RGC Ref. No.:
UGC/FDS14/H16/18
(please insert ref. above)

RESEARCH GRANTS COUNCIL
COMPETITIVE RESEARCH FUNDING SCHEMES FOR THE LOCAL SELF-FINANCING DEGREE SECTOR

FACULTY DEVELOPMENT SCHEME (FDS)

Completion Report
(for completed projects only)

Submission Deadlines:

1. *Auditor's report with unspent balance, if any: within **six** months of the approved project completion date.*
2. *Completion report: within **12** months of the approved project completion date.*

Part A: The Project and Investigator(s)

1. Project Title

A Hybrid Approach to the Translation of Government Press Releases: Integration of Translation Memories and Neural Machine Translation

2. Investigator(s) and Academic Department(s) / Unit(s) Involved

Research Team	Name / Post	Unit / Department / Institution
Principal Investigator	Dr SIU Sai Cheong / Associate Professor	School of Translation and Foreign Languages/ The Hang Seng University of Hong Kong
Co-Investigator(s)	Dr SIU Sai Yau / Assistant Professor	School of Translation and Foreign Languages/ The Hang Seng University of Hong Kong
Others		

3. Project Duration

	Original	Revised	Date of RGC / Institution Approval *(must be quoted)*
Project Start Date	1 January 2019	-	N/A
Project Completion Date	31 December 2021	30 June 2022	Approved by HSUHK on 6 October 2021

Duration *(in month)*	36 months	42 months	
Deadline for Submission of Completion Report	31 December 2022	30 June 2023	

Part B: The Final Report

5. Project Objectives

5.1 Objectives as per original application

1. study the major problems of the out-of-domain English/Chinese neural machine translation (NMT) of HK government press releases;
2. propose an integrated approach to the English/Chinese NMT of government press releases, which will involve (i) the pre-translation of the input using a translation memory (TM) that contains of aligned bilingual sentences from government press releases and (ii) the use of specialised NMT (SNMT), which consists of in-domain word-embeddings and attention-based encoder-decoder recurrent neural networks trained on both out-of-domain and in-domain government documents; and
3. evaluate the integrated approach.

5.2 Revised objectives

Date of approval from the RGC: N/A

Reasons for the change: N/A

5.3 Realisation of the objectives

(Maximum 1 page; please state how and to what extent the project objectives have been achieved; give reasons for under-achievements and outline attempts to overcome problems, if any)

All three project objectives have been fully achieved. The details are as follows:

Objective 1: Studying the major problems of out-of-domain E-C neural machine translation (NMT) of Hong Kong government press releases. Bilingual government press releases from 2016 to 2018 were collected and pre-processed, comprising 12 million English tokens and 10 million Chinese tokens. Randomly selected press releases were translated using an online MT system for general texts. This was done to study the features and issues of out-of-domain English/Chinese MT of government press releases. The MT results were compared with the official translation, and translation errors were analyzed. MT issues such as omission, mistranslated proper nouns, mistranslated technical terms, word selection, and word order/sentence structure were identified.

Objective 2: Proposing an integrated approach to E-C NMT for government press releases. This involved the following tasks: (1) building a translation memory (TM) based on government press releases; (2) developing Chinese and English word embeddings using government documents; (3) designing attention-based encoding-decoding neural networks for translation; (4) training the networks using out-of-domain, partially in-domain, and fully in-domain data; and (5) integrating the TM in (1) and the trained networks in (4). More specifically, for (1), the TM comprised various types of translation units, not only sentences but also common stock phrases, job titles, names of units, policies or initiatives, and other technical terms. Unlike translation memories for translators that mainly consist of bilingual sentences, this translation memory was more diverse. For (2), Chinese and English word embeddings were developed using government documents and general texts containing 604 million English tokens and 440 million Chinese tokens. For (3), we adopted a bidirectional recurrent neural network architecture with long-short term memory, consisting of an encoder and a decoder. The encoder converts tokens in the source language into an intermediate representation, which is then used by the decoder to generate tokens in the target language. The input tokens for both 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 for English and 59,440 for Chinese. For (4), we trained three models based on (3): Model 1 was trained on out-of-domain data comprising 24 million sentence pairs from general documents. It was then further trained on text data related to government and public affairs to build Model 2. Model 3 was built by training Model 2 on the bilingual sentences extracted from the press releases mentioned in Objective 1 above. In this regard, Model 1, Model 2, and Model 3 can be considered out-of-domain, partially in-domain, and in-domain translation models, respectively. Technical details are available on our project website. For (5), a module was designed to first pre-translate the input using examples from the translation memory in Task 1 and then send the pre-translated text to the neural machine translation model in (4), with demo systems developed with HTML, CSS and JS.

Objective 3: Evaluating the integrated approach. A test set was built for evaluation consisting of sentences randomly selected from bilingual press releases published in January 2019, containing 156,000 English tokens and 130,000 Chinese tokens. This test set was used to evaluate the out-of-domain NMT, partially in-domain NMT, and fully in-domain NMT, with or without the TM. The BLEU framework was adopted to evaluate the three platforms. Our integrated approach using the TM in combination with the fully in-domain NMT performed the best (for both recurrent and self-attention neural networks), with encouraging results that even outperformed Google Translate. The details of the evaluation results, together with technical information about the integrated approach and its models, are presented in the journal article “Where Neural Machine Translation and Translation Memories Meet: Domain Adaptation for the Translation of HKSAR Government Press Releases” (please refer to Part C of this completion report).

5.4 Summary of objectives addressed to date

Objectives *(as per 5.1/5.2 above)*	**Addressed** *(please tick)*	**Percentage Achieved** *(please estimate)*
1.Studying the major problems of out-of-domain E-C NMT of HK government press releases	✓	100%
2.Proposing an integrated approach to the E-C NMT of government press releases	✓	100%
3.Evaluating the integrated approach	✓	100%

6. Research Outcome

6.1 Major findings and research outcome

(Maximum 1 page; please make reference to Part C where necessary)

The major findings of this project are summarized below. For details, please refer to the journal article "Where Neural Machine Translation and Translation Memories Meet: Domain Adaptation for the Translation of HKSAR Government Press Releases" (see Part C) and the project website.

We identified major problems in out-of-domain English/Chinese neural machine translation of government press releases, including issues with word choices, sentence structure, terminology, names of persons and places, formatting, and writing style. Refer to Table 4 of the article for specific examples and explanations. Given the above, it would 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. Our approach comprises three steps: (1) pre-translation of the source text with a translation memory designated for machine translation, resulting in a partially translated source text comprising placeholders representing pre-translated expressions (see Table 6 of the article for sample entries of the translation memory); (2) machine translation of the partially translated text using a neural network progressively trained on out-of-domain, partially in-domain, and in-domain data; and (3) post-translation, which restores the target text by replacing the placeholders in the machine translation output with the corresponding target language expressions using the translation memory. See Section 4.1 of the paper for a full description of our integrated approach, and Figure 1 therein for an illustration of the process.

We created a test dataset to evaluate the results of neural machine translation using our models, with and without pre-translation. We included Google Translate for comparison, along with the adoption of the BLEU evaluation framework. As summarized in the table below, there are two key findings: (1) The use of a translation memory can boost the performance of Google and all three of our neural models (refer to Section 5.3 and Section 4.4 of the journal article for details of the three models); (2) Our specialized model (i.e. Model 3) outperformed less specific ones, including Model 1, Model 2, and Google Translate.

MT Model	**BLEU Score**	
	Without TM	**With TM**
Google Translate	23.73	**35.51**
Model 1 (Out-of-domain)	15.56	29.28
Model 2 (Partially in-domain)	18.09	31.23
Model 3 (In-domain)	**38.75**	**44.71**

For translation examples illustrating the performance of Google Translate and Model 3 with and without the use of the translation memory for pre-translation, refer to Tables 8 and 9 (in Sections 4.3 and 4.4) of the article. One advantage of our approach is its capability to enhance the accuracy of technical terminology in the target language and adopt an appropriate writing style that better aligns with that of press releases.

This study demonstrates that a combination of translation memories and additional training using partially in-domain and fully in-domain data could contribute to the customization of models for specialized translation of press releases. In particular, the integration of translation memory and machine translation through pre-translation offers a simple solution for adapting general neural machine translation to domain-specific tasks, without requiring fine-tuning of existing models.

6.2 Potential for further development of the research and the proposed course of action *(Maximum half a page)*

Our project demonstrates promising results in terms of adapting general machine translation systems for specialized translation tasks by integrating translation memories and machine translation. For future research, it would be helpful to explore ways to enhance this integration, including the matching mechanism, for example, by better utilizing internal representations of source language sentences in the form of sentence embeddings.

Expanding this approach to the translation of other types of government documents beyond press releases would also be desirable. To achieve this, we could build a larger translation memory that incorporates expressions and sentences from other sources and investigate whether we can further improve the quality of NMT of these documents by means of pre-translation. Additionally, we may need to consider how to design translation memories for pre-translation to optimize the overall matching rate for better translation performance. Key issues to consider might 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., proper noun phrases versus common noun phrases), and the use of better alignment methods to reduce the number of misaligned translation units, which had an impact on the overall translation quality in our experiment.

It is also noteworthy that neural network architectures such as the Transformer, which is the basis of the ChatGPT chatbot that went viral in recent months, have emerged. We tested the use of a Transformer model trained on both general and specialized data and integrated it with our translation memory, resulting in more encouraging initial results with an even higher BLEU score of 46.1 points. This indicates that our approach is applicable not only to recurrent neural networks but also to other architectures, and further exploration of this could enhance the quality of automatic translation. We could also study the robustness of our approach and conduct a comparative study of the performance of integrated translation in various settings, such as other language pairs and hyperparameters (e.g., beam size for decoding).

7. Layman's Summary

(Describe in layman's language the nature, significance and value of the research project, in no more than 200 words)

HKSAR Government Press Releases in both English and Chinese serve as an important communication channel between the government and the public. The publication of the bilingual press releases could benefit from translation technology, such as neural machine translation (NMT). Given that general NMT engines have issues translating government texts in terms of style and terminology, this project proposes "pre-translation with translation memories", together with the use of NMT trained on general and specialized documents, for English/Chinese press release translation. Unlike human translators' translation memories, ours complements NMT by pre-translating source text with length-ranked bilingual sentences and sub-sentential units before machine translation. Results show that our method outperforms the NMT-only baseline, suggesting translation memories provide a simple solution to adapting general NMT to domain-specific translation without fine-tuning models. Higher quality automatic translation from integrating translation memories into specialized NMT could reduce post-editing and improve bilingual press release production and news dissemination, helping citizens of different linguistic backgrounds better understand public policies. In addition, our work could serve as a starting point to develop specialized translation engines for other government documents (e.g., papers and reports) to further promote communication between the government and Hong Kong citizens and raise the city's profile internationally.

Part C: Research Output

8. Peer-Reviewed Journal Publication(s) Arising Directly From This Research Project

(Please attach a copy of the publication and/or the letter of acceptance if not yet submitted in the previous progress report(s). All listed publications must acknowledge RGC's funding support by quoting the specific grant reference.)

The Latest Status of Publications				**Author(s)** *(denote the correspond-ing author with an asterisk***)*	**Title and Journal / Book** *(with the volume, pages and other necessary publishing details specified)*	**Submitted to RGC** *(indicate the year ending of the relevant progress report)*	**Attached to this Report** *(Yes or No)*	**Acknowledged the Support of RGC** *(Yes or No)*	**Accessible from the Institutional Repository** *(Yes or No)*
Year of Publication	**Year of Acceptance** *(For paper accepted but not yet published)*	**Under Review**	**Under Preparation** *(optional)*						
2022				SIU Sai Cheong	"Where Neural Machine Translation and Translation Memories Meet: Domain Adaptation for the Translation of HKSAR Government Press Releases". In: *International Journal of Techno-Humanities*, 1, pp. 45-66.	No	Yes	Yes Appendix I	Yes

9. Recognized International Conference(s) In Which Paper(s) Related To This Research Project Was / Were Delivered

(Please attach a copy of each conference abstract)

Month / Year / Place	**Title**	**Conference Name**	**Submitted to RGC** *(indicate the year ending of the relevant progress report)*	**Attached to this Report** *(Yes or No)*	**Acknowledged the Support of RGC** *(Yes or No)*	**Accessible from the Institutional Repository** *(Yes or No)*
Jan 2022 HK	"Where Neural Machine Translation and Translation Memories Meet: Domain Adaptation for the Translation of HKSAR Government Press Releases"	International Conference on "Building a Techno-Humanities Culture in Hong Kong"	No	Yes	Yes Appendix II	Yes

10. Whether Research Experience And New Knowledge Has Been Transferred / Has Contributed To Teaching And Learning

(Please elaborate)

The research findings have been incorporated into the PI's language technology modules for undergraduate and postgraduate students: TRA3105 Computer and Business Translation 1 (a core undergraduate module) and TRA6001 Translation Technology: Knowledge and Skills (a core postgraduate module). The modules aim to teach students how to use computers effectively in translation. Students have hands-on experience in using the state-of-the-art computer / computer-aided systems (e.g., electronic dictionaries, corpora, concordancers, machine translation systems, translation memories and terminology databases). Our discussion of the use of SNMT models, the possibility of integrating TM into NMT, and the importance of using specialized data for training NMT models has helped to increase students' awareness of the limitations of free online translation engines that are often used for general translation. It has also emphasized the significance of domain adaptation to achieve better results in computer-aided specialized translation projects.

The recruitment of part-time/student research assistants contributed to advancing research on language technology. The participants gained hands-on experience that familiarized them with data preprocessing and evaluation methods for translation technologies.

11. Student(s) Trained

(Please attach a copy of the title page of the thesis)

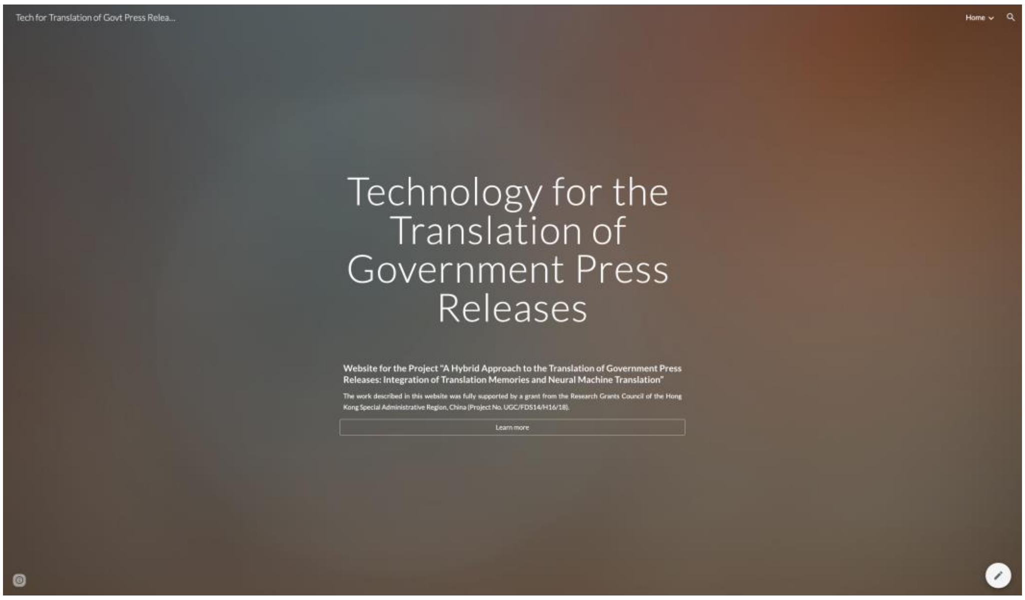
Name	Degree Registered for	Date of Registration	Date of Thesis Submission / Graduation
N/A			

12. Other Impact

(e.g. award of patents or prizes, collaboration with other research institutions, technology transfer, teaching enhancement, etc.)

In addition to the research deliverables outlined in the previous sections, we have also developed the following portals to showcase our project and provide additional technical details about our work:

"Technology for the Translation of Government Press Releases": This gives an overview of our project and technical information about the design and training of our models, with sample files illustrating our TM and datasets.



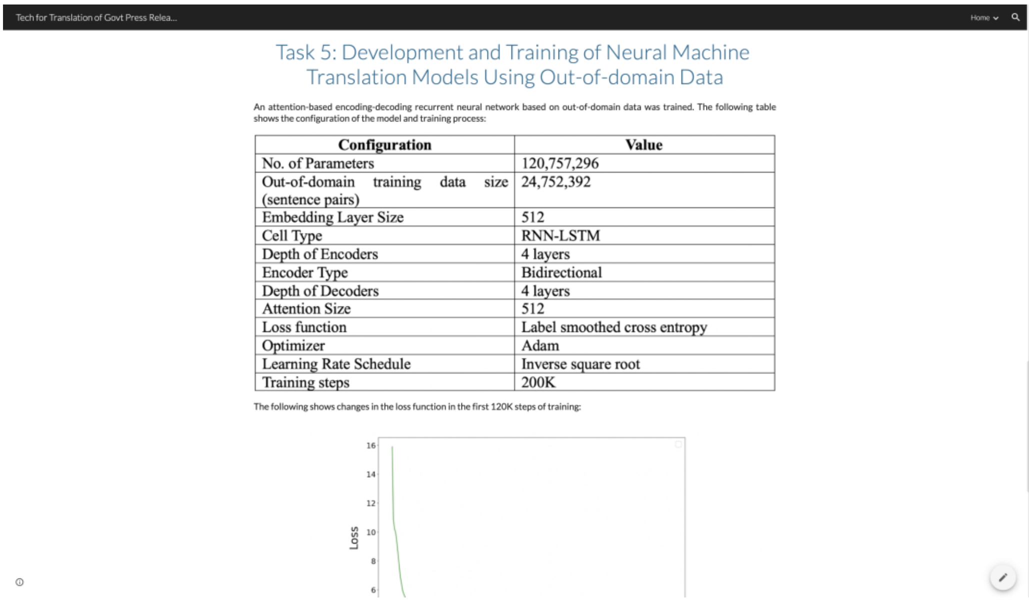
Technology for the Translation of Government Press Releases

Website for the Project "A Hybrid Approach to the Translation of Government Press Releases: Integration of Translation Memories and Neural Machine Translation"

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

Learn more

Main Page of Project Portal



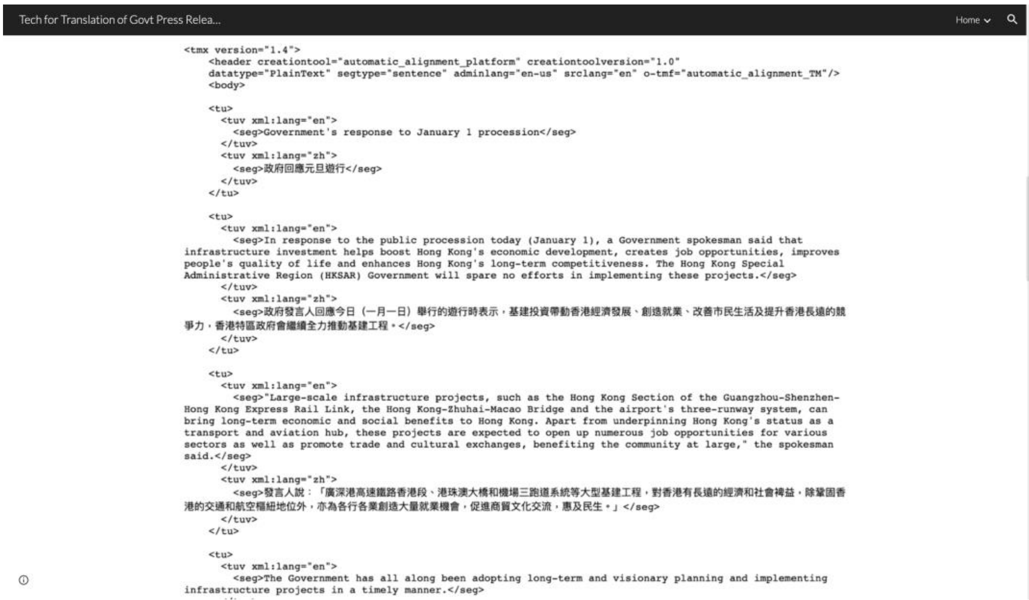
Task 5: Development and Training of Neural Machine Translation Models Using Out-of-domain Data

An attention-based encoding-decoding recurrent neural network based on out-of-domain data was trained. The following table shows the configuration of the model and training process:

Configuration	Value
No. of Parameters	120,757,296
Out-of-domain training data size (sentence pairs)	24,752,392
Embedding Layer Size	512
Cell Type	RNN-LSTM
Depth of Encoders	4 layers
Encoder Type	Bidirectional
Depth of Decoders	4 layers
Attention Size	512
Loss function	Label smoothed cross entropy
Optimizer	Adam
Learning Rate Schedule	Inverse square root
Training steps	200K

The following shows changes in the loss function in the first 120K steps of training:

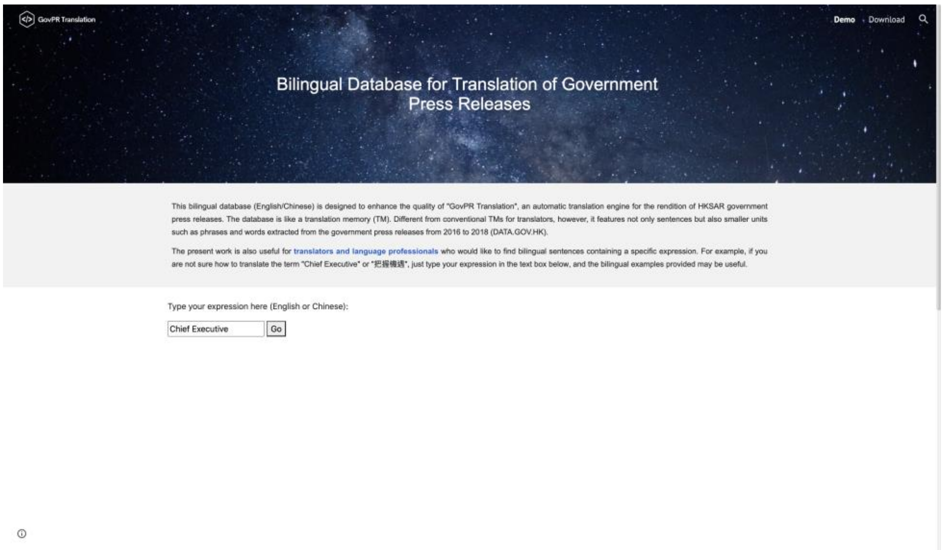
Technical details of Model Design and Training



```
<tmx version="1.4">
    <header creationtool="automatic_alignment_platform" creationtoolversion="1.0"
    datatype="PlainText" segtype="sentence" adminlang="en-us" srclang="en" o-tmf="automatic_alignment_TM"/>
    <body>

    <tu>
      <tuv xml:lang="en">
        <seg>Government's response to January 1 procession</seg>
      </tuv>
      <tuv xml:lang="zh">
        <seg>政府回應元旦遊行</seg>
      </tuv>
    </tu>

    <tu>
      <tuv xml:lang="en">
        <seg>In response to the public procession today (January 1), a Government spokesman said that
infrastructure investment helps boost Hong Kong's economic development, creates job opportunities, improves
people's quality of life and enhances Hong Kong's long-term competitiveness. The Hong Kong Special
Administrative Region (HKSAR) Government will spare no efforts in implementing these projects.</seg>
      </tuv>
      <tuv xml:lang="zh">
        <seg>政府發言人回應今日（一月一日）舉行的遊行時表示，基建投資帶動香港經濟發展、創造就業、改善市民生活及提升香港長遠的競
爭力，香港特區政府會繼續全力推動基建工程。</seg>
      </tuv>
    </tu>

    <tu>
      <tuv xml:lang="en">
        <seg>"Large-scale infrastructure projects, such as the Hong Kong Section of the Guangzhou-Shenzhen-
Hong Kong Express Rail Link, the Hong Kong-Zhuhai-Macao Bridge and the airport's three-runway system, can
bring long-term economic and social benefits to Hong Kong. Apart from underpinning Hong Kong's status as a
transport and aviation hub, these projects are expected to open up numerous job opportunities for various
sectors as well as promote trade and cultural exchanges, benefiting the community at large," the spokesman
said.</seg>
      </tuv>
      <tuv xml:lang="zh">
        <seg>發言人說：「廣深港高速鐵路香港段、港珠澳大橋和機場三跑道系統等大型基建工程，對香港有長遠的經濟和社會裨益，除鞏固香
港的交通和航空樞紐地位外，亦為各行各業創造大量就業機會，促進商貿文化交流，惠及民生。」</seg>
      </tuv>
    </tu>

    <tu>
      <tuv xml:lang="en">
        <seg>The Government has all along been adopting long-term and visionary planning and implementing
infrastructure projects in a timely manner.</seg>
```

Sample TM Entries

"Bilingual Database for Translation of Government Press Releases": We built a search platform using a subset of the TM for pre-translation because the TM could not only complement NMT but also facilitate professional translators' work by retrieving bilingual reference sentences or segments matching a query expression provided by users.



Bilingual Database for Translation of Government Press Releases

This bilingual database (English/Chinese) is designed to enhance the quality of "GovPR Translation", an automatic translation engine for the rendition of HKSAR government press releases. The database is like a translation memory (TM). Different from conventional TMs for translators, however, it features not only sentences but also smaller units such as phrases and words extracted from the government press releases from 2016 to 2018 (DATA.GOV.HK).

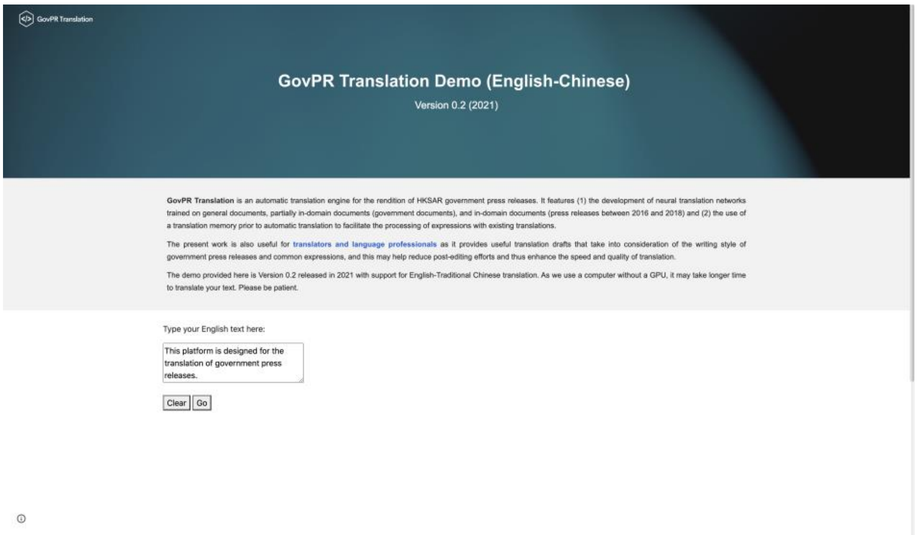
The present work is also useful for translators and language professionals who would like to find bilingual sentences containing a specific expression. For example, if you are not sure how to translate the term "Chief Executive" or "把握機遇", just type your expression in the text box below, and the bilingual examples provided may be useful.

Type your expression here (English or Chinese):

Chief Executive

Bilingual Database for Translation of Government Press Releases

GovPR Translation Demo Systems: Demo systems utilizing our SNMT models trained on general, partially in-domain, and/or fully in-domain data were built, generating translation results that can be compared against those from other models for evaluation.



GovPR Translation Demo (English-Chinese)

Version 0.2 (2021)

GovPR Translation is an automatic translation engine for the rendition of HKSAR government press releases. It features (1) the development of neural translation networks trained on general documents, partially in-domain documents (government documents), and in-domain documents (press releases between 2016 and 2018) and (2) the use of a translation memory prior to automatic translation to facilitate the processing of expressions with existing translations.

The present work is also useful for translators and language professionals as it provides useful translation drafts that take into consideration of the writing style of government press releases and common expressions, and this may help reduce post-editing efforts and thus enhance the speed and quality of translation.

The demo provided here is Version 0.2 released in 2021 with support for English-Traditional Chinese translation. As we use a computer without a GPU, it may take longer time to translate your text. Please be patient.

Type your English text here:

This platform is designed for the translation of government press releases.

An Example of GovPR Translation Demo Systems

To disseminate our findings more widely, we shared our preliminary findings with a government department during the project period. We were also approached by companies in the translation industry for the exploration of the use of the TM and/or application of the integrated method.

13. Statistics on Research Outputs

	Peer-reviewed Journal Publications	Conference Papers	Scholarly Books, Monographs and Chapters	Patents Awarded	Other Research Outputs (please specify)	
No. of outputs arising directly from this research project	1	1			Type	No.
					Project portal	1
					Dataset search platform	1
					Demo systems	2
					Online article	1

14. Public Access Of Completion Report

(Please specify the information, if any, that cannot be provided for public access and give the reasons.)

Information that Cannot Be Provided for Public Access	**Reasons**
N/A	