7-25-2023 1:30 PM

Enhancing the Performance of NMT Models Using the Data-Based Domain Adaptation Technique for Patent Translation

Maimoonah Ahmed, *Western University*

Supervisor: Ouda, Abdelkader, *The University of Western Ontario*
Co-Supervisor: Abusharkh, Mohamed, *Ferris State University*

A thesis submitted in partial fulfillment of the requirements for the Master of Engineering Science degree in Electrical and Computer Engineering

Abstract

During today's age of unparalleled connectivity, language and data have become powerful tools capable of enabling effective communication and cross-cultural collaborations. Neural machine translation (NMT) models are especially capable of leveraging linguistic knowledge and parallel corpora to increase global connectivity and act as a tool for the transmission of knowledge. In this thesis, we apply a data-based domain adaptation technique to fine-tune pre-existing NMT transformer models with attention mechanisms for the task of patent translation from English to Japanese. Languages, especially in the context of patents, can be very nuanced. A clear understanding of the intended meaning requires comprehensive domain knowledge and expert linguistic abilities, which may become expensive and time-consuming. Automating the translation process is helpful; however, commercially available NMT models perform poorly for this task as they are not trained on highly technical words whose meaning may depend on the relevant domain in which they are used. Our aim is to enhance the performance of translation models on highly technical inputs using a range of essential steps, focusing on data-based domain adaptation. These steps collectively enhance the NMT model's performance and increase the model's baseline BiLingual Evaluation Understudy (BLEU) score by 41.22%.

Keywords: Machine translation, NLP, neural machine translation, patent translation, domain adaptation, self-attention, Transformer architecture, low-resource domains, technical information, translating legal documents

Lay Summary

In an age of innovation driven by technology, a globally increasing number of patent applications are being filed, according to the World Intellectual Property Organization (WIPO). In 2022, the UN reported that patent applications increased to more than 278,000 patent applications, and according to Carsten Fink, the chief economist at WIPO, 2022 ”represents the 13th year of uninterrupted growth” [1]. With patent applications and global connections around the world increasing at a steady rate, the need for patent translation using machine translation (MT) systems also increases. Translation in any field is a complex problem requiring a deep knowledge of the natural language pairs involved. Simply applying language rules to translate text does not return accurate or acceptable translations because language is a complex, nuanced system that is affected by different cultural, social, and historical factors. Automated translation is even more complex for patent documents as they are highly technical documents, often containing legal terminology. To translate a patent, the translator must be well-versed in subjects pertaining to legal jargon and the relevant technical domain.

Due to advancements in the field of MT, especially in neural machine translation (NMT), the field of patent translation has seen a growth in interest. Currently, the most common architectures of NMT used for translation are the transformer model, recurrent neural networks (RNNs), and the encoder-decoder architecture, or variants of the three. This study uses a pre-trained transformer model with an encoder-decoder architecture and attention mechanisms. Another source of challenge stems from low-resource domains where the translation model is not able to learn highly technical words and phrases. One technique that aims to remedy that is data-based domain adaptation. This technique leverages the value of data in order to train the model to perform well in its target domain.

References: [1] United Nations. (2022). Patent filings hit a record high in 2022, Un Agency reveals. Retrieved April 30, 2023, from https://news.un.org

Acknowlegements

First and foremost, I would like to express my sincerest and heartfelt gratitude to my supervisors, Dr. Abdelkader Ouda and Dr. Mohamed Abusharkh. Your invaluable guidance, support, and encouragement throughout my research journey made this thesis not only possible but easier during stressful days.

I would also like to thank my closest friends, Rijaa, Chhavi, Ashfia, Dalia, Alyssa, and Kate, who have inspired me and provided their unwavering support throughout my time at Western. From celebrating small victories and proofreading to lending a listening ear and making me laugh during challenging times, your friendship has made this possible.

And, of course, a big thank you to my parents, Shakil and Seemi, and my brother Umer for their trust and faith in my dreams. I could not accomplish this without your support, care, and love. Thank you for all the million little things.

I also had the pleasure of working with the Data Science team at XLScout, our Mitacs' industry partner. Thank you for kindly welcoming me to your team and for all the help you provided to achieve the research goals. I would especially like to thank Khushwant Rai and Sandeep Kohli for their expertise and mentorship.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

When thinking about the process of translation, it may seem ostensibly simple: first, a translator, whether human or machine, must decode the meaning of the original text and then encode that meaning back into the target language specified. However, when we break down these two steps, the hidden complexities beneath the surface become clearer. To achieve an accurate translation, the translator must have a comprehensive understanding of the components representing languages, such as syntax, semantics, lexicology, etc., of both the source and target languages as well as the culture of the speakers of both languages. In the practice of machine translation, we refer to the language of the original input text as the source language (SL), and the language being translated into is referred to as the target language (TL) [118].

Additionally, the task of translating includes having to navigate lexical ambiguity. Many words that are either homonyms or polysemous cause lexical ambiguity because they are words that have the same spelling but their meanings vary and may be dependent on the context that they are used in. Lexical ambiguity requires a translator to identify the expected meaning behind the words in order to choose the right word in the target language and translate correctly [1]. Inflectional morphemes, a source of lexical ambiguity, further confound the translation process. For example, the word number in the English language may be a noun, or it may be the inflected form of numb [1]. This makes selecting the correct translation meticulous work for a machine as the correct word choice is often dependent on the context. On the other hand, words that are not ambiguous in the source language may be open to more than one interpretation in its target language. This lack of one-to-one correspondence between words in different languages makes translation a complex, time-consuming, and costly task.

A different origin of complexity in translation is syntactic ambiguity, where the syntax of a sentence leads to more than one meaning. This type of ambiguity is particularly more complex to a computer because humans are able to pick up on the intended meaning through context, while computers have a difficult time discerning multiple possible meanings [1]. For instance, it is clear to humans that the sentence *"the stolen wallet was found by the fire hydrant"* means that the wallet was found next to the fire hydrant. However, a computer might interpret it as the fire hydrant finding the stolen wallet. These ambiguities, among many others, make it very difficult for a computer to represent the structure of a language in the form of rules.

Translation becomes even more complex when patent documentation is introduced into the mix, which involves patent conditions, correspondence with lawyers, and a unique writing style [73]. The translator requires deep knowledge of technical terms and a comprehensive understanding of the legal language to achieve high-quality translations. As powerful documents for encouraging innovation, patents require excellent language pair expertise. Compromising on the accuracy of the translation may create a case for fraudulence or potentially interrupt the patent filing process, which may lead to other consequences such as theft of inventions, expenses, and delays, among many others [74]. To overcome the aforementioned limitations, researchers have employed automation systems as tools for translation, known as machine translation models.

1.1 Context and Motivation

Machine translation (MT hereafter) has been an active and rapidly evolving technology in today's software engineering scene, even though the idea of using automatic translation predates the invention of computers by a few hundred years. Notable mathematicians and philosophers such as Leibniz and Descartes put forth the idea of using numerical codes as a universal language in the seventeenth century [1]. Though this idea has been around for a long time, the emergence of machine translation, in the modern sense, is said to have taken place in 1949 when Warren Weaver published a memorandum titled ”Translation” [2]. In the ”Translation” memorandum, Weaver formulated specific goals and methods that overcame the substantial limitations created by the straightforward method of word-to-word translation and presented a set of ideas for machine-aided translation based on the principle of information theory. In essence, the objective of MT is to translate text or speech from a source natural language (SL) to a target natural language (TL) through the use of a computer, and with or without the need for direct human intervention. Early research done in the field of computational linguistics steadily led to progress in linguistic and computational techniques, and eventually, the rule-based machine translation (RBMT) system was developed.

RBMT is a system that heavily depends on language theory since it is formed using a collection of grammatical and linguistic rules. In order to translate text using RBMT, extensive syntactic knowledge of both the SL and TL is required by a linguist so that they may define rules for the system to follow using information such as lexical, syntactic, semantic, and morphological congruities of both languages. Simply put, the process of translation involves applying predetermined rules to the SL in order to output a translation in the TL. RBMT comprises three different types of systems: direct, transfer, and interlingua systems [13]. The most rudimentary method of machine translation is direct MT. To translate text, direct translation models often rely on a vast collection of language pair-dependent rules that connect different grammatical and lexical events in the SL to their realizations in the TL [13]. As seen in Figure 1.1, the direct translation method requires a minimal structural analysis of input text (SL) that is needed for translation [64] and is akin to that of using a dictionary for word-to-word translation. Its directness, however, limits its ability to carry the nuances of the SL over to the translated TL as it does not consider a key component of the translation process: context. Indirect RBMT approaches, such as transfer and interlingua approaches, remedied this limitation to a degree.

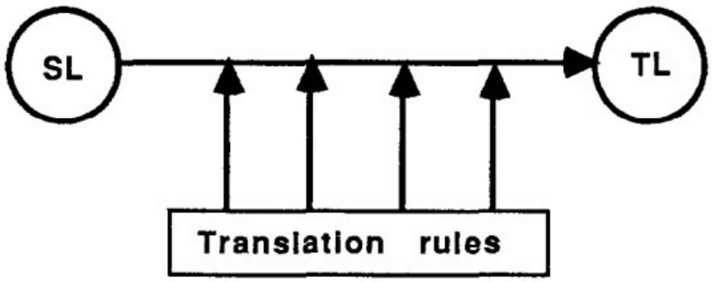


Figure 1.1: A schematization of the direct translation system [3]

For the former indirect RBMT approach, analysis, transfer, and synthesis are constituents of the process of transfer translation systems. This process entails the analysis of the SL into TL-independent representations [88, 89]. These representations are then transferred into syntactic structures dependent on the TL via translation rules and then synthesized to produce a translated output in the TL [88]. In interlingua translation systems, two main stages are involved. First, the SL is translated into an intermediate abstract representation independent of a natural language, also called an interlingua. This representation of the SL is then decoded and translated into the TL [90].

A comparison of Figures 1.3 and 1.4 with Figure 1.1 highlights the differences in complexity between the direct and indirect systems. To further understand the relationship between the three RBMT approaches, the Vauquois triangle illustrated in Figure 1.2 shows how the systems are related to one another as well as RBMT's evolution. The triangle, established by Bernard Vauquois in 1968 [4], displays the trade-off between the depth of analysis and the amount of transfer knowledge used by the three systems. The base of the triangle, where the direct method is found, requires minimal depth in analysis but the most amount of transfer knowledge. The

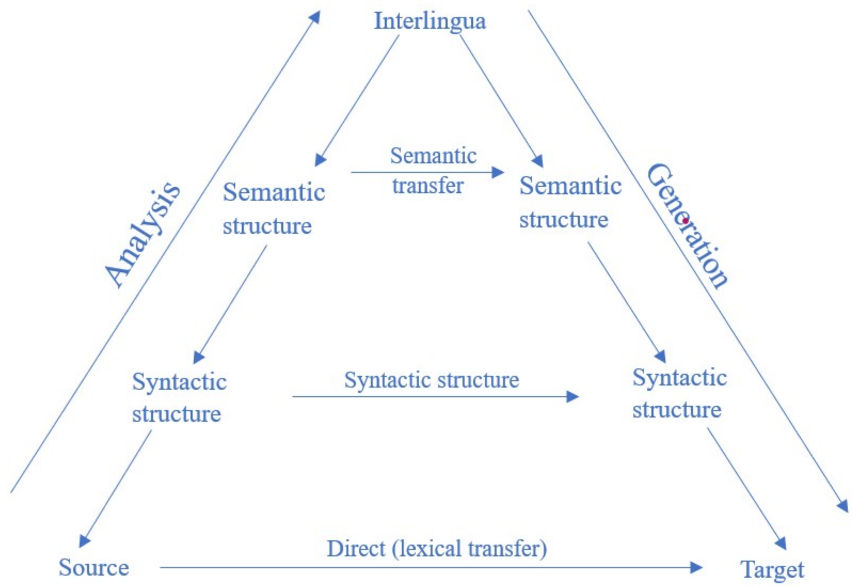


Figure 1.2: The Vauquois triangle adapted from [4]

depth of analysis increases moving up the triangle, and the amount of transfer knowledge decreases moving down the triangle. The interlingua system is found at the top with the most amount of transfer knowledge as well as the most depth of analysis. Though indirect systems are more advantageous than direct translation, they still lack accuracy in translated text, and they are more computationally expensive. The manual creation of rules that map syntax and grammar for both the SL and TL also requires a thorough knowledge of both languages. The limitations of both the direct and indirect methods called for a more robust system which led to the emergence of statistical machine translation (SMT).

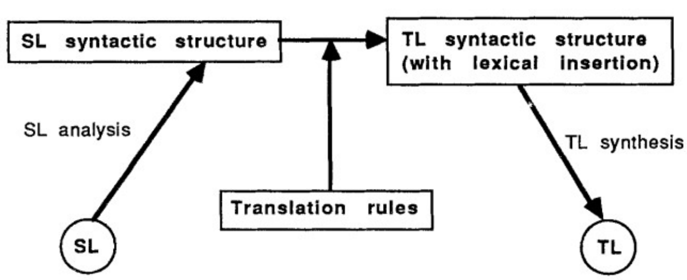


Figure 1.3: A schematization of the transfer translation system [3]

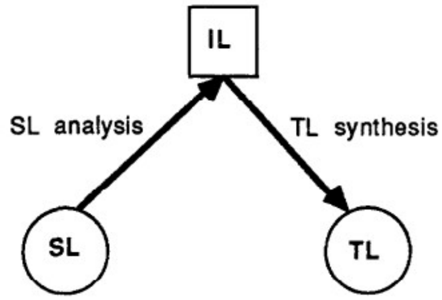


Figure 1.4: A schematization of the interlingua translation system [3]

As suggested by the name, statistical machine translation models are statistical models trained on large parallel text corpora that align the sentences or phrases of the SL and TL. The main idea behind SMT is to approximate the translation task. Utilizing probabilistic methods given a source input, SMTs search for the most likely translation. Statistically, modern SMT models are defined as a $y^* = argmaxyP(x|y)P(y)$ where y^* is the target text given some source text x with the objective of maximizing the two feature functions: the translation model $P(x|y)$ and the language model $P(y)$. The argmax operation searches for the target sentence with the greatest predicted probability [4]. This method is referred to as the source-channel approach and is regarded as one of the earliest SMT methods, however, it does not come without limitations. Because the source-channel approach depends on calculating the probability of two independent functions and then calculating the product between them, this method is computationally very expensive, especially with very large corpora. They are difficult to optimize for the same reason [92]. Additionally, the search process or decoding of the argmax is an NP-hard problem, and thus very complex [93]. Och and Ney [91] presented a framework of SMT that extends the baseline source-channel method and mitigates the limitations of the aforementioned method. Their framework uses a log-linear model and is given by the following equation:

$$y^* = argmax_y(\sum_i f_i(y, x)\lambda_i) \tag{1.1}$$

where $f_i(y, x)$ are the feature functions, and λ_i are the associated weights of the features. During training, the maximum likelihood estimation adjusts the weights of the features to maximize the probability of determining the correct translated target text given a source text [92] and uses log probabilities to represent the translation and language models [95]. The approach with the log-linear model improves the source-channel method by being more powerful and flexible as it permits the amalgamation of different features such as a translation and language function, lexical translation and alignment probabilities, as well as phrase tables [94]. Although the source-channel approach is still used for low-resource languages and noisy data, most modern SMT systems use the log-linear model because of its ability to incorporate multiple features for better translation results than the source-channel. However, by eliminating the limitations of the source-channel method, the log-linear SMT model raises its own problems. Namely, the incorporation of multiple features is considered to be a double-edged sword since each feature must be optimized individually before being combined together.

Possessing the potential for growth, machine translation moved from log-linear SMT models to neural machine translation (NMT) models. The basic idea behind NMT is to use end-to-end trained neural networks to encompass the entire process of machine translation without the need for feature engineering [15]. The similarities between SMT and NMT lie in the fact that they both utilize and rely on large corpora of sentences in the source language and their corresponding target sentences. Their similarities diverge with the implementation of continuous vector representations of linguistic units by NMT models. This differs from SMT models, which implement discrete symbolic representations [16] and rely on lexical and alignment units; a radical improvement over earlier MT techniques. A detailed description of the architecture and components of NMT models is provided in Sec. 2.1.

The world has grown increasingly connected and more technologically advanced, and this is reflected by the increase in patent applications in new and different technology spaces over the years. In 2020, the reported number of patents filed worldwide increased by 1.6% with 3.3 million patents filed, where approximately 85% of the total number of patents filed were accounted for by five national/international patent offices [28]. The National Intellectual Property Administration of the People's Republic of China (CNIPA) received upwards of 1.5 million applications, followed by the United States Patent and Trademark Office (USPTO), which received 597,172 applications. Ranked third, the Japan Patent Office (JPO) had 288,472 applications, the Korean Intellectual Property Office (KIPO) had 226,759, and finally, the European Patent Office (EPO) had 180,346 [28]. Figure 1.5 below demonstrates the growth of patent applications from 2006-2020 worldwide. MT of patents becomes an important problem as it is useful, in terms of industrial use, for countries to be able to file patents in foreign languages. The translation must also be as accurate as possible since even a small variation of the intended meaning may lead to legal loopholes that will be taken advantage of to exploit intellectual properties [29].

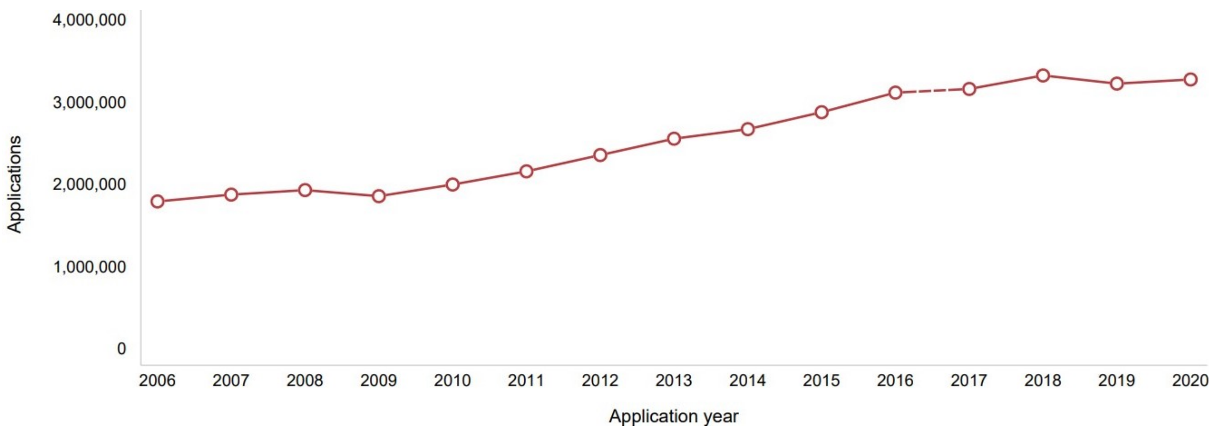


Figure 1.5: Number of Patent Applications Submitted Each Year From 2006-2020 [28]

Although commercially available translators like Google Translate, Microsoft Translator, and DeepL offer machine translation with a high level of accuracy for generic and nontechnical texts, more specialized domains rely on industry-specific training data to translate the text so that the translation may be relevant in their context [30]. While Google's Cloud Translation and Microsoft's Translator allow for custom translation and model training to counter this issue, the cost of using them risks getting expensive since more complex and large training data, as well as custom translations, lead to high costs, which may be prohibitive. Custom translations with

Microsoft are done using their C2-C4 instances where the cost of using those instances ranges from \$2,055 / month to \$45, 000 / month respectively [31]. This incurs a large expense, especially for companies with many patent grants. For instance, IBM, which ranked as 2021's most innovative company in the US IP space, had $9,130$ and $8,682$ patent grants in 2020 and 2021, respectively [32]. Since IBM offers a range of products and patents corresponding to different industries, the use cases would require multiple model training, which further increases the cost.

Given the difficulty of MT and the high cost associated with it, continued research in the field of MT is extremely important, and therefore, the motivation of this research is to demonstrate the enhancement of recent pre-trained NMT models using the application of deep learning and computational linguistics for the purpose of patent translation from English to Japanese.

Why English-Japanese?

In general, NMT systems are trained on parallel corpora consisting of tens to hundreds of millions of sentences, and a dataset of this enormity is most commonly accessible for only a few highly resourced language pairs such as English paired with some European languages, Arabic, and Chinese [124]. Though the English-Japanese pairing is not considered to be a low-resource language, to the best of our knowledge — only one English-Japanese parallel corpus has been created for the purpose of patent translation by the Japanese Patent Office for the Workshop on Asian Translation (WAT). The training corpus is made up of one million parallel sentences from Japanese patent descriptions which were sourced from only four out of eight of the IPC sections [121]. Though sufficiently large, the corpus risks overfitting and catastrophic forgetting due to domain mismatch.

Additionally, the motivation for selecting the English-Japanese (en-ja) pair specifically was a direct result of the business objectives of our industry partner Xlscout, and the need for an efficient patent translator within the industry, as shown in our discussion.

1.2 Objective and Contribution

As discussed by Saunders in [80], in situations where the domain is known and the training data relates closely to the test set, such as for the WMT shared tasks [119], sentences belonging to the known domain will be best translated by the NMT system as the system has been adapted to that domain. Knowing the domain is not a common scenario, however, and it is especially not common for patent translation. Scenarios, where a corpus does contain known-domain labels, may also prove to be unhelpful as the labels may not necessarily be indicative of the entire text. For example, this thesis may be filed and labelled as a document belonging to the domain of computer engineering, but this broad domain could encompass machine learning and network security, consequently leaving the exact domain that the thesis belongs to vague to an NMT system.

The objective of the work presented in this thesis is to enhance the performance of neural machine translation systems for the task of patent document translation using domain adaptation. Specifically, we work to improve the results obtained from translating patent documents of highly technical domains, from English to Japanese. We collect and generate a bilingual parallel corpus using $120,000$ patents across 8 technical domains from an IPC database. We then fine-tune and adapt 3 pre-trained transformer models, such as MarianMT, and evaluate their results with the aim of increasing the degree of similarity between the SL and TL sentences from patent documents. Adapting pre-existing models also allows us to avoid reinventing the wheel, and lets us stand on the shoulders of past NMT researchers [57, 6, 21, 8] to gain integral insights for the task of translating complex language structures. The specific interest in using the English-Japanese language pair for translation is motivated by the business objectives of our industry partner, XLScout.

1.3 Thesis Structure

To provide a comprehensive exploration of the thesis topic, the structure of the thesis from this point onward will be as follows:

- Chapter 2 provides background information on NMT systems as well as a literature review of the different applications of machine translation. We also cover the nature of patent translation from English to Japanese, as well as the challenges that may arise from the en-ja language pair.
- Chapter 3 covers the methodology of implementing a domain adaptation technique. We also discuss the architecture of our baseline models, including hyperparameters, and explain how our research constitutes a domain adaptation problem.
- In Chapter 4, we explain the details of our experimental setup and discuss the results.
- Chapter 5 concludes the work of this research by summarizing key points and discussing possible future directions, including some limitations.

CHAPTER 2

Neural Machine Translation: A Review

This chapter offers an in-depth review of the literature pertaining to recent developments in NMT models and explores existing studies that have tackled similar challenges in various domains using domain adaptation techniques. By examining these studies, our objective is to identify state-of-the-art techniques, comprehend their limitations, and adapt them to the domain of patent translation. By doing so, we aim to contribute to the advancement of NMT models in translating patent documents, thereby facilitating effective communication within the realm of intellectual property.

The chapter begins with an overview of NMT (Sec. 2.1) and its applications (Sec. 2.2). Subsequently, it explores the challenges associated with patent translation and the need for domain adaptation techniques to improve translation quality (Sec. 2.3). Additionally, the literature review covers studies that have addressed similar problems in different domains and highlights their contributions and limitations (Sec. 2.4).

2.1 Neural machine translation techniques

While many neural language models have emerged for the task of machine translation, the encoder-decoder model is considered to be the foundational model used for most NMT systems [81]. As such, the NMT models used for our research all comprise the encoder-decoder baseline architecture. The following paragraphs will focus on discussing its architecture.

In essence, an NMT model with the encoder-decoder architecture accepts a source text input x through the encoder, encodes it as a vector known as a context vector, and passes it to the decoder which decodes the vector to output a translated text y. The most widely used method for encoders and decoders that utilizes two recurrent neural networks (RNNs) was first proposed by Cho et al [7]. Since then, many variations in the choice of neural network used within the architecture have emerged and will be briefly reviewed in Sec. 2.1.1. We first explain the architecture of the encoder-decoder framework statistically.

Within the framework, the encoder sequentially reads a fixed-length input $x = (x_1, .., x_n)$ and calculates the hidden state such that

$$h_{(t)} = f(h_{(t-1)}, x_t), \tag{2.1}$$

where $h(t)$ is the hidden state at each time step t and f is a nonlinear activation function such as a logistic sigmoid function or a long short-term memory (LSTM) network which was first used by Sutskever *et al.* [8]. Once the entire sequence has been read, the encoder generates a context vector c that holds the hidden states of the entire input sentence. The decoder, another RNN, predicts the symbol y_t given $h_{(t)}$. In the decoder, the hidden state is given by Equation 2.2,

$$h_{(t)} = f(h_{(t-1)}, y_{t-1}, c). \tag{2.2}$$

The conditional probabilities of both RNNs are given by Equation 2.3,

$$P(y_t|y_1, .., y_{t-1}, c) = g(h_{(t)}, y_{t-1}, c), \tag{2.3}$$

where g is another activation function. Together, the encoder and decoder are trained so that they maximize the conditional log-likelihood. Bahdanau *et al.* [6] proposed a novel architecture that extended and improved the basic architecture so that the encoder was made up of a bidirectional RNN and the decoder that (soft-)searched through the source sentence during the decoding process. They argued that using a fixed-length input vector serves as an obstacle and showed that relieving the encoder of encoding the entire information within the source sentence into a vector of fixed length improves the model. To achieve this, the authors defined the conditional probability of the RNNs so that the probability is conditioned on c_t for each target output y_t. The equation is as follows:

$$P(y_t|y_1, .., y_{t-1}, x) = g(y_{t-1}, s_t, c_t), \tag{2.4}$$

where s_t is the hidden state given by

$$s_t = f(s_{t-1}, y_{t-1}, c_t). \tag{2.5}$$

The context vector here is dependent on annotations $(h_1, .., h_t)$ that contain information on the source sentence, focusing on the t-th word and its surroundings within the source sentence. The context vector c_t is generated using the weighted sums of h_t. The weights of each annotation are computed using an alignment model which calculates how well the input and output in the ith and jth position match. This provides the decoder with an attention mechanism to allow it to focus on parts of the source sentence deemed important and thus removing the need for a fixed-length vector [6].

2.1.1 Modelling variants of the encoder-decoder framework

As a collection of algorithms inspired by the function and structure of the human brain with the ability to identify patterns and make predictions among other uses, neural networks (NN) have emerged as powerful tools for modeling conditional probability distributions with more than one input. NNs have traditionally been divided into three distinct classifications: recurrent neural networks, long short-term memory neural networks, and feed-forward neural networks.

Recurrent neural network

In a recurrent neural network (RNN), each node receives both the output and the input values of the previous layer's node, i.e. the network's neurons send feedback to each other. Feedback in RNNs is referred to as recurrent loops over time. Not only does this overcome the limitations of FFNNs and allow the network to process inputs of arbitrary lengths but it is able to work with sequential data such as time series data [96].

The architecture of a simple RNN is made up of an input layer, as well as recurrent hidden and output layers. The hidden layer consists of mathematical equations, which can be referred back to in Sec 1.1, and a set of values that compiles all required information about the network's previous layers over a number of timesteps. The collection of integrated information helps the network to learn and output precise future behavior of the network [97]. Similar to FFNNs, the hidden layers of RNNs contain an activation function. The graphs of common activation functions can be seen in Figure 2.1 [98]. Linear functions may also be used in place of nonlinear activation functions, however, a polynomial of degree one is very simple to solve and lacks the complexity required to learn the complex mappings from the network [99].

The capacity of the RNN to represent intricate temporal and sequential dependencies within the input data is dependent on the activation function selected. Widely used functions such as the rectified linear unit function, $ReLU$, capture nonlinear interactions between the input values and their relationship over a period of time, allowing the network to model complex relationships, and the use of each function depends on the type of problem at hand.

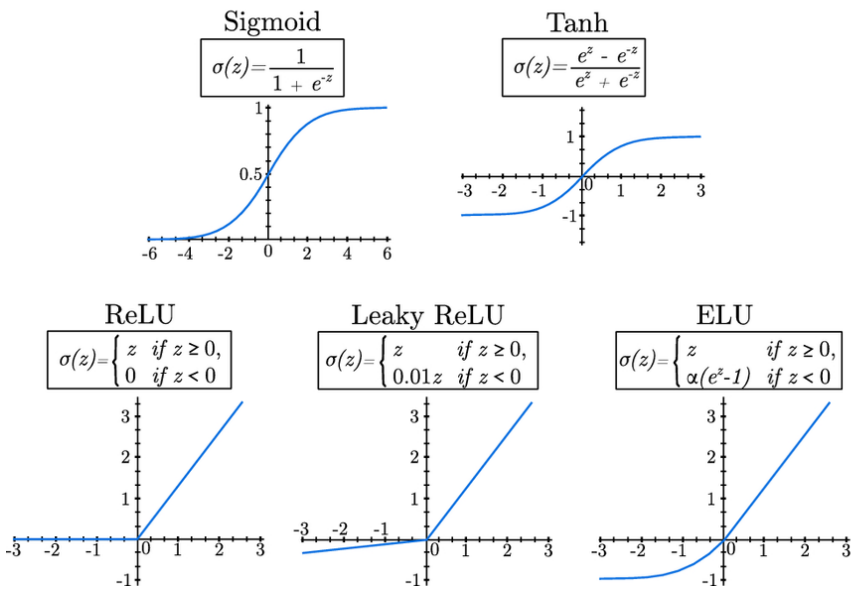
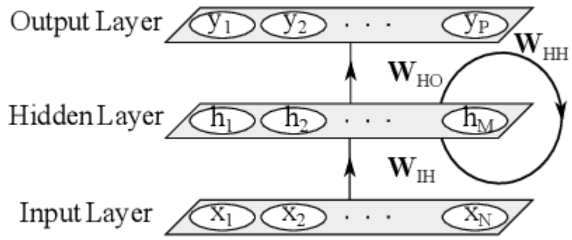


Figure 2.1: Activation functions and their graphs [98]

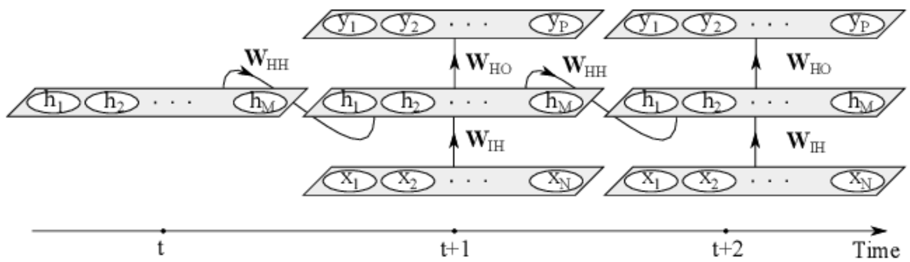
For instance, the range of the sigmoid function, over its entire domain, is between 0 and 1 so it may be selected where the output of the network must be within that range like in binary classification problems. The disadvantage of the sigmoid function is that it is susceptible to the vanishing gradient problem, as is the tanh function [99]. The ReLU function thresholds the range at 0 and thus is computationally cheap. Additionally, the rate at which the stochastic gradient descent (SGD) converges is faster than it is for sigmoid and tanh [96]. Figure 2.2 illustrates an overview of the RNN architecture (2.2a) and an unfolded structure over time of the same RNN (2.2b).

Long short-term memory neural networks

Feedback is also very prevalent in the brain [96]. In generality, its functionality may be able to include some of the brain's dynamics, such as retaining memory of past inputs [100]. Although recurrent connections allow a network to understand complex temporal dependencies, the algorithms used in the training of the network may profoundly induce a threshold for the memory produced by the recurrent connections. All RNN models are susceptible to both vanishing and exploding gradients and thus, they are only capable of learning short-term temporal



(a) The architecture of an RNN (folded)



(b) The architecture of an RNN (unfolded)

Figure 2.2: RNN Architecture with a visualization of recurrent cycles [96]

dependencies [96]. For the network to retain long-term temporal dependencies within the data, the architecture of the hidden units in the network is changed from activation functions to memory cells. Within the memory cells are gates that control the flow of input and output data to neurons in the hidden layers. The gates determine what information should be retained by the network and what information should be forgotten [101]. Resultantly, the consequence of vanishing and exploding gradients is lessened. Networks that implement the use of memory cells are called long short-term memory, or LSTM networks and they are among the most efficient for the use of long-term dependencies.

The architecture of an LSTM cell is composed of input gates, forget gates, output gates, and a cell activation element. If the input and output gates are closed, the contents of the memory cell stay the same between time steps [102]. Mathematically, the components that make up the LSTM cell can be defined by the equations given in Table 2.1.1 below:

Table 2.1: LSTM units [102]

	Components of LSTM cell	Equation *
(a)	Block input	$z^t = g(W_z x^t + R_z y^{t-1} + b_z)$
(b)	Input gate	$i^t = \sigma(W_i x^t + R_i y^{t-1} + p_i \odot ct - 1 + b_i)$
(c)	Forget gate	$f^t = \sigma(W_f x^t + R_f y^{t-1} + p_f \odot ct - 1 + b_f)$
(d)	Cell state	$c^t = i^t \odot z^t + f^t \odot c^{t-1}$
(e)	Output gate	$o^t = \sigma(W_o x^t + R_o y^{t-1} + p_o \odot ct + b_o)$
(f)	Block gate	$y^t = o^t \odot h(c^t)$

* where, x^t is an input vector at time t
W is a rectangular input weight matrix
R is a square recurrent weight matrix
p is a peephole weight vector
b is a bias vector

Feed-forward neural networks

Feed-forward neural networks (FFNN) get their name from the fact that information flows in one direction with no feedback from the outputs to the inputs [103]. Structurally, they are made up of three components: the input layer, the hidden layer(s), and the output layer. Unlike the latter two layers, the input layer does not perform any computations. Instead, it is responsible for receiving input data in the form of a vector or matrix. The length of the input vector for a typical FFNN must be a specified value because the input layer is made up of a fixed number of nodes or neurons. In other words, the total number of input values will be determined by the size of the FFNN's input layer. This serves as a limitation in the network's performance, especially for tasks like natural language processing [103].

Connected to the input layer is the hidden layer. Depending on the number of hidden layers in the structure, FFNNs are referred to as single-layer FFNNs if there is only an input and output layer, and as multi-layer FFNNs if there are one or more hidden layers. Composed of a predetermined number of nodes, the task of the hidden layers is to receive input from the nodes of the previous layer, and compute the weighted sums of the inputs that are then passed to a nonlinear activation function. Finally, the resulting output from each node is passed as input to the nodes of the next layer. The nodes of the last hidden layer transfer the outputs of the layer to the output layer which produces the final output [103]. The activation function, a non-linear mathematical function, applies its nonlinearity in order to model complex relationships between the inputs and the outputs. Several activations are commonly used within the neural network including the sigmoid function, rectified linear unit (ReLU), the tanh function, and softmax [104]. Figure 2.3 below illustrates a simple fully connected multi-layer FFNN.

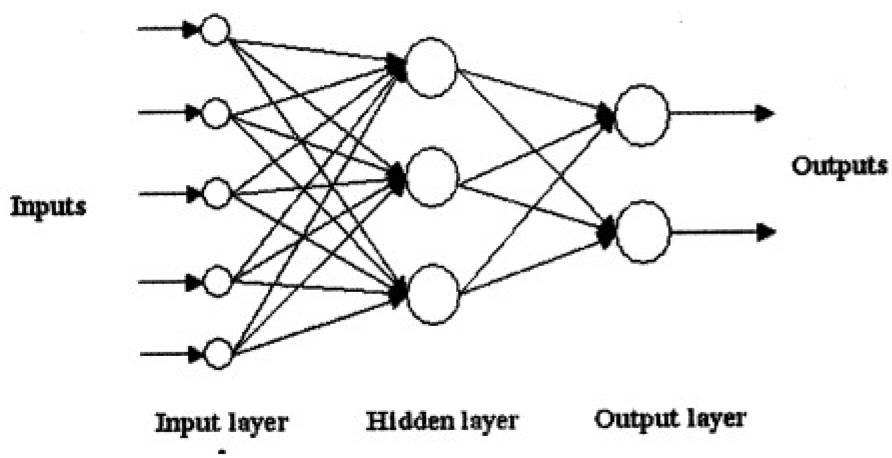


Figure 2.3: Architecture of a multi-layer FFNN [103]

2.1.2 Attention mechanism

Attention is an important mechanism in NMT systems as it is able to dynamically highlight the most pertinent parts of a given input. The basic idea is that the NMT model ”pays attention” to relevant parts of the data to improve performance. For the task of translation, the attention mechanism allows the NMT system to focus its attention on specific words within the input text in order to understand the context and derive the meaning of the text [20]. Generally, each word is mapped to a weight by the mechanism where the value of the weight is determined by its relevancy within context. The more significant the word, the higher the value. In this section, we cover a brief overview of the different types of attention mechanisms, including dot product attention, scaled dot product attention, additive attention, multi-head attention, and self-attention.

Dot product attention

Dot product attention was first introduced by Luong *et al.* in [21] and is used in NNs so that the network can learn to selectively pay attention to different parts of the input sequence with the goal of improving the network's performance. The dot product attention mechanism comprises three main steps. First, the queries, keys, and values are computed and transformed into three distinct vectors; Q, K, and V respectively. Once the vectors have been calculated, the dot product between the query vector Q and the key vector K is passed as input to a softmax function to obtain a probability distribution over all the values called the attention scores. Next, the mechanism computes the weighted sum of the value vector V to generate the final output, where each weight is given by the attention scores calculated in the previous step [21].

Scaled dot product attention

Vaswani *et al.* introduced their attention mechanism in "Attention Is All You Need" [57] called the scaled dot product attention, the architecture of which can be seen in Figure 2.4. The authors used this particular mechanism within the Transformer model, which will be discussed in the following section, to enable the model to pay attention to various input sequence segments during the encoding and decoding process. The authors found that the efficiency and dependability of the attention mechanism during training improved by scaling the dot product by the square root of the query vector, $\sqrt{d_k}$, where d_k is the dimension of the keys. The dimensions of values is denoted by d_v. Within the encoder-decoder architecture, the queries were obtained from the previous layer of the decoder while the keys and values came from the output of the encoder [57]. Formally, they define the mechanism using the following mathematical equation:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{2.6}$$

where Q is a matrix of queries, K is a matrix of keys, and V is a matrix of values [57].

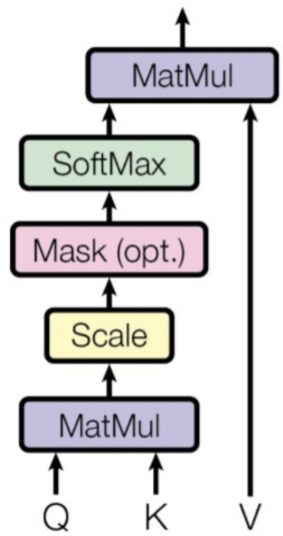


Figure 2.4: The scaled dot product attention mechanism [57]

Additive attention

Building upon the work of Cho *et al.* [7] and Sutskever *et al.* [8], the authors of [6] implemented the FFNN encoder-decoder framework that allowed the system to receive a source sequence of variable length unlike the earlier models which could only accept an input of fixed length. The authors hypothesized that fixed-length input vectors serve as limitations and cause the performance of the encoder-decoder model to deteriorate at an increasing rate proportionate to the increasing length of the source sequence. To prove this, they replaced the fixed-length input vector with one that varied in length.

As discussed previously and in more detail in Section 1.1, the additive attention mechanism calculates a context vector that is able to retain the most important parts of a source sequence for the decoder to pay attention to. This allows the model to align the source and target sequences better, improving on the quality of translation and allowing more flexibility in choosing an input length. The context vector is a result of the weighted sums of the hidden units within the encoder. The weight applied to each node of the hidden layers is determined by the softmax function which outputs a range between 0 and 1. Figure 2.5 provides a graphical illustration of the encoder-decoder model with attention.

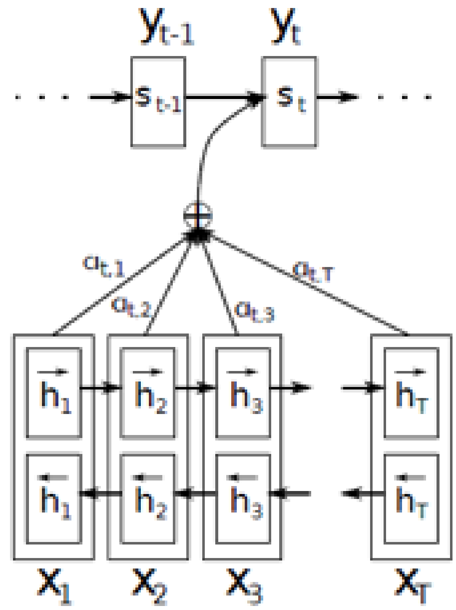


Figure 2.5: Architecture of the encoder-decoder NMT model with attention mechanism [6]

Multi-head attention

Another attention mechanism proposed by Vaswani *et al.* in [57] is the multi-head attention mechanism, a variant of the scaled dot product attention and a powerful tool for understanding complex relationships between the different segmented source sequences. The mechanism repeatedly projects the queries, keys, and values, h times instead of using a single attention mechanism in the network. Each of the projected queries, keys, and values are applied to the mechanism in parallel first which outputs values that are then concatenated. The final values are obtained once the concatenated values have been projected in parallel a second time. An overview of its functionality is illustrated in Figure 2.6.

With multi-head attention, the model can learn in parallel and jointly focus on data from different parts of the input sequence.

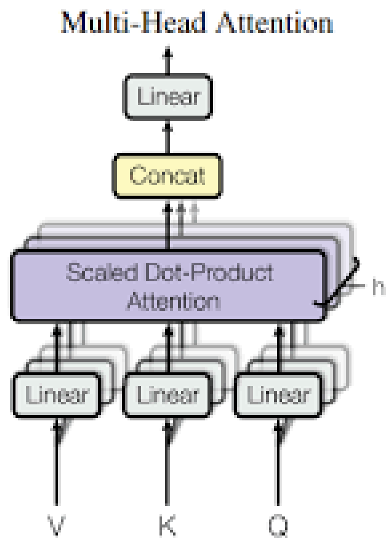


Figure 2.6: Architecture of the encoder-decoder NMT model with multi-head attention mechanism [6]

Self-attention

Put simply, self-attention refers to an attention mechanism that models the relationship of the different positions of the source sequence with the aim of calculating and outputting a target representation of the input. Although defined very similarly to the other attention mechanism discussed, the main difference found in self-attention is the source of the attention scores. In contrast to [6] and [21], the computation of attention scores does not depend on external context but on the inputs. Additionally, traditional attention mechanisms are only able to pay attention to one position of the input sequence at a time. Self-attention on the other hand is able to "multi-task" and focus on multiple positions at once, allowing it to capture more complex relationships between the different segments of the input sequence. This allows the mechanism to model long-term dependencies. Lin *et al.*, authors of [41] show that the introduction of self-attention allows the last sentence representation to access previous hidden states meaning that the model is not required to carry all the information to the last hidden layer. Applied to an LSTM model, the authors conclude that self-attention is capable of successfully encoding a variable-length input vector into a fixed-length representation without facing the consequences of long-term temporal dependencies [41].

Vaswani *et al.* [57] use a self-attention mechanism in conjunction with multi-head attention to provide a transduction model, based completely on attention. Since its inception, the Transformer model has become the conventional standard for NMT purposes. Following this trend, the NMT models used for the experiments of this thesis project also implement the Transformer architecture, the details of which will be covered in Sec. 3.1.1.

2.2 Translation and domain mismatch

2.2.1 Challenges of English-Japanese translation

One source of distinctions other than the obvious difference in characters of both languages is the word order in English and Japanese. In English, the word order is in the form of Subject-Verb-Object (SVO) while the word order in Japanese is SOV [23]. This means that topics are at times expressed in entirely different sentence structures and thus there is a lack of correspondence between the structures of sentences that poses a challenge for the machine to translate accurately [24]. Additionally, being a high-context language, it is common in Japanese for much of the information to be implied [25]. For example, the subject is usually dropped in situations where the context is clear [26]. If a person were to say "I am going to bake a cake" in English, the direct Japanese translation would be "bake a cake" since it is understood by the listener that the speaker is the subject and thus "I" is implied. In contrast, English, as a low-context language, communicates content explicitly and writing is understood very literally. This demonstrates that there is more to translation than ensuring the machine translator chooses the correct words; the cultural context must also be taken into account to make sure that the intended meaning is delivered correctly [27]. This could prove an unfeasible task for computers; being a cultural mediator requires a certain level of refined reasoning where one must be able to deduce what meanings might be extracted by a reader so that the translation may be adjusted as needed [27]. Without the explicit specification of a large number of language and cultural context-based rules, this task is difficult for machines. And if such intricate rules, which may be in the tens of thousands, were to be constructed and documented with the system, the effort expended may be prohibitive and contradictory to the purpose of utilizing machine learning to automate the MT task altogether. A system that overcomes the need to explicitly set language-dependent or culture-dependent rules is the overall goal.

The problem of context sensitivity is more or less avoided when considering the source and target text that is shared by readers with the same background knowledge [27], for example, readers may be part of the same scientific discipline or industry. The premise of this work will focus on the translation of scientific patents from English to Japanese using machine translation. Although the complexity of context sensitivity is reduced, we are faced with another challenge: patents of a scientific nature contain many technical words that are domain-dependent and may be homonyms with more than one interpretation depending on the subject. Instead of running into the problem of deciphering cultural context, the machine must determine the correct translation of a word depending on the domain it is used in. For example, "arm" is a homonym for both the biological human arm as well as a robotic arm in English. In Japanese, it becomes sensitive to the domain it is used in; "腕 " refers to a human arm while "アーム" refers to a mechanical arm.

2.2.2 Patent translation

A patent is a form of intellectual property that confers the patent owner the legal right to produce, sell, exchange, or give an invention away, for a certain period of time. An approved patent application grants an inventor exclusive ownership over their invention(s) and the legal right to exclude others from selling or remaking their product(s) [105]. In patent law, prior art is generally defined as any information that is publicly disclosed including granted and published patent applications, research publications, or product descriptions [120]. Secondarily, the prior art encompasses all patents around the globe as well as public disclosures that are relevant industrially, and thus includes content in any language [120]. This makes the contents and details of a patent extremely important as mistakes and insufficient information may lead to legal disputes and the potential loss of the invention.

The aforementioned point is one, among many, of the reasons why the task of patent translation, including research conducted in patent MT, tends to be very meticulous. In addition, patent translation is also complex and challenging because the combination and use of legal and scientific terminology make patents highly technical in language. In [106], Rossi and Higgins discuss the syntactic complexities that are a product of the technical, legal, and linguistic rules. Specifically, they point out that patent sentences ”make massive use of nominal style, relative clauses, formal constructions, and huge long-distance relationships among constituents” (Rossi & Higgins, 2013). Additionally, the authors argue that from the perspective of MT, patents have a tendency to encompass convoluted and inconsistent text due to the freedom given to patent drafters in writing the description of the invention, idiosyncratic legal writing styles, and the lack of usage of authoring tools that are specific to patents [106].

To compound the challenge of patent translation, inconsistencies are also present within the same domains and this is in part due to the use of nonstandard terminologies by investors to disconnect the invention from prior art [107], in hopes of increasing the patentability of their inventions. And as chain reactions go, this may lead to false similarities between patents based on style, not content [107]. For NMT models that are heavily reliant on the domain of the training dataset that they are trained on.

For instance, in [108], Hiroyuki Kaji studies domain dependence, specifically the dependence of translations of nouns on its domain, in English-Japanese patent translation. To do this, Kaji essentially first calculates the ratios of the number of associated words that suggest each translation to the total number of associated words and uses each ratio to rank the translations of the target word [108]. Table 2.2 by Kaji, demonstrates the results obtained from experiments ran, details of which can be found in [108]. These findings show that translations of five target words differ between subdomains and that a target word may only have one translation in some subdomains [108].

Target word	Translation*	*RAW*† (%)								
		ALL	A	B	C	D	E	F	G	H
administration	管理 (management, control)	*50.7*	7.8	*100*	14.3	-	-	-	*73.9*	*96.2*
	行政 (government)	-	-	-	-	-	-	-	4.3	-
	局 (government, department)	-	-	-	-	-	-	-	-	3.8
	経営 (management of an organization)	3.1	-	-	-	-	-	-	8.4	-
	運営 (operation)	-	-	-	-	-	-	-	9.7	-
	掌 (conducting, management)	-	3.6	-	-	-	-	-	-	-
	投与 (giving medication)	39.7	*88.6*	-	*85.7*	-	-	-	2.7	-
column	柱 (pillar)	*62.5*	12.4	-	-	-	*94.3*	*52.6*	3.7	16.4
	支柱 (prop, support)	6.2	*62.2*	9.8	-	*100*	5.7	29.3	-	5.0
	円柱 (cylinder)	-	-	-	-	-	-	11.1	-	-
	列 (line, array)	17.5	22.1	-	3.0	-	-	2.6	*67.8*	*64.5*
	ライン (line)	3.8	2.7	15.8	*97.0*	-	-	4.0	4.4	-
	コラム (newspaper column)	4.5	-	*60.3*	-	-	-	-	9.4	-
	欄 (section, blank)	3.4	-	8.4	-	-	-	-	13.1	9.7
culture	培養 (growing of bacteria)	*70.9*	16.4	-	*100*	-	-	-	-	-
	栽培 (growing of plants)	22.4	*76.9*	-	-	-	-	-	-	-
	養殖 (raising of animals)	5.4	6.7	-	-	-	-	-	-	-
	訓練 (training)	-	-	-	-	-	-	*100*	*63.1*	-
	教育 (education)	-	-	-	-	-	-	-	36.9	
nail	釘 (fastener)	*79.3*	*79.2*	22.8	-	-	*96.6*	*92.4*	-	-
	爪 (body structure)	20.7	20.8	*77.2*	-	-	3.4	7.6	*100*	*100*
plant	植物 (flora)	*46.5*	*88.3*	*31.8*	*56.8*	-	*67.2*	-	-	-
	植木 (garden plant)	-	5.0	-	-	-	-	-	-	-
	プラント (industrial plant)	21.1	-	31.1	2.8	-	-	*85.7*	*81.6*	21.5
	装置 (instrument, device)	-	-	-	22.5	*87.9*	12.4	5.5	3.0	*46.2*
	工場 (factory, works)	-	-	8.0	5.3	-	-	-	-	-
	設備 (apparatus, facilities)	26.4	2.7	28.8	12.6	12.1	14.4	8.9	9.4	29.9
	建物 (building)	-	-	-	-	-	5.5	-	-	-

* English translations other than the target word are given in parentheses.
† Italicized *RAW* values indicate the most major translations; a hyphen (-) means that *RAW* is less than 2.5%.

Table 2.2: Translations of 5 target words from the whole IPC* domain and 8 subdomains [108]

When the in-domain training data is unavailable or inaccurate, the model is trained using corpora unrelated to the target domain, which lowers the model's NMT performance. A domain shift problem is caused by the mismatch between the domains of the training and test sets of data [109]. Domain adaptation techniques have been introduced to address the difficulty of the domain shift/mismatch issue in NMT.

Before we begin the discussion on domain adaptation methods, we first discuss what constitutes a *domain* in the context of machine translation. Summarized in [110] by Pham *et al.*, a domain d is statistically defined as a distribution, given by $D_d(x)$, over a feature space H, where H ranges over all possible domains. In NMT, H is defined as the representation space for source sentences. Each domain in H is associated with a specific data source, and it differs from other sources of data in terms of ”textual genre and thematic content” [110]. For the purpose of our research, our feature space H is the International Patent Classification (IPC) system and each of the eight main sections of the IPC is referred to as a domain. The IPC is further subdivided into sections, classes, subclasses, groups, and subgroups [78], however, we do not define any ‘subdomains’. A detailed summary of the IPC system is discussed in Sec. 3.3 along with the details of our data collection process.

To briefly reiterate, the performance of MT models can significantly degrade when applied to a target domain that differs from the domain on which the model was trained. This domain shift often leads to a distribution mismatch between the training data and the target domain, resulting in suboptimal translations. Domain adaptation (DA) techniques in machine translation address this challenge by adapting or fine-tuning the MT model to improve its performance in the target domain.

2.2.3 Domain adaptation techniques

Several techniques leverage domain adaptation principles to align the source and target domains. Within the NLP context, the techniques are categorized into model-centric, data-centric, and hybrid approaches. Figure 2.7, taken from [111], provides an illustrative overview of each

DA approach.

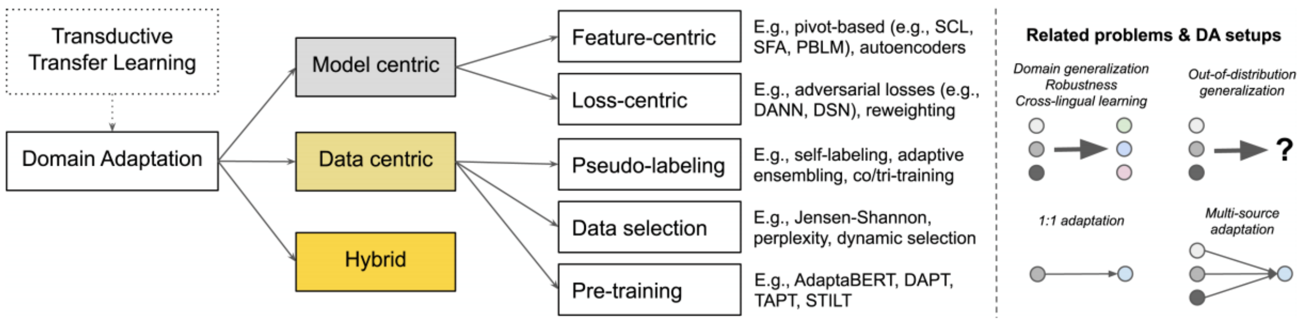


Figure 2.7: Classification of the DA approaches used in NLP cases (left). Related problems and DA setups (right) [111]

Model-centric

This approach redesigns components of the model and is divided into two methods; feature-based and loss-based methods. By learning a transformation that extracts an invariant feature representation across domains, feature-based methods try to map the source data into the target data. Typically, they transform the original features into a new feature space, then, using an optimization procedure, minimize the gap between domains in the new feature representation space while maintaining the underlying structure of the original data [112]. For example, in [113], Kouw *et al.* proposed a feature-level domain adaptation (FLDA) approach that fits a probabilistic model. In order to capture the transfer between the source and target data, the method trains a domain-adapted classifier by minimizing the expected loss on the source data under the transfer model [113]. Following their experiments, the authors concluded that the results obtained using the FLDA approach are comparable to that of advanced DA approaches [113]. Additionally, this approach proves to be advantageous in some scenarios as FL transfer models are very competent at modeling variances in the marginal distribution of words between the source and target domain, for example, when training an NLP model on news articles for the purpose of applying it to Twitter data [113]. The latter type of model-centric DA method, the loss-based method is used to adapt the model by changing the loss function in particular ways. In general, to make accurate predictions on text from the target domain, model-centric methods alter the feature space, loss function, and model architecture and parameters [111].

Data-centric

Hinted within its name, the data-centric approach strategically utilizes certain components of data rather than adjusting the model itself. Based on which aspect of the data would benefit the NLP model, this approach is divided into three main methods. These are *pseudo-labeling*, which refers to the process of using the labeled data to predict labels for the unlabelled data, otherwise known as pseudo labels [114], *data selection*, *data selection*, and *pre-training* [111]. In [115], the authors, Wotherspoon *et al.*, use the process of data selection as a critical step for the DA of automatic speech recognition systems (ACR). ACR systems are extremely sensitive to the domain mismatch problem and thus require ample amounts of training data that is transcribed. However, due to costly transcriptions, finding a large dataset of transcribed in-domain data becomes an impractical task. To overcome this challenge, the authors take advantage of the readily available nature of untranscribed out-of-domain audio and implement data selection methods to train a recognizer model on the out-of-domain labeled data in order to more accurately transcribe audio from the target domain. [115]. Their methodology was successfully able to achieve up to 57% improvements over the baseline model. The data selection technique specifically, outperformed, if not matched, ”word-level confidence selection across six separate domain shift conditions” (Wotherspoon, 2021). Finally, a hybrid DA approach includes methods that are a permutation of the model and data-centric methods.

The relevance of domain adaptation (DA) in patent translation is crucial due to the unique characteristics and specialized language used in patent documents. Patent translation requires not only accurate translation of technical terms but also capturing the legal, scientific, and precise nuances specific to patent language, and thus domain adaptation techniques play a significant role in addressing the challenges associated with patent translation and improving the quality and effectiveness of the translation process. DA is especially useful for the purpose of using NMT models to translate patent documents containing legal and formal language elements, such as claims, legal provisions, and precise descriptions. These aspects require adherence to specific linguistic conventions and accuracy in translation, thus DA helps the system comprehend and reproduce the legal and formal language features inherent in patent documents. By training the model on patent-specific data, it can better handle the formal language requirements and ensure the translations maintain legal and technical accuracy. For these reasons, the central method that motivates our experimental setup in this thesis is the data-centric, specifically the data selection technique. To the best of our knowledge, this area of research has been relatively unexplored for the problem of English-Japanese patent translation. A thorough discussion of our methodology and the experimental setup is conducted in Chapters 3 and 4. In the next section, we review recent efforts in MT.

2.3 Recent efforts in NMT

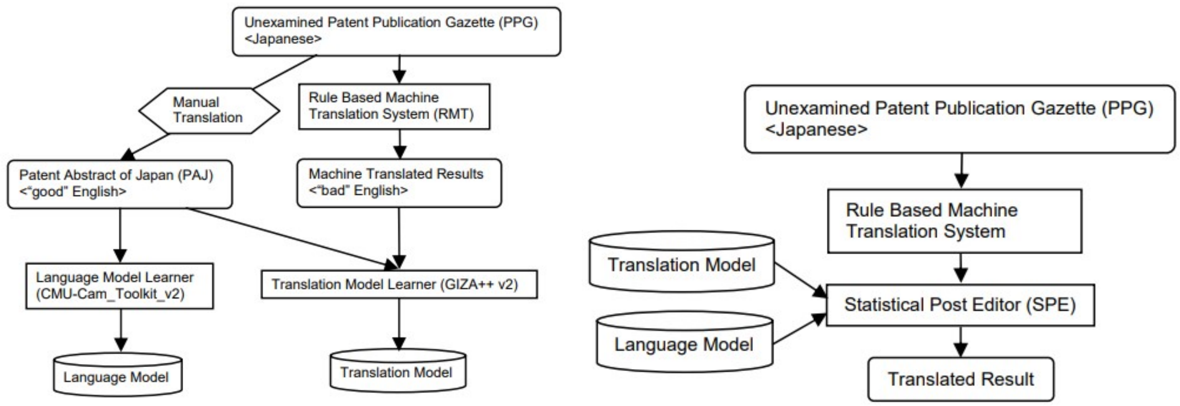
In [51], the authors propose to extend the use of a rule-based technique to simplify sentences before using an RBMT system to translate those sentences from English to Tamil. Complex sentences are made simpler using connectives such as relative pronouns, coordinating conjunctions, and subordinating conjunctions [51]. Table 2.3 below shows the words that were used by the system as connectives.

Table 2.3: Connectives used to simplify sentences in [51]

Relative pronouns	Who, which, whose, whom
Coordinating conjunction	For, and, not, but, or, yet, so
Subordinating conjunction	After, although, because, before, if, since, that, though, unless, where, wherever, when, whenever, whereas, while, why

Additionally, delimiters such as '.' and '?' are used to divide long and complex sentences into sub-sentences where the meaning of the sentence remains the same [51]. The authors chose to ignore the ‘,’ delimiter. The authors lay out the framework as follows: first, the initial splitting of the sentences from paragraphs is done using delimiters. Each sentence obtained after the initial splitting is then parsed using the Stanford parser [51]. The next round of splitting is completed using the coordinating and subordinating conjunctions in each sentence. Then, the sentence is further simplified if it contains a relative pronoun [51]. To compare the system’s accuracy, 200 sentences were first given to the RBMT system to translate from English to Tamil. Due to syntax and reordering, 70% of the translated sentences were incorrect. Then, the same 200 sentences were simplified using the outlined framework and given again to the RBMT system. After simplification, 57.5% of the sentences were translated correctly. The authors concluded that longer sentences that are given to the MT result in a low translation accuracy while simplifying the sentences increase the accuracy significantly when translating from English to Tamil. Although an accuracy of 100% is not possible to achieve in MT, the authors prove that the splitting and simplification technique can notably improve MT systems.

In another paper titled “Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation”, the authors also tackle the problem of decreased accuracy due to long sentences used in machine translation. Their hypothesized solution is to use a statistical post editor in conjunction with the RBMT system to improve accuracy. The data used by the authors was collected from the Patent Abstract of Japan or PAJ, and the abstract of the Patent Publication Gazette (PPG) of Japan in which the former was used as the Japanese corpus and the latter was used as the English corpus [52]. The sentences were manually translated. Using the corpi, the training and test datasets were created using the following steps: first, the number of words in sentences was counted and if the number exceeded 90, the sentence was rejected. Next, if the ratio of the number of words in sentences from both the reference text and the translated source text did not fall between 0.5 and 2, inclusively, the sentence would also be rejected [52]. To evaluate the RBMT system combined with the statistical post editor (SPE), the authors propose a new metric of evaluation: an n-gram-based NMG measure. NMG, or normalized mean grams, counts the total number of words within



(a) The learning process of a statistical post editor (b) The process of translation

Figure 2.8: System architecture of a statistical post editor [52]

the "longest word sequence matches between the test sentence and the target language reference corpus" [52]. The architecture of the system can be seen in the figures below. Figure 2.8a illustrates the learning process of the SPE while Figure 2.8b demonstrates the process of translation. Their results concluded that for patent translation where sentences are long and complex, the RBMT provided an advantage for structural transfer [52]. Additionally, since patents are made up of very technical terms, the SPE provided improved lexical transfer [52].

Before NMT systems became popularized, SMT models were widely used for translation purposes. SMT systems use a phrase-based approach which allows them to reduce the restrictions of RBMT's word-based approach by translating sequences of words at a time. NMT systems further improve on this because of their ability to learn and improve independently, however, they are still far from perfect. In [53], the authors isolate one of the shortcomings of NMT and propose a novel approach to improve NMT by using SMT. NMT systems have a tendency to forgo accuracy for fluent translation and to improve on this, the authors introduce a hybrid approach using SMT [53]. To attain the goal of improvement, the authors implement the following steps. After the SMT and NMT models have been trained on parallel corpora, the SMT system receives the source text first as input and the translated output of the SMT is encoded. The authors then modify the NMT beam search algorithm which gives a chance to related SMT tokens during NMT decoding [53]. After the modified beam search has been con-

```
Begin
   SMT_token_index = -1
   for beam_index 1 to beam_size
      if(beam_tokens[beam_index] isEqual SMT_tokens[token_step])
            SMT_ token _index = beam_index
            breakTheLoop
      end
   end
   if SMT_ token _index isNotEqual -1
      beam_probabilities[SMT_token_index] = beam_probabilities[1]
   end
End
```

Figure 2.9: Pseudocode of modified beam search algorithm by Şatır *et al.* [52]

ducted, the translation sentence from the modified beam algorithm, as seen in Figure 2.9, goes into the decoder which then outputs the translated target text [53]. The intent of this improved algorithm is to increase the probability of SMT tokens that are in lower positions within the beam to be chosen in the NMT decoding step [53]. The algorithm is used in three different approaches in the paper. First, it is applied for a certain number of tokens from the start of the decoding, e.g., the first, second, or third tokens. Second, the algorithm is applied for each sentence as long as the SMT token is discovered in the beam at some part of the decoding. Lastly, it is applied until the decoding ends [53]. The authors use automatic evaluation metrics, specifically BLEU and METEOR, to draw conclusions from the experiment. They conclude that the first and second approaches provide good results in the translation quality while the third approach, however, performs poorly. Nevertheless, they are able to successfully show that using the phrase-based SMT system can provide improvements in NMT decoding which ultimately leads to a higher quality of translations [53].

As mentioned previously, NMT models have become a dominant approach to machine translation problems and have improved on several shortcomings of SMT over the years. Though a promising technology, NMT still faces many hurdles as its accuracy in translating several language pairs significantly depends on the availability and use of large parallel corpora [54]. For a large group of languages, however, obtaining a large parallel corpus proves to be difficult. For example, language isolates such as Basque, or macaronic languages such as German-Russian [54] do not have enough data available to train an MT model efficiently. Research has been conducted to overcome this problem and techniques such as triangulation techniques and semi-supervised methods have been proposed, however, there is still a need for strong cross-lingual learning [54]. To overcome the need of using cross-lingual learning, the authors of [54] propose a novel solution to train NMT models. They suggest relying on a monolingual corpus for entirely unsupervised model training.

The authors used a standard model architecture: an encoder-decoder system with attention mechanisms where the encoder and decoder contained a two-layer bi-directional recurrent neural network, and the attention mechanism used was a global attention method with the general alignment function [54]. Three critical aspects allowed the MT system to be translated with an unsupervised approach and this included: a dual structure, a shared encoder, and fixed embeddings within the encoder. Furthermore, two strategies allowed the NMT system to predict translations in a monolingual corpora which would have been otherwise impracticable since the authors did not use a parallel corpus. First, they use the principle of denoising autoencoders that train the system to reconstruct a corrupted input to its original form. More specifically, they switch the word order of the input sentences so that the system can learn to retrieve the correct order [54]. Next, they use an adjusted on-the-fly back translation method so that given a particular input sentence in the source language, the system can use inference mode with greedy decoding to translate the input to the target language [54].

After conducting both automatic and human evaluations, the experiments conclude that there is a significant improvement in translation over a baseline system that performed word-by-word substitution. The system was also able to effectively model cross-lingual relations and output accurate and excellent quality [54]. They further showed that moving on from a strictly unsupervised case by incorporating a small parallel corpus has the potential to further improve the quality of translations [54].

Another translator that successfully utilizes neural network architecture for translation is

DeepL. DeepL is a neural machine translation service that advertises its enhanced performance compared to competitor tech companies such as Google Translate and Microsoft Translator. It separates itself from the competitors by improving the neural network methodology in four different areas: the network architecture, training data, training methodology, and the size of the network [14].

2.4 Current applications of machine translation

2.4.1 In healthcare

Since the development of machine translation, many sectors have started to rely on its technology. One crucial application of the technology can be found in healthcare where language barriers contribute to the disruption of patient-clinician communication which plays an essential part in the quality of healthcare provided [33]. Although a cost-effective and efficient way to communicate with patients, MT struggles with reliability and accuracy and studies have found that medical discharge information that had been translated by general purpose translators such as Google Translate was inaccurately and incorrectly translated in Spanish 8% of the time while Chinese was incorrectly translated 19% of the time. Of the incorrect Spanish translations, 2% posed the risk of harm while 8% of the incorrect Chinese translations posed harm [34]. The authors of [33] found that MT provides assistance to healthcare workers, specifically by saving clinician time and by assisting medical interpreters with translation services for languages that are less widely spoken. However, in a field where accuracy is of utmost importance and should not be compromised, the improvement of MT in highly technical fields is necessary. The authors of [33] establish opportunities where MT can dependably support cross-lingual communication [33]. MT systems must improve in ways that support domain-specific languages and thus MT systems that are to be used in the healthcare sector must be trained to translate medical language as accurately as possible and the system must also have the capacity to support the various dialects of languages [33]. One approach proposed is to amalgamate NMT with phrases that are professionally translated. Although this method would shrink the range of communication for users, accurate translation of those phrases is assured [33, 35, 36]. Translation quality assurance is also a main concern when implementing MT in healthcare spaces. In [37], the authors conducted 34 interviews and addressed the anxiety felt by clinicians and staff members over the quality of translation, which was attributed to apprehension over losing credibility and being held responsible for negative outcomes caused by incorrect translation results. The performance of MT systems varies depending on language pairs, training data, and the investment put forth in developing the model, and thus, developers of MT systems meant for the healthcare sector must consider conveying the limitations of the system in a clear manner to its users [33]. Materials for on-boarding and guidelines for use would help clinicians feel comfortable about using the technology safely [33].

2.4.2 In business

Another field that stands to significantly benefit from the introduction of MT is business. According to the European Parliament, “language knowledge can contribute to the creation of added value because it can make the process of purchase, production, and sales more efficient” [38]. Thanks to an increasingly global economy, businesses stand to benefit greatly from creating international connections [39]. Research conducted by the authors of [38] concludes that on average a shared language has the capability to increase trade flows by 44% [38]. Additionally, the internationalization of several companies that depend on communicating in several languages to market their products and give their administrations has incited a myriad of interest. For instance, in 2022, the market for MT was estimated at 153.8 million USD and is expected to grow to 230.69 million USD by 2028 with a compound annual growth rate of 7.3% between 2022 and 2030 [40].

2.5 Challenges in machine translation

Defined by Philipp Koehn in [84], the task of translation is always an approximation. Rather than just a simple code, languages are complicated systems of communication that include a wide range of social, cultural, and cognitive characteristics. Each language has its own set of norms and structures, as well as a unique cultural and historical context, all of which influence their usage and interpretation. Context and target audience also influence interpretation and so there may be multiple possible translations for a given sentence. The creative process for which translators require linguistic, cultural, and communicative competence make translation a very difficult and complex problem that linguists, computer scientists, and researchers have studied and made notable advancements in machine translation for 74 years and continue to do so today. In this section, we cover some common challenges encountered within the realm of machine translation.

2.5.1 Out-of-domain performance

Translating is a balancing act between adequacy and fluency. With many different ways of translating a single sentence, as shown in Table 2.4, translators must make the decision to preserve the meaning of the original input text, or to output text that is interpreted articulately in the target language but sacrifices the literal meaning of the source text. The former choice aims to achieve adequacy while the latter choice prioritizes fluency. Both of these goals of translation exist simultaneously but lack harmony as there is a trade-off between the two. Consider the following examples. Translations of literature, specifically those of poetry, are often focused on capturing the theme of a story which the author may build using literary devices such as metaphors, word choice, and irony among many others. These devices must be interpreted by the translator and the literal meaning may need to be changed completely so that the intended meaning is faithfully translated into the target language [85]. Here, fluency is given precedence. On the flip side, translations of legal text or medical documents require rigid rules as the smallest mistranslation of the documents in either profession may likely lead to serious consequences for the participating parties [86]. To reduce the risk of mistranslations or key information being left out of the translated text, adequacy is prioritized despite the increased risk of outputting awkward sentence structures in the target language as a result.

Assessment Correct/Wrong	Translation
1/3	*Without fail, he has been concise and accurate.*
4/0	*Without getting flustered, he showed himself to be concise and precise.*
4/0	*Without falling apart, he has shown himself to be concise and accurate.*
1/3	*Unswayable, he has shown himself to be concise and to the point.*
0/4	*Without showing off, he showed himself to be concise and precise.*
1/3	*Without dismantling himself, he presented himself consistent and precise.*
2/2	*He showed himself concise and precise.*
3/1	*Nothing daunted, he has been concise and accurate.*
3/1	*Without losing face, he remained focused and specific.*
3/1	*Without becoming flustered, he showed himself concise and precise.*

Table 2.4: 10 different ways of translating the same sentence [84]

Out of the many challenges NMT faces, its performance out-of-domain has been a persistent one. NMT systems falter when faced with circumstances that do not reflect the training conditions and will forego adequacy for the sake of fluency. In [87], authors Koehn and Knowles tested the performance of NMT and SMT systems, trained on five distant domains. These domains included law, medical, IT, Koran, and subtitles. Additionally, the systems were also trained on the combined training data from all five domains. Figure 2.10 presents an overview of the results achieved by Koehn and Knowles. Their results show that the in-domain performance of NMT and SMT are alike, however, the performance of NMT out-of-domain is drastically worse in most occurrences [87]. They reported that the NMT system would output a translation that was fluent in its TL but would be unrelated to the input. The SMT system on the other hand would not translate some words to preserve adequacy relative to the input [87]. These results are far from desirable when considering the need for NMT to translate legal and medical document translations.

System ↓	Law		Medical		IT		Koran		Subtitles	
All Data	30.5	32.8	45.1	42.2	35.3	44.7	17.9	17.9	26.4	20.8
Law	31.1	34.4	12.1	18.2	3.5	6.9	1.3	2.2	2.8	6.0
Medical	3.9	10.2	39.4	43.5	2.0	8.5	0.6	2.0	1.4	5.8
IT	1.9	3.7	6.5	5.3	42.1	39.8	1.8	1.6	3.9	4.7
Koran	0.4	1.8	0.0	2.1	0.0	2.3	15.9	18.8	1.0	5.5
Subtitles	7.0	9.9	9.3	17.8	9.2	13.6	9.0	8.4	25.9	22.1

Figure 2.10: Quality of NMT and SMT systems, represented as green and blue bars respectively, when trained on one domain (rows) and tested on another domain (columns) [84].

2.5.2 Rare word problem

Within the realm of the previous challenge exists another problem posed by NMT: the problem of rare words. NMT systems struggle to perform well when faced with rare words because of their use of smaller vocabularies during training. This difficulty arises from the fact that the computational complexity is directly proportional to the number of target words used in training. Additionally, memory requirement is also linearly dependent on the number of target words involved and thus grows in proportion. To confront this issue and make training more feasible, recent NMT systems use 30k to 80k most frequent words in the target language [6, 8]. However, by tackling this problem, it raises another. In a situation where the translation of the SL depends on numerous words that are not included in the list of most frequent words, the quality of the model's output will significantly decline [128]. This problem is amplified with highly inflected languages such as Japanese which has several inflected verbs and adjectives. Additionally, according to [87], verbs and adjectives across nearly all word frequencies have a higher rate of deletion and a lower rate of accuracy than nouns.

2.5.3 Long sentences

An initial problem faced by the encoder-decoder NMT model that resulted in poor translation quality was its effort against the translation of long sentences. Although this difficulty has been resolved to a limited extent by the introduction of the attention mechanism, authors Koehn and Knowles [87] tested how well and to what extent the problem had been rectified. As seen in Figure 2.11, their results show that though overall NMT performs better than SMT, its quality of translation drastically decreases when longer sentences of length 80 and more are introduced to the system. The context vector in the attention mechanism is responsible for calculating and delivering predictive support to the network that determines which target word should be chosen next. When very long sentences are presented to the system, the values of the hidden states become too distributed for the context vector to carry out its function efficiently [129]. This produces translations that are too short to be accurate; Koehn and Knowles [87] reported that the length ratio was 0.859 (versus 1.024).

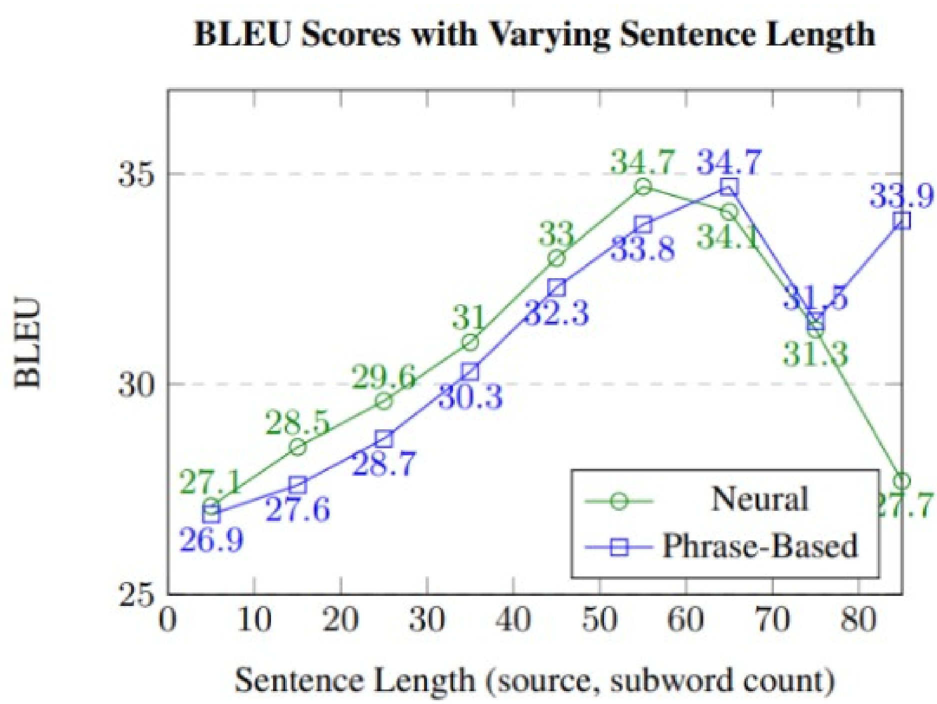


Figure 2.11: Quality of translation as longer sentences are introduced to NMT and SMT systems [84].

CHAPTER 3

Methodology

The aim of this chapter is to present a comprehensive framework that was used for our research project, as well as to provide an in-depth exploration of the methods and procedures employed to gather, analyze, and interpret the data. Specifically, our methodology aims to overcome domain mismatch in the realm of patent document translations using the English-Japanese (en-ja) language pair. We begin by discussing the architecture of the three chosen baseline models and begin an initial evaluation of translation results from patent data. Successively, we determine the right domain adaptation techniques needed for the enhancement of model performances. The data collection process implements critical and meticulous data-centric methods to create a patent corpus using language resources provided to us by our industry partner. We end the chapter with the pre-processing steps used to create our training dataset. Figure 3.1 presents a UML activity diagram that provides an overview of our applied approach.

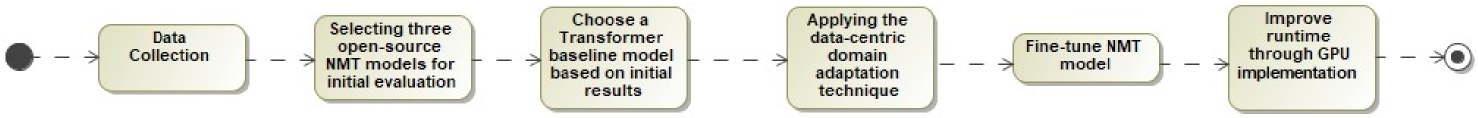


Figure 3.1: UML activity diagram of our research methodology

3.1 Baseline model architecture

Before beginning the methodology section, we present some background information on the Transformer NMT architecture, which was the baseline architecture of the open-sourced NMT models chosen for experimentation, and domain adaptation techniques.

3.1.1 Transformer encoder-decoder models

Introduced by Vaswani et al. [57] in 2017, the Transformer NMT architecture revolutionized the field of machine translation and became the state-of-the-art approach for several translation tasks. This architecture, based on attention mechanisms, addresses the limitations of traditional RNNs and CNNs by leveraging the power of self-attention and parallel processing and leaving out recurrences and convolutions altogether [57]. The Transformer model employs self-attention to capture dependencies between different words in the input sentence. This allows the Transformer to effectively model long-range dependencies and capture contextual information more efficiently. Large language models (LLMs) such as ChatGPT also use the attention mechanism for the purpose of translation, however, similar to the limitations of Google translate, ChatGPT translation ability may struggle with specific types of text or context [127].

The architecture of the Transformer consists of an encoder and a decoder. The encoder processes the source language sentence, while the decoder generates the translated target language sentence. Both the encoder and decoder consist of multiple layers of self-attention and feed-forward neural networks. Specifically, each layer in the encoder and decoder has two sub-layers: a multi-head self-attention mechanism and a position-wise feed-forward neural network

[57]. The self-attention mechanism enables the model to attend to different positions within the input sentence to capture the most relevant information. Additionally, it also allows the model to consider the context of each word based on its relationship with other words within the sentence. To accomplish this, the attention mechanism computes attention weights for each word in the input sentence by taking the dot product of the query, key, and value vectors. This process is repeated multiple times using different linear projections, called attention heads, to capture diverse relationships between words and determine how much each word contributes to the representation of other words in the sentence [57].

After the self-attention sub-layer, a position-wise feed-forward neural network is applied to each word individually. This network consists of two linear transformations followed by a non-linear activation function, such as the ReLU, which helps the model capture more complex and non-linear relationships between words [57]. The encoder uses multiple layers consisting of self-attention and feed-forward sub-layers to process the source sentence. Subsequently, the encoder outputs a sequence of vector representations for each word in the input sentence that capture the semantic and syntactic information of the source sentence. The decoder takes these representations as input and generates the target sentence word by word. It also incorporates an additional attention mechanism called the encoder-decoder attention which allows the decoder to attend to the relevant parts of the encoded source sentence when generating each word in the target sentence [57]. Because of its reputation as the state-of-the-art foundational NMT model, the models used for the purpose of our experiments have a transformer-based architecture. Further details of the encoder-decoder architecture are provided in Sec. 2.1. Figure 3.2, originally presented in [57], provides a visual of the architecture of the transformer model.

3.2 Data collection and processing

In [117], Saunders refers to fine-tuning in the context of domain adaptation for NMT models as the default approach and highlights that though it is an efficient and uncomplicated method

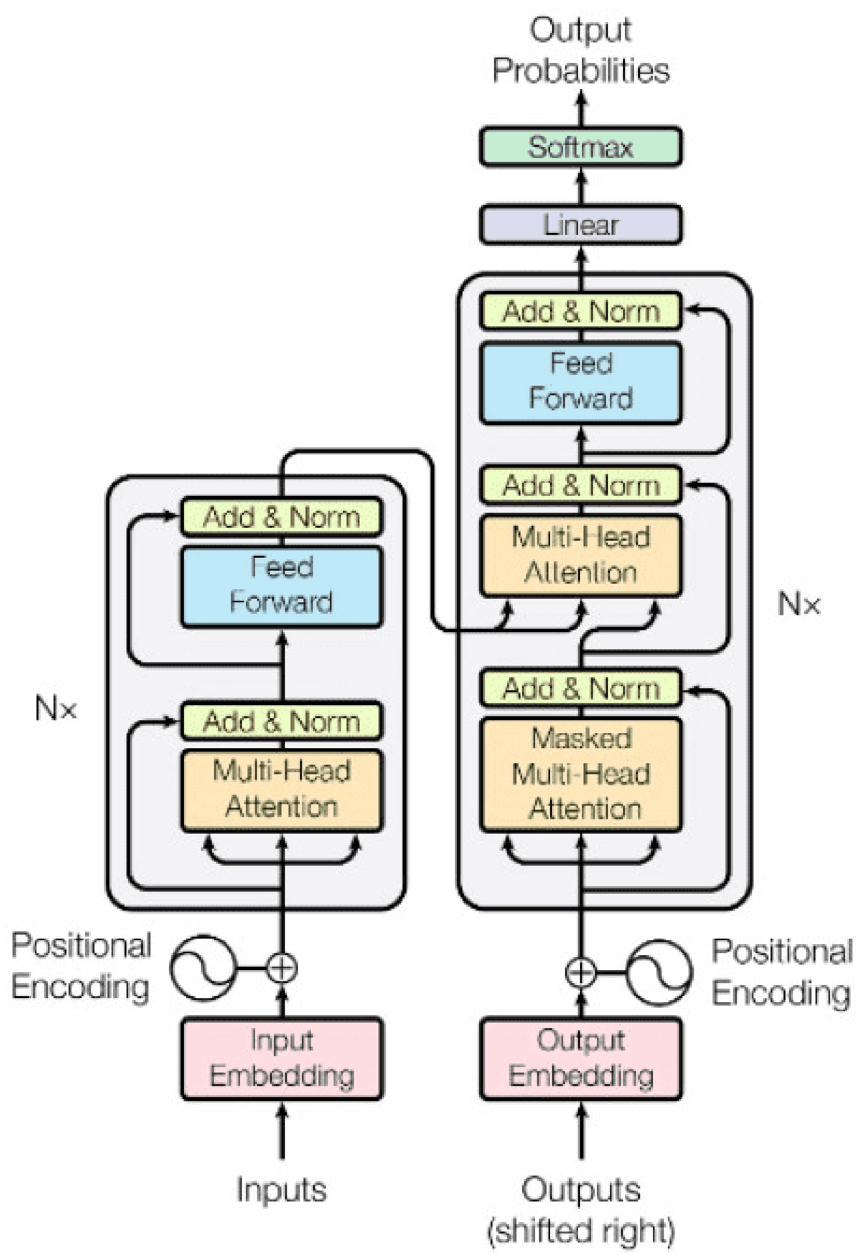


Figure 3.2: The transformer model architecture from Vaswani *et al.* [57]

there are three main difficulties related to fine-tuning that negatively affect the quality of models; overfitting, catastrophic forgetting, and insufficient training data [117]. Overfitting is a problem that commonly occurs when the dataset used for fine-tuning is not large enough or the contents of the dataset are repetitive. Overfitting is also affected when there exists a mismatch between the test-sentence and fine-tuning domains and becomes especially relevant in cases of patent translation due to its susceptibility to domain mismatch. For instance, an NMT model trained on a specific adaptation domain, say for example the domain encompassing IPC section C, may perform exceptionally when translating patents relating to inorganic chemistry but a slight deviation to domain section D raises overfitting difficulties. We discuss what constitutes a *domain* over our feature space of interest in the following section.

The second difficulty is catastrophic forgetting, or catastrophic interference, which also stems from the same reasons associated with overfitting; forgetting also occurs in instances when the system must translate a domain that is outside the feature space of the training domain. However, in this scenario, forgetting occurs when a system trained on domain A is fine-tuned using domain section B, causing the system to 'forget' its training and ensuing performance depreciation on A for the benefit of B [117].

Regardless of their differences, the prevailing difficulties can be mitigated by expanding the in-domain corpus as explained by Saunders in [117]. However, it cannot be ensured that there exists an in-domain dataset that is large enough and of high quality.

In this section, we discuss our approach of leveraging data-centric methods to reduce domain mismatch by expanding the in-domain corpus to include all eight IPC sections. We meticulously select our data to create training, testing, and validation datasets that are representative of the target domain: the entire IPC feature space. The dataset used for our experiments was created with the help of our industry partner, XLScout. By using their database, we were provided access to all patents classified within the international patent classification (IPC) system.

3.2.1 IPC feature space

The international patent classification system, abbreviated to IPC, is a hierarchical system that offers a practical and effective way of classifying and retrieving patent documents according to a specific domain [78]. The classification is a language and terminology-independent system that is maintained by the World Intellectual Property Organization (WIPO) and is used by patent offices around the world.

The IPC system is organized into sections, classes, subclasses, groups, and subgroups, where each level corresponds to a progressively more specific technological domain. Depending on the technical content of the invention, each patent is given one or more IPC codes which assist in identifying and categorizing the invention, making it simpler to search for. There are **eight** main sections:

(A) Human Necessities

(B) Performing Operations; Transporting

(C) Chemistry; Metallurgy

(D) Textiles; Paper

(E) Fixed Constructions

(F) Mechanical Engineering; Lighting; Heating; Weapons; Blasting

(G) Physics

(H) Electricity

Each section is divided into classes. Classes are further divided into over 600 subclasses. In total, about 70,000 classification codes can be assigned to patent documents in the classification system [78]. Table 3.1 provides an example of each division using domain (C) Chemistry, and Figure 3.3 provides a visual representation of the IPC hierarchy.

Table 3.1: An example demonstrating each level in the IPC system

	IPC Divisions	
	Symbol	Description
Section	C	Chemistry; Metallurgy
Class	C01	Inorganic Chemistry
Subclass	C01C	Ammonia; Cyanogen; Compounds Thereof
Group	C01C3/00	Cyanogen; Compounds thereof
Subgroup	C01C3/02	Preparation of hydrogen cyanide

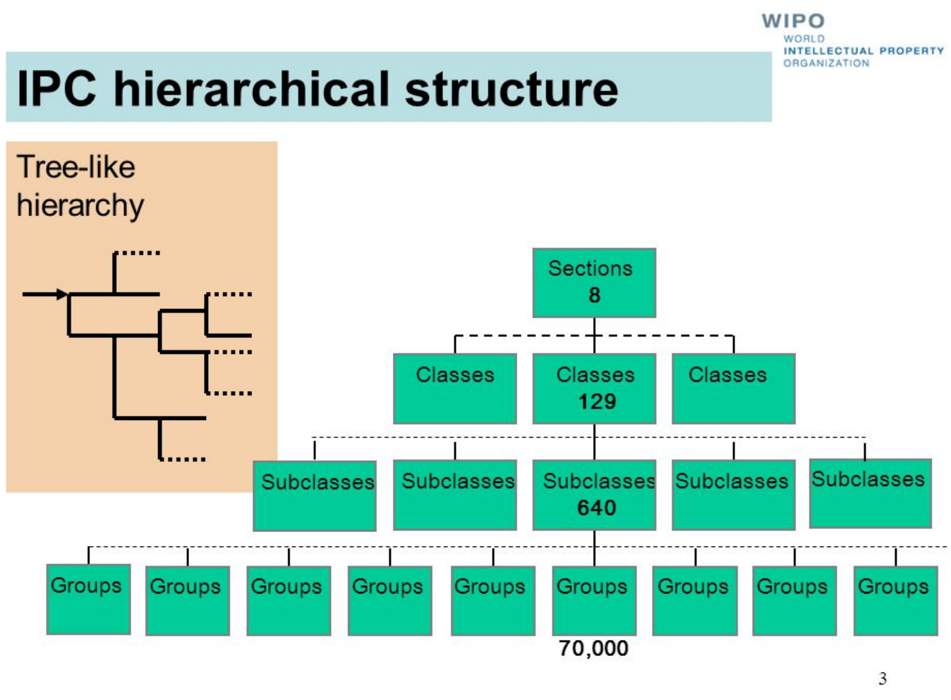


Figure 3.3: A tree diagram representation of the IPC system hierarchy [78]

To begin our research methodology, we searched through the space of all eight domains resulting in a total of 144, 789, 078 patents found and refined the data by filtering through the patents based on three things: the publication country, which was set to Japan; the patent's abstract in both its original language (Japanese) and English; and the IPC section that the patent belongs to. Although the task is to translate from English to Japanese, the data was filtered in this manner to guarantee we extracted abstracts in both languages.

We formally define our feature space to be the entire IPC hierarchical system and each section of the system is defined as a domain. Thus, we have a total of 8 domains from which we needed sufficient data. To ensure a high-quality dataset, we decided to use patent abstracts to create our parallel corpus because they offer a comprehensive summary of the invention including technical aspects and potential applications that are specific to the domain they belong to.

3.2.2 Pre-processing

We exported 15,000 random publication numbers, each referencing a distinct patent, from each domain to begin the data preprocessing steps, and using 120,000 publication numbers in total, the patent abstracts in English and Japanese were extracted, cleaned, and tokenized. After extracting the patent abstracts from the IPC database, we cleaned the data by removing noisy elements such as '< .*? >' that were included as part of the raw data. The abstracts in both languages were then tokenized into sentences and aligned to create a bilingual parallel corpus. To split the English abstract, we used the '.' delimiter to iteratively split sentences and store them in a list, however, tokenizing Japanese sentences as a non-Japanese speaker is much more difficult to achieve without the help of a program as there is no space between words and though the delimiter '。' may intuitively seem as the Japanese equivalent of a full stop, English and Japanese punctuation vary greatly. The Japanese full stop is used to separate sentences rather than finish a sentence. Additionally, the Japanese language also uses symbols such as the kanji character for "stop" (止) to indicate the end of a sentence, and thus, splitting the Japanese sentences based on this delimiter would prove to create a misaligned corpus. Instead, we used the *fugashi* tokenizer to segment our data and similarly stored the segmented Japanese sentences into a new list. Next, using the Enchant spellchecking library, we created an English dictionary and applied it to spellcheck and clean the English data. We then iteratively combined the two lists to create a bilingual parallel corpus of 188,169 parallel sentences using the eight sections of the IPC system.

3.2.3 Augmentation using back translation

After the preprocessing steps, the dataset created still contained grammatical errors and thus reduced the accuracy of the performance. To rectify this, an additional preprocessing step was performed: data augmentation using back translation. Back translation is a technique of three steps to improve the quality and fluency of translations and involves temporary translation, back translation, and duplicate removal [123]. Temporary translation is the process of translating the original text to the target language, however, since we had extracted the original Japanese text, we skipped to the next step; back translation. Back translation translates the data in the target language to the source language and uses either the original source sentences or back-translated sentences in an effort to train the model. The last step, duplicate removal, aims to ensure that only one instance of the SL text is used. Keita demonstrates in [123] how back translation can increase the quality of data as seen in Figure 3.4.

3.2.4 Training, testing, and validation

In many machine learning applications, including MT applications such as this research, a large dataset is desirable to ensure that sufficient data is available for splitting into subsets for training, testing, and validation. Common split ratios include 70/15/15 or 80/10/10 and in general, best practices within ML research and development include using a split ratio such that more data is used for training than for testing and validation, given a corpus of adequate quality. We make no exception to this rule as can be seen in Figure 3.5. Prior to splitting the

Original Texts

```
['This article aims to perform back translation for text data increase',
 "This is Zoumana's 25th article on Medium. He likes to give back to the community.",
 'The first model translates from English to French, which is a temporary process',
 'The second model finally translates all the temporary French text into English']
```

Final Augmentation

```
{'It is the 25th article by Zoumana on Medium. He loves to give back to the community',
 'The first model translates from English to French, which is a temporary process',
 'The second model finally translates all the temporary French text into English',
 'The second model finally translates back all the temporary french text into English',
 'This article aims to perform back translation for text data increase',
 'This article aims to perform the back translation for text data augmentation',
 "This is Zoumana's 25th article on Medium. He likes to give back to the community."}
```

Figure 3.4: Results from [123] demonstrating the results of using back translation

corpus, we shuffle the data to implement randomization using the Pandas and NumPy libraries. This is done to ensure generalizability to the feature space, help eliminate bias present in the data, and maintain statistical validity. We then split the corpus using the split ratio of 60/20/20. Table A.1 in Appendix A presents a subset of the bilingual parallel corpus.

```
DatasetDict({
    train: Dataset({
        features: ['english', 'japnese', '__index_level_0__'],
        num_rows: 112901
    })
    validation: Dataset({
        features: ['english', 'japnese', '__index_level_0__'],
        num_rows: 37634
    })
    test: Dataset({
        features: ['english', 'japnese', '__index_level_0__'],
        num_rows: 37634
    })
})
```

Figure 3.5: Features of our bilingual training, testing, and validation datasets.

3.2.5 Machine translation post-editing (MPTE)

In many professional settings, including patent law, post-editing machine-translated text is a required and crucial step of the research methodology as the desired standards of quality are high. Post-editing is the process of reviewing and amending text that has been translated by a machine, used by professional and qualified translators. In the case of patent translations, it is also imperative that the translators are knowledgeable in the patent laws of both of the countries involved, as the patent must be as accurate as possible.

As non-Japanese speakers, we needed to confirm the quality of our translations as our quality check was limited to Google Translate and BLEU score calculations (Sec. 3.5.1). To do this, a Japanese translator at XLScout provided post-editing feedback on the quality of each machine's performance so that we may be able to make an informed decision before selecting the model for adaptation and fine-tuning.

3.3 Pre-trained transformer models

Within machine learning research and development, open-sourced codes are commonly used as a tool to advance the different fields of study by encouraging replication and experimentation. In general, the majority of MT systems rely on specialized setups that require extensive engineering or that might only be effective for a particular issue [82], and in scenarios where computational resources and the original dataset are not accessible, building a model from scratch becomes a virtually impossible task [117]. To combat the challenges that come from conducting novel experiments and to accelerate research within DL, the Google Brain Team introduced an open-source system called Tensor2Tensor (T2T) that enabled researchers to run their own experiments with deep learning models [82]. Our experiment consists of three open-sourced and pre-trained T2T models. The first NMT model [65] we looked at is an implementation of the Transformer model introduced by Vaswani *et al.*, and the remaining two models have a baseline Transformer architecture that has been adapted using the T2T library by HuggingFace. For commercial confidentiality reasons and for the sake of consistency, the names of the HuggingFace models will be hidden and all the models will be referred to as NLP-Model-I ([65]), NLP-Model-II (HuggingFace model I), and NLP-Model-III (HuggingFace model II).

The authors of [65] implemented the baseline architecture of the Transformer model outlined in *"Attention is All You Need"* [57] to make NLP-Model-I. Details of the base model's architecture were given by the authors of [57] and are presented in Table 3.2.

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps
base	6	512	2048	8	64	64	0.1	0.1	100K

Table 3.2: * Values of the baseline transformer model architecture presented by [57].
* *where N* is the number of layers in the encoder/decoder,
d_{model} is the dimensionality of the input and output,
d_{ff} is the dimensionality of the inner-layer,
h is the number of parallel attention layers or heads,
d_k are the queries and keys of dimension d_k,
d_v are the values of dimension d_v,
P_{drop} is the residual dropout rate, and
ϵ_{ls} is the label smoothing value

The model had not been fine-tuned but had been trained on a bilingual parallel corpus for the purpose of general Japanese-English translation. The corpus was made up of two merged datasets; the first is a corpus that contains approximately 500,000 pairs of sentences that cover the topics of Japanese religion, culture, and history [68], and the second is a collection of bilingual sentence pairs created by [55] that comprise of Japanese sentences used in daily conversations. The model was trained using 68,674 rows of the dataset and then evaluated using the BLEU score. The authors were able to achieve a BLEU score of 41.49 [65] as seen in Figure 3.6, suggesting a good accuracy of the translated text. To use this model for the purpose of patent translation, first, the model was adjusted so that the source text was English and the target text was Japanese. The same dataset used in [65] was then used to train the new model. Once trained, we tested the translation accuracy on sentences extracted from technical patents and then evaluated the results using both automatic (BLEU) and human expert evaluations.

NLP-Model-II was developed for the Tatoeba translation challenge which aims to serve as a catalyst for the development of open translation models [56]. The dataset used is an

	BLEU Score
Baseline Model (kyoto Lexican Dataset)	4.86
Transformer Model (kyoto Lexican Dataset)	14
Transformer Model (Anki Dataset)	61
Transformer Model (Merged Datasets)	41

Figure 3.6: The authors of [65] implemented the base transformer architecture for Japenese-English MT and achieved a BLEU score of 41.49

amalgamation of the Open Parallel Corpus (OPUS) [62], an open collection of parallel corpora, and test data extracted from [66]. Similarly, the dataset used to train NLP-Model-III was built using various datasets which consisted of a total of approximately 6.6 million bilingual pairs. Of these many datasets, the following were used: the Japanese-English Subtitle Corpus [67], the Kyoto Free Translation Task (KFTT) [68], the Tanaka Corpus [69], the Japanese SNLI dataset [70], and finally WikiMatrix [71]. Each model's hyperparameters remained unchanged from the original baseline model's hyperparameters.

We evaluated the initial performance of each model to determine which NMT model would provide a high baseline BLEU score on test data representative of the target domain. The model with the highest BLEU score was then chosen for data-centric adaptation and fine-tuning. In Chapter 4, we present and discuss the results of each model.

3.4 Evaluation and Post-Editing

The evaluation of NMT models is extremely important as it measures the degree of reliability of the output from the MT model and it also informs us when a model requires improvement [42]. Over the years, many evaluation methods have emerged, many of which fall into two categories: human evaluation and automatic evaluation. Automatic evaluation metrics work by comparing the output of an MT system to a set of references generated by humans also called the gold standard references [43], and then making use of statistical calculations to compute how different the machine-translated output is from the reference translation [43]. The quality of the translation is considered better if the difference between the output and reference is smaller. Automatic metrics use n-grams to calculate the precision scores, where an n-gram is a sequence of n words [43]. In the next section, we go over some commonly used automatic metrics that are used to evaluate MT and discuss why the BLEU score is a commonly used metric of evaluation and why we chose it as part of our automatic evaluation.

The cost of automating the evaluation process is that the quality of MT systems cannot be measured with vigor because automatic evaluations are complicated and opaque [125]. It is essential to complement metrics like the BLEU score with human evaluation, also called post-editing. Thus, it is an imperative step of our evaluation process and we employ the expertise of a Japanese translator to evaluate the results of the NMT models.

3.5 Evaluation metrics

3.5.1 Bilingual Evaluation Understudy score

A BLEU score, also known as the Bilingual Evaluation Understudy score, is a metric used to assess machine-translated text and evaluate how accurate it is compared to a set of references. More specifically, it is the product of the geometric mean of precision scores given N, where N is the n-gram, and a brevity penalty, BP. The brevity penalty refers to the penalty applied to the BLEU score when the translated text is much shorter than the reference text and is also used to restitute the BLEU score for not having a recall term [44]. The n-gram precision is calculated by counting the total number of word sequences from the MT system output that is also in the set of references [45]. An n-gram, put simply, is the set of n consecutive words within a given sentence [46]. For example, considering the sentence “the wall is white”, a 1-gram or unigram is a set that consists of “The”, “wall”, “is”, “white” and a 2-gram or bigram consists of “The wall”, “wall is”, “is white”. It is important to note that the words within an n-gram must be taken in consecutive order [46].

$$\begin{aligned} Geometric\ Mean\ Precision\ (N) &= exp\,(\sum_{n=1}^{N} w_n log p_n) \\ &= \prod_{n=1}^{N} p_n^{w_n} \\ &= p_1 \cdot p_2 \cdot p_3 \cdot p_4 \end{aligned} \tag{3.1}$$

$$Brevity\ Penalty\ = \begin{cases} 1 & if\ c > r \\ e^{(1-r)/c} & if\ c \leq r, \end{cases} \ where\ c\ is\ the\ predicted\ length, and\ r\ is\ the\ target\ length \tag{3.2}$$

$$BLEU\ (N)\ = GMP\ (N) \cdot BP \tag{3.3}$$

The product of (3.1) and (3.2), given by Equation (3.3) (when $N = 1$), returns a BLEU score that falls within the range of 0 to 1 (or more commonly, 0-100%), where 0 indicates no overlap between the machine-translated text and the reference text [45]. A score of 100 indicates that the machine-translated text perfectly matches the reference text. Since even linguistic consultants or human translators do not achieve a perfect translation, a BLEU score of 100 is almost impossible. However, as a rough guideline, a score between 60-70 is generally the best a model can achieve. The n-gram precision of BLEU depends on exact word matches between the output and references. However, since a specific reference may not be the only correctly translated option, a good translation may be scored lower [47]. Despite the transparency of the flaws associated with using the BLEU score, it continues to be widely used in MT research mainly due to its high correlation with human judgment of accuracy [47].

3.5.2 National Institute of Standards and Technology

Another metric commonly used in MT evaluation is the National Institute of Standards and Technology or NIST. A variant of BLEU, NIST assigns a higher weight to more informative n-grams and uses the arithmetic mean instead of the geometric mean used by BLEU [45]. The calculation of the BP is also where NIST and BLEU diverge; the variation in length between the translated text and the reference text does not affect NIST as it does with the BLEU score [48]. This is because the precision scores that are calculated in BLEU are replaced with the information gained from each n-gram [49]. This enables the system to get more credit or weight if the n-gram match is difficult to obtain, or less credit if the match is easier [49].

3.5.3 Word Error Rate

The Word Error Rate (WER) is one of the earlier metrics used for evaluating MT [45] and it examines the accuracy based on the Levenshtein distance. The Levenshtein distance between two words from the translated output and the set of reference text, refers to the minimum number of edits that are required to change a word from the translated output to the word from the reference text [45]. The edits allowed are: substitutions (S), insertions (I), and deletions (D). Equation 1 is used to calculate WER, where N is the total number of words in the reference text:

$$WER = (S + I + D)/N \tag{3.4}$$

3.5.4 Metric for Evaluation of Translation with Explicit Ordering

As mentioned in Section 3.5.1, the precision-oriented nature of BLEU is the source of a few weaknesses and so the Metric for Evaluation of Translation with Explicit Ordering (METEOR), a recall-oriented metric, is used to tackle these shortcomings [45]. METEOR calculates the harmonic mean, as opposed to the geometric mean, by combining precision and recall with a greater bias towards recall [45]. The computation of the final METEOR score requires multiple stages; the first stage is exact matching where sentences in the translated output and reference text that are completely alike are aligned [45]. The next stage called stem matching refers to the process of aligning words that have the same morphological stem [45]. Finally, in the synonym matching stage, words that are synonyms of each other (according to WordNet, a lexical database of the English language [50]) are aligned [45]. At each stage, only words that are not aligned are allowed to be matched in the succeeding stage. Furthermore, a fragmentation penalty (FP) is applied to account for the differences in word order [45]. The METEOR score is then calculated by taking the product of the harmonic mean and $(1-FP)$ which outputs a score in the range of 0-1.

3.5.5 Why BLEU?

The BLEU score is a popular choice of evaluation metric in MT mainly due to its simplicity and computational efficiency. Additionally, choosing to use a unigram ($N = 1$) for the calculation of the BLEU score allows for a more focused evaluation that addresses some of the issues related to longer sentences and rare words. Since patents contain highly technical terminology which is often domain-specific, setting the n-gram = 1 may highlight the importance of translating each word as accurately as possible and increase the likelihood of lexical accuracy.

We also use the BLEU score metric to compare the performance of various MT models since BLEU is widely used in the literature and this aids in the comparison of results. Table 3.3 below provides a summary of the advantages and disadvantages of the metrics covered above.

Table 3.3: Advantages and Disadvantages of MT Evaluation Metrics

Evaluation Metric	**Advantages**	**Disadvantages**
BLEU	• High correlation with human judgment of translation accuracy [47].	• Does not account for synonyms [47]. • Penalizes longer sentences [47].
NIST	• No penalty applied when translation text is shorter in length [48].	• Does not consider synonyms, syntactic structure, or word order [48].
WER	• Not difficult to understand or implement [48]. • Reproducible [48].	• Does not consider semantic similarity, syntactic structure, or word order of sentences [48].
METEOR	• Incorporates some linguistic knowledge using stem and synonym matching [48]. • Doesn't penalize longer sentences unlike BLEU [48].	• WordNet does not support most languages well except for English [48]. • Relies on WordNet so not suitable for morphologically rich languages [48].

CHAPTER 4

Experimental Setup and Results

In this chapter, we discuss into the experimental setup employed to investigate and enhance the performance of NMT models for patent translation. The experimental setup plays a crucial role in ensuring the reliability and validity of our research findings. We provide an overview of the experimental design, including the configuration, evaluation, and selection of the NMT models with the primary objective to present a comprehensive outline of the decisions and considerations made during the experimental setup. We aim to lay a solid foundation for the subsequent analysis and interpretation of our results by outlining the decisions made.

4.1 Computing Setup

All training and testing of the NMT models covered in the subsequent subsections were completed using 4 NVIDIA Tesla P100-SXM2 GPUs with 17 GB RAM, as well as 24 CPUs/cores.

4.2 Configuration and Parameter Settings

The "max length" parameter defines the maximum length of input sequences that the model can handle during translation. This parameter ensures that excessively long sequences are truncated or split into smaller segments to maintain computational efficiency. Additionally, the "attention dropout" parameter controls the dropout rate for attention layers, helping prevent overfitting and enhancing model generalization.

"Num beams" specifies the number of beams used during beam search decoding. Beam search is a common technique in MT that explores multiple potential translations simultaneously, allowing the model to generate diverse output translations. By adjusting the value, users can control the trade-off between translation quality and decoding speed.

Beam search decoding is a widely used technique in Neural Machine Translation (NMT) that plays a crucial role in generating accurate and fluent translations. It is employed during the decoding phase of the NMT model to explore multiple potential translation candidates and select the most suitable output. During beam search decoding, the model generates translations by iteratively predicting the next token based on the previously generated tokens. It starts with an initial seed token, usually the start-of-sentence token, and proceeds to generate subsequent tokens by considering the probabilities assigned to different target language tokens by the model. These probabilities are typically obtained through the softmax function applied to the model's output logits.

The beam search algorithm maintains a beam width, which determines the number of potential translation candidates to consider at each step. This width represents the number of branches or paths the algorithm explores during decoding. The larger the beam width, the more candidates are considered, increasing the likelihood of finding high-quality translations. However, larger beam widths also require more computational resources. At each decoding step, the algorithm computes the scores for all the candidate translations based on a scoring function that takes into account the model's predicted probabilities and various other factors. The scoring function helps evaluate the quality and fluency of the translations. The candidate translations with the highest scores are retained, while the rest are pruned to keep the beam width intact. As the decoding process continues, the algorithm proceeds to the next token, expanding the search space further. The scores of the candidate translations are updated at each step, considering both the current token's probability and the accumulated scores from previous steps. This allows the algorithm to capture the most promising translation hypotheses.

Beam search decoding continues until a predetermined stopping criterion is met, such as reaching the maximum length of the output sequence or encountering an end-of-sentence token. Once the decoding process is complete, the algorithm selects the translation candidate with the highest overall score as the final translation. Beam search decoding provides several advantages in NMT. It allows for the exploration of multiple translation hypotheses, enabling the model to generate diverse translations and avoid getting stuck in local optima. Additionally, it helps address the issue of exposure bias by considering the model's own predictions during decoding, leading to more consistent and coherent translations.

However, beam search decoding is not without limitations. One major challenge is the risk of ending up with suboptimal translations due to the greediness of the algorithm. Since beam search only considers the most likely candidates at each step, it may overlook less probable, yet better, translations. Furthermore, beam search tends to favor shorter translations, as they are more likely to achieve higher scores earlier in the decoding process.

Overall, the configurations and architecture of the model highlight the utilization of a Transformer-based neural network, specifically designed for English-to-Japanese machine translation. The model leverages self-attention mechanisms, beam search decoding, and adjustable parameters to achieve accurate translations while maintaining efficiency. These configurations and architecture choices contribute to the model's performance in translating English sentences into their Japanese counterparts.

4.3 Performance comparison of the NLP models

4.3.1 Result I: NLP-Model-I

To implement NLP-Model-I, we followed the code provided by its authors in [65] and switched the *source* and *target* languages so that our source language was set to English and our target language to Japanese. We also used the same training dataset to train our model. As their dataset was not reflective of our target domain, we decided to test the model with shorter and more general sentences from the corpus, to begin with. Though the BLEU score achieved originally by the authors was a desirably high score (see Figure 3.6 in Sec. 3.3.), the lowest level of accuracy achieved out of the three models was by NLP-Model-I, where approximately 1% of the translated text was relatively accurate. Figure 4.1 provides the results of the translations. As shown in the figure, the back-translated results were very poor and nowhere near the original text. Though unpromising, these results reinforced the need to create a training dataset that was composed of parallel sentences and technical terms that are reflective of the test domain, and why general translation systems are not sufficient for patent translation.

4.3.2 Result II: NLP-Model-II

Using test data consisting of technical terminology/jargon from the aforementioned domains, we tested both NLP-Model-II and NLP-Model-III. The results were very promising for NLP-Model-II when compared to NLP-Model-I as the BLEU score was much higher than it was for NLP-Model-I. However, it still could not perform acceptably when translating text from scientific patents. The BLEU score calculated for a sample of paragraphs (of 2-3 sentences) was around 40-50% and many sentences were cut off mid-sentence after the translation as seen in Table 4.1. Moreover, there were occasional punctuation and grammar that were less than desired. We then used post-editing to confirm the results with a human evaluator to help us determine whether the model was performing below par or accurate in its translation of the scientific text. It was confirmed that the accuracy was much lower than expected using

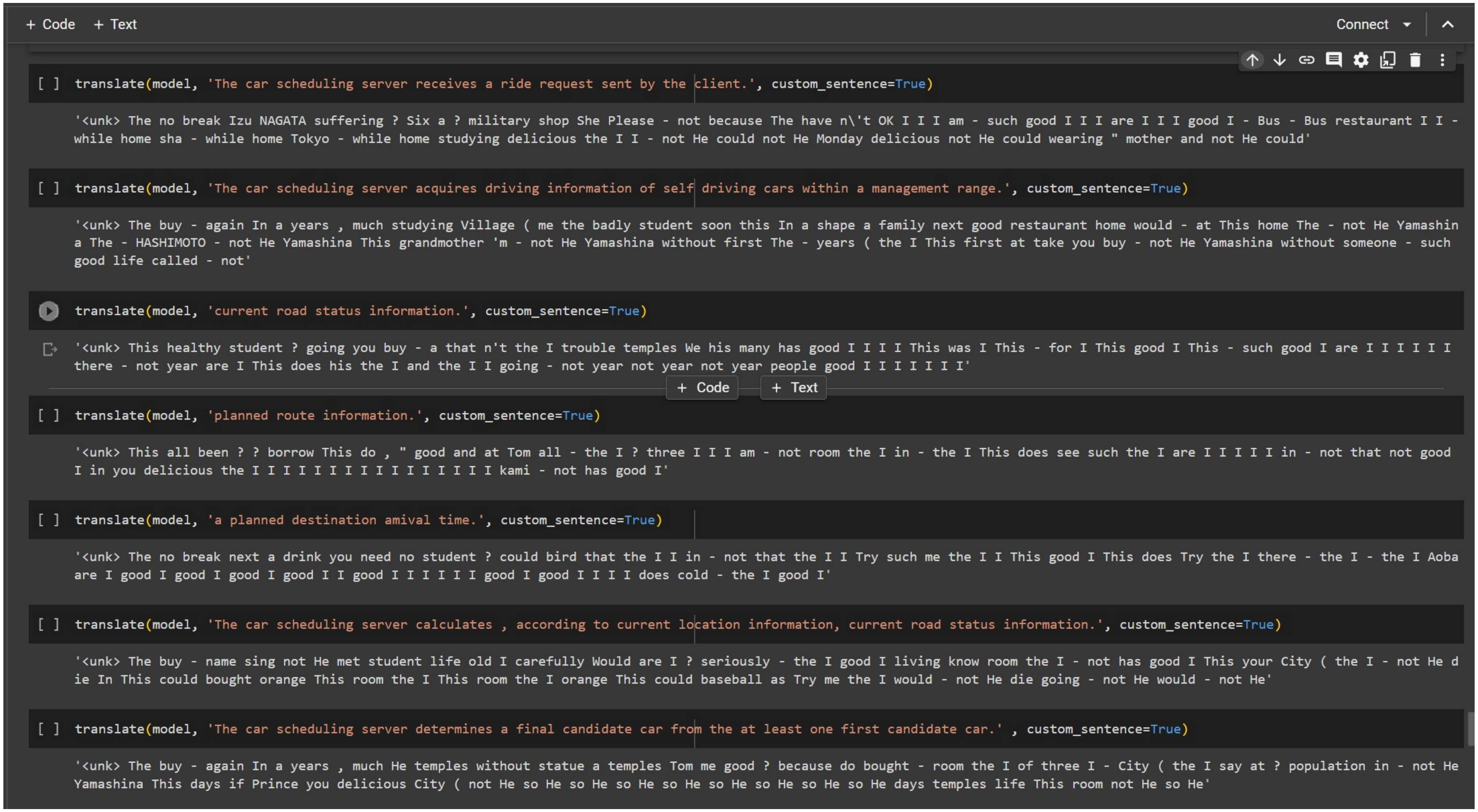


Figure 4.1: NLP-Model-I test results

post-editing; roughly 10% of the translated text was correct. This demonstrates one of the disadvantages when it comes to using only the BLEU score as an evaluation metric.

Model	Original Text	Backtranslated Text
NLP-Model-II	The present invention provides a carbon dioxide absorbent during combustion of fossil fuels comprising a pressed dry powder of plant fibers, and a fossil fuel characterized by comprising such a pressed dry powder of plant fibers.	The present invention is characterized by absorbing carbon dioxide during the combustion of fossil fuels comprising a pressed dry powder of plant fibers , drying the plant fibers in such a way as to produce fossil fuels . We supply."

Table 4.1: NLP-Model-II test results

4.3.3 Result III: NLP-Model-III

As with the previous models, we used the BLEU evaluation metric which was calculated using the *sacreBLEU* function. The function accepted the original text and the back-translated text and produced their BLEU score using n-grams where n ranges from 1 to 4. Detailed features of sacreBLEU are provided in the list below [126].

Features of sacreBLEU:

- "automatically downloads common WMT test sets and processes them to plain text."
- "properly computes scores on detokenized outputs."
- "supports different tokenizers for BLEU including support for Japanese and Chinese."
- "supports chrF, chrF++, and Translation error rate (TER) metrics."
- "performs paired bootstrap resampling and paired approximate randomization tests for statistical significance reporting."

An example of the output displaying the n-gram BLEU scores for a translated patent abstract can be seen in Figure 4.2. To calculate the average BLEU score of each test sentence, we use the 1-gram results. We multiply the BLEU score results by 100 to measure using a consistent scale as for the previous models.

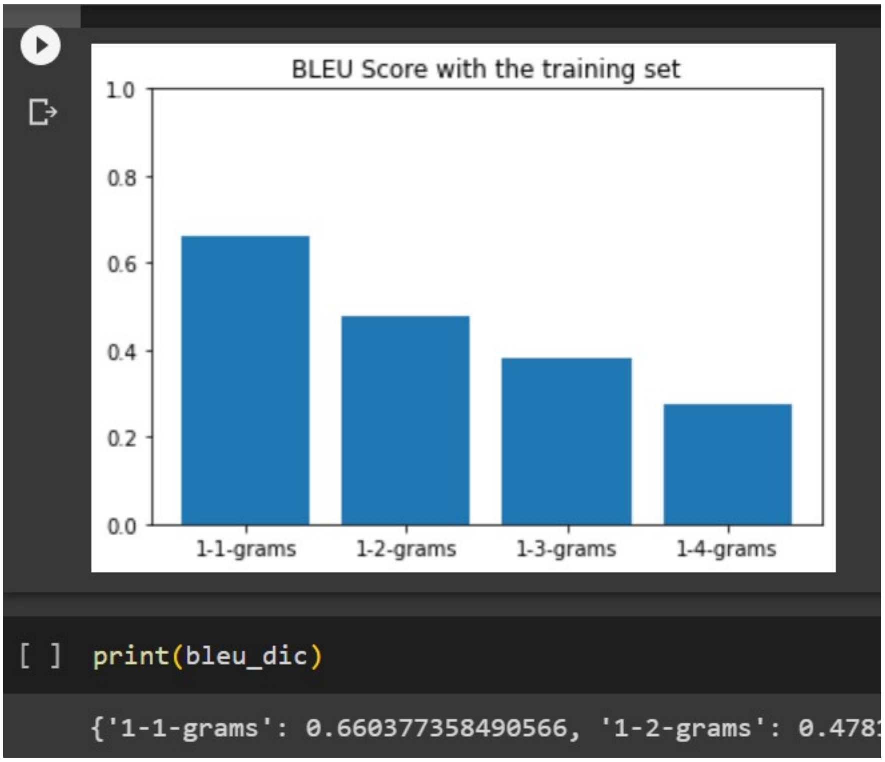


Figure 4.2: Example of the BLEU score of a sample test sentence input

As seen in Table 4.2, the model provided the closest and most accurate translation without compromising on technical words and preserving their meaning according to their context. To strive to achieve further improvement we tested three different values for three different hyperparameters of the model using a manual search based on our judgment. Although there was some improvement from the baseline model, ultimately, the best-performing model was the one fine-tuned using our data set and the default parameters. Given these outcomes, it is clear that NLP-Model-III is the best model to focus on as it translated scientific patent abstracts the most accurately and showed promising results.

Table 4.3 depicts the BLEU score of the post-adapted NLP-Model-III as well as a comparison of the % increase from the original BLEU scores of each pre-trained transformer models chosen for initial evaluation. NLP-Model-I, the implementation of a Japanese-English trans-

Model	Original Text	Backtranslated Text
NLP-Model-III	The present invention provides a carbon dioxide absorbent during combustion of fossil fuels comprising a pressed dry powder of plant fibers, and a fossil fuel characterized by comprising such a pressed dry powder of plant fibers.	The present invention provides a carbon dioxide absorbent at the time of combustion of fossil fuels containing a compressed dry powder of plant fibers , and such a compressed dry powder of plant fibers ."

Table 4.2: NLP-Model-III test results

Model	BLEU	*% increase*
Post-adapted BLEU of NLP-Model-III:	**46.8**	
Base model BLEU of NLP-Model-I:	**41.49**	*+12.8%*
Base model BLEU of NLP-Model-II:	**15.2**	*+207.9%*
Base model BLEU of NLP-Model-III:	**32.7**	*+43.12%*

Table 4.3: A comparison between our post-adapted model and the baseline models of NLP-Model-I [65], NLP-Model-II [59], NLP-Model-III [59]

former by Bharadwaj *et al.* in [65], achieved a BLEU of 41.49, the highest BLEU score among the baseline models. However, our adaptation of the model for the purpose of English-Japanese patent translation (as opposed to general translation) yielded much lower and negligible results as seen in Figure 4.1. NLP-Model-II, pre-trained by the community on HuggingFace [59], achieved a BLEU score of 15.2, which though a good score, fell short of translating highly technical language correctly. NLP-Model-III, also pre-trained on HuggingFace, had the most promising baseline BLEU score results which we were successfully able to increase to 48.8 through data-based domain adaptation techniques.

4.4 Domain adaptation and post-editing

In the previous chapter, we explained the meticulous process of building our bilingual corpus that followed the data-centric domain adaptation technique for fine-tuning and enhancing the performance of our chosen NMT model. Since NLP-Model-III performed the best (before fine-tuning) out of the three models we experimented with as seen in the previous section, we used the corpus to train the model and test it on a subset of our target domain. Due to our computational setup as given in Section 4.1, training the model took approximately 3 hours.

The BLEU score was calculated on the entire test corpus and the final results were sent to a professional Japanese translator to confirm the accuracy of the results.

After testing and evaluating NLP-Model-III, we determined that our proposed technique for machine translation of Japanese patents which encompasses NLP-Model-III performed the best for this particular problem. Post-fine-tuned results also confirmed the improvement. The average BLEU score of the three epochs was determined to be 46.18. This is a 41.22% increase in model performance from the original model BLEU score of 32.7. Table B.1 in Appendix B provides sample results from the post-edited NLP-Model-III experiments and depicts their level of accuracy in translation.

CHAPTER 5

Conclusion

In this thesis, we proposed a comprehensive multi-step technique to work on the problem of English-Japanese patent translation. The unique challenges of patent translation stem from the legal nature of the patent document in comparison to general Japanese-to-English translation and from existing challenges of the English-Japanese language pair which raises the complexity of the models that could be successful. Our technique included prepossessing steps, data preparation and processing, data-centric domain adaptation and fine-tuning, enlisting human expert feedback, and linguistic analysis to refine the machine translation model performance. The results section includes evaluation results of 3 major alternatives for the transformer model to depend on for the last step. The aim was to achieve an output from the models that would fall in the range of 0.5-0.7 of the BLEU score which is the current state of the art. Our technique which encompassed a variation of NLP-Model-III achieved the best performance for the problem at hand reaching a BLEU score of 46.8. It is noticed also that fine-tuning hyperparameters yields up to a 3-point improvement in the BLEU Score. This work included developing a novel dataset that consisted of data collected from patent documents.

This work also examined the domain-specific challenges of patent translation in Japanese from English. Challenges like syntactic ambiguity, out-of-domain data, and others were analyzed and tackled. This would represent a base for future scientific patent translation efforts.

Moreover, the interest in commercial use motivates the further study of MT as the initial findings of this work were demoed in Japan at the PIFC commercial conference where companies tested the performance of the model and were satisfied with the level of translation provided. Due to the results of this research, Xlscout plans on implementing a translation tool to their services to translate patents from English to Japanese. Along with the quality of translation, the speed of translation is also of great importance. Moving forward, it is important to study how to implement data parallelism to use multiple GPUs at once to increase the speed at which translation occurs.

5.1 Future Research Directions and Limitations

One consideration to further examine in the future was the performance speed of the model. For large paragraphs, it would take the model approximately 19 seconds to deliver an output of translated text and sentences took approximately 3 seconds to translate. Because of the specialized nature of the patent translation problem and the fact that it is used by domain-specific enterprise users, there is tolerance for that translation speed level. However, it would enhance the model adoption if we are able to achieve a translation result in milliseconds so that the translated content would be able to load faster on a webpage. Ideas to consider are optimizing hardware performance through GPU parallel processing and experimenting with pre-quantized models. Additionally, experimenting with different domain adaptation techniques and evaluating the effects of the difference in nuances between English and Japanese would be interesting experiments as discussed in [117].

Additionally, the performance of the NMT system should also be good for general sentences and thus extending the corpus to include non-technical sentences and experimenting to improve the model further should also be another next step in enhancing the performance of

NMT systems.

Bibliography

[1] Somers, Harold, Machine Translation: History, Development, and Limitations, in Kirsten Malmkjær, and Kevin Windle (eds), The Oxford Handbook of Translation Studies (2011; online edn, Oxford Academic, 18 Sept. 2012), https://doi.org/10.1093/oxfordhb/9780199239306.013.0029, accessed 9 Jan. 2023.

[2] Weaver, W. (1952). Translation. In Proceedings of the Conference on Mechanical Translation.

[3] Nirenburg, S. (1989). Knowledge-based machine translation. Machine Translation, 4(1), 5–24. https://doi.org/10.1007/bf00367750

[4] Vauquois, B., Veillon, G., & Veyrunes, J. (1965). Syntaxe et interprétation. In COLING 1965.

[5] Vauquois, B. 1968. A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Machine Translation. Proceedings of the IFIP Congress-6. 254–260.

[6] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[7] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[8] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.

[9] Novak, M. (2012, May 29). The Cold War origins of Google Translate. BBC Future. Retrieved January 9, 2023, from https://www.bbc.com/future/article/20120529-a-cold-war-google-translate

[10] Okpor, M.D. (2014). Machine Translation Approaches: Issues and Challenges.

[11] Poibeau, T. (2017). Machine translation. MIT Press.

[12] Garg, A., & Agarwal, M. (2018). Machine translation: a literature review. arXiv preprint arXiv:1901.01122.

[13] Concise history of the language sciences: from the Sumerians to the cognitivists. Edited by E.F.K.Koerner and R.E.Asher. Oxford: Pergamon Press, 1995. Pages 431-445

[14] Tsarfaty, R., Seddah, D., Goldberg, Y., Kübler, S., Versley, Y., Candito, M., ... & Tounsi, L. (2010, June). Statistical parsing of morphologically rich languages (spmrl) what, how and whither. In Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages (pp. 1-12).

[15] Stahlberg, F. (2020). Neural machine translation: A review. Journal of Artificial Intelligence Research, 69, 343-418.

[16] Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. AI Open, 1, 5-21.

[17] Cifka, O. (2018). Continuous Sentence Representations in Neural Machine Translation

[18] Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167).

[19] Yousuf, H., Lahzi, M., Salloum, S. A., & Shaalan, K. (2021). A systematic review on sequence-to-sequence learning with neural network and its models. International Journal of Electrical and Computer Engineering (IJECE), 11(3), 2315. https://doi.org/10.11591/ijece.v11i3.pp2315-2326

[20] Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. Neurocomputing, 452, 48-62.

[21] Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

[22] Brownlee, J. (2019, August 7). How does attention work in encoder-decoder recurrent neural networks. MachineLearningMastery.com. Retrieved January 9, 2023, from https://machinelearningmastery.com/how-does-attention-work-in-encoder-decoder-recurrent-neural-networks/

[23] Compton, E. (2018, June 19). Analyzing Japanese with natural language processing and go. Medium. Retrieved January 9, 2023, from https://medium.com/@enocom/analyzing-japanese-with-natural-language-processing-and-go-64b26cb2436

[24] Nagao, M. (n.d.). A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle. Retrieved January 9, 2023, from https://aclanthology.org/www.mt-archive.info/70/Nagao-1984.pdf

[25] Barchas-Lichtenstein, J. (2016, March 1). Low and high context communication. CultureReady. Retrieved January 9, 2023, from https://www.cultureready.org/blog/low-and-high-context-communication

[26] Shih, J. (2020, October 13). Linguistic differences between Japanese and English. JBI Localization. Retrieved January 9, 2023, from https://jbilocalization.com/linguistic-differences-japanese-english/

[27] Somers, H. (2003). Computers and translation. Computers and Translation, 1-365.

[28] WIPO (2021). World Intellectual Property Indicators 2021. Geneva: World Intellectual Property Organization.

[29] Nitoń-Greczuk, E. (2022, December 16). Things to know about patent translation: Textology. WeAreTextoloxy. Retrieved January 9, 2023, from https://wearetextology.com

[30] Madhavan, R. (2019, November 22). Machine translation - 14 current applications and services. Emerj Artificial Intelligence Research. Retrieved January 10, 2023, from https://emerj.com/ai-sector-overviews/machine-translation-14-current-applications-and-services/

[31] Pricing translator: Microsoft Azure. (n.d.). Retrieved January 10, 2023, from https://azure.microsoft.com/en-us/pricing/details/cognitive-services/translator/#pricing

[32] 2021 top 50 US patent assignees. IFI CLAIMS® Patent Services. (n.d.). Retrieved January 10, 2023, from https://www.ificlaims.com/rankings-top-50-2021.htm

[33] Mehandru, N., Robertson, S., & Salehi, N. (2022). Reliable and safe use of machine translation in medical settings. 2022 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3531146.3533244

[34] Khoong, E. C., Steinbrook, E., Brown, C., & Fernandez, A. (2019). Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. JAMA internal medicine, 179(4), 580-582.

[35] Panayiotou, A., Hwang, K., Williams, S., Chong, T. W., LoGiudice, D., Haralambous, B., Lin, X., Zucchi, E., Mascitti-Meuter, M., Goh, A. M., You, E., & Batchelor, F. (2020).

The perceptions of translation apps for everyday health care in healthcare workers and older people: A multi-method study. Journal of Clinical Nursing, 29(17-18), 3516–3526. https://doi.org/10.1111/jocn.15390

[36] Turner, A. M., Choi, Y. K., Dew, K., Tsai, M.-T., Bosold, A. L., Wu, S., Smith, D., & Meischke, H. (2019). Evaluating the usefulness of Translation Technologies for Emergency Response Communication: A scenario-based study. JMIR Public Health and Surveillance, 5(1). https://doi.org/10.2196/11171

[37] Turner, A. M., Brownstein, M. K., Cole, K., Karasz, H., & Kirchhoff, K. (2015). Modeling workflow to design machine translation applications for public health practice. Journal of biomedical informatics, 53, 136-146.

[38] Benefits and Costs of the European Strategy for Multilingualism. (2016). Retrieved January 10, 2023, from https://www.europarl.europa.eu

[39] Building Global Connections. Teaching for the Asia-Pacific. (n.d.). Retrieved January 10, 2023, from https://teachapac.nz/resources/building-global-connections

[40] Newsmantraa. (2022, June 14). Machine translation market to observe exponential growth by 2022 to 2030: Microsoft Corporation, IBM. Digital Journal. Retrieved January 10, 2023, from https://www.digitaljournal.com/pr/machine-translation-market-to-observe-exponential-growth-by-2022-to-2030-microsoft-corporation-ibm

[41] Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130.

[42] Chatzikoumi. (2020). How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering, 26(2), 137–161. https://doi.org/10.1017/S1351324919000469

[43] Panić, M. (2020, July 22). Automated Mt Evaluation Metrics. TAUS. Retrieved January 11, 2023, from https://www.taus.net/resources/blog/automated-mt-evaluation-metrics

[44] (n.d.). Evaluating models. Google. Retrieved January 11, 2023, from https://cloud.google.com/translate

[45] Dorr, B., Snover, M., & Madnani, N. (2011). Part 5: machine translation evaluation. Handb. Nat. Lang. Process. Mach. Transl. DARPA Glob. Auton. Lang. Exploit, 936.

[46] Doshi, K. (2021, May 11). Foundations of NLP explained. Medium. Retrieved January 11, 2023, from https://towardsdatascience.com

[47] Vashee, K. (2019, April 12). Understanding machine translation Quality: BLEU scores. Understanding MT Quality: BLEU Scores. Retrieved January 11, 2023, from https://www.rws.com/blog/understanding-mt-quality-bleu-scores/

[48] Chauhan, S., Daniel, P. A Comprehensive Survey on Various Fully Automatic Machine Translation Evaluation Metrics. Neural Process Lett (2022). https://doi.org/10.1007/s11063-022-10835-4

[49] Zhang, Y., Vogel, S., & Waibel, A. (2004, May). Interpreting bleu/nist scores: How much improvement do we need to have a better system?. In LREC.

[50] The Trustees of Princeton University. (n.d.). WordNet. Princeton University. Retrieved January 12, 2023, from https://wordnet.princeton.edu/

[51] Poornima, C., Dhanalakshmi, V., Anand, K. M., & Soman, K. P. (2011). Rule based sentence simplification for english to tamil machine translation system. International Journal of Computer Applications, 25(8), 38-42.

[52] Terumasa, E. (2007). Rule based machine translation combined with statistical post editor for japanese to english patent translation. In Proceedings of the Workshop on Patent translation.

[53] Şatır, E., & Bulut, H. (2021, August). A Novel Hybrid Approach to Improve Neural Machine Translation Decoding using Phrase-Based Statistical Machine Translation. In 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) (pp. 1-5). IEEE.

[54] Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2017). Unsupervised neural machine translation. arXiv preprint arXiv:1710.11041.

[55] [dataset] Team AI, 2017. Japanese-English bilingual corpus. Kaggle. Retrieved January 12, 2023, from https://www.kaggle.com/datasets/team-ai/japaneseenglish-bilingual-corpus

[56] [dataset] Tab-delimited Bilingual Sentence Pairs from the Tatoeba Project, (n.d.). Retrieved January 12, 2023, from https://www.manythings.org/anki/

[57] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[58] Neural Machine Translation Background. Tutorial: Neural machine translation - seq2seq. (n.d.). Retrieved January 12, 2023, from https://google.github.io/seq2seq/nmt/

[59] Helsinki-NLP. Opus-tatoeba-en-ja · Hugging Face. (n.d.). Retrieved January 12, 2023, from https://huggingface.co/Helsinki-NLP/opus-tatoeba-en-ja

[60] Stakata-NLP. Fugumt. Hugging Face. (n.d.) Retrieved January 12, 2023, from https://huggingface.co/staka/fugumt-en-ja?text=My+name+is+Sarah+and+I+live+in+London

[61] Tiedemann, J. (2020). The Tatoeba Translation Challenge–Realistic Data Sets for Low Resource and Multilingual MT. arXiv preprint arXiv:2010.06354.

[62] [dataset] The Open Parallel Corpus. (n.d.). Retrieved January 12, 2023, from https://opus.nlpl.eu/

[63] [dataset] Tiedemann, J. (2012, May). Retrieved January 12, 2023, Parallel data, tools, and interfaces in OPUS. (Vol. 2012, pp. 2214-2218).

[64] Li, P. (2013). A survey of machine translation methods. TELKOMNIKA Indonesian Journal of Electrical Engineering, 11(12), 7125-7130.

[65] Bharadwaj, S., Janardhanan , M. T., & Minkuri , P. (2020, December 27). Translation of Japanese to English language. GitHub. Retrieved 2023, from https://github.com/bharadwaj1098

[66] [dataset] Collection of sentences and translations, Tatoeba. (n.d.). Retrieved January 12, 2023, from https://tatoeba.org/en/

[67] [dataset] Japanese-English Subtitle Corpus. JESC. (2019, May 12). Retrieved January 12, 2023, from https://nlp.stanford.edu/projects/jesc/

[68] [dataset] Neubig, G. (n.d.). The Kyoto Free Translation Task (KFTT). The Kyoto Free Translation Task. Retrieved January 12, 2023, from http://www.phontron.com/kftt/

[69] [dataset] Tanaka Corpus - EDRDG Wiki. (n.d.). Retrieved January 12, 2023, from http://www.edrdg.org/wiki/index.php/TanakaCorpus

[70] [dataset] Japanese SNLI dataset. (n.d.). Retrieved January 12, 2023, from https://nlp.ist.i.kyoto-u.ac.jp

[71] [dataset] Facebook research. (2019). GitHub. Retrieved January 12, 2023, from https://github.com/facebookresearch/LASER/tree/main/tasks/WikiMatrix

[72] David. (2022, December 23). Alternative to colab pro: Comparing Google's Jupyter notebooks to gradient notebooks (updated!). Paperspace Blog. Retrieved January 19, 2023, from https://blog.paperspace.com/alternative-to-google-colab-pro/

[73] Bianchi, C. (2020, October 20). Everything you should know about patent translation, the Professional Translation Blog. Language Buró. Retrieved February 8, 2023, from https://languageburo.com/blog/everything-to-know-about-patent-translation

[74] Galvani, T. W. (2019, August 30). Accuracy and precision in patent writing. Thomas W Galvani . Retrieved February 11, 2023, from https://galvanilegal.com/accuracy-and-precision-in-patent-writing/

[75] Bawden, R., Cohen, K. B., Grozea, C., Yepes, A. J., Kittner, M., Krallinger, M., ... & Navarro, M. V. (2019, August). Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2) (pp. 29-53).

[76] Chu, C., & Wang, R. (2018). A survey of domain adaptation for neural machine translation. arXiv preprint arXiv:1806.00258.

[77] Chu, C., Dabre, R., & Kurohashi, S. (2017). An empirical comparison of simple domain adaptation methods for neural machine translation. arXiv preprint arXiv:1701.03214.

[78] International Patent Classification (IPC). (2023). Retrieved 2022, from https://www.wipo.int/classifications/ipc/en/

[79] About CPC. Cooperative Patent Classification. (n.d.). Retrieved April 30, 2023, from https://www.cooperativepatentclassification.org/about

[80] Saunders, D. (2021). Domain adaptation for neural machine translation (Doctoral dissertation, University of Cambridge).

[81] Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

[82] Kaiser, Ł. (2017, June 19). Accelerating deep learning research with the tensor2tensor library. Google AI. https://ai.googleblog.com/2017/06/accelerating-deep-learning-research.html

[83] Dorr, B. J., Hovy, E. H., & Levin, L. S. (2004). Natural language processing and machine translation encyclopedia of language and linguistics, (ell2). machine translation: Interlingual methods. In Proceeding International Conference of the World Congress on Engineering.

[84] Koehn, P. (2020). The Translation Problem. In Neural Machine Translation (pp. 3-18). Cambridge: Cambridge University Press. doi:10.1017/9781108608480.002

[85] Brooks, R. (2020, January 30). The challenges of translating literature. K International. https://www.k-international.com

[86] Most difficult documents to translate. London Translation Agency. (2018, April 25). https://www.languagereach.com

[87] Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. arXiv preprint arXiv:1706.03872.

[88] Bennett, W. S., & Slocum, J. (1985). The LRC machine translation system. Computational linguistics, 11(2-3), 111-121.

[89] Trujillo, A. (1999). Transfer Machine Translation. In: Translation Engines: Techniques for Machine Translation. Applied Computing. Springer, London. https://doi.org/10.1007/978-1-4471-0587-96

[90] Nirenburg, S. (1989). Knowledge-based machine translation. Machine Translation, 4(1), 5-24.

[91] Och, F. J., & Ney, H. (2002, July). Discriminative training and maximum entropy models for statistical machine translation. In Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (pp. 295-302).

[92] Osborne, M. (2011). Statistical Machine Translation. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8783

[93] Udupa, R., & Maji, H. K. (2006, April). Computational complexity of statistical machine translation. In 11th Conference of the European Chapter of the Association for Computational Linguistics (pp. 25-32).

[94] Neves, M. (2015). Lexical Translation and Alignment [PowerPoint slides]. Hasso Plattner Institute. https://hpi.de/fileadmin/userupload/fachgebiete/plattner/teaching/MachineTranslation/MT2015/MT04-LexicalTranslationAlignment.pdf

[95] Z. El Maazouzi, B. E. El Mohajir, and M. Al Achhab, ”A technical reading in statistical and neural machines translation (SMT & NMT),” 2017 8th International Conference on Information Technology (ICIT), Amman, Jordan, 2017, pp. 157-165, doi: 10.1109/ICITECH.2017.8079994.

[96] Salehinejad, H., Sankar, S., Barfett, J., Colak, E., & Valaee, S. (2017). Recent advances in recurrent neural networks. arXiv preprint arXiv:1801.01078.

[97] Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 1017-1024).

[98] Celard, P., Iglesias, E. L., Sorribes-Fdez, J. M., Romero, R., Vieira, A. S., & Borrajo, L. (2023). A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. Neural Computing and Applications, 35(3), 2291-2323.

[99] Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. Towards Data Sci, 6(12), 310-316.

[100] Stephen Grossberg (2013) Recurrent neural networks. Scholarpedia, 8(2):1888.

[101] Q. V. Le, N. Jaitly, and G. E. Hinton, “A simple way to initialize recurrent networks of rectified linear units,” arXiv preprint arXiv:1504.00941, 2015.

[102] Grolinger, K. (2021). Foundations of Data Analytics [PowerPoint slides]. ECE Department, Western University.

[103] Sazli, M. H. (2006). A brief review of feed-forward neural networks. Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering, 50(01).

[104] Bebis, George, and Michael Georgiopoulos. ”Why network size is so important.” Ieee Potentials 13, no. 4 (1994).

[105] Patents. WIPO. (2023). https://www.wipo.int/patents/en/

[106] Rossi, L., & Wiggins, D. (2013). Applicability and application of machine translation quality metrics in the patent field. World Patent Information, 35(2), 115-125.

[107] Larkey, L. S. (1999, August). A patent search and classification system. In Proceedings of the fourth ACM conference on Digital libraries (pp. 179-187).

[108] Kaji, H. (2005). Domain Dependence of Lexical Translation: A Case Study of Patent Abstracts. In Workshop on patent translation (pp. 43-49).

[109] Kumar, A., Pratap, A., Singh, A. K., & Saha, S. (2022). Addressing domain shift in neural machine translation via reinforcement learning. Expert Systems with Applications, 201, 117039.

[110] Pham, M., Crego, J. M., & Yvon, F. (2021). Revisiting multi-domain machine translation. Transactions of the Association for Computational Linguistics, 9, 17-35.

[111] Ramponi, A., & Plank, B. (2020). Neural unsupervised domain adaptation in NLP—a survey. arXiv preprint arXiv:2006.00632.

[112] Kundu, R. (2022). Domain adaptation in Computer Vision: Everything you need to know. V7. https://www.v7labs.com/blog/domain-adaptation-guide

[113] Kouw, W. M., Van Der Maaten, L. J., Krijthe, J. H., & Loog, M. (2016). Feature-level domain adaptation. The Journal of Machine Learning Research, 17(1), 5943-5974.

[114] van der Meer, A.-M. (2022, December 19). Domain adaptation: Types and methods. TAUS Language Data Solutions. https://www.taus.net

[115] Wotherspoon, S., Hartmann, W., Snover, M., & Kimball, O. (2021, June). Improved data selection for domain adaptation in asr. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7018-7022). IEEE.

[116] Zhang, K., Gong, M., & Schölkopf, B. (2015, February). Multi-source domain adaptation: A causal view. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1).

[117] Saunders, D. (2022). Domain adaptation and multi-domain adaptation for Neural Machine Translation: A survey. Journal of Artificial Intelligence Research, 75, 351–424. https://doi.org/10.1613/jair.1.13566

[118] Hatim, B., & Munday, J. (2004). Translation: An advanced resource book. Psychology Press.

[119] WMT. Machine Translate. (n.d.). https://machinetranslate.org/wmt#shared-tasks

[120] Patent FAQ (Ca/US/PCT). Heer Law. (n.d.). https://www.heerlaw.com/patent-faq

[121] JPO Patent Corpus for Wat2018-2021. JPO. (n.d.). http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/jpc2018.html

[122] Monck, S. (2014). IPC - International Patent Classification Basics. SlidePlayer. https://slideplayer.com/slide/1688565/

[123] Keita, Z. (2022, November 21). Data Augmentation in NLP using back translation with marianmt. Medium. https://towardsdatascience.com/data-augmentation-in-nlp-using-back-translation-with-marianmt-a8939dfea50a

[124] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, Alexandra Birch; Survey of Low-Resource Machine Translation. Computational Linguistics 2022; 48(3) : 673–732.*doi* : *https* : //*doi.org*/10.1162/*colia*00446

[125] Vashee, K. (2022, April 21). Understanding MT Quality: BLEU Scores. ModernMT Blog. https://blog.modernmt.com/understanding-mt-quality-bleu-scores

[126] Mjpost. (n.d.). Sacrebleu. GitHub. https://github.com/mjpost/sacrebleu

[127] Can chat GPT translate text and languages?. Can Chat GPT Translate Text and Languages? (2023, April 17). https://www.awesomescreenshot.com/blog/knowledge/can-chat-gpt-translate

[128] Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

[129] Tien, H. N., & Minh, H. N. T. (2019, March). Long sentence preprocessing in neural machine translation. In 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF) (pp. 1-6). IEEE.

APPENDIX A

Subset of English-Japanese Bilingual Parallel Corpus (Section 3.2.4)

Table A.1: This table presents a sample of the English-Japanese bilingual parallel corpus created to enhance the performance of our adapted and fine-tuned NMT model.

English	Japanese
The modulation units (101-1 to 101-L) modulate the multicarrier signals included in each block using a different modulation scheme for each block selected by the allocation unit (108)	） から 入力 さ れる ブロック 毎 の ＳＮＲ の 平均 と ＳＮＲ の 分散 に 基づい て 、 ブロック 毎 に 変調 方式 を 選択 し 、
Passivation films 3a and 3b are formed so as to cover both surfaces of the semiconductor substrate 1 having the terminal pads 2a and 2b on both surfaces	両面 に 端子 パッド ２ ａ ， ２ ｂ を 有する 半 導体 基板 １ の 両面 を 覆う ように パッシベーション 膜 ３ ａ ， ３ ｂ
Openings 3c and 3d are provided in positions where the passivation films 3a and 3b overlap the terminal pads 2a and 2b	が 形成 さ れ て いる この パッシベーション 膜 ３ ａ ， ３ ｂ の 、 端子 パッド ２ ａ ， ２
A through hole 9 is formed inside the openings 3c and 3d so as to penetrate the terminal pad 2a, the semiconductor substrate 1, and the terminal pad 2b	ｂ と 重なる 位置 に 、 開口 部 ３ ｃ ， ３ ｄ が 設け られ て いる 開口 部 ３ ｃ ， ３ ｄ の 内側 に 、
An insulating layer 4 made of SiO 2, SiN, SiO, or the like is formed on the inner surface of the through hole 9	端子 パッド ２ ａ と 半 導体 基板 １ と 端子 パッド ２ ｂ を 貫通 する 貫通 孔 ９ が 形成 さ れ

Table A.1 continued from previous page

English	Japanese
A buffer layer 5 made of a conductive adhesive is formed so as to cover the insulating layer 4 and the terminal pads 2a and 2b in the openings 3c and 3d	て いる 貫通 孔 ９ の 内面 に 、 Ｓｉ Ｏ ２ 、 Ｓｉ Ｎ 、 また は Ｓｉ Ｏ 等 から なる 絶縁 層 ４ が 形成 さ れ て いる 絶縁 層 ４
Furthermore, a conductive layer 6 made of a metal film is formed on the buffer layer 5 by electrolytic plating or electroless plating	と 、 開口 部 ３ ｃ ， ３ ｄ 内 の 端子 パッド ２ ａ ， ２ ｂ と を 覆う よう に
An object is to provide a semiconductor device including an oxide semiconductor with stable electrical characteristics and high reliability	課題 酸化 物 半 導体 を 用い た 半 導体 装置 に 安定 し た 電気 的 特性 を
In this method, a second insulating film is formed over an oxide semiconductor film supplied with oxygen atoms, and a gate electrode is formed in a region overlapping with the oxide semiconductor film over the second insulating film	原子 を 除去 し 、 水素 原子 が 除去 さ れ た 酸化 物 半 導体 膜 に 酸素 ドープ 処理 を 行っ て 、 酸化 物 半 導体 膜 中 に 酸素 原子 を 供給 し 、 酸素
[Selection] Figure 2	原子 が 供給

Table A.1 continued from previous page

English	Japanese
A first insulating film is formed, a source electrode and a drain electrode, and an oxide semiconductor film electrically connected to the source electrode and the drain electrode are formed over the first insulating film, and the oxide Heat treatment is performed on the semiconductor film to remove hydrogen atoms in the oxide semiconductor film, oxygen doping treatment is performed on the oxide semiconductor film from which hydrogen atoms are removed, and oxygen atoms are supplied into the oxide semiconductor film	付与 し 、 高 信頼 性 化 する こと を 目的 の 一 と する 解決 手段 第 1 の 絶縁 膜 を 形成 し 、 第 1 の 絶縁 膜上 に 、 ソース 電極 および ドレ イン 電極 、 ならび に 、 ソース 電極 および ドレ イン 電極 と 電気 的 に 接続 する 酸化 物 半 導体 膜 を 形成 し 、 酸化 物 半 導体 膜 に 熱 処理 を 行っ て 、 酸化 物 半 導体 膜中 の 水素
PROBLEM TO BE SOLVED: To provide a semiconductor device having a structure that can be formed on-chip with other elements, capable of controlling a large current even though the element area is small, and having a small on-resistance and a high breakdown voltage, and a manufacturing method thereof	課題 他 の 素子 と オン チップ 化 できる 構造 で あり ， 素子 面積 が 小さい に も かかわら ず 大 電流 を 制御 でき ， オン 抵抗 が 小さく 高 耐圧 を 可能 に する 半 導体 装置 及び その 製造 方法 を 提供 する 解決

Table A.1 continued from previous page

English	Japanese
In the case of an N-type LDMOS, an N-well layer 102 formed on a P-type semiconductor substrate 101, a P-well layer 103 formed in the N-well layer 102, and a source trench hole in the P-well layer 103 Source electrode 107a formed in 105a, gate electrode 107b formed in at least one gate trench hole 105b in P well layer 103 through oxide film 106, and formed in drain trench hole 105c in N well layer 102 The N + diffusion layers 108a and 108c are formed around the source trench hole 105a and the drain trench hole 105c	手段 Ｎ 型 の ＬＤＭＯＳ の 場合 に は ， Ｐ 型 半導体 基板 １ ０ １ に 形成 し た Ｎ ウェル 層 １ ０ ２ と ， Ｎ ウェル 層 １ ０ ２ 内 に 形成 し た Ｐ ウェル 層 １ ０ ３ と ， Ｐ ウェル 層 １ ０ ３ 内 の ソーストレンチ 孔 １ ０ ５ ａ に 形成 さ れ た ソース 電極 １ ０ ７ ａ と ， Ｐ ウェル 層 １ ０ ３ 内 の 少なく とも １ つ の ゲート トレンチ 孔 １ ０ ５ ｂ に 酸化 膜 １ ０ ６ を 介し て 形成 さ れ た ゲート 電極 １ ０ ７ ｂ と ， Ｎ ウェル
[Selection] Figure 1 (a)	層 １ ０ ２ 内 の
PROBLEM TO BE SOLVED: To provide a circuit board comprising a via structure with a fine pitch, and a method for manufacturing the same	課題 微細 ピッチ の ビア 構造 物 を 備える 回路 基板 及び その 製造 方法 を 提供 する 解決 手段 本 発明 の 回路

Table A.1 continued from previous page

English	Japanese
SOLUTION: A circuit board 100 of the present invention includes: a base substrate 110; an interlayer insulating layer 120 covering the base substrate 100; a via structure 140 passing in the vertical direction through at least the interlayer insulating layer 120 among the base substrate 110 and the interlayer insulating layer 120; and an etch stop pattern 130 disposed on the interlayer insulating layer 120 in the horizontal direction, which is perpendicular to the vertical direction, to surround the via structure 140 and made of an insulating material	基板 １ ０ ０ は 、 ベース 基板 １ １ ０ と 、 ベース 基板 １ ０ ０ を 覆う 層間 絶縁 膜 １ ２ ０ と 、 ベース 基板 １ １ ０ 及び 層間 絶縁 膜 １ ２ ０ の うち の 少なく とも 層間 絶縁 膜 １ ２ ０ を 上下 方向 に 貫く ビア 構造 物 １ ４ ０ と 、 層間 絶縁 膜 １ ２ ０ 上 に 上下 方向 に 鉛直 な 水平 方向 に 配置 さ れ 、 ビア 構造 物 １ ４ ０ を 取り囲み 、 絶縁 材料 から 成る エッチング 防止 パターン １ ３ ０ と を 含む 選択 図 図 １
The present technology relates to a semiconductor device and electronic equipment that can suppress the generation of noise in a signal	本 技術 は 、 信号 に おける ノイズ の 発生 を 抑制 する こと が できる よう に する 半
A semiconductor device includes a first semiconductor substrate on which at least a part of a first conductor loop is formed, and a first conductor layer having a conductor and a second conductor layer forming a second conductor loop	導体 装置 および 電子 機器 に 関する 半 導体 装置 は 、 第 １ の 導体 ループ の 少なく とも 一部 が 形成 さ れる 第 １ の 半 導体 基板 と 、 第 ２ の 導体 ループ を

Table A.1 continued from previous page

English	Japanese
A second semiconductor substrate, wherein the first conductor layer and the second conductor layer have a loop surface direction in which a magnetic flux is generated from the second conductor loop, and a loop which generates an induced electromotive force in the first conductor loop	形成 する 、 導体 を 有する 第 1 の 導体 層 及び 第 2 の 導体 層 を 含む 第 2 の 半 導体 基板 と を 備え 、 第 1 の 導体 層 と 第 2 の 導体 層 は 、 第 2 の
The direction of the plane is different from that of the plane	導体 ループ から 磁束 が 発生 する ループ 面 の 方向 と
The present technology can be applied to, for example, a CMOS image sensor	、 第 1 の 導体 ループ に 誘導 起電 力 を 発生 さ
This disclosure describes methods and structures for three-dimensional memory devices	本 開示 は 、 三 次元 メモリ デバイス の ため
The method includes providing a bottom substrate and forming a plurality of doped layers on the bottom substrate	の 方法 および 構造 を 説明 する この 方法 は 、 底部 基板 を 提供 する こと と

Table A.1 continued from previous page

English	Japanese
The plurality of doped layers have a top surface of the plurality of doped layers substantially flat, and a doping concentration of each of the plurality of doped layers is substantially along a direction substantially perpendicular to the top surface of the plurality of doped layers	、底部 基板 上 に 複数 の ドープ 層 を 形成 する こと と を 含む 複数 の ドープ 層 は 、 複数 の ドープ 層 の 上面 が 実質 的 に 平坦 で あり 、 かつ 複数 の ドープ 層 の 各々 の ドーピング 濃度
Having a total thickness within the thickness range that is uniform	が 複数 の ドープ 層 の 上面 に 実質 的 に
<P>PROBLEM TO BE SOLVED: To give a high heat radiating property to a semiconductor device having a plurality of laminated semiconductor elements with respect to the heat generated from the semiconductor elements when the elements are operated	課題 積層 した 複数 の 半導体 素子 を 有する 半導体 装置 に おいて 、 半導体 素子 の 動作 時 の 発熱 に 対して 、 高い 放熱 性 を 持たせる 解決
<P>SOLUTION: The semiconductor device has the plurality of laminated semiconductor elements 2 and resin films 3 provided among the semiconductor elements 2 and having high water absorption	手段 積層 した 複数 の 半導体 素子 ２ と 、 各 半導体 素子 間 に 設け られ た 高 吸水 性 樹脂 膜 ３

Table A.1 continued from previous page

English	Japanese
It is preferable that the resin films 3 contain water or a low-melting point organic solvent	とを 有する ここ で 、 高 吸水 性 樹脂 膜 3 は 、 水 又
Alternatively, the resin film 3 contains an organic solvent having a boiling point higher than the reflow temperature of solder or is made to contain the organic solvent after mounting	は 低 沸点 の 有機 溶媒 を 含ん でいる こと が 好ましい あるいは 、 高 吸水 性 樹脂 膜 は 、 半田 の リフロー 温度 以上 の 沸点 を
<P>COPYRIGHT: (C)2005,JPO&NCIPI	有する 有機
To provide a semiconductor device capable of suppressing injection of hot carriers into a gate insulation layer in a simple manufacturing process and improving a high off voltage, and a manufacturing method for the same	課題 簡易 な 製造 工程 で 、 ゲート 絶縁 層 へ の ホット キャリア の 注入 を 抑制 でき 、 かつ オフ 耐圧 を 向上 可能 な 半 導体 装置 および その 製造 方法 を
SOLUTION: In a plan view, a first comb section of an n-type well region NWL and a second comb section of a pdrift region DFT engage with each other	提供 する 解決 手段 平面 視 に おいて 、 ｎ 型 ウエル 領域 ＮＷＬ の 第 １ 櫛部 と ｐ - ドリフト 領域 ＤＦＴ の 第 ２ 櫛部
Therefore in a plan view, a p-n junction of the n-type well region NWL and the pdrift region DFT has a zigzag configuration	とは 互いに 噛み合っ ている これ により 平面 視 に おいて 、 ｎ 型 ウエル 領域 ＮＷＬ と ｐ

Table A.1 continued from previous page

English	Japanese
The p-n junction configured by the n-type well region NWL and the pdrift region DFT extends from a main surface MS to a bottom face BWS of an isolation groove TNC along a source side sidewall SWS of the isolation groove TNC	-ドリフト領域ＤＦＴとのｐｎ接合はジグザグ形状を有しているｎ型ウエル領域ＮＷＬとｐ-ドリフト領域ＤＦＴとにより構成されるｐｎ接合は、分離溝ＴＮＣのソース側
SELECTED DRAWING: Figure 6	壁面ＳＷＳに沿っ
A semiconductor layer on which a semiconductor element is formed, a first conductor film formed on the upper surface of the semiconductor layer and electrically connected to the semiconductor element, and formed on a side surface of the semiconductor layer, and electrically connected to the semiconductor element A second conductor film connected to the second conductor film; and a first protective film formed on the first conductor film and having an opening exposing the first conductor film	半導体素子が形成された半導体層と、半導体層の上面上に形成され、半導体素子と電気的に接続された第１の導体膜と、半導体層の側面上に形成され、半導体素子と電気的に接続された第２の導体膜と、第１の導体膜上に

Table A.1 continued from previous page

English	Japanese
The semiconductor device has a height up to the top surface of the film that is the same as or lower than a height from the top surface of the semiconductor layer to the top surface of the first conductor film	形成 さ れ 、 第 １ の 導体 膜 を 露出 する 開口 部 を 有す る 第 １ の 保護 膜 と を 備え 、 半 導体 層 の 上面 から 第 ２ の 導体 膜 の 上面 まで の 高
PROBLEM TO BE SOLVED: To provide a method that controls a base station in a cellular wireless communications network and comprises within the base station, autonomously and dynamically adapting a maximum value for a total transmit power of the base station, such that interference between the base station and other access points in the vicinity is minimized	課題 セルラー 式 無線 通信 ネットワーク に おけ る 基地 局 を 制御 する 方法 で あっ て 、 基地 局 と 近く の 他 の アクセス ポイント と の 間 の 干渉 が 最低 限 に 抑え られ る よう に 、 基地 局 内 で 、 基地 局 の 総 送信 電力 の 最 大 値 を
SOLUTION: The base station selects a carrier frequency and a scrambling code from lists provided from a management system that generally controls the operation of the femtocell base stations in the network	自律 的 かつ 動的 に 適応 さ せる こと を 含む 方法 を 提供 する 解決 手段 ネットワーク に おけ る フェムト セル 基 地 局 の 動作 を 一般 に 制御 する

Table A.1 continued from previous page

English	Japanese
The carrier frequencies and the scrambling codes on the lists are shared with other base stations in the network, including nodeBs of the macro layer and other femtocell base stations	管理 システム から 提供 さ れ た リスト から 、 搬送 波 周 波 数 及び スクランブル コード を 選択 する リスト に お け る 搬送 波 周波 数 及び スクランブル
In response to an error condition related to the radio environment, a message may be sent to the user of the base station, requesting that the base station be repositioned	コード は 、 マクロ レ イヤ の ｎｏｄｅＢ 及び 他 の フェ ムト セル 基地 局 を 含め て 、 ネットワーク に おけ る 他 の 基地 局 と 共有 さ れる
The image data corresponding to each of the left and right view-points is thinned out by the thinning unit (101)	左右 の 視点 の 各々 に 対応 し た 画像 データ を 間引き 部 （ １０１ ） で 間引く 合成
When synthesizing the thinned image data, the synthesis method selection unit (104) selects a synthesis method that minimizes the discontinuity in the boundary portion of the synthesized image	方法 選択 部 （ １０４ ） が 間引か れ た 画像 データ を 合 成 する 場合 に 合成 画像 の 境界 部 の 不 連続 性 が 最も
A combining unit (102) combines the plurality of image data us-ing the selected combining method	小さく なる 合成 方法 を 選択 する 合成 部 （ １０２ ） は 、 選択
The encoding unit (103) encodes the combined image data, and the combining method encoding unit (105) encodes information on the combining method	し た 合成 方法 を 用い て 前記 複数 の 画像 データ を 合成 する 符号 化 部 （ １０３ ） は

Table A.1 continued from previous page

English	Japanese
A multiplexing unit (106) multiplexes these encoded data	合成 さ れ た 画像 データ を 符号
In this way, the continuity of the composite image is increased and the encoding efficiency is increased	化 し 、 合成 方法 符号 化 部 （ １ ０ ５ ） は 合成 方法 の 情報 を
PROBLEM TO BE SOLVED: To provide a semiconductor device which can prevent generation of voids at a through electrode and which has higher reliability than in the past; and provide a manufacturing method of the semiconductor device and provide an electronic component	課題 貫通 電極 に おけ る ボイド の 発生 を 防止 する こと が でき 、 従来 に 比べ て 信頼 性 の 高い 半 導体 装置 および その 製造 方法 、 ならびに 電子 部品 を 提供 する こと 解決 手段

Table A.1 continued from previous page

English	Japanese
SOLUTION: A semiconductor manufacturing method comprises: forming an electrode layer 51 on a gate insulation film 30 on a Si substrate 29; forming an interlayer insulation film 31 on the gate insulation film 30, and subsequently forming a lower pad 40 including lower wiring 42 of the same pattern with the electrode layer 51 and a lower side insulation film 43 of a negative pattern by a damascene method; subsequently, forming a through hole 59 and simultaneously exposing, in the through hole 59, a first interlayer insulation film 32 on which projections 60 of the same pattern with the lower side insulation film 43 are formed; forming a via insulation film 38 after etching the first interlayer insulation film 32 so as to leave a part of the projections 60 as an etching residue, and etching the via insulation film 38 at a bottom surface of the through hole 59; and forming a through electrode 17 by growing an electrode material by plating on an inner side of the via insulation film 38 of the through hole 59	Ｓｉ 基板 ２９ 上 の ゲート 絶縁 膜 ３０ 上 に 電極 層 ５１ を 形成 する ゲート 絶縁 膜 ３０ 上 に 層間 絶縁 膜 ３１ を 形成 し た 後 、 ダマシン 法 に より 電極 層 ５１ と 同一 パターン の 下側 配線 ４２ と 、 反対 パターン の 下側 絶縁 膜 ４３ を 含む 下側 パッド ４０ を 形成 する 次に 、 貫通 孔 ５９ を 形成 し 、 同時 に 、 貫通 孔 ５９ 内 に 下側 絶縁 膜 ４３ と 同一 パターン の 突出 部 ６０ が 形成 さ れ た 第 １ 層間 絶縁 膜 ３２ を 露出 さ せる そして 、 突出 部 ６０ の 一部 が エッチング 残渣 と し て 残る よう に 第 １ 層間 絶縁 膜 ３２ を エッチング し た 後 、 ビア 絶縁 膜 ３８ を 形成 し 、 貫通 孔 ５９ の 底面 の ビア 絶縁 膜 ３８ を エッチング する 次に 、 貫通 孔 ５９ の ビア 絶縁 膜 ３８ の 内側 に 電極 材料 を めっき 成長 さ

Table A.1 continued from previous page

English	Japanese
A wireless communication method and apparatus for selecting and reselecting cells used by a wireless transmit / receive unit (WTRU) in a wireless multi-cell communication system	無線 マルチ セル 通信 システム に おける 無線 送受信 ユニット （ＷＴＲＵ） に よって 使用 さ れる 、 セル を 選択 し 、
The WTRU includes a switched beam antenna configured to form a plurality of directional beam patterns and omnidirectional beam patterns	および 再 選択 する ため の 無線 通信 方法 および 装置 ＷＴＲＵ は 、 複数 の 指向 性 ビーム パターン
The WTRU measures signals from multiple cells using directional and omnidirectional beam patterns	および 無 指向 性 ビーム パターン を 形成 する ように 構成 さ
The WTRU selects the cell with the strongest signal and registers with that cell	れ た 切り替え ビーム アンテナ を 含む ＷＴＲＵ は 、 指向 性 ビーム パターン
In one embodiment, the WTRU selects the directional beam with the strongest signal, uses that beam as the active beam, and communicates with the selected cell	および 無 指向 性 ビーム パターン を 用い て 複数 の セル から の 信号 を 測定 する ＷＴＲＵ は 、 最強 の 信号 を 有する
In another embodiment, the WTRU selects a cell and beam combination and registers with the selected cell using the selected beam	セル を 選択 し 、 その セル に 登録 する 一 実施 形態 では 、 ＷＴＲＵ は 、 最強 の

Table A.1 continued from previous page

English	Japanese
In yet another embodiment, the WTRU initiates a handoff to a neighboring cell that has better signal measurement results than the selected cell	信号 を 有する 指向 性 ビーム を 選択 し 、 その ビーム を アクティブ ビーム と し て 使用 し 、 選択 セル
A semiconductor device with good reliability is provided	良好 な 信頼 性 を 有する 半 導体
A first insulator, a second insulator, and a third insulator are formed over the first conductor, and microwave-excited plasma treatment is performed on the third insulator to form an island-shaped first oxidation A first semiconductor, a second conductor on the first oxide semiconductor, and a third conductor, and the first oxide semiconductor, the second conductor, and the third conductor	装置 を 提供 する 第 1 の 導電 体上 に 第 1 の 絶縁 体 、 第 2 の 絶縁 体 、 及び 第 3 の 絶縁 体 を 形成 し 、 第 3 の 絶縁 体 に マイクロ 波 励起 プラズマ 処理 を 行い 、 島 状 の 第 1 の 酸化 物 半 導体 と 、 第
Then, an oxide semiconductor film, a first insulating film, and a conductive film are formed, a part of the first insulating film and the conductive film is removed, and a fourth insulator and a fourth conductor are formed	1 の 酸化 物 半 導体 上 の 第 2 の 導電 体 、 及び 第 3 の 導電 体 と 、 を 形成 し 、 第 1 の 酸化 物 半 導体 、 第 2 の 導電

Table A.1 continued from previous page

English	Japanese
A second insulating film is formed so as to cover the oxide semiconductor film, the fourth insulator, and the fourth conductor, and the oxide semiconductor film and the second insulating film And the second oxide semiconductor and the fifth insulator are formed to expose the side surface of the first oxide semiconductor and to contact the side surface and the side surface of the second oxide semiconductor	体 、 及び 第 ３ の 導電 体上 に 、 酸化 物 半 導体 膜 、 第 １ の 絶縁 膜 、 及び 導電 膜 を 形成 し 、 第 １ の 絶縁 膜 、 及び 導電 膜 の 一部 を 除去 し 、 第 ４ の 絶縁 体 、 及び 第 ４ の 導電 体 を 形成 し 、 酸化 物 半 導体 膜 と 、
, Forming a sixth insulator in contact with the sixth insulator, forming a seventh insulation, heat treatment is performed	第 ４ の 絶縁 体 と 、 第 ４ の 導電 体 と 、 を 覆う ように 、
An integrated circuit package substrate, such as a package substrate or an interposer substrate, wherein the capacitor structure is formed on a sintered ceramic base structure	課題 パッケージ 基板 または インターポーザ 基板 などの 基板 などの 集積 回路 パッケージ の 基板 で 、 焼結 セラミック ベース 構造 の 上 に
A base structure 12 is formed from a raw material 12 having a plurality of via openings 22P, 22G, 22S therein	キャパシタ 構造 が 形成 さ れ た 集積 回路 パッケージ の 基板 を 提供 する 解決 手段 ベース 構造 １ ２ は

Table A.1 continued from previous page

English	Japanese
The green material becomes a sintered ceramic material and is sintered such that a base structure with a plurality of via openings becomes a sintered ceramic base structure	、複数 の ビア 開口 ２２Ｐ、２２Ｇ、２２Ｓ を そこ に 備える 未 加工 の 物質 １２ から 形成 さ れる 未 加工
Conductive vias 14P, 14G, 14S are formed in respective via openings of the sintered ceramic base structure	の 物質 は 、 焼結 セラミック 物質 に なり および 複数 の ビア 開口 を 備える ベース
Capacitor structure 16 is formed on a sintered ceramic base structure	構造 が 焼結 セラミック ベース 構造 に なる よう に 焼結
The plurality of power sources 24P and the ground plane 26 having a capacitor structure are connected to a plurality of vias	さ れる 導電 性 ビア １４Ｐ、１４Ｇ、１４Ｓ は 、 焼結 セラミック ベース 構造 の それぞれ の
Multiple via openings can be connected to multiple vias without the need to drill through multiple brittle substrates such as multiple silicon substrates	ビア 開口 の 中 に 形成 さ れる キャパシタ 構造 １６ は 、 焼結 セラミック ベース 構造 の 上 に 形成 さ れる
[Selection] Figure 1	キャパシタ 構造 の

Table A.1 continued from previous page

English	Japanese
In the wireless network including the first coordinator and at least one device, the channel change method of the first device searches for availability of a channel other than the first channel being used in the wireless network; As a result of the search, changing the first channel to a second channel among at least one usable channel, transferring data to the second device through the second channel, and transferring data from the second device Receiving	第 １ 調整 器 と 少なく とも 一 つ の デバイス と を 含ん で なる 無線 ネットワーク に おい て 第 １ デバイス の チャネル 変更 方法 は 、 前記 無線 ネットワーク で 使用 中 の 第 １ チャネル 以外 の 他 の チャネル の 使用 可否 を 探索 する 段階 と 、 前記 探索 結果 、 使用 可能 な 少なく とも 一 つ の チャネル の うち の 第 ２ チャネル へ と 前記
Enhancement mode GaN MOSFET (100) is formed by using Al-GaN (or InAlGaN) barrier layer (118) on SiO 2 / Si 3 N 4 gate insulation layer (124)	エンハンスメント・モード ＧａＮ ＭＯＳＦＥＴ （ １ ０ ０ ） が 、 ＡｌＧａＮ （ 又 は ＩｎＡｌＧａＮ ） 障壁 層 （ １ １ ８ ） 上 の ＳｉＯ ２ ／ Ｓｉ ３ Ｎ
The Si 3 N 4 portion (120) of the SiO 2 / Si 3 N 4 gate insulating layer (124) reduces the formation of interface states at the junction between the gate insulating layer (124) and the barrier layer (118)	４ ゲート 絶縁 層 （ １ ２ ４ ） を 用い て 形成 さ れる Ｓｉ Ｏ ２ ／ Ｓｉ ３ Ｎ ４ ゲート 絶縁 層 （ １ ２ ４ ） の Ｓｉ ３ Ｎ ４ 部分 （ １ ２ ０ ） は 、 ゲート 絶縁 層
The SiO 2 portion (122) of the SiO 2 / Si 3 N 4 gate insulating layer (124) significantly reduces the leakage current	（ １ ２ ４ ） と 障壁 層 （ １ １ ８ ） と の 間 の 接合 で の 界面 準位 の 形成 を 低減 さ
[PROBLEMS] To improve reliability in a semiconductor device	課題 半 導体 装置 に おい て 、

Table A.1 continued from previous page

English	Japanese
A plurality of substrates each having a semiconductor substrate on which a circuit having a predetermined function is formed, and a multilayer wiring layer laminated on the semiconductor substrate are stacked	信頼 性 を より 向上 さ せる こと を 可能 に する 解決 手段 所定 の 機能 を 有する 回路 が 形成 さ れ た 半 導体 基板 と 、 前記
A structure for electrically connecting the at least two substrates to each other, wherein the electrodes formed on the bonding surface are bonded in a state of being in direct contact with each other	半 導体 基板 上 に 積層 さ れる 多層 配線 層 と 、 を それ ぞれ 有する 複数 の 基板 が 積層 さ れ て 構成 さ れ 、 前記 複数 の 基板 の うち
And an electrode forming the electrode bonding structure and / or a via for connecting the electrode to a wiring in the multilayer wiring layer on at least one of the two substrates	の 少なく とも ２ つ の 基板 間 の 貼り 合わせ 面 に は 、 当該 ２ つ の 基板 間 を 電気 的 に 接続 する ため の 構造 で あっ て 、
The present invention provides a semiconductor device having a structure in which a protective film for preventing diffusion of the conductive material is embedded in a conductive material forming the electrodes and the vias	前記 貼り 合わせ 面 に それぞれ 形成 さ れる 電極 同士 が 直接 接触 し た 状態 で 接合 し て いる 電極 接合 構造 が 存在 し 、 前記 ２ つ の 基板 の
[Selection diagram] FIG	の 少なく とも

Table A.1 continued from previous page

English	Japanese
To suppress a voltage of a fuel cell from being negative during rapid warm-up operation	課題 急速 暖機 運 中 に 燃料 電池 の 電圧 が 負 電圧 に なる
SOLUTION: A fuel cell system 100 comprises: a fuel cell 10 that generates power by an electrochemical reaction of a fuel gas and an oxidant gas; and a control device 200	の を 抑制 する 解決 手段 燃料 電池 システム １ ０ ０ は 、 燃料 ガス と 酸化 剤 ガス と の 電気 化学 反応 に より 電力 を 発生 する 燃料 電池
The control device 200 comprises a low-efficiency power generation execution unit that performs low-efficiency power generation with a larger power generation loss compared with normal power generation	１ ０ と 、 制御 装置 ２ ０ ０ と 、 を 備える 制御 装置 ２ ０ ０ は 、 通常 発電 と 比べ て 発電 損失 の 大きい 低 効率 発電

Table A.1 continued from previous page

English	Japanese
The low-efficiency power generation execution unit is configured to make the fuel cell 10 generate power so that a calorific value of the fuel cell 10 caused by the power generation loss becomes a first calorific value when a temperature of the fuel cell 10 at power generation start of the fuel cell 10 is less than a reference temperature, and to make the fuel cell 10 generate power so that the calorific value becomes a second calorific value larger than the first calorific value when a current integrated value during a time period when the fuel cell 10 is made to generate power so that the calorific value becomes the first calorific value becomes equal to or more than a predetermined integrated value	を 実施 する 低 効率 発電 実施 部 を 備える 低 効率 発電 実施 部 は 、 燃料 電池 １ ０ の 発電 開始 時 に おける 燃料 電池 １ ０ の 温度 が 基準 温度 未満 の とき は 、 発電 損失 に 伴う 燃料 電池 １ ０ の 発熱 量 が 第 １ 発 熱量 と なる ように 燃料 電池 １ ０ を 発電 さ せ 、 発熱 量 が 第 １ 発 熱量 と なる ように 燃料 電池 １ ０ を 発電 さ せて いる 期間 の 電流 積算 値 が 所定 積算 値 以上 に なった とき は 、 発熱 量 が 第 １ 発 熱量 より も 大きい 第 ２ 発熱 量 と なる ように 燃料 電池 １ ０
SELECTED DRAWING: Figure 1	を 発電 さ せる

Table A.1 continued from previous page

English	Japanese
<P>PROBLEM TO BE SOLVED: To provide an electronic component mounting structure wherein an electronic component is embedded with face up in an insulation film on a wiring board, and via holes are formed on connection pads of the electronic component without causing any defects	課題 配線 基板 上 に 電子 部品 が 絶縁 膜 内 に 埋設 さ れ 、 かつ フェイスアップ で 実装 さ れ た 電子 部品 実装 構造 に おい て 、 何 ら 不具合 が 発生 する こと なく 、 電子 部品
<P>SOLUTION: The electronic component mounting structure comprises a body 26a to be mounted with the electronic component 20, the electronic component 20 mounted on the body 26a with the connection pads 18 of the electronic component 20 which have an etch stop layer 16 (a copper film, gold film, silver film, or conductive paste film) on the top with the pads 18 being faced up, interlayer insulation film 28a which covers the electronic component 20, the via holes 28y formed in the interlayer insulation film 28a on the connection pads 18 of the electronic component 20, and wiring pattern 26b connected to the connection pads 18 via the via holes 28y	の 接続 パッド 上 に ビア ホール が 形成 さ れる 電子 部品 実装 構造 を 提供 する 解決 手段 電子 部品 ２０ が 実装 さ れる 被 実装 体 ２６ ａ と 、 被 実装 体 ２６ ａ の 上 に 、 最上 に エッチング ストップ 層 １６ （ 銅 膜 、 金 膜 、 銀 膜 又 は 導電 性 ペースト 膜 ） を 備え た 接続 パッド １８ を 有する 電子 部品 ２０ の 接続 パッド １８ が 上向き に なっ て 実装 さ れ た 電子 部品 ２０ と 、 電子 部品 ２０ を 被覆 する 層間 絶縁 膜 ２８ ａ と 、 電子 部品 ２０ の

Table A.1 continued from previous page

English	Japanese
<P>COPYRIGHT: (C)2004,JPO	接続 パッド
The present invention relates to a battery pack for preventing a heat transfer phenomenon and a device including the battery pack, the battery module frame accommodating the battery cell laminate, the battery pack frame to which the battery module frame is mounted, and the above	本 発明 は 熱 伝達 現象 を 防止 する 電池 パック および 該 電池 パック を 含む デバイス に 関する もの で あっ て 、 電池 セル 積層 体 を 収容 する 電池 モジュール フレーム と 、 前記 電池 モジュール フレーム が 装着 さ れる
A battery module mounting portion formed on the battery pack frame so that the battery module frame is mounted on the battery pack frame is included, and an insulating member is formed between the battery module mounting portion and the battery module frame	電池 パック フレーム と 、 前記 電池 パック フレーム に 形成 さ れ て 前記 電池 モジュール フレーム が 前記 電池 パック フレーム に 装着 さ れる よう に する 電池 モジュール マウンティング 部 と 、 を 含み 、 前記 電池 モジュール マウンティング
Has been done	部 と 前記
To suppress deterioration of reliability of a semiconductor device	課題 半 導体 装置 の 信頼 性 が 低下

Table A.1 continued from previous page

English	Japanese
SOLUTION: A first gate electrode is formed on a semiconductor layer SM located in an SOI region 1A of a substrate 1 having a semiconductor substrate SB, an insulating layer BX, and a semiconductor layer SM via a first gate insulating film, and a second gate electrode is located on a first region 1Ba of a bulk region 1B, and formed on the semiconductor substrate SB subjected to epitaxial growth treatment via a second gate insulating film, and a third gate electrode is located in a second region 1Bb of the bulk region 1B, and formed on the semiconductor substrate SB not subjected to the epitaxial growth treatment via a third gate insulating film	する の を 抑制 する 解決 手段 半 導体 基材 ＳＢ 、 絶縁 層 ＢＸ 、 半 導体 層 ＳＭ を 有する 基板 １ の ＳＯＩ 領域 １ Ａ に 位置 する 半 導体 層 ＳＭ 上 に 第 １ ゲート 絶縁 膜 を 介し て 第 １ ゲート 電極 を 、 バルク 領域 １ Ｂ の うち の 第 １ 領域 １ Ｂａ に 位置 し 、 かつ 、 エピタキシャル 成長 処理 が 施さ れ た 半 導体 基材 ＳＢ 上 に 第 ２ ゲート 絶縁 膜 を 介し て 第 ２ ゲート 電極 を 、 バルク 領域 １ Ｂ の うち の 第 ２ 領域 １ Ｂｂ に 位置 し 、 かつ
SELECTED DRAWING: Figure 8	、 エピタキシャル 成長 処理
PROBLEM TO BE SOLVED: To provide a solid state imaging device which achieves high functionality while maintaining high accuracy of a wire bonding process	課題 高い ワイヤ ボンディング プロセス の 精度 を 維持 し つつ 、 高 機能 化 さ れ た 固体 撮像 素子 を 提供 する

Table A.1 continued from previous page

English	Japanese
SOLUTION: A semiconductor device comprises: a substrate composed of a semiconductor material; and layers which are composed of a plurality of kinds of materials and constructed on the substrate; and an opening which is provided for exposing a surface of an electrode pad and pierces the layer constructed as at least an insulation film, and filled with aluminum or an aluminum alloy	こと が できる よう に する 解決 手段 半 導体 材料 から 成る 基板 と 、 前記 基板 上 に 構成 さ れ た 複数 種類 の 材料 から なる 層 と を 有し 、 電極 パッド の 表面 を 露出 さ せる ため の 開口 部 で あっ て 、 前記 基板 上 に 構成 さ れ た 層 の うち
<P>PROBLEM TO BE SOLVED: To provide a picture encoding method, wherein intensity values of luminance or color difference are quantized into integers, capable of preventing accumulation of rounding errors caused by bilinear interpolation processing in motion compensation	課題 輝度 または 色差 の 強度 値 が 整数 に 量子 化 さ れる 画像 符号 化 方法 に おい て 、 動き 補償 に おい て 行わ れる 共 1 次 内挿 の 処理 に
<P>SOLUTION: Two kinds of P+ frames 503, 507 adopting positive rounding where a value resulting from adding 0	よっ て 発生 する 丸め込み 誤差 の 蓄積 を 防ぐ 方法 を 提供 する 解決 手段 片方 向
5 to an integer is rounded up and P- frames 505, 509 adopting negative rounding where a value resulting from adding 0	の 動き 補償 予測 を 行う P フレーム に 関し 、 整数 に 0 . 5 を 加え た 値 を 切り上げる

APPENDIX B

Results of NLP-Model-III (Section 4.3)

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
This application is a continuation of International Application No. PCT/CN2015/078259, filed on May 5, 2015, which claims priority to Chinese Patent Application No. 201410187463.3, filed on May 6, 2014.	この申請は、2015年5月5日に提出された国際出願番号PCT/CN2015/078259の継続であり、2014年5月6日に提出された中国特許出願番号10187463.3の優先権を主張しています。	This application is a continuation of International Application No. PCT / CN 2015/078259 filed May 5, 2015 and claims priority to Chinese Patent Application No. 10187463.3 filed May 6, 2014.	0.77
The disclosures of the aforementioned applications are hereby incorporated by reference in their entireties.	上記の申請書の開示は、その全体への参照によってここに組み込まれています。	The disclosure of the above application is incorporated here by reference in its entirety.	0.5

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
The present invention relates to the field of computer technologies, and in particular, to a self-driving car scheduling method, a car scheduling server, and a self-driving car.	本発明は、コンピュータ技術、特に、自動運転車スケジューリング方法、自動車スケジューリングサーバ、および自動運転車の分野に関する。	The present invention relates to computer techniques, in particular the fields of self-driving car scheduling methods, self-driving car scheduling servers, and self-driving cars.	0.58

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
A self-driving car is also referred to as a driverless car, a computer driving car, or a wheeled mobile robot, and is an intelligent car that implements unmanned driving by using a computer system.	自動運転車は、無人運転車、コンピュータ駆動車、または車輪付き移動ロボットとも呼ばれ、コンピュータシステムを使用して無人運転を実装するインテリジェントな車である。	Self-driving cars, also called unmanned vehicles, computer-driven vehicles, or wheeled mobile robots, are intelligent vehicles that use computer systems to implement unmanned driving.	0.24

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
The self-driving car relies on coordination among an artificial intelligence system, a visual computation system, a radar system, a monitoring system, and a global positioning system, so that a computer can automatically and safely operate the self- driving car without an active operation of a person.	自動運転車は、人工知能システム、視覚計算システム、レーダーシステム、監視システム、およびグローバルポジショニングシステム間の協調に依存しているため、コンピュータは人のアクティブ操作なしで自動運転車を自動的に安全に操作できる。	Self-driving cars rely on coordination between artificial intelligence systems, visual computing systems, radar systems, surveillance systems, and global positioning systems, so computers automatically and safely drive self-driving cars without active human intervention. Can be operated.	0.35

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
As self-driving cars become popularized, at present it becomes an emerging means of travel to use a self-driving car as a taxi.	自動運転車が普及するにつれて、現在、自動運転車をタクシーとして使用する旅行の新たな手段となっています。	As self-driving cars become more widespread, they are now a new way of traveling to use self-driving cars as taxis.	0.45
To enable self-driving cars used as taxis to maximally meet different ride requirements, how to schedule the self-driving cars becomes a problem to which a person skilled in the art pays relatively much attention.	タクシーとして使用される自動運転車が異なる乗車要件を最大に満たすために、どのように自動運転車をスケジュールするかは、芸術に熟練した人が比較的注意を払う問題となる。	How to schedule self-driving cars so that the self-driving cars used as taxis can meet the different riding requirements to the maximum is a matter of relative attention to art-savvy people.	0.43

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
In the prior art, cars are usually scheduled in the following manner: A car scheduling server receives each ride request by means of a telephone, where a ride request includes a location of the ride request.	自動車スケジューリングサーバは、乗車要求の場所を含む電話によって各乗車要求を受信する。	The vehicle scheduling server receives each boarding request by telephone including the location of the boarding request.	0.21

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
Every time when a new ride request is received, the car scheduling server acquires a location of each car within a management range, and determines, from assigned cars, cars located within a preset range of a location corresponding to the new ride request.	車両スケジューリングサーバは、新規の乗車依頼を受けるたびに、管理範囲内の各車両の位置を取得し、割り当てられた車両から、当該新規の乗車要求に対応する位置の予め設定された範囲内に位置する車両を決定する。	Each time a new boarding request is received, the vehicle scheduling server acquires the position of each vehicle within the management range, and from the assigned vehicle, the position is within the preset range of the position corresponding to the new boarding request. Decide which vehicle to use.	0.54

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
And finally, determines, from the cars within the preset range, a car having a minimum increase in a driving distance after the new ride request is added, and assigns the new ride request to the car.	最後に、プリセット範囲内の車から、新規乗車要求を追加した後の走行距離の最小増加を有する車を決定し、新規乗車要求を車両に割り当てる。	Finally, from the cars within the preset range, the car having the minimum increase in mileage after adding the new ride request is determined, and the new ride request is assigned to the vehicle.	0.69

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
In the process of implementing the present invention, the inventor finds that the prior art at least has the following problems: In the process of car scheduling, because car scheduling is performed according to only locations of assigned cars and a ride request, reference factors for scheduling are undiversified, causing low scheduling precision.	本発明の実施過程において、発明者は、先行技術に少なくとも次のような問題があることを発見した:自動車スケジューリングの過程において、割り当てられた車の位置と乗用車要求のみに応じて自動車スケジューリングが行われるため、スケジューリングの基準因子が多様化せず、スケジューリング精度が低下する。	In the process of implementing the present invention, the inventor has found that the prior art has at least the following problems: In the process of vehicle scheduling, vehicle scheduling is performed only according to the assigned vehicle position and passenger vehicle requirements. Therefore, the reference factors for scheduling do not diversify, and the scheduling accuracy decreases.	0.63

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
For example, after a car that responds to a new ride request is determined, if there is a real-time severe traffic jam at a location of the car or a traffic accident occurs on a driving course, the selected car cannot necessarily arrive at a location corresponding to the new ride request soonest, and scheduling precision is affected.	例えば、新規乗車要求に応答する車両が判定された後、走行コースにおいて、車両の場所におけるリアルタイムの重大渋滞や交通事故が発生した場合、選択された車両は、新乗車要求に対応する場所に早急に到着することができず、スケジューリング精度が影響を受ける。	For example, if a vehicle that responds to a new ride request is determined and then a real-time serious traffic jam or traffic accident occurs at the vehicle location on the driving course, the selected vehicle will immediately move to the location that corresponds to the new ride request. Cannot arrive at, and scheduling accuracy is affected.	0.69

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
To resolve the problem in the prior art, embodiments of the present invention provide a self-driving car scheduling method, a car scheduling server, and a self-driving car.	本発明の実施形態は、先行技術における問題を解決するために、自動運転車スケジューリング方法、自動車スケジューリングサーバ、および自動運転車を提供する。	Embodiments of the present invention provide an autonomous vehicle scheduling method, an automobile scheduling server, and an autonomous vehicle in order to solve problems in the prior art.	0.46

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
The technical solutions are as follows: According to a first aspect, a self-driving car scheduling method is provided, including: receiving a ride request, where the ride request includes at least a ride starting location, a ride destination, and an expected destination arrival time; determining, according to the ride request and driving information of self-driving cars within a management range, at least one first candidate car from the multiple self-driving cars; calculating according	技術的解決方法は、第1の側面に従って、乗用車要求が少なくとも1つの乗車開始場所、乗用目的地、および予想目的地到着時刻を含む乗用車要求の受信、管理範囲内の自動運転車の乗用要求および運転情報に応じて、複数の自動運転車から少なくとも1つの第1の候補車を決定すること、現在位置情報、現在の道路状況情報、および各第1の候補車の計画経路情報に応じて、各第1の候補車が乗用目的地に到着するために必	The technical solution is to receive a passenger vehicle request including at least one boarding start location, passenger destination, and expected destination arrival time, a vehicle request for an autonomous vehicle within control, and a vehicle request according to the first aspect. Determining at least one first candidate vehicle from multiple autonomous vehicles according to driving information, current location information, current road condition information	0.59

Table B.1 continued from previous page

(A):Original English text	(B):Translation of (A)	(C):Backtranslation of (B)	(D):BLEU
Average BLEU Score:			51%

Table B.1: Average BLEU scores of a sample of our test data using NLP-Model-III

Curriculum Vitae

Name:	Maimoonah Ahmed
Post-Secondary Education and Degrees:	Western University London, ON 2021 - 2023 MESc., Software Engineering, ECE Department Vector Institute Accredited Collaborative Specialization in Artificial Intelligence
	University of Guelph Guelph, ON 2016 - 2021 B.A., Honours Mathematical Science Area of Emphasis in Computer Science
Related Work Experience:	Teaching Assistant The University of Western Ontario 2021 - 2023

Publications:

Maimoonah Ahmed, Abdelkader Ouda, Mohamed Abusharkh, Sandeep Kohli, Khushwant Rai, 2023, An Optimized Approach to Translate Technical Patents from English to Japanese Using Machine Translation Models. MDPI Applied Sciences

Maimoonah Ahmed, Abdelkader Ouda, Mohamed Abusharkh, 2022, An Analysis of the Effects of Hyperparameters on the Performance of Simulated Autonomous Vehicles. International Telecommunications Conference (ITC-Egypt)

Maimoonah Ahmed, Ben Cameron, 2021, The node cop-win reliability of unicyclic and bicyclic graphs. Networks, DOI: http://doi.org/10.1002/net.22055.