich sind... werden wir nicht zögern, sie zu ergreifen;.	"während wir bereits unter den strikten testsystemen zu diesen produkten sind, wenn die beweise zeigen, dass weitere maßnahmen erforderlich sind, werden wir nicht zögern, sie zu ergreifen".	B
"The findings are some of the most distressing and shocking I have ever read."	„Die Ergebnisse gehören zu den erschütterndsten und schockierendsten, die ich je gelesen habe.“	die ergebnisse sind einige der belastendsten und schockierendsten ergebnisse, die ich je gelesen habe.	"die ergebnisse sind einige der verstreblichsten und <unk> chsten ergebnisse, die ich je gelesen habe".	a
Cleveland is one of 15 forces that has been recently inspected by HMICFRS inspectors, and the only one rated inadequate in all areas.	Cleveland ist eine von 15 Polizeieinheiten, die kürzlich von HMICFRS-Inspektoren inspiziert wurden, und die einzige, die in allen Bereichen als	cleveland ist eine von 15 kräften, die kürzlich von hmicfrs-inspektoren überprüft wurden, und die einzige hat sich in allen bereichen als unzureichend eingestuft.	15 kräfte, die kürzlich von den hmicfrs inspektoren kontrolliert wurden, und die einzige, die in allen bereichen unzureichend sind, ist eine der 15 kräfte,	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
	unzulänglich eingestuft wurde.		die in jüngster zeit geprüft wurden.	
When she asked United Healthcare why she was receiving a mountain of letters, she was told it was a coding issue, she said.	Als sie United Healthcare fragte, warum sie einen Berg von Briefen erhalten habe, wurde ihr gesagt, dass es sich um ein Codierungsproblem handele, sagte sie.	als sie die geeinte gesundheitsversorgung fragte, warum sie einen berg von briefen erhielt, wurde ihr gesagt, es sei ein codierendes thema, sagte sie.	als sie die united health gefragt hat, warum sie einen berg von briefen empfangen hat, wurde gesagt, dass es sich um eine kodierende frage handele, sie habe gesagt, dass es sich um eine kodierende frage handele.	a
The crowd was completely supportive, cheering the 74-year-old as he explained what had happened and how the band hoped to return to	Die Menge unterstützte den 74-Jährigen voll und ganz und jubelte ihm zu, als er erklärte, was passiert war und wie die Band hoffte, zurückzukehren, um ihr	die menge war völlig unterstützend und betrog die 74-jährigen, als er erklärte, was passiert war und wie die band hoffte, ihr engagement für die show einzuhalten.	die menge wurde vollkommen unterstützt, indem er die 74-jährigen, die er erklärt hat, was passiert ist, und wie sich das band gehofft hat, ihre	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
honour their commitment to the show.	Versprechen für die Show einzuhalten.		zusage an die show zurückzugeben.	
"I swear to you we will come back and we will honour your tickets. I love you so much and I'm sorry to let you down."	„Ich schwöre euch, dass wir zurückkommen werden und ihr eure Tickets einlösen könnt. Ich liebe euch sehr und es tut mir leid, euch zu enttäuschen.“	ich schwöre dir, dass wir zurückkommen werden und wir deine tickets. ich liebe dich so sehr und es tut mir leid, dich runterlassen zu lassen;...	"ich <unk> dir, wir kommen zurück, und wir werden eure titten <unk>, und ich liebe dich so sehr und es tut mir leid, dich runter zu lassen".	a
The automaker is expected to report its quarterly vehicle deliveries in the next few days.	Es wird erwartet, dass der Automobilhersteller seine Quartalsauslieferungen in den nächsten Tagen melden wird.	es wird erwartet, dass der automaker seine vierteljährlichen fahrzeuglieferungen in den nächsten tagen melden wird.	der maschinenführer wird voraussichtlich in den nächsten tagen seine vierteljährlichen fahrzeuglieferungen in den nächsten tagen <unk>.	a

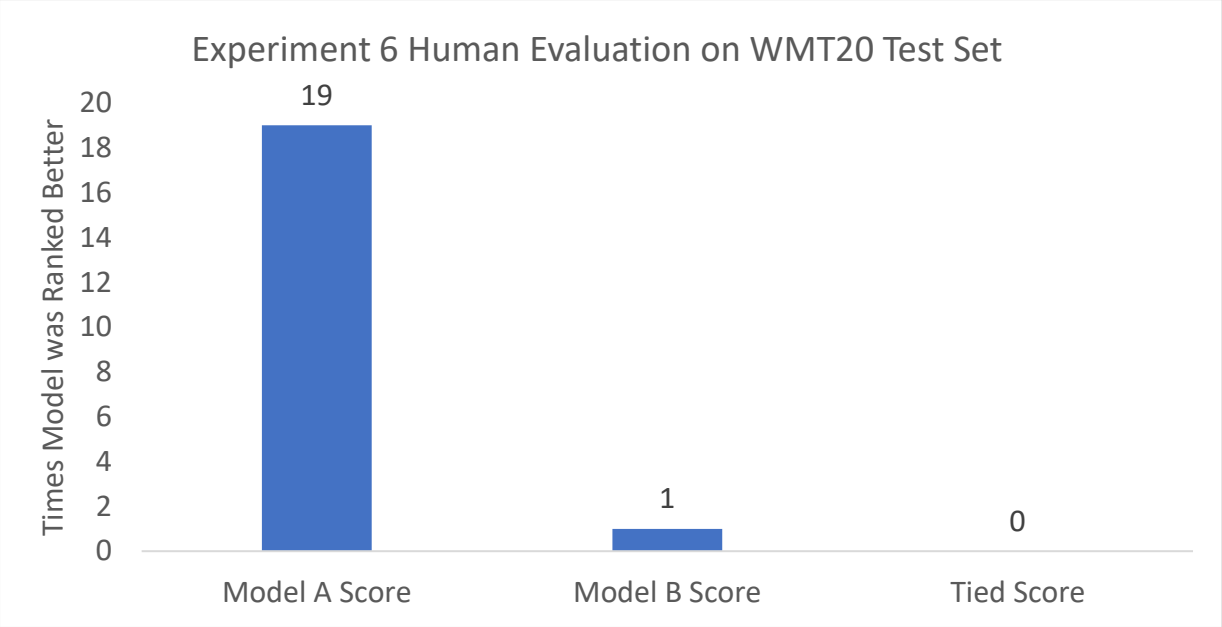
Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
She added that having Meghan present 'didn't feel hierarchical', as it was 'women gathered together and a talk about the struggles we have, as well as the things we need to do to move forward and grow and change our societies'.	Sie fügte hinzu, dass sich die Anwesenheit von Meghan „nicht hierarchisch anfühlte“, da es sich um „eine Versammlung von Frauen und ein Gespräch über die Kämpfe, die wir haben, sowie über die Dinge, die wir tun müssen, um vorwärts zu kommen und zu wachsen und unsere Gesellschaften zu verändern“ handelte.	sie fügte hinzu, dass es sich bei der anwesenheit von meghan um meghan handelte, sich hierarchische unterzuckerung fühlte, da es sich um eine rolle handelte; frauen versammelten sich zusammen und sprachen über die kämpfe, die wir haben, sowie über die dinge, die wir tun müssen, um voranzukommen und unsere gesellschaften zu vergrößern und zu verändern;	sie fügte hinzu, dass sie meghan heute nicht "gefühlte" gefühlte, da es "frauen waren, die sich zusammengeschlossen haben, und sprach über die stru<unk>, die wir haben, sowie über die dinge, die wir tun müssen, um unsere gesellschaften voranzutreiben und zu vergrößern und zu ändern".	a
But Mr Puglia insists he was just doing his job.	Aber Joseph Puglia besteht darauf, dass er nur seinen Job gemacht hat.	aber herr puglia besteht darauf, dass er gerade seinen job gemacht hat.	aber herr puglia bestand darauf, dass er seinen job gemacht hat.	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
Mr Haile-Gabriel stated that most stray animals carried rabies and re-echoed the need for the country to supplement the United against Rabies 2030 strategy that requires the pet owners to vaccinate animals that carried zoonotic diseases.	Herr Haile-Gabriel erklärte, dass die meisten streunenden Tiere die Tollwut tragen und wiederholte die Notwendigkeit, dass das Land die „United against Rabies 2030“-Strategie ergänzt, die von den Haustierbesitzern verlangt, Tiere zu impfen, die Zoonoseerkrankungen tragen.	herr haile-gabriel erklärte, dass die meisten streunenden tiere tollwut beförderten und wiederholte die notwendigkeit, dass das land die strategie gegen tollwut 2030 ergänzt, die von den heimtierbesitzern verlangt wird, tiere zu impfen, die zoonotische krankheiten beförderten.	haile-gabriel hat gesagt, dass die meisten erdbebentiere rabies mitgebracht und die notwendigkeit für das land erneut erkannt haben, die vereinigten staaten gegen rabies 2030 strategie zu ergänzen, die es dem haustier erfordert, tiere, die zoonotische krankheiten tragen, zu <unk>.	a
He is in hospital in a critical but stable condition after suffering serious burns.	Er ist nun in einem kritischen, aber stabilen Zustand im Krankenhaus, nachdem er sich schwere Verbrennungen zugezogen hatte.	er ist in einem kritischen, aber stabilen zustand im krankenhaus, nachdem er schwere verbrennungen erleidet hat.	er ist in einem kritischen, aber stabilen zustand, nachdem er schwere verbrennungen <unk> hat.	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
However, whether a murder is defined as a crime of honour is left to the judge's discretion, meaning that killers can theoretically claim a different motive and still be pardoned.	Es liegt jedoch im Ermessen des Richters, ob ein Mord als Ehrenverbrechen definiert wird, was bedeutet, dass Mörder theoretisch ein anderes Motiv geltend machen und trotzdem begnadigt werden können.	ob jedoch ein mord als ehrenverbrechen definiert ist, bleibt dem ermessensspielraum des richters überlassen, was bedeutet, dass mörder theoretisch ein anderes motiv beanspruchen und trotzdem begnadigt werden können.	ob jedoch ein mord als verbrechen der ehre definiert ist, ist der diskretion des richters überlassen, was bedeutet, dass mörder theoretisch ein anderes motiv verlangen können und immer noch pardoniert werden können.	a
She added that there is still effort to reach the 14 per cent of the people, who are living with HIV but do not know their HIV status.	Sie fügte hinzu, es gebe nach wie vor Bemühungen, die 14 Prozent der Menschen zu erreichen, die mit HIV leben, aber ihren HIV-Status nicht kennen.	sie fügte hinzu, dass es immer noch bemühungen gibt, die 14 prozent der menschen zu erreichen, die mit hiv leben, aber ihren hiv-status nicht kennen.	hinzu kommt, dass sie nach wie vor anstrengungen unternimmt, um die 14 prozent der menschen, die mit hiv leben, zu erreichen, aber ihren hiv-status nicht kennen.	a

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
"I'm really sorry this has happened to you guys tonight... I don't think there is much point trying to put anything together that makes any sense because this is Roger's show and he's not in good shape."	„Es tut mir wirklich leid, dass das heute Abend gerade euch passiert ist ... Ich glaube nicht, dass es viel bringt, etwas zusammenzuschustern, das Sinn macht, denn das ist Rogers Show und er ist nicht in guter Verfassung.“	ich tut mir wirklich leid, dass dir das heute abend passiert ist... ich glaube nicht, dass es viel sinn hat, irgendetwas zusammenzubringen, was sinnvoll ist, weil es roger tollinavirs show ist und er es nicht in guter form ist.	"es tut mir wirklich leid, dass euch heute abend das passiert ist... ich glaube nicht, dass es viel sinn gibt, alles miteinander zu verbinden, denn roger zeigt es wirklich, dass er nicht gut ist".	a

Model A Score	Model B Score	Tied Score
19	1	0



Human evaluation of experiment 6 deemed NMTs trained on synthetically trained data have superior outputs on the WMT20 test set.

Experiment 7 – Results

Experiment 7 NMT Training Charts

Figures below illustrate experiment 7 validation BLEU score which calculated during training of the model.

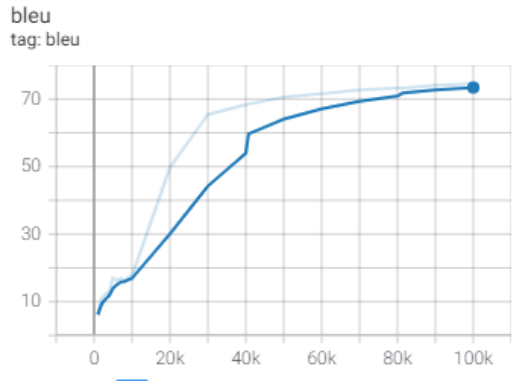


Figure 45 Experiment 7 Synthetic Dataset NMT validation BLEU score during training optimizer run 1

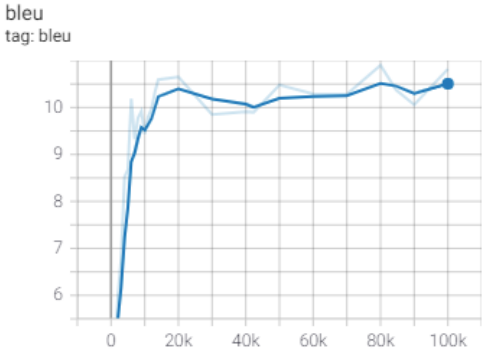


Figure 44 Experiment 7 Original Dataset NMT validation BLEU score during training optimizer run 1

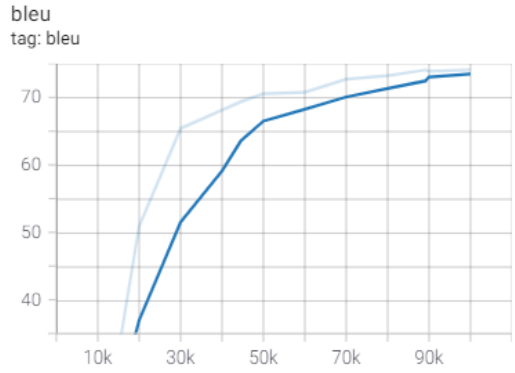


Figure 47 Experiment 7 Synthetic Dataset NMT validation BLEU score during training optimizer run 2

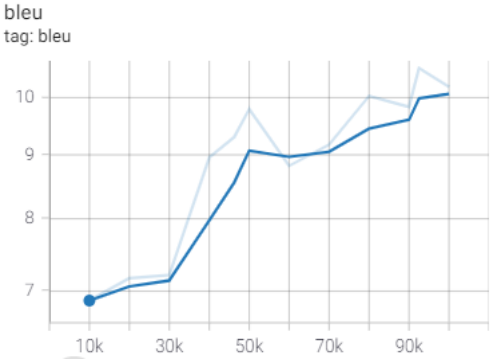


Figure 46 Experiment 7 Original Dataset NMT validation BLEU score during training optimizer run 2

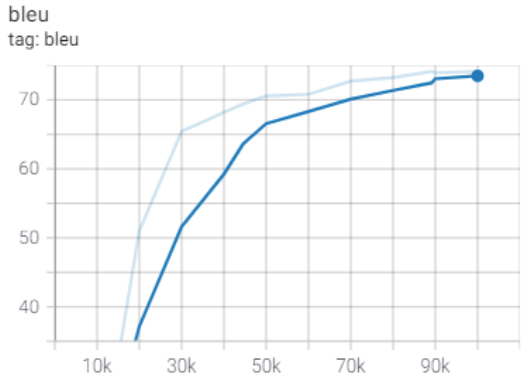


Figure 49 Experiment 7 Synthetic Dataset NMT validation BLEU score during training optimizer run 3

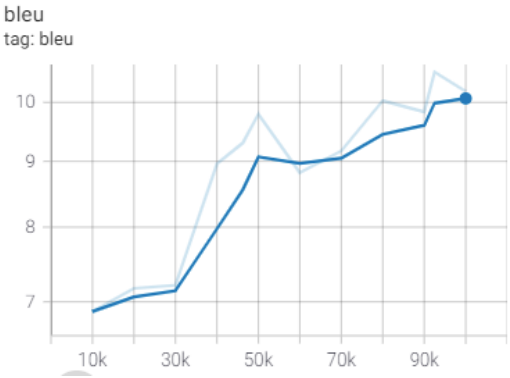


Figure 48 Experiment 7 Original Dataset NMT validation BLEU score during training optimizer run 3

Experiment 7 BLEU Score Test Results

Experiment	Corpus	Optimizer Run	Test Data Set	Original Data BLEU Score	Synthetic Data BLEU Score	Difference in score (Original data score - Synthetic data Score)
experiment7	UFAL	optimizer_run1	Withheld Corpus Data	6.23	14.06	-7.83
experiment7	UFAL	single_run	wmt17	0.3	8.7	-8.4
experiment7	UFAL	single_run	wmt18	0.4	12.7	-12.3
experiment7	UFAL	single_run	wmt19	0.7	11.9	-11.2
experiment7	UFAL	optimizer_run1	wmt20	0.3	9.5	-9.2
experiment7	UFAL	optimizer_run2	Withheld Corpus Data	6.78	13.78	-7
experiment7	UFAL	optimizer_run2	wmt20	0.4	9.7	-9.3
experiment7	UFAL	optimizer_run3	Withheld Corpus Data	6.78	13.78	-7
experiment7	UFAL	optimizer_run3	wmt20	0.4	9.7	-9.3

Table 19 Experiment 7 BLEU score results from Test Data Set Translations

Average difference between original data and synthetic data BLEU scores: -9.06

Standard deviation between original data and synthetic data BLEU scores: 1.69

Experiment 7 Character and word F Score results

Experiment	Corpus	Optimizer Run	Test Data Type	Original Data c6+w2-F2 Score	Synthetic Data c6+w2-F2 Score	Difference in score (Original data score - Synthethic data Score)
experiment7	UFAL	single_run	wmt17	15.2669	41.485	-26.2181
experiment7	UFAL	single_run	wmt18	15.1969	47.2741	-32.0772
experiment7	UFAL	single_run	wmt19	14.8276	45.2199	-30.3923
experiment7	UFAL	optimizer_run1	wmt20	13.2791	42.5597	-29.2806
experiment7	UFAL	optimizer_run1	Withheld Corpus Data	18.1014	27.9949	-9.8935
experiment7	UFAL	optimizer_run2	wmt20	13.2013	42.1123	-28.911
experiment7	UFAL	optimizer_run2	Withheld Corpus Data	16.785	27.8322	-11.0472
experiment7	UFAL	optimizer_run3	wmt20	13.2013	42.1123	-28.911
experiment7	UFAL	optimizer_run3	Withheld Corpus Data	16.785	27.8322	-11.0472

Table 20 Experiment 7 character and word F score results from Test Data Set Translations

Average difference between original data and synthetic data F scores: -23.1

Standard deviation between original data and synthetic data F scores: 8.91

Experiment 7 Statistical Significance Test

Baseline represents model trained on original data.

System 1 represents model trained on synthetic data.

Withheld Corpus Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	11.0 (0.1/0.3/-)	15.7 (0.1/0.4/-)	85.5 (0.4/9.6/-)	94.7 (0.5/15.7/-)
system 1	30.4 (0.2/0.2/0.0001)	27.5 (0.1/0.2/0.0001)	55.1 (0.2/0.2/0.0001)	95.8 (0.1/0.6/0.0001)

WMT20 Test Dataset Descriptive Statistics and P-values

n=3	BLEU (s_sel/s_opt/p)	METEOR (s_sel/s_opt/p)	TER (s_sel/s_opt/p)	Length (s_sel/s_opt/p)
baseline	0.5 (0.1/0.0/-)	5.3 (0.1/0.1/-)	100.6 (0.7/5.4/-)	73.4 (1.2/4.7/-)
system 1	23.5 (0.4/0.1/0.00)	25.6 (0.2/0.2/0.00)	62.4 (0.5/2.4/0.00)	98.8 (0.6/4.2/0.00)

Experiment 7 Human Evaluation Test Results

Results of relative ranking human evaluation performed by native German speaker on experiment 7 NMT model outputs of WMT20 test set.

Model A = NMT trained on Synthetic data.

Model B = NMT trained on Original data.

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
But, heartbroken at the thought of losing him also, they changed their minds and said they wanted him to be forgiven.	Doch aus Verzweiflung bei dem Gedanken, auch ihn zu verlieren, änderten sie ihre Meinung und sagten, sie wollten, dass ihm vergeben wird.	aber herzzerreißend bei dem gedanken, ihn auch zu verlieren, änderten sie ihre meinung und sagten, sie wollten, dass ihm vergeben wird.	er sagte, er sei in der lage, die dinge zu ändern, aber...	A
Cleveland is one of 15 forces that has been recently inspected by HMICFRS inspectors, and the only one rated inadequate in all areas.	Cleveland ist eine von 15 Polizeieinheiten, die kürzlich von HMICFRS-Inspektoren inspiziert wurden, und die einzige, die in allen Bereichen als unzulänglich eingestuft wurde.	cleveland ist eine von 15 kräften, die kürzlich von hmicfrs-inspektoren kontrolliert wurde, und die einzige, die in allen bereichen unzureichend bewertet wurde.	es gibt keine anzeichen dafür, dass sich die europäische union in der lage sieht, sich mit der frage der sicherheit und des gesundheitsschutzes auseinanderzusetzen.	A

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
She said the duchess pushed them to plough ahead with their new ideas and spoke about how they have to keep pushing and going forward, knowing they are doing it for others who will follow.	Sie sagte, die Herzogin habe sie dazu gedrängt, ihre neuen Ideen zu verwirklichen, und sprach darüber, wie sie als Frauen ständig nach vorne streben müssen, im Bewusstsein, dass sie es für andere tun, die folgen werden.	sie sagte, die herzogin drängte sie, ihre neuen ideen voranzutreiben und sprach darüber, wie sie weiter vorankommen und voranschreiten müssen, wohl wissend, dass sie es für andere tun, die folgen werden.	sie wissen, wie wichtig es ist, dass sie sich an die regeln halten, und zwar nicht nur an die stelle der kommission, sondern auch an die stelle der kommission.	A
The officers shot Clark seven times as he approached them.	Die Polizisten feuerten sieben Schüsse auf Clark ab, als er sich ihnen näherte.	die offiziere erschossen clark siebenmal, als er sich ihnen näherte.	er hat die sieben polizisten angegriffen.	A
In 2018 after he held up the placard calling for Xi's resignation, he wrote online of how police stormed into his home, ordering him to write a confession letter and a statement promising he would stop.	Im Jahre 2018, nachdem er das Plakat mit der Aufforderung zum Rücktritt von Xi hochgehalten hatte, schrieb er online darüber, wie die Polizei in sein Haus stürmte und ihm den Befehl gab, ein Geständnis und eine Stellungnahme zu schreiben, in	2018, nachdem er die placard hochgehalten hatte, in der er den rücktritt des ausrufs gefordert hatte, schrieb er online darüber, wie die polizei in sein haus stürzte, und befahl ihm, einen geständnisbrief und eine erklärung zu schreiben, in der er versprach, er würde aufhören.	er sagte, dass er sich in der lage sieht, sich mit der polizei zu treffen,..... und dass er sich von der polizei erkundigt hat.	A

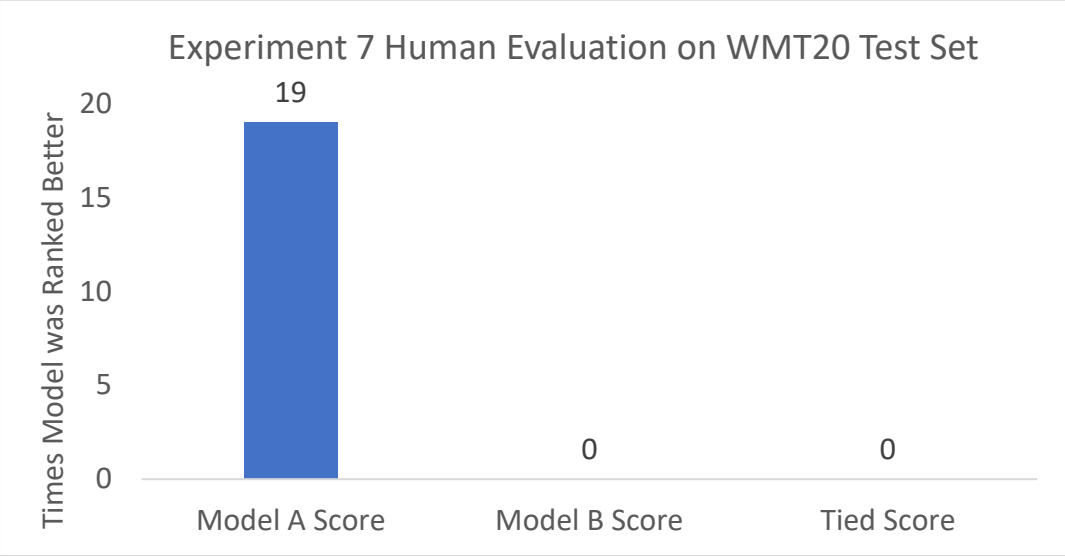
Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
	der er versprach, dass er aufhören würde.			
But Mr Puglia insists he was just doing his job.	Aber Joseph Puglia besteht darauf, dass er nur seinen Job gemacht hat.	aber mr puglia besteht darauf, dass er gerade seine arbeit gemacht hat.	aber ich habe den eindruck, dass er sich in der lage sieht, mit ihnen zusammenzuarbeiten.	A
He then tied her to a tree and sexually assaulted her.	Dann fesselte er sie an einen Baum und missbrauchte sie sexuell.	dann fesselte er sie an einen baum und griff sie sexuell an.	und sie hat ein mädchen gefangen.	A
"We decided not to put her in that position so I never asked."	„Wir haben uns entschieden, sie nicht in diese Situation zu bringen, also habe ich sie nie gefragt.“	wir haben beschlossen, sie nicht in diese position zu bringen, also habe ich nie darum gebeten, sie zu zitieren?	ich habe mich nicht gefragt, ob wir uns in die luft jagen können.	A
When she asked United Healthcare why she was receiving a mountain of letters, she was told it was a coding issue, she said.	Als sie United Healthcare fragte, warum sie einen Berg von Briefen erhalten habe, wurde ihr gesagt, dass es sich um ein Codierungsproblem handele, sagte sie.	als sie die gemeinsame gesundheitsversorgung fragte, warum sie einen berg briefe erhielt, wurde ihr gesagt, es sei ein codierungsthema, sagte sie.	sie hat mich gefragt, ob es eine gute idee ist, sie zu töten.	A
The automaker is expected to report its quarterly vehicle deliveries in the next few days.	Es wird erwartet, dass der Automobilhersteller seine	es wird erwartet, dass der automaker seine vierteljährlichen	die kommission hat im laufe des jahres einen bericht über den stand der technik veröffentlicht.	A

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
	Quartalsauslieferungen in den nächsten Tagen melden wird.	fahrzeuglieferungen in den nächsten tagen meldet.		
Brexit can get plenty more toxic from here.	Brexit kann jetzt noch viel giftiger werden.	brexit kann von hier aus viel giftiger werden.	hier ist ein bißchen vergiftet.	A
Why viruses like Herpes and Zika will need to be reclassified: Biotech impact	Wieso Viren wie Herpes und Zika neu klassifiziert werden müssen: Biotechnische Auswirkungen	warum viren wie herpes und zika neu klassifiziert werden müssen: biotechnologische auswirkungen	wie ist bristol aufzubewahren?	A
He is in hospital in a critical but stable condition after suffering serious burns.	Er ist nun in einem kritischen, aber stabilen Zustand im Krankenhaus, nachdem er sich schwere Verbrennungen zugezogen hatte.	er befindet sich in einem kritischen, aber stabilen zustand im krankenhaus, nachdem er schwere verbrennungen erlitten hat.	er ist in der lage, sich in den ruhestand zu setzen, und zwar unter anderem durch die tatsache, dass er in den ruhestand getreten ist.	A
She added that there is still effort to reach the 14 per cent of the people, who are living with HIV but do not know their HIV status.	Sie fügte hinzu, es gebe nach wie vor Bemühungen, die 14 Prozent der Menschen zu erreichen, die mit HIV leben, aber ihren HIV-Status nicht kennen.	sie fügte hinzu, dass es immer noch anstrengungen gibt, die 14 prozent der menschen zu erreichen, die mit hiv leben, aber ihren hiv-status nicht kennen.	es gibt keine anzeichen dafür, dass sich die eu mit dem problem auseinandersetzen wird, aber es gibt auch keine anzeichen dafür, dass sie mit dem problem konfrontiert ist.	A

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
She added that having Meghan present 'didn't feel hierarchical', as it was 'women gathered together and a talk about the struggles we have, as well as the things we need to do to move forward and grow and change our societies'.	Sie fügte hinzu, dass sich die Anwesenheit von Meghan „nicht hierarchisch anfühlte“, da es sich um „eine Versammlung von Frauen und ein Gespräch über die Kämpfe, die wir haben, sowie über die Dinge, die wir tun müssen, um vorwärts zu kommen und zu wachsen und unsere Gesellschaften zu verändern“ handelte.	sie fügte hinzu, dass wir uns, nachdem wir meghan anwesend gewesen seien, hierarchisch nicht hierarchisch fühlten, wie es war; frauen versammelten sich zusammen und sprachen über die kämpfe, die wir haben, sowie über die dinge, die wir tun müssen, um voranzukommen und unsere gesellschaften zu wachsen und zu verändern;	& quot; in der zwischenzeit haben wir eine gute idee, und zwar aus der gruppe, die aus folgendem besteht?	A
This is the heroic moment a police officer ran out onto a busy motorway to rescue an injured dog.	Das ist der heroische Moment, in dem ein Polizist auf eine dicht befahrene Straße läuft, um einen verletzten Hund zu retten.	dies ist der heroische moment, in dem ein polizist auf eine beschäftigte autobahn rannte, um einen verletzten hund zu retten.	das ist eine gute nachricht, dass er nach hause kam.	A
"I wish I could trade places with you son, I wish I could trade places with you," Guy Alford, the victim's father, said.	„Ich wünschte, ich könnte mir dir tauschen, mein Sohn, ich wünschte, ich könnte mit dir tauschen“, sagte der Vater des Opfers, Guy Alford.	ich wünschte, ich könnte orte mit dir tauschen, ich wünschte, ich könnte orte mit dir tauschen;, guy alford, der vater des opfers, sagte er.	ich habe mich gefragt, ob du mit mir zusammen bist, aber...	A

Source_Sentence	Reference_Sentence	Model_A_Hypothesis	Model_B_Hypothesis	Better Translation (Model A or Model B)
"The Undergraduate Awards are fiercely contested by thousands of students around the world and it is a great reflection of the quality of our students, and the education they receive here, that Dundee is so strongly represented amongst the prize winners.	„Die Undergraduate Awards werden von tausenden von Studenten überall auf der Welt hart umkämpft, und es zeugt von der Qualität unserer Studenten und der Bildung, die sie hier erhalten, dass Dundee bei den Preisträgern so stark vertreten ist.	die absolventen auszeichnungen werden von tausenden studenten auf der ganzen welt heftig angefochten, und es ist ein großer spiegelstrich der qualität unserer studenten und der ausbildung, die sie hier erhalten, dass dundee unter den preisträgern so stark vertreten ist.	& quot; in der zwischenzeit gibt es eine große zahl von personen, die sich an der spitze des projekts befinden, und zwar nicht nur an der spitze des projekts, sondern auch an der spitze des projekts & quot;.	A
They used a generalization of the quasiequivalence principle to see how proteins can wrap around an icosahedral capsid.	Sie verwendeten eine Verallgemeinerung des Prinzips der Quasi-Äquivalenz, um zu sehen, wie sich Proteine um ein ikosaedrischer Kapsid wickeln können.	sie verwendeten eine verallgemeinerung des quasiäquivalenzprinzips, um zu sehen, wie proteine sich um einen icosahedralen capsid wickeln können.	so ist es möglich, dass sich die gfs auf eine bestimmte art und weise mit der frage befasst.	A

Model A Score	Model B Score	Tied Score
19	0	0



Human evaluation of experiment 7 deemed NMTs trained on synthetically trained data have superior outputs on the WMT20 test set.

Statistical Significance Test Applied to Combined Experimental Results

All Scores Combined - Shapiro Wilk Test for Normality Results

As the experiments had significant variations between experiments it is assumed that metric scores are not normally distributed. To validate this assumption a Shapiro-Wilk test was performed on the combined score results of all experiments. Results of Shapiro-Wilk test are shown below.

```
Shapiro-Wilk test for normality on charcter and word F scores from orignal data.
Statistics0.940, p=0.0147731788
Orginal dataset scores  do not look normally distributed (reject H0)

Shapiro-Wilk test for normality on scharcter and word F scores from synthethic data.
Statistics=0.961, p=0.0998097658
Synthethic dataset scores looks normally distributed (fail to reject H0)

Shapiro-Wilk test for normality on original data scores.
Statistics0.811, p=0.0000019104
Orginal dataset scores  do not look normally distributed (reject H0)

Shapiro-Wilk test for normality on synthethic data scores.
Statistics=0.823, p=0.0000037173
Synthethic dataset scores  do not look normally distributed (reject H0)
```

All BLUE Scores Combined - Wilcoxon Signed Rank Test Tailed Hypothesis Test Results

Results

```
Wilcoxon signed rank test: Two-Tailed Hypothesis Test
null hypothesis = median of the original data is zero against the synthethic data
null hypothesis rejected
Original and Synthethic data have different distributions
Statistics: t=95.000, p=0.0000003
Original data mean: 10.63, standard deviation 10.39
Synthetic data mean: 14.48, standard deviation 9.40
```

All BLEU Scores Combined Except UFAL- Wilcoxon Signed Rank Test Tailed Hypothesis Test Results

```
Wilcoxon signed rank test: Two-Tailed Hypothesis Test
null hypothesis = median of the original data is zero against the synthethic data
null hypothesis rejected
Original and Synthethic data have different distributions
Statistics: t=33.000, p=0.00004
Original data mean: 14.05, standard deviation 10.25
Synthetic data mean: 16.74, standard deviation 11.09
```

All character and word F Scores Combined - Wilcoxon Signed Rank Test Tailed Hypothesis Test Results

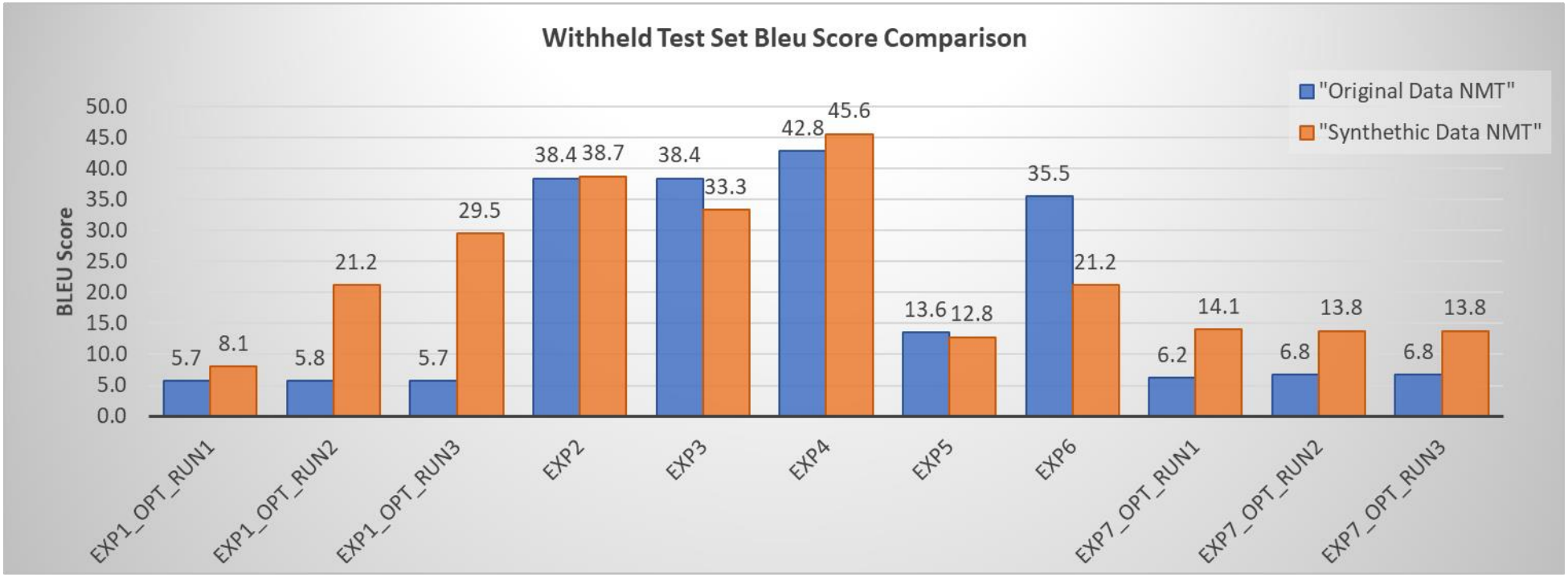
```
Wilcoxon signed rank test: Two-Tailed Hypothesis Test
null hypothesis = median of the original data is zero against the synthethic data
null hypothesis rejected
Original and Synthethic data have different distributions
Statistics: t=95.000, p=0.0000003
Original data mean: 34.17, standard deviation 12.88
Synthetic data mean: 41.49, standard deviation 8.48
```

All character and word F Scores Combined Except UFAL- Wilcoxon Signed Rank Test Tailed Hypothesis Test Results

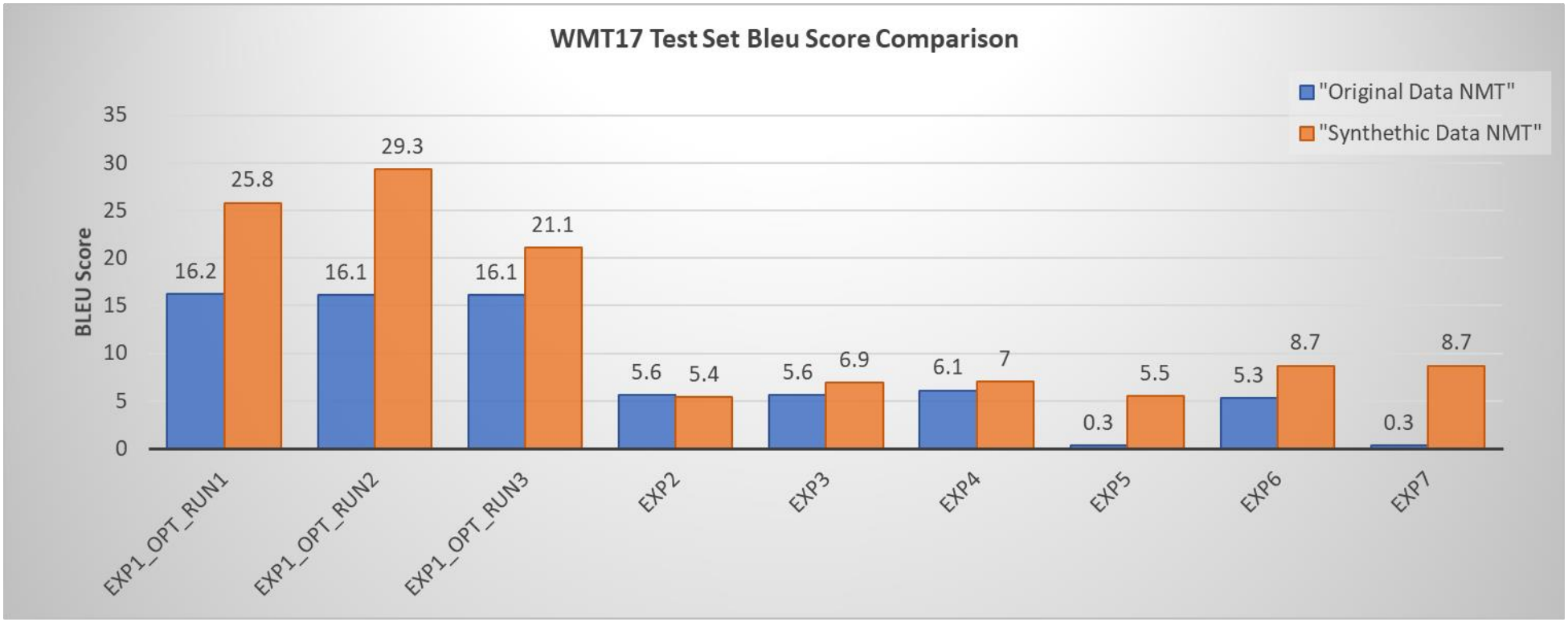
```
Wilcoxon signed rank test: Two-Tailed Hypothesis Test
null hypothesis = median of the original data is zero against the synthethic data
null hypothesis rejected
Original and Synthethic data have different distributions
Statistics: t=31.000, p=0.00003
Original data mean: 41.18, standard deviation 8.07
Synthetic data mean: 44.12, standard deviation 8.57
```

Comparison of Test Results Across Experiments

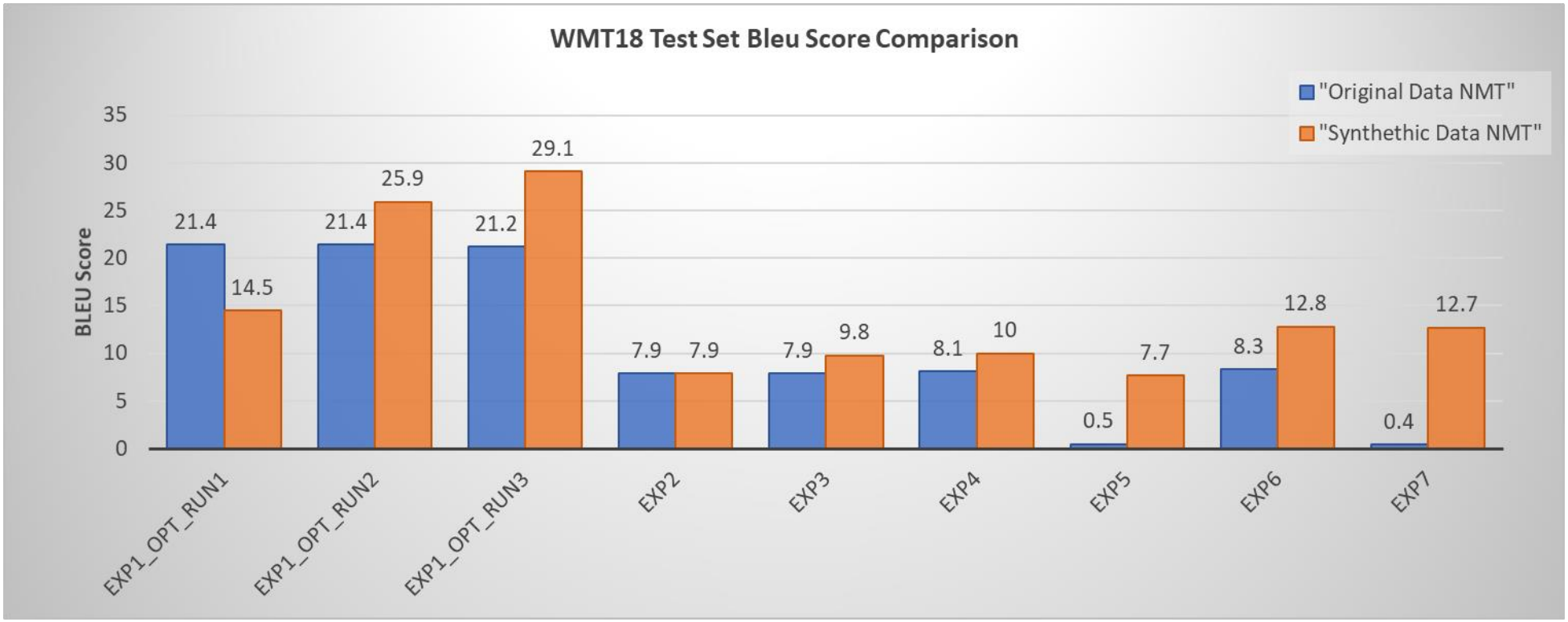
Comparison of Withheld Data Test Set BLEU scores across experiments



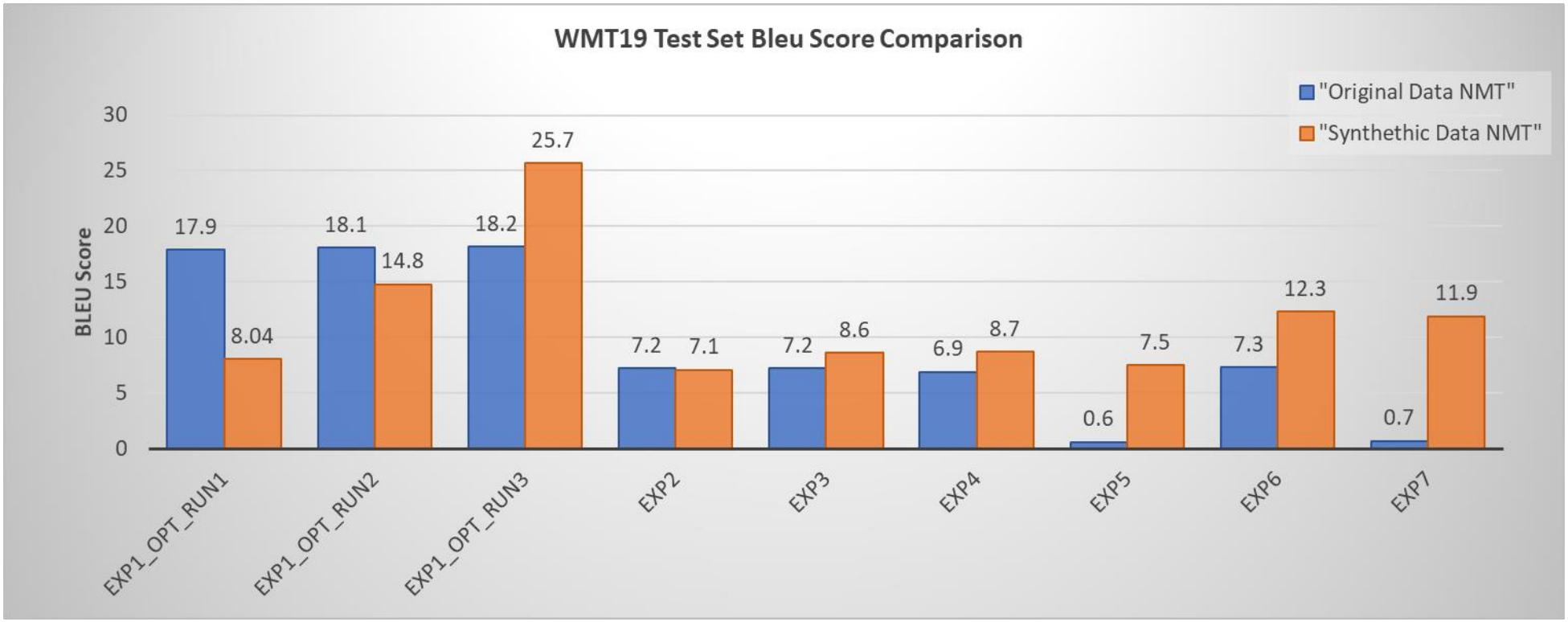
Comparison of WMT17 Test Set BLEU scores across experiments



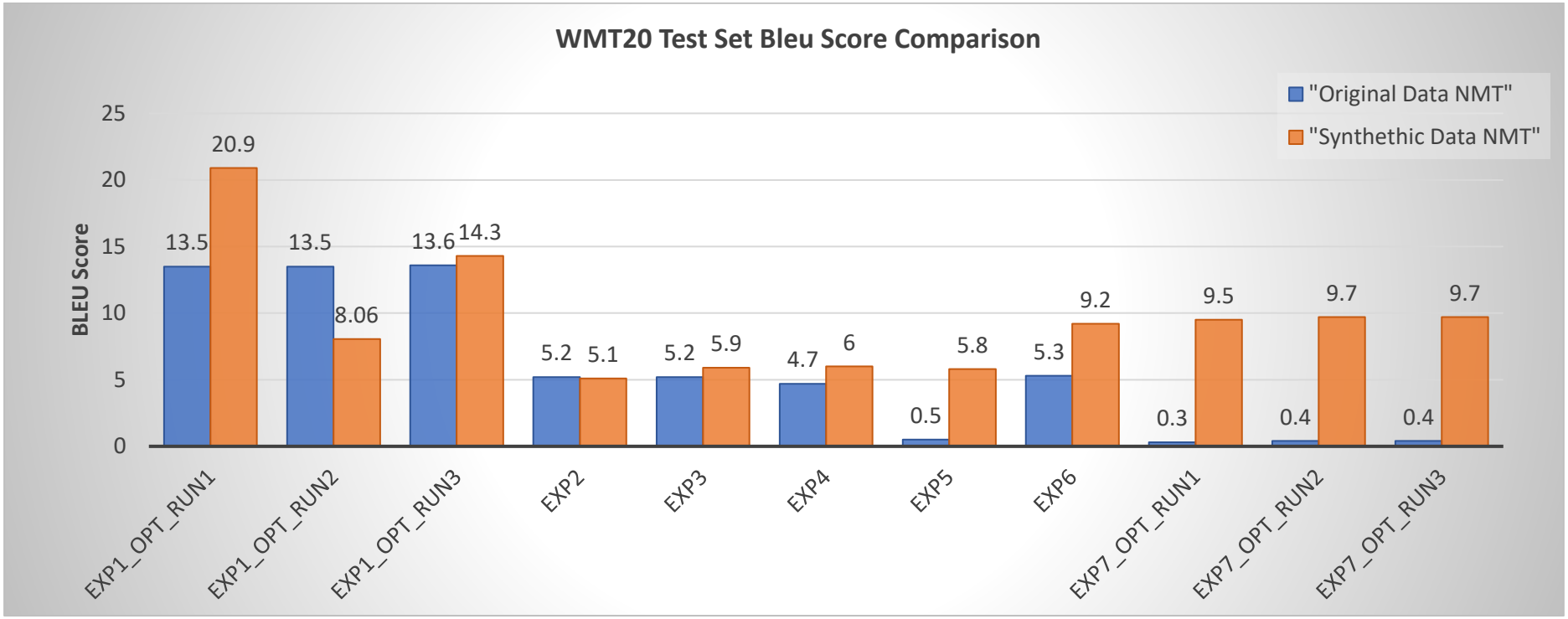
Comparison of WMT18 Test Set BLEU scores across experiments



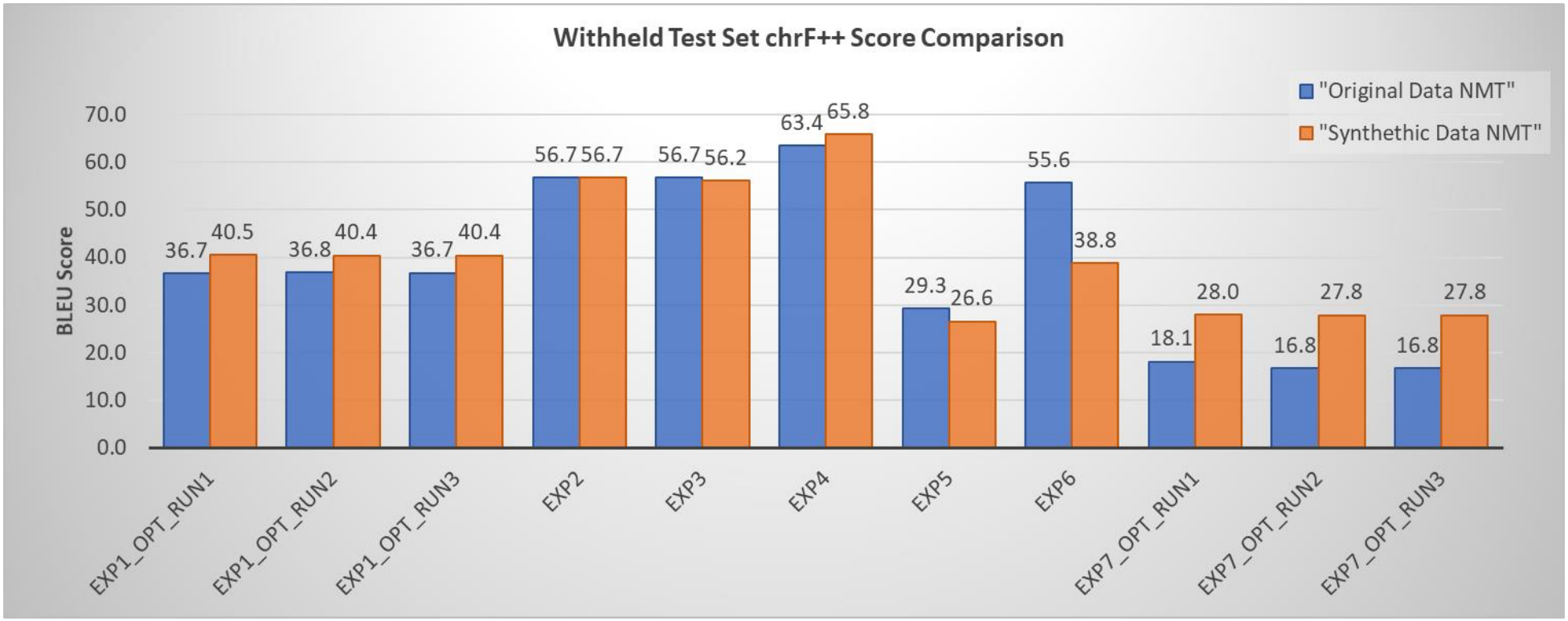
Comparison of WMT19 Test Set BLEU scores across experiments



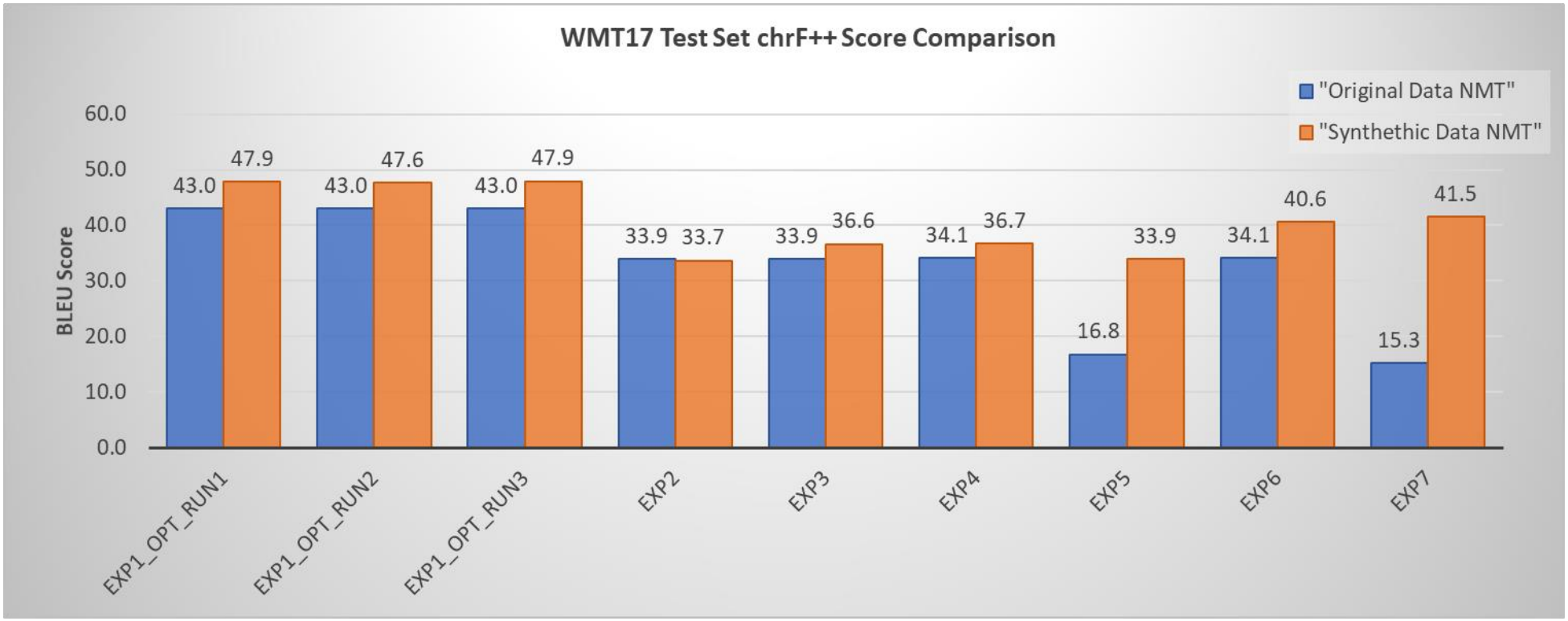
Comparison of WMT20 Test Set BLEU scores across experiments



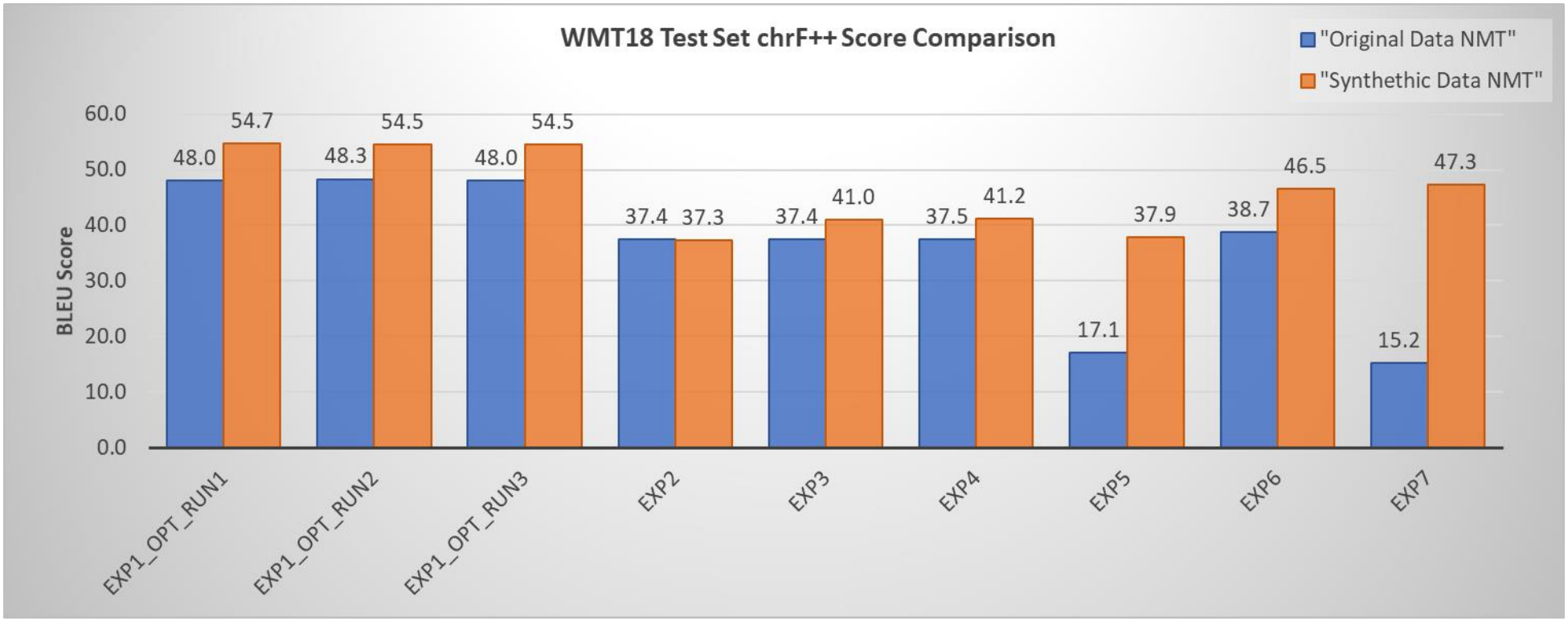
Comparison of Withheld Data Test Set chrF++ scores across experiments



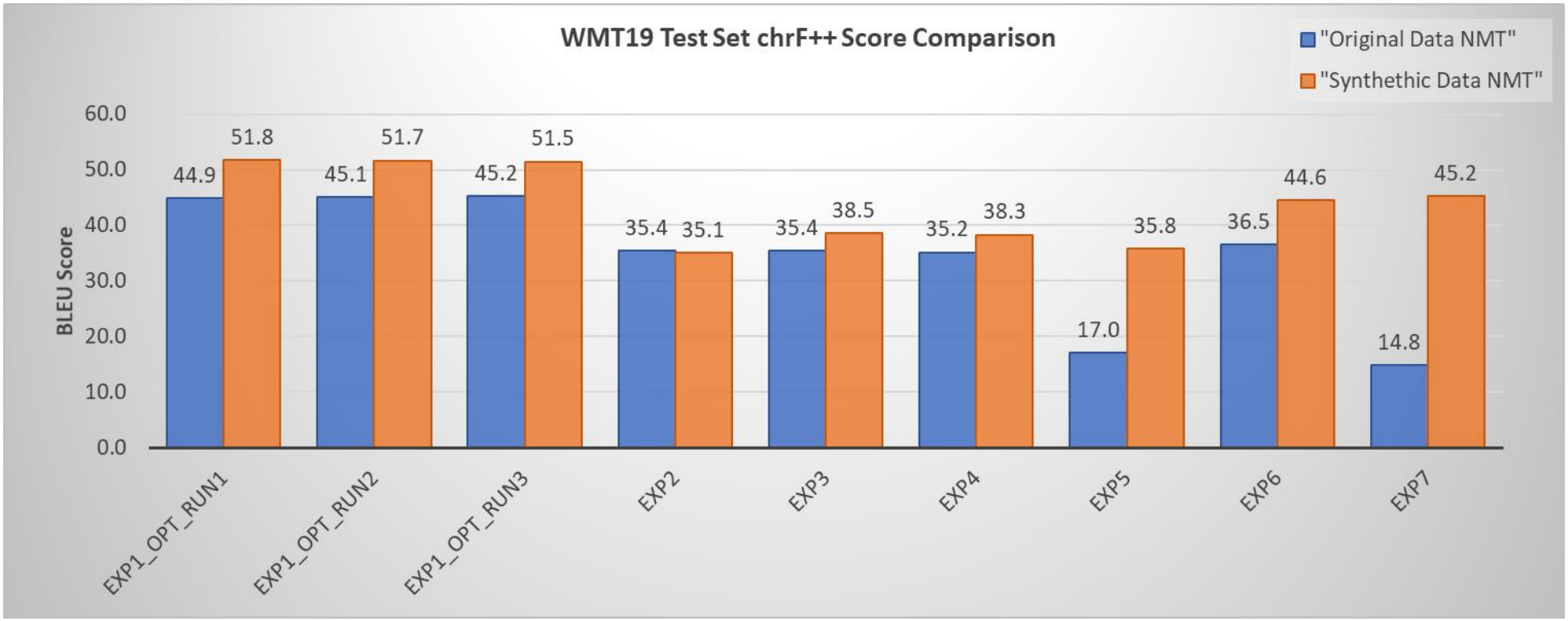
Comparison of WMT17 Test Set chrF++ scores across experiments



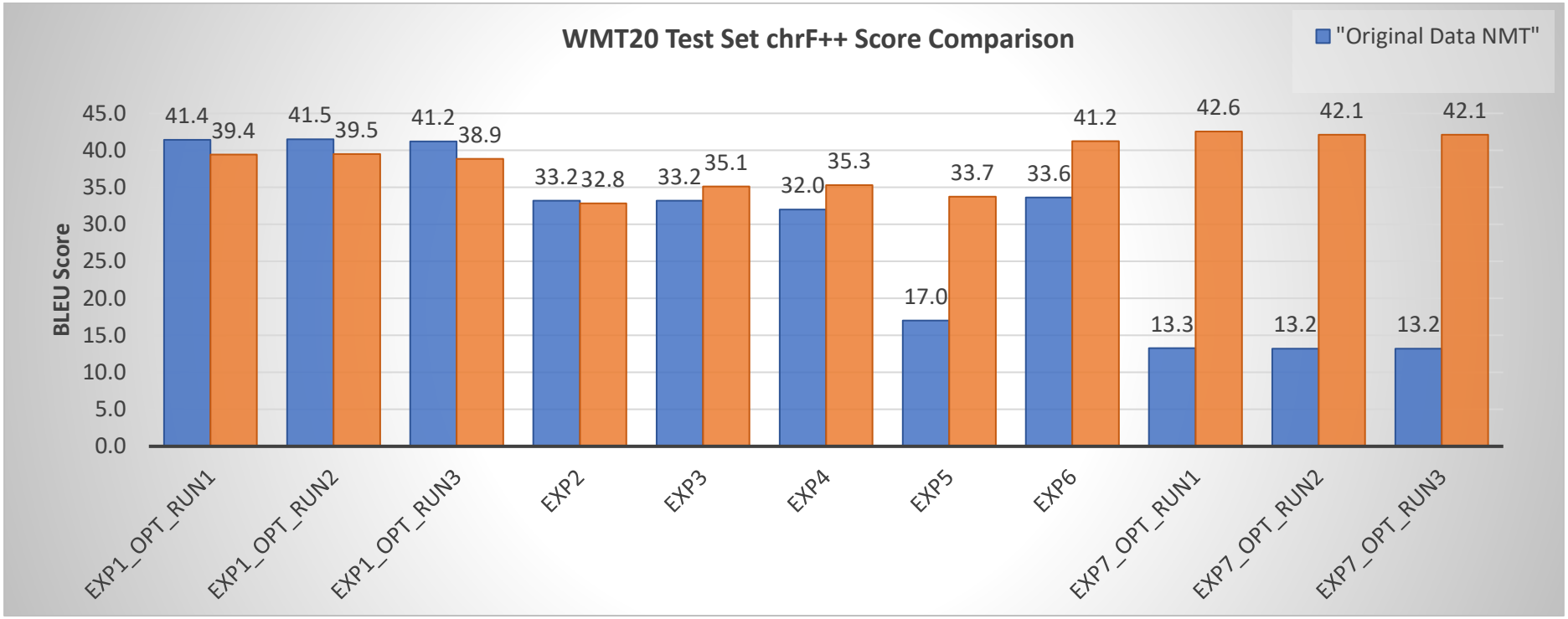
Comparison of WMT18 Test Set chrF++ scores across experiments



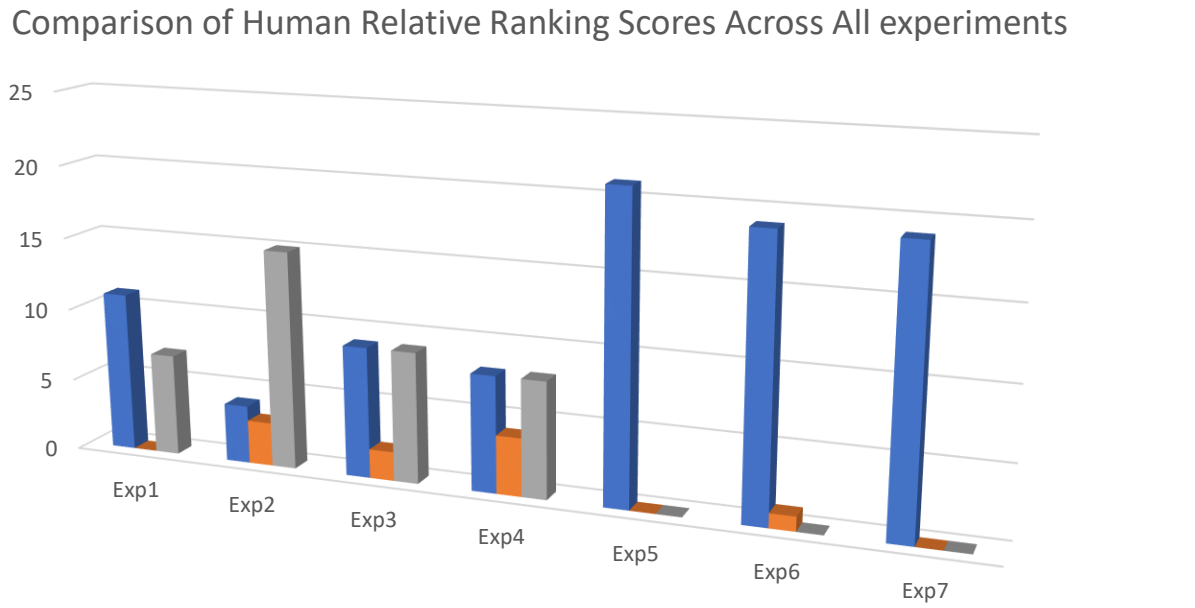
Comparison of WMT19 Test Set chrF++ scores across experiments



Comparison of WMT20 Test Set chrF++ scores across experiments



Comparison of Human Evaluation Relative Ranking scores across experiments



	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7
Synthetic Model Score	11	4	9	8	21	19	19
Original Model Score	0	3	2	4	0	1	0
Tied Score	7	15	9	8	0	0	0

Samples of Sentences Used to Train Models.

Experiment 1: Sample of Sentences Used to Train Models.

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
But it was. Specifically, what was broken was the story assigning value to mortgage-backed securities and other derivatives based on unrepayable loans.	Aber sie war es. Im Besonderen bestand der Bruch in dieser Geschichte darin, dass sie hypothekarisch gesicherten Wertpapieren und anderen Derivaten einen Wert zuschrieb.	Was jedoch konkret gebrochen wurde, war die Geschichte, die hypothekenbesicherten Wertpapieren und anderen Derivaten auf der Grundlage nicht rückzahlbarer Kredite Wert zuweist.
To optical physicists, though, light is a series of electromagnetic waves that vary widely in their frequency and wavelengths.	Für Physiker im Bereich Optik hingegen ist Licht eine Reihe elektromagnetischer Wellen, die in ihrer Frequenz und Wellenlänge stark variieren.	Für optische Physiker ist Licht jedoch eine Reihe elektromagnetischer Wellen, die in ihrer Frequenz und Wellenlänge sehr unterschiedlich sind.
• Orient the learner and concentrate upon the details of biomechanics within the Carvinggolf technique.	• Den Schüler ausrichten und konzentrieren auf die Vorgaben der Biomechanik zur Carvinggolf Technik.	• Orientieren Sie den Lernenden und konzentrieren Sie sich auf die Details der Biomechanik im Carvinggolf.
A main point is the security about the suspended loads on the kite line.	Ein besonderer Augenmerk wegen der Sicherheit ist auf die Behandlung von schwebenden Lasten zu richten.	Ein Hauptpunkt ist die Sicherheit der Hängelasten auf der Kitelinie.
He was already very sportive in his younger days which inured to the benefit of his future film career.	Er war bereits in jungen Jahren sportlich sehr aktiv, was ihm für seine spätere Filmkarriere zugute kommen sollte.	Er war schon sehr sportlich in seinen jüngeren Tagen, die sich zum Wohle seiner zukünftigen Filmkarriere.
The advantages and disadvantages for agriculture will maintain a balance.	Die Vor- und Nachteile für die Landwirtschaft werden sich in etwa die Waage halten.	Die Vor- und Nachteile für die Landwirtschaft werden ein Gleichgewicht erhalten.
Those who try to lead a spiritual life have always been compared to warriors (there are classic writings on this subject), and one must truly be a fighter – “fighter ” is more exact than“ warrior ” because you wage war against no one: everything wages war against you! Everything... (Mother makes a gesture like an avalanche falling upon her) and with such savage opposition!...	Man hat diejenigen, die ein spirituelles Leben führen wollen, immer mit Kriegern verglichen (darüber gibt es altüberlieferte Texte). Man muß wirklich ein Kämpfer sein – "Kämpfer" ist richtiger als "Krieger"; man bekriegt niemanden: alles bekriegt euch! Alles... (Geste wie eine Lawine, die auf Mutters Kopf fällt) mit einem derart ungebändigten Widerstand!...	Diejenigen, die versuchen, ein spirituelles Leben zu führen, sind immer mit Kriegern verglichen worden (es gibt klassische Schriften zu diesem Thema), und man muss wirklich Kämpfer sein – „Kämpfer“ ist genauer als „Krieger“, weil man Krieg gegen niemanden führt: Alles zieht Krieg gegen euch.
special unusual trumpet construction, artificial prolongation of the trumpet with a garden hose, "remote playing" through a metal box	Besonderheit besondere Trompetenkonstruktion, künstliche Verlängerung der Trompete mit Gartenschlauch, "Fernspielen" aus Metallkasten	besondere ungewöhnliche Trompete Konstruktion, künstliche Verlängerung der Trompete mit einem

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
		Gartenschlauch, "Entfernung" durch eine Metallbox.
And the Lord shall deliver me from every evil work, and will preserve unto his heavenly kingdom.	Der Herr wird mich auch von jedem boshaften Werk erlösen und mich in sein himmlisches Reich retten.	Und der Herr wird mich von jedem bösen Werk befreien und seinem himmlischen Reich bewahren.
“This kind of use of APIs in this way where you use the minimum you need to be compatible is fair use, ” Van Nest declared.	“Diese Art der Nutzung von APIs auf diese Weise, wo Sie das Minimum Sie brauchen, um kompatibel sein ist fair use, ” Van Nest erklärt.	„Diese Art der Verwendung von APIs in dieser Weise, in der Sie das Minimum verwenden, das Sie benötigen, um kompatibel zu sein, ist fairer Gebrauch“, erklärt Van Nest.
Other events that may interested in: Doubleday Mystery Pot 3. March 15: 00 Kitzbühel Additional chances to win await you on the slots at Casino Austria on the doubledays.	Weitere Veranstaltungen die Sie interessieren könnten: Doppeltag Mystery Pot 3. März 15: 00 Kitzbühel An den Automaten von Casinos Austria erwarten Sie an den Doppeltagen zusätzliche Gewinnchancen!	Weitere Events, die vielleicht interessiert sind: Doubleday Mystery Pot 3. März 15: 00 Kitzbühel Weitere Gewinnchancen erwarten Sie an den zweitägigen Slots bei Casino Austria.
Request here your reservation for the holiday-home holiday home "Lild Strand"	Stellen Sie hier Ihre Reservierungsanfrage für Ferienhaus "Lild Strand"	Fordern Sie hier Ihre Reservierung für das Ferienhaus "Lild Strand" an.
Through it they confirm their belief in a pre-christian Godhead while being guided by him.	Hiermit bestätigen sie ihren Glauben an dem vor-christlichen Gottheit und lassen sich von ihm führen.	Durch sie bestätigen sie ihren Glauben an einen vorchristlichen Gott, während sie von ihm geführt werden.
Consequently, you need to have publishing permissions for the corresponding dictionary categories, if you want to revoke the publication of a diagram that references dictionary entries.	Folglich brauchen Sie Veröffentlichungsrechte für die dazugehörigen Glossarkategorien, wenn Sie die Veröffentlichung eines Diagramms, welches Glossareinträge referenziert, widerrufen möchten.	Folglich benötigen Sie die Veröffentlichungsrechte für die entsprechenden Wörterbuchkategorien, wenn Sie die Veröffentlichung eines Diagramms widerrufen möchten, das Wörterbucheinträge enthält.
National presentations and participations in international exhibitions at all the editions of the Venice Biennale.	Nationale Präsentationen und Teilnahmen an Ausstellungen bei allen Editionen der Biennale Venedig.	Nationale Präsentationen und Teilnahme an internationalen Ausstellungen auf allen Editionen der Biennale Venedig.
Do not forget to bring your friends, cameras and to put your signature on the wall in front of the studios (everybody else does it, and you are perfectly safe to do it).	Vergessen Sie nicht, Ihre Freunde, Kameras zu bringen und zu Ihrer Unterschrift an die Wand hängen vor der Studios (jeder tut es, und Sie sind absolut sicher, es zu tun).	Vergessen Sie nicht, Ihre Freunde, Kameras und Ihre Unterschrift an die Wand vor die Studios zu bringen (jeder andere tut es, und Sie sind vollkommen sicher, es zu tun).
The Ariane 5 launch window opens at 22: 54 CEST and closes 57 minutes later at 23: 51 CEST.	Das Fenster für den Start der Ariane öffnet sich um 22: 54 MESZ und schließt sich 57 Minuten später um 23: 51 MESZ.	Das Startfenster der Ariane 5 öffnet sich um 22: 54 CEST und schließt 57 Minuten später um 23: 51 CEST.

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
The Austrian anthropologist S.F. Nadel reported that there should have existed 105 different Nuba languages, not dialects but different languages as there are in Europe.	Der österreichische Anthropologe S.F. Nadel berichtete, daß es 105 verschiedene Nubasprachen gegeben haben soll, keine Dialekte, sondern so verschieden, wie es die Sprachen in Europa sind.	Der österreichische Anthropologe S.F. Nadel berichtete, dass es 105 verschiedene Nuba-Sprachen geben sollte, nicht Dialekte, sondern verschiedene Sprachen wie in Europa.
New Hyundai Solaris 2011: fuel consumption is 6.4 liters per 100 km	Neue Hyundai Solaris 2011: Verbrauch liegt bei 6.4 Liter pro 100 km	Neue Hyundai Solaris 2011: Kraftstoffverbrauch 6,4 Liter pro 100 km.
4. Click the blue button to create your Stripe account.	5. Klicken Sie auf das blaue Feld und erstellen Sie Ihr Stripe-Konto.	Klicken Sie auf die blaue Schaltfläche, um Ihr Stripe-Konto zu erstellen.

Experiment 2 : Sample of Sentences Used to Train Models.

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
General product information about 44A0111-20-0(1200)\n	Allgemeine Produktinformationen zu 44A0111-20-2(1200)\n	Allgemeine Produktinformationen zu 44A0111-20-2(1200)\n
24 232 owners on the site\n	24 230 Besitzer auf der Website\n	24 230 Besitzer auf der Website\n
Number of cells: 13\n	Anzahl der Zellen: 13\n	Anzahl der Zellen: 13\n
2. How did you get the role of Eve?\n	2. Wie hast du die Rolle der Eve bekommen?\n	2. Wie hast du die Rolle der Eve bekommen?\n
General product information about 55A0111-24-5(100)\n	Allgemeine Produktinformationen zu 55A0111-24-5(100)\n	Allgemeine Produktinformationen zu 55A0111-24-5(100)\n
The street of 57Hotel a Sydney\n	Die Straße, 57Hotel in Sydney\n	Die Straße, 57Hotel in Sydney\n
1 person recommends this bar\n	1 Person empfiehlt diese Bar\n	1 Person empfiehlt diese Bar\n
Turkish - 55 million in Turkey\n	Türkische - 55 Million in Türkei\n	Türkische - 55 Million in Türkei\n
10000 people lacking 11000 People needed\n	10000 Menschen fehlen 11000 Menschen benötigt\n	10000 Menschen fehlen 11000 Menschen benötigt\n
Marmot's Summer 2015 catalog!\n	Der neue Sommer 2015 Katalog!\n	Der neue Sommer 2015 Katalog!\n
Date added: June 23, 2006\n	Datum hinzugefügt: 23 Juni 2006\n	Datum hinzugefügt: 23 Juni 2006\n
Property of about 14.000 m²\n	Grundstück von ca. 14.000 m²\n	Grundstück von ca. 14.000 m²\n
9 Does it make sense to buy?\n	9 Sollte ich es kaufen?\n	9 Sollte ich es kaufen?\n
Machine weight : 140 000 kg\n	Maschinengewicht : 140 000 kg\n	Maschinengewicht : 140 000 kg\n
Some of them are: 1.\n	Einige von ihnen sind: 1.\n	Einige von ihnen sind: 1.\n
Advertisement Back Advertisement Azimut 90 £591,058 Listed price: €675,000 Print Facebook Twitter PREVIOUS NEXT - of - images Like this boat?\n	Anzeige Zurück Anzeige Azimut 90 €675.000 Print Facebook Twitter Vorherige Nächste - von - Fotos Gefällt Ihnen dieses Boot?\n	Anzeige Zurück Anzeige Azimut 90 €675.000 Print Facebook Twitter Vorherige Nächste - von - Fotos Gefällt Ihnen dieses Boot?\n
Completion: from 2011 to 2013\n	Fertigstellung: von 2011 bis 2013\n	Fertigstellung: von 2011 bis 2013\n
Any questions on eDressit New Arrivals Strapless Sweetheart Neckline Evening Dress (00096306), welcome to contact us!\n	Bei weiterem Fragen zu eDressit Trägerlos Süße Ausschnitt Abendkleid (00096306) dann kontaktieren Sie uns bitte!\n	Bei weiterem Fragen zu eDressit Trägerlos Süße Ausschnitt Abendkleid (00096306) dann kontaktieren Sie uns bitte!\n
Death: after September 1642\n	Gestorben: nach September 1642\n	Gestorben: nach September 1642\n
From 18 October 2012 to 22 October 2012\n	Von 18 Oktober 2012 bis 22 Oktober 2012\n	Von 18 Oktober 2012 bis 22 Oktober 2012\n

Experiment 3 : Sample of Sentences Used to Train Models.

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Tra Inputs
and collected in such manner as may be prescribed .	und auf die vorgeschriebene weise gesammelt werden .	und so gesammelt , wie vorgeschrieben ist .
with the exlibris	mit dem exlibris	mit den exlibris
for it turns earth , and even manure , into gold ,	denn es macht die erde und sogar in scharen zu gold ,	denn es verwandelt erde und gülle in gold ,
and we will send you an offer .	und wir senden ihnen ein angebot zu .	und wir unterbreiten ihnen angebot .
we do not inherit the earth from our ancestors ;	wir erben die erde nicht von unseren vorfahren ,	wir erben die erde nicht von uns vorfahren .
there is not a moving creature on earth , nor a bird that flies with its two wings , but are communities like you. we have neglected nothing in the book. then to their lord they will be gathered .	es gibt e kein gehendes tier auf der erde und keinen vogel , die auf ihren zwei flügeln fliegen , die keine gemeinschaften sind wie ihr. wir haben im buch nichts unvollständig gelassen. dann werden sie zu ihrem herrn versammelt werden .	es gibt kein bewegliches geschöp erden noch einen vogel , der mit se beiden flügeln fliegt , sor gemeinschaften wie ihr. wir h nichts im buch vernachlässigt , werden sie zu ihrem herrn versam werden .
nights 123456789101112131415161718192O	nächte 123456789101112131415161718192O	nächte 12345678910111213141516171
and the tables of the tora , and the book of the koran .	und die tafeln der tora und das buch des korans .	und die tabellen der tora und das des korans .
or also those who would rather secure , , these circumstances , , ,	oder auch diejenigen , die diese , , umstände , , ein bischen versichern ,	oder auch diejenigen , die umstände lieber absichern würde
until then we wish you a successful hunting year ,	bis dahin wünschen wir ihnen ein erfolgreiches jagdjahr ,	bis dahin wünschen wir ihnen erfolgreiches jagdjahr ,
alif , lam , ra. these are the verses of the book and a clear qur 'an .	alif-lam-ra. dies sind die zeichen des buches und eines deutlichen qur ' ans .	alif , lam , ra. dies sind die verse buches und ein klarer koran .
the creation of the	die erschaffung der	die entstehung der
with every pedal spin ,	mit jeder pedalumdrehung ,	mit jeder pedalumdrehung ,
" so bring back our forefathers , if you are truthful ! "	so bringt doch unsere väter (zurück) , wenn ihr die wahrheit redet ! "	" bringt also unsere vorfa zurück , wenn ihr wahrhaftig s "
or : where does the koran come from ?	oder : woher kommt der koran ?	oder : woher kommt der koran ?
posted on 23 / 04 / 2018 at 13 : 35 .	veröffentlicht am 23 / 04 / 2018 um 13 : 35 .	veröffentlicht am 23.04.2018 um 35 uhr
we welcome you warmheartedly !	seien sie uns herzlich willkommen !	wir heißen sie herzlich willkomme

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Tra Inputs
and let me guide you to your lord so you would fear [him] ? ' "	und daß ich dich zu deinem herrn rechtleite , so daß du gottesfürchtig wirst ? "	und lass mich dich zu deinem h führen , damit du [ihn & fürchtest ? "
verily this qur 'an doth explain to the children of israel most of the matters in which they disagree .	wahrlich , dieser koran erklärt den kindern israels das meiste von dem , worüber sie uneins sind .	wahrlich , dieser koran erklärt kindern israels die me angelegenheiten , in denen sie and meinung sind .
do you have any questions ?	haben sie irgendwelche fragen ?	haben sie fragen ?

Experiment 4 : Sample of Sentences Used to Train Models.

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Tra Inputs
and collected in such manner as may be prescribed.	und auf die vorgeschriebene weise gesammelt werden .	und so gesammelt , wie vorgeschrieben ist .
with the exlibris	mit dem exlibris	mit den exlibris
for it turns earth , and even manure , into gold ,	denn es macht die erde und sogar in scharen zu gold ,	denn es verwandelt erde und gülle in gold ,
and we will send you an offer .	und wir senden ihnen ein angebot zu .	und wir unterbreiten ihnen angebot .
we do not inherit the earth from our ancestors ;	wir erben die erde nicht von unseren vorfahren ,	wir erben die erde nicht von uns vorfahren .
there is not a moving creature on earth , nor a bird that flies with its two wings , but are communities like you. we have neglected nothing in the book. then to their lord they will be gathered .	es gibt e kein gehendes tier auf der erde und keinen vogel , die auf ihren zwei flügeln fliegen , die keine gemeinschaften sind wie ihr. wir haben im buch nichts unvollständig gelassen. dann werden sie zu ihrem herrn versammelt werden .	es gibt kein bewegliches geschöp erden noch einen vogel , der mit se beiden flügeln fliegt , sor gemeinschaften wie ihr. wir h nichts im buch vernachlässigt , werden sie zu ihrem herrn versam werden .
nights 123456789101112131415161718192O	nächte 123456789101112131415161718192O	nächte 12345678910111213141516171
and the tables of the tora , and the book of the koran .	und die tafeln der tora und das buch des korans .	und die tabellen der tora und das des korans .
or also those who would rather secure , , these circumstances , , ,	oder auch diejenigen , die diese , , umstände , , ein bischen versichern ,	oder auch diejenigen , die umstände lieber absichern würde
until then we wish you a successful hunting year ,	bis dahin wünschen wir ihnen ein erfolgreiches jagdjahr ,	bis dahin wünschen wir ihnen erfolgreiches jagdjahr ,
alif , lam , ra. these are the verses of the book and a clear qur 'an .	alif-lam-ra. dies sind die zeichen des buches und eines deutlichen qur ' ans .	alif , lam , ra. dies sind die verse buches und ein klarer koran .
the creation of the	die erschaffung der	die entstehung der
with every pedal spin ,	mit jeder pedalumdrehung ,	mit jeder pedalumdrehung ,
" so bring back our forefathers , if you are truthful ! "	so bringt doch unsere väter (zurück) , wenn ihr die wahrheit redet ! "	" bringt also unsere vorfa zurück , wenn ihr wahrhaftig s "
or : where does the koran come from ?	oder : woher kommt der koran ?	oder : woher kommt der koran ?
posted on 23 / 04 / 2018 at 13 : 35 .	veröffentlicht am 23 / 04 / 2018 um 13 : 35 .	veröffentlicht am 23.04.2018 um 35 uhr
we welcome you warmheartedly !	seien sie uns herzlich willkommen !	wir heißen sie herzlich willkomme

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Tra Inputs
and let me guide you to your lord so you would fear [him] ? ' "	und daß ich dich zu deinem herrn rechtleite , so daß du gottesfürchtig wirst ? "	und lass mich dich zu deinem h führen , damit du [ihn & fürchtest ? "
verily this qur 'an doth explain to the children of israel most of the matters in which they disagree .	wahrlich , dieser koran erklärt den kindern israels das meiste von dem , worüber sie uneins sind .	wahrlich , dieser koran erklärt kindern israels die me angelegenheiten , in denen sie and meinung sind .
do you have any questions ?	haben sie irgendwelche fragen ?	haben sie fragen ?

Experiment 5 : Sample of Sentences Used to Train Models.

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
o-a-s-l-s . \ n	o-a-s-e . \ n	o-a-s-l-s . \ n
• national accounts esa aggregates \ n	• volkswirtschaftliche gesamtrechnungen : esvg aggregate \ n	• volkswirtschaftliche gesamtrechnungen esa aggregate \ n
locally , the development will provide a tangible focus around which pride , confidence and public support can grow . \ n	die autoren des lehrbuchs setzen bei den beteiligten interesse an der feststellung der probleme in einem raum und den politischen willen zur lösung dieser probleme voraus . \ n	auf lokaler ebene wird die entwicklung einen greifbaren schwerpunkt bilden , um den herum stolz , vertrauen und öffentliche unterstützung wachsen können .
the equipment (100) according to claim 12 , wherein a part of a flat surface (101 & apos ;) is provided with a through hole (102) in correspondence with which is found a salting device for the bases (20) of a whole cheese contained within a band mould . \ n	vorrichtung (100) nach anspruch 12 , wobei ein teil einer ebenen fläche (101 & apos ;) mit einer durchbrechung (102) versehen ist , mit der in übereinstimmung eine salzvorrichtung für die böden (20) eines in einer bundform enthaltenen käses bereitgestellt ist . \ n	die ausrüstung (100) gemäß anspruch 12 , bei der ein teil einer ebenen oberfläche (101 & apos ;) mit einer durchgangsbohrung (102) versehen ist , in deren zusammenhang eine salzvorrichtung für die basen (20) eines ganzen käses gefunden wird , der in einer bandform enthalten ist .
equipment for the preparation of a hot beverage \ n	ausrüstung zur zubereitung eines heissgetränks \ n	ausrüstung für die zubereitung eines heißen getränks
antigelling agent for hydrocarbon mixture containing conjugated diene and method of preventing clogging of apparatus for separating and purifying the mixture \ n	konjugierte diene enthaltendes mittel zur verhinderung der ausbildung von gel in einer kohlenwasserstoffmischung und verfahren zur verhinderung von verstopfungen in einer vorrichtung zur trennung und säuberung der kohlenwasserstoffmischung \ n	entzündungshemmer für kohlenwasserstoffgemisch mit konjugiertem dien und methode zur verhinderung von verstopfungen der vorrichtung zur trennung und reinigung der mischung
and- and she was definitely giving me the vibe . \ n	ach , nun werd doch erwachsen , monica . \ n	und sie gab mir definitiv den vibe . \ n
the cegb and sseb have in effect thereby confirmed their policy of reprocessing . \ n	cegb und sseb haben damit ihre wiederaufarbeitungs- politik bekraeftigt . \ n	cegb und sseb haben damit ihre politik der wiederaufbereitung bestätigt . \ n

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
the most pronounced changes in the unemployment rate from april 1999 to april 2000 were infour greek regions . \ n	die veränderung der arbeitslosenquotevon april1999 auf april 2000 war am stärksten in vier griechischen regionen . \ n	die deutlichsten veränderungen der arbeitslosenquote von april 1999 bis april 2000 gab es in vier griechischen regionen .
method according to any of claims 1 to 3 , characterized in that during the introduction of the second end-piece (13) into the body portion (10) a quantity of cooling gas is provided to the region of said second end-piece (13) adjacent the product (11) . \ n	verfahren nach einem der ansprüche 1 bis 3 , dadurch gekennzeichnet , daß während der einführung des zweiten endstücks (13) in das körperteilstück (10) eine kühlgasmenge zu dem bereich des zweiten endstücks (13) neben dem produkt (11) geliefert wird . \ n	verfahren gemäß anspruch 1 bis 3 , dadurch gekennzeichnet , dass bei der einführung des zweiten endstücks (13) in den körperteil (10) eine kühlgasmenge in den bereich des zweiten endstücks (13) neben dem produkt (11) geleitet wird .
the correct solution is : yes to stress tests , yes to supervision and no to demonising nuclear energy . \ n	desmin-gen mit neuer punktmutation , die dilatationscardiomyopathie verursacht \ n	die richtige lösung lautet : ja zu stresstests , ja zur aufsicht und nein zur dämonisierung der atomenergie .
that & apos ; s nowhere near the full story . \ n	das ist nicht einmal ansatzweise die komplette geschichte . \ n	das & apos ; ist noch lange nicht die ganze geschichte . \ n
the back pressure cylinder and the forme inking unit (17) and / or the substrate are subdivided in the axial direction into a plurality of thermic zones (z1 , z2 ... \ n	der gegendruckzylinder bzw. das farbauftragswerk (17) und / oder das substrat sind in axialer richtung in mehrere thermische zonen (z1 , z2 ... \ n	der gegendruckzylinder und die formfarbeinheit (17) und / oder das substrat sind in axialer richtung in mehrere thermische zonen unterteilt (z1 , z2 ... \ n

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
a process for extending carbon chain by the use of allylic sulfide of chemical formula 1 , which comprises the steps of (a-2) deprotonating allylic sulfone compound (b) , and reacting the resultant compound with allylic sulfide of chemical formula 1 to obtain thio-sulfone compound (c) ; and (b-2) selectively oxidizing the thio-sulfone compound (c) to obtain the corresponding allylic sulfone compound (d) . \ n	verfahren zur verlängerung der kohlenstoffkette durch die verwendung von allylsulfid der chemischen formel 1 , das die folgenden schritte umfasst : (a-2) deprotonierung der allylsulfonverbindung (b) und zur reaktion bringen der sich ergebenden verbindung mit allylsulfid der chemischen formel 1 , um die thiosulfonverbindung (c) zu erhalten ; und (b-2) selektive oxidierung der thiosulfonverbindung (c) , um die entsprechende allylsulfonverbindung (d) zu erhalten . \ n	ein verfahren zur verlängerung der kohlenstoffkette durch verwendung von allylsulfid der chemischen formel 1 , das die schritte der (a-2) deprotonierenden allylsulfonverbindung (b) umfasst und die resultierende verbindung mit allylsulfid der chemischen formel 1 zur gewinnung der thio-sulfonverbindung (c) und (b-2) zur selektiven oxidation der thio-sulfonverbindung (c) zur gewinnung der entsprechenden allylsulfonverbindung (d) .
everyone in the village knows ... \ n	jeder im dorf weiß ... \ n	jeder im dorf weiß ... \ n
a method as claimed in any preceding claim characterised in that the glass film is in the form of hollow fibres . \ n	verfahren nach einem der vorhergehenden ansprüche , dadurch gekennzeichnet , daß der glasfilm aus hohlfasern besteht . \ n	ein verfahren wie in jedem vorhergehenden anspruch beansprucht , das dadurch gekennzeichnet ist , dass die glasfolie in form von hohlfasern ist . \ n
what & apos ; s the matter , brain ? \ n	was ist los , genie ? \ n	what & apos ; s the matter , brain ?

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
the article of claim 34 , wherein the thermoplastic foam is comprised of a polymeric blend including a major amount of a homopolymer or copolymer in which the majority of monomeric units are propylene monomeric units and a minor amount of a polyethylene type resin modifier in which a majority of monomeric units are ethylene monomeric units . \ n	gegenstand nach anspruch 34 , bei welchem der thermoplastische schaum aus einer polymeren mischung besteht , die eine größere menge eines homopolymeren oder copolymeren , in denen der hauptanteil der monomeren einheiten monomere propyleneinheiten sind , und eine kleinere menge eines harzmodifikationsmittels vom polyethylentyp , in dem ein hauptanteil von monomeren einheiten monomere ethyleneinheiten sind , einschließt . \ n	der artikel des anspruchs 34 , wonach der thermoplastische schaum aus einer polymeren mischung besteht , die eine große menge eines homopolymers oder copolymers enthält , bei dem die mehrheit der monomeren einheiten aus propylenmonomeren einheiten besteht , und eine geringe menge eines polyethylenharzmodifikators , bei dem die mehrheit der monomeren einheiten aus ethylenmonomeren einheiten besteht .
although the exclusive purchasing provisions , described in recieal 14 , foreclose sales opportunities for third party manufacturers of motors and hydraulics , this restriction resulta from the setting up of the joint venture and appears to be reasonably necessary to its operation . \ n	obwohl die in absatz 14 beschriebenen alleinbezugsverpflichtungen dritten herstellern von motoren und hydraulikgeräten keine absatzmöglichkeiten lassen , hängt diese einschränkung mit der gründung des gemeinschaftsunternehmens zu sammen und scheint für seinen betrieb mehr oder weniger notwendig zu sein . \ n	obwohl die in abschnitt 14 beschriebenen exklusiven einkaufsbestimmungen vertriebschancen für drittanbieter von motoren und hydraulik ausschließen , ergibt sich diese beschränkung aus der gründung des gemeinschaftsunternehmens und scheint für seinen betrieb angemessen notwendig zu sein .

Experiment 6 : Sample of Sentences Used to Train Models.

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
- it & apos ; s urgent . \ n	doch , es ist dringend . \ n	- es ist dringend . \ n
in ireland 80 % of beef calves normally come from the dairy herd , and milk and beef represent more than 70 % of total agricultural output . \ n	in irland stammen 80 % der kälber normaler weise aus der milchwirtschaft , und milch und rindfleisch stellen insgesamt einen anteil von mehr als 70 % aller landwirtschaftlichen erzeugnisse . \ n	in irland stammen 80 % der rindfleischkälber normalerweise aus der milchviehherde , und milch und rindfleisch machen über 70 % der gesamten landwirtschaftlichen produktion aus .
as chairman of the committee on fisheries , i warmly congratulate mr provan for the standard of his work and the clear-sightedness of his report . \ n	der herr berichterstatter hat in seiner entschließung , in seinem kommentar und in seinen schlußfolgerungen einen hervorragenden weg aufgezeigt . \ n	als vorsitzender des fischereiausschusses gratuliere ich herrn provan herzlich zu der qualität seiner arbeit und der weitsichtigkeit seines berichts .
show me . \ n	zeig es mir . \ n	zeig mir . \ n
on the one hand in the form of the usual tables supplied for previous surveys using national computing programmes already in existence and designed to meet community standards , \ n	bei dieser arbeit würde eurostat die von den mitgliedataaten geäusserte besorgnis berücksichtigen und entsprechend versuchen , die anzahl der nimexe-rubriken in vernünftigen grenzen zu halten . \ n	- einerseits in form der üblichen tabellen , die für frühere erhebungen unter verwendung bereits bestehender nationaler computerprogramme bereitgestellt wurden und die den gemeinschaftsstandards entsprechen ,
you owe me four dinners , three breakfasts in bed , a bunch of lap dances , and a car wash . \ n	du schuldest mir vier essen , drei frühstücke im bett , einen haufen lap dances und eine autowäsche . \ n	sie schulden mir vier abendessen , drei frühstücke im bett , ein paar schoßtänze und eine autowaschanlage . \ n
(i) to improve the quality and promote the development and restructuring of high er education in the eligible countries by developing and reshaping course pro grammes and reforming the structures and management of higher education institutions ; \ n	verbesserung der qualität sowie förderung von aus und umbau des hoch schulwesens in den förderungsberechtigten ländern durch entwicklung und überarbeitung von lehrplänen sowie reform der strukturen und verwaltungs verfahren der hochschuleinrichtungen : \ n	i) verbesserung der qualität und förderung der entwicklung und umstrukturierung der höheren bildung in den förderfähigen ländern durch die entwicklung und umgestaltung von studienprogrammen und die reform der strukturen und des managements von hochschuleinrichtungen ;
well , i want to understand their relationship . \ n	weil ich ihre beziehung verstehen will . \ n	nun , ich will ihre beziehung verstehen . \ n
what a character ! \ n	was für ein halunke ! \ n	was für ein charakter ! \ n

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
macro-economic progress has yet to trickle down into tangible economic benefits for the population at large . \ n	vorerst hat der makroökonomische fortschritt noch nicht zu spürbaren verbesserungen für die bevölkerung geführt . \ n	der makroökonomische fortschritt muss sich erst noch in greifbare wirtschaftliche vorteile für die gesamtbevölkerung niederschlagen .
has adopted this regulation : \ n	dib kommission der europaischen gemeinschaften - \ n	hat diese verordnung angenommen : \ n
a system as claimed in any preceding claim , wherein the prepayment unit (64) is arranged to access the meter measurements periodically , and to calculate after each such access , on the basis of price information stored in the prepayment unit (64) , the amount of credit remaining . \ n	ein system nach einem der vorangehenden ansprüche , bei dem die vorauszahlungseinheit (64) ausgebildet ist zum periodischen zugänglichmachen der messungen des meßgerätes und zum berechnen nach jedem solchen zugang der menge an verbleibendem kredit auf der basis von preisinformation , die in der vorauszahlungseinheit (64) abgespeichert ist . \ n	ein system , wie es in jedem vorhergehenden anspruch beansprucht wird , bei dem die vorauszahlungseinheit (64) so angeordnet ist , dass sie periodisch auf die zählermessungen zugreift und nach jedem derartigen zugriff auf grundlage der in der vorauszahlungseinheit (64) gespeicherten preisinformationen den verbleibenden kreditbetrag berechnet .
that & apos ; s a stupid thing to say . \ n	du hast deinen teil beigetragen ! du hast deinen teil beigetragen ! \ n	das & apos ; ist eine dumme sache . \ n
a method as claimed in claim 28 , wherein the exotherm depressant comprises alpha-methyl styrene in an amount of about 1 % to about 20 % by weight of the composition . \ n	verfahren nach anspruch 28 , worin das die exotherme reaktion unterdrückende mittel a-methylstyren in einer menge von ca. 1 gew.- % bis ca. 20 gew.- % der zusammensetzung umfasst . \ n	ein verfahren gemäß anspruch 28 , bei dem das exotherme depressivum alpha-methylstyrol in einer menge von etwa 1 bis etwa 20 gewichtsprozent der zusammensetzung enthält .
display apparatus according to any preceding claim 3 to 9 , chacaterised in that an air gap is provided between the top mounted electromagnet (3) of the outer fixed frame (1) and the top mounted permanent magnet (5) of the picture frame (2) of between 0.5cm to 2.5cm. \ n	darstellungsvorrichtung nach einem der vorhergehenden ansprüche 3 bis 9 , gekennzeichnet durch einen 0,5 bis 2,5 cm betragenden luftspalt zwischen dem oben am äußeren ortsfesten rahmen (1) angeordneten elektromagneten (3) und dem oben am bildrahmen (2) angeordneten dauermagneten (5) . \ n	anzeigegerät gemäß anspruch 3 bis 9 , das dadurch gekennzeichnet ist , dass ein luftspalt zwischen dem oben angebrachten elektromagneten (3) des äußeren festen rahmens (1) und dem oben angebrachten permanentmagneten (5) des bilderrahmens (2) von 0,5 bis 2,5 cm vorgesehen ist . \ n

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
method according to one of claims 19 to 21 , characterised in that following incorporation of the gas distributing structures an elevation which surrounds the reaction areas is applied so that only one respective recess develops in the area of the reaction areas . \ n	verfahren nach einem der ansprüche 19 bis 21 , dadurch gekennzeichnet , dass nach dem einarbeiten der gasverteilerstrukturen eine die reaktionsräume umgebende erhebung aufgebracht wird , so dass im bereich der reaktionsräume jeweils eine vertiefung entsteht . \ n	verfahren gemäß einem der ansprüche 19 bis 21 , das dadurch gekennzeichnet ist , dass nach einbeziehung der gasverteilenden strukturen eine die reaktionsbereiche umgebende erhebung aufgebracht wird , so dass sich im bereich der reaktionsbereiche nur eine entsprechende vertiefung bildet .
the apparatus is intended for extraction of material from a reactor (1) , in particular , a sprinkle roast reactor . \ n	die erfindung betrifft eine vorrichtung zum austrag von material aus einem reaktor , insbesondere sprühröstreaktor (1) . \ n	das gerät ist für die gewinnung von material aus einem reaktor (1) , insbesondere einem streubratenreaktor , bestimmt .
a glass ceramic article as claimed as in claim 1 , wherein said infrared transmitting glass ceramics is adjusted so that a precipitated crystal and a heterogeneous particle have particle sizes not greater than 3 µm . \ n	glaskeramischer gegenstand nach anspruch 1 , wobei die infrarotdurchlässige glaskeramik so eingestellt wird , dass ein abgeschiedener kristall und ein heterogener partikel partikelgrößen von nicht größer als 3 µm aufweist . \ n	ein glaskeramisches erzeugnis gemäß anspruch 1 , bei dem die infrarotübertragende glaskeramik so eingestellt wird , dass ein gefällter kristall und ein heterogener partikel eine partikelgröße von nicht mehr als 3 µm aufweisen .
collins was alive when his feet were cut off ? \ n	collins lebte , als man ihm die füße abschnitt ? \ n	collins lebte , als ihm die füße abgeschnitten wurden ? \ n
a filterbank as claimed in claim 4 , wherein the multiplier means (28) comprises one or more dedicated multiplier resources incorporated on the application specific integrated circuit (16) . \ n	filterbank nach anspruch 4 , worin das vervielfachungsmittel (28) ein oder mehrere zweckbestimmte multiplikatorbetriebsmittel umfasst , die im anwendungsspezifischen integrierten schaltkreis (16) enthalten sind . \ n	eine filterbank gemäß anspruch 4 , wobei das multiplikatormittel (28) eine oder mehrere dedizierte multiplikatorressourcen umfasst , die in den anwendungsspezifischen integrierten schaltkreis (16) integriert sind . \ n

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
infrastructure investments to get rid ofbottlenecks in the rail network and developa transeuropean rail freight network.thisrequires novel financial solutions , e.g.usinginvestment funds built up using tolls oncompeting road routes . \ n	die investitionen in die infrastruktur müssensich auf die beseitigung von engpässen im schienennetz und auf die schaffung eines transeuropäischen schienengüterverkehrsnetzes konzentrieren.dies erfordert innovative finanzierungslösungen , zum beispiel durch investitionsfonds , die aus gebühren für die nutzung konkurrierender straßenverkehrswege gespeist werden . \ n	infrastrukturinvestitionen zur beseitigung von engpässen im schienennetz und zum aufbau eines transeuropäischen schienengüterverkehrsnetzes erfordern neuartige finanzlösungen , z. b. die verwendung von investitionsfonds , die über mautgebühren auf konkurrierenden straßen aufgebaut werden . \ n
the multi-chamber container of claim 8 wherein the alkaline bicarbonate solution comprises : sodium chloride ; sodium lactate ; and sodium bicarbonate . \ n	mehrkammerbehälter gemäß anspruch 8 , wobei die alkalische bicarbonatlösung umfaßt : natriumchlorid , natriumlactat und natriumbicarbonat . \ n	der mehrkammerbehälter nach anspruch 8 , in dem die alkalische bikarbonatlösung besteht aus : natriumchlorid , natriumlaktat und natriumbicarbonat . \ n
good . \ n	gut . \ n	gut . \ n
• certain electronic scales assembled or produced in the community \ n	• bestimmten in der gemeinschaft montierten oder her gestellten elektronischen waagen \ n	• bestimmte elektronische waagen , die in der gemeinschaft zusammengebaut oder hergestellt werden
god help me . \ n	gott hilf ' mir ! \ n	gott hilf mir . \ n
you & apos ; re not allowed to smoke . \ n	du hast keine erlaubnis zu rauchen . \ n	you & apos ; re nicht erlaubt , zu rauchen . \ n
since the region underneath the run-in guard (9) is also closed by the closing part (38) , the insertion opening (11) is completely closed by the run-in guard (9) and this closing part , so that chips , splinters and the like generated during machining of the workpiece are not spewed out through the insertion opening (11) . \ n	da auch der bereich unterhalb des einlaufschutzes (9) durch das schließteil (38) verschlossen ist , ist die einführöffnung (11) vollständig durch den einlaufschutz (9) und dieses schließteil geschlossen , so daß bei der bearbeitung des werkstückes entstehende späne , splitter und dgl. nicht durch die einführöffnung (11) herausgeschleudert werden . \ n	da der bereich unterhalb der einlaufsicherung (9) auch durch das schließteil (38) verschlossen ist , wird die einführöffnung (11) durch die einlaufsicherung (9) und dieses schließteil vollständig verschlossen , so dass bei der bearbeitung des werkstücks anfallende späne , splitter und dergleichen nicht durch die einführöffnung (11) ausgespuckt werden . \ n
- i & apos ; m curious . \ n	- ich bin neugierig . \ n	- ich bin neugierig . \ n

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
they & apos ; re nice kids. all nice . \ n	das sind nette kinder . \ n	sie & apos ; re nice kids. alle nett . \ n
the lamp unit of any one of claims 9 to 13 , wherein one electrode supporting closure (22) has a tubular configuration affording a storage reservoir (23) for excessive sodium and mercury . \ n	lampeneinheit nach einem der ansprüche 9 bis 13 , worin ein elektroden tragender verschluß (22) eine rohrförmige konfiguration hat , der ein speicherreservoir (23) für überschüssiges natrium und quecksilber bietet . \ n	die lampeneinheit eines beliebigen der ansprüche 9 bis 13 , wobei eine elektrode , die den verschluss (22) trägt , eine röhrenförmige konfiguration aufweist , die einen speicher (23) für übermäßiges natrium und quecksilber bietet .
mixed seasoning on basis of monosodium l-glutamate . \ n	mischgewürz auf der basis von mononatrium-l-glutamat . \ n	gemischte würze auf basis von mononatriuml-glutamat . \ n
the use according to claim 5 , wherein the neuronal disorder is an alzheimer & apos ; s disease . \ n	verwendung gemäß anspruch 5 , wobei die neuronale störung alzheimer-krankheit ist . \ n	die anwendung gemäß anspruch 5 , wobei die neuronale störung eine alzheimer & apos ; s-krankheit ist . \ n
the reader wishing to gain an overall view of the progress of the work and procedures relating to the budget should turn to p . \ n	(") leser , die sich einen globalen überblick über den verlauf aller haushaltsarbeiten und -verfahren verschaffen möchten , ver weisen wir auf s. 102 . \ n	der leser , der sich einen gesamtüberblick über den fortschritt der arbeiten und verfahren im zusammenhang mit dem haushalt verschaffen möchte , sollte sich an s. \ n wenden
please send me a free copy of & apos ; eurostat mini-guide & apos ; (catalogue containing a selection of eurostat products and services) language required : π de □ en π fr \ n	bitte schicken sie mir ein gratisexemplar des " minikatalogs voh eurostat " (eine auswahl der produkte und dienstleistungen von eurostat) gewünschte sprache : p de p e? p fr a \ n	bitte senden sie mir ein kostenloses exemplar von & apos ; eurostat mini-guide & apos ; (katalog mit einer auswahl von eurostat-produkten und -dienstleistungen) sprache erforderlich : p de ? en p fr \ n
- gite woch . \ n	gite woch . \ n	- ferienhaus woch . \ n
i will take drakkar noir . \ n	ich werde " drakkar noir " nehmen . \ n	ich nehme drakkar noir . \ n

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
anti-theft case (101) according to claim 8 or 9 , characterized in that each one of said arms (123 , 123 & apos ;) is further equipped with at least one reinforcing rib (127 , 127 & apos ;) adapted to slidingly cooperate with a corresponding rib (129 , 129 & apos ;) formed in said box (105) of said case (101) . \ n	diebstahlsicherungshülle (101) nach anspruch 8 oder 9 , dadurch gekennzeichnet , dass jeder der beiden arme (123 , 123 ') weiterhin mit mindestens einem verstärkungssteg (127 , 127 ') versehen ist , der mit einem entsprechenden , in der schachtel (105) der hülle (101) vorgesehenen steg (129 , 129 ') schiebend zusammenarbeitet . \ n	diebstahlschutz-fall (101) gemäß anspruch 8 oder 9 , dadurch gekennzeichnet , dass jedes der arme (123 , 123 & apos ;) zusätzlich mit mindestens einer verstärkungsrippe (127 , 127 & apos ;) ausgestattet ist , die dazu geeignet ist , gleitend mit einer entsprechenden rippe (129 , 129 & apos ;) zusammenzuarbeiten , die in besagter box (105) des falles (101) gebildet wird .
the government favours moving rapidly to emu , for both political and economic reasons . \ n	teile der oppositionsparteien lehnen den vorschlag des de lors-berichts für eine einheitliche währung in der dritten phase der wwu ab . \ n	die regierung befürwortet aus politischen und wirtschaftlichen gründen einen raschen übergang zum emu . \ n
yeah , i bring them out here and i walk them through the process . \ n	ja , ich bringe sie her und zeige ihnen , wie das funktioniert . \ n	ja , ich bringe sie hier raus und gehe sie durch den prozess . \ n
if i were to give some advice on moving toanother country , i would say it & apos ; s good to havea job before you go , or at least have somecontacts and a plan . \ n	wenn ich jemandem , der in ein anderes landziehen möchte , einen rat geben sollte , würdeich sagen , sieh zu , dass du vor der abreiseeinen job in der tasche hast oder wenigstens einpaar kontakte und einen plan . \ n	wenn ich ein paar ratschläge zum umzug in ein anderes land geben würde , würde ich sagen , dass es gut ist , einen job zu haben , bevor man geht , oder zumindest kontakte und einen plan zu haben . \ n

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
the planarizing machine of claim 23 wherein the first arm (132) is coupled to an actuator (131) to sweep the substrate (12) across the planarizing path and the second remains stationary with respect to the sweep axis so that the distance measuring device periodically impinges the beam against the backside (138) of the substrate (136) holder when the substrate holder is beneath the distance measuring device . \ n	maschine nach anspruch 23 , bei der erste arm (132) mit einem aktuator (131) gekoppelt ist , um das substrat (12) über den planarisierweg streichen zu lassen , während der zweite arm in bezug auf die schwenkachse ortsfest verbleibt , so dass die abstandsmesseinrichtung den strahl periodisch auf die rückseite (138) des substrathalters (36) auftreffen lässt , wenn der substrathalter sich unterhalb der abstandsmesseinrichtung befindet . \ n	die planiermaschine nach anspruch 23 , bei der der erste arm (132) an einen aktuator (131) gekoppelt ist , um das substrat (12) über den planierpfad zu fegen , und der zweite in bezug auf die sweep-achse stationär bleibt , so dass das entfernungsmessgerät den strahl periodisch gegen die rückseite (138) des substratträgers (136) drückt , wenn sich der substrathalter unter dem entfernungsmessgerät befindet .
catalyst according to claim 10 , characterised in that the proportion by weight of the second functional layer constitutes 10 to 50 % of the overall coating . \ n	katalysator nach anspruch 12 , dadurch gekennzeichnet , daß der gewichtsanteil der zweiten funktionsschicht 10 bis 50 % der gesamtbeschichtung ausmacht . \ n	katalysator nach anspruch 10 , dadurch gekennzeichnet , dass der gewichtsanteil der zweiten funktionsschicht 10 bis 50 % der gesamtbeschichtung ausmacht .
what are you worried about ? \ n	worüber machen sie sich sorgen ? \ n	worüber machen sie sich sorgen ?
- where ? \ n	- wohin ? \ n	- wo ? \ n
method according to claim 11 , characterised in that the temperature of the polyhydroxyether in the melt is 180 to 280 ° c and preferably 190 to 240 ° c. method according to one of claims 12 or 13 , characterised in that the retention time of the polyhydroxyether in the melt is less than 10 minutes and preferably less than 8 minutes . \ n	verfahren nach anspruch 11 , dadurch gekennzeichnet , dass die temperatur des polyhydroxyethers in der schmelze 180 - 280 ° c und bevorzugt 190 - 240 ° c beträgt. verfahren nach einem der ansprüche 12 oder 13 , dadurch gekennzeichnet , dass die verweilzeit des polyhydroxyethers in der schmelze weniger als 10 minuten und bevorzugt weniger als 8 minuten beträgt . \ n	verfahren nach anspruch 11 , dadurch gekennzeichnet , dass die temperatur des polyhydroxyethers in der schmelze 180 bis 280 ° c und vorzugsweise 190 bis 240 ° c beträgt. verfahren nach anspruch 12 oder 13 , dadurch gekennzeichnet , dass die verweilzeit des polyhydroxyethers in der schmelze weniger als 10 minuten und vorzugsweise weniger als 8 minuten beträgt .

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthethic Model Training Inputs
in the eu15as a whole , those without qualification havean employment rate of 55 % ; havingachieved tertiary level qualification , this rateraises to 84.5 % . \ n	in der eu-15 als ganzer liegt die erwerbsquote von unqualifizierten bei 55 % ; bei personen mit einer abgeschlossenen ausbildungim tertiärbereich liegt sie bei 84,5 % . \ n	in der eu15 insgesamt liegt die beschäftigungsquote der menschen ohne qualifikation bei 55 % , mit hochschulabschluss bei 84,5 % .
the element of claim 1 wherein said chlorinated paraffin is c 24 h 43 cl 7 . \ n	element nach anspruch 1 , in dem das chlorierte paraffin die formel c 24 h 43 cl 7 aufweist . \ n	das element des anspruchs 1 , in dem das chlorierte paraffin c 24 h 43 cl 7 beträgt .
exposure apparatus as claimed in claim 21 wherein said mask (17 ; 137) comprises the reflective optical element . \ n	belichtungsvorrichtung gemäß anspruch 21 , wobei die maske (17 ; 137) das reflektierende optische element aufweist . \ n	expositionierungsvorrichtung gemäß anspruch 21 , in der die maske (17 ; 137) das reflektierende optische element enthält .
the microturbine power generating system (10) of claim 6 , wherein the conditioner segment (11) includes a preheater (15) in communication with the recuperator (22) for heating air to a temperature sufficient to ensure fuel evaporation . \ n	mikroturbinen-energieerzeugungssystem (10) nach anspruch 6 , wobei das konditionierelement (11) einen vorwärmer (15) in verbindung mit dem rekuperator (22) aufweist zum erwärmen von luft auf eine ausreichende temperatur , um brennstoffverdampfung sicherzustellen . \ n	das mikroturbinenförmige stromerzeugungssystem (10) nach anspruch 6 , bei dem das klimagerät-segment (11) einen vorwärmer (15) enthält , der mit dem rekuperator (22) kommuniziert , um luft auf eine temperatur zu erhitzen , die ausreicht , um die brennstoffverdunstung zu gewährleisten .

Experiment 7 : Sample of Sentences Used to Train Models.

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthetic Model Training Inputs
the first is that establishing a european financial area , which implies a worldwide financial area , without acquiring the means with which to set up a social area in the community or to operate a properly coordinated community monetary policy is to expose our young community to a considerable risk of an early demise . \ n	der zuständige landesminister matthiesen hat sich in dieser tatkraft gesonnt. ein kleiner schatten fällt jedoch auf diese tatkraft , wenn ich mir vor stelle , daß er immerhin zwei wochen gewartet hat , bis er seine erkenntnisse an die zuständigen stellen weitergeleitet hat. damit sie handeln konnten . \ n	die erste ist , dass die schaffung eines europäischen finanzraums , der einen weltweiten finanzraum impliziert , ohne die mittel zu erwerben , um einen sozialen raum in der gemeinschaft zu schaffen oder eine angemessen koordinierte gemeinschaftliche geldpolitik zu betreiben , unsere junge gemeinschaft einem erheblichen risiko eines frühen niedergangs aussetzt .
first were gonna stop at a cemetery . \ n	aber zuerst halten wir auf einem friedhof . \ n	die ersten würden auf einem friedhof halt machen . \ n
use according to any one of claims 26 , 27 or 36 to 40 wherein the compound is administered in an amount sufficient to inhibit at least one plk enzyme . \ n	verwendung nach einem der ansprüche 26 , 27 oder 36 bis 40 , wobei die verbindung in einer menge verabreicht wird , die ausreichend ist , um wenigstens ein plk-enzym zu hemmen . \ n	verwendung gemäß den ansprüchen 26 , 27 oder 36 bis 40 , wobei die substanz in einer menge verabreicht wird , die ausreicht , um mindestens ein plk-enzym zu hemmen .
apparatus according to one of the preceding claims , characterised in that steam overflowing under a slight excess pressure from the steamer (90) into the application container (40) is used for steam rinsing of the application container (40 , 94) . \ n	vorrichtung nach einem der vorangehenden ansprüche , dadurch gekennzeichnet , daß zum dampfspülen des applikationsbehälters (40 , 94) aus dem dämpfer (90) in den applikationsbehälter (40) unter leichtem überdruck überströmender dampf verwendet wird . \ n	vorrichtung gemäß einem der vorstehenden ansprüche , die dadurch gekennzeichnet ist , dass dampf , der unter einem leichten überdruck aus dem dampfer (90) in den anwendungsbehälter (40) strömt , zur dampfspülung des anwendungsbehälters (40 , 94) verwendet wird .
- alonzo stevens . \ n	- alonzo stevens . \ n	- alonzo stevens . \ n
- well , it would be . \ n	- es stimmt doch . \ n	- ja , das wäre es . \ n
reclusive front man homer simpson has holed himself up in his mansion , as rumors swirled about an out-of-control narcotics addiction . \ n	der zurückgezogen lebende frontsänger homer simpson hat ... ... sich in seiner villaeingeschlossen , gerüchte kusieren ... ... über ein unkontrolliertebetäubungsmittelinjektion . \ n	der zurückgezogen lebende frontmann homer simpson hat sich in seiner villa verschanzt , als gerüchte über eine außer kontrolle geratene drogensucht die runde machten .

Common Source Text Training Inputs	Sample of Original Model Training Inputs	Sample of Synthetic Model Training Inputs
i & apos ; m not. i can & apos ; t wait till i get to fuck a girl while she rests her tits on my forehead . \ n	eigentlich kann ich es gar nicht erwarten , ein mädchen zu ficken , während sie mir ihre titten ins gesicht presst . \ n	i & apos ; m nicht. i & apos ; t wait until i get to fuck a girl while she ruht her tits on my stirn . \ n
not i said that he / she does not do this . \ n	- ich hab nie gesagt , ich tue es nicht . \ n	nicht ich sagte , dass er / sie dies nicht tut . \ n
adjustable support leg for a casing \ n	aufstellfuss für ein gehäuse \ n	verstellbares standbein für ein gehäuse \ n
oxygen permeable polymeric materials \ n	sauerstoffdurchlässige polymermaterialien \ n	sauerstoffdurchlässige polymere materialien \ n
- it is . \ n	- stimmt . \ n	- ist es . \ n
chang himself just said that they never venture beyond that point . \ n	chang selbst sagte , dass sie nie über diesen ort hinausgegangen seien . \ n	chang selbst sagte nur , dass sie sich nie über diesen punkt hinaus wagen würden .
a total slut fuck . \ n	eine totales volksloch . \ n	eine totale schlampe ficken . \ n
charles athanase walckenaer \ n	charles athanase walckenaer \ n	charles athanase walckenaer \ n
we have three main obstacles - two doors and a guard . \ n	wir haben drei haupthindernisse , zwei türen und eine wache . \ n	wir haben drei haupthindernisse - zwei türen und eine wache . \ n
67 france laboratoire glaxosmithkline tél : + 33 (0) 1 39 17 84 44 diam @ gsk. com \ n	france laboratoire glaxosmithkline tél : + 33 (0) 1 39 17 84 44 diam @ gsk.com \ n	67 france laboratoire glaxosmithkline tél : + 33 (0) 1 39 17 84 44 diam @ gsk. com \ n
the absorbent product of claim 7 wherein said polymeric material (16) is derived from a plastisol . \ n	absorbierendes produkt nach anspruch 7 , wobei das polymere material (16) von einem plastisol abgeleitet ist . \ n	das absorbierende produkt nach anspruch 7 , bei dem das polymere material (16) aus einem plastisol gewonnen wird .
method of transferring connection management information in world wide web requests and responses \ n	verfahren zur übertragung von verbindungsverwaltungsinformationen in world wide web anforderungen und antworten \ n	methode der übertragung von verbindungsmanagement-informationen in anfragen und antworten im world wide web

References

Banerjee, S., & Lavie, A. (2005). *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments.* Paper presented at the Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.

Banón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Espla-Gomis, M., . . . Koehn, P. (2020). *ParaCrawl: Web-scale acquisition of parallel corpora.* Paper presented at the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Bawden, R., Di Nunzio, G., Grozea, C., Unanue, I., Yepes, A., Mah, N., . . . Oronoz, M. (2020). *Findings of the WMT 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional Languages.* Paper presented at the 5th Conference on Machine Translation.

Birch, A., Abend, O., Bojar, O., & Haddow, B. (2016). HUME: Human UCCA-based evaluation of machine translation. *arXiv preprint arXiv:1607.00030*.

Callison-Burch, C., Fordyce, C. S., Koehn, P., Monz, C., & Schroeder, J. (2007). *(Meta-) evaluation of machine translation.* Paper presented at the Proceedings of the Second Workshop on Statistical Machine Translation.

Clark, J. H., Dyer, C., Lavie, A., & Smith, N. A. (2011). *Better hypothesis testing for statistical machine translation: Controlling for optimizer instability.* Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

El-Kishky, A., Chaudhary, V., Guzman, F., & Koehn, P. (2019). CCAligned: A massive collection of cross-lingual web-document pairs. *arXiv preprint arXiv:1911.06154*.

Farajian, M. A., Lopes, A. V., Martins, A. F., Maruf, S., & Haffari, G. (2020). *Findings of the wmt 2020 shared task on chat translation.* Paper presented at the Proceedings of the Fifth Conference on Machine Translation.

Fisher, R. A. (1937). The design of experiments. *The design of experiments.*(2nd Ed).

Graham, Y., Baldwin, T., & Mathur, N. (2015). *Accurate evaluation of segment-level machine translation metrics.* Paper presented at the Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Haffari, G., Roy, M., & Sarkar, A. (2009). *Active learning for statistical phrase-based machine translation.* Paper presented at the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Huck, M., Braune, F., & Fraser, A. (2017). *Lmu munich's neural machine translation systems for news articles and health information texts.* Paper presented at the Proceedings of the Second Conference on Machine Translation.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kiyono, S., Ito, T., Konno, R., Morishita, M., & Suzuki, J. (2020). *Tohoku-AIP-NTT at WMT 2020 News Translation Task.* Paper presented at the Proceedings of the Fifth Conference on Machine Translation.

Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P.-J., & Guzmán, F. (2020). *Findings of the WMT 2020 shared task on parallel corpus filtering and alignment.* Paper presented at the Proceedings of the Fifth Conference on Machine Translation.

Krishna, K., Tomar, G. S., Parikh, A. P., Papernot, N., & Iyyer, M. (2019). Thieves on sesame street! model extraction of bert-based apis. *arXiv preprint arXiv:1910.12366*.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Kudo, T., & Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Liu, M., Buntine, W., & Haffari, G. (2018). *Learning to actively learn neural machine translation.* Paper presented at the Proceedings of the 22nd Conference on Computational Natural Language Learning.

Lo, C.-k. (2020). *Extended study on using pretrained language models and YiSi-1 for machine translation evaluation.* Paper presented at the Proceedings of the Fifth Conference on Machine Translation.

Mandl, T., Modha, S., Kumar M, A., & Chakravarthi, B. R. (2020). *Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german.* Paper presented at the Forum for Information Retrieval Evaluation.

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*.

Mathur, N., Wei, J., Freitag, M., Ma, Q., & Bojar, O. (2020). *Results of the WMT20 Metrics Shared Task.* Paper presented at the Proceedings of the Fifth Conference on Machine Translation.

Mubarak, H., Darwish, K., Magdy, W., Elsayed, T., & Al-Khalifa, H. (2020). *Overview of osact4 arabic offensive language detection shared task.* Paper presented at the Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019). Facebook FAIR's WMT19 News Translation Task Submission. *arXiv preprint arXiv:1907.06616*.

Niu, X., Denkowski, M., & Carpuat, M. (2018). Bi-directional neural machine translation with synthetic parallel data. *arXiv preprint arXiv:1805.11213*.

Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses*: Wiley New York.

Orekondy, T., Schiele, B., & Fritz, M. (2019). *Knockoff nets: Stealing functionality of black-box models.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Ott, M., Edunov, S., Grangier, D., & Auli, M. (2018). Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). *Practical black-box attacks against machine learning.* Paper presented at the Proceedings of the 2017 ACM on Asia conference on computer and communications security.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *Bleu: a method for automatic evaluation of machine translation.* Paper presented at the Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., & Žabokrtský, Z. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications, 11*(1), 1-15.

Popović, M. (2015). *chrF: character n-gram F-score for automatic MT evaluation.* Paper presented at the Proceedings of the Tenth Workshop on Statistical Machine Translation.

Popović, M. (2017). *chrF++: words helping character n-grams.* Paper presented at the Proceedings of the second conference on machine translation.

Post, M. (2018). A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771*.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Riezler, S., & Maxwell III, J. T. (2005). *On some pitfalls in automatic evaluation and significance testing for MT.* Paper presented at the Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.

Sennrich, R., Haddow, B., & Birch, A. (2015a). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sennrich, R., Haddow, B., & Birch, A. (2015b). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.
Settles, B. (2009). Active learning literature survey.
Settles, B., & Craven, M. (2008). *An analysis of active learning strategies for sequence labeling tasks.* Paper presented at the Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.
Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3/4), 591-611.
Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). *A study of translation edit rate with targeted human annotation.* Paper presented at the Proceedings of association for machine translation in the Americas.
Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability, 57*, 1-436.
Tiedemann, J. (2016). *OPUS–parallel corpora for everyone.* Paper presented at the Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.
Wang, W., Peter, J.-T., Rosendahl, H., & Ney, H. (2016). *Character: Translation edit rate on character level.* Paper presented at the Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers.
Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin, 1*(6), 80-83. doi:10.2307/3001968
Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics* (pp. 196-202): Springer.
Zhao, Y., Zhang, H., Zhou, S., & Zhang, Z. (2020). *Active Learning Approaches to Enhancing Neural Machine Translation: An Empirical Study.* Paper presented at the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., . . . Kurakin, A. (2019). On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*.
Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*.
Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
Hu, X., Liang, L., Li, S., Deng, L., Zuo, P., Ji, Y., . . . Sherwood, T. (2020). *Deepsniffer: A dnn model extraction framework based on learning architectural hints.* Paper presented at the Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems.
Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., & Papernot, N. (2020). *High accuracy and high fidelity extraction of neural networks.* Paper presented at the 29th {USENIX} Security Symposium ({USENIX} Security 20).
Krishna, K., Tomar, G. S., Parikh, A. P., Papernot, N., & Iyyer, M. (2019). Thieves on sesame street! model extraction of bert-based apis. *arXiv preprint arXiv:1910.12366*.
Orekondy, T., Schiele, B., & Fritz, M. (2019). *Knockoff nets: Stealing functionality of black-box models.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Pal, S., Gupta, Y., Shukla, A., Kanade, A., Shevade, S., & Ganapathy, V. (2020). *Activethief: Model extraction using active learning and unannotated public data.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). *Practical black-box attacks against machine learning.* Paper presented at the Proceedings of the 2017 ACM on Asia conference on computer and communications security.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *Bleu: a method for automatic evaluation of machine translation.* Paper presented at the Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). *Stealing machine learning models via prediction apis.* Paper presented at the 25th {USENIX} Security Symposium ({USENIX} Security 16).