HYBRID MACHINE TRANSLATION USING MALAY-ENGLISH LANGUAGE PARALLEL TEXT EXTRACTION FROM COMPARABLE TEXT

YEONG YIN LAI

UNIVERSITI SAINS MALAYSIA

2024

HYBRID MACHINE TRANSLATION USING MALAY-ENGLISH LANGUAGE PARALLEL TEXT EXTRACTION FROM COMPARABLE TEXT

by

YEONG YIN LAI

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

December 2024

ACKNOWLEDGEMENT

Firstly, I would like to thank my supervisor, Associate Professor Doctor Tan Tien Ping for his experienced opinion, continuous support, guidance, motivation, passion, well-targeted questions and careful analysis. Besides that, he also often helped me to focus on the relevant matters. A big thank you to my beloved family for their financial support, moral support and encouragement for me to pursue my studies this far. Furthermore, I would like to thank my friends and schoolmates who have accompanied me all the time and given me valuable opinions and moral support. Furthermore, I would like to thank Malaysia's government who providing a scholarship for helping me to pay the registration fees. Last but not least, I would like to thank all staff in the School of Computer Sciences, USM for providing a comfortable environment and perfect equipment for all students.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BBC	British Broadcasting Corporation
BLEU	BiLingual Evaluation Understudy
CS	Computer Sciences
DT	Decision Tree
EBMT	Example-based Machine Translation
EM	Expectation Maximization
GRU	Gated Recurrent Unit
HTML	HyperText Markup Language
IPS	Institut Pengajian Siswazah
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
L2	Second Language
MT	Machine Translation
NB	Naïve Bayes
NLP	Natural Language Processing
NMT	Neural Machine Translation
NN	Neural Network
OOV	Out-of-vocabulary
POS	Part-of-speech
RBMT	Rule-based Machine Translation
RNN	Recurrent Neural Network
SL	Source Language
SMT	Statistical Machine Translation
SVM	Support Vector Machine
T5	Text-to-Text Transfer Transformer
TL	Target Language
URLs	Uniform Resource Locators
USM	Universiti Sains Malaysia
UNK	Unknown
WMT	Workshop on Machine Translation

LIST OF APPENDICES

TERJEMAHAN MESIN HIBRID MENGGUNAKAN PENGEKSTRAKAN TEKS SELARI BAHASA MELAYU-INGGERIS DARIPADA TEKS SEBANDING

ABSTRAK

Terjemahan mesin (MT) menyiasat pendekatan untuk menterjemah teks daripada bahasa sumber (SL) kepada bahasa sasaran (TL). Teks selari ialah sumber yang penting untuk membina model terjemahan sistem MT. Teks selari ialah teks dan terjemahannya dalam satu atau lebih bahasa. Namun begitu, teks selari yang tersedia adalah terhad. Oleh itu, beberapa arahan telah diterokai dan disiasat dalam tesis ini untuk meningkatkan kualiti terjemahan walaupun teks selari terhad. Pertama, kami menganalisis menggunakan maklumat linguistik dalam terjemahan mesin untuk mengimbangi kekurangan data untuk latihan. Kedua, kami mengkaji masalah memperoleh teks selari daripada teks yang setanding. Teks setanding ialah teks serupa dalam bahasa berbeza yang mungkin dihasilkan secara bebas. Ketiga, kami menyiasat seni bina terjemahan mesin statistik (SMT) dan terjemahan mesin saraf (NMT) untuk menggabungkan kekuatan kedua-dua sistem. Kajian ini dijalankan menggunakan terjemahan mesin bahasa Inggeris-Melayu dalam domain berita dan domain sains komputer. Untuk masalah pertama, kami mencadangkan untuk menggunakan maklumat tatabahasa untuk membina model terjemahan. Kami meningkatkan skor BLEU daripada 13.40% kepada 15.41% menggunakan 315,194 teks selari. Dalam masalah kedua, kami mencadangkan algoritma untuk mengekstrak ayat selari dan serpihan/subayat selari daripada teks yang setanding. Pendekatan ini menemui teks yang setanding yang sepadan. Kemudian, penyelaras ayat dan pengelas digunakan untuk menyelaraskan ayat dalam teks yang setanding. Serpihan selari kemudiannya

diekstrak menggunakan pengelas. Kami berjaya mengekstrak 446,065 ribu teks selari daripada pelbagai sumber dan meningkatkan skor BLEU terjemahan mesin dalam domain sains komputer dan domain berita daripada 20.33% dan 35.38% bagi NMT, 16.63% dan 41.81% bagi SMT kepada 21.17% dan 41.75% bagi NMT, 18.56% dan 48.35% bagi SMT. Bagi masalah ketiga pula, SMT dan NMT hibrid telah dicadangkan untuk meningkatkan kualiti terjemahan. Pendekatan kami menggunakan seni bina ingatan jangka panjang pendek (LSTM) berbilang sumber pengekod-penyahkod. Seni bina menggunakan dua pengekod, satu untuk membenamkan ayat yang hendak diterjemahkan, dan satu lagi pengekod untuk membenamkan terjemahan yang dihasilkan oleh SMT. MT hibrid meningkatkan skor BLEU terjemahan mesin dalam domain sains komputer dan domain berita daripada 21.21% dan 40.92% bagi NMT, 21.13% dan 48.35% bagi SMT kepada 35.97% dan 61.81% masing-masing.

HYBRID MACHINE TRANSLATION USING MALAY-ENGLISH LANGUAGE PARALLEL TEXT EXTRACTION FROM COMPARABLE TEXT

ABSTRACT

Machine translation (MT) investigates the approaches to translate a text from a source language (SL) to a target language (TL). Parallel text is the resource that is essential for building the translation model of an MT system. A parallel text is a text and its translation in one or more languages. Nevertheless, there are not many parallel texts that are freely available. Thus, a few directions were explored and investigated in this thesis to improve the translation quality despite the limited parallel text. Firstly, we analysed using linguistic information in machine translation to compensate for the lack of data for training. Secondly, we studied the problem of acquiring a parallel text from comparable texts. Comparable texts are similar texts in different languages that may be independently produced. Thirdly, we investigated the architecture of statistical machine translation (SMT) and neural machine translation (NMT) to combine the strength of both systems. This study was carried out using English-Malay machine translation in the news domain and computer science domain. For the first problem, we propose to use affixation and part-of-speech information to build a translation model. We improve the BLEU score from 13.40% to 15.41% using 315,194 parallel texts. In the second problem, we propose an algorithm to extract parallel sentences and parallel fragments/subsentences from comparable texts. The approach finds matching comparable texts. Then, a sentence aligner and a classifier are used to align the sentences in the comparable text. The parallel fragments are then extracted using a classifier. We managed to extract 446,065 thousand parallel texts from multiple

sources and improved the BLEU scores of machine translation in the computer science domain and news domain from 20.33% and 35.38% in NMT, 16.63% and 41.81% in SMT to 21.17% and 41.75% in NMT, 18.56% and 48.35% in SMT. As for the third problem, a hybrid SMT and NMT was proposed to improve the translation quality. Our approach uses the multi-source encoder-decoder long short-term memory (LSTM) architecture. The architecture uses two encoders, one to embed the sentence to be translated, and another encoder to embed the translation produced by SMT. The hybrid MT increased the BLEU score of machine translation in the computer science domain and news domain from 21.21% and 40.92% in NMT, 21.13% and 48.35% in SMT to 35.97% and 61.81% respectively.

CHAPTER 1

INTRODUCTION

1.1 Machine Translation

Machine translation (MT) investigates the approaches to translating text from a source language (SL) to a target language (TL). The maturity of the technology has made it possible to use MT in many situations, such as when a tourist visits a foreign country, or a user wants to understand a foreign language web page etc. Among the online systems providing machine translation services are Google Translate, Bing Translation, Babel Fish Translate, etc. Although at present the quality of the translation produced automatically is still far from what is produced by translation experts, especially in some specialized domains, MT provides the convenience of translation that is fast and at a low cost to consumers. Besides that, the quality of translation is also improving every year.

As early as the 1960s, MT systems were rule-based machine translation (RBMT). RBMT used linguistic information such as dictionaries and grammar on SL and TL to generate the translation sentences. After almost twenty years, a new method of MT has been introduced called “corpus-based” by a Japanese group by using methods based on corpora and called example-based machine translation (EBMT). People break the sentence into phrases and use phrasal translation by analogy to previous translations as example translations. EBMT requires linguists to provide a descriptive tree for the SL and TL, and analysis needs to be done to determine the alignment between words or phrases between the SL and the TL. The principal feature of this approach is that no or less syntactic or semantic rules are used in the analysis of texts or the selection of lexical equivalents. The more example translation available, the better the translation produced by an EBMT (Makoto, 1984).

Around the 1980s, a statistical machine translation (SMT) learns a statistical translation model from the analysis of bilingual texts for translation, also known as parallel text (Brown et al., 1988), and it also captures the word distribution from a target language text corpus in a statistical language model.

Early 2000s, the advent of Neural Machine Translation (NMT) began to shift focus from traditional phrase-based SMT to neural network-based approaches. Artificial neural networks have been applied in many fields, for example in image recognition, face recognition, speech synthesis, and so on. One of the successes of neural networks is due to the advancement in the RNN that models very complex patterns using deep layers of neural networks.

1.2 Motivation

Research in low-resource machine translation (MT) is essential to bridge the linguistic digital divide, ensuring that speakers of underrepresented languages have access to information, services and opportunities. With most MT technologies focused on high-resource languages like English and French, millions of people speaking low-resource languages face barriers in education, healthcare, disaster response and global communication. By improving translation for these languages, low-resource MT can help promote linguistic diversity, cultural preservation and inclusion, providing people with better access to digital content, economic opportunities and critical information in their native languages.

On the technical front, low-resource MT poses significant challenges, such as data scarcity and complex language structures. Addressing these challenges motivates the development of innovative approaches like transfer learning, multilingual models and unsupervised learning, advancing the field of artificial intelligence (AI) and NLP.

Additionally, this research has ethical and social implications, particularly for marginalized communities and indigenous populations where MT can play a vital role in promoting linguistic rights and social justice. Overall, low-resource MT holds the potential to enhance global communication, preserve endangered languages and push the boundaries of current machine translation technology.

1.3 Problem Statements

Malay, like other languages, is also very much influenced by English. A lot of English words have been absorbed into Malay, especially in the field of science and technology **(**Ranaivo-Malancon & Tan, 2009). For example, the English word "system" is absorbed into Malay as "*sistem*" which has the same pronunciation but with the letter "y" changed to "i". While most Malaysians can also understand English, MT is an essential tool for transferring knowledge, especially from English to Malay. The translation of English literature can bring the latest knowledge to the people who speak Malay. Iskandar (2016) stated the differences between Malay and English that Malay has no tenses, no subject-verb agreement, no difference in the number of speakers and no singular-plural forms.

English-Malay machine translation is considered a low-resource task primarily due to the limited availability of parallel corpora. This is the main definition of low-resource in this thesis. High-quality parallel datasets, which are essential for training effective neural machine translation models, are scarce for this language pair. Unlike high-resource language pairs such as English-French or English-Spanish, there are fewer aligned sentence pairs for English and Malay. The existing datasets tend to be small, domain-specific (e.g., religious texts or legal documents) and may contain noise, making it difficult to build robust translation models.

Another significant factor contributing to the low-resource status of English-Malay translation is the linguistic differences between the two languages. English, a Germanic language, has a relatively rigid word order and complex morphological structures. In contrast, Malay, an Austronesian language, has a more flexible word order, simpler morphology and unique grammatical rules. These differences make it more challenging for translation models to generalize effectively from limited training data, especially when trying to capture nuances like context and idiomatic expressions across the two languages.

In general, one of the essential resources for building a machine translation system is a parallel text corpus. A parallel text is a text and its translation in one or more languages. An entry in the parallel text corpus consists of a source language sentence and its translation (in the target language) language sentence. An entry can also be a pair of source and target language words or phrases. Table 1.1 shows examples of English-Malay parallel sentences and phrases.

Table 1.1 English-Malay parallel sentences and phrases.

English sentences	Malay sentences
profitability .	*kemungkinan mendatangkan keuntungan .*
go abroad .	*pergi ke luar negeri .*
furthermore , there is a need to widen the channel to prevent congestions .	t*ambahan pula , terdapat keperluan meluaskan saluran untuk mengelakkan kesesakan .*
all kinds of things .	macam-macam .

Other factors that affect the size of parallel text required are the domain, for example, open domain versus close domain translation, language pairs and MT model. In addition, some popular language pairs may have more parallel text that are available than others. For example, the parallel text for the language pairs such as English-

French, English-Chinese and English-Japanese are more on the Internet compared to English-Malay.

The main research problem investigated in this study is low-resource machine translation. Studies have shown a positive correlation between the size of the parallel text used for training and the quality of the translation produced. A significant amount of research has proven that translation systems tend to produce erroneous and disfluent translations when out-of-vocabulary (OOV) words appear in the test data (Corston-Oliver & Gamon,2004; El-Kahlout & Oflazer, 2006; Habash & Sadat, 2006; Virpioja et al., 2007). To compensate for the lack of parallel text, linguistic knowledge can be applied, just like in other NLP problems. Word morphology is a fundamental aspect of linguistic knowledge that deals with the structure and formation of words. It examines how words are formed from smaller meaningful units called morphemes, which are the smallest grammatical units in a language. Word morphology is a critical aspect of linguistic knowledge that provides insights into how words are structured, formed and function within a language. Within this framework, a root word serves as the core element that carries the primary meaning of a word, while affixation involves adding affixes (prefix and suffix) to the root to modify its meaning or grammatical function. This relationship allows for the creation of new words and variations in meaning. For example, adding the prefix “un-” to the root word “happy” forms “unhappy” altering its meaning to “not happy”. In this work, the usage of word morphology in MT will be investigated to improve the translation quality. One of the purposes of using the word morphology is to reduce OOV words. OOVs are a ubiquitous and difficult problem in MT. The translation systems will produce erroneous and disfluent translations if an OOV or a word that was not observed in the

training data. So, knowing the affixation information of a word can be useful to solve OOV affixed words or improve the overall translation.

The second problem addressed in this research involves retrieving comparable text from multiple sources and extracting parallel text from the comparable data. In this study, 'multiple sources' refers to raw data gathered from various areas, fields or domains. For example, this may include data from daily work conversations, online news articles and other forms of communication, whether formal or informal. Additionally, sources from the education sector include books, articles, dictionaries, theses and journals. The education field itself spans different domains such as law, medicine, social sciences and computer science. As a result, the vocabulary, sentences and terminologies used in these contexts can vary significantly.

Even though parallel text is scarce for most language pairs, comparable text for language pairs is more readily available. Comparable text is a similar text in a different language that may be independently produced, for example, news that reports an event that happens, or Wikipedia articles on a topic in different languages. Compared to the parallel text, the sentences in the translated text may not be aligned, and only some or part of the sentences are translated. One of the interesting problems in this study is to find the comparable text and then extract the parallel text from the comparable text. With the additional parallel text, we will test on different domain-specific test sets: the general domain (daily news) and the specific domain (exam papers).

The third problem addressed in this research is proposing a model that combines the strengths of SMT and NMT. While SMT has been successfully deployed in many commercial systems, it does not work very well and suffers from the following two major drawbacks. First, translation decisions are locally determined as we

translate phrase-by-phrase and long-distance dependencies are often ignored. More problematically, the entire MT pipeline is becoming increasingly complex as more and more features are added to the log-linear framework such as in recent MT systems (Galley and Manning, 2008; Chiang et al., 2009; Green et al., 2013). Many different components need to be tuned separately, e.g., translation models, language models, reordering models, etc., which makes it difficult to combine them and to innovate. As a result, the translation quality has saturated for SMT and big changes to the existing framework were in dire need. Thus, the quality of a machine translation system largely depends on the availability of a large number of resources to build a robust language model and translation model.

However, if compared with SMT, NMT can train multiple features jointly and does not need prior domain knowledge, which enables zero-shot translation (Johnson et al., 2016). In addition, NMT can also help reduce morphology errors, syntax errors, and word order errors which were commonly seen on SMT. On the other side, there are still problems and challenges of NMT that need to be tackled such as the training and decoding process is quite slow, the style of translation can be inconsistent for the same word, there exists an "out-of-vocabulary" problem on the translation results, the "black-box" neural network mechanism leads to poor interpretability, etc. Thus, the parameters for training are mostly picked based on experience.

Furthermore, SMT and NMT have their strengths and weaknesses. SMT used more training data and spaces in a disk to calculate the probability value for the words. Moreover, NMT solves the MT problems such as long-distance reordering, morphology, syntax and agreement errors, tolerance to noisy data, and multilingual or multi-domain translation whereas SMT faces fewer problems with interpretability and

vocabulary or rare word problems. Due to the mechanism used in NMT is sentence-by-sentence, attentional encoder-decoder networks (optimization) and training multiple features jointly whereas SMT uses word-by-word or phrase-by-phrase, statistical analysis (probability) and feature engineering required.

However, NMT systems have outperformed the state-of-the-art SMT model on various language pairs in terms of translation quality. However, recent studies show that NMT generally produces fluent yet sometimes inaccurate translations, mainly due to three problems. The first problem is the low MT resources problem, NMT lacks a mechanism to record the source words that have been translated or need to be translated, resulting in either "over-translation" or "under-translation" (Cohn et al. 2016; Tu et al. 2016). The second problem is the incorrect translation problem, NMT is prone to generate words that seem to be natural in the target sentence, but do not reflect the original meaning of the source sentence (Arthur et al. 2016). The third problem is the unknown (UNK) word problem, NMT uses a fixed modest-sized vocabulary to represent the most frequent words and replaces other words with a UNK word (Jean et al. 2015; Luong et al. 2015). Experimental results show that translation quality degrades rapidly with the number of UNK words increasing (Cho et al. 2014). In the other way, SMT translated every source word into the target word and SMT treated words as discrete symbols, which ensures that a source word will be translated into a target word which has been observed at least once in the training data. Furthermore, SMT memorizes all the translations where NMT is taken as UNK words. Thus, the last problem in this research work is to find ways to combine the strengths of SMT and NMT. The intuition of the hybrid MT approach is to present the SMT decoding to NMT, and let the NMT to learn/decide which part of the SMT decoding it wants to keep or modify.

In summary, below are the research questions for our studies:

1) How to use linguistic knowledge in word morphology for low-resource MT?

2) How to extract parallel text from the comparable text for training MT models?

3) How to combine the strength of SMT and NMT to improve the translation quality?

1.4 Objectives

The objectives of the research are as follows.

1) To propose an approach to improve English-Malay MT using linguistic word morphology knowledge in SMT.
2) To propose a method to extract English-Malay parallel text from the comparable text and then use them to train MT models.
3) To propose a hybrid translation model that combines SMT and NMT to improve the translation quality.

1.5 Research Contribution

The first contribution of this research is to propose an approach for improving English-Malay MT by incorporating linguistic knowledge. Specifically, the approach uses word morphological structures to address out-of-vocabulary (OOV) words, particularly those with the same base word. For example the words ‘correctly’, ‘incorrect’, and ‘correction’ all share the same base word, ‘correct’. By reintroducing words with affixations, such as prefixes and suffixes, this approach aims to increase

the vocabulary size and improve word alignment during the MT decoding process. This method is particularly useful for low-resource language translation, especially when the language has a high degree of affixation.

The second contribution of this research is to propose a method for extracting English-Malay parallel text from comparable texts using both SMT and NMT toolkits. This method takes advantage of existing sources to obtain useful data for our research, offering an effective and cost-saving approach, especially in the field of translation, which often requires professional translators to ensure sentence accuracy and fluency. By using the proposed method, both the size of the training data and the quality of the MT can be improved.

Last but not least, the third contribution of this research is to propose a hybrid MT model that combines the strengths of both SMT and NMT for English-Malay translation. By applying the parallel text extracted from the method proposed in the second objective, the hybrid MT approach aims to achieve better English-Malay translation results. In conclusion, in this research leverages available resources, such as existing sources and toolkits, to improve the quality of English-Malay language translation.

1.6 Research Scope and Limitation

This research focuses on the English-Malay translation. The study of English-Malay in machine translation (MT) research is crucial for promoting linguistic representation among Malay speakers, who number in the millions across Malaysia, Indonesia, Brunei and Singapore. Despite the widespread use of the Malay language, there is a significant lack of digital resources, particularly the parallel text and the methods for modeling the language pair. Developing effective translation systems can

help bridge the digital divide, enabling greater access to essential information and services in education, healthcare and government. Furthermore, English-Malay MT presents unique technical challenges due to differences in grammar, syntax and vocabulary, making it an area ripe for innovation. By focusing in this language pair, researchers can explore novel approaches such as transfer learning and multilingual models, contributing valuable insights to the broader field of machine translation. Additionally, investing in translation technology supports the preservation of the Malay language and culture, safeguarding it against marginalization in the face of globalization.

1.7 Organization of Research Thesis

This research thesis is organized as follows:

Chapter 2 presents the literature background of the methods and the sources of data collection and parallel corpus acquisition, low-resource MT is one of the problems faced by many researchers and how their approach or method worked to solve this problem will also be discussed in this chapter. Besides that, current research works in hybrid MT, the comparison between SMT and NMT and the evaluation method of MT will also be discussed.

Chapter 3 presents the approaches for low-resource language translation, data collection from multi-sources, data extraction and English to Malay language translation approach using SMT and NMT. A detailed description of the research methodology and research procedures is presented. Implementation steps are also described for the proposed approach to be reproduced.

In Chapter 4, data collection, experiments, results and discussion are presented. This chapter also highlights the steps taken to improve accuracy when compared to the other baseline approaches.

Finally, Chapter 5 summarizes and concludes this research work. Future work to extend the proposed method to other research applications is also suggested.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The challenges in machine translation stem from the limited language resources, particularly parallel texts, available for training translation models. To address this constraint, several approaches that incorporate linguistic knowledge into building translation models will be explored. Additionally, we will investigate methods for acquiring parallel text corpora from existing resources such as comparable texts. Finally, we will examine improved translation models that can better capture the relationship between the source and target languages.

2.1.1 Evaluation Metric

Automatic evaluation of translation involves the use of parallel text that can be done quickly. The reference or gold standard is a translation that has been examined by human translators. BLEU scores are the most widely used assessment because the tests performed show a positive correlation between BLEU scores and translation quality. BLEU scores can be assessed with the formula below. BLEU scores are between 0 and 1 or in percentages. The higher the score obtained the better the translation output obtained compared to the reference.

$$BLEU = minimum\left(1, \frac{length\ output}{length\ reference}\right) . \ (\prod_{ngram=1}^{4} precision_{ngram})^{1/4} \quad (2.1)$$

An example is given below to show the BLEU score calculation. Let's say the English sentence to be translated is “please serve the rice hot .”. The translation reference is “*sila sajikan nasi yang panas* .”, and the translation hypothesis produced is “*sila berkhidmat nasi yang panas* .”. So the BLEU MT score is 0.5373 or 53.73%. Table 2.1 below shows the BLEU calculations for the examples given.

Sentence: please serve the rice hot .

Hypothesis MT English Malay: *sila berkhidmat nasi yang panas .*

Match: (1-gram) (4-gram)

Reference: *sila* *nasi yang panas .*

Table 2.1 Example evaluation for BLEU score.

Source	Score
Precision 1-gram	5/6
Precision 2-gram	3/5
Precision 3-gram	2/4
Precision 4-gram	1/3
BLEU	**$(1/12)^{1/4}$ or 53.73%**

2.1.2 Statistical Machine Translation

Statistical machine translation (SMT) consists of two models, which are the translation model and language model. The SMT translation model is phrase-based, mapping phrases from the source language to the target language using alignment probabilities learned from parallel texts. Even when data is limited, SMT can still rely

on these learned mappings and direct word-to-word or phrase-to-phrase translations, which do not require a lot of parallel text to train. This makes SMT relatively robust when parallel data is scarce, as it focuses on localized correspondences rather than global patterns. SMT's language model, typically based on n-grams, also plays a crucial role in low-resource scenarios. This model captures the likelihood of word sequences in the target language and can be trained using monolingual text corpus, which is often easier to obtain than parallel data. In general, it models the syntax of the target language. Unlike NMT, which requires large amounts of parallel data to learn the syntax, SMT's approach allows it to produce fluent and locally coherent translations even with limited resources.

SMT takes a source sentence, S = [s_1 s_2 … s_n] in the source language, and generates a target sentence, T* = [t_1 t_2 … t_n] in the target language, where s_1 s_2 s_3 … s_n are phrases / null in the source language, and t_1 t_2 t_3 … t_n are phrases/null in the target language. Many possible target sentences can be translated from a source sentence. The idea is to find the most probable sentence as follows:

$$T^* = argmax\big(P(T|S)\big) \tag{2.2}$$

The equation will be decomposed using Bayes theorem as follows:

$$T^* = argmax\left(\frac{P(S|T)\times P(T)}{P(S)}\right) \tag{2.3}$$

Since P(S) is always constant, thus it can be removed. The equation can be simplified as follows:

$$T^* = argmax(P(S|T) \times P(T)) \tag{2.4}$$

P(T) is the probability of a target language sentence, which is modelled by a

language model. The language model for the target language can be built with a target language text corpus. The monolingual language model can be based on bigram or trigram models (Brown et al., 1992). From which the likelihood of a string of words is a valid sentence can be computed. By contrast, the translation model uses the frequency of co-occurrence of source and target words, the length of sentences in which they appear, their positions within their respective sentences, the fertility of the TL word (the number of SL words from which it arises), the actual words from which a TL word derives, and the position of these SL words in the SL string. Brown et al., (1993) propose a series of increasingly more sophisticated models that include more and more of those features.

On the other hand, P(S|T) is the probability of a source sentence given the target sentence, which is modelled by a translation model. The model is built using a parallel text corpus. The formula can be interpreted as saying that to translate source sentence s, we search for the target word string t that maximizes the value of the whole formula. The idea is that given sufficiently accurate statistics, the P(t) term biases the search towards grammatical TL word strings, while the P(s|t) term biases the search towards strings that are likely translations of the source sentence. This last conditional probability may appear confusing. Conditioning is on the target word string, as it is easier to estimate the probability of a given source sentence from a TL word string than the other way round. To appreciate this, we can think of the source sentence as giving us hints about the TL sentence. Then it is simpler to estimate from corpora the probability pf a set of hints (s) given a TL sentence (t) than it is to estimate the probability of a TL sentence from the hints alone. For example, Figure 2.1 shows the example sentence “serve the rice hot” by using SMT to select the most probable word or phrase in the Malay language. 882 different combinations of Malay sentences can be

produced.

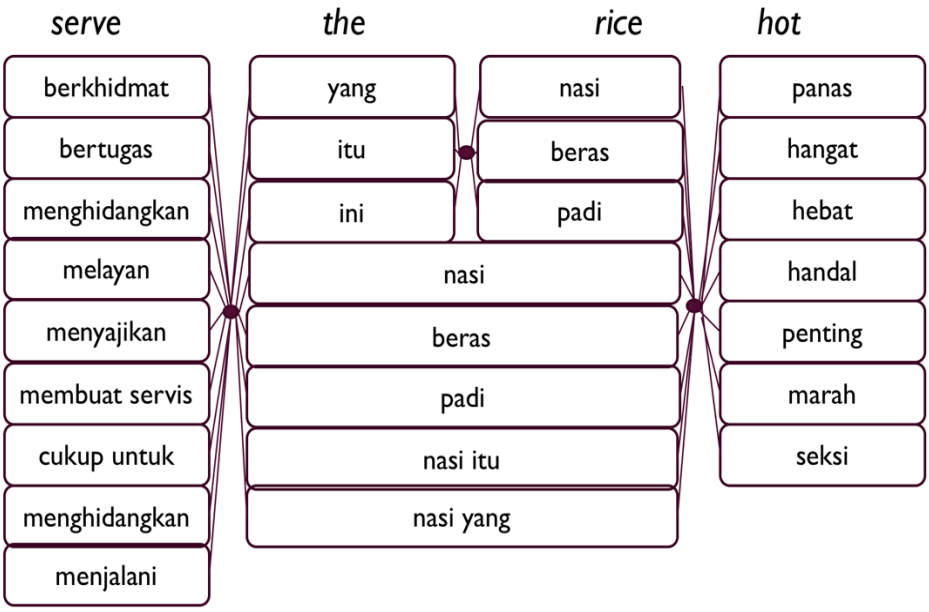


Figure 2.1 Example translation sentence "serve the rice hot" to Malay language using SMT.

Most of the sentences that are produced grammatically are wrong or the verses do not carry meaning. This example has been simplified with the assumption that there is no change in the position of the target language/phrase t_j to facilitate discussion. Note that there are several translations for each word/phrase. The word "serve" has nine translations in languages other than English for this example. The translation for each word/phrase and its probability or score (not shown in the diagram) can be obtained through the process of aligning the words from the parallel corpus. The automatic word alignment approach uses the expectation-maximization approach (Koehn, 2010). Pairs of target words and source words in parallel corpora that are often aligned together will have a high score, P (s_i | t_j). Also, there is another score, a language model score (for the Malay language). The Malay language model must be built with a Malay text corpus. The order of the Malay words that are often found in the text will have a high score. For

example, the order of the words "*menyajikan nasi*" and "*menghidangkan nasi*", "*nasi panas*" and "*nasi hangat*" and so on, will have a score (2-grams) are high because of the order of these two words can be found in Malay. Meanwhile, 2-grams sequences such as "*bertugas nasi*", "*padi marah*", "*nasi seksi*" will have an empty or very low score because it is very rarely found in the corpus of text. Scores from the translation model and language model for each word/phrase will be multiplied, and the combination that gives the highest value will be selected.

Popular implementations of SMT are Pharaoh (Koehn, 2004), which was succeeded by the open-source Moses (Koehn et al., 2007). Phramer (Olteanu et al., 2006) is a Java implementation of a phrase-based statistical system. Thot (Ortiz-Martinez et al., 2005) is a toolkit to train phrase-based models for these decoders. Available implementations of tree-based decoders are Hiero (Chiang et al., 2005) and SAMT (Zollmann et al., 2007). GIZA++ (Och and Ney, 2003) and MTTK (Deng and Byrne, 2006) are tools for word alignment; Hunalign (Varge et al., 2005) is a popular tool for sentence alignment. Cattoni et al. (2006) present a web demo for their phrase-based system. An online tool for machine translation evaluation is presented by Eck et al. (2006). Yawat is a web-based tool to view and create word alignments (Germann, 2007, 2008).

2.1.2(a) Word-based Translation SMT

Word-based SMT uses the word as the translation unit. This approach was first introduced by IBM (Brown et al., 1993). Vogel et al. (1996) proposed HMM-based models or IBM-1 to IBM-5 to estimate the word alignment from a large number of parallel texts by using the expectation maximization (EM) algorithm. These alignments are called "Viterbi alignments". Word-based models are the starting point for translation

models. EM algorithm consists of two steps which are the expectation step and the maximization step. The expectation step applies the model to the data using the model and then assigns the probabilities to possible values. The maximization step is to estimate the model from data which is to collect counts or weigh by probabilities and estimate the model from word counts. Figure 2.2 shows an example sentence "serve the rice hot" and the translation of the sentence in Malay. The initial step initializes alignments between the source word and target word that are equally likely. In a subsequent iteration, the model learns that word pairs that often appear together in parallel sentences, e.g., the word "serve" is often seen with "*menghidangkan*" will be aligned together. The expectation step and maximization step repeat until convergence and the inherent hidden structure is revealed by EM. Finally, parameter estimation will be formed from the aligned corpus.

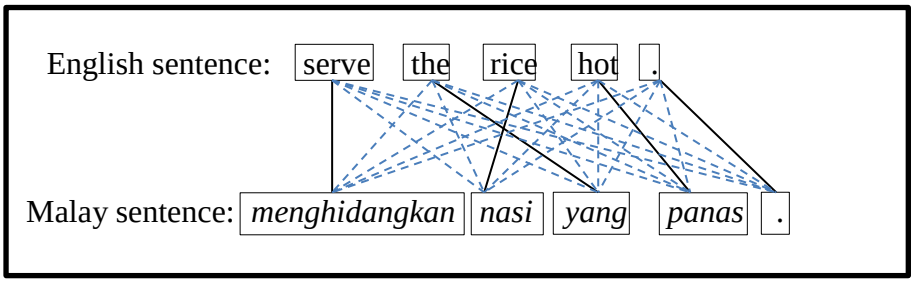


Figure 2.2 Example of word-based translation.

The word-based translation (one source word aligned to (→) one target word) alignment results for the example sentence pair above:

- serve → *menghidangkan*
- the → *yang*
- rice → *nasi*
- hot → *panas*
- . → .

2.1.2(b) Phrase-based Translation SMT

The phrase-based SMT uses a phrase instead of a word as a unit of translation. Phrase alignments are produced from word alignments. One of the ideas in phrase alignment is to use the word alignment produced to find neighbouring source language words and neighbouring target language words that always appear together. Schwenk (2012) used neural networks to directly learn the translation probability of phrase pairs using continuous representations. Koehn (2004) used Pharaoh as a decoder for phrase-based SMT. Wu et al. (2008) used out-of-domain parallel corpora to improve in domain translation model through interpolation. Figure 2.3 shows the example sentence "serve the rice hot" by using phrase-based translation to select the most probable word in the Malay language.

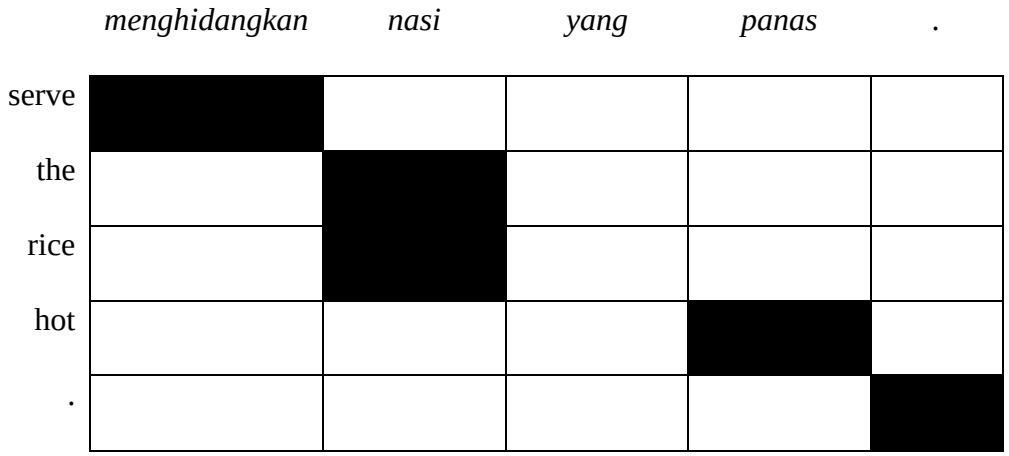


Figure 2.3 Example of phrase-based translation.

The phrase-based translation (more than one source word aligned to (→) more than one target word) alignment results for the example sentence pair above:

- serve → *menghidangkan*
- null → *yang*
- the rice → *nasi*
- hot → *panas*

- . → .

2.1.3 Neural Machine Translation (NMT)

In general, a neural network consists of connected neurons. A basic neuron n_i with input x_j is multiplied by the weight w_{ij} and summed as z. The value of z is normalized to a value between 0 and 1 using a logistic function (like tanh function and ReLU function) to get output h. Typically, more than one neurons are used for modelling and prediction.

$$z = \sum w_i x_i \tag{2.5}$$

$$h = \text{logistics}(z) \tag{2.6}$$

Besides, the output of a neuron can be input into the next neuron, and this forms a feedforward neural network. The modelling of neural network parameters is done by using the backpropagation algorithm. In the beginning, the neural network will be initialized with the appropriate values. Then, training data (for example x_1 and x_2) are inserted into the neural network from the outer neural layer, and the predictions (for example h_1, h_2 and h_3) produced are compared to the expected values found in the training data. The difference between predictions and expectations in training data will be calculated (usually with mean square errors or cross-entropy) and backpropagated so that the neural network parameters are altered to reduce the difference between forecasts and expectations. This process is repeated until converges.

On the other hand, a recurrent neuron looks like a typical neuron, but it has an additional feedback loop to allow present information to be used for the subsequent neuron to make decisions. An RNN can be considered as multiple copies of the same recurrent neurons that convey information to themselves as shown in Figure 2.4. There

are many variations of a recurrent neuron. Figure 1.4 shows a simple recurrent neuron where x_t is inputted and produces output y_t. The output y_i is also feedback to the subsequent neuron. In some other types of recurrent neurons known as a recurrent cell, a state hi instead of output y_i, which is the function of input x_t and previous state h_{i-1} is feedback to the next neuron.

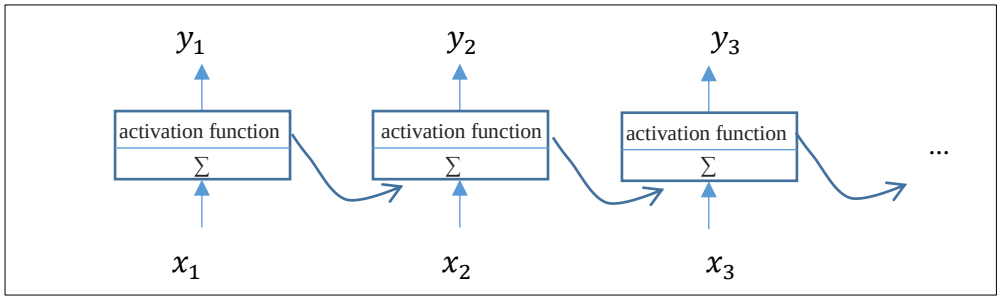


Figure 2.4 A layer of the recurrent neuron (Geron, 2017).

2.1.3(a) Encoder-Decoder Model

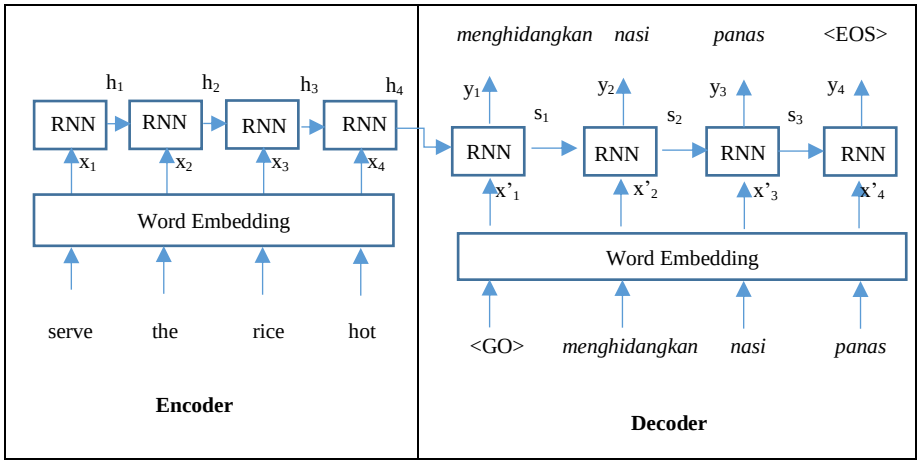
NMT aims to directly model the conditional probability p(y|x) of translating a source sentence, $x_1,\ldots,x_n$, to a target sentence, $y_1,\ldots,y_m$. It accomplishes this goal through an encoder-decoder framework (Kalchbrenner and Blunsom, 2013; Sutskever er al., 2014; Cho et al., 2014). The encoder computes a representation $\boldsymbol{\delta}$ for the source sentence. Based on that source representation, the decoder generates a translation, one target word at a time, and hence, decomposes the log conditional probability as:

$$\log p(y|x) = \sum_{t=1}^{m} \log p(y_t|y_{<t}, \delta) \tag{2.7}$$

NMT collects information from the entire source sentence before translating. NMT can capture long-range dependencies in languages. The RNN encoder-decoder architecture is widely used in NMT (Sutskever et al., 2014). Figure 2.5 shows an example where it is used for decoding.

Figure 2.5 An encoder-decoder RNN for decoding (Sutskever et al., 2014).

The encoder-decoder RNN can be visualized as two RNNs, namely the RNN



encoder and RNN decoder. The RNN of the encoder looks like a typical RNN except that the output is ignored. A word embedding module will convert a word to a vector. The vectors are then inputted to the encoder. The RNN of the decoder converts the vector h received from the encoder to words. Notice that the vector of the tag <GO> will be entered as the first input, x_1 into the decoder to initiate the generation of a translation. During testing, the decoder will predict vector y_t. The output vector will be used as an input, x_{t+1} to the next cell in the decoder. This process repeats until the <EOS> tag is generated.

During training, the decoder generates a sequence of predictions ($\hat{y}_1$, $\hat{y}_2$,..., $\hat{y}_T$),

where T is the length of the output sequence. Each $\hat{y}_t$ is the predicted probability distribution over the entire vocabulary, typically obtained using a softmax function. The loss function is encoder-decoder models is typically cross-entropy loss computed at each time step of the output sequence. The goal is to minimize the difference between the predicted distribution $\hat{y}_t$ and the actual ground truth token y_t. The cross-entropy loss for each time step is given by:

$$L_t = -\sum_{i=1}^{V} y_{t,i}\, log(\hat{y}_{t,i}) \tag{2.8}$$

Where $y_{t,i}$ is the one-hot encoded ground truth token at time step t, with i indicating the correct token. $\hat{y}_{t,i}$ is the predicted probability of token i from the softmax output of the decoder. The total loss for the entire sequence is the sum of the losses at each time step:

$$L = \sum_{t=1}^{T} L_t \tag{2.9}$$

2.1.3(b) Long Short-Term Memory (LSTM)

LSTM is a type of RNN proposed by Hochreiter and Schmidhuber (1997) to improve the modelling of long-term dependencies and reduce the vanishing gradients problem. Figure 1.6 shows an LSTM cell. In the LSTM cell, there is a hidden state, $h^{(t)}$ and a cell state, $c^{(t)}$ and both are vectors length n. The cell stores long-term information and LSTM can erase, write and read information. The selection of which information is erased/written/read is controlled by three corresponding gates. The gates are dynamic where their value is computed based on the current context.

i) Forget gate: controls what is kept and what is forgotten from the previous cell state. The equation of calculation is shown below: