Sheffield Hallam University | College of Business, Technology and Engineering

MSc Dissertation Report

"*Document-level Translation and Evaluation of Parallel-Bilingual Texts using Deep Learning Techniques*"

A dissertation submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of Master of Science in Big Data Analytics

Student Name	Brian Davis
Student ID	B7015685
Supervisor	Bayode Ogunleye
Date of Submission	13th September 2021

This dissertation does NOT contain confidential material and thus can be made available to staff and students via the library.

Abstract

Neural Machine Translation is one of the most prevalent areas of research, with the main focus being translation of a source language to a target language. The aim of this paper is to understand what impact the adjustment of hyper-parameters in the training phase of model development can have on the validity and accuracy of translations. Making use of a dataset containing up to 100,000 sentence pairs, model development shows that the adjustment of hyper-parameters is not the only aspect that needs to be considered. Whilst these adjustments show small differences in training model accuracy, the use of automated evaluation metrics don't help to understand the full scale of model quality. To understand model strengths and weaknesses in further detail, the utilisation of a questionnaire to gather human feedback is created, to help show that a model that looks good when considering automated scoring, the reality shows that native speakers of the target language do not agree that the context is being maintained within these translations. This confirms that human evaluation, although seen as a time-consuming endeavour, is still a necessity. The project was able to develop a translation model with a score of 53% via Word Error Rate, with this result showing that a model with a relatively low training time can still achieve a satisfactory accuracy.

Acknowledgements

I would like to thank the tutors at Sheffield Hallam University, in particular my Supervisor Bayode Ogunleye, who supported me throughout the project at each stage. Also, to my friends and family, whose support throughout and help proofreading the report helped in finalising the project.

Contents

List of Tables

List of Figures

1.0 Introduction

One of the earliest challenges for AI was the translation of text from one language (source) to another (target). Neural Machine Translation (NMT) is an approach to Automated Machine Translation that utilises the strengths of Neural Networks to create models able to translate sentences and phrases from a source language to a target language, most commonly in a bilingual nature (one-to-one). An approach developed to improve upon Statistical Translation systems, NMT systems remain at the forefront of current advances in the Machine Translation spectrum. Machine Translation is a difficult task for an AI, due to the differential nature of human language, and the inability for an AI to understand and maintain the contextual meaning of a sentence being translated.

How can adjustments to hyper-parameters have an impact on the translation accuracy of a Neural Machine Translation model?

1.1 Project Rationale

Translation is a key tool in the modern world, but current offerings continue to miss the mark when it comes to achieving human parity (Läubli et al., 2018). One of the reasons behind this lack of clarity is Context, which is a key aspect of written communication. The problem is simple, AI is unable to understand context in a way that can be easily maintained within a translation task (Precup-Stiegelbauer, 2013). The premise of document-level NMT originally focused on sentence-to-sentence or phrase level (Bahdanau et al., 2015; Cho, van Merrienboer, Gulcehre, et al., 2014; Sutskever et al., 2014).

However, automated translation doesn’t always factor in the differences in language structure. Languages differ to us in various aspects, such as sentence structure (syntax), word structure (morphology) and vocabulary (lexicon) (Universitet Utrecht, n.d.). The problem of context arises in documents with many sentences that use similar words in the source language, where the context of the word is different over multiple sentences, but the same word is used. This could be due to the gender neutrality of the word within the sentence, or simply through the fact that some languages do not directly translate to another language, due to differences in morphology or syntax. This is a problem because translation tasks are based at the sentence level and only recently did translation tasks consider the inter-dependencies among these sentences, thus leading to the first translation meaning being considered the absolute truth (Maruf et al., 2021).

When we consider why these groups of sentences are important, this all comes down to their Discourse, which is defined as a group of sentences that are coherent, structured and contiguous in nature (Jurafsky & Martin, 2008). The research problem being addressed is, how can we utilise Deep Learning Techniques to help us improve upon the issues of contextual awareness in NMT, and how can we make use of the processes that have been considered towards the solution of this problem of contextual mis-translation.

The justification for this research focuses on Deep Learning techniques and the process of utilising human evaluation for machine translation. Automated evaluation methods such as BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) are two of the most frequently used forms of automated evaluation, due to the lack of time available to human translators (Papineni et al., 2002). Research shows that human parity hasn't been reached, and some people still wonder if the crowd alone can effectively evaluate an NMT system (Graham et al., 2017). As a result, it is imperative that further research is conducted into evaluation of an NMT system via the crowd.

1.2 Project Scope

Based on the bilingual corpora that has been obtained for this research, the project scope begins its focus on the creation of an existing Deep Learning Neural Machine Translation model architecture with utilisation of key Deep Learning techniques being a defining factor in evaluation. The scope has been refined from the initial proposal, with a key adjustment being a stronger focus on hyper-parameter manipulation to discover the best settings for the task.

Consideration for the contextual validity of the sentences must be at the forefront of the evaluative step. It is no secret that translation can lose the context of the sentence in translation, and thus the decision to maintain translation at the sentence level has been decided, as translating on a word-by-word level simply elevates the issue. The project focuses on sentence-to-sentence translation, which will utilise bilingual parallel datasets obtained from the OPUS website (OPUS, n.d.). The key aspects of the evaluation process are to be explored, utilising popular automated translation evaluation methods, and a feedback focused qualitative evaluation method.

1.3 Project Aims

The main aims of the project are defined as research questions. This project contains a main research question, along with a secondary sub-question. These are.

1) How can we effectively use existing Deep Learning techniques to improve translation quality of Bilingual Parallel Corpora?
2) How can we make use of Qualitative techniques to verify Neural Machine Translation model performance?

Considering the key areas of this project based on initial research undertaken and the initial research questions defined above, the following have been decided upon for the project. These are to.

- Implement an existing Neural Machine Translation model architecture
- Enhance performance through utilisation of Deep Learning technologies
- Evaluate the model to discover strengths and weaknesses
- Consider adjustments required to improve the model
- Obtain Qualitative feedback to evaluate performance

The overall purpose of the project is to develop a Machine Translation model that can successfully translate input from source to target. The chosen source language for this is English, with the chosen target language being Chinese. The primary output is a usable Translation model that can translate input in the form of English text, into accurate Chinese text. The final aim is to evaluate the model based on the feedback of users, utilising online questionnaires to gather feedback of translation quality, before presenting the results of this feedback-focused evaluation method, and providing recommendations for continued future improvements.

1.4 Project Objectives

To enable the project aims to be realised, the project objectives are:

- Identify current research / current challenges of Neural Machine Translation
- Identify valid sources of usable parallel bilingual data for model development
- Develop model utilising model properties to enhance outcome
- Evaluate performance of the model utilising deep learning techniques
- Obtain Qualitative feedback from participants to measure accuracy of developed model
- Perform final evaluation and discuss key findings

The initial objective is to identify and review current research, current approaches, and existing gaps where new techniques could be utilised to improve translation accuracy. The next objective is to identify valid sources for data and validate these within a

Python IDE for their suitability and general accuracy. Suitable data then enables progress toward the next objective, which is to develop an NMT, with performance evaluation being the next objective. The feedback generation through conducting a survey will help to fulfil the next objective, which is to measure accuracy of the developed model via the crowd. The feedback will be gathered by utilising translations to see how accurate the model has been. The final objective is to perform final evaluations and present the key findings within this paper.

1.5 Project Benefits

As this project will be completed in a short period of time, the scale of the model will be unable to match the entirety of current research within the field. Therefore, the project offers the following short-term benefits:

- Offer new approach to fill knowledge gap
- Use alternative hyper-parameters to discover how accuracy and speed of learning differs
- Utilise Qualitative research methodology to evaluate a translation model
- Identify possibilities for further research in the field

The benefits listed here focus on the development of the solution, looking at some alternative approaches to those being utilised in the field at the time of this project. Small adjustments can make a big difference in the world of Data Science, which is what this research wants to discover. If small changes are made, how will the results reflect on this change, and will there be an improvement or a decline. Once the model is created, using a qualitative style of evaluation will allow real world feedback to help understand possible shortcomings.

1.6 Achieving the Research Objectives

The initial proposal for this project can be found within the appendix of the paper (Appendix A), where appropriate ethical approval was obtained before any data collection commenced (Appendix B). The project flow (figure 1) follows a similar design to that seen in the time plan attached within the project proposal. The project begins with research and resource gathering, which underpins the literature search and data collection. The flow then continues with the model development, which involves the creation, validation, implementation, and evaluation of the NMT. After this is done, the project then moves on to the Surveying, which involves questionnaire research,

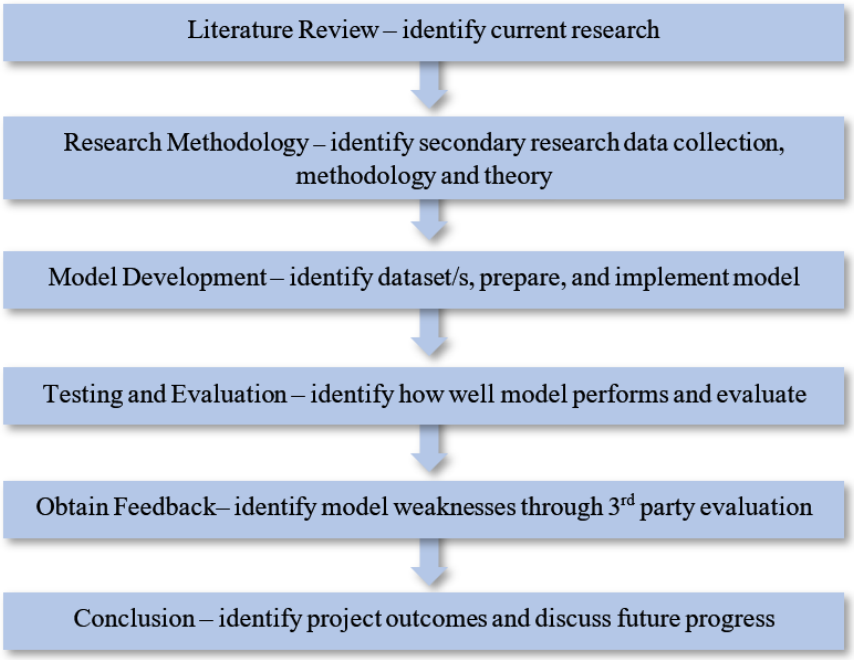
design, and distribution, to gather feedback from participants regarding the suitability of the created model and the decided upon implementation methods. Evaluation follows on from this stage, which concludes with the findings and recommendations.

1.7 Thesis Overview

This chapter defines the structure of the paper. Section 1 introduces the project, covering the rationale for conducting this research, the scope of the project, as well as the aims, objectives, and benefits, concluded with how these will all be met within the project. Section 2 contains the literature review that discusses the background of Machine Translation (MT), focused on Text Alignment, Encoder-Decoder Framework, and the Out-of-Vocabulary (OOV) words problem.

Section 3 discusses the research methodology, diving deeper into the secondary research data chosen to be utilised, covering some ethics procedures that must be followed and talking about the methodology. Section 4 aims to discuss the development of the NMT model, how this model was trained and evaluated. Section 5 aims to present the results and analysis of the NMT system. Section 6 will present the user feedback collected through surveying means, present the survey ethics considered during the survey creation and distribution, and will discuss the evaluation of the model based on the findings of the survey. Section 7 will conclude the paper, with a critical analysis, evaluation of the research, and the conclusion. Finally, the next steps for the research based on the findings this paper has presented will be discussed.

Figure 1 – Project Flow Overview Diagram



2.0 Literature Review

The literature review focuses on the process of effective model development and methods for evaluation of Translation systems, with a focus being on Neural Network based approaches. The background for Machine Translation systems was researched as part of this process, with the history and progression within Machine Translation being a key area of interest. Text alignment, model development which focused on encoder-decoder and transformer, attention-based development, and out-of-vocabulary words were the most important considerations.

2.1 Text Alignment

2.1.1 Statistical Alignment

When we consider the earliest work in machine translation, it was P. F. Brown et al. (1990) that began the fundamental idea of using statistics within a model to translate text from one language to another. This initial approach made use of sentences, as it was thought that "*every sentence in one language is a possible translation of any other sentence in the other*". Earliest statistical models required the use of probability to accurately estimate, choosing the translation with the highest probability score as the one with the highest chance of being the successful translation of the source.

This probability was used to align sentences to their equivalent expected translation through a language model, but at times no equivalent word would be found in the translated text, meaning that some words weren't aligned accurately. Words near the beginning of the source sentence had tendency to align with those at the beginning of the target sentence (Brown et al., 1990). The flaw in this early work is that probability can be skewed toward a particular direction, which means incorrect translations are entirely possible, especially if the linguistic of the sentence mean the beginning of the source sentence doesn't translate in the same order, due to structural differences.

Gale & Church (1993) introduced a model that was based on character lengths to create the alignment. The alignment processes took place in two stages. First, paragraphs are aligned, and then the sentences within the paragraph are aligned further. This method heralded a 96% success rate, with only a meagre 4% of sentences not being aligned correctly. The end goal is to find correspondence between words, but to do this, they had to first align at a higher level.

2.1.2 Alignment with Mixture Distribution

Considerations for alignment continued on, with Vogel et al. (1996) creating a solution involving a Hidden Markov Model (HMM) first derived by Jelinek (1976) toward the task of speech recognition. It is possible to notice that a lot of the work originally carried out for the task of speech recognition was able to be transferred over to statistical modelling solutions. This HMM model is very similar to those derived for speech recognition, but not entirely identical (Vogel et al., 1996).

Alignment originally focused on alignment probabilities with consideration of the absolute position, but Vogel et al. (1996) decided to focus efforts on consideration for the relative position instead. The HMM model was compared with a mixture-based model introduced by P. F. Brown et al. (1990). The mixture-based model utilised sentence length probability, mixture alignment probability and translation probability, which when trained, amounted to a sequence of positional alignment and parameter estimation as its two steps for training criterion.

The motivation for the HMM model was that there is typically a strong localisation effect in the aligning of words within parallel texts, and whilst not distributed arbitrarily, they do tend to form clusters. Whilst the idea of the HMM model was an interesting concept, the results of it didn't better the current mixture-based model of the time.

2.1.3 Inverted Alignment Model

Earlier work by Tillmann et al. (1997) had proposed a search algorithm, based on dynamic programming that examines the source string sequentially (Nießen et al., 1998). A so called inverted alignment model was thought of and experimented with by Nießen et al. (1998) and was found to be reasonable. With its recursive formulae and acceleration techniques, this IA model, tested on the Verbmobil Corpus, proved this style of model was applicable to real-world translation ability, but better perplexity scores were discovered through the use of the HMM model created by Vogel et al. (1996) on the same Corpus.

2.1.4 Extending the Baseline

Och et al. (1999) improved on the work of Nießen et al. (1998) and Vogel et al. (1996), with the prime focus being Word-to-Word statistical translation models utilising two differing approaches. The first approach was based on dependencies between single

words, with a second approach focusing on phrase-based alignment (between phrases) and word level alignment (between words within the phrases).

The problem with word-to-word correspondence within Statistical Machine Translation, is that a word in the source is assigned exactly *one* target word. This is evident in models by (Vogel et al., 1996), (Tillmann et al., 1997) and (Brown et al., 1993). This problem is due to the way mapping is done within the models, which is mapped as $j \rightarrow i = aj$, from source position j to target position $i = aj$ (Och et al., 1999).

The possible alignments that could be utilised by the HMM can be seen in figure 2. The concept of monotone alignments allows a search procedure to be formulated, equivalent to finding the best path through a translation lattice (Och et al., 1999).

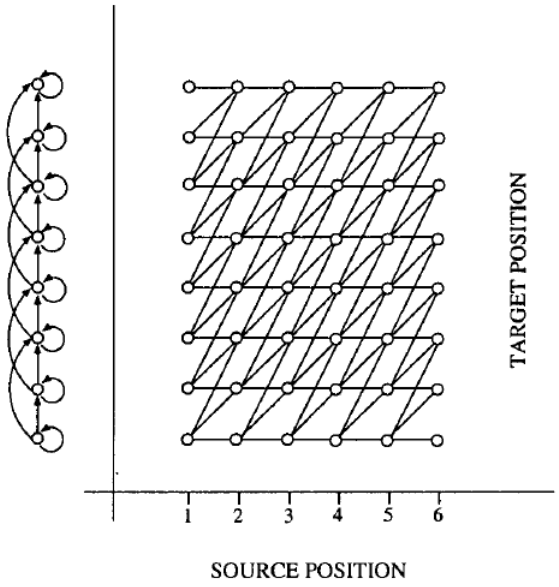


Figure 1 – Illustration of alignments for the monotone HMM (Och et al., 1999; Tillmann et al., 1997)

Monotonicity is a problem for translation, as there may be instances where a single word in one language, translates to multiple words in the other language. To counteract this, Och et al. (1999) created an extension that could handle non-monotonicity, which first had to assume the alignment to be monotone and respect the word order for the majority share of the alignments.

The DP search utilised a left-to-right beam search, first used within speech recognition by Lowerre (1976). Two HMM alignment models like Vogel et al. (1996) were used for the two translation directions, but maximum approximation was not applied, leading to a slightly improved alignment on both sides.

2.1.5 Sentence Length-Based Alignment

P. F. Brown et al. (1991) and Gale & Church. (1993) took the idea of modelling the relationship between the length of sentences that are considered mutual translations. A new method was then created by Chen (1993) that was based on optimising word-translation probabilities, but the result was much slower than that of P. F. Brown and

Gale & Church. Another version was then researched by Wu (1994) that utilised lexical cues on Chinese translation, as the use of only sentence lengths doesn't work too well on Chinese text, as Chinese is not a language where words are split with spaces like in many other common western languages such as French or English.

The issue with earlier alignment techniques is that they require some particular knowledge of the corpus or the languages involved, which is not always going to be the case (Moore, 2002). This is evident in the use of anchor points through the implementations of (Brown et al., 1991; Gale & Church, 1993), where either this or prior alignment is required.

P. F. Brown et al. (1991) believed that alignment of sentences is entirely possible through making use of 'beads', which is the idea that sentences align 1-to-1, 1-to-2, 2-to-1, or in some cases 2-to-2 or even higher, but Wu (1994) showed that length based alignment using beads for English -> Chinese resulted in 100% incorrect alignment for any sentences where the beads were 2-to-2, 1-to-3, 3-to-1 or 3-to-3, meaning that the upper limit for this language pairing was 2-to-1 to obtain highly accurate length based alignment (95.2%).

Moore (2002) developed an algorithm which made use of many of the previous work done before and adapted it into a single combined approach. This was done using a three-step process which involved sentence-length-based-alignment, word-translation-model, and word-correspondence-based-alignment. The result of the implementation was the ability to use the highest probability 1-to-1 beads from the initial alignment to train a word-translation model. It was shown that word correspondence can be utilised to produce higher-accuracy sentence alignment, and that the older approaches of anchor points and a bilingual lexicon, along with a prior knowledge corpus is not entirely required if the computational cost is increased in the development phase of model implementation.

2.1.6 Attention for Alignment

Global Attention is an attention-based approach to alignment that has become popularised through the creation of the Neural Machine Translation (NMT) model. First introduced by Kalchbrenner & Blunsom (2013) as a recurrent continuous translation model, the NMT has quickly become the go-to standard for machine translation. This first model didn't rely on alignments to generate results, but was sensitive to word order, syntax, and the meaning of the source sentence.

Global Attention helps the decoder to decide which part of the sentence it needs to pay attention to. It can consider all hidden states from the encoder when deriving context, which is something that can be easily lost in translation otherwise. Context is important, since if contextual meaning is not at the forefront of the translation, words with multiple meanings can be translated incorrectly, and the meaning can be lost (Bahdanau et al., 2015; Luong, Pham, et al., 2015; Precup-Stiegelbauer, 2013).

Most recent alignment came within the Transformer model. Vaswani et al. (2017) introduced a transduction model able to align based on a recurrent attention mechanism, instead of the more traditional approach of sequence-alignment seen in earlier work. Comparing attention to traditional alignment helps us to understand what attention is paying attention to. Attention is in some ways different to alignment, but is also capturing other useful information than just alignments (Ghader & Monz, 2017).

2.2 Translation System Frameworks

2.2.1 LSTM based Neural Machine Translation systems

The Neural Machine Translation model was first envisioned within a translation task by Sutskever et al. (2014). Working with the WMT ’14 dataset of English to French, this first work made use of the LSTM (Long Short-Term Memory) architecture (Hochreiter & Schmidhuber, 1997), with a key aspect being the orientation of the input sentence, which was read by the LSTM in reverse to introduce short term dependencies which were deemed to improve upon the optimisation problem. This architecture can be seen in figure 3.

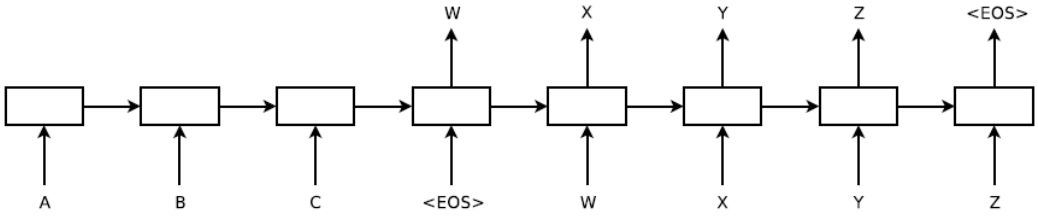


Figure 2 - LSTM model to read in an input sentence and output the target sentence (Sutskever et al., 2014)

Kalchbrenner & Blunsom (2013) did offer a similar approach to LSTMs, but utilised Convolutional n-grams for their model, which mapped the sequence to vectors. The issue with this approach was that order of the sentence was lost due to the mapping procedure chosen. The utilisation of two LSTMs, one for the source and another for the target, began the idea of the Encoder-Decoder.

2.2.2 NMT Encoder-Decoder

The Encoder-Decoder framework didn't really come to light until the work of Cho et al. (2014). The focus of this approach was phrase representations within Statistical Machine Translation systems. The first Encoder-Decoder model uses Recurrent Neural Networks (RNNs) that act as the Encoder-Decoder pair. With the strength of the RNN being its ability to remember sequences, it is a useful and preferred approach for many Natural Language Processing (NLP) tasks (MIAP, 2021).

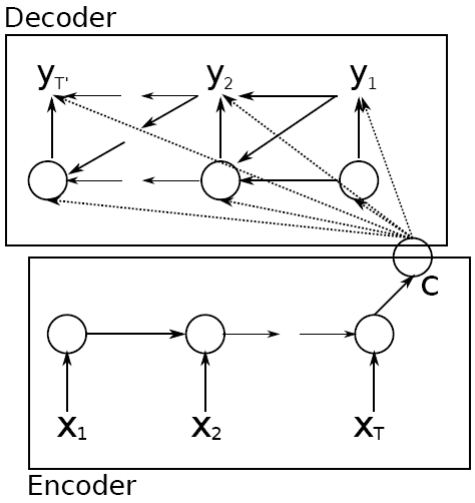


Figure 3- Illustration of the RNN Encoder - Decoder (Cho et al., 2014)

Figure 4 shows the initial proposed design of the Encoder-Decoder. The model takes in a phrase sequence and then uses an RNN to exhibit a behaviour like the human brain, where a prediction is produced in sequential order. The decoder is trained to predict the next word given its context vector, taking into consideration all previously predicted words (Bahdanau et al., 2015).

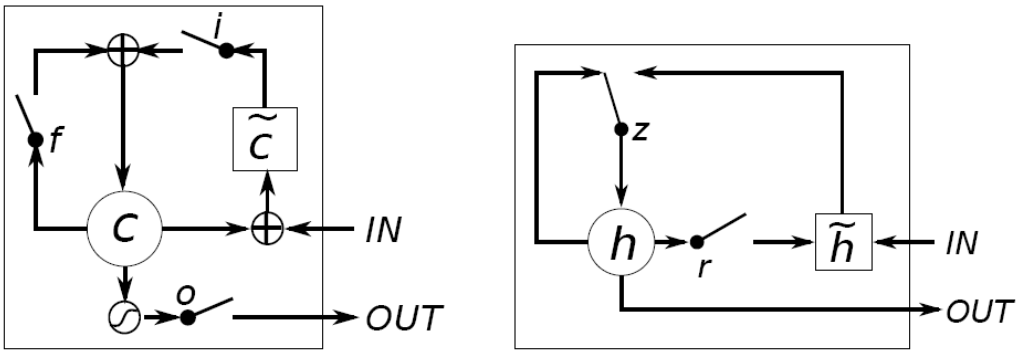
2.2.3 Creation of the Attention Mechanism

One of the big issues with Translation models is model degradation. It is possible to overcome this issue using a neural model of attention. By allowing the decoder to have an attention mechanism, this relieves the encoder from having to encode all the information within the source sentence into a fixed-length vector (Bahdanau et al., 2015).

The grConv model was an attempt to create a system that was based purely on neural networks. Despite its radical differences, the grConv still suffers from what can be called the 'curse of sentence length', where model performance degrades as the sentence length increases. This new style of model was able to mimic grammatical structure without supervision (Cho, van Merrienboer, Bahdanau, et al., 2014).

Although the LSTM was a popular choice for RNNs, it wasn't the only choice that was available for the gated mechanism. The Gated Recurrent Unit (GRU), found to be comparable to that of the LSTM, was first experimented with by Cho, van Merrienboer, Gulcehre, et al. (2014). The GRU has gating units that modulate the flow of information inside the unit, without the need for separate memory cells (Chung et al., 2014).

The evaluation of LSTM vs GRU on sequential modelling yielded interesting results. Figure 5 shows an illustration to showcase the similarities of the two units. Whilst it is possible to see the improvement of the LSTM and GRU vs a more traditional recurrent unit, the choice of which performs better depends upon the task at hand, where the choice really depends on the dataset being used for model training (Chung et al., 2014).



(a) Long Short-Term Memory (b) Gated Recurrent Unit

Figure 4 - Illustration of LSTM (a) vs GRU (b) (Chung et al., 2014)

The concept of Attention is something that helped to springboard new ideas for NMT alignment architecture, which can be achieved through different modalities. First used for image objects and agent actions by Mnih et al. (2014), this concept was then successfully applied to NMT by Bahdanau et al. (2015), where the attentional mechanism was applied to teach the model to jointly align and translate.

2.2.4 Different Approaches to Attention

Attention models can be categorised as Global and Local. The difference is whether the attention is placed on all source positions or on a few of the source positions. Global Attention is an attentional model that considers all hidden states of the encoder when it derives the context vector, inferring a variable length alignment weight based on the current target state. However, Local Attention differs in that it first predicts a single aligned position for the current target word and then computes the weighted average of the context vector (Luong, Pham, et al., 2015). Comparisons between the Encoder-Decoder (Cho, van Merrienboer, Gulcehre, et al., 2014) and RNNsearch (Bahdanau et al., 2015) were done through the use of sentences of length up to 30 words and 50 words respectively, with model translation accuracy being initially ascertained by the use of beam search similar to that used by Sutskever et al (2014).

The RNNsearch model (Bahdanau et al., 2015) was initially proposed due to the use of a fixed-length context vector being applied within Encoder-Decoder (Cho, van Merrienboer, Gulcehre, et al., 2014) at the time. As the length of the sentences in the training set increase, the accuracy of Encoder-Decoder begins to degrade over time, a property that was improved upon by RNNsearch due to its fixed-length context vector limitation.

GNMT was designed with the aim to bridge the gap between humans and machines. Since inception the NMT has some obvious flaws, such as slower training and inference speed, alongside severe weakness to rare words and occasional inability to translate all words within a sequence. Utilizing 8 deep LSTM layers within the encoder and decoder respectively, whilst also connecting the decoders bottom layer directly to the encoder, the GNMT was able to severely improve arithmetic time and inference speed (Y. Wu et al., 2016).

GNMT was able to achieve competitive results, reducing translation errors by 60% with help from a human evaluation alongside automated methods. A novel Neural Network, called ByteNet, was introduced by (Kalchbrenner et al., 2017) shortly after. The advantage of ByteNet was the ability of the decoder to be able to generate variable-length target sequences through dynamic unfolding, which allows the network to process source and target sequences more efficiently.

2.2.5 The Transformer Model

Work begun to slowly move away from working with sole RNN based approaches that use attention as a primary way to align and began to take advantage of Convolutional layers. One of the first examples of this was the Transformer Model (Vaswani et al., 2017), seen in figure 6. This was designed to remove the computational shortcomings of RNNs, as RNNs are computationally ineffective in comparison due to sequential issues. This was done by removing the need for the encoder and decoder and focusing solely on a model that uses the attention mechanism, using stacked self-attention layers.

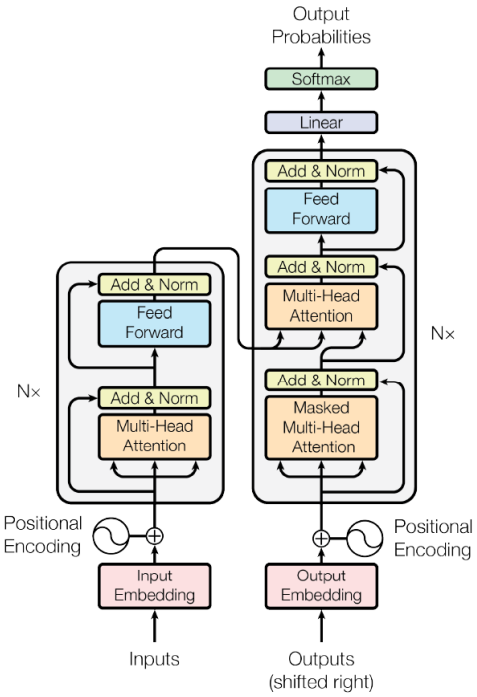


Figure 5 - Transformer Model architecture (Vaswani et al., 2017)

To further improve upon sequential issues encountered within the RNN models, sequence-to-sequence learning was introduced. Although sequence learning is not an entirely new concept in itself, the use of Convolutional layers within a sequence-to-sequence model hadn't been fully researched. Gehring et al. (2017) introduced the architecture of convolutional sequence-to-sequence learning, which utilised Convolutional Neural Networks (CNNs) to create a fully parallelised approach.

The biggest issue with the sequence-to-sequence learning models, is the length of time (in days) they typically take to complete, with the model from Gehring et al. (2017) taking 6 to 7.5 days to complete training on a single GPU. The work of Ott et al. (2018) mainly aimed to show that it is possible to increase time taken to achieve state-of-the-art models simply by upscaling resources. The adjustments made allowed training of the Transformer model (Vaswani et al., 2017) to convergence in 85 minutes, generating a new state-of-the-art BLEU (Papineni et al., 2002) score.

Development of the Tensor2Tensor library, developed within Tensorflow (Abadi et al., 2016) began to make deep learning research both more accessible and faster for the creation of deep learning models. With a library containing many of the most popular NMT models and datasets, research into NMT was able to advance at a much faster pace than ever before (Vaswani et al., 2018).

2.2.6 The Focus becomes Contextual

K. Chen et al. (2019) proposed a double context NMT architecture, consisting of a global context vector and a syntax-directed local context vector, to help improve translation performance via source representation. This was done via extension of the local attention with syntax distance constraints, to allow the model to capture context related source words. Syntax-directed attention (SDAtt), when compared to Global and Local attention (Bahdanau et al., 2015; Luong, Pham, et al., 2015), produced improved results, although these were only minimal when it came to BLEU evaluation.

To continue to improve NMT and reduce inconsistencies that can crop up during training due to a model learning errors from sequence generation tasks, where early errors in generation harm further sequence generation (Zhang et al., 2019), a technique called Target Bidirectional Agreement was introduced by L. Liu et al. (2016), which attempts to leverage an additional NMT model that trains Right-to-Left (R2L) instead of the traditional written direction of Left-to-Right (L2R).

Although the work of L. Liu et al. (2016) only used this additional R2L model to re-rank the translations that were generated by the L2R model, Zhang et al. (2019) took into consideration the agreement between the L2R and R2L models into their training objectives to generate good prefixes and suffixes. The use of an optimisation regulator to enable fast training is proposed to jointly optimise L2R and R2L models.

2.2.7 Deep Models

The Deep Transformer model (Bapna et al., 2018) is a continuation of the Transformer model (Vaswani et al., 2017). Development of deep networks like the Deep Transformer model are difficult to optimise, a problem which can't be simply solved through stacking more layers. Primary focus began on the decoder, but recently the focus shifted toward the encoder, deemed to require a lower computation cost (Domhan, 2018). Optimisation can go smoothly by simply moving the layer normalisation unit within the model, which is key to successful optimisation when many layers are present. A 30-layer Deep Transformer was established by Wang et al. (2019), matching and in some cases managing to surpass that of (Bapna et al., 2018), requiring 3x fewer training epochs and performing 10% faster for inference.

The use of an ensemble to train a Transformer model has shown that the L2RxR2L approach for joint optimisation has the potential to improve upon translation quality going forward (S. Wu et al., 2020; Zhang et al., 2019). Work moves toward continued research of deeper models, utilising a deep encoder with shallow decoders, which is a requirement to continue to develop faster and more accurate multilingual translation models.

2.3 Out-of-vocabulary words

2.3.1 The Rare-Word Problem

Out-of-vocabulary (OOV) words refer to the words that can be found in one language but not in the other. These can cause a problem for Translation systems, as OOV words can hold key information into the context of the sentence or phrase, and without them the context can be lost (Arthur et al., 2016). Automated translation that doesn't factor in OOV words can lead to issues with translation quality due to loss of context, something that a machine can't factor into translation like a human translator is able to (Precup-Stiegelbauer, 2013).

The factorisation of OOV words began due to the emergence of NMTs. Utilisation of a large vocabulary in NMT was the aim of Jean et al. (2015), who made of this within

NMT systems to create a model that had no real performance drop, and led to performance on par with the NMT systems of the time (Bahdanau et al., 2015; Cho, van Merrienboer, Bahdanau, et al., 2014; Sutskever et al., 2014). The results were obtained through model training, as the use of a large vocabulary meant that Jean et al. (2015) didn't require a large change in model complexity.

To factor in OOV words into the model, it is required to have some sort of placeholder variable that will allow the model to establish which words are OOV and which ones are not. Use of the UNK variable is preferred to allow a model to understand which words in the vocabulary fall into this category (Luong, Sutskever, et al., 2015; Sutskever et al., 2014). Implementation of an NMT system with a word alignment algorithm which allowed for emittance of the position of the OOV word in the source sentence, helps the NMT system achieve good results whilst also providing some mitigation to the OOV issue (Luong, Sutskever, et al., 2015).

2.3.2 Low Context Words

Whilst OOV is an issue that continues to be evaluated, the frequency of low context words also has its place as one of the challenges in the creation of the NMT system. Lexicons offer a way to help predict the next word in a sequence, by making use of the attention vector of the NMT model (Arthur et al., 2016).

NMT systems are vulnerable to rare words (OOV) but the plethora of rare words we can expect to find, all depends on the domain we are working in, as you would expect more rare words in a domain such as medicine than you would in news commentary, due to the vocabulary of the domain (B. Liu & Huang, 2020). A higher frequency of rare words will result in a bigger issue of generalisation. An attempt to utilise OOV word recovery, where the words that are deemed missing are recovered in a written form and changed back into vocabulary words is a process that has been explored as a possible solution by Qin (2013), who discussed the process of identifying concurrent OOV words through bottom clustering, and then attempted to recover these through phenome-to-grapheme (P2G) conversion. (Need to find more info on what this is, isn't explained in paper)

A fixed vocabulary is common within NMT system translation, however, the inclusion of OOV words creates this issue where these OOV words can't be ignored. A simpler and more effective approach, which considers the problem of translation as more of an open-vocabulary problem, allows these words to be encoded as words that contain sub-

word units (Sennrich et al., 2016). Copying unknown words into the target text (Jean et al., 2015) is a reasonable approach, but transliteration and morphological changes are required for this strategy to be effective.

Berrichi & Mazroui (2021) recently explored an interesting approach, considering a series of different experiments where morphosyntactic features were factoring into the model training step. Features such as Part-of-Speech (POS) were used to help with the OOV problem. Morphosyntactic features are an interesting approach that is an interesting idea to consider focused interest on, to see if further improvements can be made from this angle.

2.4 Literature Review Summary

The review summarises the key focus areas that must be considered when developing an NMT model. The Encoder-Decoder (Cho, van Merrienboer, Bahdanau, et al., 2014) was the first real approach at the development of an NMT system, with enhancements taking us ever closer to reaching human parity. The Attention mechanism (Bahdanau et al., 2015) followed by the Transformer (Vaswani et al., 2017) have shown what is possible with the NMT.

The issue of OOV words is another area that requires attention. With NMT systems showing real weakness to rare words, further consideration must always be at the forefront of any model development, but the domain must be considered as well when taking this into account (Arthur et al., 2016; B. Liu & Huang, 2020). Utilisation of the UNK token to define OOV words is a necessary step to help alleviate issues caused by rare words in the vocabulary (Luong, Sutskever, et al., 2015).

3.0 Research Methodology

This section will introduce the research methodology. The focus of discussion is the chosen secondary dataset/s, how these were evaluated before being pre-processed, alongside further discussion involving the key processes involved in the task of Natural Language Processing with Neural Machine Translation in mind, alongside further discussion regarding the overall theory of the topic and the model chosen for implementation.

3.1 Research Data Collection

The primary requirement when looking for datasets for this project was sentence maintained parallel raw accuracy deemed suitable for training a translation model. Due to the length of time of this project, the dataset itself could not be to the same scale as some other NMT tasks have been in the past, due to the time required for model training. A few options were found and analysed toward their suitability. Alongside secondary data collection, the process of primary data collection was also considered.

Consideration had to be made for how primary data was going to be obtained for the human evaluation of the model. It was deemed to be the best approach to go along with an anonymous questionnaire, allowing respondents to answer a series of questions regarding the validity of translation results obtained by the model, with a Likert scale being the best option for this. Secondary datasets of varying lengths were obtained from the OPUS open parallel corpus (OPUS, n.d.), with additional sentence pairs found from the Tatoeba Project via manythings.org, containing a large variety of languages to a much lower and richer scale. The selected datasets for this project are sentence pairs with no primary domain focus, with variable lengths.

3.2 Research Method and Design

When we evaluate NMT systems, the most common approach utilised is automated, due to the time-limited availability of interpreters, and the duration required for human translation evaluation (Papineni et al., 2002), which was the driving force behind the creation of automated approaches. However, it is shown that we can also use the crowd to help evaluate the work of the translation model, and that the results can be useful in the context of evaluation (Graham et al., 2017).

This research project utilises the Mixed Methods Research methodology, which allows for the combination of Qualitative (QUAL) and Quantitative (QUAN) research

techniques. This research method is deemed the most appropriate due to the origins of the data sources that will be used within this research. The mixed methods methodology is being used here to coincide with the use of the Pragmatism philosophy, as it helps to enable practicality and awards multiple viewpoints (Creswell, 2012; Fischler, n.d.), which is crucial for the evaluation steps of the research, which focus on automated practices, deep learning evaluative techniques, and utilisation of a questionnaire for via-the-crowd evaluation.

The Methodology is justified by the approach of the research, which is to first run performance tests to understand the model outcome against the training and test data, with a follow up evaluation done using a questionnaire, as a way to build up from one phase to another, or in the case of this project, to answer the research question and

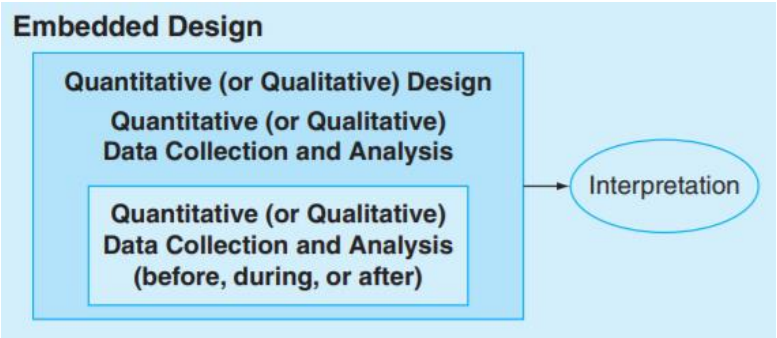


Figure 6 - Embedded Design model (Creswell, 2012)

accompanying sub-question (Creswell, 2012). This research focuses on the Embedded Design as the mixed method model deemed most suitable.

The Embedded Design mixed methods model (see figure 7) as the chosen approach lends itself well with the goal of making data collection flexible and adaptive (Creswell, 2012; Fischler, n.d.). The separation of the QUAN and QUAL data helps to improve the project flow significantly when preparing and conducting data collection. Much like an Explanatory Sequential Design in that our QUAN data is collected sequentially with QUAL data, however, the Embedded Design has the added benefit of being flexible and adaptable if it is deemed necessary for the research to require a secondary dataset to be collected in a parallel nature, also requiring less resources to accomplish this additional data collection (Fischler, n.d.).

To enable the use of an Embedded Design model, the research question was devised as a single entity, and then split into a research question and an accompanying sub-question, which allows for a mixed methods methodology to be utilised over a sequential or parallel strategy, dependant on the needs of the research. The two datasets

will be collected at different intervals in the project, and then analysed separately to create a clear distinction between what problem each dataset is attempting to answer (Creswell, 2012).

3.3 The Model

This research focuses on the Transformer model (Vaswani et al., 2017) as the baseline model in use. The Transformer is chosen due to its strength as an all-round solution for the NMT task, remaining as one of the most popular options for NMT research. For this research, due to the short time constraints, it was decided that this model would be the best choice, as the speed of implementation and testing could be streamlined to allow for completion in the necessary timeframe. As this model was implemented to promote a more parallelized approach, there was interest to implement a model that could be improved by going wider, rather than going deeper, as is the case with other Neural Network based approaches.

Most research has maintained its focus on how to utilise the Transformer and other seq2seq models to achieve higher accuracy translation solutions through sophisticated adjustments to current implementations (Bao et al., 2021; Mallick et al., 2021; Wang et al., 2019) and through further implementation of deep layers (Bapna et al., 2018; Kong et al., 2021; Wang et al., 2019).

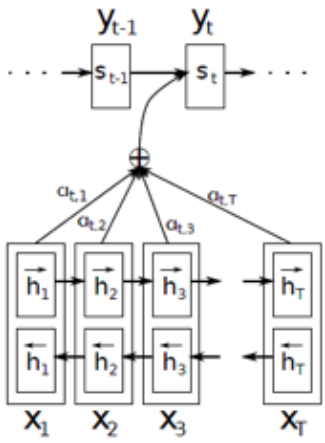


Figure 8 - The Attention Mechanism (Bahdanau et al, 2015)

The Attention Mechanism is one of the primary reasons behind the success of the Transformer model. Introduced by Bahdanau et al, (2015), the Attention Mechanism (see figure 8) is the machine equivalent of cognitive attention. Due to the problem known as “long-range dependency” within RNN/LSTM based NMT systems, Attention was introduced to solve the length dependent issue plagued by the RNN Encoder-Decoder network (Cho, van Merrienboer, Bahdanau, et al., 2014).

The attention mechanism found within the Transformer is “self-attention”, a mechanism for relating different positions of a single sequence or sentence to gain a more vivid representation. Shown in figure 6, the structure of the Transformer includes three Multi-

Headed Attention layers within the architecture, utilising stacked self-attention and point-wise fully connected layers (Vaswani et al., 2017).

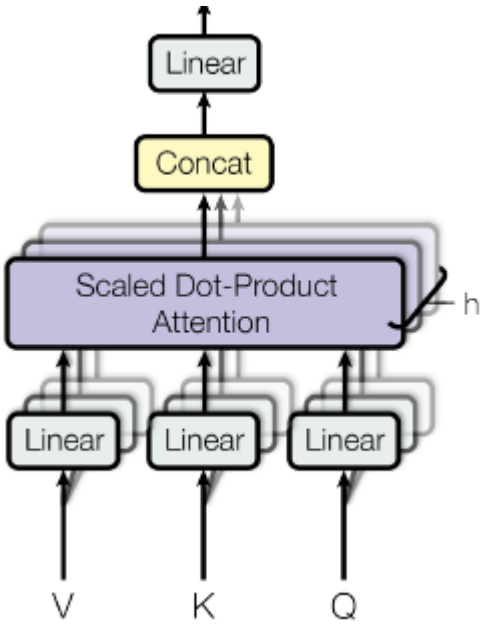


Figure 9 – Multi-Head Attention (Vaswani et al, 2017)

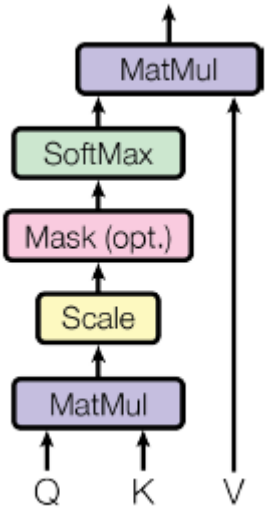


Figure 10 - Scaled Dot-Product Attention (Vaswani et al, 2017)

3.4 In the Context of Natural Language Processing

The Multi-Head Attention (figure 9) consists of several attention layers running in parallel. To calculate the attention for a word, such as 'it' within the context of a sentence that includes the word 'animal' at the start, the mechanism will take the dot product of the embedding of 'it' with the embedding of each word before it, including 'animal' and look at the corresponding vectors, which are *Key*, *Value*, and *Query*, which are derived through matrix multiplication.

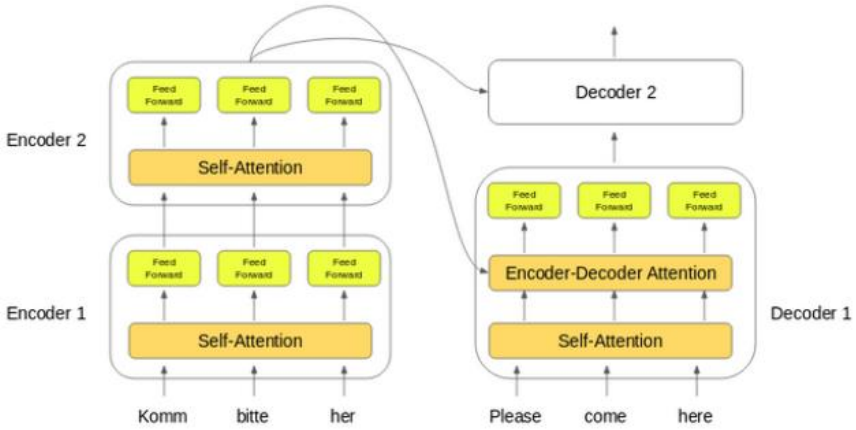


Figure 11 - The Transformer Encoder-Decoder Stack (Joshi, 2019)

The Encoder-Decoder within the Vaswani et al, (2017) Transformer consists of 6 identical stacked layers. Figure 11 shows a representation of these stacked layers.

Shown in figure 10, the Scaled Dot-Product Attention ('*Self Attention'*) consists of Queries and Keys of dimension d_k and values of dimension d_v. After this, the dot-product of the queries is computed with all keys, dividing each by $\sqrt{dk}$, then the application of a softmax function is used to obtain the weights of the values. The keys and values are packed together into matrices K and V, as is the computation of the attention function into a matrix Q (Vaswani et al., 2017).

The computation of the matrix of outputs is:

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

Multi-Head Attention (figure 9) is performed via linear projection of the queries, keys, and values *h* times, using differently learnt linear projections to d_k, d_k and d_v dimensions respectively. The attention function is performed in parallel, with which the dimensional values are concatenated and again projected. The score for each word in the sequence is calculated through taking the dot-product of the query vector Q, with the key vectors **k1…kx** of all other words. These scores are then divided by the square root of the dimension of the key vector, and then normalised with the softmax activation function. The normalised scores are multiplied then by the value vectors **v1…vx**, where the sum of these vectors arrives at the final vector. This is finally passed to the feed forward network as the input. This is shown in more detail in figure 12.

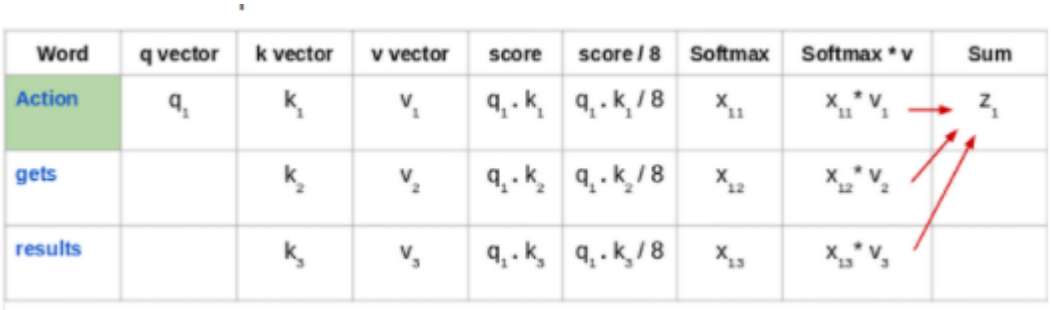


Figure 12 - Self-Attention Example (Joshi, 2019)

3.5 Data Pre-Processing

When working with Neural Networks for the task of Natural Language Processing (NLP), we can't use plain text to train a model, as computers don't understand human-readable text. We must first engage in data pre-processing to prepare our data for the purpose of model training and testing. We can think about the process of text pre-processing as *Task = Approach + Domain*. As an example, we want to extract the top

keywords from a document using an *approach* such as TF-IDF (Sammut & Webb, 2010) from Tweets via Twitter (*domain*), which classifies as our *Task*.

3.5.1 Must-Do Pre-processing steps

The two key ***must-do*** pre-process steps for NLP are Noise Removal and Lowercasing. Something that is sometimes overlooked, the process of lowercasing all of text within your corpus is incredibly important, as without this, some words may be perceived as another word entirely, especially if your analysis involves counting the frequency of words within your corpus. Figure 13 shows an example, where the word Canada, is perceived as three different words before it is lowercased.

Raw	Lowercased
Canada CanadA CANADA	canada
TOMCAT Tomcat toMcat	tomcat

Figure 13 – Raw vs Lowercased Plain Text (Ganesan, 2019)

Noise removal is the process of removing characters that might interfere with your analysis. This step is fairly domain dependant, as what you would deem to be noise all depends on the domain itself. A good example of noise is the "#" symbol found in tweets when extracting data from twitter, or text such as "**2.blossoms***",* where a number and dot precede the word.

3.5.2 Other Pre-Processing Steps Worth Considering

Some of the other pre-processing techniques commonly utilised for NLP are dependent on the task, such as Stopwords Removal, which is a key technique used when conducting Sentiment Analysis. For the process of NMT, the other key techniques that are commonly used, are Normalisation and Text Enrichment. The process of Normalisation involves the standardisation of words that are familiar to other words in the domain and is generally undertaken with the help of dictionary mappings and spelling correction indexing. The technique of Text Enrichment is another powerful pre-processing technique, but fairly task dependant as well. It is the process of providing stronger semantics to your texts, enriching it with information it didn't previously have.

3.6 Making Our Text Computer-Readable

NMT models, such as the Transformer, process raw text in a tokenised form. Tokenization is the process of converting raw text input into machine-readable numeric values. The processing of raw text at the token level is a procedure done by the most popular NLP architectures, such as RNN, GRU and LSTM. Tokens themselves, can be

either words, characters or subwords (n-gram characters). We can classify the process of Tokenization into these three sub-categories.

3.6.1 Word Tokenization

Word tokenization is the most common tokenization strategy implemented into NLP, where the sentences are split at the word-level, with each individual word being assigned a numeric token. This information is then saved as a vocabulary within the Tokenizer and used as a map to allow for reverse-tokenization to be implemented, which allows for the presentation of human-readable text back to a user. The biggest issue of word tokenization is OOV words, which is a major problem in the testing phase of NLP modelling.

3.6.2 Character Tokenization

When the number of unique words being used within a corpus starts to become too large, character tokenization becomes a better choice. This variation of tokenization splits the text into a set of characters, which allows for the ability to overcome some of the issues with word tokenization. Although this technique can improve upon some of the issues with word-based tokenization, it isn't without its own flaws too, as we are splitting many words into character representation, which can create a very large vocabulary to work with.

3.6.3 Subword Tokenization

This technique splits words on the subword level e.g., *smarter* would become *smart-er*, as the subword would split the word on the n-gram character level, instead of on a character or word level. One of the most popular subword based tokenisation processes is known as Byte Pair Encoding (BPE) (Gage, 1994), which is a word segmentation algorithm that merges frequently occurring character or character sequences iteratively.

3.7 Model Evaluation

Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) is the most popular and most commonly used metric for the evaluation of NMT systems. The concept of evaluation for machine translation is a difficult task, due to the ambiguity of the sentence pairs within the dataset being used. It is true that human evaluation is the best way to evaluate model performance, which is the reason it is the preferred methodology for evaluation, but a computer-based evaluation approach should also be considered.

Alongside BLEU, some investigation into METEOR (Banerjee & Lavie, 2005) is also explored. METEOR is another approach to NMT evaluation which takes precision and recall into account. Alongside this, we have the more common metrics of evaluation that are seen in many machine learning and deep learning tasks such as Sentiment Analysis and Prediction Modelling. Human evaluation will be conducted after this, with the findings being discussed further. Other metrics such as NIST, GLEU and Word Error Rate will also be discussed and explored.

4.0 Model Development

This section discusses the process of the model development phase of the project. The main focuses refer to data loading and pre-processing, the different options explored for tokenisation, the different hyper-parameters that were experimented with and finally, discussion regarding the overall model architecture and how it all comes together.

4.1 Data Loading and Pre-Processing Steps

The data used for the research comes in the form of the tmx format and the txt format. Tmx (Translation Memory eXchange) is an XML data structure typically used for translation memory data. It allows for the ability to retain the source and target texts in a parallel nature. To use this tmx data in the research, it must be extracted, converted, and then loaded into a new data structure, done by looping through the iteration property of the node. After this is done, the data for each of the source and target languages are added into two separate lists, and then added into a two column pandas data frame to maintain the parallel nature of the texts. The ***translate.storage.tmx*** python library is used for the extraction. The data in the txt format is much simpler, with the properties declaring the separator, header as None, and making use of only columns one and two of the file, as the third column housed information on the person who had translated the text from source to target and isn't needed.

For pre-processing, there were a few aspects that were considered, but the most important aspect concerned the removal of punctuation. In some ways, punctuation such as ! and ? are necessary to understand if the sentence is a question, an expression of anger, or just a general sentence. This means that consideration for the removal of punctuation must be considered an important part of the pre-processing stage, especially for translation texts. Some initial training is done on text without punctuations, to see if the removal of punctuation on the source and target text has any real noticeable difference in the final translation accuracy and overall quality of predictions. It is decided later that punctuation will remain.

Chinese text, such as **你用跑的**, is difficult to tokenise in its raw form. To separate the Chinese characters into their word meaning, a library called ***Jieba*** is used to create a segmented representation of all text within the target side of the data frame. After this is implemented and added back into a list, this sequence becomes **你 用 跑 的**, where the space between the characters is noticeable, allowing for tokenisation to take place on the target texts. The final step for pre-processing at this stage involved lowercasing all

words in the data. The data is split into a training set and testing set, utilising a random state to maintain the same data-split each run, using a 75/25 ratio.

4.2 Start and End Tokens

In some models that use a decoder, the process of helping the model understand when to start and stop predicting the input within the decoder is important. Choosing to have these special **<s>** and **</s>** tokens in your data is a personal preference when it comes to the source, but for the target, there is need for these tokens for the process of prediction.

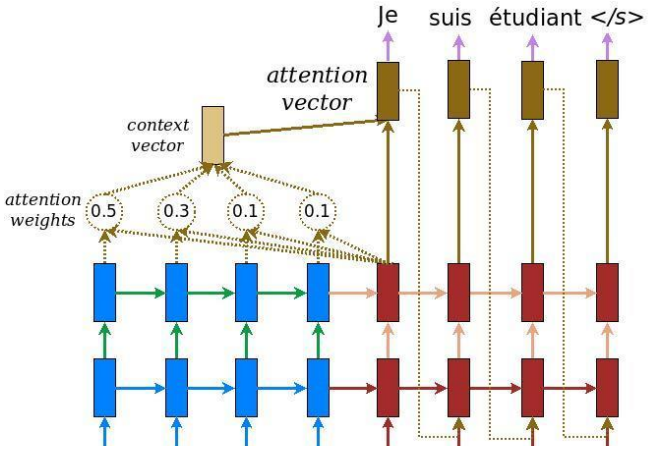


Figure 14 – Encoder / Decoder showing the use of the start and end tokens for the decoder (red)

Figure 14 shows the Encoder / Decoder, with visual representation of start and end tokens in action for the decoder input and output. The start tag **<s>** helps the decoder understand the input, and helps in the process of prediction, it helps the decoder to predict what will be the first token in the sentence, with the accompanying end **</s>** token used to mark the ending within the decoder. The choice of implementation of these tokens within the source target is purely based on the researcher, and for this research the use of the input start token is fully utilised and implemented.

4.3 Options for Tokenization

The process of tokenization is a crucial step in processing the data for use in machine translation, as without this step, the model will be unable to read or understand any of the raw text that is fed into it. Section 3.6 outlines the three options available for this process, being word, subword, and character based. For this research the focus of tokenization makes use of the subword tokenization strategy to split up words and form numeric equivalents for them. This form of tokenization was decided upon due to the idea of stemming, which is where the form of a word can be based on its stem. For example, let's consider the words "**love**" and "**loved**". If we perform a word-based

tokenization strategy on these words, the context could be lost, as they would be given different numbers, let's say [2] for "**love**", and [17] for "**loved**".

If we want to maintain the context of these words and allow the model to learn that the word "**loved**" stems from the word "**love**", we can perform subword tokenization, which will mean the word "**love**" would still have a token form of [2], but the word "**loved**" would now be split up into two words, "**love**" and "**d**" with token form of [2] and [91] as an example. By performing subword tokenisation, the model can capture the context better, and will learn that "**loved**" stems from the word "**love**", rather than considering it to be a new word entirely.

4.4 Hyper-Parameters

When it comes to the hyper-parameters for the Transformer model, there are only so many options that can be adjusted to increase, or decrease the performance and time needed to train the model. The first set of hyper-parameters that can be adjusted refer to the data used for input, which are maximum vocabulary length, maximum length of sequences, or batch size.

A batch, not to be confused with an epoch, is a hyper-parameter of gradient descent, that allows for control over the number of training samples to work through before the internal parameters of the model are updated. This is not the same as the epoch, which is the hyper-parameter of gradient descent that allows control over the number of complete passes the model takes through the training dataset before it is marked as complete (Brownlee, 2018).

The other set of hyper-parameters available for the model are the number of samples, d_model, number of layers (encoder/decoder), feed-forward neuron units (ffn units), number of heads for attention and the dropout rate. By default, the d_model is the dimensionality of the representations used as the input and output for the multi-head attention and is set with the value of 512. For this research, there are no real adjustments made to this value during implementation for all, except for one model tested on a smaller amount of data. The main hyper-parameters that see changes here, are number of layers, ffn units and the dropout rate.

In the standard Transformer, there are 6 layers for each of the encoder and decoder. The focus of model implementation for this research took place via Google Collaboratory, which is an online platform where you can utilise GPUs on virtual instances. Due to the

RAM limitations of the GPU memory available (16GB via Colab), these layers remained at 6 (total of 12), due to the size of the dataset being used for training, as setting hyper-parameters too high either resulted in memory allocation resource errors, or led to memory leaks, causing eventual model crashes between epochs.

4.5 Model Architecture Implementation

For the model implementation, the architecture follows the representation of the Vaswani et al. (2017) Transformer model. The Transformer implemented contains the Scaled Dot-Product Attention, Multi-Head Attention, Encoder and Decoder layers, and finally the Positional Encoding layer. More information on the structure of the Transformer can be found in the paper by Vaswani et al. (2017).

The creation of the model follows several stages, which involves the implementation of the different layers for the model in incremental code blocks, to distinguish the different required sections of the model. The first stage involves the creation of the Attention layers, Scaled Dot-Product and Multi-Head. These are the necessary attention layers needed to calculate dot-product, get the scale factor, and apply masking.

After the creation of the attention layers, next was the creation of the Positional Encoding. This is a vector that is added to the embedding vector to help represent a token in n-dimensional space. Tokens with similar meaning will be close together, whilst those with vastly different meanings will be further away. The addition of the Positional Encoding vector helps to keep tokens close to each other, based on their similarities. The next phase refers to the Encoder / Decoder layers. The key attributes for these are the number of ffn units and the necessary dropout rate. The encoder and decoder require a layer to be created first, which is then applied to the final model n number of times, with n being a hyper-parameter decided beforehand as a global variable.

With all these created, the final implementation of the Transformer architecture can be constructed. The number of layers, ffn units, dropout rate, vocab size and d_model to the encoder and decoder are decided here, as it is important that these layers share these attributes. After this, the creation of a padding mask and look ahead mask for the padding and causal attention requirements of the model is required. Creation and application of a loss function is done after, passing values for the mask and loss for this, and finally building the training function to allow the ability to run the model with the hyper-parameters chosen. To make sure long epoch runs aren’t lost in the case of power

loss or other unforeseen errors, a checkpointing system is applied after every epoch, meaning that it is possible to return to the model in the case of a fault, or reload the model for further use at its last final state, without needing to return to the very start of the model training.

5.0 Results and Analysis

The initial results and analysis of the completed NMT models are collated and put through automated evaluative methods to evaluate their formation at a quick pace. A suitable number of automated methods must be utilised, as to make sure the model/s developed has been tested and evaluated thoroughly before being used within a survey for human evaluation purposes. The development code for this project can be found at:

https://github.com/BrianALDavis/MSc-Machine_translation_en-zh

5.1 Methods for Analysis

The analysis of a translation model includes the use of automated methods, as they are faster to implement, and require no true human interaction from anyone but the researcher. There are many options that can be utilised for this, such as BLEU (Papineni et al., 2002), Word Error Rate (WER)(Morris et al., 2004) and METEOR (Banerjee & Lavie, 2005). It is important to evaluate using at least three automated methods, to thoroughly test the implementation of the model through a variety of methodology, as these automated methods each target different aspects of the models output.

5.1.1 Model Training Performance Analysis

The performance of the model in training is highly dependent on the hyper-parameters. During the training phase, the training accuracy and loss are returned at each epoch. Utilisation of early hyper-parameters showed results that only managed a high of 7.6% accuracy through a total of 100 epochs. Neural Machine Translation models require hundreds, and sometimes, thousands of epochs to become fluent toward the target language or reach a high accuracy, but due to the computation available and the time taken to train these models on the local implementation, the number of epochs is reduced substantially.

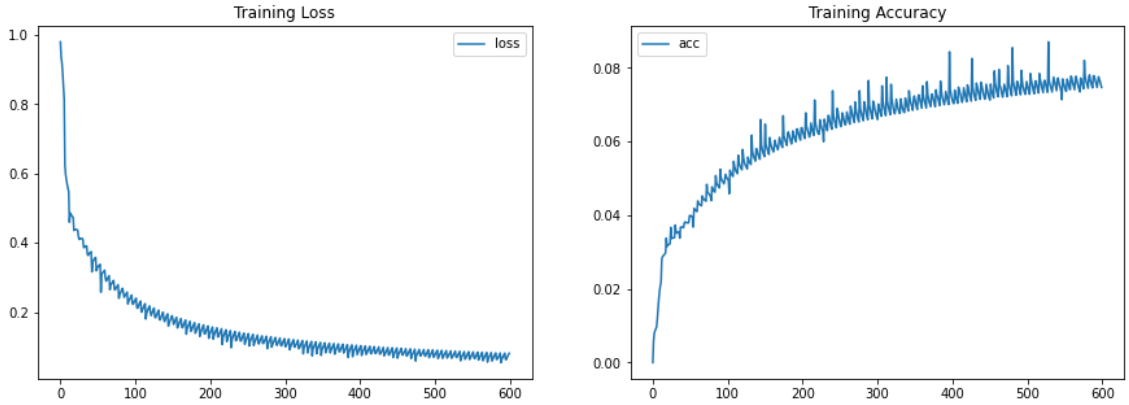


Figure 15 – Training and Loss of Model CKPT-3 over 100 epochs

Figure 15 illustrates this point. Utilising a parallel dataset of approximately 78,000 sentence pairs, batch size of 128 and 40000 samples, this model was able to reach a max accuracy against the training set of 7.61%. Whilst the results against the training data were respectable, results on unseen training data were prone to grammatical errors, and thus changes had to be made to improve training speed and increase accuracy gain over time. Although training accuracy numbers are nowhere near what is wanted, the loss of 0.0639 is very promising on a naturally difficult target language.

A smaller batch size is tested (64), but the findings suggest the reduction in batch size doesn't have a positive effect, as results appear to be the same in terms of accuracy, however, training time almost doubles in terms of minutes per epoch, leading to a slower training time for a similar outcome. A test is conducted next on a smaller dataset, totalling 26,388 sentence pairs, with a much smaller d_model and reduced number of feed-forward units, to see how adjusting the size of the dataset along with changes to the dropout rate, slightly increasing this to 0.2, instead of the default 0.1, and other parameters influence the model performance. Results can be seen in figure 16.

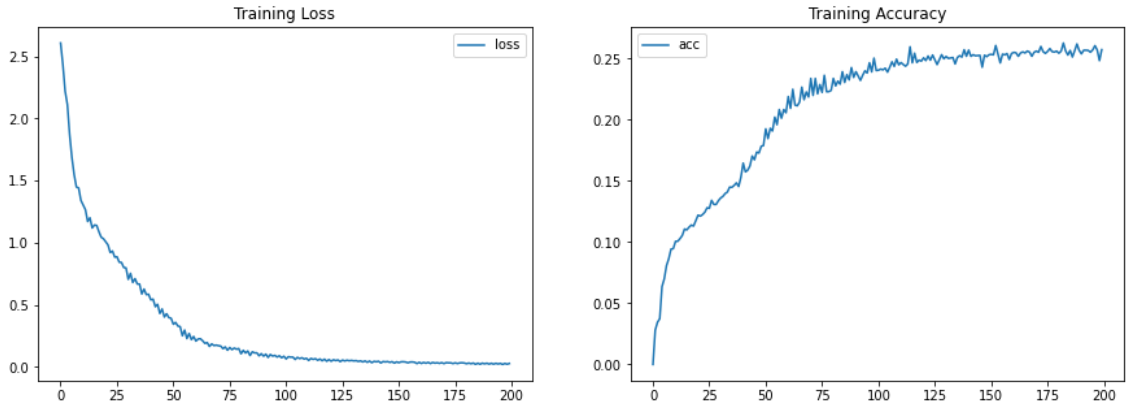


Figure 16 – Training and Loss of Model CKPT-2 over 100 epochs – smaller dataset

It is possible to observe an accuracy of 24.84% for epoch 100, with a loss of 0.0222 obtained for this model. Increasing the samples to 50000 and reducing the training data led to drastically faster epoch training cycles of 1.50 mins per epoch, but also the added ability to gain accuracy at a much faster pace than that seen in figure 15. With changes made to parameters in the form of increasing the d_model back to 512, and with feed-forward units set to 2048, a new model, titled model 6, showed vast gains over its training cycle, reaching 14.88% accuracy and 0.2715 loss.

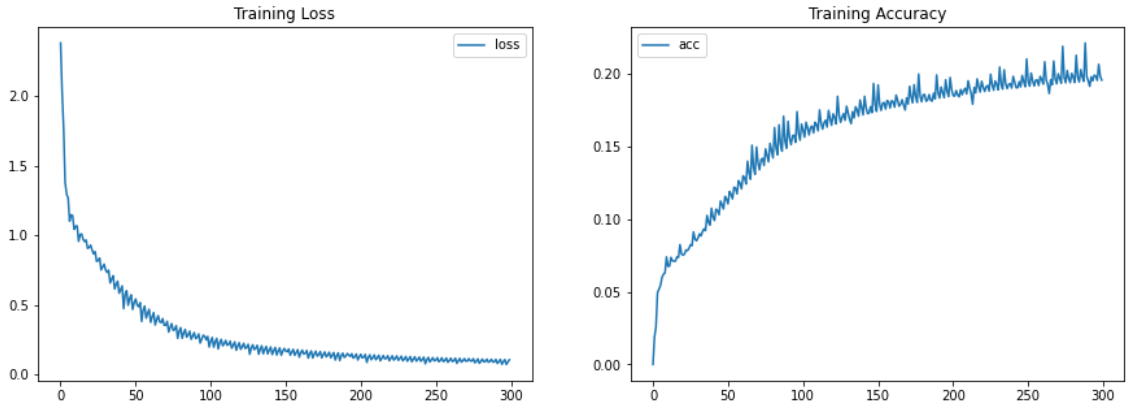


Figure 17 – Training and Loss of Model 7 over 100 epochs

With a slightly higher sentence pair count of 79,818 sentence pairs in the training set, and a reduction in feed-forward units by 50% to 1024, model-7 is trained and managed to vastly improve over model-6. Seen in figure 17, model-7 has a similar accuracy to that of model CKPT-2, topping out at 20.65% accuracy and 0.0725 loss, however with a bigger model and a larger training dataset. Totalling almost 13 hours to train, Model-7 was one of the better models to be trained based on accuracy and loss pairing, only rivalled by CKPT-2 and Model-5, but whilst accuracy is a great metric, it is not the be all end all metric, with tests being undertaken now to see how well the model performs on real world unseen translation tasks.

5.1.2 Hyper-parameters adjustments findings

It is difficult to make too many adjustments in terms of the hyper-parameters for this research that have a large influence over the training of the model. The parameters that led to the biggest changes, however, were the d_model, feed-forward units, dropout size and layers. Initially tests focused on how increasing the number of layers impacted training but found that increasing the number of layers didn't have noticeable impact. However, changes to d_model, feed-forward units and dropout rate had more substantial effects when it came to model training and accuracy gains / losses.

When comparing model-6 and model-7, where the only difference between these models is the number of feed-forward units, shows a much faster accuracy gain in training when ffn units were halved. Model-6 manages 14.88% when reaching 50 epochs, whilst model-7 reaches 19.32%, which is a gain of 4.44% over the same number of epochs, with an equal dataset size. When considering loss, model-6 has a loss of 0.2715 whilst model-7 has a loss of 0.1387, which is almost 50% lower.

Most models reach a high of 7% approx. accuracy over 50 to 100 epochs, with a change in number of layers, d_model, dropout rate and batch size having only incredibly marginal effects. The biggest impact came from changing feed-forward units, which were better at a lower number for a small dataset. It would be interesting to see how a larger dataset has an impact on these hyper-parameters, as a smaller dataset leads to more difficulty for NMT models to learn the language.

5.2 Model Testing

The generation of predictions is required to understand how well each model has learnt the target language. This uses a training and test dataset for predictions across models. Other testing means such as by-sentence and automated methods are discussed.

5.2.1 Against the Dataset

The generation of predictions comes from the creation of the translate and predict functions, to allow the feeding of individual sentences into the model to generate a by-sentence prediction. Once this is created, it must be adapted to be able to translate the first 15000 sentence pairs of the training set. Taking the Reference and Source values from the training data array, this is fed into an if statement, which loops through all the values and generates a prediction for each. At the end, this is saved to a CSV to capture the predictions for later use. The code to generate the predictions is shown in figure 18.

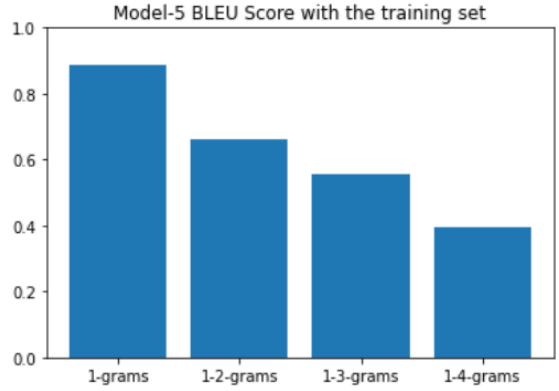
Initial predictions are analysed using n-gram weighting of BLEU and graphed to show performance of the model when n-gram weighting is adjusted. Figure 19 shows the graph of Model-5 and how n-gram weighting adjustments have an impact on the overall returned score of the model.

```
train_list = []
sentences = en_array[:15000]

counter = 0
n = 500
for i in sentences:
    counter += 1
    if counter % n == 0:
        print("counter has reached:", counter)
    translation = translate(i)
    train_list.append(translation)
```

Figure 18 – Corpus level Predictions on Training Data

Alongside the visualisation, the scores for BLEU are also shown below the chart. This shows that scores decrease the more n-gram weighting is factored in, with the 1-4 n-gram score reaching 39.56 for Model-5, putting it very close to 40, which is deemed to be a good BLEU score, to be expected against the training data that the model has gotten familiar with.




```
Corpus BLEU Score Train 1-1 N-Gram Weighting: 88.64
Corpus BLEU Score Train 1-2 N-Gram Weighting: 66.32
Corpus BLEU Score Train 1-3 N-Gram Weighting: 55.48
Corpus BLEU Score Train 1-4 N-Gram Weighting: 39.56
```

Figure 19 – Model-5 N-Gram Weighting Chart

As scores can be undesirable for short length sentences, a smoothing function, first discussed by B. Chen & Cherry (2014), is used to alleviate biases in scores, and also to avoid unwanted and possibly incorrect scores for BLEU here. Method 7 is chosen as the smoothing function of choice, with more information on these functions found at (NLTK, 2009/2021). Other tests conducted include by-sentence prediction, which involves taking a random value from the data, and asking the model in question to translate this, to see how close this is to the truth. This is commonly done against the testing data, rather than against the training dataset.

5.2.2 Against Unseen Input

Most of the focus when it comes to model testing focuses on the strength of each model on unseen data from the test dataset. One of the first methods utilised to test unseen input is through by-sentence translations. The output of a sentence translation can be seen in figure 20. A randomiser is created, to take a random index and pull a sentence from the test dataset, which is then fed to the translate method, which tokenises, feeds the tokenised input to the model, and generates an output in readable text. It is possible to observe a low score for this example sentence prediction from GLEU of 0.18, however, it is also possible to see that the output sentence has captured almost all aspects of the expected sentence here, with some additional words added in, with others missed out. It appears harsh that this prediction receives such a low score.

```
Input sentence: please help yourself to the cake.
Expected sentence: 你们自己吃蛋糕。
Output sentence:   你自己拿蛋糕吃吧。

GLEU Sentence Score: 0.18
Reference: ['你们', '自己', '吃', '蛋糕', '。']
Candidate: ['你', '自己', '拿', '蛋糕', '吃', '吧', '。']

Sentence Index: 101
```

Figure 20 – A By-Sentence Example, utilising a random sample

Figure 21 shows the code used to create by-sentence predictions to test models on a sentence-by-sentence basis. Alongside utilisation of by-sentence predictions, the automated metrics of BLEU, GLEU, WER, NIST and METEOR are utilised on a corpus level, adapted where this is not possible through use of the translate library via NLTK.

```
actual  = prediction_test_df_model7['Reference']
predict = prediction_test_df_model7['Candidate']
source  = prediction_test_df_model7['Source']

# lets find a random example from the test dataset using the
# random module to test the accuracy of the sentence
index = random.randint(0, len(actual))

# show a translation from the test dataset
sentence = source[index]
print("Input sentence: {}".format(sentence))
predicted_sentence = predict[index]
#predicted_sentence = translate(sentence)
print("Expected sentence: {}".format((actual[index].replace(" ", ""))))
print("Output sentence:   {} \n".format(predicted_sentence.replace(" ","")))

# set-up for BLEU sentence level eval
reference = actual[index].split()
candidate = predicted_sentence.split()

score = sentence_gleu([reference], candidate)
print("GLEU Sentence Score: {:.2f}".format(score))
print("Reference:", reference)
print("Candidate:", candidate)
print("\nSentence Index:", index)
```

Figure 21 – By-Sentence predictions for Model-7 using GLEU

The Word Error Rate (WER) is a useful metric to gain a reasonable representation of the accuracy of the model against the corpus. Whilst accuracy is seen in section 5.1, the word error rate provides a better representation of accuracy, by checking the reference and candidate sentences, and finding the number of sentences that have zero-word errors after predictions.

```
# count how many sentences in the corpus have at least 1 incorrect word
count = 0
for i, j in zip(actual, predict):
    if wer_score(i, j, print_matrix=False) == 0:
        pass
    else:
        count += 1
print("Number of sentences with errors:", count)
print("Percentage of correct sentences: {:.2f}%".format(100-(count / len(actual) * 100)))
```

```
Number of sentences with errors: 7745
Percentage of correct sentences: 48.37%
```

Figure 22 – Word Error Rate of Model-5 predictions against the test corpus

Figure 22 shows the code that creates and runs the WER for the Model-5 Test corpus. As WER is not typically used on a corpus level, and rather more on a word level, this must be adapted, so that it is possible to get the accuracy of the corpus translation through a loop and counter method. Each time the WER finds a candidate sentence that differs from its reference sentence, the score would be greater than 0, which means the

counter would be incremented on the pass of this sentence. The final tally is that of all the sentences that have at least one error, with the percentage total being the percentage of correct sentences translated by the model, deemed to be the model accuracy against the corpus.

5.3 Results

Once the models have been tested and predictions have been obtained, they are put against the automated metrics to evaluate and obtain metric score, utilising automated methods to understand the overall accuracy and legibility of each model. In the end, only three models are used for the automated evaluation, with their results presented. The models chosen are used based on their accuracy obtained through training, with a minimum of 100 epochs, which excludes model-6.

5.3.1 BLEU

As one of the most popular and well-known automated metrics for NMT, it is impossible to ignore BLEU for first evaluation of each model. However, weighting is a big part of BLEU, and its performance can be measured based on the n-gram weighting that is applied during evaluation. Due to this, weighting of [1.0,0,0,0], [0.5,0.5,0,0], [0.3,0.3,0.3,0] and [0.25,0.25,0.25,0.25] are applied and output into bar charts for visual comparison.

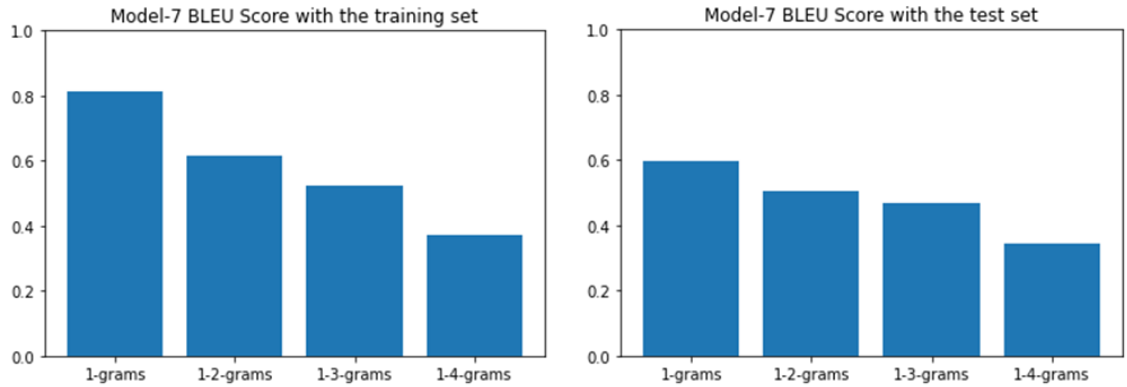


Figure 23 – BLEU Weighting Comparison – Train (left) and Test (right)

One of the biggest limitations with BLEU is that shorter sentences require a smoothing function, as BLEU as a metric is not well designed to handle short sentences and is better utilised on a corpus of larger sentences. Table 5.1 shows the BLEU scores across the three tested models. This shows that, as expected, the smaller model with smaller parameters achieved a slightly higher score than the other models and is the best performer against the Training data. However, when it comes to unseen input, Model-5

performs better, with a BLEU score of 35.40. All scores shown in table 5.1 utilise the [0.25,0.25,0.25,0.25] weighing, which is the default weighting.

Table 5.1 – BLEU Score Comparison Across Models

	CKPT-2	MODEL-5	MODEL-7
BLEU Train	40.02	39.56	37.31
BLEU Test	33.17	35.40	34.38

5.3.2 GLEU

Generalised BLEU (Napoles et al., 2015) is a variant of BLEU that uses n-gram overlap instead of precision/recall of specific annotated errors. GLEU is a harsher scoring system compared to BLEU, which is evidenced in the scores given to each model for their predictions on the test data. The model CKPT-2 achieves a score of 83.45 on the training data but achieves a lowly score of 7.73 on the test data. Model-5 is the best performing model on the test data according to GLEU, achieving 22.04. This is in stark comparison to BLEU, which rates the former models with a lot more praise than GLEU does, which is why numerous methods are undertaken here due to some metrics giving higher preference to certain errors than others.

Table 5.2 – GLEU Score Comparison Across Models

	CKPT-2	MODEL-5	MODEL-7
GLEU Train	83.45	60.27	54.89
GLEU Test	7.73	22.04	0.46

5.3.3 Word Error Rate

Word Error Rate is useful to obtain the accuracy of the model, but it also has the added benefit of being able to obtain the number of sentences that are predicted with no errors, or with a flexible error count. This is much faster than manually checking each individual sentence prediction against the reference sentence in the target language. Whilst BLEU and GLEU score the model based on how many n-grams are precise, the WER helps to understand only where words were either right or wrongly placed in the prediction. The reported scores found in Table 5.3 allow some leeway for each model, where the percentage for WER Test is gathered of how many predictions have less than or equal to 3 errors in the candidate sentence compared to the reference.

Table 5.3 – WER Score Comparison Across Models

	CKPT-2	MODEL-5	MODEL-7
WER Train	92.97%	81.13%	77.34%
WER Test (<= 3)	25.07%	53.79%	4.78%

5.3.4 NIST

NIST scores the predictions with a different brevity penalty than that of BLEU and GLEU. NIST takes the list of references and compares the candidate against them, presenting the user with a NIST score at the end for the corpus, based on the presented candidates. If NIST is unable to find a good match, then the score will be decreased, where a higher score means a better matching rate in terms of NIST. Table 5.4 shows the results of NIST scoring on the predictions of the three models on train and test data. Model CKPT-2 performs the best on the train data, but all models perform poorly when it comes to the test dataset. For these models, it seems that the use of NIST is not a good choice, as the scores seem to be a lot worse than is to be expected, and the perception of NIST scores is not as easy to understand, but the results are still presented here for comparison.

Table 5.4 – NIST Score Comparison Across Models

	CKPT-2	**MODEL-5**	**MODEL-7**
NIST Train	1.46	4.27	5.10
NIST Test	0.0003	0.0001	0.0003

5.3.5 METEOR

The METEOR score is produced by finding the closest possible candidate sentence within the corpus, and then scoring this result between 0 and 1 based on how close to the reference it matches the candidate. METEOR is very similar to WER, except METEOR makes use of precision and recall for scoring the resulting matched sentence pair. Table 5.5 shows the results of the corpus-based METEOR adaptation undertaken, due to METEOR being primarily focused on a sentence level. When limiting accepted results to above 0.5, after rounding, it is possible to see that Model-5 has the best overall accuracy for METEOR scores found within prediction pairs at 8773, which is over half of the test dataset, with Model-7 performing the worst for METEOR on the test data.

Table 5.5 – METEOR Score Comparison Across Models

	CKPT-2	**MODEL-5**	**MODEL-7**
METEOR Train	14517	13704	13598
METEOR Test	3774	8773	1085

5.3.6 Final Comparison

For the comparison, utilising the scores obtained from all the automated metrics, table 5.6 shows that considering the training data, all the models perform as well as each other, but table 5.7 shows that when considering the unseen input of the test data,

Model-5 shows the best performance, with a WER score almost double that of its closest competitor model CKPT-2. Across the board Model-5 performs better and scores higher in all categories except NIST, where all models achieve a low score from this metric.

Table 5.6 – Overall Train Score Comparison Across Models

Train	CKPT-2	MODEL-5	MODEL-7
BLEU	40.02	39.56	37.31
GLEU	83.45	60.27	54.89
WER (correct)	92.97%	81.13%	77.34%
NIST	1.46	4.27	5.10
METEOR	14517	13704	13598

Table 5.7 – Overall Test Score Comparison Across Models

Test	CKPT-2	MODEL-5	MODEL-7
BLEU	33.17	35.40	34.38
GLEU	7.73	22.04	0.46
WER (<= 3)	25.07%	53.79%	4.78%
NIST	0.0003	0.0001	0.0003
METEOR	3774	8773	1085

6.0 Survey Evaluation and Feedback

To evaluate the model found to be the best from automated evaluations, surveys were distributed to native speakers of the target language to measure how well they thought the model did over three different aspects. The main tests from the surveying are based on Accuracy of translation, Fluency of translation, and finally, Context maintenance after translation.

6.1 Primary Research

When conducting primary research, it is of key importance that the target audience is correctly decided upon and targeted for the needs of the research toward successful data collection. Purposive Sampling (Lavrakas, 2008) is a type of non-probability sampling, where the main objective is to produce a sample that can be logically assumed to be representative of the population. The selection of a non-random sample of participants from the population is important to make sure the collection of feedback from those that are deemed to be the target audience for our sampling needs, is correct and suits the research.

The survey that is created for this research focuses on native Chinese speakers who have at least a partial understanding of the English language, to generate feedback on the quality of predictions taken from the NMT model. Convenience sampling is also considered when deciding on the best type of non-probability sampling to apply, but it is found to be unnecessary for this research, as Convenience sampling leverages individuals for approach, but the concept of Purposive Sampling fits the needs of the research here more closely and allows for faster collection of feedback in the short period of time allotted for surveying.

6.2 Survey Design

To design the survey, it’s necessary to understand exactly what elements of the predictions will be the most adequate to evaluate via the crowd. The main aspects of translation output for evaluation are Fluency, Accuracy of Translation and Context Maintenance, which means how well has the translation maintained the original context of the sentence in its translation output. It’s important not to make the survey too long, as a survey that takes too much time deters respondents from reaching the end. The chosen length is a maximum of 15 questions, taking around 10 minutes to complete, with questions as evenly spread as possible over the three aspects. The questions are chosen through a random selection of predictions, focusing on the aspects to be covered.

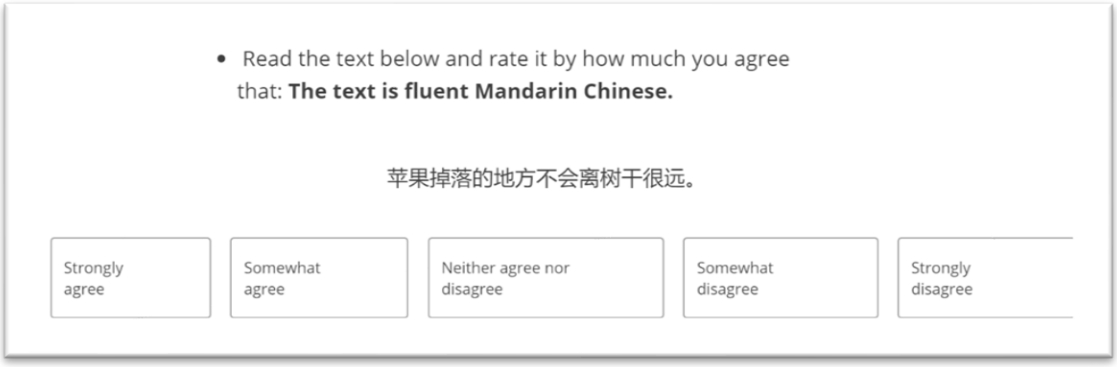


Figure 24 – Survey Question regarding Fluency

Figure 24 shows a Fluency based question from the survey. It's important to make sure that the content shown makes sense for the question, and regarding Fluency, showing the respondent the English text that has been translated isn't important. The questions make use of a Likert scale, which is a similar style to that of Graham et al. (2017), who also made use of a Likert scale and bold text to emphasise what the respondent should be agreeing, or disagreeing with.

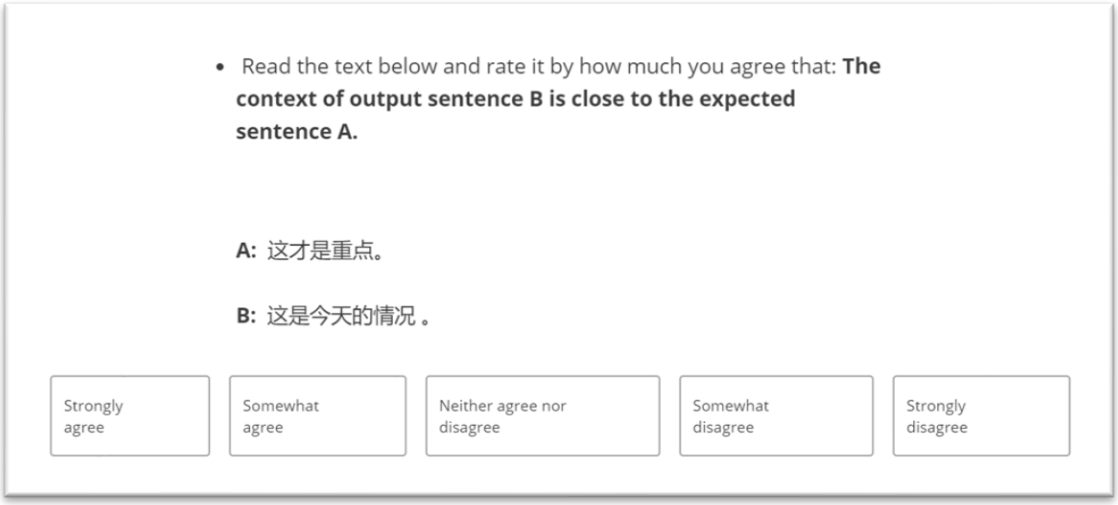


Figure 25 – Survey Question regarding Context Maintenance

Figures 25 and 26 each show an example of the other aspects being presented for response. Figure 25 is a context maintenance question, which shows the respondent two sentences, labelled A and B. The idea here is to have the reference and candidate sentence and ask the respondent how well the context of sentence A (reference) is being represented by sentence B (candidate). For figure 26, as accuracy is the key focal point of the question, it's important to use the input and output formatting for the question, asking the user how well the output matches the input in terms of its translation accuracy.

Figure 26 – Survey Question regarding Accuracy

6.3 Survey Ethics

Following SHU Guidelines (Sheffield Hallam University, n.d.), ethical practice was conducted and enforced for data collection, analysis, and presentation of this data. It was important to make sure that the process of data collection was thorough, and the chosen datasets collected from external means was checked by a third party who was familiar with the target language for this research, to clarify accuracy was to at least a 95% level. Official permission was obtained before any search and collection of primary data was conducted, with a consistent record of all data being held on cloud-based storage facilities with encryption present, although any data collected did not in any way identify a user, as this research doesn't focus on individuals, and thus all collected data had no issue in reference to the Data Protection Act 2018 (*Data Protection Act 2018*, n.d.). The completed ethics checklist can be found within Appendix B.

6.4 Survey Findings

The survey resulted in 50 unique respondents. The findings of this can be found in the following tables. Table 6.1 shows the results from the fluency-based questions (4). The idea of fluency for Chinese speakers, are sentences that not only sound like something a Chinese speaker would say, but also follow the same punctuation that the speaker would use themselves. For the first three questions of the survey, agreement percentiles were 78% for question 1, 69.5% for question 2 and 67.4% for question 3, showing that the

model had predicted in a way that sounded and looked fluent, with most respondents agreeing with this outcome. For question 15, just over 50% of respondents agreed with this statement, however a much higher number chose neither agree nor disagree at 29.3%, which could be due to this question being the last one on the survey.

Table 6.1 – Fluency Question Responses

User Response	Fluency Questions			
	Q1	Q2	Q3	Q15
Strongly Agree	32%	30.4%	41.3%	24.4%
Somewhat Agree	46%	39.1%	26.1%	26.8%
Neither	6%	17.4%	15.2%	29.3%
Somewhat Disagree	10%	8.7%	4.3%	14.6%
Strongly Disagree	6%	4.3%	13%	4.9%

After testing fluency, the next aspect for evaluation is accuracy. Accuracy is a metric that is quite subjective, so it's important to measure accuracy from the crowd, rather than just relying on the Word Error Rate. This is because sometimes a translation is not the same, but the meaning is maintained with slightly different wording. Table 6.2 shows the responses regarding the accuracy of the translations presented. For questions 6 (72.8%) and 11 (73.8%), most respondents agreed that the output, although not the same, still conveyed the same meaning. Questions 5 (54.5%) and 14 (51.2%) were closer to 50% in terms of respondents who agreed that accuracy was maintained. Question 4 contained a word that was not present in the input, and this inclusion is evident and misplaced, as only 34.1% selected a positive response to this question.

Table 6.2 – Accuracy Question Responses

User Response	Accuracy Questions				
	Q4	Q5	Q6	Q11	Q14
Strongly Agree	13.6%	38.6%	61.4%	50%	27.9%
Somewhat Agree	20.5%	15.9%	11.4%	23.8%	23.3%
Neither	13.6%	15.9%	6.8%	7.1%	18.6%
Somewhat Disagree	22.7%	13.6%	11.4%	7.1%	16.3%
Strongly Disagree	29.5%	15.9%	9.1%	11.9%	14%

The biggest area of contention for the survey was context maintenance. Six questions were presented to the respondents that concerned context, with the results presented in table 6.3. Question 12 had the highest positive agreement response at 72.1%, where the sentences A and B were a complete match. All other questions received a higher number of negative response than positive when it came to context, with only question 9 being close to 50% at 43.3%. This tells us that the model is fluent and accurate, but

when it comes to maintaining context, it is still not doing a good job, and further improvements are still necessary to improve context maintenance in predictions.

Table 6.3 – Context Maintenance Question Responses

User Response	**Context Maintenance Questions**					
	Q7	**Q8**	**Q9**	**Q10**	**Q12**	**Q13**
Strongly Agree	16.3%	16.3%	23.3%	16.3%	62.8%	18.6%
Somewhat Agree	25.6%	9.3%	20.9%	23.3%	9.3%	23.3%
Neither	18.6%	18.6%	7%	20.9%	11.6%	20.9%
Somewhat Disagree	20.9%	32.6%	16.3%	20.9%	9.3%	16.3%
Strongly Disagree	18.6%	23.3%	32.6%	18.6%	7%	20.9%

7.0 Findings and Conclusions

Upon completion of this research project, it can be said that the initial aims have been achieved:

- Implement an existing Neural Machine Translation model architecture
- Enhance performance through utilisation of Deep Learning technologies
- Evaluate the model to discover strengths and weaknesses
- Consider adjustments required to improve the model
- Obtain Qualitative feedback to evaluate performance

Conducting primary research with native speakers through Qualitative feedback enhanced the evaluation of the developed model and has helped to understand the weakness of the model, to a much stronger degree than through the automated methods. Although they presented scores that could be understood, they didn't do enough to outline the improvements that would be needed to continue to further develop the model in the future. Furthermore, the key hyper-parameters were understood and tweaked where necessary, with different training performance being apparent in the developed models.

Despite performance of the model/s not reaching the levels that would have been more beneficial due to lower epoch runs than was possible in the time available, and with the computation at hand, it was still possible to see how the adjustments impacted model training, and how the utilisation of a variety of automated metrics helped to enhance the overall quality of machine translation evaluation, utilising more than just BLEU.

7.1 Research Discussion

The primary deliverable for this research is a Neural Machine Translation model, able to translate the source language to the target language with a moderate level of precision. The project, although only achieving a 53.9% Word Error Rate accuracy in the strongest achieved model, can still be considered a success. The fulfilment of project aims has been attained through the initial project objectives:

- Identify current research / current challenges of Neural Machine Translation
- Identify valid sources of usable parallel bilingual data for model development
- Develop model utilising model properties to enhance outcome
- Evaluate performance of the model utilising deep learning techniques

- Obtain Qualitative feedback from participants to measure accuracy of developed model
- Perform final evaluation and discuss key findings

The literature review was a very important piece of the puzzle when it came to scoping out the best approach to follow, and to understand where the research should be focused. The research project would have been a lot more difficult without the understanding gained from the literature and helped to steer the research onto the right track in the initial stages of development. Evaluation via the crowd is a key focal point in the evaluation of the model, which would have been much more difficult to conduct without the primary feedback.

7.2 Recommendations from Research

Upon completion of the research, the findings suggest that:

- The systematic benefits of hyper-parameter adjustments require further testing on a much larger system to truly understand how changes can further improve translation quality
- NMT systems can translate fluently and accurately after a small amount of training with a rich dataset
- Dropout rate adjustment is a key metric that should be tested further
- Short sentence datasets work well for translation, but a mixture of short and long sentences should be avoided
- Context maintenance through translations still requires a lot of work to be recognised by native speakers
- Via the crowd evaluation is required, regardless of the time commitment necessary to implement it on larger scale research

The evaluation via primary research indicates that a translation model with a good BLEU score is still not really matching the quality that is required to reach a level deemed successful against human translators. Many translation models train over many days to reach a level of accuracy deemed to be successful enough for deployment, but this research has shown that with correct parameter adjustment, even a model with a small amount of training can still maintain close to 50% accuracy on unseen input. More research is required to understand the correlation between training time and parameter tuning, along with the richness of the training data being used.

7.3 Problems during Research

Throughout the research project, the following issues occurred:

- Local computation was lacking, which led to a smaller scale model being created, hampering the strength of results
- Primary research respondents required English to answer the survey, this could have been avoided with a better mix of questions, or a survey written entirely in the target language
- The dataset used contained mostly below 50 words per sentence, a bigger dataset for training would enhance findings
- More exploration is required to understand why the context wasn't maintained from the primary research, presentation of text boxes for comments would have enhanced primary data feedback, although this would have increased survey time required for completion

With the following considered, offering comments to users who disagreed with any of model outputs would have allowed for further understanding as to why these respondents felt that the context was not being maintained in the predictions presented by the model.

7.4 Conclusion

This project has demonstrated that the adjustment of hyper-parameters sometimes only shows minimal difference. At times, adjustments don't look to make a huge difference in the earliest stages of model development, but the longer the model trains, the more influence these hyper-parameters begin to have. The main angle that must be utilised for evaluation of model prowess is that of the crowd. It can be argued that the time commitments required to do human evaluation isn't worthwhile, but this research shows that the evaluation of the NMT is highly successful with the inclusion of by the crowd evaluation.

Considering scope for future work, the model developed has potential to overcome the issue of context maintenance through longer training times and larger training data, which requires further research and testing. This project has provided a good opportunity to further understand the need for human evaluation in the development of NMTs going forward, regardless of scale.

7.5 Limitations of Research

The following limitations were discovered throughout the research:

- The dataset was small, and a much larger scale of sentence pairs is needed to create a more comprehensive system
- The computation was limited, larger scale computation via the cloud or other local sources was not considered in time, thus limiting the scope of development
- Time restrictions meant model training had to be limited in epochs (100/200)
- Limited time to conduct primary research, more time would have allowed for a larger pool of respondents to have been contacted
- Limited scope of primary respondents due to some use of source language on survey

The findings are relevant for the research but having more time and stronger computation would have allowed the research to scale up massively, and stronger findings may have been possible. In addition to this, the use of the source language on the survey, with the limited time frame to have the collect responses, may have limited the ability to collect feedback. However, the feedback achieved reached a good number in the time the survey was active.

7.6 Scope for Further Research

When considering the limitations of the research project, the primary focus should be on expanding the research scope, with larger computation, larger training data and a longer period of time for model development and training. Future researchers should:

- Utilise larger computation for parameter adjustment
- Experiment with a wider model, rather than a deeper model
- Conduct a long-term study to establish long term influence of parameter adjustment
- Use a wider variety of data to train and evaluate models
- Utilise a larger array of automated methods
- Train a model for a much longer period of time to achieve better accuracy
- Conduct large scale human evaluation, over a wide geographic scale

If this project was taken further, with a large budget over a large period of time available, then more comprehensive evaluation and adjustments of models through development would be utilised. The strength of NMTs has always been their

development, with deeper models being created, but also the creation and development of the Attention mechanism really pushing forward the possibilities of machine translation. But something as small as a parameter adjustment, change of dataset, or utilisation of human evaluators can really strengthen the understanding towards finding the inherent weakness of the models in development before they are published and used in the real world. It is important we utilise all the enhancements available at all times, but never discount the need for humans to help translate language.

8.0 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *ArXiv:1603.04467 [Cs]*. http://arxiv.org/abs/1603.04467

Arthur, P., Neubig, G., & Nakamura, S. (2016). Incorporating Discrete Translation Lexicons into Neural Machine Translation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1557–1567. https://doi.org/10.18653/v1/D16-1162

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv:1409.0473 [Cs, Stat]*. http://arxiv.org/abs/1409.0473

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv:1409.0473 [Cs, Stat]*. http://arxiv.org/abs/1409.0473

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. https://aclanthology.org/W05-0909

Bao, G., Zhang, Y., Teng, Z., Chen, B., & Luo, W. (2021). G-Transformer for Document-level Machine Translation. *ArXiv:2105.14761 [Cs]*. http://arxiv.org/abs/2105.14761

Bapna, A., Chen, M., Firat, O., Cao, Y., & Wu, Y. (2018). Training Deeper Neural Machine Translation Models with Transparent Attention. *Proceedings of the*

2018 Conference on Empirical Methods in Natural Language Processing, 3028–3033. https://doi.org/10.18653/v1/D18-1338

Berrichi, S., & Mazroui, A. (2021). Addressing Limited Vocabulary and Long Sentences Constraints in English–Arabic Neural Machine Translation. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-020-05328-2

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, *16*(2), 79–85.

Brown, P. F., Lai, J. C., & Mercer, R. L. (1991). Aligning sentences in parallel corpora. *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics* -, 169–176. https://doi.org/10.3115/981344.981366

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, *19*(2), 50.

Brownlee, J. (2018, July 19). Difference Between a Batch and an Epoch in a Neural Network. *Machine Learning Mastery*. https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/

Chen, B., & Cherry, C. (2014). A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 362–367. https://doi.org/10.3115/v1/W14-3346

Chen, K., Wang, R., Utiyama, M., Sumita, E., & Zhao, T. (2019). Syntax-Directed Attention for Neural Machine Translation. *ArXiv:1711.04231 [Cs]*. http://arxiv.org/abs/1711.04231

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics* -, 9–16. https://doi.org/10.3115/981574.981576

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *ArXiv:1409.1259 [Cs, Stat]*. http://arxiv.org/abs/1409.1259

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *ArXiv:1406.1078 [Cs, Stat]*. http://arxiv.org/abs/1406.1078

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *ArXiv:1412.3555 [Cs]*. http://arxiv.org/abs/1412.3555

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed). Pearson.

Data Protection Act 2018. (n.d.). Queen's Printer of Acts of Parliament. Retrieved 4 September 2021, from https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the Second International Conference on Human Language Technology Research*, 138–145.

Domhan, T. (2018). How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1799–1808. https://doi.org/10.18653/v1/P18-1167

Fischler, A. S. (n.d.). *Mixed Methods*. Retrieved 20 July 2021, from https://education.nova.edu/Resources/uploads/app/35/files/arc_doc/mixed_methods.pdf

Gage, P. (1994). *A New Algorithm for Data Compression*. http://www.pennelynn.com/Documents/CUJ/HTML/94HTML/19940045.HTM

Gale, W. A., & Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, *19*(1), 75–102.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional Sequence to Sequence Learning. *ArXiv:1705.03122 [Cs]*. http://arxiv.org/abs/1705.03122

Ghader, H., & Monz, C. (2017). *What does Attention in Neural Machine Translation Pay Attention to?* 10.

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, *23*(1), 3–30. https://doi.org/10.1017/S1351324915000339

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2015). On Using Very Large Target Vocabulary for Neural Machine Translation. *ArXiv:1412.2007 [Cs]*. http://arxiv.org/abs/1412.2007

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*(4), 532–556. https://doi.org/10.1109/PROC.1976.10159

Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Vol. 2).

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent Continuous Translation Models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709. https://www.aclweb.org/anthology/D13-1176

Kalchbrenner, N., Espeholt, L., Simonyan, K., Oord, A. van den, Graves, A., & Kavukcuoglu, K. (2017). Neural Machine Translation in Linear Time. *ArXiv:1610.10099 [Cs]*. http://arxiv.org/abs/1610.10099

Kong, X., Renduchintala, A., Cross, J., Tang, Y., Gu, J., & Li, X. (2021). Multilingual Neural Machine Translation with Deep Encoder and Multiple Shallow Decoders. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1613–1624. https://www.aclweb.org/anthology/2021.eacl-main.138

Läubli, S., Sennrich, R., & Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4791–4796. https://doi.org/10.18653/v1/D18-1512

Lavrakas, P. (2008). *Encyclopedia of Survey Research Methods*. Sage Publications, Inc. https://doi.org/10.4135/9781412963947

Liu, B., & Huang, L. (2020). NEJM-enzh: A Parallel Corpus for English-Chinese Translation in the Biomedical Domain. *ArXiv:2005.09133 [Cs]*. http://arxiv.org/abs/2005.09133

Liu, L., Utiyama, M., Finch, A., & Sumita, E. (2016). Agreement on Target-bidirectional Neural Machine Translation. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 411–416. https://doi.org/10.18653/v1/N16-1046

Lowerre, B. T. (1976). The Harpy speech recognition system. In *Ph.D. Thesis*. https://ui.adsabs.harvard.edu/abs/1976PhDT.....81L

Lumen. (n.d.). *Attention | Boundless Psychology*. Retrieved 26 July 2021, from https://courses.lumenlearning.com/boundless-psychology/chapter/attention/

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. *ArXiv:1508.04025 [Cs]*. http://arxiv.org/abs/1508.04025

Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., & Zaremba, W. (2015). Addressing the Rare Word Problem in Neural Machine Translation. *ArXiv:1410.8206 [Cs]*. http://arxiv.org/abs/1410.8206

Mallick, R., Susan, S., Agrawal, V., Garg, R., & Rawal, P. (2021). Context- and Sequence-Aware Convolutional Recurrent Encoder for Neural Machine Translation. *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 853–856. https://doi.org/10.1145/3412841.3442099

manythings.org. (n.d.). *Tab-delimited Bilingual Sentence Pairs from the Tatoeba Project (Good for Anki and Similar Flashcard Applications)*. Retrieved 19 July 2021, from http://www.manythings.org/anki/

Maruf, S., Saleh, F., & Haffari, G. (2021). A Survey on Document-level Neural Machine Translation: Methods and Evaluation. *ACM Computing Surveys*, *54*(2), 1–36. https://doi.org/10.1145/3441691

MIAP, C. T. Bs. H. (2021, April 19). *Recurrent Neural Networks and Natural Language Processing.* Medium. https://towardsdatascience.com/recurrent-neural-networks-and-natural-language-processing-73af640c2aa1

Mnih, V., Heess, N., & Graves, A. (2014). *Recurrent Models of Visual Attention*. 9.

Moore, R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In S. D. Richardson (Ed.), *Machine Translation: From Research to Real Users* (Vol.

2499, pp. 135–144). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45820-4_14

Morris, A. C., Maier, V., & Green, P. (2004). *From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition*. 4.

Napoles, C., Sakaguchi, K., Post, M., & Tetreault, J. (2015). Ground Truth for Grammatical Error Correction Metrics. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 588–593. https://doi.org/10.3115/v1/P15-2097

Nießen, S., Vogel, S., Ney, H., & Tillmann, C. (1998). A DP based Search Algorithm for Statistical Machine Translation. *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*. COLING 1998. https://www.aclweb.org/anthology/C98-2153

NLTK. (n.d.). *Natural Language Toolkit—NLTK 3.5 documentation*. Retrieved 6 December 2019, from https://www.nltk.org/

NLTK. (2021). *Natural Language Toolkit (NLTK)* [Python]. Natural Language Toolkit. https://github.com/nltk/nltk/blob/f989fe65d421e7ea4d1037a00f07eaeee3ad6a29/nltk/translate/bleu_score.py (Original work published 2009)

Och, F. J., Tillmann, C., & Ney, H. (1999). Improved Alignment Models for Statistical Machine Translation. *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. https://www.aclweb.org/anthology/W99-0604

OPUS. (n.d.). *OPUS - an open source parallel corpus.* Retrieved 19 July 2021, from https://opus.nlpl.eu/

Ott, M., Edunov, S., Grangier, D., & Auli, M. (2018). Scaling Neural Machine Translation. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 1–9. https://doi.org/10.18653/v1/W18-6301

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Precup-Stiegelbauer, L.-R. (2013). Automatic Translations Versus Human Translations in Nowadays World. *Procedia - Social and Behavioral Sciences*, *70*, 1768–1777. https://doi.org/10.1016/j.sbspro.2013.01.252

Qin, L. (2013). *Learning Out-of-Vocabulary Words in Automatic Speech Recognition*. /paper/Learning-Out-of-Vocabulary-Words-in-Automatic-Qin/c24551688e3da4f292106a21859b8e4ccc7557fc

Sammut, C., & Webb, G. I. (Eds.). (2010). TF–IDF. In *Encyclopedia of Machine Learning* (pp. 986–987). Springer US. https://doi.org/10.1007/978-0-387-30164-8_832

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research methods for business students*. Pearson.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *ArXiv:1508.07909 [Cs]*. http://arxiv.org/abs/1508.07909

Sheffield Hallam University. (n.d.). *Guidance | Sheffield Hallam University*. Retrieved 4 September 2021, from https://www.shu.ac.uk/research/excellence/ethics-and-integrity/guidance

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *ArXiv:1409.3215 [Cs]*. http://arxiv.org/abs/1409.3215

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). *Accelerated DP Based Search for Statistical Translation*.

Universitet Utrecht. (n.d.). *Language structure: Variation and change*. Retrieved 11 July 2021, from https://www.uu.nl/en/research/utrecht-institute-of-linguistics-ots/research/language-structure-variation-and-change

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., & Uszkoreit, J. (2018). Tensor2Tensor for Neural Machine Translation. *ArXiv:1803.07416 [Cs, Stat]*. http://arxiv.org/abs/1803.07416

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-Based Word Alignment in Statistical Translation. *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*. COLING 1996. https://www.aclweb.org/anthology/C96-2141

Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., & Chao, L. S. (2019). Learning Deep Transformer Models for Machine Translation. *ArXiv:1906.01787 [Cs]*. http://arxiv.org/abs/1906.01787

Wu, D. (1994). Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. *ArXiv:Cmp-Lg/9406007*. http://arxiv.org/abs/cmp-lg/9406007

Wu, S., Wang, X., Wang, L., Liu, F., Xie, J., Tu, Z., Shi, S., & Li, M. (2020). Tencent Neural Machine Translation Systems for the WMT20 News Translation Task. *Proceedings of the Fifth Conference on Machine Translation*, 313–319. https://www.aclweb.org/anthology/2020.wmt-1.34

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv:1609.08144 [Cs]*. http://arxiv.org/abs/1609.08144

Zhang, Z., Wu, S., Liu, S., Li, M., Zhou, M., & Xu, T. (2019). Regularizing Neural Machine Translation by Target-Bidirectional Agreement. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 443–450. https://doi.org/10.1609/aaai.v33i01.3301443

Appendix A – Research Project Plan

Document-level Translation and Evaluation of Parallel-Bilingual Texts using Deep Learning Techniques

Brian Davis
Student Number: 27015685

INTRODUCTION AND JUSTIFICATION

Machine Translation enables us to understand the context of works that are not written in our native language, using automated software that can translate text without human intervention. This is incredibly useful when we have large amounts of user-generated content. Recent technological advances in Deep Learning have made the creation and distribution of Translation systems much easier than ever before.

However, most evaluations are based on automated scoring systems, rather than by that of the crowd, which is due to the length of time and human labour required to evaluate these systems (Papineni et al., 2002). Real-world performance is harder to dictate through these methods, which is where the need for human intervention is necessary. The aim is to make use of Deep Learning technologies to create an effective document-level model, performing evaluations using a mixed methodology with qualitative research methods to verify the integrity of evaluations done through these automated methods.

RESEARCH QUESTION, AIMS & OBJECTIVES

The project aim is to develop a Neural Machine Translation model making use of Deep Learning technologies to enhance translation quality. The optimal choice for the characteristics of this model will be explored throughout the research, making use of previous work to clarify and enhance choices made, to help find the most effective characteristics that lead to the highest quality model.

Research question /s

- How can we effectively use existing Deep Learning techniques to improve translation quality of Bilingual Parallel Corpora?
- How can we make use of Qualitative techniques to verify Neural Machine Translation model performance?

Objectives

Deep Learning model:

1. Identify the current research and existing gaps in knowledge / current challenges of Neural Machine Translation.

2. Identify valid sources of data that can be used. Perform validity tests on the data to ensure quality of source data.

3. Develop the model, making use of hyper-parameter tuning and utilising the best software for the task. Review the model and check the quality of the outputs.

4. Evaluate the performance of the model, using popular automated sources found in previous research. These will be BLEU and/or NIST.

Feedback Generation:

5. Develop and conduct timely research in the form of Questionnaires and possibly Interviews, to gauge feedback for the model.

6. Utilise this feedback to make improvements to the model where possible to enhance the strength of translation.

Evaluation:

7. Design a suitable process for evaluating the feedback regarding the performance of the translation model.

8. Perform final evaluation.

Deliverable

The output of this research will be a Neural Machine Translation model that can accurately translate from the source language to the target language. This system will aim to help those who want to translate to the target language but have limited experience of anything but the source language. Results of the surveys and Evaluation will also be included.

LITERATURE REVIEW

For the project, the key areas are Neural Machine Translation Systems themselves along with the areas of interest within the development of these systems. These involve attention-based approaches, Bilingual text alignment, the development of the encoder-decoder framework and the ongoing issue of out-of-vocabulary words.

Bilingual Text Alignment

Gale & Church (1993) described the creation of a program that could align sentences within a Bilingual Corpora. The alignment processes take place in two stages. First, paragraphs are aligned, and then the sentences within the paragraph are aligned further. Och et al. (1999) improved on the work of Nießen et al. (1998) and Vogel et al. (1996), with the prime focus being Word-to-Word statistical translation models. The focus on

using words for alignment is restrictive, due to the fact source words are only assigned to one target word at a time, and this can mean words with multiple meanings within a sentence can potentially lose their contextual meaning (Precup-Stiegelbauer, 2013). Gale & Church (1993) used humans to evaluate the accuracy and legibility of the results.

Improvements to the accuracy of NMT systems has been ongoing since the creation of the first production scale NMT system designed by Google (2016), but the focus of evaluation has continued to make use of automated frameworks. A reliable method is necessary to assess the quality of translations, Graham et al. (2017) showcased the use of a Likert scale fluency assessment, which shows an alternative approach to the popular automated approaches. Figure 1 shows an example of the Likert scale assessment interface used by Graham et al. (2017).

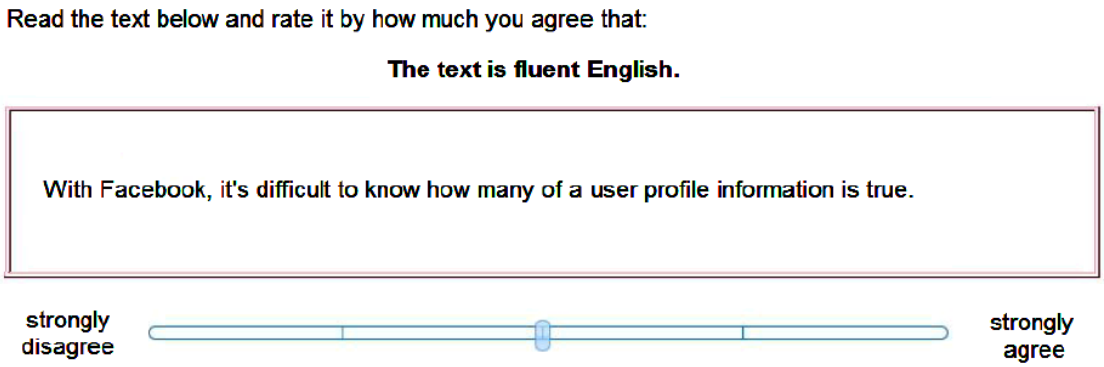


Figure 7 – Screenshot of fluency assessment interface (Graham et al., 2017)

Recent work now shifts to attention-based approaches to Alignment. Defined as Global Attention, this is an attention mechanism implemented within the decoder, that decides which parts of the source sentence it needs to pay attention to. This model is able to consider all hidden states of the encoder when deriving context, which is an important concept, as earlier work shows that word-to-word context is easily lost in translation (Bahdanau et al., 2014; Luong, Pham, et al., 2015).

Encoder-Decoder Framework

The Encoder-Decoder Framework was originally proposed by Cho et al. (2014) with the main focus being to learn phrase representations within Statistical Machine Translation systems. Figure 2 shows the initial proposed design of the Recurrent Neural Network (RNN) based Encoder-Decoder. In Statistical Machine Translation, the goal of the Decoder is to find a translation for a given source sentence (Cho et al., 2014).

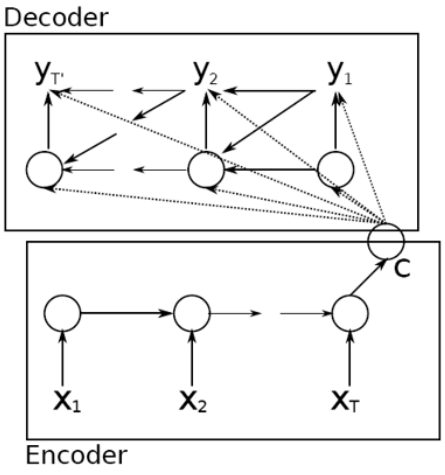


Figure 2 – Illustration of the RNN Encoder-Decoder (Cho et al., 2014)

Due to the way Recurrent models factor computation, Vaswani et al. (2017) created the Transformer model, a model that is designed to remove the short-comings of the RNN-based approach, due to sequential issues.

Attention-based NMTs have become a large focus on current research trends. Bahdanau et al. (2014) found a way to overcome the issue of model degradation by making use of a neural model of attention. Work begun to move away from sole RNN based approaches with attention in mind, leading to a stark increase in the use of Convolutional layers.

Gehring et al. (2017) introduced convolutional sequence to sequence learning. The architecture is based on Convolution Neural Networks (CNNs), allowing for a fully parallelised approach to be utilised. Whilst convolutions appear to improve substantially over RNNs, the Transformer Model introduced by Vaswani et al. (2017) depends on the self-attention mechanism, and has significantly improved translation quality (Wu et al., 2020).

We can continue to consider the approach of left-to-right (L2R) training for translation, but a better approach moving on in the future is the new approach of right-to-left (R2L) training. The idea is that translation with the R2L approach can significantly reduce inconsistencies between training and inference. Transformer model ensemble training done by Wu et al. (2020) proves that this approach has the potential to drastically improve quality going forward (Zhang et al., 2019). Fully convolutional model (Gehring et al., 2017) and transformer model (Vaswani et al., 2017) are architectures that are continuing to be focal points in potential quality improvements over the older RNN based NMT (Cho et al., 2014).

Out-of-vocabulary words

Rare words, or Out-of-vocabulary (OOV) words, are words that are found in the source language, but might not be contained in the target language. It is important that these OOV words are factored into model training, as the words that might be missed could hold key information into the context and meaning of the sentence (Arthur et al., 2016).

Input:	I come from Tunisia.
Reference:	チュニジア の 出身です。 Chunisia no shusshindesu. *(I'm from Tunisia.)*
System:	ノルウェー の 出身です。 Noruue- no shusshindesu. *(I'm from Norway.)*

Figure 3 – Example of a mistake made by NMT on OOV content words (Arthur et al., 2016).

Making use of a very large target vocabulary, Jean et al. (2015) was able to train a model without the need for a large change in training complexity. Making use of previous models developed by (Cho et al., 2014), (Sutskever et al., 2014) and (Bahdanau et al., 2014), the first results show model performance is on par with those that didn't factor unknown words into training.

The approach to factor in unknown words into the model, is the use of the UNK token (Luong, Sutskever, et al., 2015). Figure 3 shows the result of a translation from an NMT system, where the word for Tunisia has been replaced by the word for Norway in the vocabulary, meaning the result is now incorrect and considered an error (Arthur et al., 2016).

NMT systems are vulnerable to rare words, therefore their ability to generalise depends on the domain (Liu & Huang, 2020). If the domain has a plethora of OOV words, then the chance that the model will have issues with generalisation will be more apparent. Another consideration for OOV words causing issues with translation quality is where the source and target languages differ in their linguistic structure. Berrichi & Mazroui

(2021) considered a series of experiments for English -> Arabic translation, where they factored in morphosyntactic features such as Part-of-speech tagging (POS) to help address OOV words.

It can be argued that the use of these morphosyntactic features could be an interesting development into further improvements in the OOV problem domain.

RESEARCH DESIGN

Research Philosophy, Approach and Methodology

For the Deep Learning model development section of the project, a Pragmatistic philosophy is preferred to develop a model that can fulfil the objectives and successfully address the research problem. It is the most suitable philosophy for this project, as pragmatism allows for the use of methods that embody the researcher to collect credible, well-founded, reliable, and relevant data that can help to advance the research problem they are working towards addressing (Kelemen & Rumens, 2008; Saunders et al., 2016).

The research starts off with the analysis of existing data and works, which is the characteristic of a deductive approach, but then seeks to create a new model to explore development of an existing technique using some differing approaches. This is more akin to the abductive approach, where the aim is to identify themes and patterns, which makes this approach more attractive. The abductive approach goes back and forth, which in effect allows for a combination of deduction and induction, whereas deduction and induction are not as flexible when considered alone (Saunders et al., 2016; Suddaby, 2006).

The two main sections of this project aim to answer different research questions and will require a mixture of independent and combined evaluation against the research objectives.

- **Deep Learning model**

Deep Learning model assessment lends itself to an approach that is assessed almost entirely using classification metrics, such as precision, recall and confusion matrices (Minaee, 2019). Making use of automated evaluation metrics such as BLEU (Papineni et al., 2002), alongside traditional classification metrics, allows for sufficient performance comparison.

- **Feedback Generation & Evaluation**

Quantitative and Qualitative data could be used to evaluate the final model/s. To reduce bias of the results gathered from questionnaires, questions need to be written in a clear and concise manner, and each participant should see the same examples, so that all data is collected equally with the same measure of validity. Any survey questions need to be written in a manner that supports the philosophy.

As the sources of these data will be from different origins, this means that the appropriate methodology for this project would be a mixed method methodology.

For this project, the software development approach will follow the popular Waterfall Model introduced by Royce (1987). This is a classic approach to software engineering

where all sections of the development lifecycle are separated into more manageable chunks, which rely on sequential completion of tasks.

The plan for this project is to take advantage of the Waterfall Model but use it with a more adapted agile approach. Whilst the Agile methodologies of Kanban and SCRUM are much more flexible and focus on fast delivery of functionalities, Waterfall itself allows for a more structured process where each step must be completed in sequence. Combining elements of Agile and Waterfall provide a flexible but structured approach to the methodology (Andrei et al., 2019). This adapted approach can be seen in figure 4.

When considering the Research Strategy for this project, the strategy that appears the most reasonable and makes sense for this project is a Case Study. This type of study sets out to understand the dynamics of the topic at hand within its real-life setting. Importance is dictated by the determining of boundaries of the study as a key factor (Flyvbjerg, 2011; Saunders et al., 2016).

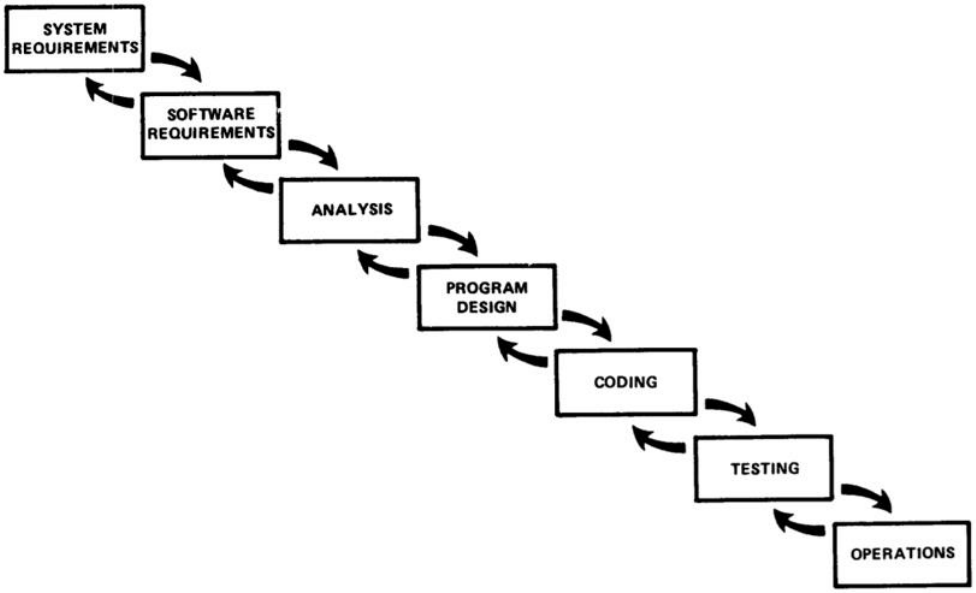


Figure 4: Implementation steps using Waterfall Model with Iterative Interaction (Royce, 1987)

Tools and Techniques

- **Deep Learning**

There are a variety of software packages available for Deep Learning, with the most popular ones being used as part of the extensive Python suite. There are other options worth considering for this project, which includes Amazon ML, R and perhaps even C or Java. Library use comes down to the developers of those resources and their preference, but Python is one of the best for AI and Machine Learning workloads due to its simplicity and consistency. It's ease of learning and extensive library make it the most popular choice amongst developers (Beklemysheva, n.d.).

As computation will be expected to be high for this project, appropriate hardware will need to be acquired, if the physical hardware currently being used is not found to be adequate. This hardware must have the capacity and computational power to allow for successful implementation and testing to be accomplished. As time goes on, constant

evaluation of the methods and software used for the project will be considered, and changes and improvements will be utilised where those are deemed necessary.

The most popular library for Deep Learning at the time of writing is Tensorflow (Abadi et al., 2016). This is due to its extensive tutorials available online and for its accessibility, capability, and integration with other tools, such as Keras, Apache Spark and others. As the requirements for the model are developed, further research will be conducted to discover if a more suitable alternative is available.

- **Feedback Generation & Evaluation**

As the feedback generation of the research will be collected electronically, it is important to use a reputable software to create and distribute the questionnaire. Qualtrics will be utilised as the preferred platform of choice for the creation of the Questionnaire, due to its ease of use and entirely web-based platform.

All the details regarding the questionnaire content are not known as this time, as this will be planned and created within the project timeline. The types of questions to use for the questionnaire will be researched through exploration of existing work, as to make sure the questions will give reliable and useful feedback that can be used to enhance the performance or accuracy of the model.

Evaluation will be conducted using appropriate Data Analysis elements of Python to understand and transform the survey results into a format that can be visualised and analysed further. The findings will then be used to evaluate the performance of the model, and the perception of the participants and how they thought the model met their specific requirements. Automated Evaluation of the model will take place in the environment that is deemed to be the most suitable at the time.

ETHICS, RISKS, AND ISSUES

Ethics and Legal Issues

For this project, there are expected to be no issues in relation to the ethics checklist documentation. Whilst this study will be conducting qualitative feedback research in the form of questionnaires, there are some unique ethical challenges that need to be considered.

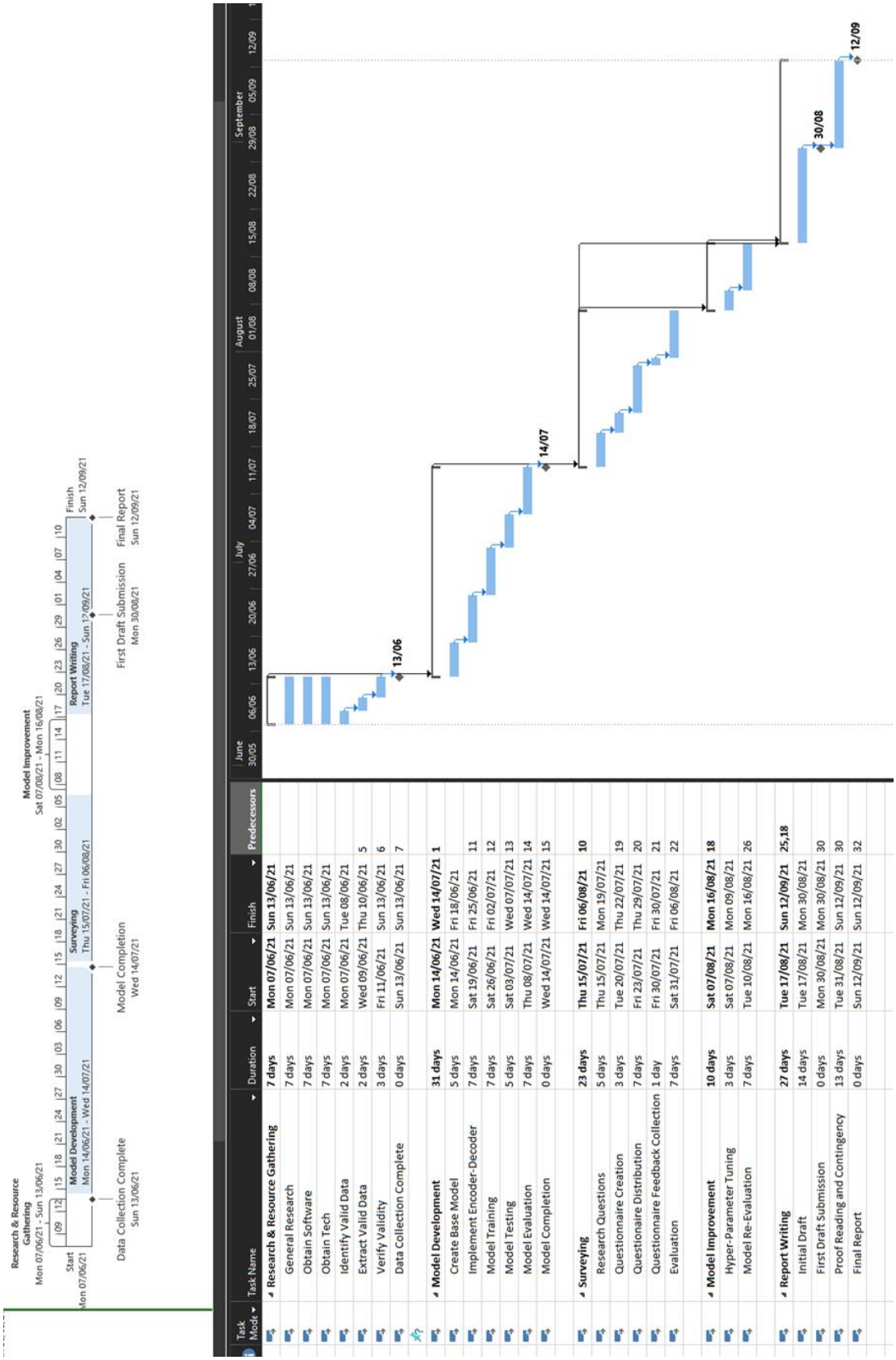
A complete ethics checklist can be found in Appendix A.

Risk Analysis

To achieve project objectives effectively, it is necessary to consider the risks that the proposed research carries, and how to mitigate these risks to avoid them hampering the successful completion of the project aim.

A detailed Risk Assessment Table is included in Appendix B.

TIME PLAN



Task Mode	Task Name	Duration	Start	Finish	Predecessors
	Research & Resource Gathering	**7 days**	**Mon 07/06/21**	**Sun 13/06/21**	
	General Research	7 days	Mon 07/06/21	Sun 13/06/21	
	Obtain Software	7 days	Mon 07/06/21	Sun 13/06/21	
	Obtain Tech	7 days	Mon 07/06/21	Sun 13/06/21	
	Identify Valid Data	2 days	Mon 07/06/21	Tue 08/06/21	
	Extract Valid Data	2 days	Wed 09/06/21	Thu 10/06/21	5
	Verify Validity	3 days	Fri 11/06/21	Sun 13/06/21	6
	Data Collection Complete	0 days	Sun 13/06/21	Sun 13/06/21	7
	Model Development	**31 days**	**Mon 14/06/21**	**Wed 14/07/21**	**1**
	Create Base Model	5 days	Mon 14/06/21	Fri 18/06/21	
	Implement Encoder-Decoder	7 days	Sat 19/06/21	Fri 25/06/21	11
	Model Training	7 days	Sat 26/06/21	Fri 02/07/21	12
	Model Testing	5 days	Sat 03/07/21	Wed 07/07/21	13
	Model Evaluation	7 days	Thu 08/07/21	Wed 14/07/21	14
	Model Completion	0 days	Wed 14/07/21	Wed 14/07/21	15
	Surveying	**23 days**	**Thu 15/07/21**	**Fri 06/08/21**	**10**
	Research Questions	5 days	Thu 15/07/21	Mon 19/07/21	
	Questionnaire Creation	3 days	Tue 20/07/21	Thu 22/07/21	19
	Questionnaire Distribution	7 days	Fri 23/07/21	Thu 29/07/21	20
	Questionnaire Feedback Collection	1 day	Fri 30/07/21	Fri 30/07/21	21
	Evaluation	7 days	Sat 31/07/21	Fri 06/08/21	22
	Model Improvement	**10 days**	**Sat 07/08/21**	**Mon 16/08/21**	**18**
	Hyper-Parameter Tuning	3 days	Sat 07/08/21	Mon 09/08/21	
	Model Re-Evaluation	7 days	Tue 10/08/21	Mon 16/08/21	26
	Report Writing	**27 days**	**Tue 17/08/21**	**Sun 12/09/21**	**25,18**
	Initial Draft	14 days	Tue 17/08/21	Mon 30/08/21	
	First Draft Submission	0 days	Mon 30/08/21	Mon 30/08/21	30
	Proof Reading and Contingency	13 days	Tue 31/08/21	Sun 12/09/21	30
	Final Report	0 days	Sun 12/09/21	Sun 12/09/21	32

RISK ANALYSIS TABLE

Risk	Likelihood	Severity	Impact	Mitigation	Adjusted Likelihood	Adjusted Severity	Adjusted Impact
Data Difficulty	Low(2)	High(4)	8	Ensure collected data is backed up correctly as access to the site may be lost. Collected data for model building must be kept safe	Low(1)	Low(1)	4
Inability to get Data	Low(2)	Very High(5)	5	Seek alternatives if data can't be found online, maybe from an academic source who has done this work before	Low(1)	Medium(3)	3
Hardware Failure	Medium(3)	Very High(5)	9	Ensure replacement hardware is available if any issues arise with the Virtual Machine that will be used for the task	Low(2)	Medium(3)	5
Personal Illness	Low(1)	Medium(3)	3	Time for each section of project plan will include leeway for extra time to mitigate time lost for sudden illness	Low(1)	Low(1)	1
Model Inaccuracy Issues	Medium(3)	Medium(3)	10	Allow additional time for the model to be developed, possibly alongside other parts of the project to keep project flowing smoothly	Low(2)	Low(2)	8
Time Allocation Issues	High(4)	Very High(5)	10	Ensure time management is planned with realistic timescales to allow for work to be complete on time alongside other commitments	Medium(3)	Medium(3)	7
Software Problems	Low(2)	High(4)	8	Use Git backups to make sure all data and progress is always backed up so that work can be resumed elsewhere	Low(1)	Medium(3)	5
Questionnaire Distribution	Low(2)	High(4)	8	Distribute through efficient channels, make use of resources available to the Supervisor to distribute to a variety of participants	Low(1)	Low(2)	4
Lack of relevant results for Manual Evaluation	Medium(3)	Medium(3)	7	Conduct as high a volume of questionnaire research as possible to collect a high variety of results that can help to aid in further evaluation.	Low(2)	Medium(3)	5
Inaccurate Automated Evaluation	Low(2)	Medium(3)	6	Use a variety of popular automated evaluation systems to allow for plenty of results to cross examine, t ensure model is being evaluated correctly	Low(2)	Low(2)	4

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. 21.

Andrei, B.-A., Casu-Pop, A.-C., Gheorghe, S.-C., & Boiangiu, C.-A. (2019). A STUDY ON USING WATERFALL AND AGILE METHODS IN SOFTWARE PROJECT MANAGEMENT.

Arthur, P., Neubig, G., & Nakamura, S. (2016). Incorporating Discrete Translation Lexicons into Neural Machine Translation. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 1557–1567. https://doi.org/10.18653/v1/D16-1162

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv:1409.0473 [Cs, Stat]. http://arxiv.org/abs/1409.0473

Beklemysheva, A. (n.d.). Why Use Python for AI and Machine Learning? Retrieved 4 May 2021, from https://steelkiwi.com/blog/python-for-ai-and-machine-learning/

Berrichi, S., & Mazroui, A. (2021). Addressing Limited Vocabulary and Long Sentences Constraints in English–Arabic Neural Machine Translation. Arabian Journal for Science and Engineering. https://doi.org/10.1007/s13369-020-05328-2

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. ArXiv:1406.1078 [Cs, Stat]. http://arxiv.org/abs/1406.1078

Flyvbjerg, B. (2011). Case Study (SSRN Scholarly Paper ID 2278194). Social Science Research Network. https://papers.ssrn.com/abstract=2278194

Gale, W. A., & Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19(1), 75–102.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional Sequence to Sequence Learning. ArXiv:1705.03122 [Cs]. http://arxiv.org/abs/1705.03122

Google. (2016, September 27). A Neural Network for Machine Translation, at Production Scale. Google AI Blog. http://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. Natural Language Engineering, 23(1), 3–30. https://doi.org/10.1017/S1351324915000339

Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2015). On Using Very Large Target Vocabulary for Neural Machine Translation. ArXiv:1412.2007 [Cs]. http://arxiv.org/abs/1412.2007

Kelemen, M., & Rumens, N. (2008). An Introduction to Critical Management Research. SAGE Publications, Ltd. https://doi.org/10.4135/9780857024336

Liu, B., & Huang, L. (2020). NEJM-enzh: A Parallel Corpus for English-Chinese Translation in the Biomedical Domain. ArXiv:2005.09133 [Cs]. http://arxiv.org/abs/2005.09133

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. ArXiv:1508.04025 [Cs]. http://arxiv.org/abs/1508.04025

Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., & Zaremba, W. (2015). Addressing the Rare Word Problem in Neural Machine Translation. ArXiv:1410.8206 [Cs]. http://arxiv.org/abs/1410.8206

Minaee, S. (2019, October 28). 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics. Medium. https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce

Nießen, S., Vogel, S., Ney, H., & Tillmann, C. (1998). A DP based Search Algorithm for Statistical Machine Translation. COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics. COLING 1998. https://www.aclweb.org/anthology/C98-2153

Och, F. J., Tillmann, C., & Ney, H. (1999). Improved Alignment Models for Statistical Machine Translation. 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. https://www.aclweb.org/anthology/W99-0604

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. https://doi.org/10.3115/1073083.1073135

Precup-Stiegelbauer, L.-R. (2013). Automatic Translations Versus Human Translations in Nowadays World. Procedia - Social and Behavioral Sciences, 70, 1768–1777. https://doi.org/10.1016/j.sbspro.2013.01.252

Royce, W. W. (1987). Managing the development of large software systems: Concepts and techniques. Proceedings of the 9th International Conference on Software Engineering, 328–338.

Saunders, M., Lewis, P., & Thornhill, A. (2016). Research methods for business students. Pearson.

Suddaby, R. (2006). From the editors: What grounded theory is not. Academy of Management Journal, 49(4), 633–642. https://doi.org/10.5465/AMJ.2006.22083020

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. ArXiv:1409.3215 [Cs]. http://arxiv.org/abs/1409.3215

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. ArXiv:1706.03762 [Cs]. http://arxiv.org/abs/1706.03762

Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-Based Word Alignment in Statistical Translation. COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics. COLING 1996. https://www.aclweb.org/anthology/C96-2141

Wu, S., Wang, X., Wang, L., Liu, F., Xie, J., Tu, Z., Shi, S., & Li, M. (2020). Tencent Neural Machine Translation Systems for the WMT20 News Translation Task. Proceedings of the Fifth Conference on Machine Translation, 313–319. https://www.aclweb.org/anthology/2020.wmt-1.34

Zhang, Z., Wu, S., Liu, S., Li, M., Zhou, M., & Xu, T. (2019). Regularizing Neural Machine Translation by Target-Bidirectional Agreement. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 443–450. https://doi.org/10.1609/aaai.v33i01.3301443

Appendix B – Completed Research Ethics Checklist

Sheffield Hallam University

UREC2 RESEARCH ETHICS PROFORMA FOR STUDENTS UNDERTAKING LOW RISK PROJECTS WITH HUMAN PARTICIPANTS

This form is designed to help students and their supervisors to complete an ethical scrutiny of proposed research. The University Research Ethics Policy should be consulted before completing the form. The initial questions are there to check that completion of the UREC 2 is appropriate for this study. The final responsibility for ensuring that ethical research practices are followed rests with the supervisor for student research.

Note that students and staff are responsible for making suitable arrangements to ensure compliance with the General Data Protection Act (GDPR). This involves informing participants about the legal basis for the research, including a link to the University research data privacy statement and providing details of who to complain to if participants have issues about how their data was handled or how they were treated (full details in module handbooks). In addition the act requires data to be kept securely and the identity of participants to be anonymized. They are also responsible for following SHU guidelines about data encryption and research data management. Information on the Ethics Website

The form also enables the University and College to keep a record confirming that research conducted has been subjected to ethical scrutiny.

The form may be completed by the student and the supervisor and/or module leader (as applicable). In all cases, it should be counter-signed by the supervisor and/or module leader, and kept as a record showing that ethical scrutiny has occurred. Some courses may require additional scrutiny. Students should retain a copy for inclusion in their research projects, and a copy should be uploaded to the relevant module Blackboard site.

Please note that it may be necessary to conduct a health and safety risk assessment for the proposed research. Further information can be obtained from the College Health and Safety Service.

Checklist Questions to ensure that this is the correct form

1. Health Related Research with the NHS or Her Majesty's Prison and Probation Service (HMPPS)or with participants unable to provide informed consent

Question	Yes/No
1. Does the research involve? • Patients recruited because of their past or present use of the NHS	No
• Relatives/carers of patients recruited because of their past or present use of the NHS	No
• Access to data, organs or other bodily material of past or present NHS patients	No
• Foetal material and IVF involving NHS patients	No
• The recently dead in NHS premises	No
• Prisoners or others within the criminal justice system recruited for health-related research*	No
• Police, court officials, prisoners or others within the criminal justice system*	No
• Participants who are unable to provide informed consent due to their incapacity even if the project is not health related	No
2. Is this a research project as opposed to service evaluation or audit?	No

For NHS definitions of research etc. please see the following website http://www.hra.nhs.uk/documents/2013/09/defining-research.pdf	

If you have answered **YES** to questions **1 & 2** then you **MUST** seek the appropriate external approvals from the NHS, Her Majesty's Prison and Probation Service (HMPPS) under their independent Research Governance schemes. Further information is provided below.
https://www.myresearchproject.org.uk

NB College Teaching Programme Research Ethics Committees (CTPRECS) provide Independent Scientific Review for NHS or HMPPS research and initial scrutiny for ethics applications as required for university sponsorship of the research. Applicants can use the IRAS proforma and submit this initially to their CTPREC.

1. Checks for Research with Human Participants

Question	**Yes/No**
1. Will any of the participants be vulnerable? *Note: Vulnerable' people include children and young people, people with learning disabilities, people who may be limited by age or sickness, people researched because of a condition they have, etc. See full definition on ethics website*	No
2. Are drugs, placebos or other substances (e.g. food substances, vitamins) to be administered to the study participants or will the study involve invasive, intrusive or potentially harmful procedures of any kind?	No
3. Will tissue samples (including blood) be obtained from participants?	No
4. Is pain or more than mild discomfort likely to result from the study?	No
5. Will the study involve prolonged or repetitive testing?	No
6. Is there any reasonable and foreseeable risk of physical or emotional harm to any of the participants? *Note: Harm may be caused by distressing or intrusive interview questions, uncomfortable procedures involving the participant, invasion of privacy, topics relating to highly personal information, topics relating to illegal activity, or topics that are anxiety provoking, etc.*	No
7. Will anyone be taking part without giving their informed consent?	No
8. Is it covert research? *Note: 'Covert research' refers to research that is conducted without the knowledge of participants.*	No
9. Will the research output allow identification of any individual who has not given their express consent to be identified?	No

If you have answered **YES** to any of these questions you are **REQUIRED** to complete and submit a UREC 3 or UREC4). Your supervisor will advise. If you have answered **NO** to all these questions then proceed with this form (UREC 2).

General Details

Name of student	Brian Davis
SHU email address	

Course or qualification (student)	MSc Big Data Analytics
Name of supervisor	Bayode Ogunleye
email address	
Title of proposed research	Document-level Translation and Evaluation of Parallel-Bilingual Texts using Deep Learning Techniques
Proposed start date	14th June 2021
Proposed end date	13th September 2021
Background to the study and scientific rationale for undertaking it.	Neural Machine Translation is the current focus of translation systems, where our need for accurate translation systems is key to allowing communication to people who cannot speak either the source or the target language. Many techniques have been applied to improve research, and this research aims to take a slightly different approach to some of the methods done already.
Aims & research question(s)	1) How can we effectively use existing Deep Learning techniques to improve translation quality of Bilingual Parallel Corpora? 2) How can we make use of Qualitative techniques to verify Neural Machine Translation model performance? Aims 1) Identify the current research and existing gaps in knowledge / current challenges of Neural Machine Translation. 2) Identify valid sources of data that can be used. Perform validity tests on the data to ensure quality of source data. 3) Develop the model, making use of hyper-parameter tuning and utilising the best software for the task. Review the model and check the quality of the outputs. 4) Evaluate the performance of the model, using popular automated sources found in previous research. These will be BLEU and/or NIST. 5) Develop and conduct timely research in the form of Questionnaires and possibly Interviews, to gauge feedback for the model. 6) Utilise this feedback to make improvements to the model where possible to enhance the strength of translation. 7) Design a suitable process for evaluating the feedback regarding the performance of the translation model. 8) Perform final evaluation.
Methods to be used for: 1.recruitment of participants,	1. Participants will be recruited based on their fluency of the output language; this will be in the form of an online questionnaire or a live

2.data collection, 3. data analysis.	demonstration. 2. Data will be collected from 3rd party sites and quality checked to make sure it is suitable for purpose beforehand. The data must be accessible to anyone and can be gained without the need for any 3rd party tools. 3. Analysis will be conducted through the tools and techniques defined within the proposal. All analysis will take place within the same location as other work.
Outline the nature of the data held, details of anonymisation, storage and disposal procedures as required.	All data stored will either be translation data for model evaluation and training or will be fully anonymized feedback data from questionnaires. Data will be stored on the local machines, and disposal of the data will occur once the assignment is submitted for marking.

3. Research in Organisations

Question	**Yes/No**
1. Will the research involve working with/within an organisation (e.g. school, business, charity, museum, government department, international agency, etc.)?	No
2. If you answered YES to question 1, do you have granted access to conduct the research? *If YES, students please show evidence to your supervisor. PI should retain safely.*	N/A
3. If you answered NO to question 2, is it because: A. you have not yet asked B. you have asked and not yet received an answer C. you have asked and been refused access. *Note: You will only be able to start the research when you have been granted access.*	N/A

4. Research with Products and Artefacts

Question	**Yes/No**
1. Will the research involve working with copyrighted documents, films, broadcasts, photographs, artworks, designs, products, programmes, databases, networks, processes, existing datasets or secure data?	Yes

2. If you answered YES to question 1, are the materials you intend to use in the public domain? *Notes: 'In the public domain' does not mean the same thing as 'publicly accessible'.* • *Information which is 'in the public domain' is no longer protected by copyright (i.e. copyright has either expired or been waived) and can be used without permission.* • *Information which is 'publicly accessible' (e.g. TV broadcasts, websites, artworks, newspapers) is available for anyone to consult/view. It is still protected by copyright even if there is no copyright notice. In UK law, copyright protection is automatic and does not require a copyright statement, although it is always good practice to provide one. It is necessary to check the terms and conditions of use to find out exactly how the material may be reused etc.* *If you answered YES to question 1, be aware that you may need to consider other ethics codes. For example, when conducting Internet research, consult the code of the Association of Internet Researchers; for educational research, consult the Code of Ethics of the British Educational Research Association.*	Yes
3. If you answered NO to question 2, do you have explicit permission to use these materials as data? *If YES, please show evidence to your supervisor.*	N/A
4. If you answered NO to question 3, is it because: A. you have not yet asked permission B. you have asked and not yet received and answer C. you have asked and been refused access. *Note You will only be able to start the research when you have been granted permission to use the specified material.*	**A/B/C**

Adherence to SHU policy and procedures

Personal statement	
I can confirm that: YES - I have read the Sheffield Hallam University Research Ethics Policy and Procedures YES - I agree to abide by its principles.	
Student	
Name: Brian Davis	Date: 04/05/21
Signature:	
Supervisor or other person giving ethical sign-off	
I can confirm that completion of this form has not identified the need for ethical approval by the FREC or an NHS, Social Care or other external REC. The research will not commence until any approvals required under Sections 3 & 4 have been received and any necessary health and safety measures are in place.	

Name: Bayode Ogunleye	Date: 09/06/21
Signature:	
Additional Signature if required by course:	
Name:	Date:
Signature:	

Please ensure the following are included with this form if applicable, tick box to indicate:

	Yes	No	N/A
Research proposal if prepared previously	☒	☐	☐
Any recruitment materials (e.g. posters, letters, etc.)	☐	☐	☐
Participant information sheet	☒	☐	☐
Participant consent form	☒	☐	☐
Details of measures to be used (e.g. questionnaires, etc.)	☐	☐	☐
Outline interview schedule / focus group schedule	☐	☐	☐
Debriefing materials	☐	☐	☐
Health and Safety Project Safety Plan for Procedures	☐	☐	☐

Appendix C – Primary Data and Secondary Data

A – Sample of Primary Data

R	S	T	U	V	W	X
Q0	Q1_1	Q2_1	Q3_1	Q4_1	Q5_1	Q6_1
By answering yes to t 1. I have read the inf 2. My questions abou 3. I understand that I 4. I agree to provide 5. I wish to participat 6. I consent to the in Thank you very muc	Read the text below	Read the text below	Read the text below	Read the text below Input: I want to know	Read the text below Input: Which book is	Read the text below Input: Come in, the c
I wish to participate i	苹果掉落的地方不会	会出现什么情况呢？	天气预报说明天会下	Output: 我想知道汤姆	Output: 哪本是你们的	Output: 进来，门开着
{"ImportId":"QID31"}	{"ImportId":"QID53_1	{"ImportId":"QID57_1	{"ImportId":"QID58_1	{"ImportId":"QID60_1	{"ImportId":"QID62_1	{"ImportId":"QID63_1
2	1	1	1	4	2	1
2	2	1	1	4	2	1
2	1	1	1	1	1	4
2	2	2	3	2	2	2
2	1	3	1	3	1	1
2	2	3	3	3	3	3
2	3	2	2	4	1	5
2	2	4	2	2	4	4
2	2	2	1	2	1	1
2	2	2	1	5	1	4

B – Sample of Secondary Data

```
  <tuv xml:lang="en"><seg>And those simple themes aren't really themes about the complex science of what's going on, but things t
  <tuv xml:lang="zh"><seg>这些简单的话题确实不是 有关那复杂的科学有了怎样的发展， 而是一些我们都恰好知道的事情。</seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg>And I'm going to start with this one: If momma ain't happy, ain't nobody happy.</seg></tuv>
  <tuv xml:lang="zh"><seg>接下来我就来说一个， 如果老妈不高兴了，大家都别想开心。</seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg>We know that, right? We've experienced that.</seg></tuv>
  <tuv xml:lang="zh"><seg>我们都知道，不是吗？我们都经历过。</seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg>And if we just take that and we build from there, then we can go to the next step, which is that if the
  <tuv xml:lang="zh"><seg>接下来如果我们能理解这一点 从这里出发， 可以得出下一步的， 那就是如果海洋不高兴了 大家也都别想开心。</s
</tu>
<tu>
  <tuv xml:lang="en"><seg>That's the theme of my talk.</seg></tuv>
  <tuv xml:lang="zh"><seg>这就是我演讲的主题。</seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg>And we're making the ocean pretty unhappy in a lot of different ways.</seg></tuv>
  <tuv xml:lang="zh"><seg>我们正在通过许多不同的方法惹怒海洋。</seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg>This is a shot of Cannery Row in 1932.</seg></tuv>
  <tuv xml:lang="zh"><seg>这是1932年在坎纳里鲁夫拍的一副照片</seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg>Cannery Row, at the time, had the biggest industrial canning operation on the west coast.</seg></tuv>
  <tuv xml:lang="zh"><seg>那时的坎纳里鲁夫， 有着西海岸最大的 工业化罐头工厂。</seg></tuv>
```

Note: Please see the GitHub repository found in Appendix G to see the full data.

Appendix D – Information Sheet and Consent via Survey

A – Participant Information Sheet

PARTICIPANT INFORMATION SHEET

Title of Project: Document-level Translation and Evaluation of Parallel-Bilingual Texts using Deep Learning Techniques

Translation is a key tool in the modern world. One of the earliest challenges for AI was the translation of text from one language to another. Neural Machine Translation, also known as NMT, is an approach to Automated Machine Translation that utilises the strengths of Neural Networks to create functioning models able to translate sentences and phrases from a source language to a target language.

In this project, the aim to create a functioning NMT model has been realised on text based on English as the source, and Mandarin Chinese as the target. Through the adjustment of model parameters, many versions of these NMTs have been tested and analysed via automated methods, but methods that rely on automated metrics can sometimes miss the mark.

This study aims to collect feedback from the crowd, to discover just how well the model has learnt the ability to translate custom inputs, different to those it has been trained on. This is important to understand if the model is learning, and if the context is able to be maintained within the models Attention. The project will then explore these results to discover better ways to further improve the model towards final conclusion of findings.

We are looking for those familiar with the target language of Mandarin Chinese to help us to evaluate the strengths of the model through a variety of questions that aim to evaluate the models over a number of different angles.

It is up to you to decide if you want to take part. Your participation in this survey is optional and completely anonymous, no information that can link back to you will be obtained through your participation. This information will be used to improve the developed model, to help continue the drive for better computer-based translation solutions.

You can decide to withdraw during the survey without giving a reason - simply by closing the browser tab or window - and you can decide not to answer a particular question. However, once you have completed the survey you cannot withdraw your data because it is stored in a fully anonymous form.

If you choose to take part in the survey, it is expected that the survey should take no more than 10 minutes of your time.

Once the study is complete, the data will be stored for a period of up to one year, and will be deleted within one year of the completion of the project. The data will be redacted and anonymised before being added to the Appendix of the research paper - which may or may not be published. If you wish to see a copy of this before it is uploaded please inform the researcher.

The findings from the study will be primarily used to improve the final publishable model, but may also be used in the findings and conclusion section of the research paper.

If you would like to know more information about the study before making a decision, please feel free to contact the researcher.

Researcher Details:

The University undertakes research as part of its function for the community under its legal status. Data protection allows us to use personal data for research with appropriate safeguards in place under the legal basis of **public tasks that are in the public interest.** A full statement of your rights can be found at https://www.shu.ac.uk/about-this-website/privacy-policy/privacy-notices/privacy-notice-for-research. However, all University research is reviewed to ensure that participants are treated appropriately and their rights respected. This study is undertaken in partial fulfilment of the requirements of Sheffield Hallam University for the degree of Master of Science in Big Data Analytics.
Further information regarding ethics of this research can be found at https://www.shu.ac.uk/research/excellence/ethics-and-integrity/policies.

You should contact the Data Protection Officer if:

- you have a query about how your data is used by the University;
- you would like to report a data security breach (e.g. if you think your personal data has been lost or disclosed inappropriately);
- you would like to complain about how the University has used your personal data;

DPO@shu.ac.uk

You should contact the Head of Research Ethics (Professor Ann Macaskill) if:

- you have concerns with how the research was undertaken or how you were treated.

a.macaskill@shu.ac.uk

Postal address: Sheffield Hallam University, Howard Street, Sheffield S1 1WBT Telephone: 0114 225 5555

B – Consent Form via Survey

By answering yes to the below question and proceeding to the survey you confirm that you consent to take part in survey in accordance with the following conditions:

1. I have read the information above which details what this study is about.
2. My questions about the study have been answered to my satisfaction and I understand that I may ask further questions at any point.
3. I understand that I am free to withdraw from the study within the time limits outlined in the information above, without giving a reason for my withdrawal or to decline to answer any particular questions in the study without any consequences to my future treatment by the researcher.
4. I agree to provide information to the researchers under the conditions of confidentiality set out in the information above
5. I wish to participate in the study under the conditions set out in the information above.
6. I consent to the information collected for the purposes of this research study, once anonymised (so that I cannot be identified), to be used for any other research purposes.

Thank you very much in advance.

I wish to participate in the survey:

Yes ○

Appendix E – Questionnaire

Survey Body

Q0

The first 3 questions will test translation accuracy in regards to Fluency.

Page Break

Q1

- Read the text below and rate it by how much you agree that: **The text is fluent Mandarin Chinese.**

苹果掉落的地方不会离树干很远。

Strongly agree | Somewhat agree | Neither agree nor disagree | Somewhat disagree | Strongly disagree

Page Break

Q2

- Read the text below and rate it by how much you agree that: **The text is fluent Mandarin Chinese.**

会出现什么情况呢?

Strongly agree | Somewhat agree | Neither agree nor disagree | Somewhat disagree | Strongly disagree

Page Break

Q3

- Read the text below and rate it by how much you agree that: **The text is fluent Mandarin Chinese.**

天气预报说明天会下雪。

Strongly agree | Somewhat agree | Neither agree nor disagree | Somewhat disagree | Strongly disagree

Page Break

Q0

These next 3 questions will test the accuracy of translations against the input.

Page Break

Q4

- Read the text below and rate it by how much you agree that: **The output sentence accurately translates the input sentence.**

Input: I want to know why Tom wants us to do that.

Output: 我想知道汤姆为什么要帮我们那样做。

Strongly agree | Somewhat agree | Neither agree nor disagree | Somewhat disagree | Strongly disagree

Page Break

Q5

- Read the text below and rate it by how much you agree that: **The output sentence accurately translates the input sentence.**

Input: Which book is yours?

Output: 哪本是你们的书?

Strongly agree | Somewhat agree | Neither agree nor disagree | Somewhat disagree | Strongly disagree

Page Break

Q6

- Read the text below and rate it by how much you agree that: **The output sentence accurately translates the input sentence.**

Input: Come in, the door's open.

Output: 进来，门开着。

Strongly agree | Somewhat agree | Neither agree nor disagree | Somewhat disagree | Strongly disagree

Page Break

Q7

- Read the text below and rate it by how much you agree that: **The context of output sentence B is close to the expected sentence A.**

A: 这才是重点。

B: 这是今天的情况 。

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
○	○	○	○	○

Page Break

Q8

- Read the text below and rate it by how much you agree that: **The context of output sentence B is close to the expected sentence A.**

A: 我试着不去想了。

B: 我不想那样想。

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
○	○	○	○	○

Page Break

Q9

- Read the text below and rate it by how much you agree that: **The context of output sentence B is close to the expected sentence A.**

A: 我早上会在家。

B: 我明天早上在家。

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
○	○	○	○	○

Page Break

Page Break

Q10

- Read the text below and rate it by how much you agree that: **The quality of the translation B closely matches A.**

A: 我们给大家看一段视频。

B: 我们一起给你看一个视频。

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
○	○	○	○	○

Page Break

Q11

- Read the text below and rate it by how much you agree that: **The output sentence accurately translates the input.**

Input: This is the office in which he works.

Output: 那是他工作的办公室。

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
○	○	○	○	○

Page Break

Q12

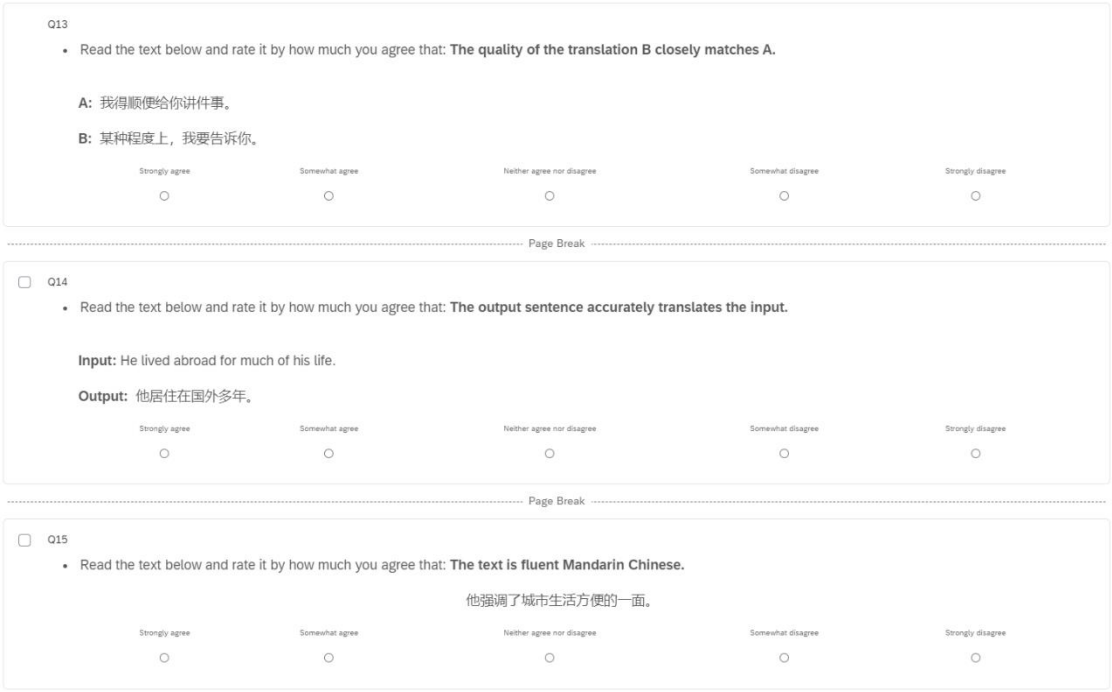
- Read the text below and rate it by how much you agree that: **The quality of the translation B closely matches A.**

A: 他放学后打棒球。

B: 他放学后打棒球。

Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
○	○	○	○	○

Page Break


Q13
Read the text below and rate it by how much you agree that: The quality of the translation B closely matches A.
A: 我得顺便给你讲件事。
B: 某种程度上，我要告诉你。
Strongly agree
Somewhat agree
Neither agree nor disagree
Somewhat disagree
Strongly disagree
Page Break
Q14
Read the text below and rate it by how much you agree that: The output sentence accurately translates the input.
Input: He lived abroad for much of his life.
Output: 他居住在国外多年。
Strongly agree
Somewhat agree
Neither agree nor disagree
Somewhat disagree
Strongly disagree
Page Break
Q15
Read the text below and rate it by how much you agree that: The text is fluent Mandarin Chinese.
他强调了城市生活方便的一面。
Strongly agree
Somewhat agree
Neither agree nor disagree
Somewhat disagree
Strongly disagree

Appendix F – Publication Procedure Form

Sheffield Hallam University | College of Business, Technology and Engineering

Research Skills and Dissertation Module (55-706556).

PUBLICATION PROCEDURE FORM

In this module, while you create your own research question or topic area, your supervisor makes a significant intellectual contribution to this work as the research progresses. Your supervisor will make the decision on whether your work merits publication based on the quality of the work you have produced. Your supervisor will co-author the paper for publication with you and your supervisor will both be listed as authors. You are required to sign the declaration below to confirm that you understand and will follow this procedure.

Declaration:

I Brian Davis confirm that I understand will comply with the Publication Procedure outlined in the Module Handbook and the Blackboard Site.		
Student: Brian Davis	Signature	Date 10/09/2021
Supervisor: Bayode Ogunleye	Signature	Date 10/09/2021

Appendix G – GitHub Repository Location

Please see the following link to access the code used for this project.

https://github.com/BrianALDavis/MSc-Machine_translation_en-zh