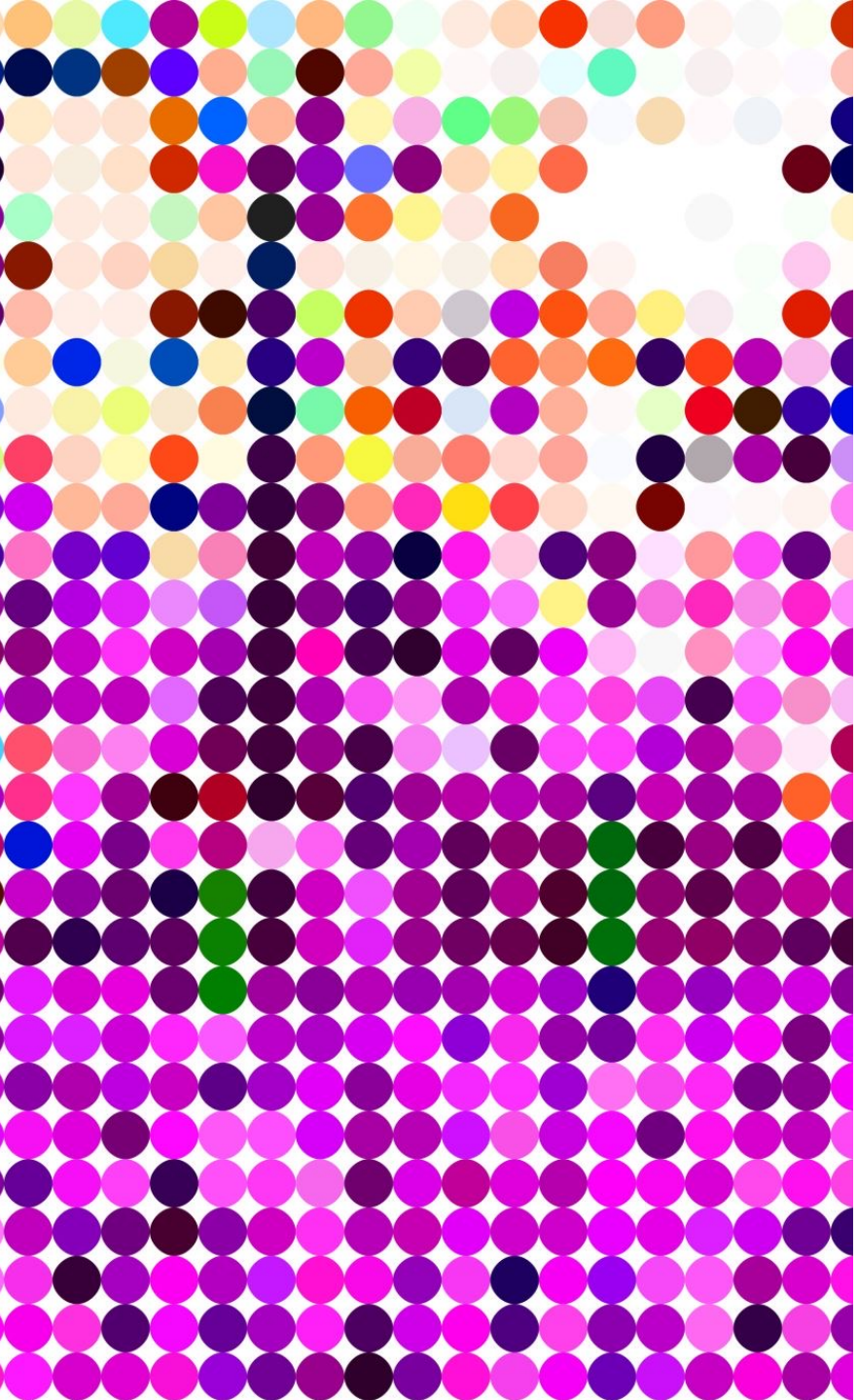


Where Neural Machine Translation and Translation Memories Meet

Domain Adaptation for the Translation of HKSAR Government Press Releases

SIU Sai Cheong

The work described in this presentation 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).

Outline

1. An Overview of HKSAR Government Press Releases
2. Machine Translation of Government Press Releases
3. Our Approach: Translation Memory + Machine Translation
4. Our Experiments
5. The Way Forward

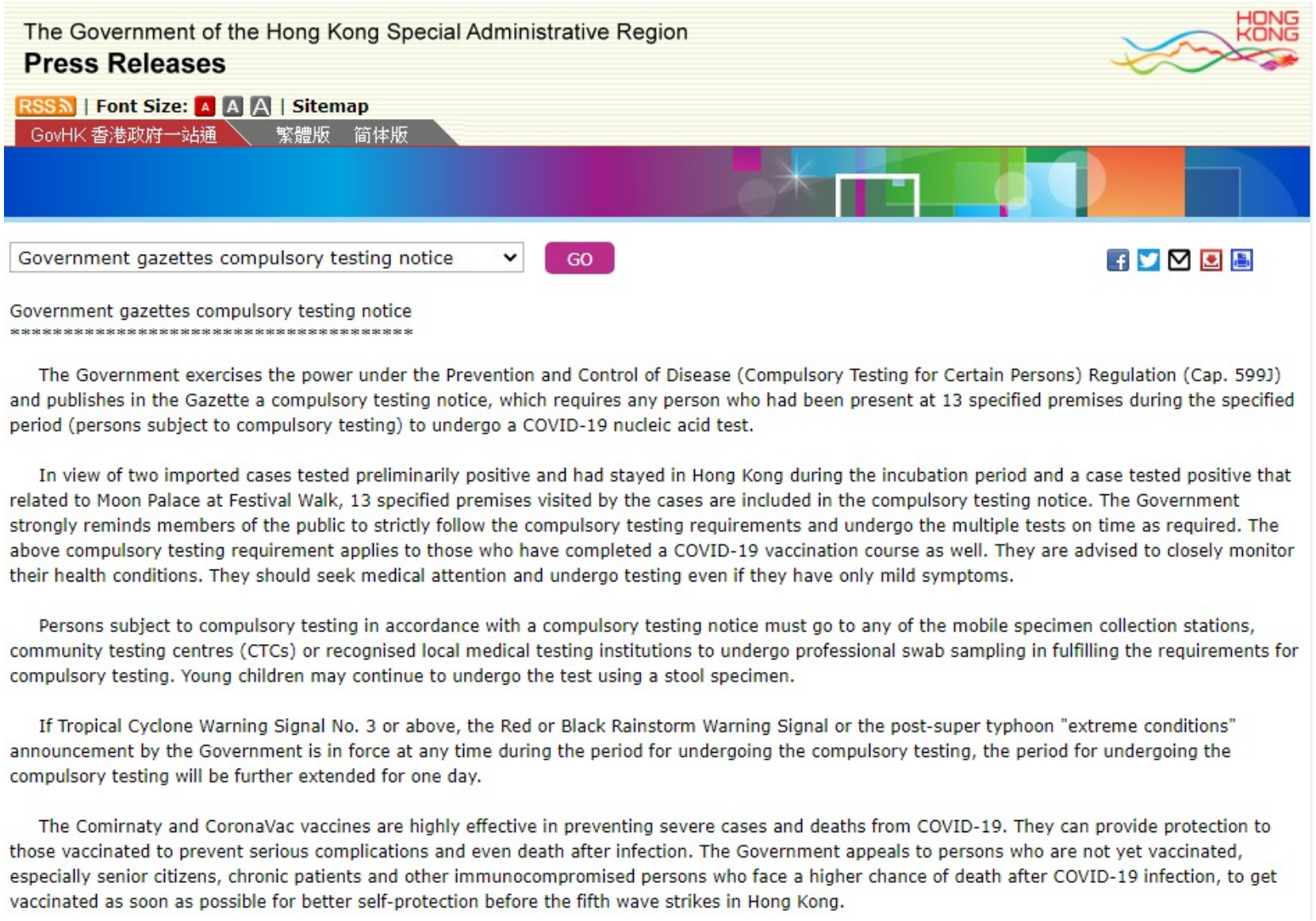
1. An Overview of HKSAR Government Press Releases

Hong Kong Government Press Releases

- An important communication channel between the government and the public
- Information provided by different bureaus and departments

HKSAR Government Press Releases

- Available online: https://www.info.gov.hk/gia/general/today.htm
- Three versions: English, Traditional Chinese, and Simplified Chinese



The Government of the Hong Kong Special Administrative Region

Press Releases

RSS | Font Size: A A A | Sitemap

GovHK 香港政府一站通 繁體版 简体版

Government gazettes compulsory testing notice GO

Government gazettes compulsory testing notice

****

The Government exercises the power under the Prevention and Control of Disease (Compulsory Testing for Certain Persons) Regulation (Cap. 599J) and publishes in the Gazette a compulsory testing notice, which requires any person who had been present at 13 specified premises during the specified period (persons subject to compulsory testing) to undergo a COVID-19 nucleic acid test.

In view of two imported cases tested preliminarily positive and had stayed in Hong Kong during the incubation period and a case tested positive that related to Moon Palace at Festival Walk, 13 specified premises visited by the cases are included in the compulsory testing notice. The Government strongly reminds members of the public to strictly follow the compulsory testing requirements and undergo the multiple tests on time as required. The above compulsory testing requirement applies to those who have completed a COVID-19 vaccination course as well. They are advised to closely monitor their health conditions. They should seek medical attention and undergo testing even if they have only mild symptoms.

Persons subject to compulsory testing in accordance with a compulsory testing notice must go to any of the mobile specimen collection stations, community testing centres (CTCs) or recognised local medical testing institutions to undergo professional swab sampling in fulfilling the requirements for compulsory testing. Young children may continue to undergo the test using a stool specimen.

If Tropical Cyclone Warning Signal No. 3 or above, the Red or Black Rainstorm Warning Signal or the post-super typhoon "extreme conditions" announcement by the Government is in force at any time during the period for undergoing the compulsory testing, the period for undergoing the compulsory testing will be further extended for one day.

The Comirnaty and CoronaVac vaccines are highly effective in preventing severe cases and deaths from COVID-19. They can provide protection to those vaccinated to prevent serious complications and even death after infection. The Government appeals to persons who are not yet vaccinated, especially senior citizens, chronic patients and other immunocompromised persons who face a higher chance of death after COVID-19 infection, to get vaccinated as soon as possible for better self-protection before the fifth wave strikes in Hong Kong.

(Source: https://www.info.gov.hk/gia/general/202201/02/P2022010200040.htm)

香港特別行政區政府
新聞公報

RSS | 字型大小： A A A | 網頁指南

GovHK 香港政府一站通 简体版 English

政府就強制檢測公告刊憲 去

政府就強制檢測公告刊憲

＊＊＊＊＊＊＊＊＊＊

政府引用《預防及控制疾病（對若干人士強制檢測）規例》（《規例》）（第599J章），就《規例》下的強制檢測公告刊憲，要求於指定期間曾身處13個指明地方的人士（下稱「受檢人士」）接受2019冠狀病毒病核酸檢測。

因應兩宗於潛伏期間曾於香港逗留的初步陽性檢測輸入個案，以及一宗與又一城望月樓相關的陽性檢測個案，有13個指明地方被納入強制檢測公告。政府強烈提醒市民必須嚴格遵守強制檢測要求，按時完成須多次進行的強制檢測，而即使已經接種2019冠狀病毒病疫苗亦必須接受上述強制檢測。他們應同時密切留意身體狀況，如有任何輕微病徵必須立即求醫並接受檢測。

因應強制檢測公告而須進行強制檢測的人士必須到流動採樣站、社區檢測中心或認可本地醫療檢測機構進行專業拭子採樣檢測，以符合強制檢測的要求，而幼童則可繼續以糞便樣本進行檢測。

若在進行強制檢測期間的任何時間，三號或以上熱帶氣旋警告信號、紅／黑色暴雨警告信號或政府公布的「超強颱風後的極端情況」生效，進行強制檢測的期限將會延長一天。

復必泰和克爾來福疫苗對於預防2019冠狀病毒病重症和死亡情況高度有效，能為接種人士提供有效保護，免於感染後併發重症甚至死亡。政府呼籲仍未接種疫苗的市民，特別是感染新冠病毒後死亡風險極高的長者、長期病患者及其他免疫力較弱人士，為自己健康着想應在本港第五波疫情來臨前盡快接種疫苗。

(Source: https://www.info.gov.hk/gia/general/202201/02/P2022010200039.htm)

Examples

Announcements

Statements

Speeches

Event promotional materials

Themes


Immigration
Labour
Urban Planning
Finance
Law and Order
Housing and Land
Recreation
Public Health

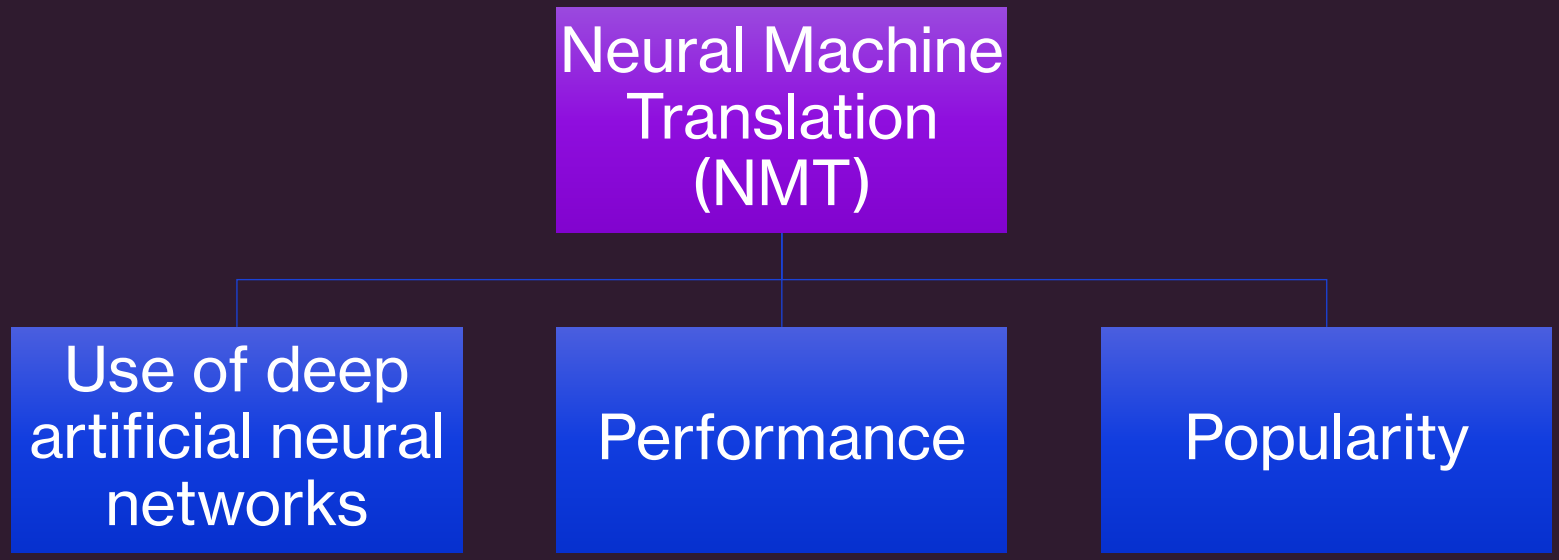
Bilingual Press Releases

- **English:** 4 million tokens/year
- **Chinese:** 6 million characters/year

(Based on the bilingual press releases (2016-2018) collected automatically)

2. Machine Translation of Government Press Releases

Machine Translation of Government Press Releases



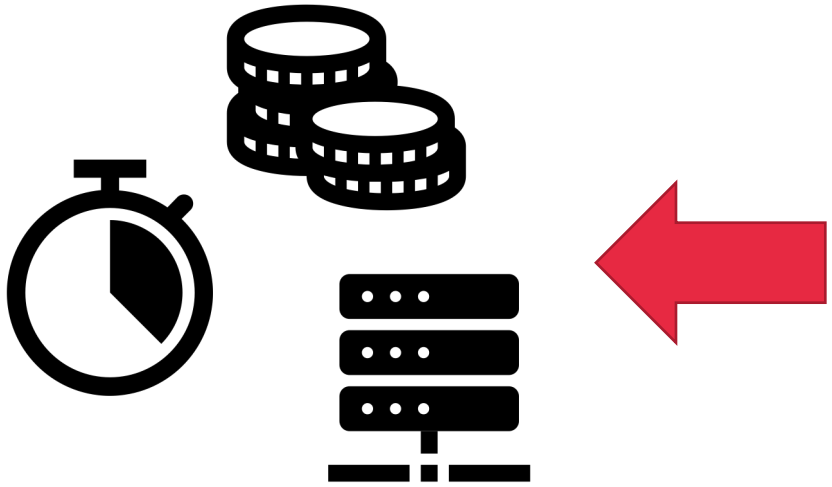
Issues

- Word Choices
- Sentence Structure
- Terminology
- Proper Nouns
- Formatting
- Writing Style

Domain Adaptation: General MT→ Specialized MT?

3. Our Approach: TM + MT

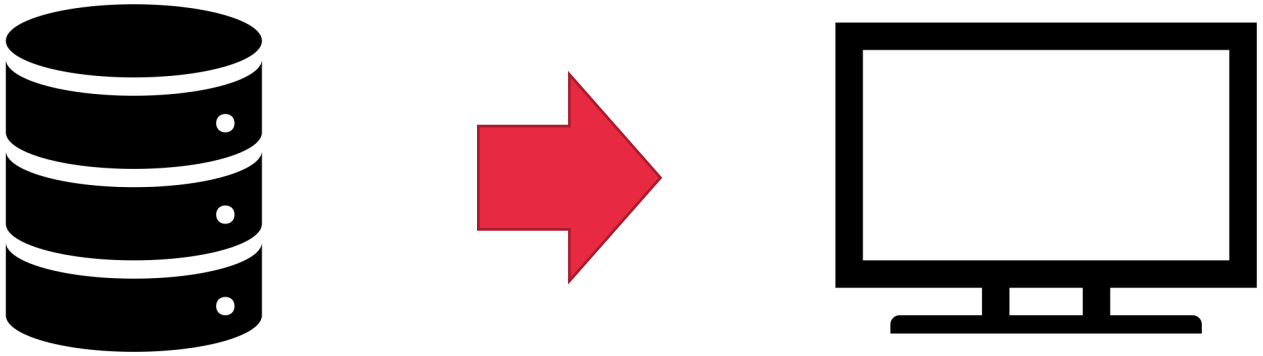
A Common Approach



We can use domain-specific data.

However, this often requires re-training or additional training of the model.

Our Approach



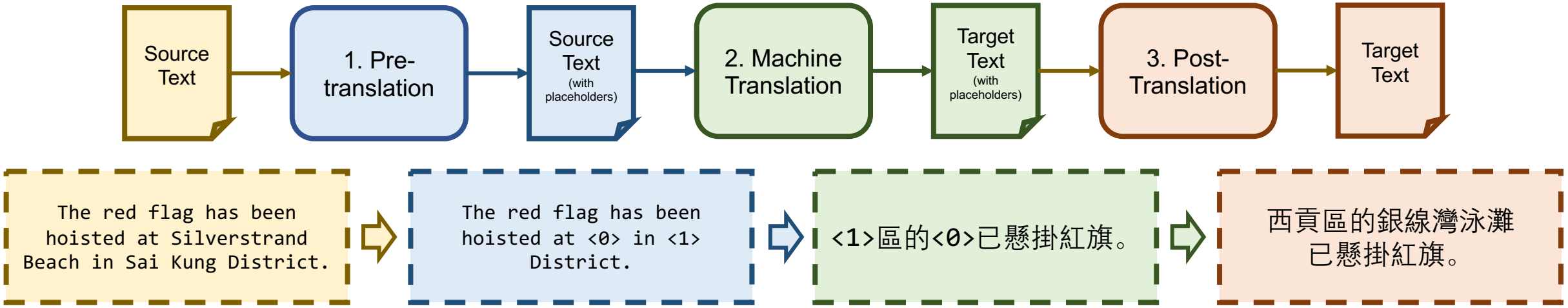
Translation Memory

- Designed for MT (rather than HT)
- Containing phrases and sentences
- Facilitating the translation of domain-specific content

Machine Translation

- No re-training or additional training of the model

Our Approach: A Closer Look



Our Sample TM

- Bilingual government press releases (2016-2018)
- 408,000 sentence pairs (after automatic alignment)
- 270,000 translation units (after removal of segments with low interlingual similarity)

TM 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，政府有決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

TM Examples

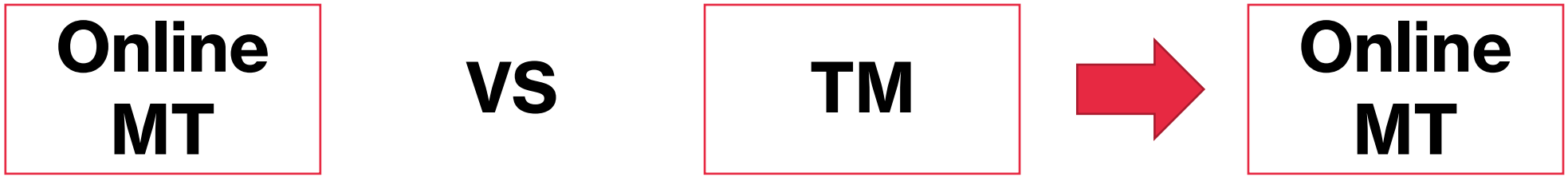
Names of units	Architectural Services Department 建築署 Business Environment Council 商界環保協會 Sha Tau Kok District Rural Committee 沙頭角區鄉事委員會
Names of policies or initiatives	Qualifying Debt Instrument Scheme 合資格債務票據計劃 Redevelopment of Tai Hang Sai Estate 大坑西邨重建計劃 SME Financing Guarantee Scheme 中小企融資擔保計劃
Other 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

4. Our Experiments

Experiments

- Experiment 1: A Popular Online Engine
- Experiments 2 and 3: Our Own NMT Models

Experiment 1



Our Test Dataset

- Press releases in January 2019
- 5,000 bilingual sentences
- 156,473 English tokens and 130,646 Chinese tokens

Experiment 1: Results

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

$$BLEU = 100 \cdot \min(1, e^{1-\frac{r}{c}}) \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Example 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。

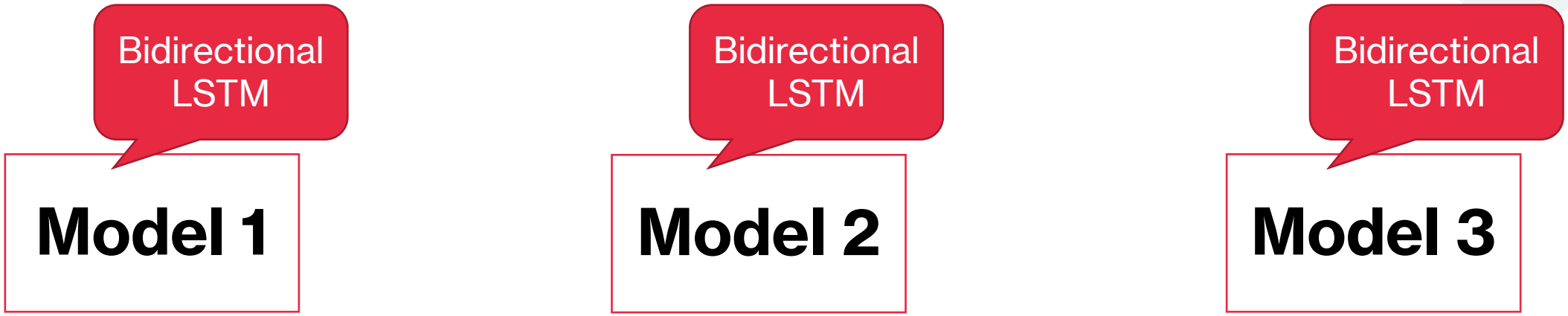
Example 2

- **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。

Example 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。

Experiment 2

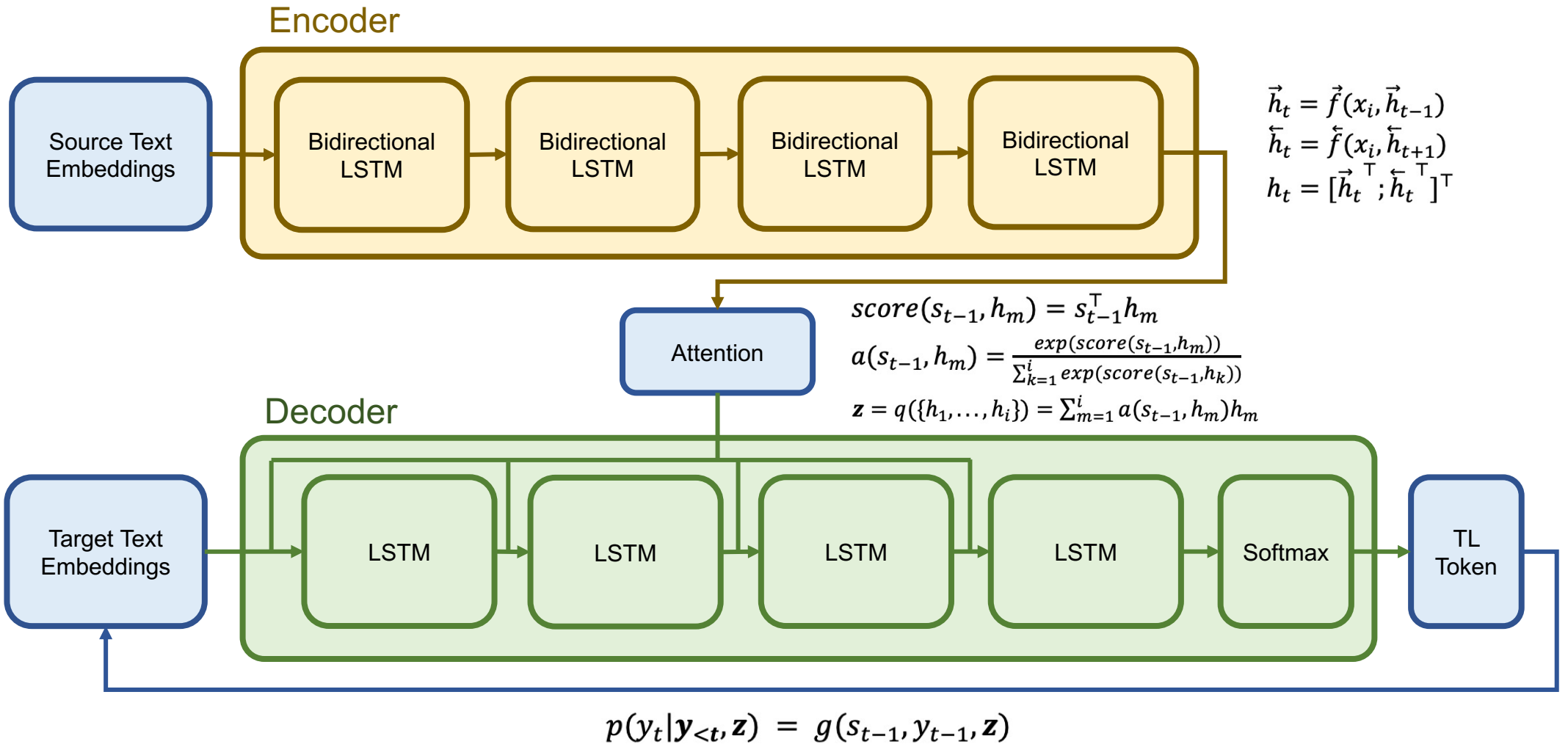


With/out TM

Model Architecture

- 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.

Our Bidirectional LSTM Model



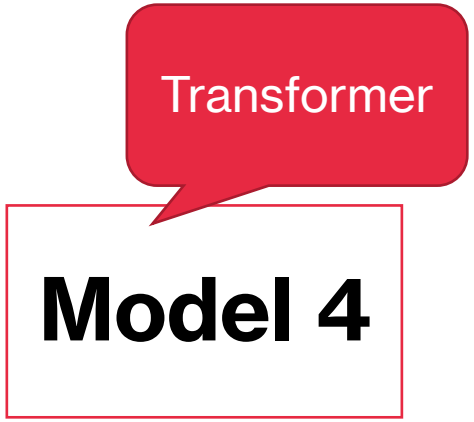
Experiment 2: Results

	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

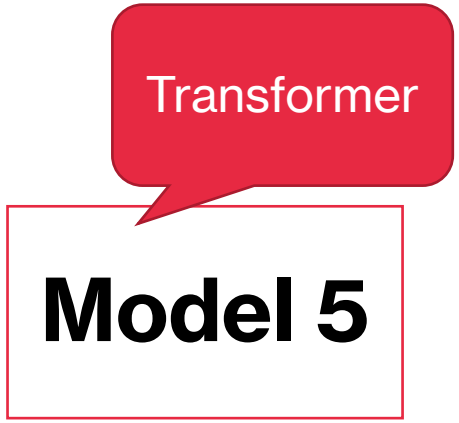
Experiment 3

- 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.

Experiment 3



General documents



General documents +
Government documents

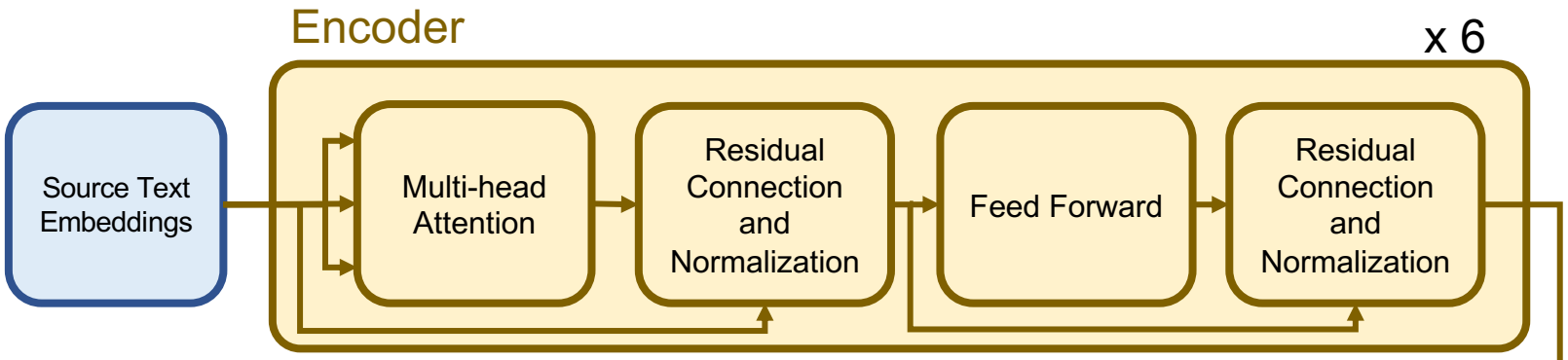
General documents +
Government documents +
Government Press Releases

With/out TM

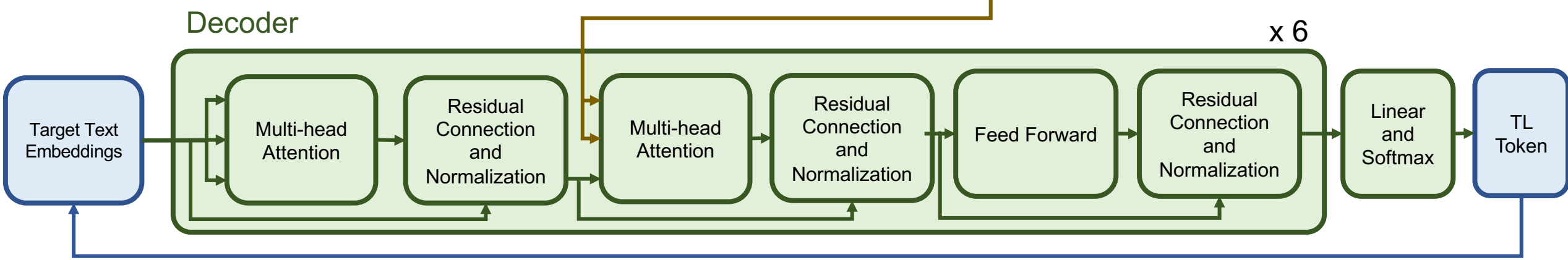
Model Architecture

- The configuration of each of the models 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.

Our Transformer Model



$$scoremat(Q,K) = softmax(QK^{\top} \cdot scale)$$
$$attention(Q,K,V) = (scoremat(Q,K))V$$
$$head_h = attention(W_O Q, W_K K, W_V V) = attention(Q_h, K_h, V_h)$$
$$multihead(Q,K,V) = W_{multihead}[head_1; head_2; \ldots; head_H]$$
$$X_{intermediate} = layernorm(multihead(X_{input}, X_{input}, X_{input}) + X_{input})$$
$$X_{output} = layernorm(F_{x2}(ReLU(F_{x1}(X_{intermediate})) + X_{intermediate})$$



$$Y_{intermediate(1)} = layernorm(multihead(Y_{input}, Y_{input}, Y_{input}) + Y_{input})$$
$$Y_{intermediate(2)} = layernorm(multihead(Y_{intermediate(1)}, X_{output(encoder)}, X_{output(encoder)}) + Y_{intermediate(1)})$$
$$Y_{output} = layernorm(F_{y2}(ReLU(F_{y1}(Y_{intermediate(2)}))) + Y_{intermediate(2)})$$

Experiment 3: Results

	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

Future Research

- Design of translation memories for pre-translation
- Matching and restoration methods
- Ways to represent pre-translated expressions in the source text for MT
- Applicability to other NMT models or language pairs

References

- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- 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*.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

References

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