# Machine translation models for Cantonese-English translation
# Project Plan

Liu Hey Wing (3035474016)

Supervisor: Dr. Yip Beta

4 October, 2020

# Contents

# 1 Introduction

## 1.1 Background

Machine Translation (MT) had been developed since 1950s [1, Chapter 1.3], and had been using in practise in recent years. However, numerous limitations had 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 (e.g. Megatron-LM[2]) and evaluating method (automatic or human).

An abstract concept of MT consist of 2 parts, 'declarative' and 'procedural' information, which can be simply understood as data and algorithms [1, Chapter 3]. The combination of these information construct the basis of an MT model, and will be elaborated in the following content.

### 1.1.1 Linguistic aspect

Necessary linguistic information are the 'declarative' in MT systems, which are based on facts, definitions and relationships of natural languages. These information may include grammatical description (dependency or generative), a corpus of orthography[1] or pronunciation of the source and target languages.

Generative grammar was proposed back in 1960s by Noam Chomsky in his book "Syntactic Structures" [3], which is a systematic approach to analyse and decompose sentences into tree structure. Chomsky argued that 'universal grammar' exist in natural languages, and introduced some parameters and methods attempting to derive such description.

Cantonese is a language ranked 18th in the world by the number of speakers [4]. The study of Cantonese linguistics traced back to 1900s with Ball [5], since then, linguists had provided different aspects about Cantonese extensively and continuously, such as phonology, grammar [6] [7] and sociologist studies.

Bernard Vauquois' pyramid is a very famous framework for preliminary MT, which focus on how linguistic information can be integrated, and describe different levels and layers of translation, such as direct and syntactic transfer approaches.

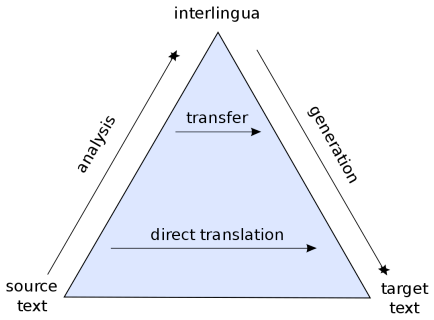



Figure 1: Bernard Vauquois' pyramid
Source: [8]

Some linguistic challenges are provided below (Figure 2), which had struggled with computational linguist for years, and some problems are still a hurdle in the development of MT.

[1] not all languages in the world uses 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
Source: [9]

### 1.1.2 Computational aspect

Algorithms, software and models are the 'procedural' information in MT, which means how the system interpret the 'declarative' information, represented as a form of program or procedures. There are different approaches for performing MT, each requires different 'declarative' information and perform different actions [10]. [11] listed out different categories of approaches, including

1. Dictionary based machine translation
2. Knowledge based machine translation
   (a) Rule-based machine translation (RBMT)
3. Corpus based machine translation
   (a) Example-based machine translation (EBMT)
   (b) Statistical Machine Translation (SMT)
   (c) Hybrid approaches
   combining SMT and EBMT to form word-based, phrase-based (PBSMT), syntax-based, forest-based and hierarchical approaches
   (d) Neural machine translation (NMT) [12]
      i. Long Short-Term Memory (LSTM) [13] [2]
      ii. Attention mechanism [14]

Although numerous approaches of MT models exist, there is a universal protocol or procedure among them. MT is always a process of input and output, where the models are the tools or intermediate between, nothing is different comparing with a human translator [14].

The development of MT changed a lot since 1960s [15], where the first 'real' MT system was introduced as RBMT. Then, the rules are getting much more complicated, where the 'pure' translation based on Vauquois's pyramid had reached some kind of limitation, and the increase in computational power leads to the development of other MT systems.

EBMT was introduced in the 1980s which focus on comparison and extraction [16]. Towards 1990s, IBM had performed a series of experiment for a new MT system based on sentence alignment which is later known as SMT, and the popularity continued until 2010s. The rise of NMT started around mid-2010s, together with the rise of AI and machine learning, which brings a lot of resources and attentions towards MT, and the relevant research is still ongoing.

[2] Google Neural machine translation (GNMT) is based on this approach, the cell was proposed back in 1997 and had been modified since then.

## 1.2 Objectives

Despite the abundant resources, there are 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, thus, the purpose of this FYP is to discover if there are any obstacles for Cantonese MT.

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.

## 1.3 Scope

MT is changing rapidly, with numerous variant proposed and some are being implemented commercially. Some of them are outdated and some have better performance, so our scope is to narrow down and focus on some categories of MT systems listed above. In the Conference on Machine Translation (WMT) 2019 [17], the most advanced and newly developed MT were presented, which are mainly SMT (PBSMT) and NMT (LSTM & Transformer). Some historically popular MT would be discussed in this project, including RBMT and EBMT.

We would implement different MT systems with a training set and a testing set of data, evaluate the accuracy with different metrics, and discover the reason for the lack of Cantonese MT.

## 1.4 Outline

In this project plan, the overall background is provided with objectives and scope. Investigation of MT systems include of 3 main components, data, algorithm and evaluation, where these components will be discussed in the methodology section. Finally, the project schedule and conclusion would be discussed.

# 2 Methodology

In this section, the information of each MT models would be listed out, including data collection, algorithms and evaluation. 5 particular MT models and a previous MT on Cantonese-English are chosen in this project, which covers the most advanced technologies, and had proved to perform certain level of accuracy.

## 2.1 MT models

### 2.1.1 RBMT

RBMT is based on linguistic knowledge, which deformat the source language, restructure the lexicons, and reformat into the target language as described in Figure 1. This was a popular model before 1980s, but it had faded out in use because the improvement of computational power had give rise to SMT and other systems.

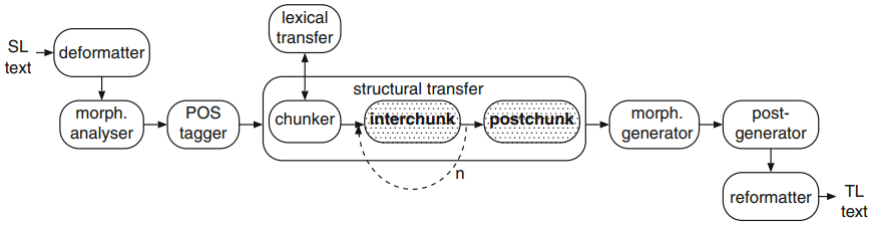



Figure 3: Structure of Apertium
Source: [18]

We adopt Apertium free/open-source (FOS) platform [18] as an example of RBMT to perform our training. It is heavily based on parallel (Bilingual) data and explicit linguistic data such as dictionaries, grammars, and structural transfer rules. The modules in Apertium perform different functions,

- *deformatter*: encapsulate input format information as superblank
- *morphological analyser*: segment surface forms (SF) (words) into lexical forms (LF) (lemma, lexical category) with dictionary
- *statistical PoS tagger*: use first-order hidden Markov model (HMM) to match LF with SF
- *lexical transfer*: translate SL LF into TL LF using bilingual dictionary
- *structural transfer (chunker)*: perform syntactic operations
- *morphological generator*: generate TL LF into TL SF
- *post-generator*: perform contraction or epenthesis (e.g. Englsih $a$ + $V$ -> $an$)
- *reformatter*: de-encapsulate format information

### 2.1.2 EBMT

EMBT is the first MT system get rid of the 'knowledge' of linguistics, and based on 'pure' corpus. Makoto Nagao [16] proposed that EMBT performs better than RBMT in language pairs with great divergence like English-Japanese. Input sentence would be searched for occurrences and similarities in the trained model, substitute the phrase or character, and generate the output sentence, as shown in example (1)-(3) [19].

(1) He buys a book on international politics.

(2) (a) He buys a notebook.
*Kare wa noto o kau.*
HE topic NOTEBOOK obj BUY.

(b) I read a book on international politics.
*Watashi wa kokusai seiji nitsuite kakareta hon o yomu.*
I topic INTERNATIONAL POLITICS ABOUT CONCERNED BOOK obj READ.

(3) Kare wa kokusai seiji nitsuite kakareta hon o kau.

On the other hand, Somers [20] had a thorough review about EBMT and addressed several limitations and difficulties[3] on further development and improvement. As a result, it introduced the idea of corpus-based MT, and give rise of SMT.

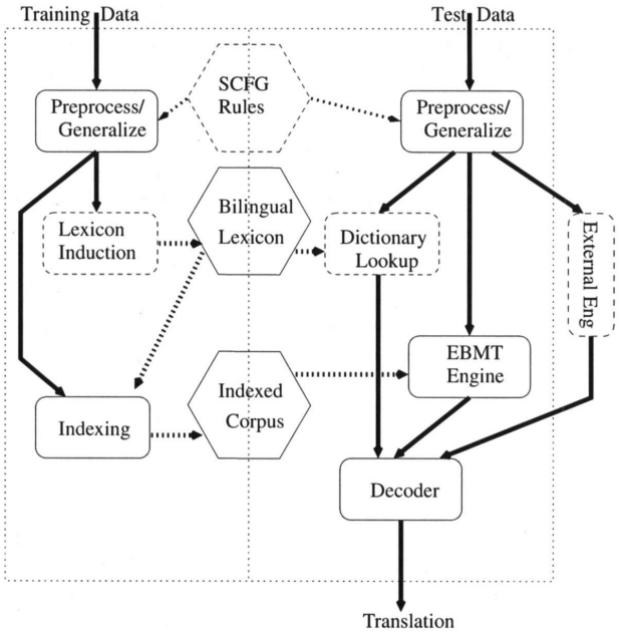



Figure 4: Structure of CMU-EBMT
Source: [21]

The rise and fall of EBMT was fast and before the era of internet, which induced low effort or resources[4] on the topic. CMU-EMBT [21] uses a nearly 'pure' EMBT, with optional components and integration to other MT models available. 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.

[3]Such as the size of example, lack of parallel corpora of language pairs, suitability of examples, matching etc.
[4]Some more known EBMT FOS platforms are already unavailable now, such as Cunei [22] and OpenMaTrEx [23] uploaded in 2010s. CMU-EBMT is the only available platform at current stage.

### 2.1.3 PBSMT

SMT uses a probabilistic approach to perform translations, which involves much less linguistic knowledge. The variant had developed from word-based to phrased-based approach and had been proven with high accuracy before the arise of NMT, however, the bottleneck for further improvement requires more involvement of linguistic analysis and complicated the structure [9].

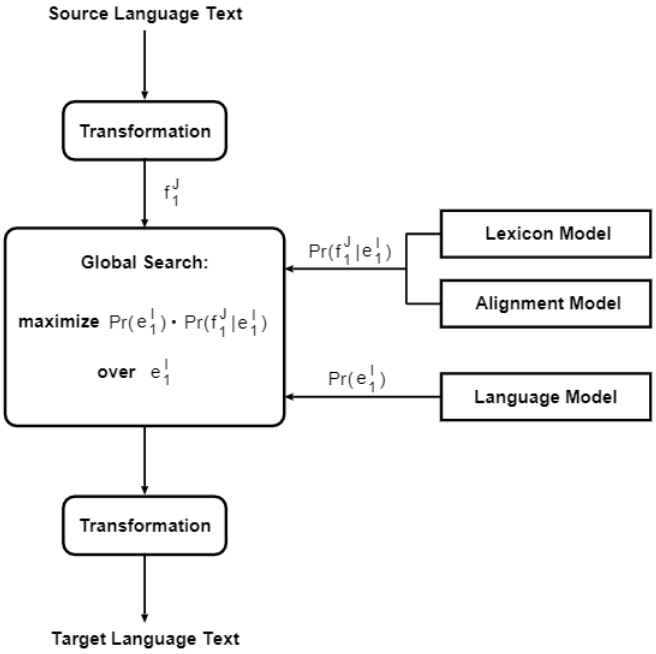



Figure 5: Structure of Bayes Decision Rule applying
Source: [24]

Bayes Decision Rule[5] describes $Pr(e_1^I|f_1^F)$ as the translation model (SMT), where $f_1^J = f^1, ..., f^j, ..., f^J$ is the source sentence needed to be translate into target sentence $e_1^I = e^1, ..., e^i, ..., e^I$. PBSMT segment words into a phrase and perform a phrase-to-phrase translation, and the mathematical description is below [24].

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

Moses [25] is one of the most popular open-source toolkit for SMT, which is famous for it flexibility. The model consist of 2 main component, *training pipeline* (language model) and *decoder* (translation model & Translation model) obeying Bayes' rule.

[5]It is a extremely brief description of Bayes decision rule, ignoring the probability of language model and other complicated information.

### 2.1.4 GRU

NMT uses 2 approaches, attention mechanism (see section 2.1.5) and RNN approach [26]. RNN approach consist of layers of feed-forward computation and recurrent neural network (RNN) between the encoder-decoder, which transform SL $x = \{x_1, ..., x_m\}$ into TL $y = \{y_1, ..., y_m\}$. The encoder would transform $x$ into a vector space of hidden layers[6], and the decoder would transform those vectors into $y$, using the likelyhood equation in the following [27].

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

NMT has been proven with significant improvement comparing with SMT, especially in WMT'16 [28] in different fields. The largest problem for NMT is the ambiguity within the hidden layers, which is not understandable by human and thus increase the difficulty for further improvement [29].

An example of RNN encoder-decoder structure is shown as below. RNN would pass information to its successor to reuse and preserve data, which also causes its low computational speed. The traditional $tanh - RNN$ cell had been proven with its unsatisfactory performance [30].

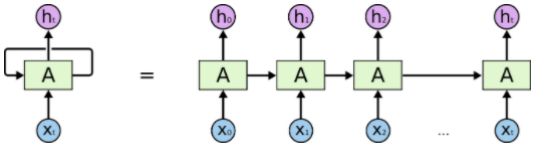
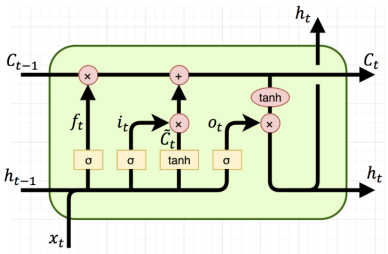



Figure 6: RNN encoder-decoder structure
Source: [31]

LSTM cell is an extension of recurrent neural network (RNN), which was proposed by Hochreiter and Schmidhuber [32] back in 1997, but the implementation in MT was introduced in 2014 [30].



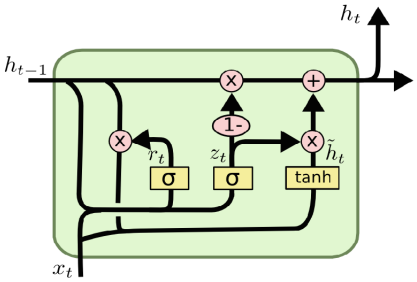

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$
$$h_t = tanh(\tilde{C}_t) * o_t$$

Figure 7: LSTM cell
Source: [33] adapted from [31]

The cell consist of 4 main component, namely *forget* gate $f_t$, *input* gate $i_t$, *output* gate $o_t$ and *memory* cell $C_t$ represented above, where $W$, $b$ are the parameter or vector, $\tilde{C}_t$ is the new memory, $x_t$ is the input and $h_t$ is the output.

[6]Hidden markov models (HMM) is a very complicated topic, which is out of the scope of this project

The crucial disadvantage for LSTM is its training speed, which is even slower than RNN, so Cho et al. [12] proposed a modified version called gated recurrent unit (GRU) with a high computational speed.




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

Figure 8: GRU cell
Source: [31]

GRU combines the input and forget gate into the *update* gate $z_t$, and introduced the *reset* gate $r_t$. Its performance in computational speed and memory space greatly improved in a study of Cho et al. [30] in most testing set.

Tensorflow [34] is an open-source library provided online providing large degree of customization. We would implement the GRU model presented by Cho et al. [12] and tested in [30], consist of 2000 hidden neurons and 620-dimensional layers.

### 2.1.5 Transformer

RNN approach faces a critical problem of mapping SL sentence into fixed-length vector, which would cause lost in input information and over-matching. Convolutional architectures using attention mechanism successfully prevents this bottle neck, and out-perform the traditional-RNN NMT in some aspect, such as long-sentence translation.

Attention mechanism, in contrast to RNN, does not uses a sigmoid function to normalize the input, rather, each segment has a vector representing the attention driven in a sentence. The encoder performs similar as in RNN, however the decoder would attempt to derive the output once at a time based on the attention towards the input, the result would be passed backwards and the model learns [35].

The implementation of attention mechanism in MT models was proposed by Bahdanau et al. [14] at 2014. Vaswani et al. [36] further extended the mechanism and proposed the Transformer model at 2017, which implemented stacked self-attention layers based on input sentence length to prevent lost in information.

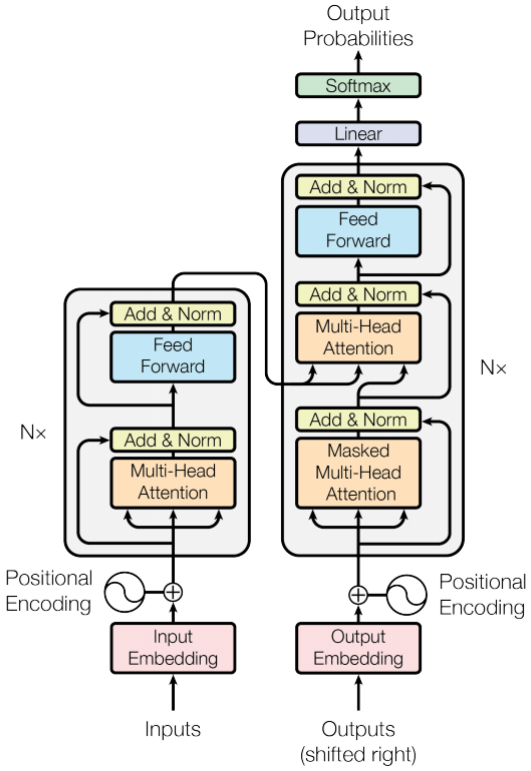



Figure 9: Transformer model
Source: [36]

The left side in Figure 9 is the encoding of input sentence, which is a attention block generating vector based on sentence length. The decoder on the right side are separated into 2 blocks, which the first attention block encodes the output sentence, and the second attention block computes the attention between the input vector and output vector. Finally, The result would be converted into probability functions.

The details of attention blocks and multi-layer is as below, where $Q$ is the attention matrices, $K$ is the key index and $V$ is the value.

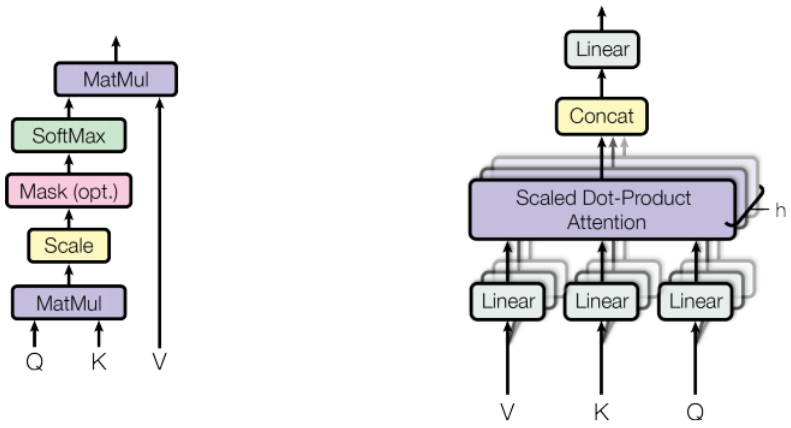



Figure 10: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

Source: [36]

We would apply the Transformer model in Tensorflow [34] with the configurations and library provided by Vaswasni [37].

### 2.1.6 Previous 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 [38].

Later in 2005, one of the author of PoluU-MT-99, Yan Wu, presented another MT system for Cantonese-English translation called LangCompMT05 [39], where the architecture is nearly identical to PoluU-MT-99 and the algorithms are extended.

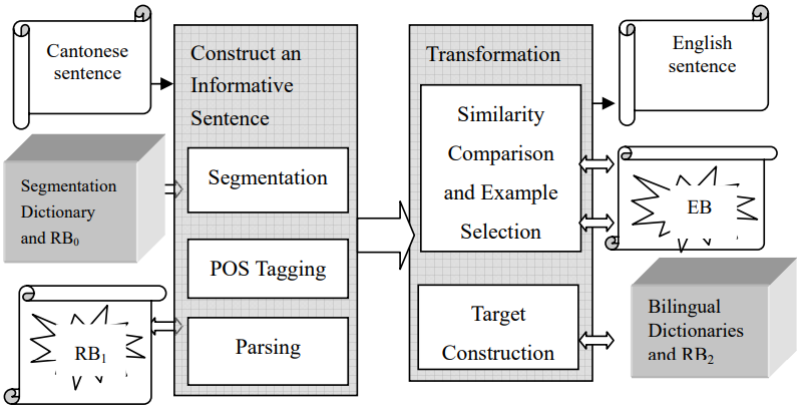



Figure 11: Structure of LangCompMT05
Source: [39]

However, the corpus data used in both PoluU-MT-99[7] and LangComp05[8] is not reliable. Some non-Cantonese words, sentences, aspect markers are used in the study, such as 有些, 的, 正在, which are from Modern Standard Chinese (MSC) or written form of Mandarin[9]. Thus, there are certain reservation towards the result of this study.

Unfortunately, the source code is not publicly available but the algorithms and structure is presented in the study. We would attempt to implement a similar model in this project.

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

[8]The example based was constructed by Shiwen Yu, but his/her corpus is not cited in [39]. The source of testing data was created by the authors, which are targeted to perform specific sentence structure translation rather than real data.

[9]Note that although Mandarin and Cantonese are both 'Chinese' as the same language and share some degree of written intelligibility of Hanzi and cognates, they are mutually unintelligible and are considered as different languages linguistically.

## 2.2 Data collection

As mentioned before, the study of Cantonese has been actively conducting for the last century. However, the lack of a unified/standard writing system in Cantonese had created various issues, but the written form is currently growing due to the development of internet in recent years [40].

Bilingual corpus is preferably used as training and testing data, as some of the MT models requires bilingual corpus to train. The written form is not restricted in this project, as long as it retains the same and compatible with dictionaries, so romanization such as Jyutping or Yale system and Hanzi are both accepted.

To my knowledge, the only available Cantonese-English bilingual corpus at current stage is ShefCE [41], which was purposely designed for speech recognition and transcription. Despite that the corpus exist, its transcription only includes Jyutping romanization without any tone marker[10]. Lexical tone in Cantonese contains significant amount of information [6], and the lack would cause information loss in MT systems, which is not tolerable in such complicated system.

In this situation, there are 2 possible solution for this problem. First, a Cantonese-English bilingual corpus called SpiCE [43] will be released in late 2020[11], which has an impact on the schedule (section 3) as planned. The second solution is to create a corpus from scratch, the possible data source includes YouTube subtitles, transcriptions of courts and implement transcribing model into Cantonese movies with English subtitles.

In extreme case, if all Cantonese-English resources cannot be obtained, the method described in [44] would be used, which used NMT to translate SL into TL without the resource (corpus) of TL.

## 2.3 Evaluating translation result

After the MT successfully translate a sentence, it is curious about how it performed, whether it translate correctly and how can we improve its result. The performance of MT systems are the key for further improvement, and several methods will be discussed in the following.

### 2.3.1 Manual evaluation

Manual evaluation is always considered time consuming and high cost, and the consistency in a large-sclae rating is difficult to maintain [45]. In current advancement in technology, machine can evaluate translation results in hours where human take days or weeks to complete, thus, manual evaluation are not preferred in most circumstances.

[46] had provided a new perspective using Amazon' s Mechanical Turk to recruit human evaluators, referred them as 'turkers'. It shows that the evaluating result of turkers are actually near to experts, where the experiment performed in several language including Chinese[12]. Considering the numbers of native speakers in Cantonese[4], the evaluation is feasible to be done using [46]'s method, however, the cost and time are not preferred in this project.

### 2.3.2 Automatic evaluation

MT systems often uses BLEU [47] to evaluate the results [8, 9, 14, 17, 26, 24, 27, 28, 30, 33, 35, 36, 39], which is a well established metric showing evaluation quality close to human. Since then, more metrics are proposed according to different dependencies, such as NIST, RED, WER, METEOR, LEPOR, NIST, ORANGE etc. The debate among metrics involves complicated computation and reasoning which are out of the scope of this project.

ASIYA [48] is a FOS platform includes some of the popular metrics currently in use, and generate scores. We would imply the system in our project to evaluate the results of our MTs.

[10]The study used a refined lexicon derived by linguistic knowledge and sub-syllable unit combinatorial constraints, and get rid of the lexical tones [42].

[11]The release had been delayed once in current situation, where the originally release date is in mid 2020.

[12]The turkers in Chinese are not native speakers with only less than 2 years of learning, which increase doubtless in their credibility.

## 2.4 Mini-conclusion

To conclude, our project will implement the software as described below.

- *MT models*:
  - *RBMT*: Apertium
  - *EBMT*: CMU-EMBT
  - *PBSMT*: Moses
  - *GNU*: Tensorflow with GRU cell
  - *Transformer*: Tensorflow with Transformer model
  - *LangComp05*: self-build
- *Data collection*: uncertainty exist
  - *Option 1*: ShefCE
  - *Option 2*: SpiCE
  - *Option 3*: self-build
- *Evaluation*: ASIYA

# 3 Schedule

| | |
|---|---|
| Oct 4, 2020 | Project Plan<br>Project Webpage |
| Oct - Nov, 2020 | Preparation for MT<br>1. Prepare training data (normalization & tokenization)<br>2. Open sources of MT models |
| Dec - Jan, 2020 | Implement evaluation system (manual or automatic) |
| Jan 24, 2021 | Interim Report |
| Feb- Apr, 2021 | Evaluation of MT<br>1. Testing data<br>2. Outline improvement |
| Apr 18, 2021 | Final report |

# 4 Conclusion

There is a lack in researches of Machine Translation between Cantonese and English. We discussed the linguistic and computational information for MT, and described details about MT systems. We will implement different MT systems, evaluate their results and outline difficulties and further improvement.

# 5 Bibliography


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

[2] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter language models using gpu model parallelism," *arXiv preprint arXiv:1909.08053*, 2019.

[3] N. Chomsky, *Syntactic structures*. Walter de Gruyter, 2002.

[4] 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/.

[5] 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

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

[7] 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.

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

[9] 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.

[10] 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.

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

[12] 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

[13] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016.

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

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

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

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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

[29] 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.

[30] 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.

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

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

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

[34] 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

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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

[42] 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.

[43] 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

[44] J. Gu, H. Hassan Awadalla, and J. Devlin, "Universal neural machine translation for extremely low resource languages," in *NAACL*, June 2018. [Online]. Available: https://www.microsoft.com/en-us/research/publication/universal-neural-machine-translation-extremely-low-resource-languages/

[45] 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.

[46] 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.

[47] 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.

[48] J. Giménez and L. Màrquez, "Asiya: an open toolkit for automatic machine translation (meta-) evaluation," *The Prague Bulletin of Mathematical Linguistics*, vol. 94, no. 2010, pp. 77–86, 2010.