Where Neural Machine Translation and Translation Memories Meet: Domain Adaptation for the Translation of HKSAR Government Press Releases

Siu Sai Cheong
School of Translation and Foreign Languages
The Hang Seng University of Hong Kong

Abstract

HKSAR Government Press Releases (https://www.info.gov.hk/gia/) by different bureaux and departments serve as an important channel of communication between the government and the public, and they are usually available in both English and Chinese, the two official languages of Hong Kong. The publication of the bilingual press releases could benefit from translation technology, especially with recent advances in neural machine translation (NMT). However, it would be suboptimal to use popular online NMT platforms as they are often designed for general texts and could have issues with terminology and style in the target language.

In this paper, we propose the method "pre-translation with translation memories" to adapt general NMT engines to the translation of the press releases from English into Chinese. Unlike conventional translation memories for human translators, our translation memory complements neural translation models, featuring length-ranked bilingual sentences and sub-sentential units for pre-translating the source text prior to NMT. Our experimental results show that our method outperforms the NMT-only baseline in terms of the BLEU score, suggesting that TMs may offer a simple solution to the adaptation of general NMT to domain-specific translation tasks, without the need of fine-tuning existing models.

1. Introduction

HKSAR Government Press Releases serve as an important communication channel between the Government and the public. They are often available in both English and Chinese, and the demand for translation has been high over the years, given the large number of announcements made by different departments and bureaux.

Recent advances in neural machine translation (NMT) may facilitate the translation process by, for example, providing translation drafts for post-editing before publication. Current NMT systems, however, tend to be developed for general purposes and may give suboptimal performance when used to translate government press releases. It is against this background that we propose the use of translation memories that pre-translate domain-specific expressions and sentences prior to machine translation, with a view to complementing NMT systems, especially those designed for general documents, and improving the translation quality.

The structure of this paper is as follows: Section 2 introduces bilingual government press releases in Hong Kong. Section 3 explores the possibilities of applying NMT to the translation of government press releases and identifies common issues. Section 4 presents our pre-translation solution, which features the integration of translation memory and machine translation as a domain adaptation method, and provides experimental results illustrating the effectiveness of the proposed approach. Section 5 discusses future research directions.

2. Publication of HKSAR Government Press Releases

HKSAR Government Press Releases include announcements, statements, speeches, event promotional materials, and other information, covering a wide spectrum of topics, such as immigration, labour, urban planning, postal services, finance, public health, recreation, and law and order. Table 1 provides examples of common types of government press releases and their corresponding areas (selected from press releases issued in early December 2021).

Table 1: Examples of government press releases

Type	Example	Area
News	Fifteen persons arrested during anti-illegal worker operations (Immigration Department, 2021)	Immigration
	Company and its director fined $65,000 for contravening Employment Ordinance (Labour Department, 2021)	Labour
Announcements	Approved Sha Tin Outline Zoning Plan amended (Town Planning Board, 2021)	Town Planning
	Speedpost services to Guadeloupe and Martinique suspended (Hongkong Post, 2021)	Postal Services
Speeches and transcripts of remarks	Speech by FS at 2021 Hong Kong Chartered Tax Adviser Conference (Financial Secretary, 2021)	Finance
	Transcript of remarks by S for S after Fight Crime Committee meeting (Secretary for Security, 2021)	Law and Order
Promotion of events	Public engagement activity on control of single-use plastics held today (Council for Sustainable Development, 2021)	Environmental Protection
	HKSAR Government to hold 2021 Constitution Day Seminar online (Constitutional and Mainland Affairs Bureau, 2021)	Constitutional Affairs
Statistics and figures	Statistics on vessels, port cargo and containers for the third quarter of 2021 (Census and Statistics Department, 2021)	Transport
	Land Registry releases statistics for November (Land Registry, 2021)	Housing and Land
Warnings and alerts	Beware of fraudsters posing as HKMA staff (Hong Kong Monetary Authority, 2021)	Finance
	Red flags hoisted at Silverstrand Beach and Clear Water Bay Second Beach (Leisure and Cultural Services Department, 2021)	Recreation
Health-related information	CHP investigates three additional confirmed cases of COVID-19 (Centre for Health Protection, 2021a)	Public Health
	CHP investigates outbreak of upper respiratory tract infection at kindergarten/nursery (Centre for Health Protection, 2021b)	Public Health

Government press releases are often issued in both English and Chinese (Traditional Chinese and Simplified Chinese), which are the two official languages of the city under Article 9 of the Basic Law (National People's Congress, 1990) and the Official Languages Ordinance (Cap. 5) (Hong Kong e-Legislation, 2021). They are published on a designated public website maintained by the Information Services Department (ISD) (https://www.info.gov.hk/gia/), as well as through other channels, such as the Government News and Media Information Platform designed for the press (https://www.isd.gov.hk/eng/gnmis.htm).

The demand for translation or preparation of the bilingual press releases has been high over the years, as evidenced by the large number of press releases issued by the government every year. Table 2 summarizes the number of press releases (in English and Chinese) published by the ISD from 2016 to 2020. The annual average was over 300,000, based on the Estimates section of the Budget from 2018/19 to 2021/22 (HKSAR Government, 2018, 2019, 2020, 2021).

Table 2: Government press releases (in English and Chinese) from 2016 to 2020

	2016	**2017**	**2018**	**2019**	**2020**
Press releases issued	328,934	329,704	320,423	294,125	251,519

The number of press releases on the ISD website (https://www.info.gov.hk/gia/) has also been high, with an annual average of 4.2 million English tokens and 6.2 million Chinese characters as shown in Table 3, which provides statistics for bilingual government press releases from 2016 to 2018 (in terms of tokens or characters) collected from the website:

Table 3: Bilingual government press releases on the ISD website

Year	**No. of English tokens**	**No. of Chinese characters**	**No. of Chinese tokens**
2016	4,237,046	6,272,943	3,586,224
2017	4,012,945	5,817,084	3,333,634
2018	3,846,370	5,538,538	3,171,356
Total	12,810,546	18,655,966	10,683,644

3. Machine Translation of Government Press Releases

The use of translation technology, such as machine translation, may facilitate the translation or preparation of bilingual government press releases, especially in the wake of the recent boom in NMT, which features the training of deep artificial neural networks for automatic translation (see, for example, Bahdanau, Cho, & Bengio, 2014; Luong, Pham, & Manning, 2015; Wu et al., 2016; Gehring et al., 2017; Vaswani et al., 2017). NMT has shown promising results with better performance compared with its predecessors such as statistical machine translation (see Bojar et al., 2016 for some examples) and has been widely adopted by different machine translation developers, such as Google (Wu et al., 2016), Microsoft (2021), Systran (Crego et al., 2016), and Tencent (2018).

Many NMT systems, however, are designed for general translation rather than the translation of government press releases. These models may not have been trained with sufficient exposure to government documents (and thus their terminology and writing style) and may give suboptimal results. To illustrate the problems with general systems, we have translated selected government press releases using Google Translate, a popular online automatic translation platform that supports more than 100 languages and different translation modes, such as text and speech translation (Google, n.d.). We have identified a few issues, namely word choices, sentence structure, terminology, person names, place names, formatting, and writing style. Table 4 gives some examples.

Table 4: Issues of machine translation

Translation Issue	Example and Discussion
Word choices	**Source Text:** The penalty for lighting fires illegally in the countryside is $25,000 and a year's imprisonment. **Reference Translation:** 在郊野非法生火的人士，可被罰款二萬五千元及入獄一年。 **Machine Translation:** 在農村非法生火的罰款為 25,000 美元和一年監禁。 **Discussion:** In the context of government press releases in Hong Kong, the expressions "countryside" and "$" should be translated as "郊野" (instead of "農村") and "港元" (instead of "美元") respectively.
Sentence structure	**Source Text:** Eligible Mainland holders of the electronic Exit-entry Permit for Travelling to and from Hong Kong and Macao (e-EEP) can enrol for the e-Channel service by using their e-EEP and undergoing the enrolment process at a traditional entry counter on their first visit to Hong Kong. Eligible Mainland visitors who have renewed their e-EEP have to go through the aforesaid enrolment process again before they can use the e-Channel. **Reference Translation:** 持有電子往來港澳通行證的合資格內地旅客，在首次使用電子通行證或在更換新的電子通行證後第一次抵港時，需使用傳統櫃位並在成功辦理入境及登記手續後，方可使用旅客 e-道服務。 **Machine Translation:** 符合條件的內地持有《往來港澳通行證》（e-EEP）的內地持有人，可使用 e-EEP 並在其傳統入境櫃檯辦理登記手續，辦理 e-道服務。第一次去香港。已續訂 e-EEP 的合資格內地旅客須重新辦理上述登記程序，方可使用 e-道。 **Discussion:** The second sentence in the machine translation output was incomplete and should have been incorporated into the first one. In addition, the Chinese title marks in the first sentence were improperly used, and the expression "renewed" was mistranslated as "已續訂".
Terminology	**Source Text:** SHA commends volunteer leaders of youth uniformed groups and outstanding youths (with photos) **Reference Translation:** 民政事務局局長嘉許青少年制服團隊義務領袖及優秀青年（附圖） **Machine Translation:** SHA 表彰青年軍裝團體志願者帶頭人及優秀青年（附圖） **Discussion:** The expressions "SHA", "youth uniformed group", and "volunteer leaders" are common terms in government documents in Hong Kong and should have been translated as "民政事務局局長" (instead of keeping the original abbreviation "SHA"), "青少年制服團隊" (instead of "青年軍裝團體"), and "義務領袖" (instead of "志願者帶頭人") respectively.
Person names	**Source Text:** The Under Secretary for Food and Health, Dr Chui Tak-yi, will be the Acting Secretary for Food and Health during her absence. **Reference Translation:** 她離港期間，食物及衞生局副局長徐德義醫生將署任食物及衞生局局長。 **Machine Translation:** 食物及衛生局副局長崔德儀博士將在她缺席期間擔任代食物及衛生局局長。

	Discussion: The name of the Acting Secretary for Food and Health, Dr Chui Tak-yi (徐德義醫生), was mistranslated as "崔德儀博士".
Place names	**Source Text:** Southbound Wong Nai Chung Road between Village Road and the Public Stands of HKJC; **Reference Translation:** - 介乎山村道與馬會公眾看台的黃泥涌道南行線; **Machine Translation:** - 黃泥湧道南行，介乎村道與馬會公眾看台之間； **Discussion:** The street names "Wong Nai Chung Road" (黃泥涌道) and "Village Road" (山村道) were mistranslated as "黃泥湧道" and "村道" respectively.
Formatting	**Source Text:** Tickets priced at $55 are now available at URBTIX (www.urbtix.hk). For credit card telephone bookings, please call 2111 5999. **Reference Translation:** 門票五十五元，現於城市售票網(www.urbtix.hk)發售；信用卡電話購票：二一一一 五九九九。 **Machine Translation:** 票價為 $55 的門票現已於 URBTIX (www.urbtix.hk) 發售。信用卡電話訂票請致電 2111 5999。 **Discussion:** The extensive use of Chinese numerals (e.g., "五十五" for "55") in government press releases in Chinese was not retained in the machine translation output.
Writing style	**Source Text:** The following is issued on behalf of the University Grants Committee: **Reference Translation:** 下稿代大學教育資助委員會發出: **Machine Translation:** 以下是代表大學教育資助委員會發出的： **Discussion:** The expression "the following is issued on behalf of (a unit)" was translated as "以下是代表……發出的" instead of "下稿代……發出", with the latter being more common in official press releases.

From Table 4, we can see that it may be less preferable to apply general systems directly to the preparation of bilingual government press releases, as this may require extensive human editing after machine translation. A more desirable option is machine translation with domain adaptation, which identifies ways of adapting general systems to the translation of specialized documents, in our case government press releases. Our proposed solution is presented in the next section.

4. Domain adaptation with Translation Memory

4.1 Pre-translation: The Concept

There are different ways to adapt a general NMT system to a specialized domain (see, for example, Freitag & Al-Onaizan, 2016; Chu, Dabre, & Kurohashi, 2017; Chu & Wang, 2018; Saunders, 2021). Common approaches such as the use of domain-specific data may require considerable computational overhead to (re-)train models. An alternative method that does not require model (re-)training is to use a translation memory before machine translation.

Conventionally, translation memories, which are databases storing translated sentences together with their corresponding source segments (Chan, 2004: 251), are computer-aided translation applications designed for translators to retrieve and reuse translated segments,

whereas in our case we propose to use translation memories in an automated manner to complement general NMT systems for translating specialized documents. More specifically, as shown in Figure 1, the first step is to pre-translate source language segments and sentences prior to neural machine translation, with a focus on terminology and common domain-specific expressions, using a specialized translation memory comprising bilingual units across linguistic ranks. In other words, certain parts of the source text are translated using the translation memory before being fed into an NMT system in the second step, and this may allow the translation memory and NMT to focus on the translation of specialized and general content respectively. The third step "post-translation" is the restoration of the target text by replacing the placeholders in the machine translation output with the corresponding target language expressions according to the translation memory.

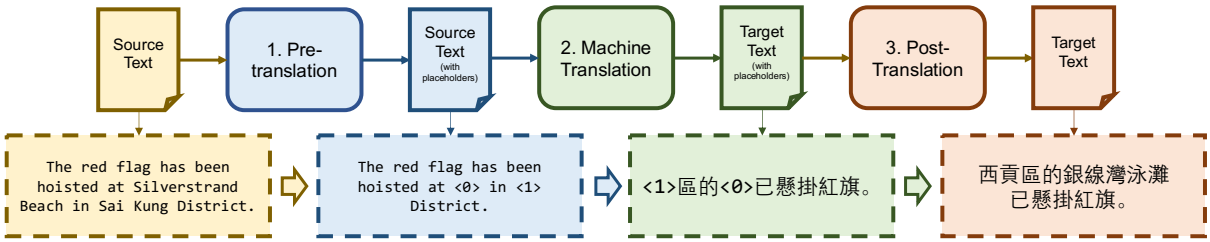

Figure 1: NMT with pre-translation and post-translation. In this illustrative example, given the source text "The red flag has been hoisted at Silverstrand Beach in Sai Kung District.", the first step is pre-translation, where the expressions "Silverstrand Beach" and "Sai Kung" are replaced by the placeholders "<0>" and "<1>". The second step is the automatic translation of the pre-translated source text "The red flag has been hoisted at <0> in <1> District.", and the result is "<1>區的<0>已懸掛紅旗。". The third step is post-translation, which replaces the placeholders in the machine translation output with the actual expressions in the target language. In our example, "<0>" and "<1>" in the machine output are replaced by "銀線灣泳灘" and "西貢" respectively, and the final result is "西貢區的銀線灣泳灘已懸掛紅旗。".

4.2 Development of Translation Memory and Test Dataset

To further illustrate this idea, we built a translation memory for NMT and compared the corresponding output of machine translation with the one without pre-translation. To develop the translation memory, we first collected and aligned bilingual government press releases between 2016 and 2018, with a total of around 408,000 translation units (see Table 5 for details). We then removed bilingual units with low interlingual similarity, and after filtering, we obtained 270,000 pairs of sentences and segments.

Table 5: Translation units extracted from government press releases (2016-2018)

Year	**No. of translation units**
2016	148,324
2017	133,384
2018	127,034
Total	**408,742**

Our translation memory for pre-translation differs from sentence-based translation memories for translators in that ours contains both bilingual sentences and phrases. Table 6 gives examples of different types of translation units we have, including (a) sentences, (b) common expressions, (c) job titles, (d) names of units, (e) names of policies or initiatives, and (f) other technical terms.

Table 6: Types of translation units for pre-translation

Translation unit type	Examples
Sentences	The HKSAR Government attaches great importance to upholding academic freedom and institutional autonomy. 香港特區政府一直致力維護學術自由及院校自主。 The Government is determined and committed to enhancing the well-being of elderly people. 此外，政府有决心和承擔為長者謀福祉。
Common expressions	LegCo committee meeting and public hearing 立法會委員會會議及公開聆訊 Public urged to stay alert against avian influenza 市民應提高警惕預防禽流感 The membership of the Advisory Committee is as follows: 諮詢委員會的成員如下：
Job titles	Secretary for Innovation and Technology 創新及科技局局長 Under Secretary for Home Affairs 民政事務局副局長 Director-General of Trade and Industry 工業貿易署署長
Names of units	Architectural Services Department 建築署 Business Environment Council 商界環保協會 Sha Tau Kok District Rural Committee 沙頭角區鄉事委員會
Names of policies or initiatives	Qualifying Debt Instrument Scheme 合資格債務票據計劃 Redevelopment of Tai Hang Sai Estate 大坑西邨重建計劃 SME Financing Guarantee Scheme 中小企融資擔保計劃
Other technical terms	Effective Exchange Rate Index 港匯指數 Free Trade Agreements (FTAs) 自貿協定 Green Bond and Green Finance 綠色債券和金融

We created a test dataset to evaluate the results of NMT with/out pre-translation. We first collected bilingual press releases published in January 2019 and randomly selected approximately 5,000 sentences, which consisted of 156,473 English tokens and 130,646 Chinese tokens. The sentence and expression match rates with respect to our translation memory were 18% and 4% respectively. The former is the percentage of test sentences matching a sentence in the translation memory, while the latter is the percentage of test sentences matching at least one sub-sentential expression in the translation memory.

4.3 Experiment 1: Google Translate

We evaluated Google Translate using the test dataset, which was translated twice: once with our translation memory for pre-translation and once with machine translation alone. We

adopted the Bilingual Evaluation Understudy (BLEU) metric (Papineni, Roukos, Ward, & Zhu, 2002), a widely-accepted measure for the automatic assessment of the performance of machine translation by calculating modified n-gram precision. We computed BLEU scores (up to 4-grams; i.e. BLEU4) by comparing the machine translation results with the reference sentences in the test dataset, both of which were tokenized using a SentencePiece tokenizer (Kudo & Richardson, 2018) trained on general and government documents. As shown in Table 7, the BLEU score of Google Translate with pre-translation was around 11 points higher than the one with machine translation alone. Table 8 provides sample translation results with/out pre-translation.

Table 7: BLEU scores of Google Translate with/out TM

	Google Translate with pre-translation	**Google Translate without pre-translation**	**BLEU difference after pre-translation**
BLEU score based on the test set (BLEU4)	23.73	**35.51**	+11.78

Table 8: Sample translation results with/out TM (Google Translate)

Sample	**Source text and translation results**
Sample 1	**Source Text:** Investigation by the Special Investigation Team of Traffic, Kowloon East is underway. **Reference Translation:** 東九龍總區交通部特別調查隊正跟進調查該宗意外。 **Output 1 (without pre-translation):** 九龍東交通特別調查組正進行調查。 **Output 2 (with pre-translation):** 東九龍總區交通部特別調查隊正跟進調查案件。 **Discussion:** Output 2 gave the correct translation of the term "Special Investigation Team of Traffic, Kowloon East" (i.e., "東九龍總區交通部特別調查隊", as opposed to "九龍東交通特別調查組" in Output 1).
Sample 2	**Source Text:** His nasopharyngeal aspirate tested positive for influenza A virus upon laboratory testing. **Reference Translation:** 他的鼻咽分泌樣本經化驗後，證實對甲型流感病毒呈陽性反應。 **Output 1 (without pre-translation):** 他的鼻咽部抽吸物經實驗室檢測呈甲型流感病毒陽性。 **Output 2 (with pre-translation):** 病人的鼻咽分泌樣本經化驗後，證實對甲型流感病毒呈陽性反應。 **Discussion:** For the translation of "nasopharyngeal aspirate", the expression "鼻咽分泌樣本" in Output 2 was clearer than the term "鼻咽部抽吸物" in Output 1 and is more commonly used in government press releases.
Sample 3	**Source Text:** Admission to the ward has been suspended and restricted visiting has been imposed. **Reference Translation:** 該病房已暫停接收新症,並實施有限度探訪安排。 **Output 1 (without pre-translation):** 該病房已暫停入場，並已實施限制探視。 **Output 2 (with pre-translation):** 有關病房已暫停接收新症，並實施有限度探訪安排。 **Discussion:** In Output 1, the word "admission" was mistranslated as "入場", which was inappropriate in the context of "admission to the ward" and less preferable to the expression "接收新症" in Output 2.

From the examples in Table 8, we can see that our integrated approach could enhance the translation of proper nouns and the selection of expressions that are more in line with the writing style of government press releases.

4.4 Experiment 2: New MT Models Based on the LSTM Architecture

To further explore how our translation memory could contribute to machine translation systems trained on different amounts of domain-specific data, we trained three new translation models: Model 1, Model 2, and Model 3. We trained Model 1 on out-of-domain data comprising 24 million sentence pairs from general documents. We continued to train Model 1 on text data relating to government and public affairs (e.g., selected sentence pairs from Hong Kong bilingual legislation and documents of the Legislative Council) to build Model 2. We then built Model 3 by training Model 2 on the raw bilingual sentences extracted from the press releases between 2016 and 2018 as discussed in Section 4.2. Given the different combinations of training data used, Model 1, Model 2, and Model 3 can be considered out-of-domain, partially in-domain, and in-domain translation models respectively.

Similar to other recurrent neural networks for translation (Sutskever, Vinyals & Le (2014), Bahdanau, Cho & Bengio (2014), Luong, Pham & Manning (2015), and Y. Wu et al. (2016)), the three models adopt the bi-directional long short-term memory (LSTM) architecture (Hochreiter & Schmidhuber, 1997) comprising the encoder and decoder (4 layers each) with the attention mechanism (see Figure 2). The encoder converts tokens in the source language into an intermediate representation, which is then used by the decoder for the generation of tokens in the target language. The encoder and decoder tokens are in the form of word embeddings trained on general and government documents, with a dimension of 512 and vocabulary sizes of 48,664 (English) and 59,440 (Chinese).

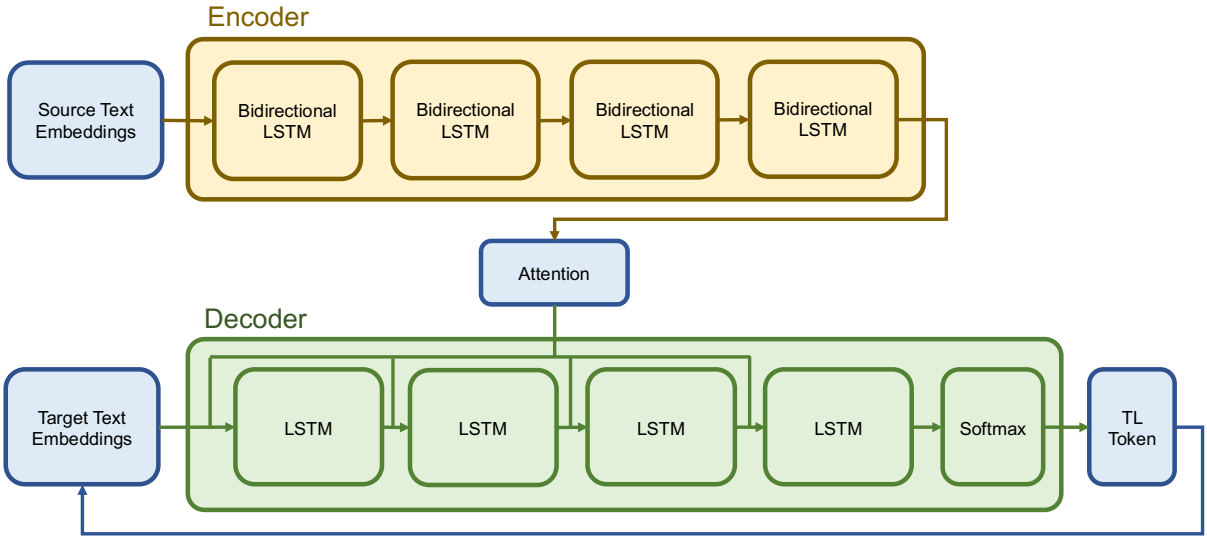

Figure 2: Architecture of LSTM translation models. Each of our LSTM models comprises an encoder and a decoder. The encoder consists of four bidirectional LSTM layers. The output of the last layer is sent to the attention module, which is connected to the decoder. The decoder is composed of four LSTM layers and a softmax layer, which predicts the next target language token.

In this experiment, the translation quality of all the three models improved after incorporating our translation memory into the translation process. As shown in Table 9, the BLEU scores of

Model 1, Model 2, and Model 3 without pre-translation were 15.56, 18.09, and 38.75 respectively, with higher scores for the two models exposed to government documents during the training phase. The scores of the three models with pre-translation went up by 13.72, 13.14, and 5.96 BLEU points respectively, and Model 3 obtained higher BLEU scores than Google Translate. Similar to the previous experiment, the results here also suggest that pre-translation with translation memory may improve the accuracy of the target text (e.g., terminology and word selection), as shown in the translation samples in Table 10.

Table 9: BLEU Scores for NMT (LSTM) with/out TM

	BLEU score of model without pre-translation	**BLEU score of model with pre-translation**	**BLEU difference after pre-translation**
Model 1	15.56	29.28	+13.72
Model 2	18.09	31.23	+13.14
Model 3	38.75	**44.71**	+5.96

Table 10: Sample translation results with/out TM (Our models)

Sample	**Source text and translation results**
Sample 1	**Source Text:** The Expo will be held from 11am to 5.30pm at Wo Hing Sports Centre, 8 Wo Ming Lane, Fanling (next to Wah Ming Estate Bus Terminus). Admission is free. **Reference Translation:** 博覽會於上午十一時至下午五時三十分在粉嶺和鳴里 8 號和興體育館（鄰近華明邨巴士總站）舉行，入場費全免。 **Output 1 (Model 3 only):** 博覽會於上午十一時至下午五時三十分在粉嶺和明里八號和興體育館(緊接華明邨巴士總站)舉行,免費入場。 **Output 2 (Model 3 with pre-translation):** 博覽會於上午十一時至下午五時三十分在粉嶺和鳴里 8 號和興體育館（鄰近華明邨巴士總站）舉行，費用全免。 **Discussion:** The street name “Wo Ming Lane” was translated correctly as “和鳴里” in Output 2, but incorrectly as “和明里” in Output 1.
Sample 2	**Source Text:** The HKG is a territory-wide major multi-sports event held biennially with the 18 District Councils as the participating units. **Reference Translation:** 港運會是兩年一度的大型綜合運動會，以全港十八區區議會為參賽單位。 **Output 2 (Model 3):** 港運會每每兩年與十八區區議會舉辦全港性大型賽事,以參與單位為參與單位。 **Output 3 (Model 3 with pre-translation):** 每兩年舉辦一屆的港運會，是以全港十八區區議會為參賽單位的大型綜合運動會。 **Discussion:** In Output 1, the expressions “held biennially” and “as the participating units” were mistranslated as “每每兩年” and “以參與單位為參與單位” and were fixed in Output 2.
Sample 3	**Source Text:** The Leisure and Cultural Services Department (LCSD) will implement special opening hours at its performance venues and URBTIX outlets during the Lunar New Year holidays, an LCSD spokesman announced today (January 21). Details are as follows: **Reference Translation:** 康樂及文化事務署（康文署）發言人今日（一月二十一日）宣布，農曆新年期間轄下各表演場地及城市售票網售票處的開放時間如下： **Output 1 (Model 3):** 康文署發言人今日(一月二十一日)宣布,康樂及文化事務署(康文署)將於農曆新年假期舉行特別開放時間及城市售票網,詳情如下:

	Output 2 (Model 3 with pre-translation): 康文署發言人今日(一月二十一日)宣布,在農曆新年假期期間,其表演場地及城市售票網售票處會實施特別開放時間,詳情如下: **Discussion:** The translation of "will implement special opening hours" in Output 2 ("實施特別開放時間") was more appropriate in terms of collocation than the translation in Output 1 ("舉行特別開放時間").

It is noteworthy that there was an increase in the BLEU score even for Model 3, which had been trained on in-domain data. This suggests that the use of translation memory may improve the translation of expressions or sentences that are present in the training data but not "learned" well in the process of training, and that our method could work jointly with not only general models but also specialized models that have already been trained on domain-specific data.

4.5 Experiment 3: New MT Models Based on the Self-attention Transformer Architecture

The recent popularity of using the Transformer model (Vaswani et al., 2017) (or its variants) for not only machine translation but also other natural language processing applications (e.g., Devlin et al., 2019) and even image processing (Dosovitskiy et al., 2020) suggests that it would be worthwhile to assess the applicability of our method to NMT engines adopting the self-attention architecture. In this experiment, we built three more MT models (Model 4, Model 5, and Model 6) in a manner similar to our second experiment, but with the Transformer architecture (see Figure 3) rather than LSTM. In other words, Models 4, 5, and 6 were out-of-domain, partially in-domain, and in-domain Transformer models respectively. The configuration of each model follows the base WMT Transformer model proposed by Vaswani et al. (2017), with a 6-layer encoder, a 6-layer decoder, 512-dimensional positional embeddings, and 2048-dimensional position-wise feed-forward networks.

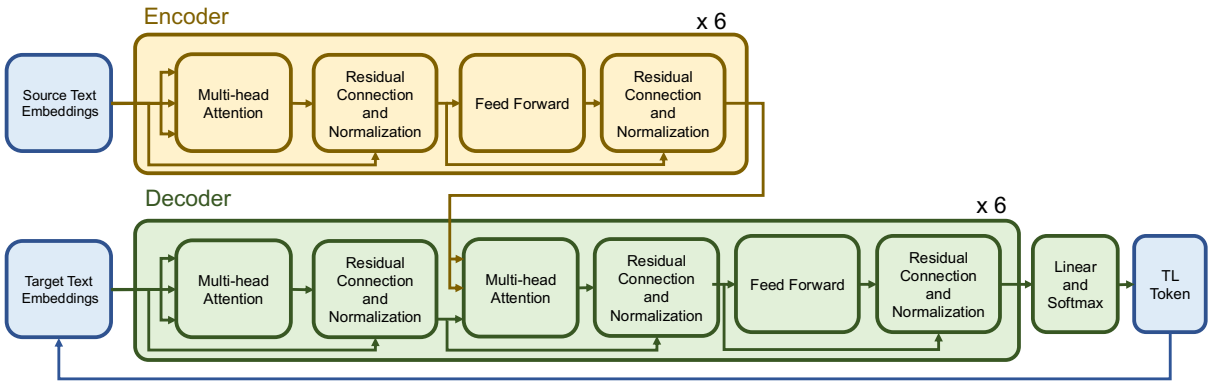

Figure 3: Architecture of self-attention translation models. Each of our self-attention models comprises an encoder and a decoder, with 6 encoding layers and 6 decoding layers respectively. Each encoding layer consists of a multi-head attention layer, a feed-forward network, and residual connection and normalization layers. Each decoding layer is similar to the encoding layer, but it is composed of two multi-head attention layers, one of which gets its input from the output of the encoder. The output of the last decoding layer undergoes linear transformation and is fed into a softmax layer for the prediction of the next token in the target language.

The results with/out pre-translation are shown in Table 11. The BLEU scores of our self-attention models without pre-translation were similar to those of the corresponding models in our second experiment, with the (partially) in-domain Transformer models achieving higher scores. We also observed a notable increase in the BLEU score for each of the Transformer models after the integration of our translation memory and NMT: 13.88 (for Model 4), 12.41

(for Model 5), and 3.3 (for Model 6). The results of this experiment suggest that our method can be extended to the Transformer models.

Table 11: BLEU Scores for NMT (Transformer) with/out TM

	BLEU score for model without pre-translation	**BLEU score for model with pre-translation**	**BLEU difference after pre-translation**
Model 4	14.40	28.28	+13.88
Model 5	18.56	30.97	+12.41
Model 6	42.86	**46.16**	+3.3

5. The Way Forward

We have explored the use of translation memory as a simple, effective tool that could complement NMT systems and facilitate their domain adaptation. It is hoped that our findings will facilitate the computer-aided translation of government press releases in Hong Kong (and perhaps other types of government documents and specialized documents in other research fields) and shed light on ways to make better use of general machine translation systems in domain-specific settings.

There are three noteworthy directions for future research: First, given that government documents are not restricted to press releases, we may build a larger translation memory that incorporates expressions and sentences from other sources and investigate whether we can further boost the quality of NMT by means of pre-translation and apply this method to the translation of not only press releases but also other texts relating to government and public affairs. We may also need to consider how to design translation memories for pre-translation to optimize the overall matching rate for better translation performance. Key issues may include the distribution of translation units across linguistic ranks (e.g., phrases and clauses), the inclusion (or exclusion) of certain categories of translation units (e.g., the inclusion of proper noun phrases and exclusion of common noun phrases might be of help), and the use of better alignment methods to reduce the number of misaligned translation units, which were also present in the translation memory in our experiment and had an impact on the overall translation quality.

Second, we may explore the impact of the matching and restoration mechanisms on the quality of integrated translation. For example, given that the present study is restricted to exact matching (i.e., replacing source language expressions or sentences only when an exact match is available), we could consider fuzzy matching and study whether it could further enhance the quality of the final translation output. The form of the pre-translated text or the representation of pre-translated segments prior to machine translation also deserves our attention. In our experiments, we used placeholders in the pre-translated text, and they contained little information about the original expressions in the source language, which could have provided useful contextual information for NMT and facilitated decoding if they had been preserved. It would therefore be useful to experiment with other ways of representing pre-translated elements, especially ways that can preserve the segments or sentences that have been replaced in pre-translation. Another challenge is that placeholders in the pre-translated text may disappear in the NMT output. This issue, which could be caused by the symbols that are used to represent the placeholders and the way in which the model is trained, could be partially addressed through proper design of the post-translation step for target text restoration. For example, given the different characteristics of the training data used for the training of Models 1-3 in our second experiment (and also Models 4-6 in the third experiment), the placeholder symbols and restoration methods were slightly adjusted accordingly. The

optimization of these placeholder-related areas may have a positive impact on the final translation results.

Third, we may also examine the robustness of our approach and conduct a comparative study of the performance of integrated translation in different settings, such as other language pairs, model types (e.g., convolutional NMT models), and hyperparameters (e.g., beam size for decoding). The findings could in turn facilitate the application of our method to other translation projects with varying models and translation directions.

Acknowledgements

The work described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. UGC/FDS14/H16/18).

References

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015, January). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

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

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., ... & Negri, M. (2016). Findings of the 2016 Conference on Machine Translation. In *ACL 2016 First Conference on Machine Translation (WMT16)* (pp. 131-198). The Association for Computational Linguistics.

Census and Statistics Department. (2021). Statistics on vessels, port cargo and containers for the third quarter of 2021. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120200282.htm

Centre for Health Protection. (2021a). CHP investigates three additional confirmed cases of COVID-19. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120300475.htm

Centre for Health Protection. (2021b). CHP investigates outbreak of upper respiratory tract infection at kindergarten/nursery. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120300609.htm

Chan, S. W. (2004). *A dictionary of translation technology*. Hong Kong: Chinese University Press.

Chu, C., & Wang, R. (2018, August). A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1304-1319).

Chu, C., Dabre, R., & Kurohashi, S. (2017, July). An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 385-391).

Constitutional and Mainland Affairs Bureau. (2021). HKSAR Government to hold 2021 Constitution Day Seminar online. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120200527.htm

Council for Sustainable Development. (2021). Public engagement activity on control of single-use plastics held today. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120300459.htm

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., ... & Enoue, S. (2016). SYSTRAN's Pure Neural Machine Translation Systems. *arXiv preprint arXiv:1610.05540*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Financial Secretary. (2021). Speech by FS at 2021 Hong Kong Chartered Tax Adviser Conference. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120300365.htm

Freitag, M., & Al-Onaizan, Y. (2016). Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017, July). Convolutional sequence to sequence learning. In *International Conference on Machine Learning* (pp. 1243-1252). PMLR.

Google. (n.d.). Translate: Explore the world in over 100 languages. *Google Translate*. Retreived from https://translate.google.com/intl/en/about/languages/

HKSAR Government. (2018). Head 74 – Information Services Department, Estimates for the year ending 31 March 2019. *The 2018-19 Budget*. Retrieved from https://www.budget.gov.hk/2018/eng/pdf/head074.pdf

HKSAR Government. (2019). Head 74 – Information Services Department, Estimates for the year ending 31 March 2020. *The 2019-20 Budget*. Retrieved from https://www.budget.gov.hk/2019/eng/pdf/head074.pdf

HKSAR Government. (2020). Head 74 – Information Services Department, Estimates for the year ending 31 March 2021. *The 2020-21 Budget*. Retrieved from https://www.budget.gov.hk/2020/eng/pdf/head074.pdf

HKSAR Government. (2021). Head 74 – Information Services Department, Estimates for the year ending 31 March 2022. *The 2021-22 Budget*. Retrieved from https://www.budget.gov.hk/2021/eng/pdf/head074.pdf

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Hong Kong e-Legislation. (2021). Official languages and their status and use. *Official languages ordinance (Cap. 5) of the Laws of Hong Kong*. Retrieved from https://www.elegislation.gov.hk/hk/cap5

Hong Kong Monetary Authority. (2021). Beware of fraudsters posing as HKMA staff. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120300603.htm

Hongkong Post. (2021). Speedpost services to Guadeloupe and Martinique suspended. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120300526.htm

Immigration Department. (2021). Fifteen persons arrested during anti-illegal worker operations. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120300469.htm

Kudo, T., & Richardson, J. (2018, November). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 66-71).

Labour Department. (2021). Company and its director fined $65,000 for contravening Employment Ordinance. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120300476.htm

Land Registry. (2021). Land Registry releases statistics for November. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/02/P2021120200347.htm

Leisure and Cultural Services Department. (2021). Red flags hoisted at Silverstrand Beach and Clear Water Bay Second Beach. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120300234.htm

Luong, M. T., Pham, H., & Manning, C. D. (2015, September). Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1412-1421).

Microsoft. (2021). Neural machine translation. *Microsoft research*. Retrieved from: https://www.microsoft.com/en-us/research/project/neural-machine-translation/

National People's Congress. (1990). Article 9. *The Basic Law of the Hong Kong Special Administrative Region of the People's Republic of China*. Hong Kong: Consultative Committee for the Basic Law of the Hong Kong Special Administrative Region of the People's Republic of China.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

Saunders, D. (2021). Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *arXiv preprint arXiv:2104.06951*.

Secretary for Security. (2021). Transcript of remarks by S for S after Fight Crime Committee meeting. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120300730.htm

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, *27*, 3104-3112.

Tencent. (2021). Machine Translation (Chinese version). *Tencent Cloud*. Retrieved from https://cloud.tencent.com/product/tmt

Town Planning Board. (2021). Approved Sha Tin Outline Zoning Plan amended. *HKSAR Government Press Releases.* Retrieved from https://www.info.gov.hk/gia/general/202112/03/P2021120300299.htm

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.