COMP4801 Final Report

Machine translation models for Cantonese-English pair

Liu Hey Wing (3035474016)

Supervisor: Dr. Yip Beta

17 April, 2021

Abstract

The studies on Cantonese language are conducted extensively and Machine Translation (MT) had been developed in the past decades, however, the study on Cantonese-based Machine Translations are under researched. This project aims to provide a foundation and preliminary research in this field, and investigate the difficulties and explanation of the absence. We had implemented different MT models, including historically and up-to-date technologies, evaluate the results and post the trained models online for future development.

Data collection of bilingual corpus is finished and MT models are successfully installed and tested. The highest BLEU score is 5.08 from Trans-4 Eng-Yue model. The results are not satisfactory because of the lack of data sources, choice of input data, choice of MT and time cost for performing translation.

Acknowledgement

I would like to thank you Dr. Yip Beta, the supervisor of this project, for his generous help in obtaining different resources and advice on this project. As a student having a double major in Computer Science and Linguistics, the opportunity of working on natural language processing and computational linguistics is very precious and cherishable, to which I am very thankful.

Secondly, I would like to thank you the CAES English lecturer, Grace Chang. She had provided sufficient support on my final year project and outlined detailed improvement for my presentation and report. This report would not be done without her.

Lastly, I am truly thankful to my family and friends for their support and help.

Contents

List of Figures

List of Tables

Listings

Abbreviations

1 Introduction

1.1 Background

Machine Translation (MT) has been developing since 1950s [9, Chapter 1.3], and started using in practice in recent years. However, numerous limitations have restricted the usage and functionalities of MT systems, such as translation direction (uni- or bi- directional), number of languages (multi- or bi- lingual), limited linguistic resources (unpopular languages), computational force and evaluating method (automatic or human).

An abstract concept of MT consists of 2 parts, ‘declarative’ and ‘procedural’ information, which can be simply understood as data and algorithms [9]. These information construct the basis of an MT model, and will be elaborated in the following content.

1.1.1 Linguistic aspect

Linguistic information is the ‘declaratives’ in MT systems, which are based on facts, definitions and relationships of natural languages. These information may include a grammatical description (dependency or generative), a corpus of orthography[1] or corpus of SL and TL.

Cantonese is a language ranked 18th in the world by the number of speakers [10]. The study of Cantonese linguistics are traced back to 1900s with Ball [11], since then, linguists had provided different aspects about Cantonese extensively and continuously, such as phonology, grammar [12] [13] and sociologist studies. However, there is a lack of a unified/standard writing system had created issues of choice in orthography. The written form is unified through the internet and forums in recent years [14].

Bernard Vauquois’ pyramid (Figure 1) is a very famous framework for preliminary MT, which focus on how linguistic information can be integrated. The pyramid is the ideological model for translation models.

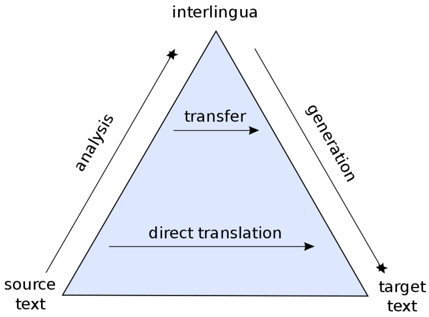
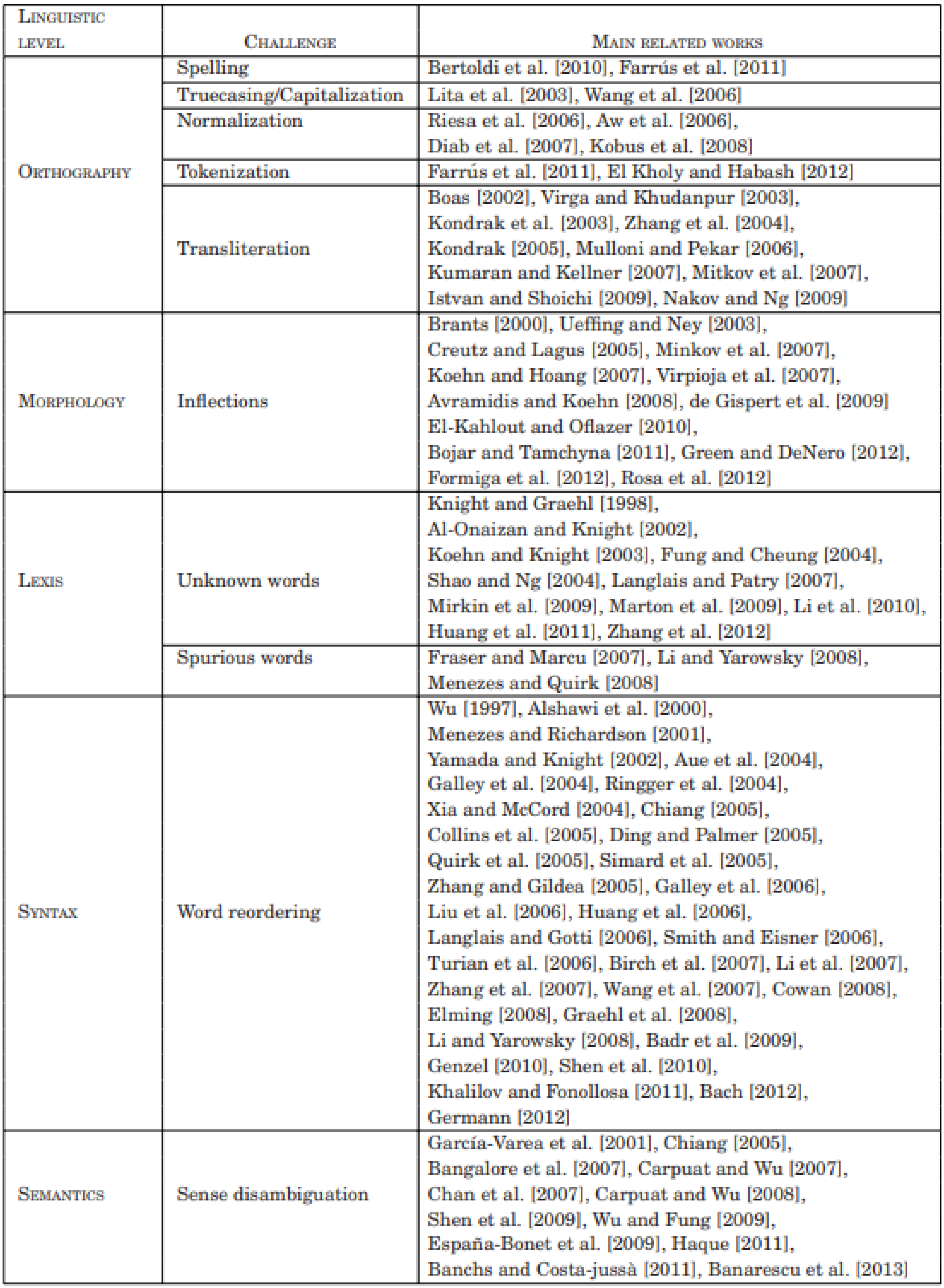


Figure 1: Bernard Vauquois’ pyramid [1].

Different levels and layers of translation are described, such as direct and syntactic transfer approaches. The concept is to decompose texts into smaller units, such as morphemes, and reconstruct them into another language.

Some linguistic challenges are described in Figure 2considering different aspects, including orthography, morphology, lexis etc. These problems had struggled with computational linguists for years, and some are still hurdles in the development of MT.

[1] Not all languages in the world use Latin alphabet, and some are lacking a unified or standard writing system like Cantonese.



Linguistic level	Challenge	Main related works
Orthography	Spelling	Bertoldi et al. [2010], Farrús et al. [2011]
	Truecasing/Capitalization	Lita et al. [2003], Wang et al. [2006]
	Normalization	Riesa et al. [2006], Aw et al. [2006], Diab et al. [2007], Kobus et al. [2008]
	Tokenization	Farrús et al. [2011], El Kholy and Habash [2012]
	Transliteration	Boas [2002], Virga and Khudanpur [2003], Kondrak et al. [2003], Zhang et al. [2004], Kondrak [2005], Mulloni and Pekar [2006], Kumaran and Kellner [2007], Mitkov et al. [2007], Istvan and Shoichi [2009], Nakov and Ng [2009]
Morphology	Inflections	Brants [2000], Ueffing and Ney [2003], Creutz and Lagus [2005], Minkov et al. [2007], Koehn and Hoang [2007], Virpioja et al. [2007], Avramidis and Koehn [2008], de Gispert et al. [2009] El-Kahlout and Oflazer [2010], Bojar and Tamchyna [2011], Green and DeNero [2012], Formiga et al. [2012], Rosa et al. [2012]
Lexis	Unknown words	Knight and Graehl [1998], Al-Onaizan and Knight [2002], Koehn and Knight [2003], Fung and Cheung [2004], Shao and Ng [2004], Langlais and Patry [2007], Mirkin et al. [2009], Marton et al. [2009], Li et al. [2010], Huang et al. [2011], Zhang et al. [2012]
	Spurious words	Fraser and Marcu [2007], Li and Yarowsky [2008], Menezes and Quirk [2008]
Syntax	Word reordering	Wu [1997], Alshawi et al. [2000], Menezes and Richardson [2001], Yamada and Knight [2002], Aue et al. [2004], Galley et al. [2004], Ringger et al. [2004], Xia and McCord [2004], Chiang [2005], Collins et al. [2005], Ding and Palmer [2005], Quirk et al. [2005], Simard et al. [2005], Zhang and Gildea [2005], Galley et al. [2006], Liu et al. [2006], Huang et al. [2006], Langlais and Gotti [2006], Smith and Eisner [2006], Turian et al. [2006], Birch et al. [2007], Li et al. [2007], Zhang et al. [2007], Wang et al. [2007], Cowan [2008], Elming [2008], Graehl et al. [2008], Li and Yarowsky [2008], Badr et al. [2009], Genzel [2010], Shen et al. [2010], Khalilov and Fonollosa [2011], Bach [2012], Germann [2012]
Semantics	Sense disambiguation	García-Varea et al. [2001], Chiang [2005], Bangalore et al. [2007], Carpuat and Wu [2007], Chan et al. [2007], Carpuat and Wu [2008], Shen et al. [2009], Wu and Fung [2009], España-Bonet et al. [2009], Haque [2011], Banchs and Costa-jussà [2011], Banarescu et al. [2013]

Figure 2: Some linguistic challenges in MT and related work [2]

Linguistics considerations and factors in an MT model are overwhelming, where each new language pair requires professional linguists to describe particular features and construct relevant data. With all the challenges described, translation frameworks and models switched from knowledge-based approach to corpus-based approach.

1.1.2 Computational aspect

Algorithms, softwares and models are the 'procedural' information in MT, which means how the system interprets the 'declarative' information, represented as a form of program or procedures. There are different approaches for performing MT, each requires different 'declarative' language information and perform different actions [15].

Figure 3 shows the development and variants of MT since 1960s [16]. The inheriting tree structure is describing the relationship between different approaches.

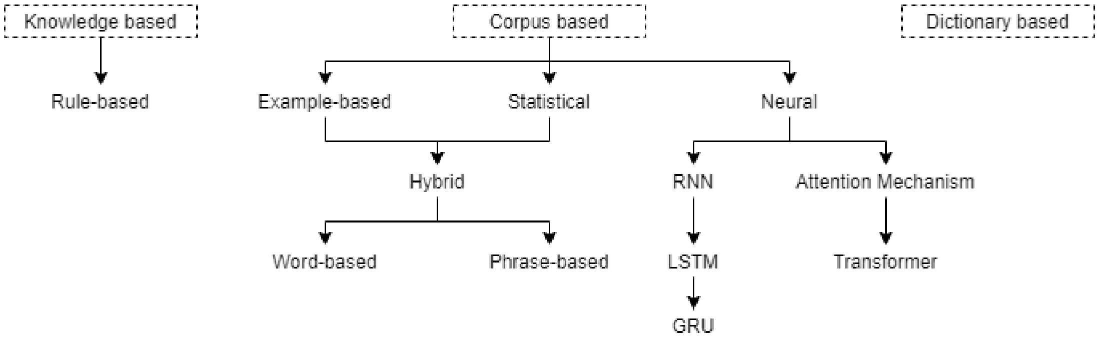


Figure 3: MT approaches and models. Based on [3]

The first MT system introduced was Rule-based machine translation (RBMT). Example-based machine translation (EBMT) was introduced in the 1980s which focus on comparison and extraction [17]. In 1990s, IBM had performed a series of experiment for a new Machine Translation (MT) system based on sentence alignment which is later known as Statistical Machine Translation (SMT). Neural machine translation (NMT), which is a kind of deep learning, started around the mid-2010s, together with the rise of AI and machine learning, and the relevant research is still ongoing.

From the history of the development of MT, the approaches switched from relying on linguistic knowledge to statistics, data science and AI. In this project, several historical and advanced MT models are implemented, and the details will be discussed.

1.2 Previous works on MTs in Cantonese-English pair

To my knowledge, PoluU-MT-99 is the first MT system implemented for Cantonese-English pair in 1999. Both RBMT and EBMT approaches are included in the system, which described the details such as Cantonese segmentation, bilingual algorithm and target construction algorithm [18].

Later in 2005, one of the authors of PoluU-MT-99, Yan Wu, presented another MT called LangCompMT05 [4], where the architecture (Figure 4) is nearly identical to PoluU-MT-99 with extended algorithms.

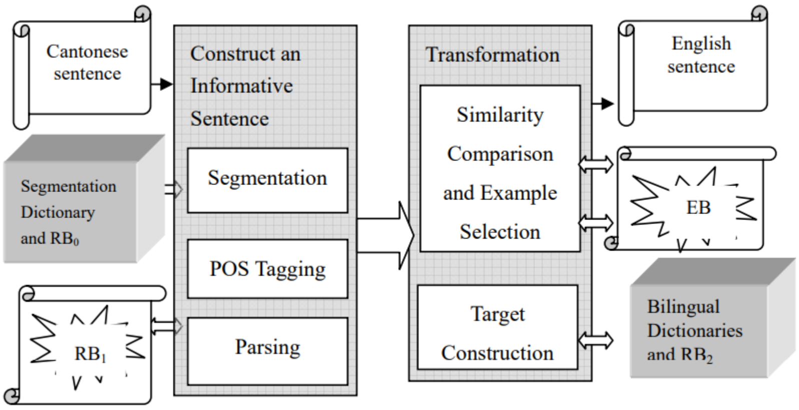

Cantonese sentence
Segmentation Dictionary and RB₀
RB₁
Construct an Informative Sentence
Segmentation
POS Tagging
Parsing
Transformation
Similarity Comparison and Example Selection
Target Construction
English sentence
EB
Bilingual Dictionaries and RB₂

Figure 4: Structure of LangCompMT05 [4].

However, the corpus data used in both PoluU-MT-99[2] and LangComp05[3] is not reliable. Some non-Cantonese words, sentences, aspect markers are used in the study, such as 有些, 的, 正在 originated from Modern Standard Chinese (MSC)[4]. Thus, we have certain reservation towards the result of this study and leave it out of our consideration.

Unfortunately, the source code is not publicly available but the algorithms and structure is presented in the study, thus we are unable to test the model. Re-implementing the models from scratch is over-complicated and out of the scope of this project.

[2] The source of its example base is not described, but the testing data was retrieved from Mingpao, a Hong Kong newspaper uses MSC rather than written form of Cantonese in their print.

[3] The example-based was constructed by Shiwen Yu, but his/her corpus is not cited in [4]. The source of testing data was created by the authors, which are targeted to perform specific sentence structure translation rather than real data.

[4] Or written form of Mandarin. Note that although Mandarin and Cantonese are both considered as the same language under the name of ‘Chinese’ and share some degree of written intelligibility of Hanzi and cognates, they are mutually unintelligible and are considered as different languages linguistically.

1.3 Objectives

Despite the abundant resources, barely any MT systems include Cantonese, with only Microsoft Bing translator and Baidu App. The lack of Cantonese MT systems (regardless of direction) is unusual. The purpose of this project is to investigate and discover such obstacles, provide possibly improvement for existing MT systems, and outline the future development for Cantonese MT.

1.4 Scope and Deliverable

MT is evolving rapidly, with numerous proposed variant and commercial implementation. Some of them are outdated and some have outstanding performance, so our scope is to narrow down and focus on some categories of MT systems listed above, including cutting-edge and historically popular models.

In The Confrence on Machine Translation (WMT) 2019 [19], the most advanced and newly developed MT were presented, which are mainly SMT (PBSMT) and NMT (LSTM and Transformer). Some historically popular MT will be discussed in this project, including RBMT and EBMT. Different MT systems will be implemented with training and testing set of data. We will evaluate the results and discover the reason for the absence of Cantonese MT.

At the end of the project, we want to deliver at least 1 MT model with a BLEU score of 10. With such system, we hope that it will increase the interest in research of Cantonese-based MT and potentially develop into a fully functional translation system.

1.5 Outline

In the Introduction (Section 1), the overall background is provided with objectives and scope. In Methodology (Section 2), we will investigate MT systems from data, algorithm and evaluation perspectives. Details of the implementation will be discussed in Experimental Setup (Section 3). The results and discussion will be presented in (Section 4). Finally, the limitations and future work will be discussed (Section 5, 6).

2 Methodology

In this section, the descriptions of each MT models are listed out, including data, algorithms and evaluation. Particular MT models are chosen, which cover historical models and advanced technologies.

2.1 Data

2.1.1 Linguistics choice

There are around 15,000 frequently used vocabularies in Hong Kong, whereas 9706 vocabularies are considered to be sufficient for primary school children, and those vocabularies are composited by 3171 Chinese characters [20]. The number of frequently used characters and vocabularies are comparable, thus, both formats are implemented to compare performances in MT models.

Chinese Character is chosen to represent Cantonese due to the high number of lexicons/cognates shared [12]. The meaning of characters is related to different vocabularies. Using romanized text would cause difficulties in data preprocessing also.

Considering the vocabularies without consensus on Chinese character and especially the polysemous, we constructed principles based on the finding [14] to transformer them, (1) preferably choose the most recognized Chinese character, (2) choose the most disambiguated Chinese character, (3) use the most widely adopted romanized form and (4) use Jyutping to transcribe[5]. All forms are chosen based on the usage in Hong Kong, but not the forms in Mainland China or overseas Cantonese community.

2.1.2 Source

Bilingual corpus is preferably used and sometimes required as training and testing data in MT. To my knowledge, the only available Cantonese-English bilingual corpus is ShefCE [21]. Despite that the corpus exists, the transcription only includes Jyutping romanization without tone markers[6], which makes it not applicable for MT.

[5] For instance, '係' is the most recognized form to represent *haai2* 'at', but it is overlapped with *haai6* 'to be', so '喺' is chosen instead. Another example is the expressions 'hea' and 'chur' without Chinese character, and they remain in this form to be recognizable.

[6] The study used a refined lexicon derived by linguistic knowledge and sub-syllable unit combinatorial constraints, and get rid of the lexical tones [22]. However, in the precise environment like translation, the

Cantonese consists of a high frequency of code-switching in daily speech or text. While Audio originated corpus encourage code-switched text, translation originated corpus discourage the phenomenon. This is an important language phenomenon in Cantonese used in Hong Kong, and possibly contribute to the absence of Cantonese-based MT, so both audio and translation originated data are included.

TED [23] is a nonprofit organization that invite speakers to give talks, namely TED talks. It is open-source for translation and subtitles in more than 100 languages, including Cantonese and English. Following web crawling/scraping method proposed in [24] Beautiful Soup [25], a total of 193 talks are available to construct the corpus. The timestamp is used to subdivide the whole talk into sentences, however, some subtitles are not divided into pieces due to formatting issue.

Netflix movies and TV shows are chosen as another data source, where they support both Cantonese and English audios and subtitles. There are 300 videos with Cantonese audio and English subtitle, while there are only 6 videos vice versa. For fairness between the language pair, it is better to include both directions. However, the Cantonese subtitles are encrypted and incomprehensible. Therefore, pyTranscriber [26], a FOS software, performs Cantonese transcription based on Google Speech-to-Text API [27].

The corpus is not exactly parallel, because of the sentence structure between Cantonese and English and the difficulties in aligning sentences. The noisiness issue will be addressed again in Section 4.2.

2.1.3 Preprocessing

The data collected are unified into the same format described above. Some of the data in TED talks are in simplified character and Mainland version of the controversial forms, which are further transformed into our standardized format.

Punctuation marks[7] are removed because full-text translation and long sentence translation is not the focus of this project [28]. Word order is relatively consistent in

disappearance of tones would cause loss in information.

[7] In addition, the output of pyTranscriber do not include any punctuation marks, and the movie subtitles contain punctuation marks seldomly.

English, which reduces ambiguities in context. While word order in Cantonese varies with right-dislocation [12], but still understandable without punctuation marks.

2.2 Description of MT models

2.2.1 RBMT

Rule-based machine translation (RBMT) is designed based on linguistic knowledge, in which decompose the Source Language (SL) into parse tree, restructure the lexicons, and reformat into the Target Language (TL). This was a popular model before the 1980s, but it had faded out in use because of the improvement of computational power, which later give rise to SMT and other systems.

Apertium [5] is a free/open source (FOS) platform and the most popular framework for RBMT. The structure is shown below (Figure 5).

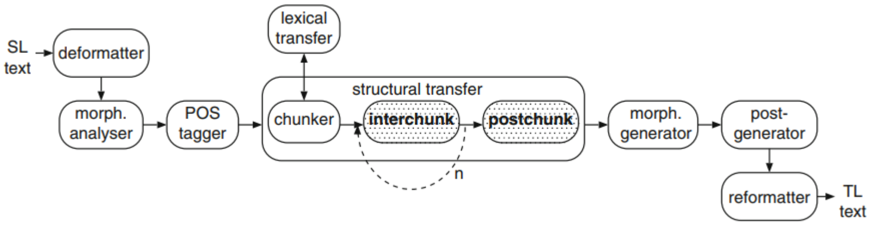


Figure 5: Structure of Apertium [5].

Apertium is heavily based on parallel (Bilingual) data, dictionary and explicit linguistic data such as grammars and structural transfer rules. The system is not based on any common linguistic framework, including generative grammar such as x-bar theory, minimalist programme, LFG, and dependency grammar. Rather, one-to-one sentence correspondence is adopted, such that every possible sentence structure are required to be defined, and all vocabularies needed to perform part-of-speech (POS) tagging.

With such constraint and no any existing framework incorporate Cantonese, the resources and efforts required are more than the scope of this project, and thus will not be implemented.

2.2.2 EBMT

Example-based machine translation (EBMT) is the first MT system to get rid of the ‘knowledge’ of linguistics and based on ‘pure’ data, which was developed by Makoto

Nagao [17] in 1984. The input sentence would be searched in the trained model, substitute the phrase or character, and generate the output sentence, as shown in example (1)-(2) [29].

(1) (a) He buys a notebook.
Kare wa noto o kau.
(b) I read a book on international politics.
Watashi wa kokusai seiji nitsuite kakareta hon o yomu.
(2) (a) He buys a book on international politics.
(b) Kare wa kokusai seiji nitsuite kakareta hon o kau.

On the other hand, Somers [30] had a thorough review about EBMT and addressed several limitations and difficulties[8] on further development and improvement. As a result, it introduced the idea of corpus-based MT and give rise to SMT.

CMU-EMBT [6] in Figure 6 uses a nearly 'pure' EMBT, with optional components and integration to other MT models available.

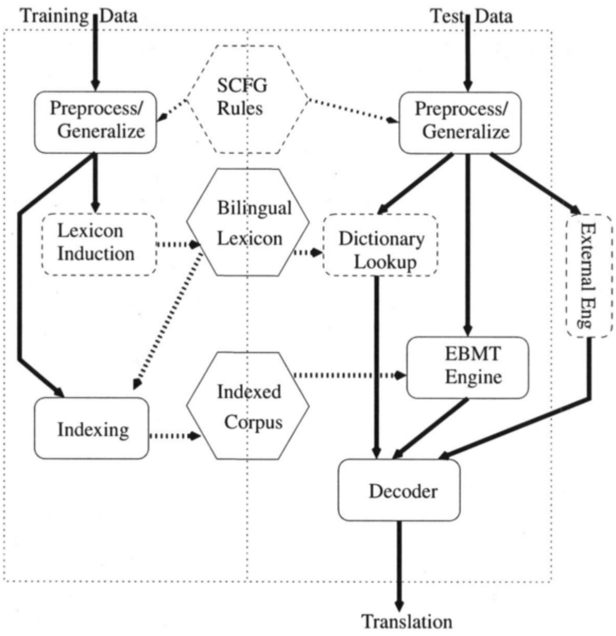


Figure 6: Structure of CMU-EBMT [6].

Input sentences are converted into lattice where character/phrase segmentation would be performed, and convert into the Language Model (LM). The decoder would perform lexicon lookup for decomposed TL sentence in the corpus, and substitute the sentence. The model is already not functional and cannot be implemented in this project, the

[8] Such as the size of example, lack of parallel corpora of language pairs, suitability of examples etc.

structure is provided to increase the understanding of EBMT.

The rise and fall of EBMT were fast, which induced low effort or resources[9] on the topic. Thus, EBMT will not be used in this project.

2.2.3 SMT

Statistical Machine Translation (SMT) uses a probabilistic approach to perform translations, which involves much less linguistic knowledge. The most popular variant is Phrase-based Statistical Machine Translation (PBSMT).

Bayes Decision Rule (Figure 7) describes $Pr(e_1^I|f_1^F)$ as the translation model (SMT), where the source sentence $f_1^J = f^1, ..., f^j, ..., f^J$ is translated into target sentence $e_1^I = e^1, ..., e^i, ..., e^I$.

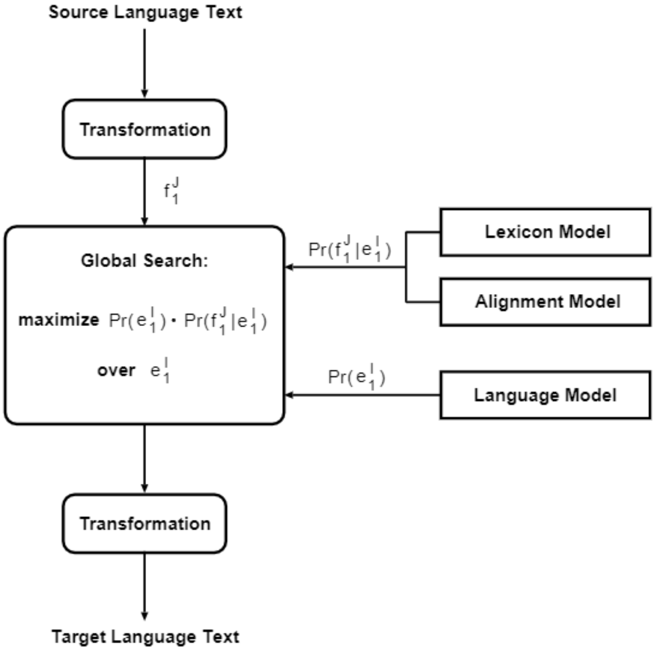


Figure 7: Structure of Bayes Decision Rule applying
Source: [33]

PBSMT segment SL into phrases and translate phrase-to-phrase with the probability, and the mathematical description is below [33].

$$\hat{e}_1^I \approx \underset{e_i^I, B}{argmax} \left\{ \prod_{i=1}^{I} p(e_i|e_{i-1} \cdot \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_k)^{\lambda} \right\}$$

[9] Some more known EBMT FOS platforms are already unavailable now, such as Cunei [31], OpenMaTrEx [32] and uploaded in the 2010s. Even if they are available, the coding is not compatible with modern technology anymore.

IBM-1 model will be implemented in our project [34]. Although there are more complex and popular model like Moses [35], the input pipeline is not accessible and difficult to build. Therefore, a public github repository [36] are modified to adapt into our setting for this purpose.

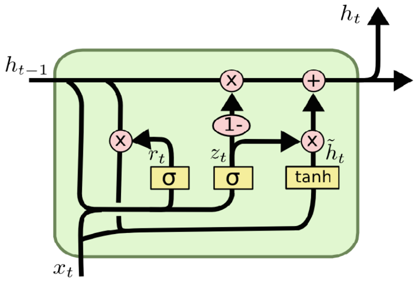
2.2.4 GRU

Neural machine translation (NMT) uses 2 approaches, attention mechanism (Section 2.2.5) and Recurrent Neutral Network (RNN) [37]. NMT has been proven with significant improvement comparing with other models, especially in WMT‘16 [38] in different fields. The limitation for NMT is the ambiguity within the hidden layers, which is not understandable by human and thus increase the difficulty for further improvement [39].

RNN approach consists of numerous layers of feed-forward computation and RNN between the encoder-decoder and sends results backwards. The encoder transforms x into a vector space of hidden layers[10], and the decoder transforms those vectors into y, using the likelyhood equation below.

$$p(y|x;\theta) = \prod_{j=1}^{m+1} p(y_j|y_{0:j-1}, x; \theta)$$

LSTM cell is an extension of RNN proposed back in 1997 [40], and implemented for MT in 2014 [41]. Cho et al. [42] proposed a modified version called Gated Recurrent Unit (GRU) as in Figure 8 with higher computational speed and low space.



$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$
$$\tilde{h}_t = tanh(W \cdot [r_t * h_{t-1}, x_t])$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Figure 8: GRU cell [7].

GRU combines the input and forget gate into the *update* gate z_t, and introduced the *reset* gate r_t, but still unable to perform multi-threading and have low training speed.

[10] Hidden Markov model (HMM) is a complicated topic, which is out of the scope of this project

The pyTorch framework provided by D2L [43] is adopted for GRU model.

2.2.5 Transformer

In the Attention mechanism, each segment has a vector of non-fixed size in contrast to RNN, which represent the attention driven in a sentence. The decoder then derives output once at a time based on the attention towards the input , and passes the results backwards. This avoids RNN‘s problem of mapping sentences into a fixed-length vector, which causes loss in input information and over-matching [8] [44] [45].

Transformer model was invented in 2017 [44], with the innovation of multi-head attention in the encoder. Figure 9 presents the structure of Transformer model.

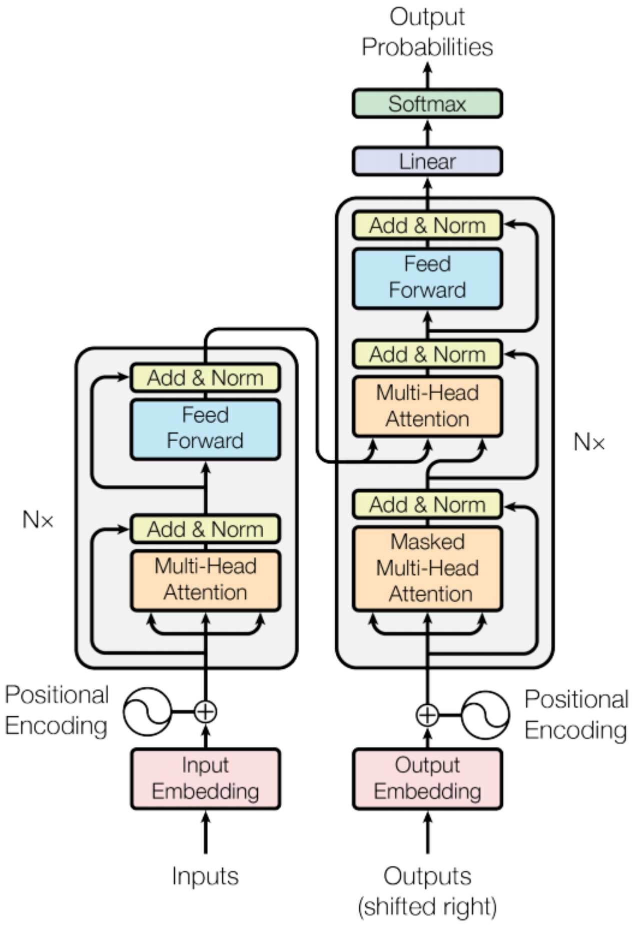


Figure 9: Structure of Transformer model [8].

The left side is the encoding of the input sentence, which is an attention block generating vector based on sentence length. The decoder on the right side is separated into 2 blocks, where the first attention block encodes the output sentence, and the second attention block computes the attention between the input vector and output vector. The details of attention blocks and multi-layer is as below (Figure 10), where Q is the attention matrices, K is the key index and V is the value.

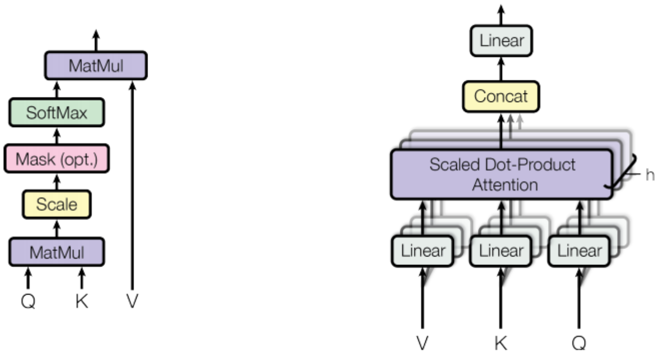


Figure 10: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention. [8]

Multi-head attention uses self-attention mechanism, which increases the quality of the system. In addition, the structure allows multi-threading in training on several GPU, and highly increase the training speed. Finally, The result would be converted into probability functions to predict the output.

The Transformer model in Tensorflow [46] with the configurations and library provided by Vaswasni [47] will be implemented in this project.

2.3 Evaluation of translation output

Each time the MT translate a sentence, it is curious about how it performed, whether it translates correctly and how can we improve its result. The performance of MT systems is the key for further improvement, and evaluation methods will be discussed.

2.3.1 Manual evaluation

Manual evaluation is always considered time-consuming and high cost, and the consistency in a large-scale rating is difficult to maintain [48]. In the current advancement in technology, the machine can evaluate translation results in hours where human take days or weeks to complete, thus, manual evaluation is not preferred in most circumstances.

[49] had provided another perspective using Amazon‘s Mechanical Turk to recruit human evaluators for numerous languages, which showed evaluating result are actu-

ally near to experts. Considering the status of Cantonese [10], the evaluation is feasible, however, the cost and time are not preferred in this project.

2.3.2 Automatic evaluation

MT systems often use automated metrics to evaluate translation results. Different metrics are proposed for better performance.

Bilingual Evaluation Understudy (BLEU) [50] is widely adopted to evaluate translation results [1, 2, 45, 19, 37, 33, 51, 38, 41, 52, 44, 8, 4], which is well-established and performs quality evaluation close to human. Since then, more metrics are proposed according to different dependencies, such as NIST, RED and WER. In this project, BLEU is adopted to conduct a comparison with previous studies.

2.4 Summary

To conclude, our project will implement the softwares / models as below.

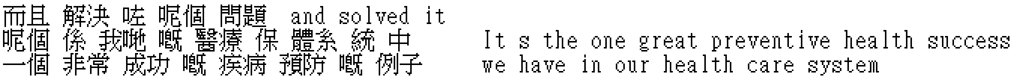
- *Data*: TED talks + Netflix in Chinese character
- *MT models*:
  - *SMT*: IBM-1
  - *GRU*: D2L pyTorch with GRU cell
  - *Transformer*: Tensorflow with Transformer model
- *Evaluation*: BLEU

3 Experimental Setup

In this section, we will describe and explain the experimental setup, including the details of the generated corpus, implemented MT models, hyperparameter settings and training interface.

3.1 Data

The corpus is generated from TED talks and Netflix with around 10000 parallel sentences. All talks in TED are extracted, while only 2 movies from Netflix are included in the corpus. The transcription accuracy of pyTranscriber for Netflix is relatively low, due to music and background noises. Thus, manual proofreading is required to preprocess the data. In Figure 11, some sample sentences from the corpus are listed out.



```
而且 解決 咗 呢個 問題  and solved it
呢個 係 我哋 嘅 醫療 保 體系 統 中      It s the one great preventive health success
一個 非常 成功 嘅 疾病 預防 嘅 例子     we have in our health care system
```

Figure 11: Example sentences from corpus.

The sentences are in the format proposed in Section 2.1.3. Some observation is that the sentence is not completely identically parallel due to the reasons mentioned in Section 2.1.2.

Through analysing the corpus composition, we investigate the difference between the choice of the input format of Cantonese. PyCantonese [28] is a library that can truncate and tokenise sentences in Cantonese. In the corpus, there are 253040 instances in the form of character where 11503 (4.55%) are unique. In the form of vocabulary/phrase, there are 151776 instances and 23422 (15.4%) are unique.

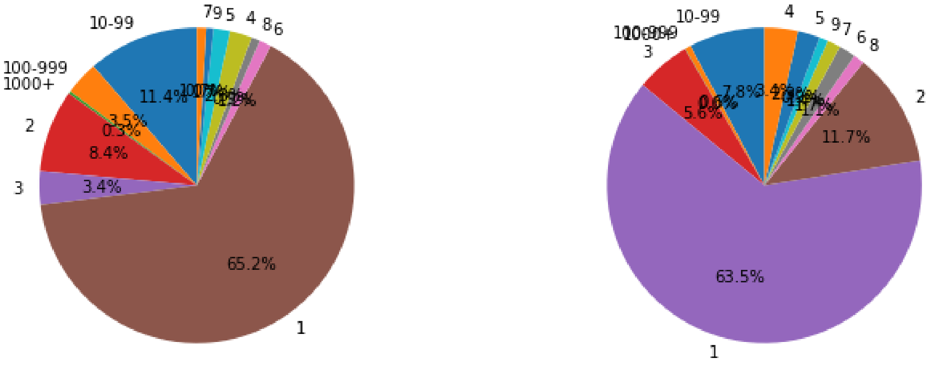


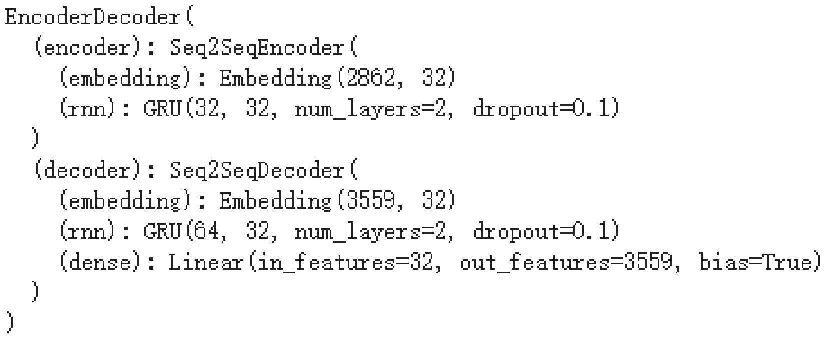
Figure 12: Distribution of occurrences of unique tokens in Character (left) and Vocabulary (right) form.

The composition of occurrences of unique tokens in both forms are shown in Figure 12. Most unique tokens only appear once in both forms (65.2% & 63.5%), and the other occurrences are similar in both forms. However, considering the percentage of unique tokens in the whole corpus, there are significantly more unique tokens in vocabulary than character form. Unique tokens increase the difficulty of translation task, since the meaning are easily misunderstood or mismatched by the system. Therefore, the character form is more suitable as the baseline input for different MT systems, and the vocabulary form is not included in this project.

Overly-long sentences of over 40 units[11] in both languages are excluded in the data, as long-range attention and documental translation are not the focus in this project. The input data is split into training and testing set, with an 80:20 ratio. Although the testing results are not fair and vary in different splittings, the computational cost and time are limited for other validation methods.

3.2 MT models

SMT model follow the default structure of IBM-1 as mentioned in Section 2.2.3. The structure of GRU and Transformer model are shown in Figure 13 and 14.

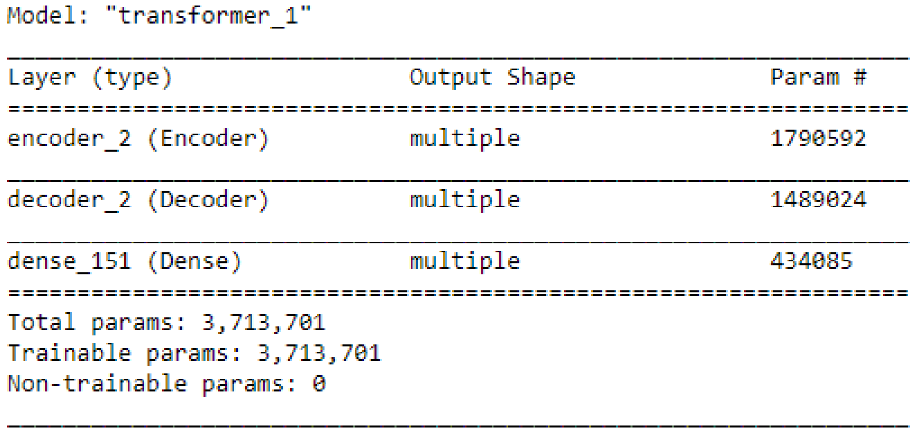


```
EncoderDecoder(
  (encoder): Seq2SeqEncoder(
    (embedding): Embedding(2862, 32)
    (rnn): GRU(32, 32, num_layers=2, dropout=0.1)
  )
  (decoder): Seq2SeqDecoder(
    (embedding): Embedding(3559, 32)
    (rnn): GRU(64, 32, num_layers=2, dropout=0.1)
    (dense): Linear(in_features=32, out_features=3559, bias=True)
  )
)
```

Figure 13: Structure summary of GRU model.

The figures are another representation of the models identical to structures discussed in Section 2.2. The Transformer model is too complicated to be shown as the form of summary, so the format below is used to summarize the structure. The term ‘params’

[11] Different sub-word encoder have a different approach for dividing word units. In English, the morphology can be decomposed like ‘-ing’ and ‘-ed’ tense morphemes. In Cantonese, we adopt a character-based approach which means the word unit is 1 Chinese character, as the input is parsed as UTF-8 coding.



```
Model: "transformer_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
encoder_2 (Encoder)          multiple                  1790592
_________________________________________________________________
decoder_2 (Decoder)          multiple                  1489024
_________________________________________________________________
dense_151 (Dense)            multiple                  434085
=================================================================
Total params: 3,713,701
Trainable params: 3,713,701
Non-trainable params: 0
_________________________________________________________________
```

Figure 14: Structure summary of Transformer model.

in Transformer is dependent on the size of data input, which is subject to change.

Considering the fact that many Cantonese speakers[12] are bilingual speakers, the demand for a Cantonese-to-English system is lower than vice versa. However, this project still implements bi-directional MT system for different needs.

3.3 Parameters

Hyperparameters of SMT, GRU and Transformer models are shown in Table 1.

	SMT-1	SMT-2	SMT-3	SMT-4
word_factor_max [13]	1	2	3	4
	GRU-1	GRU-2	GRU-3	GRU-4
embed_size	32	64	128	256
num_hiddens	32	64	128	512
num_layers	2	4	8	8
dropout	0.1	0.2	0.4	0.4
	Trans-1	Trans-2	Trans-3	Trans-4
num_layers	4	2	2	2
d_model	128	64	32	128
dff	512	256	128	256
num_heads	8	4	4	4

Table 1: Hyperparameter setting of SMT, GRU, and Transformer.

[12] Hong Kong, Macau and overseas are fluent bilingual speakers. Speakers in mainland China are required to learn English at school.

[13] This is the maximum number of words to be analysed in the SMT model. For simplicity and fairness between languages, the number of uni-gram are divided into fraction rather than a fixed number.

Different settings including the default settings (SMT-1, GRU-1 & Trans-1) are tested. SMT models had been trained until convergence where GRU and Transformer were trained on 100 epochs. These hyperparameters are subject to change based on numerous reasons, including the size of the dataset, number of epochs, quality of data etc.

3.4 Training interface

SMT models analyse the occasions of vocabulary items to learn. The NMT models learn by evaluating themselves in the training process, where either the training accuracy or training loss are used as the evaluating metric.

3.4.1 SMT

Listing 1 describes the maximum change function of SMT.

```
max_change = max(
        max_change,
        abs(
            t[(inp_word, tar_word)] − count[(inp_word, tar_word)] / total[tar_word]
        )
    )
```

Listing 1: Change function of SMT.

The change function is a simple calculation of occurrences in all sentences. Since SMT is not a stochastic model, the performances are unchanged and repeatable in a controlled environment. The maximum changes were recorded in Figure 15. The training stops when changes are lower than 0.005 or converged.

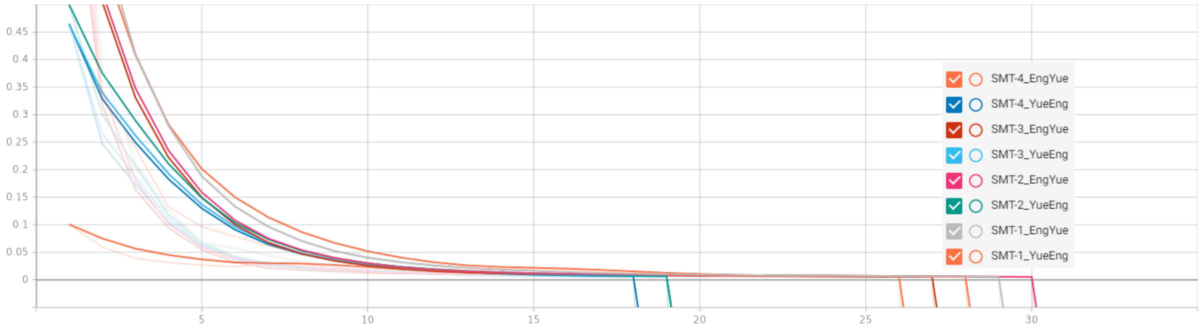


Figure 15: Change of SMT models.

All the curves are smooth and the training processes end around 10-20 iterations. SMT-4_EngYue model has a faster starting point comparing to other models.

3.4.2 GRU

The loss function defined in GRU model is cross-entropy loss (Listing 2). Using the loss function, the GRU models adjust the weighting and bias of artificial neurons.

```
class MaskedSoftmaxCELoss(nn.CrossEntropyLoss):
    ”””The softmax cross−entropy loss with masks.”””
    # ‘pred‘ shape: (‘batch_size‘, ‘num_steps‘, ‘vocab_size‘)
    # ‘label‘ shape: (‘batch_size‘, ‘num_steps‘)
    # ‘valid_len‘ shape: (‘batch_size‘,)
    def forward(self, pred, label, valid_len):
        weights = torch.ones_like(label)
        weights = sequence_mask(weights, valid_len)
        self.reduction=’none’
        unweighted_loss = super(MaskedSoftmaxCELoss, self).forward(
            pred.permute(0, 2, 1), label)
        weighted_loss = (unweighted_loss ∗ weights).mean(dim=1)
        return weighted_loss
```

Listing 2: Loss function of GRU.

The training loss through epochs, i.e. the learning progresses, are shown in Figure 16, 17, 18, and 19 for both directions (Eng-Yue & Yue-Eng).

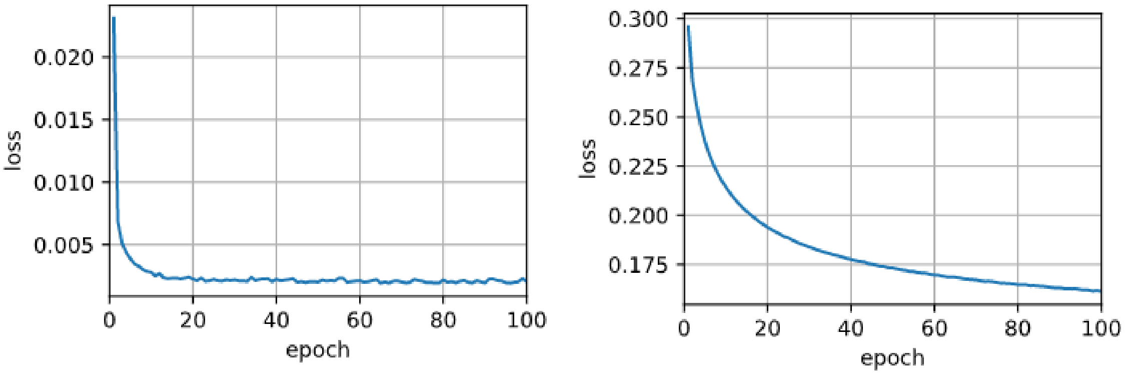


Figure 16: Training loss of GRU-1. (left) Eng-Yue. (right) Yue-Eng.

Different hyperparameters and translating directions show significantly different learning curves. The results will be discussed below (Section 4.1). Generally speaking, Cantonese to English (Yue-Eng) models demonstrated improvements gradually throughout epochs, while English to Cantonese (Eng-Yue) models had a steep learning slope.

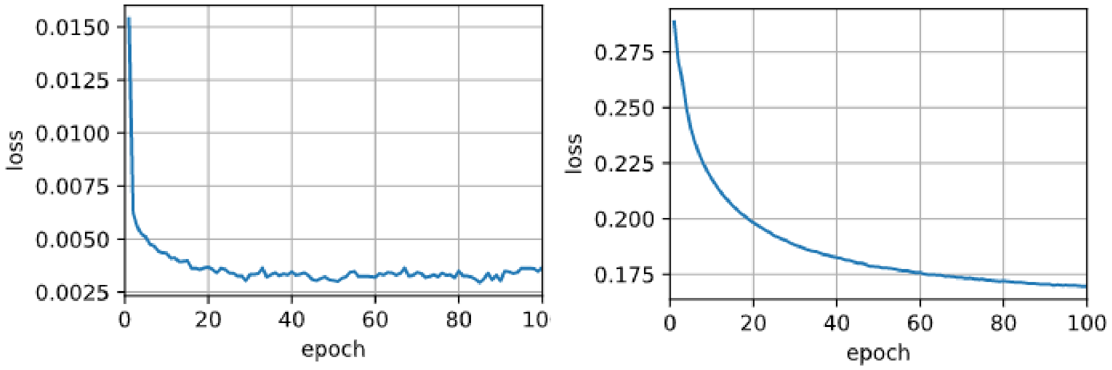


Figure 17: Training loss of GRU-2. (left) Eng-Yue. (right) Yue-Eng.

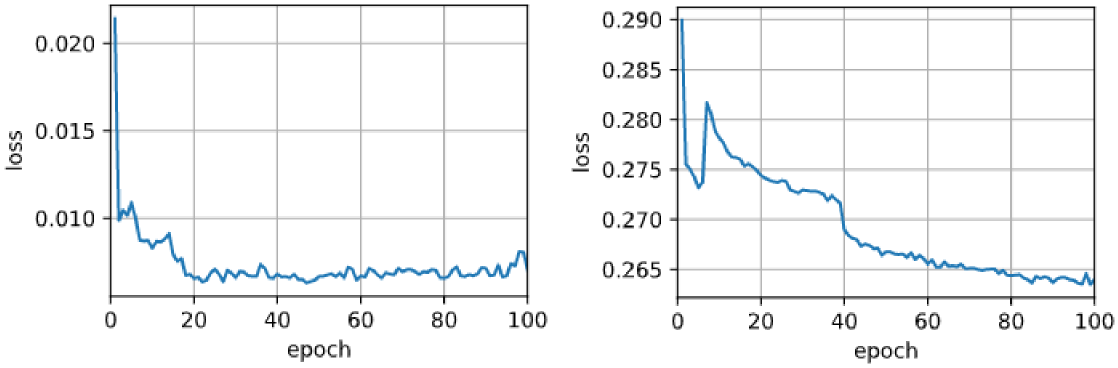


Figure 18: Training loss of GRU-3. (left) Eng-Yue. (right) Yue-Eng.

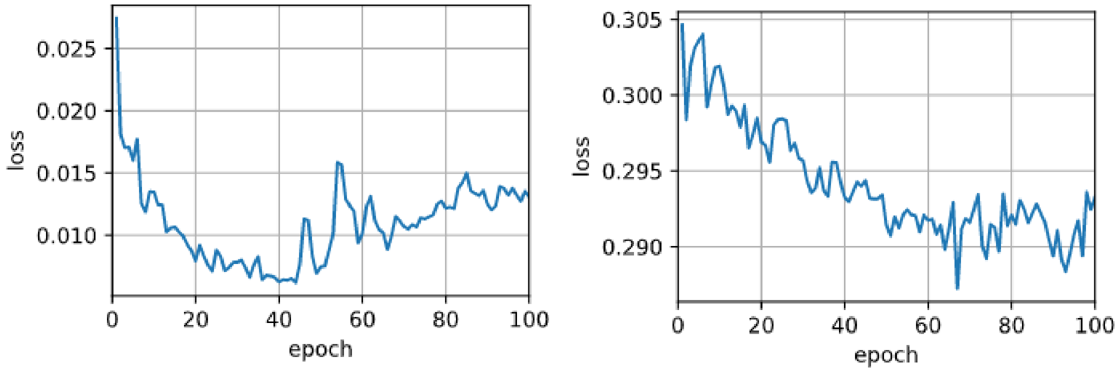


Figure 19: Training loss of GRU-4. (left) Eng-Yue. (right) Yue-Eng.

The learning curves of Eng-Yue models are likely to be a problem of overfitting, and thus having difficulty when handling new data in testing.

3.4.3 Transformer

The loss function used in the Transformer model is cross-entropy loss. Using the loss function, the Transformer model learns to pay attention towards different embedding space. Along with training loss, the training accuracy is also recorded during the epochs using the accuracy function in Listing 3.

```
def accuracy_function(real, pred):
    accuracies = tf.equal(real, tf.argmax(pred, axis=2))

    mask = tf.math.logical_not(tf.math.equal(real, 0))
    accuracies = tf.math.logical_and(mask, accuracies)

    accuracies = tf.cast(accuracies, dtype=tf.float32)
    mask = tf.cast(mask, dtype=tf.float32)
    return tf.reduce_sum(accuracies)/tf.reduce_sum(mask)

def loss_function(real, pred):
    mask = tf.math.logical_not(tf.math.equal(real, 0))
    loss_ = loss_object(real, pred)

    mask = tf.cast(mask, dtype=loss_.dtype)
    loss_ *= mask

    return tf.reduce_sum(loss_)/tf.reduce_sum(mask)
```

Listing 3: Accuracy and loss function of Transformer.

Using the two functions, we can observe the learning curve of models and contribute to fine-tune. The training processes of all trained models are recorded throughout epochs (Figure 20 and 21).

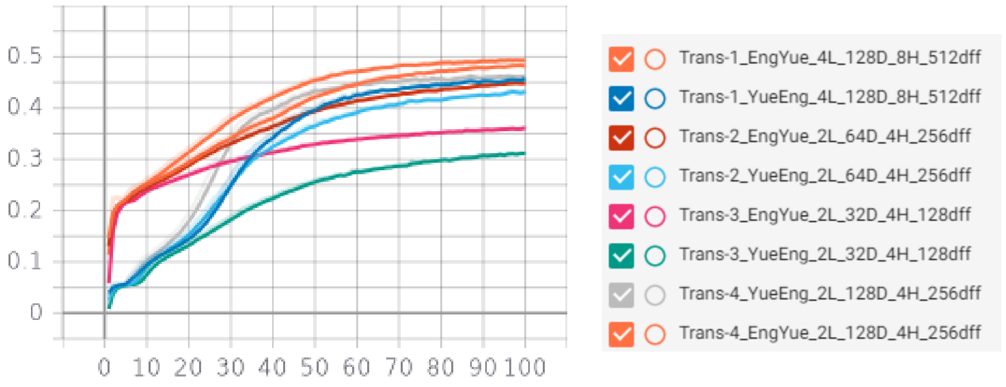


Figure 20: Training accuracy of Transformer models.

All the curves increase or decrease gently. The training accuracy and loss of models with different hyperparameters started differently and converged towards a flat

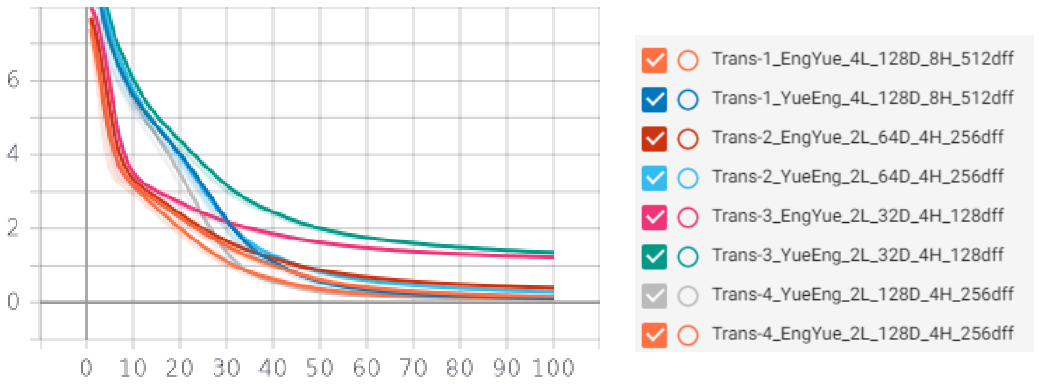


Figure 21: Training loss of Transformer models.

slope. The trend of curves is independent of different configs.

The training indicates that different direction (Eng-Yue & Yue-Eng) have different performance at the beginning during Epoch 1-30. Yue-Eng direction usually trained at a slower rate than Eng-Yue. However, both directions obtain similar metrics at the end of training. It is found that Trans-3 models are significantly different after training, that the performances are the lowest among all models.

4 Results and Discussion

In this section, the results with examples of the output of all trained models are presented. Based on the results, we will discuss and evaluate different models, and address possible obstacles for Cantonese-based MT systems.

4.1 Results

4.1.1 BLEU scores

The testing dataset (20% of corpus) is used to evaluate the performance of MT models towards unseen data. The BLEU scores of all models with different configurations (as mentioned in Table 1) are shown in Table 2. The best-performed models are shown in bold.

	Eng-Yue	Yue-Eng
SMT-1	0.167739	**0.064081**
SMT-2	0.180803	0.063924
SMT-3	**0.181737**	0.063884
SMT-4	0.173345	0.063884
GRU-1	**3.21e-11**	0.08394
GRU-2	1.44e-12	**1.31247**
GRU-3	4.49e-12	0.261526
GRU-4	0	0.228945
Trans-1	3.690426	2.173115
Trans-2	4.520353	2.516684
Trans-3	3.232586	1.864644
Trans-4	**5.076932**	**3.248945**

Table 2: BLEU scores of different models

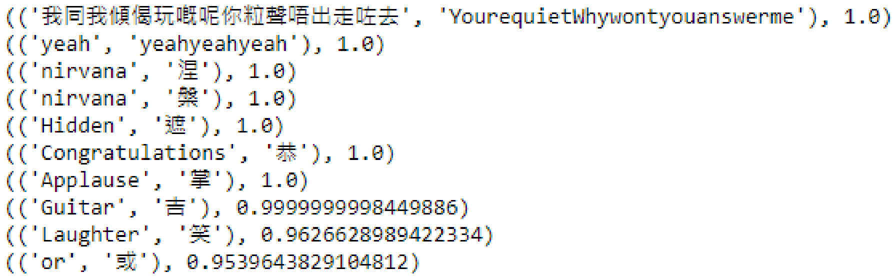
The scores of the implemented models are much lower than the original transformer (28.4 for En-De & 41.8 for En-Fr). The scores of GRU Eng-Yue models is nearly zero, which is unusual and will be further discussed below (Section 4.2.3).

Different directions of translation is a major factor of BLEU scores. Eng-Yue models have better performance than Yue-Eng direction for SMT and Transformer, while GRU models perform in reverse.

The hyperparameter settings contribute to GRU and Transformer significantly, whereas SMT models have more consistent output scores. The effect of hyperparameters and translation direction are independent of each other.

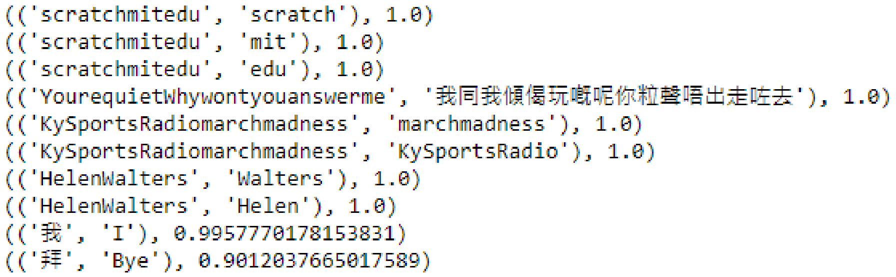
4.1.2 Output sentences

The probability table of SMT for each word are generated after training as in Figure 22 and 23. While performing translation, the model will generate the correspondences based on the probability of each word in a sentence, and then perform permutation to obtain the best arrangement.



```
(('我同我傾偈玩嘅呢你粒聲唔出走咗去', 'YourequietWhywontyouanswerme'), 1.0)
(('yeah', 'yeahyeahyeah'), 1.0)
(('nirvana', '涅'), 1.0)
(('nirvana', '槃'), 1.0)
(('Hidden', '遮'), 1.0)
(('Congratulations', '恭'), 1.0)
(('Applause', '掌'), 1.0)
(('Guitar', '吉'), 0.9999999998449886)
(('Laughter', '笑'), 0.9626628989422334)
(('or', '或'), 0.9539643829104812)
```

Figure 22: Sample probability table of SMT-1_EngYue.

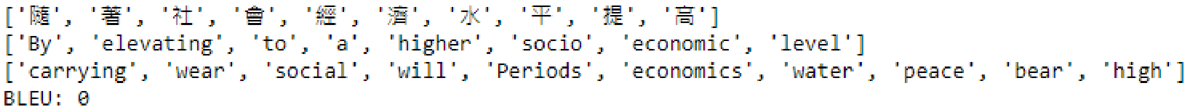


```
(('scratchmitedu', 'scratch'), 1.0)
(('scratchmitedu', 'mit'), 1.0)
(('scratchmitedu', 'edu'), 1.0)
(('YourequietWhywontyouanswerme', '我同我傾偈玩嘅呢你粒聲唔出走咗去'), 1.0)
(('KySportsRadiomarchmadness', 'marchmadness'), 1.0)
(('KySportsRadiomarchmadness', 'KySportsRadio'), 1.0)
(('HelenWalters', 'Walters'), 1.0)
(('HelenWalters', 'Helen'), 1.0)
(('我', 'I'), 0.9957770178153831)
(('拜', 'Bye'), 0.9012037665017589)
```

Figure 23: Sample probability table of SMT-1_YueEng.

This model facilitates one-to-one correspondence while performing translations, and only uni-grams and bi-grams are considered. The instances with probabilities of 1 showed the overfitting problem in SMT models, such that it learnt too well for particular training data, and possibly encountered difficulty with unseen data.

The grammar and other linguistics knowledge are not considered, so that languages with huge grammatical differences are easily misinterpreted in BLEU. Figure 24 demonstrates one example of predicted output.

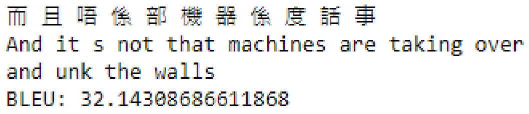


```
['隨', '著', '社', '會', '經', '濟', '水', '平', '提', '高']
['By', 'elevating', 'to', 'a', 'higher', 'socio', 'economic', 'level']
['carrying', 'wear', 'social', 'will', 'Periods', 'economics', 'water', 'peace', 'bear', 'high']
BLEU: 0
```

Figure 24: Sample output of SMT-1_YueEng.

There are possibilities for translating some of the instances correctly, but BLEU is unable to tell. For example, ‘社會經濟’ and ‘socio-economic’ are pairs in the sample, but the generated output ‘social’ and ‘economic’ are not recognised by the scoring. One-to-one correspondence causes other issues for SMT models, such as direct translation like ‘水平’ ‘level’ was literally translated into ‘water’ and ‘peace’.

Figure 25 is a sample output of GRU model.

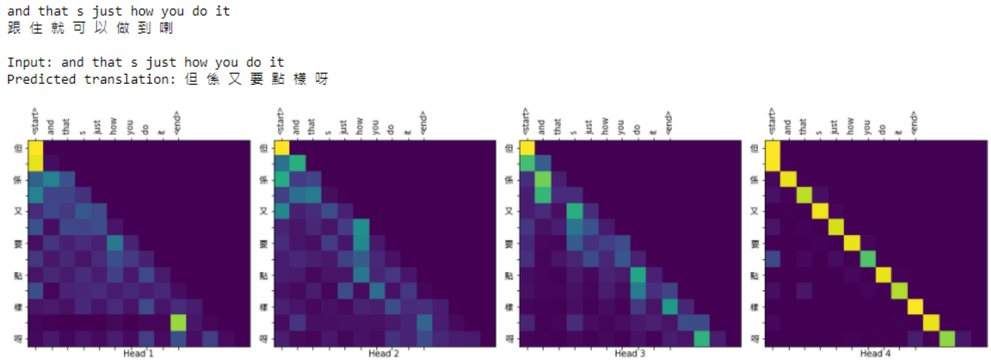


```
而 且 唔 係 部 機 器 係 度 話 事
And it s not that machines are taking over
and unk the walls
BLEU: 32.14308686611868
```

Figure 25: Sample probability table of GRU-2_YueEng.

The performance was not good enough, but some words are successfully translated. The possible explanation of low performance in GRU models will be discussed below (Section 4.2.3)

In Figure 26, an example sentence of the multi-head attention in the Transformer model is shown.



```
and that s just how you do it
跟 住 就 可 以 做 到 喇

Input: and that s just how you do it
Predicted translation: 但 係 又 要 點 樣 呀
```

Figure 26: Sample output of multi-head attention in Trans-4.

Although the translation is incorrect, the attention paid to the heads were demonstrated like head 4. The matching between input and output sentence is learnt by the model. However, different configurations had different effects on the attention model, some are messier (will be discussed in Section 4.2.1) and some are clean as in above.

4.2 Discussions

4.2.1 Overview

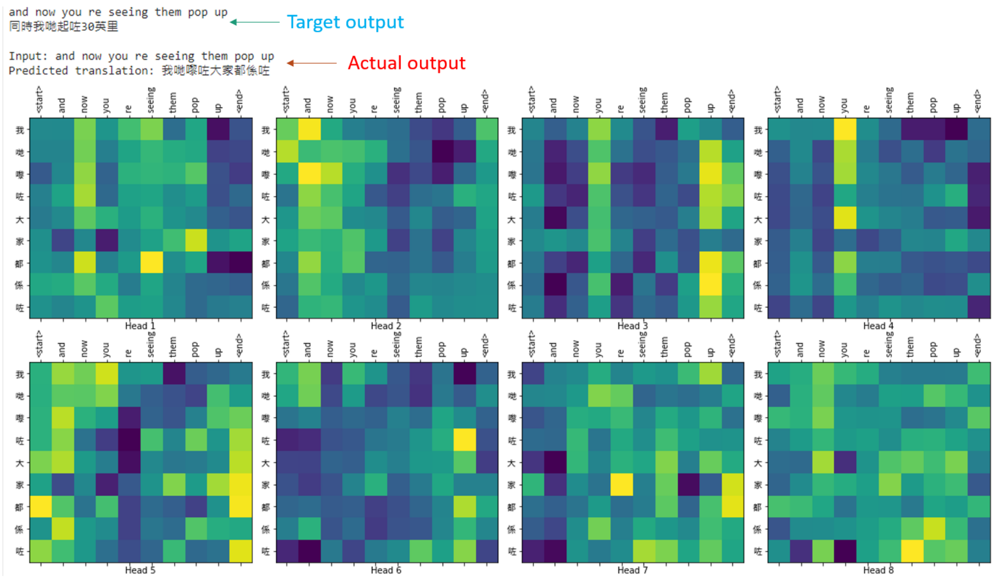
In the Transformer model, we can clearly visualise how the system maps input tokens into output tokens. Thus, attention graph similar to Figure 26is the generalization of overall performance. Some output sentences have similar meaning comparing to the input sentence, but the overall accuracy is considerably low. BLEU score lower than 10 are considered useless [53], therefore the implemented systems are not functional. The results obtained show the distance to our goal in this project.

In various NMT competition like WMT, human translators still outperform the most advanced models, especially in English-Chinese pair. Another observation is that contextual information has higher translation quality, such as news translation task, laws and documentations.

By adjusting hyperparameter settings in Table 1, it is found that stochastic deep-learning models like GRU and Transformer is able to improve their performance significantly. The results are expected to be even higher with a more fine-tuned process. However, SMT models are unable have reasonable improvement of BLEU scores.

There was an outstanding performance in the MT system. Specifically, the output result was unexpected, as revealed by the mismatch between the input sentence and the targeted output (Figure 27).

The actual input and targeted output have different meanings, which showed that BLEU scores is only one of the representation of translation quality, as different wordings can represent the same meaning in natural human language. The predicted sentence (actual output) have related meaning to the input sentence, the attention matrix does not show correlations to explain the output. This result is incredible and considered as unexpected, showing that the MT models are learning from the training.



```
and now you re seeing them pop up
同時我哋起咗30英里

Input: and now you re seeing them pop up
Predicted translation: 我哋嚟咗大家都係咗
```

Figure 27: Sample output of multi-head attention in Trans-1.

4.2.2 Corpus

In this project, the size of corpus data obtained is around 10000 pairs[14]. While MT models generally use> 100k pairs of data to perform training. Since the dataset is comparatively small, the statistical/probabilistic output is unable to obtain enough training data, which contributes to low accuracy.

As shown in Figure 27, the corpus is not exactly parallel and not completely accurate. Even when the predicted output has a similar meaning to the input sentence, the score is still low because of inaccurate targeted output. Noisy data affect the training process in MT models, and thereafter affect the evaluation result.

Efforts had been made for more data sources, and the result is expected to increase with more input data. However, there are limitations and difficulties for additional data sources, which will be discussed in Section 4.3.1.

4.2.3 MT

Overfit and underfit is one of the largest challenges in training NMT models [54], where the hyperparameter needs to match with the size of input data. Without the fine-tune of the hyperparameter described in Table 1, the performance of the MT systems would not be the best. The fine-tuning process is time-consuming requires previous ex-

[14] This is the number of pairs of the whole corpus. However, the filtering process of overly long sentences and splitting train-test datasets cause an even much lower number into training.

perience to perform efficiently, which is out of the scope of this project.

The models implemented are quantitative and statistical models. Some older and authentic models like RBMT requires too much human efforts and only language-specific. Thus, the scope of this project is limited and unable to investigate all the variants.

4.2.4 Configurations

Language imbalance is an important problem for machine translation. As discussed in Section 2.1.1 and 3.1, the word unit of Cantonese varies between different scholars and studies. The English language can also be analysed into different word units like words, morphemes, uni-gram, bi-gram etc. With restriction to the language itself, Cantonese vocabularies utilise compound words to form new meaning[15], which is less in English. The choice affects with the number of input data, which contribute to current results.

Language direction is another factor for hyperparameter settings. From the results in Section 4.1, the direction usually determined the results because of the language structure. Therefore, the adjustment of hyperparameters is required to be customised for each and every language pairs and input data to obtain optimal results.

4.2.5 Evaluation

BLEU is used as the only metric to evaluate the MT models. Although it had been used as the core evaluation metric in the past 20 years, different critiques had arisen issues around it. One of them is [55], which had shown that BLEU cannot correctly evaluate languages without word boundaries, and Cantonese is one of those languages. This problem can contribute to the current results.

As shown in Figure 24 and 25, BLEU scores sometimes have disappointing results. Even with highly similar meaning or different forms of the same vocabulary, the word would be judged as incorrect if not exact items are translated. On the other hand, the length of the translated sentences contributes to the largest determining factor. Such behaviour is not universal and adaptable for all MT models, and lead to diverse results with the same input data.

[15] For example, '電腦' literally means 'electric' and 'brain', combine to mean 'computer'.

4.3 Obstacles for Cantonese MT

Several possible explanations of the current results are discussed above (Section 4.2). These discussions are likely to correlate with the lack of Cantonese-MT.

4.3.1 Lack of data source

Despite both languages are the official languages in Hong Kong, Cantonese-English bilingual corpus and materials barely exist. Most documents are written in MSC instead of Cantonese. In Section 1.2, the study even used MSC as like 'Cantonese' for input data. The lack of written Cantonese text has indeed discouraged researchers to construct a corpus from scratch, and thus discourages the implementation of MT systems.

Even with the help of the internet, and more people are using Cantonese in daily life, researchers are still struggling to construct a large-scale data source. Older texts like missionary bibles, or early dictionaries are not transferred into computer. To my knowledge, there is no any publicly available parallel corpus for Cantonese-English pair. This is a possible explanation for the lack of a commercial translator or other studies about Cantonese MT.

4.3.2 Lack of unified representation

During the data collection and preprocessing process (Section 2.1.1, 2.1.3), various assumptions are made to handle the data. Since there is no unified representation of Cantonese writing systems, different writers and speakers would choose to use different ways to write down specific vocabularies.

In the process of computational analysis, there are conflicts in borrowing phonemic characters, transferring from simplified to traditional Chinese and resolving polysemous. The MT models require data cleansing and well-structured format as input and output, which increase the difficulty of building a functional model.

5 Limitations and recommendations

5.1 Vocabularies and genres of corpus

The vocabulary of movies and talks are heavily based on the genre. For instance, third-person narratives are more likely to appear in speeches and talks, and the content is full of technical jargons on a specific topic. Therefore, the MT systems implemented are not fully functional in real-life usage for colloquial speech.

The solution is to include as many genres as possible by manual effort. Data retrieving are limited in speech-to-text method, since the quality of transcription is required to be assured by human. So, we hope to discover other data sources to create a corpus larger in scale with ensured quality.

A brand new Cantonese-English bilingual corpus, SpiCE [56], is going to be released in 2021. The transcription is done by profession linguist and certainly have reliable and assured quality comparing to our current corpus. With the new data, the results of translation are predicted to be improved.

5.2 Choice of MT

Only 3 MT models are chosen to be implemented in this project. The coverage is smaller than the paradigm of MT, so thorough analysis and comparison cannot be done. Some models do not exist with any FOS platform as discussed in Section 2.2.2. We hope to include more alternatives in this project, however, due to the resources and cost, we had covered enough variants and had to give up less popular models.

MT is a fast-growing field, and there is an increasing number of newly invented state-of-the-art models, like Bi-LSTM and other popular methods like BERT. With the help of those models and technology, the results of this project are expected to be more conclusive and convincing.

5.3 Cost

Each training requires computational power and a lot of time in hours or even days. Restricted by the architecture, some MT models are unable to perform multi-thread training like GRU. With the number of models to be implemented and the ever-growing size of the corpus in the future, it is difficult to obtain the best performance of the MT models.

Therefore, the refinement in MT models is expected to be limited and not to be perfect. Exhaustive training is not recommended, and the hyperparameter will be adjusted based on experience rather than a blind test.

6 Future direction

This project had covered the corpus data, training of MT models and evaluating their results. Base on the results, we will raise some future directions for this project.

6.1 Improvement of corpus

The corpus size is not enough for training a functional MT model, and improvement in quality are also required. This process is considered to be continuous and without a fixed date, whenever new usable data exist.

The aggregation of corpus data will improve over time. This project hopes to be the pioneer for constructing such corpus data, and facilitate future researchers.

6.2 Trained and tested MTs

The MT models implemented are published online. Although most of the framework is adapted from different sources, the trained models are able to provide the foundation for future research.

The adjustment for models is implemented, such as character or vocabulary input, hyperparameter settings, analysing the composition of corpus etc. These are useful methods when conducting a quantitative study to obtain the best performance model.

6.3 Outcome and Value

At the end of this project, we hope to provide a foundation and preliminary study for further development on Cantonese-based MT, including the resource used and difficulties encountered. We hope to provide a functional system with BLEU score of 10. The trained models with corpus data will be available and open-source for future developers, and the code used will also be presented.

7 Conclusion

There is a lack of researches on Machine Translation between Cantonese and English. We discussed the linguistic and computational information for MT, and described details about data collection, MT systems and evaluating methods.

In this project, We implemented several Cantonese-based MT systems and investigated their performance, including SMT, GRU and Transformer. We noticed that Cantonese-English bilingual corpus does not exist, and chose to use TED talks and Netflix to construct our own corpus. Then, the well-known automated metric BLEU is used to evaluate translation outputs.

All chosen MTs are constructed with a bilingual corpus as input data. Through investing the composition of uni-gram in the corpus, we decided to use character form as Cantonese input. Each model is tested with 4 different configurations to obtain optimal results.

The results are unsatisfactory with the highest score of 5.08 of Tran-4 Eng-Yue model, which did not meet our goal of 10 in performance. However, some systems had unexpected output and incredible results. Possible reasons behind malfunctioning include corpus size, corpus quality, lack of fine-tuning, and metric that cannot evaluate correctly.

There are several limitations and difficulties found, including the vocabulary base and the choice of MT and training cost. We hope to improve these areas in manageable time and cost by manual effort and avoid exhaustive training.

At the end of this project, we aim to provide a foundation for further researches in Cantonese-based MT. We will also release the source code and trained models with corpus data for future researchers and developers.

8 Bibliography

[1] K. Shah, “Model adaptation techniques in machine translation,” Ph.D. dissertation, Université du Maine, 06 2012.

[2] M. R. Costa-Jussà and M. Farrús, “Statistical machine translation enhancements through linguistic levels: A survey,” *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, pp. 1–28, 2014.

[3] S. Tripathi and J. K. Sarkhel, “Approaches to machine translation,” *Annals of Library and Information Studies*, vol. 57, pp. 388–393, 2010.

[4] Y. Wu, X. Li, and S. C. Lun, “A structural-based approach to cantonese-english machine translation,” in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 11, Number 2, June 2006*, 2006, pp. 137–158.

[5] M. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O’Regan, S. Ortiz-Rojas, J. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. Tyers, “Apertium: A free/open-source platform for rule-based machine translation,” *Machine Translation*, vol. 25, pp. 127–144, 06 2011.

[6] R. D. Brown, “The cmu-ebmt machine translation system,” *Machine translation*, vol. 25, no. 2, p. 179, 2011.

[7] C. Olah, “Understanding lstm networks,” 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[9] W. J. Hutchins and H. L. Somers, *An introduction to machine translation*. Academic Press London, 1992, vol. 362.

[10] D. M. Eberhand, G. F. Simons, and C. D. Fenning, Eds., *Ethnologue: Languages of the World*, twenty-third ed. Dallas, TX: SIL International, 2020, https://www.ethnologue.com/.

[11] J. Ball, *Cantonese Made Easy: A Book of Simple Sentences in the Cantonese Dialect, with Free and Literal Translations, and Directions for the Rendering of*

English Grammatical Forms in Chinese. Kelly & Walsh, Limited, 1907. [Online]. Available: https://books.google.com.hk/books?id=3msuAAAAYAAJ

[12] S. Matthews and V. Yip, *Cantonese: A comprehensive grammar*. Routledge, 2013.

[13] T.-s. Wong, K. Gerdes, H. Leung, and J. Lee, "Quantitative comparative syntax on the cantonese-mandarin parallel dependency treebank," in *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università di Pisa, Italy*, no. 139. Linköping University Electronic Press, 2017, pp. 266–275.

[14] D. Snow, *Cantonese as written language: The growth of a written Chinese vernacular*. Hong Kong University Press, 2004, vol. 1.

[15] J. Oladosu, A. Esan, I. Adeyanju, B. Adegoke, O. Olaniyan, and B. Omodunbi, "Approaches to machine translation: a review," *FUOYE Journal of Engineering and Technology*, vol. 1, no. 1, 2016.

[16] T. Poibeau, *Machine Translation*, ser. The MIT Press Essential Knowledge series. MIT Press, 2017. [Online]. Available: https://books.google.com.hk/books?id=HYc3DwAAQBAJ

[17] M. Nagao, "A framework of a mechanical translation between japanese and english by analogy principle," *Artificial and human intelligence*, pp. 351–354, 1984.

[18] J. Liu and Y. Yu, "A cantonese-english machine translation system polyu-mt-99," in *MT SUMMIT VII: MT in the great translation era : proceedings of Machine Translation Summit VII*. Kent Ridge Digital Labs, Singapore: Asia-Pacific Association for Machine Translation, September 2019, pp. 481–486.

[19] L. Barrault, O. Bojar, M. R. Costa-Jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi *et al.*, "Findings of the 2019 conference on machine translation (wmt19)," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 2019, pp. 1–61.

[20] *Hong Kong Chinese Lexical Lists for Primary Learning*. Hong Kong: Chinese Language Education Section, Curriculum Development Institute, Education Bureau, The Government of the Hong Kong Special Administrative Region, 2007.

[21] W. M. Ng, A. C. Kwan, T. Lee, and T. Hain, “Shefce: A cantonese-english bilingual speech corpus – speech recognition model sets and recording transcripts,” Mar 2017. [Online]. Available: https://figshare.shef.ac.uk/articles/dataset/ShefCE_A_Cantonese-English_bilingual_speech_corpus_--_speech_recognition_model_sets_and_recording_transcripts/4522925/1

[22] R. W. Ng, A. C. Kwan, T. Lee, and T. Hain, “Shefce: A cantonese-english bilingual speech corpus for pronunciation assessment,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5825–5829.

[23] “Translate.” [Online]. Available: https://www.ted.com/participate/translate

[24] M. Cettolo, C. Girardi, and M. Federico, “Wit3: Web inventory of transcribed and translated talks,” in *Conference of european association for machine translation*, 2012, pp. 261–268.

[25] L. Richardson, “Beautiful soup documentation,” *April*, 2007.

[26] R. C. Souza, “raryelcostasouza/pytranscriber v1.4,” Jan. 2020. [Online]. Available: https://github.com/raryelcostasouza/pyTranscriber

[27] “Speech-to-text documentation nbsp;|nbsp; cloud speech-to-text documentation.” [Online]. Available: https://cloud.google.com/speech-to-text/docs/

[28] J. L. Lee, “Pycantonese: Cantonese linguistic research in the age of big data,” *Talk at the Childhood Bilingualism Research Centre, the Chinese University of Hong Kong*, 2015.

[29] S. Sato and M. Nagao, “Toward memory-based translation,” in *COLNG 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*, 1990.

[30] H. Somers, “Example-based machine translation,” *Machine translation*, vol. 14, no. 2, pp. 113–157, 1999.

[31] A. B. Phillips, “Cunei: open-source machine translation with relevance-based models of each translation instance,” *Machine Translation*, vol. 25, no. 2, p. 161, 2011.

[32] S. Dandapat, M. L. Forcada, D. Groves, S. Penkale, J. Tinsley, and A. Way, "Openmatrex: a free/open-source marker-driven example-based machine translation system," in *International Conference on Natural Language Processing*. Springer, 2010, pp. 121–126.

[33] R. Zens, F. J. Och, and H. Ney, "Phrase-based statistical machine translation," in *Annual Conference on Artificial Intelligence*. Springer, 2002, pp. 18–32.

[34] M. Collins, "Statistical machine translation: Ibm models 1 and 2," *Columbia Columbia Univ*, 2011.

[35] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, 2007, pp. 177–180.

[36] sayarghoshroy, "Statistical-machine-translation," 2020. [Online]. Available: https://github.com/sayarghoshroy/Statistical-Machine-Translation

[37] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[38] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Névéol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri, "Findings of the 2016 conference on machine translation," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 131–198. [Online]. Available: https://www.aclweb.org/anthology/W16-2301

[39] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, 2017, pp. 28–39.

[40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[41] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.

[42] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” *CoRR*, vol. abs/1409.1259, 2014. [Online]. Available: http://arxiv.org/abs/1409.1259

[43] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*, 2020, https://d2l.ai.

[44] O. Firat, K. Cho, and Y. Bengio, “Multi-way, multilingual neural machine translation with a shared attention mechanism,” *arXiv preprint arXiv:1601.01073*, 2016.

[45] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014.

[46] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, “Tensorflow: A system for large-scale machine learning,” in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. Savannah, GA: USENIX Association, Nov. 2016, pp. 265–283. [Online]. Available: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

[47] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones, Łukasz Kaiser, N. Kalchbrenner, N. Parmar, R. Sepassi, N. Shazeer, and J. Uszkoreit, “Tensor2tensor for neural machine translation,” 2018.

[48] P. Koehn and C. Monz, “Manual and automatic evaluation of machine translation between european languages,” in *Proceedings on the Workshop on Statistical Machine Translation*, 2006, pp. 102–121.

[49] C. Callison-Burch, “Fast, cheap, and creative: Evaluating translation quality using amazon’ s mechanical turk,” in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 286–295.

[50] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic

evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[51] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, "Deep recurrent models with fast-forward connections for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 371–383, 2016.

[52] S. Varsamopoulos, K. Bertels, and C. G. Almudever, "Designing neural network based decoders for surface codes," *arXiv preprint arXiv:1811.12456*, 2018.

[53] A. Lavie, "Evaluating the output of machine translation systems," *AMTA Tutorial*, vol. 86, 2010.

[54] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[55] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluation the role of bleu in machine translation research," in *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[56] K. A. Johnson, M. Babel, I. Fong, and N. Yiu, "SpiCE: A new open-access corpus of conversational bilingual speech in Cantonese and English," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4089–4095. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.503