SENTENCE-BASED ALIGNMENT FOR PARALLEL TEXT CORPORA PREPARATION FOR MACHINE TRANSLATION

BY

LEE YONG WEI

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION TECHNOLOGY (HONOURS)

COMPUTER ENGINEERING

Faculty of Information and Communication Technology
(Kampar Campus)

JANUARY 2021

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: Sentence-Based Alignment for Parallel Text Corpora Preparation for Machine Translation

Academic Session: JANUARY 2021

I LEE YONG WEI

declare that I allow this Final Year Project Report to be kept in Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,

_____ _____

(Author's signature) (Supervisor's signature)

Address:
1C, LORONG PALA 30,
TAMAN TELUK INTAN,
36000 TELUK INTAN, PERAK.

Jasmina Khaw Yen Min
Supervisor's name

Date: 15/4/2021 **Date**: 15/4/2021

SENTENCE-BASED ALIGNMENT FOR PARALLEL TEXT CORPORA PREPARATION FOR MACHINE TRANSLATION

By

Lee Yong Wei

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION TECHNOLOGY (HONOURS)

COMPUTER ENGINEERING

Faculty of Information and Communication Technology
(Kampar Campus)

JANUARY 2021

DECLARATION OF ORIGINALITY

I declare that this report entitled "**SENTENCE-BASED ALIGNMENT FOR PARALLEL TEXT CORPORA PREPARATION FOR MACHINE TRANSLATION**" is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature :

Name : LEE YONG WEI

Date : 15/4/2021

ACKNOWLEDGEMENT

I would like to express our deepest gratitude and appreciation to our supervisor, Dr Jasmina Khaw Yen Min, who had spent his valuable time and patient, standing by my side, guiding me hands to hands in completing this research. I truly appreciate her selfless guidance. A million thanks to you.

Besides, I would like to thank my beloved university, University Tunku Abdul Rahman for giving me this golden opportunity to conduct this project. It is undeniably I we had learnt a lot of knowledge that I am unable to gain from the textbooks. Moreover, it also enabled me to have better understanding of my research topic and experiences that are precious and will be useful to our future.

Last but not least, I have to thank my parents and my family for their unconditional love, full support and always staying by my side throughout this course. It is such a precious moment for me to remember deep in my heart. Once again, a million thanks to those who had given me a helping hand throughout this research.

ABSTRACT

In the age of technology, we are living in a world that is widely related to Natural Language Processing (NLP) as NLP helps in downstream applications like speech recognition, machine translation and so forth. Machine translation is important in our daily life as it is faster to translate a large number of texts compared to human translators. With the aids of machine translator, it definitely saves a lot of our times. Besides, it is also cheaper than using a human translator. In machine translation, parallel corpus plays a significant role as a resource for translation training and language teaching. A good quality of parallel corpus will greatly increase the accuracy of the machine translation. Hence, sentence-based alignment for parallel text corpora plays an important role in helping NLP especially for machine translation. However, there are limited resources on parallel corpus for some selected source language and target language. Furthermore, the accuracy of machine translation on some target languages is still low. Therefore, an approach of generating parallel corpus on source language and target language is proposed. In this study, parallel corpus of English (source language) and Malay (target language) are collected. Besides, a machine translation is developed using recurrent neural network (RNN) model of neural network translation. An accuracy of training with 0.9 is obtained from the model. Besides, the translated Malay text achieved BLEU score of 0.65 which is considered a good score.

TABLE OF CONTENTS

DECLARATION OF ORIGINALITY iv
ABSTRACT vi
LIST OF FIGURES ix
LIST OF TABLES x
LIST OF ABBREVIATIONS xi
CHAPTER 1 INTRODUCTION 1
1.1 Introduction 1
1.2 Problem Statements 1
1.3 Motivation 2
1.4 Project Scope 2
1.5 Project Objectives 3
1.6 Background Information 3
1.7 Proposed Approach 4
1.8 Highlights 5
1.9 Report Organization 5
CHAPTER 2 LITERATURE REVIEW 6
2.1 Literature Review 6
2.2 Parallel Text Collection 6
2.2.1 Parallel Text Collection Using Subtitles 6
2.2.3 Parallel Text Collection Using Transcribe and Transcription of Conversion Recording 6
2.2.4 Parallel Text Collection Using News with Same Topic in Different Languages 7
2.3 Part-of-Speech (POS) Tagging Method 7
2.3.1 Model 1: Stochastic Model 7
2.3.2 Model 2: Transformation-Based Tagging 7
2.4 Sentence Based Alignment Method 7
2.4.1 Method 1: Length-Based Algorithm 7
2.4.2 Method 2: MT-Based Algorithm 8
2.4.3 Method 3: Lexical-Based Sentence Alignment 9
2.4.4 Method 4: Champollion Parallel Text Sentence Aligner Method 9
2.4.5 Method 5: Moore Parallel Text Sentence Aligner Method 10
2.5 Summary of sentence alignment method 11
2.5 Machine Translations 12
2.5.1 Statistical Machine Translation 12

2.5.2 Rule-Based Machine Translation 12
2.5.3 Hybrid Machine Translation 13
2.5.4 Neural Machine Translation 14
CHAPTER 3 SYSTEM DESIGN 15
3.1 Design Specification 15
3.1.1 Datasets Collection 16
3.1.2 Datasets Cleaning and Sentence aligned 16
3.1.3 Import Libraries 17
3.2 Implementation Issues and Challenges 21
CHAPTER 4 SPECIFICATIONS AND SYSTEM REQUIREMENTS 22
4.1 Tools to Use 22
4.2.1 Hardware 22
4.2.2 Software 22
CHAPTER 5 EXPERIMENTS AND RESULTS 24
5.1 Grammar of Malay language 24
5.2 BLEU score of the translation of Malay sentence text 25
6.1 Project review, contributions, and conclusion 29
6.2 Future works 30
REFERENCES 31
APPENDICES 36
WEEKLY REPORT 37
POSTER 40
PLAGIARISM CHECK RESULT 41
CHECKLIST 48

LIST OF FIGURES

Figure Number	Title	Page
Figure 1.1	Proposed Method System Design	4
Figure 2.2.1	RBMT Workflow Pipeline	13
Figure 3.1	Flow Chart of the Implementation of the Model	15
Figure 3.2.1	Sample output of cleaned and sentence-aligned text	16
Figure 3.2.2	Sample of Sentence-aligned Text	17
Figure 3.2.3	Sample of Tokenized Sentence	18
Figure 3.2.4	The Data Pre-processed	18
Figure 3.2.5	Model	20
Figure 3.2.6	Sample output	21
Figure 5.1	Sample of Namanya	24
Figure 5.2	Sample of membantu	24
Figure 5.3	Sample of tinggalkan	24
Figure 5.4	Sample of buku buku	25
Figure 5.5	Sample of segala galanya	25
Figure 5.6	Sample output for calculated BLEU score	26
Figure 5.7	Result of model training	28

LIST OF TABLES

Table Number	Title	Page
Table 2.1	Summary of Sentence Alignment Method	11
Table 4.2.1	Recommended Hardware Specification	22
Table 4.2.2	Current Hardware Specification	22
Table 5.1	BLEU score of fifty translated texts	26

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
RNN	Recurrent Neural Network
BLEU	Bilingual Evaluation Understudy
SMT	Statistical Machine Translation
POS	Part of Speech Tagging
TBL	Transformation-based Learning
MT	Machine Translation
RBMT	Rule-based Machine Translation
HMT	Hybrid Machine Translation
NMT	Neural Machine Translation
LSTM	Long-Short-Term Memory
ReLU	Rectified Linear Unit
RAM	Random Access Memory
GPU	Graphic Processing Unit
GRU	Gated Recurrent Unit

CHAPTER 1 INTRODUCTION

1.1 Introduction

Literally, a parallel text is a set of text formed along with its translation which sometimes it is known as bitext (Chan, 2015). Parallel text alignment which is the process of identifying the corresponding sentence in both sides of the parallel text. For instance, the parallel text between the language of English and Malay. It is undeniable that sentence-based alignment for parallel text plays a vital role in helping natural language processing (NLP) task and any processing including translation studies, machine translation, corpus linguistic, cross-language information retrieval and so forth. The process of manual sentence alignment could be pricey, tiresome and time consuming. Thus, there is the need of automated sentence alignment techniques to ease human's life. Besides, sentence-based alignment parallel text is also significant in the approaches to the preparation for machine translation.

1.2 Problem Statements

a. Limited resources on parallel corpus for selected source language and target language

Since parallel corpus is significant for machine translation, the problem of lacking parallel corpus in some language pairs can greatly reduce the quality of machine translation (Karakanta et al, 2017). Hence, it is time and labour consuming in manual transcription to produce a huge number of parallel corpus for some of the low resource language pairs of parallel corpora. Therefore, parallel corpus is needed to be collected as training dataset for machine translation without manual transcription.

b. Low accuracy of machine translation on some target languages

Although there are a lots of free resources machine translation tools online, there is still a problem of low accuracy of machine translation on some targeted languages. For some of the uncommon and low resource of language such as Dante, the accuracy of the translation is not that good and will produce a strange translation (Shaymaa, 2016). Thus, in this project, we will develop a machine translation with the case study on English and Malay language.

1.3 Motivation

Nowadays, sentence alignment parallel text resources are the basis to some of the methods in NLP (Natural Language Processing), machine translation and so forth. Hence, there are lots of resources of sentence-based alignment parallel texts in the network. However, there are still some languages that has only little or poor and inaccurate resources. Thus, in order to provide an indispensable training dataset for the machine translation and other fields of NLP. Besides, to improve the quality in concern of the accuracy of the translation, more sentence- based alignment parallel text corpora is needed in different languages. Hence, this motivated the development of sentence-based alignment.

1.4 Project Scope

One of the focus in our project scope is the collection of parallel text to be used in sentence-based alignment for the preparation of machine translation. Firstly, the parallel text of English-Malay language pair will be collected for the preparation of the sentence-based alignment process. Next, cleaning for the datasets will be carried out. The extra timestamps, tag and other noises will be cleaned in order to get a better-quality result during sentence alignment.

After collecting the parallel text, the next focus in our project scope will be performing sentence aligning for the sentence pairs of the datasets. Later, sentence-based algorithm will be run in order to sentence align the text. After that, a machine translation will be used to train the sentence-aligned text.

Moreover, the next focus is the translation of the parallel text will be carried out using simple sentences. The accuracy of the trained model will be found out after training the model and the BLEU score is used to determine the quality of the translated text by comparing to the existing approach.

1.5 Project Objectives

The objectives of the project are as below:

a. To collect parallel corpus as training dataset for machine translation without manual transcription.
b. To develop a machine translation with case study on English and Malay language.

1.6 Background Information

In this era of technology, the field of parallel texts have a sustainable growth and these indirectly provide a plenty of resources for machine translation as sentence alignment has the ability to resolve the issue that the size of the largely translated segments forming the parallel corpora and the ambiguity of the tasks of learning. Thus, they are lots of algorithm for sentence-based alignments with the rapid rise of online parallel texts nowadays. Unfortunately, there are limitations in some of the methods. For an example, the accuracy, speed, and the robustness of the algorithm to tackle noise normally present in the datasets.

A sentence-based alignment algorithm has to be accurate and quick. In fact, both the precision and the recall rates of the result should be high to the concern of the quality of the alignment result. Besides, there is also limited resources for some languages. Thus, a sentence-based alignment for parallel text corpora preparation for machine translation with accuracy, speed and robustness of the algorithm will be developed in this project with the case study of aligning the language of English to Malay.

The interest of bitext began at around sometime in the mid-eighties, at which the time in many places were pursuing for independent efforts concurrently and this is most notably at Xerox PARC (Kay & Roscheisen, 1993). In this earlier time. The early efforts were all on sentence alignments instead of bitext. There were two that described the alignment methods that were based on a statistical translation that focused only on the length of the text segments which are the sentences and paragraphs and it was all depend on the dynamic programming in order to seek for the most similar alignment. One of them proposed the method of counted words (Brown et al, 1991) while the other one counted characters (Gale & Church, 1993).

1.7 Proposed Approach

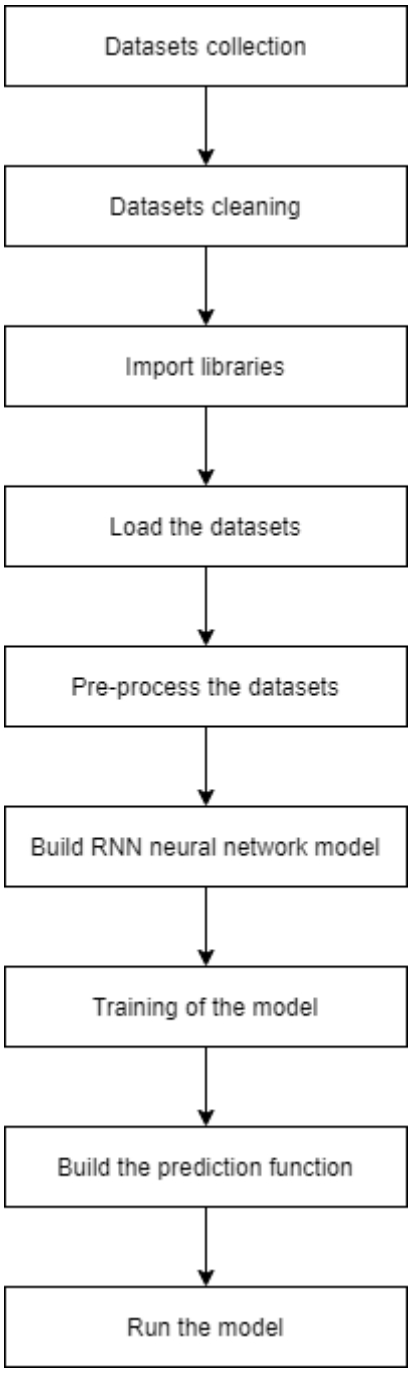


Figure 1.1 Proposed Method System Design

Figure above shows the flow of the proposed approach. Firstly, the English and Malay datasets are collected using subtitles. After, the raw datasets are collected, they are cleaned to reduce noises that will affect the performance of the model. When the sentence-aligned parallel text datasets are prepared, libraries that are needed to train the model of machine translation are imported. Later, the RNN model of machine

translation is built and compiled. We then train the model and build the prediction function. Lastly, we run the machine translation built using the sentence-aligned datasets to perform translation from English text to Malay text.

1.8 Highlights

The machine translation built using the sentence-aligned text datasets achieved a BLEU score of 0.65 which is only 0.15 lower to a well-known online machine translator, Bing Translator. The BLEU score of our machine translation is considered good. Besides, the model also achieved an accuracy of 0.9. Hence, the parallel text is able to enrich the English and Malay language resources.

1.9 Report Organization

This report is organized in a structure of 6 chapters. Literature reviews on different approaches on collecting datasets, sentence alignment and machine translation has been reviewed in chapter 2. Besides, the details of system design and implementation of the project are further discussed in chapter 3. In chapter 4, the system requirements and the specifications are discussed. In chapter 5, experiments and the results of the projects have been evaluated and justified. Lastly, we have concluded this project and the future works are also discussed in chapter 6.

CHAPTER 2 LITERATURE REVIEW

2.1 Literature Review

According to what was mentioned above, there are lots of methods used for the sentence-based alignment in the wake of advancement in technology. However, these methods have their strengths and weaknesses. Hence, in this chapter, the strength and weakness of the methods will be reviewed in order to analyse the pros and cons of the existing methods in the development of sentence-based alignment. Besides, various kinds of machine translation will also be reviewed. Nevertheless, we will also review the methods for parallel text collection in this chapter.

2.2 Parallel Text Collection

2.2.1 Parallel Text Collection Using Subtitles

Since the resources of online databases of videos subtitles from different genres are growing substantially as the advancement of technology. This is because subtitles from various languages are provided voluntarily by the Internet users. Hence, there is a large number of subtitles on the network that is free to download which ease the job of collecting parallel text. Moreover, for the same video, translators sometimes allow diverse subtitle versions of the same language. Subtitles are able to link to the actual sound signals too as most of them are transcription of spontaneous speech (Tiedemann, 2007).

2.2.3 Parallel Text Collection Using Transcribe and Transcription of Conversion Recording

Audio is definitely a good choice in the process of transcribing and translating the audio for the use of parallel text collection as it requires lower cost equipment, less expertise and it produces smaller data files. The audio obtained is then transcribed by an automatic speech recognition system and later a baseline SMT system will be used to translate it.

2.2.4 Parallel Text Collection Using News with Same Topic in Different Languages

With the advancement of news aggregator services, one can easily get the documents in the network. One of the platforms that can be used is “Google News”. With “Google News”, one can readily collect multiple new stories covering the same news item with different languages (Dolan et al., 2004). By using such resources, further aligning related documents at a finer level of resolution has to be carried out, identifying which sentences from a document is align with the sentences from the other.

2.3 Part-of-Speech (POS) Tagging Method

2.3.1 Model 1: Stochastic Model

One of the techniques for tagging the language text is Stochastic POS Tagging. It is also known as statistical model. This model assumes that every word is known and possesses a finite set of tags that is possible. These kinds of tags can be extracted from a dictionary or even from morphological analysis. This model allows us to identify the optimal sequences of POS tags $T = t_1, t_2, \ldots\ t_n$, with a given sequence of words $W = w_1, w_2, \ldots\ w_n$. This model is the simplest POS tagging as it selects the most common tags associated with a word in the corpus that is used for training.

2.3.2 Model 2: Transformation-Based Tagging

This model is an example of transformation-based learning (TBL) and is known as Brill tagging (Brill, 1992). It is a rule-based algorithm for tagging POS to the text automatically. Since this model is fully automated, it is different from the trainable stochastic tagger, where linguistic information is directly encoded in a basic non-stochastic rule.

2.4 Sentence Based Alignment Method

2.4.1 Method 1: Length-Based Algorithm

The first length-based algorithm was proposed by (Brown et al., 1991). Basically, this algorithm aligns sentences based on the length of the sentence that measured by word or character, which uses the idea of long or short sentences will be translated into long or short sentences. By assigning a probabilistic score to every proposed sentence based on the variance of the scaled difference of the lengths of both sentences, this algorithm

can provide a good alignment on those language pairs with high length correlation. The maximum likelihood for the alignment of sentences can also be found by using the probabilistic score in dynamic programming framework. For instance, French and English pairs can perform in a high-speed using this algorithm.

The advantages of using this algorithm are that it is excellent in terms of its performance and speed if there is only a minimal of noise in the input bilingual texts.

In contrast, the weakness of this algorithm is that it is not robust compared to other algorithms as it only uses the sentence length information. The sentence length will eventually become unreliable when there is too much noise in the bilingual texts.

2.4.2 Method 2: MT-Based Algorithm

Bleualign is a MT- based sentence algorithm that search for the source text on an MT translation instead of directly computing an alignment between the source text and the target text. (Senrich & Volk, 2010) showed that Bleualign can align text robustly compared to other algorithms. This algorithm applies dynamic programming to find the way to maximize the BLEU score between the translation source text and the target text. However, the quality of this algorithm depends on the performance of the SMT system used. Besides, according to Senrich and Volk (2010), a quality test on their MT-based algorithm had been demonstrated and they obtained a result that the quality of the algorithm is directly proportional to the MT system. Hence, if a superior MT system can be produced using Bleualign, the quality of the result will definitely be high.

The strength of Bleualign is it can robustly align the target text and the source text which is on top of the other algorithms.

The main weakness of Bleualign is that its performance is highly dependent on the translation provided. MT-based algorithm requires an existing MT system to make the algorithm attractive. In contrast, if the language pairs are poor in resources, this requirement will definitely degrade the quality of the result.

2.4.3 Method 3: Lexical-Based Sentence Alignment

According to the research of Chen (1993), they described a fast algorithm using lexical information for sentence alignment. This algorithm constructs a simple statistical word-to-word translation model while processing sentence alignment. This algorithm maximizes the probability of producing the corpus with this translation model. Dynamic programming with thresholding is used in the search strategy. With thresholding, a corpus need not to be subdivided into smaller chunks as the search is linear in the length of corpus.

One of the pros of this algorithm is that it is able to manage large deletions in text. Besides, this algorithm is language independent and parallelizable. In fact, lexical-based sentence alignment algorithm offers a higher accuracy in concern of the quality of the result.

However, lexical-based algorithms are not efficient enough to manage large corpora. Moreover, using lexical information in the alignment process requires a large computational cost.

2.4.4 Method 4: Champollion Parallel Text Sentence Aligner Method

Champollion is a lexicon-based method designed for robust alignment of potential noisy parallel text (Ma, 2006). There are two ways of Champollion difference from other algorithms. Firstly, Champollion assumes a noisy input. For an example, a high percentage of alignments will not be one to one alignment and the number of insertions and deletions will be vital. This assumption is against declaring a match in the absence of lexical evidence. It is claimed that non lexical measures are often unreliable when dealing with noisy data (Ma, 2006). Next, Champollion is different from other traditional lexicon-based approaches in distributing weights to translated words. The basis of Champollion is to assign greater weight to less frequent translation pairs. There is a function used by the Champollion to compute the similarity between any two segment which each of them consists of one or more sentences. Champollion has a penalty associated with alignments other than 1-1 alignment which is determined empirically. Furthermore, sentences that contain mismatching length will also be

penalized. Later, dynamic programming method is used to obtain the optimal alignment which maximizes the similarity between the source text and the translation.

The major advantage of Champollion is that it has the ability to cope with noisy data. Besides, it greatly increases the robustness of the alignment by assigning greater weights to less frequent translated words. Champollion is also a method that can be easily ported to new language pairs. Moreover, Champollion able to achieve high precision and recall on manually aligned Chinese-English parallel text corpus (Ma, 2006).

In contrast, Champollion considers a match possible only when lexical matches are present.

2.4.5 Method 5: Moore Parallel Text Sentence Aligner Method

A three-step process for aligning sentences was proposed by Moore (2002). Literally, this algorithm contains three steps. Firstly, a modified version of Brown et al.'s length-based model where search pruning techniques are used to speed up the discovery of the reliable sentence pairs is used to compute the coarse alignment of the corpus. The sentence pairs that have the highest alignment probability are collected to train a modified version of IBM Translation Model 1 (Brown et al., 1993). Lastly, by using the IBM model 1 score as an extra measure of parallelism, the whole corpus is realigned.

This method does not require any knowledge of the languages of the corpus except for breaking up the text into words and sentences. Besides, this method is able to achieve high accuracy at its modest computational cost. It is modest to sequences of inserted text, and it is fast.

However, the recall obtained is at most equal to the proportion of 1-to-1 correspondences contained in the parallel text to align especially problematic when aligning asymmetrical parallel corpora (Braune & Fraser, 2010, p.22).

2.5 Summary of sentence alignment method

Methods	Advantages	Disadvantages
Length-based algorithm	• High performance and speed	• Not robust
MT-based algorithm	• Robust	• Performance is highly dependent on the translation provided
Lexical-based sentence alignment	• Able to manage large deletions in text • Language independent • High accuracy	• Not efficient enough to manage large corpora • High computational cost
Champollion parallel text sentence aligner method	• Able to cope with noisy data • Robust • Easily ported to new language pairs • High precision and recall	• Depends on lexical information
Moore parallel text sentence aligner method	• Does not require knowledge for the language • High accuracy • Modest computational cost • Modest to sequences of inserted text • Fast	• Recall obtained is at most equal to the proportion of 1-to-1

Table 2.1: Summary of sentence alignment method

2.5 Machine Translations

2.5.1 Statistical Machine Translation

Statistical machine translation is also known as SMT. To define rules that are ideally suitable for target sentence translation, SMT implements predictive algorithm and statistical analysis (Mathias, 2015). It refers to statistical models which is based on the analysis of vast amounts of bilingual text (corpus) while performing it functions. This SMT automatically maps sentences in one language (source language) to another language (target language). Besides, SMT can be categorised into various subgroups which are phrase-based, word-based, syntax based and hierarchical phrase-based.

One of the strengths of SMT is its availability of algorithms and platforms as it can be easily found compared to other machine translation. Hence, compare to other machine translation models, one can easily train and add new languages. Besides, SMT needs less virtual space compared to other machine translation models. With this, the requirement to operate on system is smaller.

However, although SMT can perform well when the training corpora used is well defined such as technical texts, it will underperform when if the text given is in casual style like slangs. In addition, since statistical machine translation systems require bilingual co0ntent, it will be difficult when it comes to language pairs that are rare.

2.5.2 Rule-Based Machine Translation

Rule-based machine translation (RBMT) is a type of machine translation model that works based on grammatical rule. In order to perform translation, RBMT runs a grammatical analysis to the source and targeted language. However, a significant amount of human force is needed to prepare the rules and resources (Sreelekha, 2017). For instance, part-of speech tag (POS tag), morphological analysers, bilingual dictionaries and so on. It needs a long period of time to achieve efficiency since RBMT needs extensive proofreading and its dependency on lexicons is high. This model consists of three different phases which is analysis, transfer and generation. Figure 2.1 shows the workflow of the translation of RBMT.

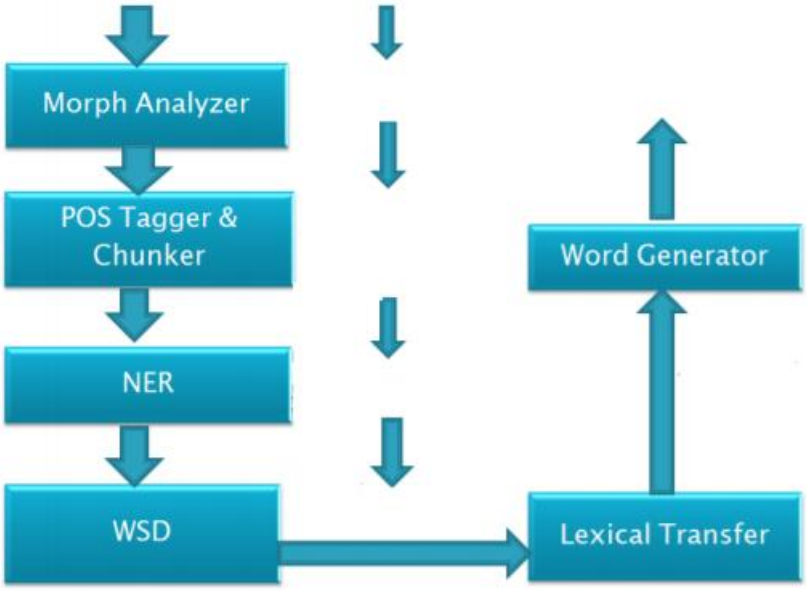


Figure 2.2.1 RBMT Workflow Pipeline

One of the advantages of RBMT system is that it can easily achieve high accuracy with only a narrow subsets of language pair. However, the construction of RBMT system requires a large number of linguistic resources and time consuming (Farrus et al, 2012). Hence, this system is expensive to build.

2.5.3 Hybrid Machine Translation

Hybrid machine translation which is known as HMT is a combination of rule-based machine translation (RBMT) and statistical machine translation (SMT). There are some reasons why HMT is implemented. According to Vilar et al. (2006), SMT which represents the empiricist method in machine translation, is more robust and able to provides translation in a better fluency due because of the use of better lexical selection and language models but this translation model faces the challenges of dealing with the requirements of linguistic knowledge like word order, morphology, and so on which causes the loss of adequacy. Besides, although RBMT is able to translate with better accuracy as it tries to represent every piece of the input during the translation, it is restricted due to the lexical selection in transferring and analysing the failure sentences. Hence, HMT is implemented to combine the strengths of both machine translation model.

The purpose of implementing HMT is to integrate the cores of machine translation models to enhance the performance of output. Besides, HMT can be built based on rule-based machine translation (RBMT). The translation process can be completed by coupling it with SMT or corpus-based machine translation, which is used to solve the problems with specific pieces (Xuan et al, 2012). As HMT leverages the translation memory, it has a better efficiency in terms of quality.

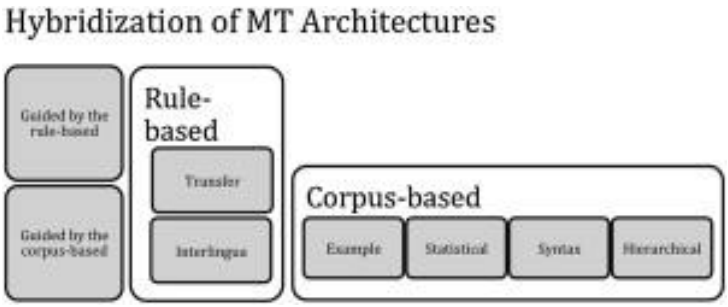


Figure 2.2.2 Architecture of HMT

2.5.4 Neural Machine Translation

Neural Machine Translation (NMT) is a machine translation model that relies on neural network models to build a statistical model for translation. This method employs an artificial neural network to predict the probability of a sequence of words by modelling entire sentences into an integrated model.

This NMT model is built as an encoder-decoder network with the recurrent neural network. This encoder works as a bidirectional neural network with gated recurrent units which reads the sentence inserted (input sentence). Later, it calculates a forward sequence of hidden state and a backward sequence. On the other hand, the decoder acts as a recurrent neural network that predict the targeted sentence. After that, each word in the target sentence is predicted based on the recurrent hidden state (Senrich et al, 2016).

The advantage of NMT is it offers a single system that can be trained to decipher both the source and target sentences. Hence, as compared to other machine translation models, it is not dependent on specialized systems.

CHAPTER 3 SYSTEM DESIGN

3.1 Design Specification

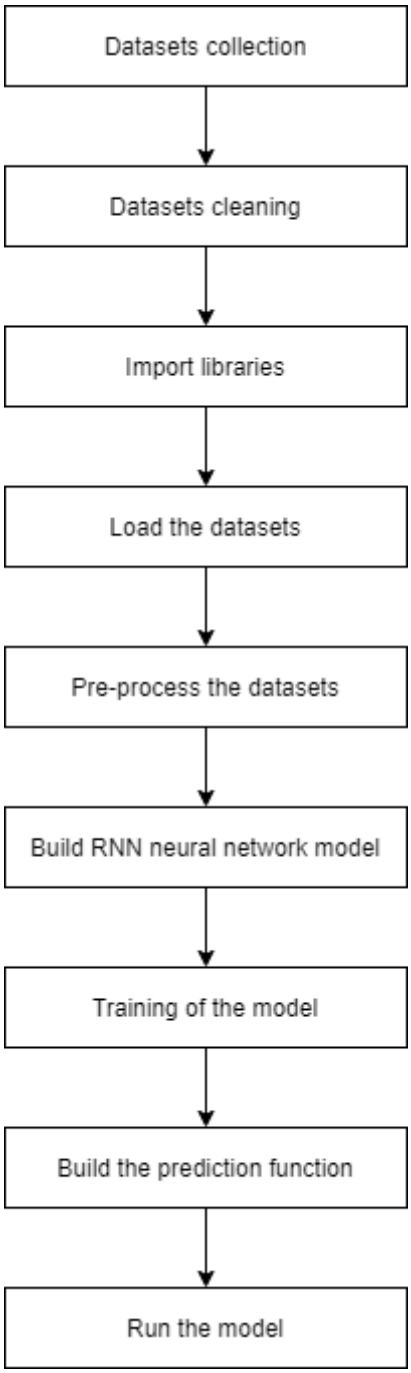


Figure 3.1: Flow Chart of the Implementation of the Model

3.1.1 Datasets Collection

Firstly, in this project, we decided to use the movies subtitles as our datasets. This is because the resources of online databases of movie subtitles from different genres are now enhancing as the advancement of technology. Subtitles from different languages are provided voluntarily by the Internet users. Thus, there is a lot of subtitles that can be download freely from the Internet. Besides, by using subtitles as our training data, it effectively decreases the effort of manual transcription.

3.1.2 Datasets Cleaning and Sentence aligned

Next, we perform datasets cleaning on the datasets collected. This process is to improve the performance as noises will greatly reduce the performance in terms of speed. The movie subtitles are downloaded in SRT format is stored in a folder. Later, the noises of the subtitles like the time stamp, scene number, HTML tags, and so on. After the noises are removed, the cleaned datasets are then output to another folder in txt format and is sentence aligned.

What if I told you...	Bagaimana jika saya memberitahu anda ...
I know what happens when you die?	Saya tahu apa yang berlaku apabila anda mati?
You become a ghost...	Anda menjadi hantu ...
trapped in a shadowland.	terperangkap di tanah bayang-bayang.
A world of whispers,	Dunia berbisik,
invisible except to other ghosts.	tidak kelihatan kecuali hantu lain.
Orphaned,	Yatim
unable to return to those you loved.	tidak dapat kembali kepada orang-orang yang anda
But with loneliness...	Tetapi dengan kesepian ...
comes freedom...	datang kebebasan ...
...to go where you please.	... untuk pergi ke mana anda sila.
Do what you want.	Buat apa yang anda mahu.
Ghosts have one power above all others:	Hantu mempunyai satu kuasa di atas semua yang la
to haunt the living.	untuk menghantui hidup.
Haunt them...	Menghantui mereka ...
for what they've done.	untuk apa yang telah mereka lakukan.
345-Tango-Tango.	345-Tango-Tango.
Tango-Tango, we have engine failure.	Tango-Tango, kami mempunyai kegagalan enjin.

Figure 3.2.1 Sample output of cleaned and sentence-aligned text

3.1.3 Import Libraries

Firstly, we need to import the libraries that are needed to build the machine translation. Below are the libraries needed:

i) OS - This OS module helps us to interact with our operating system. One of the examples of this is to interact with the file system. We need this module to import our datasets which are text files from our file.

ii) Re - This Re module helps regular expression to match the operations that is like those found in Perl. We need this module to perform the function to check whether the string matches with the given regular expression.

iii) Keras - This Keras module is a Neural Network library that runs on Tensorflow. We need this module to implement deep learning models like RNN models in our machine translation.

iv) Helper – This is a .py file created to perform the functions of loading in the datasets.

3.1.4 Loads the Datasets

Next, loads the sentence-aligned datasets prepared using the helper.py. The datasets consist of two text files which is en.txt which contains the English datasets and ma.txt which contains the Malay datasets. Both datasets consist of 138k sentence pairs and are sentence-aligned as shown in Figure 3.2.

```
Sample : 1
I know what happens when you die?
Saya tahu apa yang berlaku apabila anda mati?
---------------------------------------------------
Sample : 2
You become a ghost...
Anda menjadi hantu ...
---------------------------------------------------
Sample : 3
trapped in a shadowland.
terperangkap di tanah bayang-bayang.
---------------------------------------------------
Sample : 4
A world of whispers,
Dunia berbisik,
---------------------------------------------------
```

Figure 3.2.2 Sample of Sentence-aligned Text

3.1.5 Pre-process the Datasets

After that, we will pre-process the datasets. Firstly, we tokenize the sentences of the datasets into words and then ids using the Keras's tokenizer function. Figure 3.3 shows the example of tokenized sentence.

```
{'the': 1, 'quick': 2, 'a': 3, 'brown': 4, 'fox': 5, 'jumps': 6, 'over': 7, 'lazy': 8, 'dog': 9, 'by': 10, 'jove': 11, 'my': 1
2, 'study': 13, 'of': 14, 'lexicography': 15, 'won': 16, 'prize': 17, 'this': 18, 'is': 19, 'short': 20, 'sentence': 21}
Sequence 1 in x
  Input:  The quick brown fox jumps over the lazy dog .
  Output: [1, 2, 4, 5, 6, 7, 1, 8, 9]
Sequence 2 in x
  Input:  By Jove , my quick study of lexicography won a prize .
  Output: [10, 11, 12, 2, 13, 14, 15, 16, 3, 17]
Sequence 3 in x
  Input:  This is a short sentence .
  Output: [18, 19, 3, 20, 21]
```

Figure 3.2.3 Sample of Tokenized Sentence

Next, in order to make sure both the English and Malay sequences have the same length, we perform padding at the end of the sequences using Keras's pad_sequences function. This is done by padding 0 at the start of each sequence until each sequence achieves same length as the longest sequence.

Furthermore, after implementing the pre-process function, we are able to see the max sentence length and the vocabulary size of both languages in our datasets as show in Figure 3.4.

```
Data Preprocessed
Max English sentence length: 12
Max malay sentence length: 15
English vocabulary size: 16561
malay vocabulary size: 13636
```

Figure 3.2.4 The Data Pre-processed

3.1.6 Building of RNN Model

In our project, RNN model is used to build the machine translation. RNN which is also known as Recurrent Neural Network is a type of Neural Network that will feed the output of its previous step as an input to its current step. All the inputs and outputs of the traditional networks are mostly independent of one another. Hence, when it comes to cases like when it is needed to predict the next word of a sentence, there is a need to remember the previous words since the previous words are needed. Therefore, with the existence of RNN, this issue is solved with the help of Hidden Layer. The Hidden stated feature of RNN is able to remember some of the information of a sequence. This is because RNN has a memory that remembers all the information which was calculated. The same parameters will be used for each input since it performs the same function on every input to produce the output. With this, it has a simpler parameter compared to other neural networks.

However, this model has the disadvantages of vanishing gradient problem. Hence, we used GRU which is also known as Gated Recurrent Unit as a layer in the model to solve this problem. This GRU has fewer parameter compared to the long short-term memory (LSTM) and it has a forget gate. Thus, it is more computationally efficient compared to LSTM (Khandelwal, 2020). Besides, GRU has the ability to learn long term dependencies in a short time and remember information over a long-time step by deciding what to forget and what to remember. GRU uses reset and update gate. The reset gate plays the role to decide how to merge the new input with the memory of previous time step while the update gate decides the amount of memory to be kept. Besides, ReLU which is known as rectified linear unit is also used in building the model. This rectifier activation is easy to work with as it has a simple activation function. The advantages of ReLU compared to sigmoid and tanh activation is ReLU is not involve in expensive to compute function and is not saturate. Hence, with its simplicity, when it is used together with the dropout regularization, it is able to perform well in many tasks. This makes it suitable to work with multiple layered networks.

After that, we passed the output to the dense layer and finally the last layer with Softmax and compile with categorical cross-entropy loss. Figure 3.5 shows the model in a graphical way.

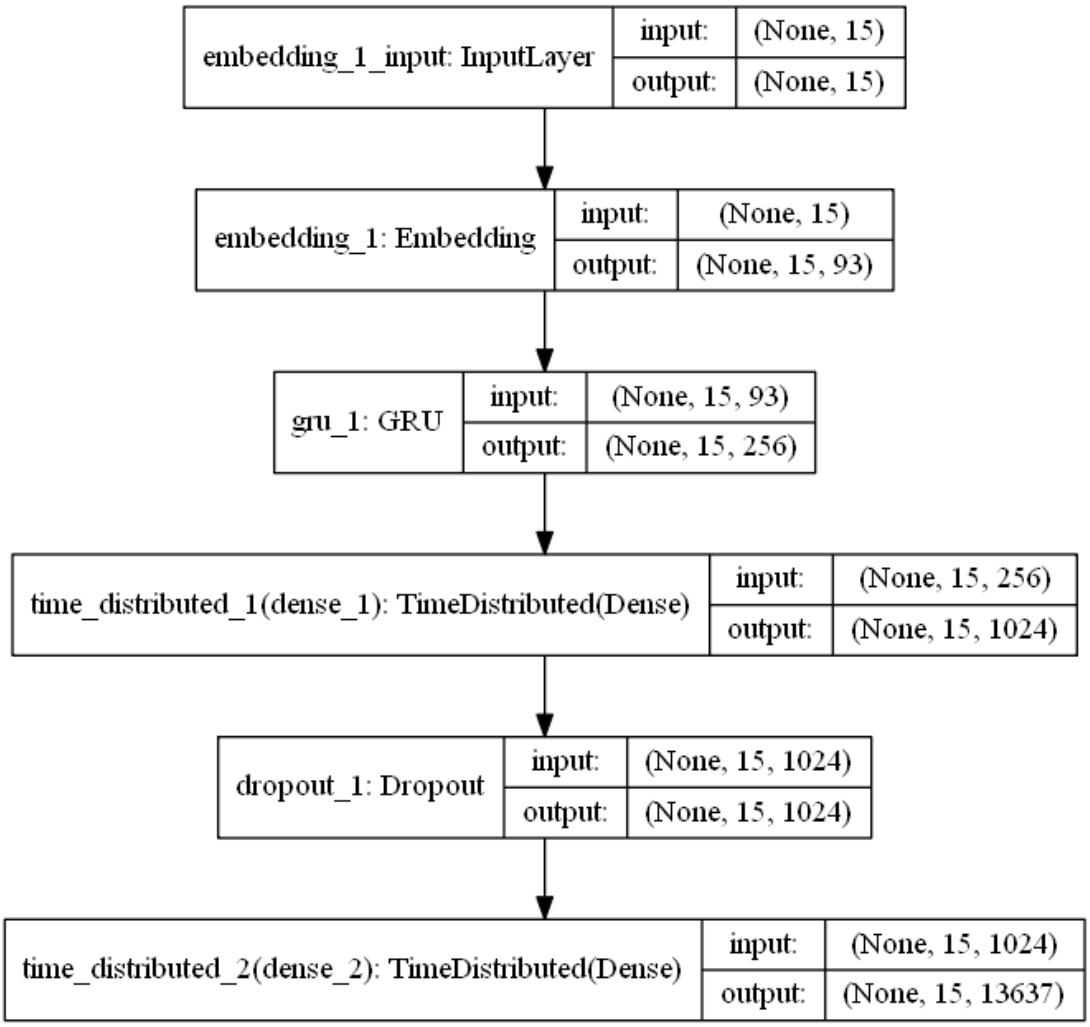


Figure 3.2.5 Model

3.1.7 Training of The Model

After building the model, we start to train the model and pass the English text and the maximum length of the sequence, with the vocabulary size for English and Malay text.

3.1.8 Build the prediction function and Run the translation

After training the model, we build the prediction function that can output the translation text (Malay text). At here, we will use the logits_to_text function as neural network will translate our English input into words ids that we cannot understand what the word actually is. This function will narrow the gap between the logits from the neural network to our Malay translation. Figure 3.6 shows the sample output of the translation.

```
In [20]: # Example of translation
         txt="I know".lower()
         final_predictions(re.sub(r'[^\w]', ' ', txt))

         (1, 15)
         saya tahu <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PA
         D> <PAD>
```

Figure 3.2.6 Sample output

3.2 Implementation Issues and Challenges

It is common that we will encounter some implementation issues and challenges while developing the system. In the stage of data collections, the challenges and issues that we encountered are the quality of the datasets. The quality of the datasets which is subtitle will be affected by the noises in the subtitle files. For instance, the time stamp, scene number, HTML tags, and so on. These noises will decrease the performance of the algorithm as those noises are not needed in the implementation of the system. Thus, the noises of the subtitle files must be reduced to produce a better quality of result.

Besides, we also encountered some challenges while dealing with the Malay datasets. This is because the resources for Malay languages is less and limited. Thus, it is definitely a challenge for us to collect either the library, dictionary or the datasets.

Moreover, in the stage of building the system, one of the challenges is the training of the RNN model is time consuming due to lacking GPU memory. It also used up a lot of random-access memory (RAM) of the device while training the model. This problem can be solved if we upgrade our hardware that have a GPU with greater internal memory.

Lastly, another challenges that encountered is overfitting. This happened when our model is trained too well with our datasets. However, the performance is not that food when testing with the new data. Hence, the datasets need to be separated into training and validation set. Our training set is used to train our model while validation set plays the role as an unseen sample which is later used to test the efficiency and validation of our model. Thus, with this, we can solve the issue of overfitting.

CHAPTER 4 SPECIFICATIONS AND SYSTEM REQUIREMENTS

4.1 Tools to Use

4.2.1 Hardware

Recommended setup to develop this system:

Operating System	Windows 7 64-bit and above
Processor	Intel(R) Core (TM) i5-460M @ 2.53GHz
RAM	8GB and above
Storage	15GB minimum (SSD is preferrable)
GPU Memory	4GB

Table 4.2.1 Recommended Hardware Specification

Current setup to develop this system

Operating System	Windows 10 Home 64-bit
Processor	Intel(R) Core (TM) i5-460M @ 2.53GHz
RAM	12GB
Storage	500GB SSD
GPU Memory	2GB

Table 4.2.2 Current Hardware Specification

4.2.2 Software

The software that has been chosen for developing this system:

1. Jupyter Notebook

   It is an open-source web application that enables us to share and create documents that contain live code, visualizations, equations and so on. It can also be used to perform machine learning, statistical modelling, data visualization and so forth.

2. Anaconda

   It is an open-source distribution of the Python and R programming languages for scientific computing such as machine learning, data science and many others. It aims to simplify the package management and deployment which is suitable for platform like Windows, Linux and MacOS.

3. Library Requirement

   TensorFlow

   Keras

   Numpy

   OS

   Re

   nltk

CHAPTER 5 EXPERIMENTS AND RESULTS

This chapter describes some of the experiments done on the machine translation and the results to access the performance of the project.

5.1 Grammar of Malay language

We have tested on some of the Malay grammar on our translation whether it is able to translate into the correct grammar. Below are some of our experiments.

1) Kata ganda nama diri orang ketiga

   Namanya

   Nama *+nya*

   Name his/her

   "his/her name"

```
# sentence 2
txt="his name is Stark".lower()
final_predictions(re.sub(r'[^\w]', ' ', txt))
```

```
(1, 15)
namanya <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
```

Figure 5.1 Sample of Namanya

2) Imbuhan(Affix) for Verb

   membantu

   mem+ *bantu*

   "help"

```
# sentence 3
txt="I can help you".lower()
final_predictions(re.sub(r'[^\w]', ' ', txt))
```

```
(1, 15)
saya boleh membantu anda <PAD> <PAD> <PAD> <PAD>
```

Figure 5.2 Sample of membantu

   Tinggalkan

   Tinggal *+kan* *dia*

   "Leave him"

```
# sentence 21
txt="So I leave him".lower()
final_predictions(re.sub(r'[^\w]', ' ', txt))
```

```
(1, 15)
jadi saya tinggalkan dia <PAD> <PAD> <PAD> <PAD>
```

Figure 5.3 Sample of tinggalkan

3) Kata Ganda

Buku buku

"Books"

```
# sentence 32
txt="I like books".lower()
final_predictions(re.sub(r'[^\w]', ' ', txt))
```

```
(1, 15)
saya suka buku buku
```

Figure 5.4 Sample of buku buku

Segala galanya

"Everything"

```
# sentence 42
txt="Thank you for everything".lower()
final_predictions(re.sub(r'[^\w]', ' ', txt))
```

```
(1, 15)
terima kasih kerana segala galanya <PAD> <PAD>
```

Figure 5.5 Sample of segala galanya

From the results above, we can conclude that the machine translation is able to translate the English text to Malay text with the correct basic grammar of Malay.

5.2 BLEU score of the translation of Malay sentence text

Bilingual Evaluation understudy which is also known as BLEU is an algorithm used to evaluate the quality of machine-translated text from a source language to a target language. The quality is determined as how similar is the machine-translated to a professional human translation. BLEU is an inexpensive and popular method to judge the quality of a translated text.

The scores are computed by comparing the candidate text with a set of high quality of reference translations of a language. Later, the scores are averaged over the whole set of corpuses in order to estimate the quality of the translation. The output of the BLEU score is always between 0 and 1. The value that is nearer to 1 represents the more similar texts while the value that is further from 1 has the lower score.

Literally, BLEU score is calculated with a couple of N-gram modified precisions as shown below:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

At here, p_n is the modified precision for the N-gram, the base for the log is e while w_n is the weight which is in between 0 and 1 for $logp_n$ and $\sum_{n=1}^{n} w_n = 1$. BP is defined as the brevity penalty which is used to penalize those short machine translations.

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases}$$

In our experiment, we have calculated the BLEU score by comparing a candidate with two references. Our candidate will be the Malay text translated by our proposed machine translator and Bing Translator. One of the references is the translated Malay text from the English language using Google Translate while the other reference is translated by human, which is by a professional of the language, a Malay student as Malay is his mother tongue.

```
# sentence 40
reference = [
    'saya tidur sepuluh jam'.split(),
    'aku tidur sepuluh jam'.split()
]
candidate = 'saya tidur sepuluh jam'.split()
print('BLEU score -> {}'.format(sentence_bleu(reference, candidate, smoothing_function=nltk.translate.bleu_score.SmoothingFuncti
```

```
BLEU score -> 1.0
```

Figure 5.6 Sample output for calculated BLEU score

Figure above shows one of the samples of the calculated BLEU score with two references where the first one is from Google Translate while the other one is translated by human. Fifty translated texts are selected to calculate the BLEU score. Table 5.1 shows the BLEU score of the selected translated text.

System	BLEU Score
Proposed Machine Translator	0.65
Bing Translator	0.80

Table 5.1 BLEU score of fifty translated texts

From the result, we found that the BLEU score of the Bing Translator is 0.15 higher than our machine translator. This can be justified by a few reasons. One of the reasons is the quality and the size of the datasets. The datasets of our project are collected from the online free resources' subtitles while the datasets of the Bing Translator are collected from various sources of bitext like articles, journals, news and so on. Moreover, the datasets collected using subtitles contain colloquial in the language. Hence, the model of the Bing Translator can learn better compared to the limited resources of our machine translator.

Besides, Bing Translator has various feedback features for the user which our machine translator does not equip with. For instance, Bing Translator introduced Collaborative Translations Framework that allows the user to suggest an alternative way of translations (Microsoft Translator 2010). With this, all the alternative translations are then integrated into their machine translation's algorithm and will be improved in the future translations. Hence, the Bing Translator is undoubtedly able to perform a better translation compared to ours due to the continuous improvement of algorithm.

Furthermore, Bing Translator is using statistical machine translation (SMT). SMT analyze the existing translations which is the bilingual text and later determining which rules are best for translating a given sentence. The more the data in the required languages are feed to the SMT, the higher the accuracy of the translation output. Since Bing Translator has a large amount of good quality of data, it fulfils the requirement to achieve a higher quality and accuracy of translations. In contrast, our approach is neural machine translation (NMT) which works by predicting a sequence of numbers when a sequence of numbers is provided (Sam, 2020). For instance, each of the word in the source language sentence is encoded as a number will be translated into an output sequence of numbers which is the translated target language sentence by the neural network. As neural network allows complex model, it has many parameters with weight

and biases to suit highly complex data and train complex model. This allows its model to generalize a large amount of training example like in our case, digesting millions of sentence language pairs. Thus, mostly NMT needs more corpus and resources when compared to SMT. However, our datasets have only around 110k training samples due to the limited free resources online. Hence, when comparing to the Bing Translator, our quality of translation will be lower.

5.3 System Performance

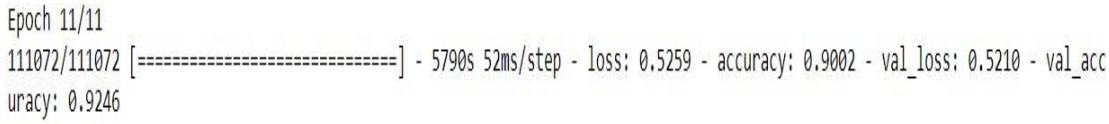


Figure 5.7 Result of model training

From the training result, it is notable that the training accuracy and the validation accuracy is high, which is 0.9002 for training accuracy and 0.9246 for validation accuracy. Besides, the difference between the training lost and the validation lost is small. Thus, the training loss and the validation loss is close to each other and we can deduce that our model is in good fit. With this, our sentence-alignment corpus can be used as a parallel corpus for machine translation.

CHAPTER 6 CONCLUSION

6.1 Project review, contributions, and conclusion

It is undoubtedly that parallel text resources play a vital role in the field of Natural Language Processing (NLP) as it is the basis of NLP. Although there are a lot of resources of parallel text available online in this advancement of technology, there are still some languages that has only limited resources compared to other languages. For instance, the language of Malay has only a little resource of parallel text online compared to languages like English, German and many other common languages. Thus, our project aims to contribute to the resources of Malay and English parallel text corpus for the preparation of performing NLP tasks like machine translation and so on. This aim is achieved as we have obtained a model with high accuracy of 0.9.

In this project, our datasets are collected from subscene.com which is a freely available subtitles download site in SRT format. As the downloaded subtitle files contain noises like HTML tag, timestamp, scene number and so on that will decrease the performance in terms of the quality of the result and the speed of the alignment process, those noises are cleaned and later output the subtitle in TXT format. At this point our objective to remove the noises of the datasets is achieved.

Besides, a lot of research on different approach of machine translations has been done to build the machine translation. In our project, RNN model neural machine translation is built to translate the English text to Malay text using our sentence-aligned parallel corpus. After that, the BLEU score for the translation text has been calculated and we obtained a score of 0.65 for our approach while 0.80 for the Bing Translator which is 0.15 lower. With this, our objective is also achieved since 0.65 is considered a good score.

Since our parallel corpus has achieved a high accuracy of 0.9 and a good BLEU score of 0.65, it is able to enrich the resources of bitext of the language of English and Malay. In short, this project has given me a golden opportunity to explore the possibilities in building the pipeline of the model of machine translation using sentence-based alignment corpus to perform translations.

6.2 Future works

The system of the project consists of limitations and hence it should not stop here and should be further improved in the future. One of the parts that can be improved is the size of the datasets. More resources of datasets can be collected in the future from time to time as more quality datasets can be collected for a longer period. With this, we can enrich the training model and further enhance the accuracy of the model and BLEU score of the translation texts.

Besides, different approach of machine translation can be used to produce a better translation text. The layers of the model can also be modified if there is any better approach. Lastly, new modules can be implemented to improve the accuracy and the context richness of current pipeline.

REFERENCES

Adafre, SF & Rijke, MD 2006, *Finding similar sentences across multiple languages in Wikipedia.* Available from <https://www.aclweb.org/anthology/W06-2810.pdf> [15 March 2021].

Braune, F & Fraser, A 2010, Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. *Coling'10: proceedings of the 23rd international conference on computational linguistic: posters.* pp.81-89.

Brown, PF, Lai, JC & Mercer, RL 1991, *Aligning sentences in parallel corpora.* Available from <https://www.aclweb.org/anthology/P91-1022.pdf> [15 March 2021].

Costa-Jussa, M.R., Farrus, M., Marino, F. & Fonollosa, J.A.R. (2012). Computing and informatics (Vol. 31). *Study and comparison of rule-based and statistical Catalan-Spanish machine translation systems*. pp.245-270.

Chan, First. The Routledge Encyclopedia of Translation Technology. New York: Routledge, 2015. Print. [15 March 2021].

Chen, SF 1993, *Aligning sentences in bilingual corpora using lexical information*. Available from <https://www.aclweb.org/anthology/P93-1002.pdf> [15 March 2021].

REFERENCES

Dimitrova, L., Koseska-toszewa, V., Roszko, D. & Roszko, R., 2015. Application of multilingual corpus in contrastive studies (on the example of the Bulgarian-Polish-Lithuanian parallel corpus). Cognitive Studies | Etudes cognitives, (10), pp. 217-239.

Dolan, B, Quirk, C & Brockett, C 2004, *Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources*. Available from <https://www.aclweb.org/anthology/C04-1051/> [15 March 2021].

Gale, W & Church, K 1993, 'A program for aligning sentences in bilingual corpora', *Computational Linguistic*, vol. 19, no. 1. Available from <https://www.aclweb.org/anthology/J93-1004.pdf> [15 March 2021].

Karakanta, A, Dehdari, J & Genabith, J. V. (2017). Neural machine translation for low-resource languages without parallel corpora. Available from <https://link.springer.com/content/pdf/10.1007/s10590-017-9203-5.pdf> [14 April 2021]

Kay, M & Roscheisen, M 1993, 'Text-translation alignment', *Computational Linguistic*, vol. 19, no. 1. Available from <https://www.aclweb.org/anthology/J93-1006.pdf> [15 March 2021].

Khandelwal, R. (2020). Intuitive explanation of neural machine translation. Towards data science. Available from

<https://towardsdatascience.com/intuitive-explanation-of-neural-machine-translation-129789e3c59f> [8 April 2021].

REFERENCES

Lei, S & Ming, Z 2008, 'Improved sentence alignment on parallel web pages using a stochastic tree alignment model', *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp.505-513.

Ma, X 2006, *Champollion: A robust parallel text sentence aligner*. Available from <http://www.lrec-conf.org/proceedings/lrec2006/pdf/746_pdf.pdf> [18 March 2021]

Mathias, J.A. (2006). An overview of statistical machine translation. Available from <https://www.researchgate.net/publication/287196051_An_Overview_of_Statistical_Machine_Translation> [12 March 2021]

Microsoft Translator 2010, Collaborative Translations: Announcing the next version of Microsoft Translator technology – V2 APIs. 15 March 2010. *Microsoft Translator: Blog.* Available from

<https://www.microsoft.com/en-us/translator/blog/2010/03/15/collaborative-translations-announcing-the-next-version-of-microsoft-translator-technology-v2-apis-and-widget/> [12 April 2021].

Mohammadi, M & Ghasem-Aghaee, N 2010, 'Building bilingual parallel corpora based on Wikipedia', *2010 Second International Conference on Computer Engineering and Applications,* 2, pp.264-268. Available from <https://ieeexplore-ieee-org.libezp2.utar.edu.my/stamp/stamp.jsp?tp=&arnumber=5445653> [18 March 2021].

REFERENCES

Peng, L, Maosong, S & Ping, X ,2010, 'Fast-champollion: a fast and robust sentence alignment algorithm', *Coling'10: proceedings of the 23rd international conference on computational linguistic: poster*, pp.710-718.

Sam, Y. 2020, What is neural machine translation & how does it work? 18 April 2020. *TranslateFX: Blog.* Available from

<https://www.translatefx.com/blog/what-is-neural-machine-translation-engine-how- does-it-work?lang=en> [5 April 2021].

Senrich, R., Haddow, B. & Birch, A. (2016). Neural machine translation of rare words with subword units. Available from

<https://arxiv.org/pdf/1508.07909.pdf> [10 April 2021].

Shaymaa, Y. (2016), Machine translation limits of accuracy and fidelity. Available from

<https://scholar.najah.edu/sites/default/files/Shaymaa%20Yousef%20Ibraheem%20Abdulhaq.pdf>

Sreelekha, S. (2017). Statistical vs rule-based machine translation; a case study on Indian language perspective. Available from

< https://arxiv.org/ftp/arxiv/papers/1708/1708.04559.pdf> [12 March 2021]

Tiedemann, J (2007). *Building a multilingual parallel subtitle corpus.* Available from <https://pdfs.semanticscholar.org/0d0e/b34ab56f7b48b6d611b5d0767bc59ba8b9fc.pdf?_ga=2.27720504.527688831.1586676870-1328383858.1586497823> [18 March 2021].

REFERENCES

Xuan, H. W., Li, W., & Tang, G. Y. (2012). An advanced review of hybrid machine translation (HMT). *Procedia Engineering, 29*, 3017-3022. Available from <https://doi.org/10.1016/j.proeng.2012.01.432> [10 April 2021]

APPENDICES

WEEKLY REPORT

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: 3,3	**Study week no.: 2**
Student Name & ID: Lee Yong Wei, 17ACB04464	
Supervisor: Miss Jasmina Khaw Yen Min	
Project Title: Sentence-based alignment for parallel text corpora preparation of machine translation	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

None

2. WORK TO BE DONE

Collecting datasets

3. PROBLEMS ENCOUNTERED

No

4. SELF EVALUATION OF THE PROGRESS

Keep going on

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: 3,3	**Study week no.: 6**
Student Name & ID: Lee Yong Wei, 17ACB04464	
Supervisor: Miss Jasmina Khaw Yen Min	
Project Title: Sentence-based alignment for parallel text corpora preparation of machine translation	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

Chapter 1

Chapter 2

Datasets collected

Datasets cleaned

2. WORK TO BE DONE

Build the model

3. PROBLEMS ENCOUNTERED

No

4. SELF EVALUATION OF THE PROGRESS

Keep going on

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project I / Project II)

Trimester, Year: 3,3	**Study week no.: 11**
Student Name & ID: Lee Yong Wei, 17ACB04464	
Supervisor: Miss Jasmina Khaw Yen Min	
Project Title: Sentence-based alignment for parallel text corpora preparation of machine translation	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

RNN model built

Training of the model

Built translation function

Chapter 4

2. WORK TO BE DONE

Complete FYP2

3. PROBLEMS ENCOUNTERED

No

4. SELF EVALUATION OF THE PROGRESS

Keep going on

Supervisor's signature

Student's signature

POSTER

SENTENCE-BASED ALIGNMENT FOR PARALLEL TEXT CORPORA PREPARATION FOR MACHINE TRANSLATION

INTRODUCTION:

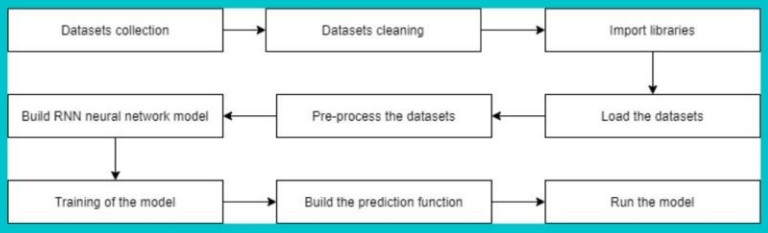
- Sentence-based alignment for parallel text corpora plays an important role in helping NLP especially for machine translation.
- This is because there are only limited resources on parallel corpus for English and Malay languages.
- There is also low accuracy of machine translation on some of the targeted languages.
- Thus, this project explores the preparation of sentence-based alignment parallel corpus and building the machine translation.

Objectives:

a. To collect parallel corpus as training dataset for machine translation without manual transcription.

b. To develop a machine translation with case study on English and Malay language.

System Design

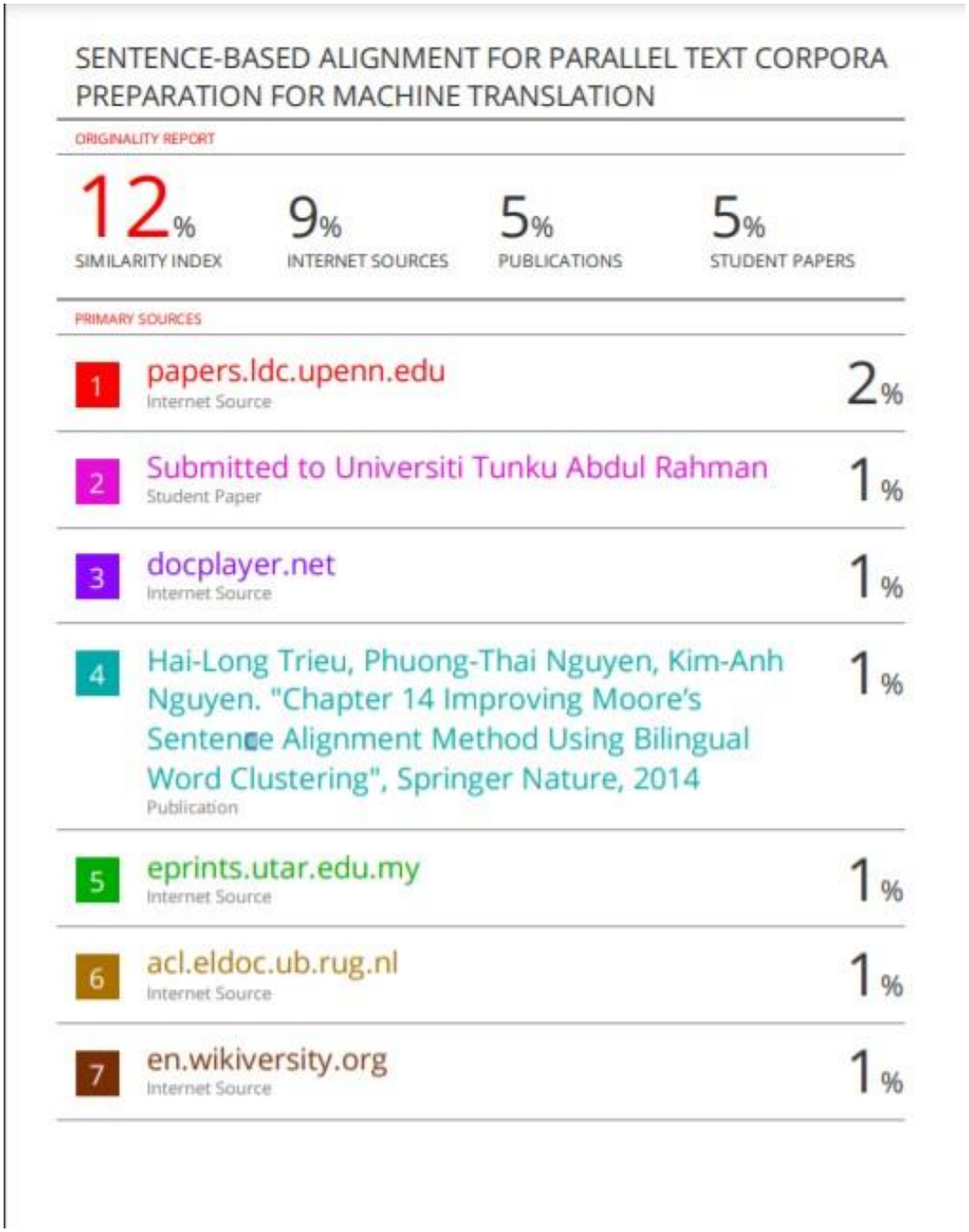


Experiments and Result

System	BLEU Score
Proposed Machine Translator	0.65
Bing Translator	0.80

- We have obtained a BLEU score of 0.65 for our proposed machine translator for an experiment of 50 translated texts.
- Our training model has achieved the accuracy of 0.9. With this, we are able to enrich the parallel text resources of English and Malay Language

PLAGIARISM CHECK RESULT



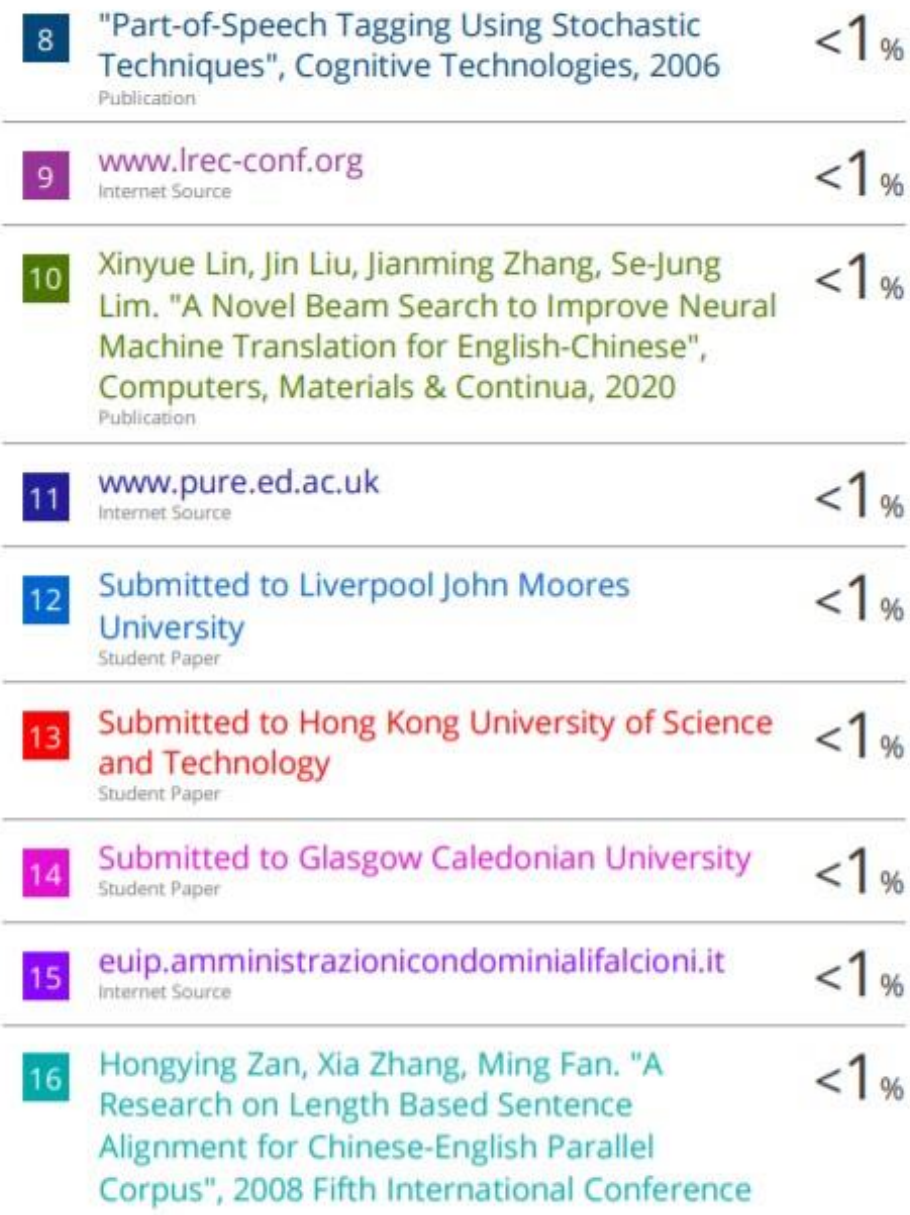
SENTENCE-BASED ALIGNMENT FOR PARALLEL TEXT CORPORA PREPARATION FOR MACHINE TRANSLATION

ORIGINALITY REPORT

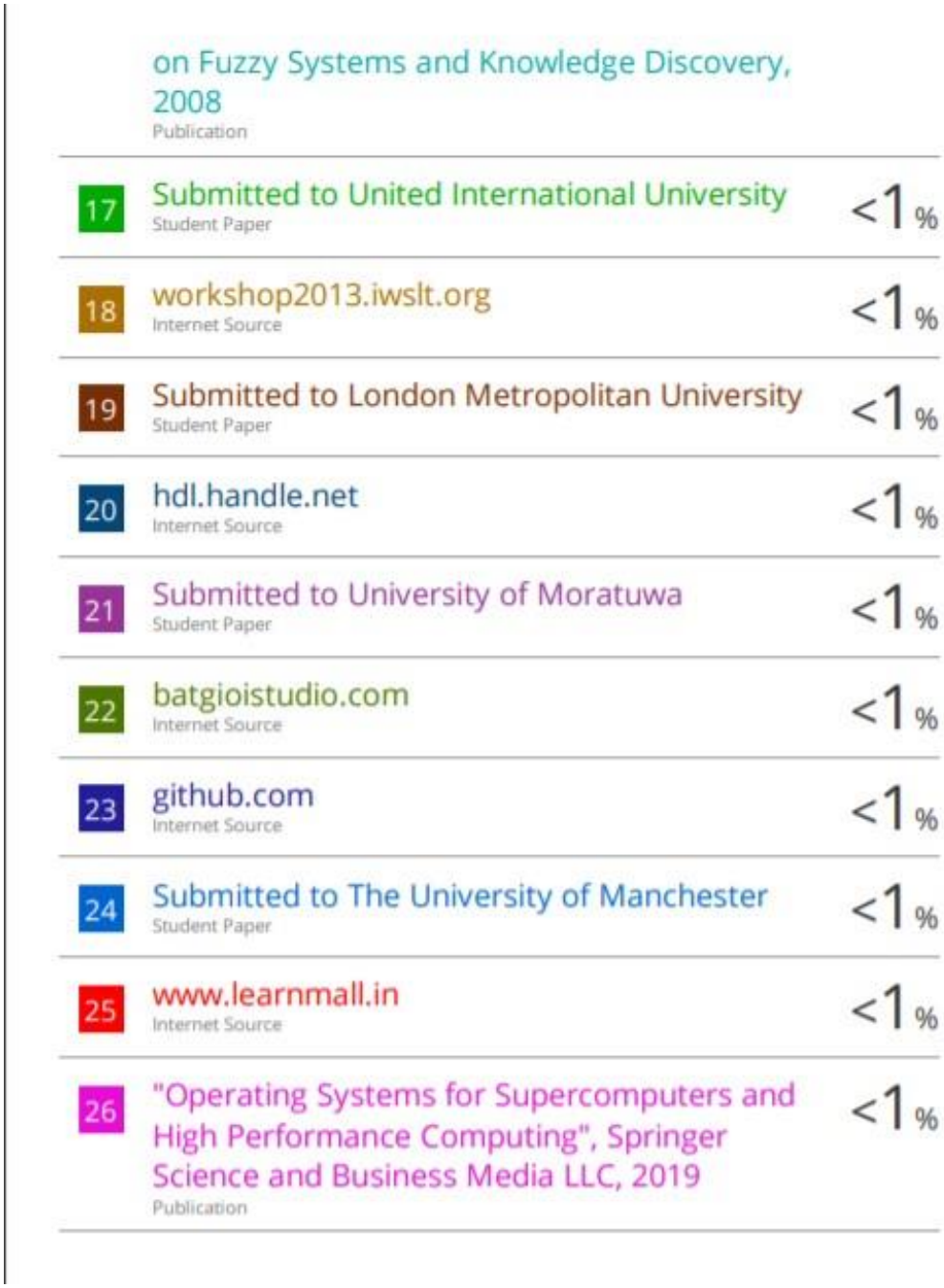
12%	9%	5%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	papers.ldc.upenn.edu Internet Source	2%
2	Submitted to Universiti Tunku Abdul Rahman Student Paper	1%
3	docplayer.net Internet Source	1%
4	Hai-Long Trieu, Phuong-Thai Nguyen, Kim-Anh Nguyen. "Chapter 14 Improving Moore's Sentence Alignment Method Using Bilingual Word Clustering", Springer Nature, 2014 Publication	1%
5	eprints.utar.edu.my Internet Source	1%
6	acl.eldoc.ub.rug.nl Internet Source	1%
7	en.wikiversity.org Internet Source	1%



8	"Part-of-Speech Tagging Using Stochastic Techniques", Cognitive Technologies, 2006 Publication	<1 %
9	www.lrec-conf.org Internet Source	<1 %
10	Xinyue Lin, Jin Liu, Jianming Zhang, Se-Jung Lim. "A Novel Beam Search to Improve Neural Machine Translation for English-Chinese", Computers, Materials & Continua, 2020 Publication	<1 %
11	www.pure.ed.ac.uk Internet Source	<1 %
12	Submitted to Liverpool John Moores University Student Paper	<1 %
13	Submitted to Hong Kong University of Science and Technology Student Paper	<1 %
14	Submitted to Glasgow Caledonian University Student Paper	<1 %
15	euip.amministrazionicondominialifalcioni.it Internet Source	<1 %
16	Hongying Zan, Xia Zhang, Ming Fan. "A Research on Length Based Sentence Alignment for Chinese-English Parallel Corpus", 2008 Fifth International Conference	<1 %



	on Fuzzy Systems and Knowledge Discovery, 2008 Publication	
17	Submitted to United International University Student Paper	<1 %
18	workshop2013.iwslt.org Internet Source	<1 %
19	Submitted to London Metropolitan University Student Paper	<1 %
20	hdl.handle.net Internet Source	<1 %
21	Submitted to University of Moratuwa Student Paper	<1 %
22	batgioistudio.com Internet Source	<1 %
23	github.com Internet Source	<1 %
24	Submitted to The University of Manchester Student Paper	<1 %
25	www.learnmall.in Internet Source	<1 %
26	"Operating Systems for Supercomputers and High Performance Computing", Springer Science and Business Media LLC, 2019 Publication	<1 %


27
en.unionpedia.org
Internet Source
<1%
28
www.ijert.org
Internet Source
<1%
Exclude quotes On
Exclude bibliography On
Exclude matches < 8 words

Turnitin Originality Report

Processed on: 14-Apr-2021 18:11 +08
ID: 1538924286
Word Count: 6768
Submitted: 1

SENTENCE-BASED ALIGNMENT FOR PARALLEL TEXT CO... By Lee Yong Wei

Similarity Index 12%

Similarity by Source
Internet Sources: 9%
Publications: 5%
Student Papers: 5%

Document Viewer

include quoted include bibliography excluding matches < 8 words mode: quickview (classic) report Change mode print download

2% match (Internet from 26-Nov-2018)
http://www.lrec-conf.org

1% match (student papers from 09-Apr-2021)
Submitted to Universiti Tunku Abdul Rahman on 2021-04-09

1% match (Internet from 05-Oct-2020)
http://docplayer.net

1% match (publications)
Hai-Long Trieu, Phuong-Thai Nguyen, Kim-Anh Nguyen. "Chapter 14 Improving Moore's Sentence Alignment Method Using Bilingual Word Clustering", Springer Nature, 2014

1% match (Internet from 19-Aug-2020)
https://en.wikiversity.org/wiki/Localization

1% match (Internet from 22-Sep-2005)
http://acl.eldoc.ub.rug.nl

<1% match (publications)
"Part-of-Speech Tagging Using Stochastic Techniques", Cognitive Technologies, 2006

<1% match (Internet from 07-Mar-2017)
http://eprints.utar.edu.my

<1% match (Internet from 03-Nov-2014)
http://hnk.ffzg.hr

<1% match (Internet from 18-Jun-2019)
http://eprints.utar.edu.my

<1% match (Internet from 04-Dec-2020)
https://www.pure.ed.ac.uk/ws/files/24570564/sent_align.pdf

<1% match (publications)
Xinyue Lin, Jin Liu, Jianming Zhang, Se-Jung Lim. "A Novel Beam Search to Improve Neural Machine Translation for English-Chinese", Computers, Materials & Continua, 2020

<1% match (student papers from 28-Jan-2021)
Submitted to Liverpool John Moores University on 2021-01-28

Universiti Tunku Abdul Rahman			
Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1of 1

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	LEE YONG WEI
ID Number(s)	17ACB04464
Programme / Course	CS
Title of Final Year Project	Sentence-based alignment for parallel text corpora preparation for machine traslation

Similarity	**Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)**
Overall similarity index: 12 % **Similarity by source** Internet Sources: 9 % Publications: 5 % Student Papers: 5 %	
Number of individual sources listed of more than 3% similarity: 0	

Parameters of originality required and limits approved by UTAR are as Follows:
- **(i) Overall similarity index is 20% and below, and**
- **(ii) Matching of individual sources listed must be less than 3% each, and**
- **(iii) Matching texts in continuous block must not exceed 8 words**

Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

_____	_____
Signature of Supervisor	Signature of Co-Supervisor
Name: Jasmina Khaw Yen Min	Name: _____
Date: 14/5/2021	Date: _____

CHECKLIST

UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY (KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	17ACB04464
Student Name	LEE YONG WEI
Supervisor Name	DR. JASMINA KHAW YEN MIN

TICK (√)	DOCUMENT ITEMS Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
√	Front Cover
√	Signed Report Status Declaration Form
√	Title Page
√	Signed form of the Declaration of Originality
√	Acknowledgement
√	Abstract
√	Table of Contents
√	List of Figures (if applicable)
√	List of Tables (if applicable)
	List of Symbols (if applicable)
	List of Abbreviations (if applicable)
√	Chapters / Content
√	Bibliography (or References)
√	All references in bibliography are cited in the thesis, especially in the chapter of literature review
√	Appendices (if applicable)
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.	Supervisor verification. Report with incorrect format can get 5 mark (1 grade) reduction.

 _____ (Signature of Student) Date: 15/4/2021	 _____ (Signature of Supervisor) Date: 15/4/2021