SG Translate Together - Uplifting Singapore's translation standards with the community through technology

Lee Siew Li LEE_Siew_Li@mci.gov.sg
Adeline Sim Adeline_SIM@mci.gov.sg
Gowri Kanagarajah Gowri_KANAGARAJAH@mci.gov.sg
Siti Amirah Siti_AMIRAH@mci.gov.sg
Foo Yong Xiang FOO_Yong_Xiang@mci.gov.sg
Gayathri Ayathorai Gayathri_AYATHORAI@mci.gov.sg
Sarina Mohamed Rasol Sarina_MOHAMED_RASOL@mci.gov.sg
Translation Department, Ministry of Communications and Information (MCI), Singapore
Aw Ai Ti aaiti@i2r.a-star.edu.sg
Wu Kui wuk@i2r.a-star.edu.sg
Zheng Weihua zhengw@i2r.a-star.edu.sg
Ding Yang ding_yang@i2r.a-star.edu.sg
Tarun Kumar Vangani vangani_tarun_kumar@i2r.a-star.edu.sg
Nabilah Binte Md Johan nabilah@i2r.a-star.edu.sg
Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (A*STAR), Singapore

Abstract

The Singapore's Ministry of Communications and Information (MCI) has officially launched the SG Translate Together (SGTT) web portal on 27 June 2022, with the aim of partnering its citizens to improve translation standards in Singapore.

This web portal houses the Singapore Government's first neural machine translation (MT) engine, known as SG Translate, which was jointly developed by MCI and the Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (A*STAR). Adapted using localised translation data, SG Translate is able to generate translations that are attuned to Singapore's context and supports Singapore's four (4) official languages – English (Singapore), Chinese (Singapore), Bahasa Melayu (Singapore) and Tamil (Singapore). Upon completion of development, MCI allowed all Government agencies to use SG Translate for their daily operations.

This presentation will briefly cover the methodologies adopted and showcase SG Translate's capability to translate content involving local culture, everyday life and government policies and schemes. This presentation will also showcase MCI's sustainable approach for the continual training of the SG Translate MT engine through citizenry participation.

1. Introduction

SG Translate is a customised Neural Machine Translation (MT) engine jointly developed by the Ministry of Communications and Information (MCI), Singapore and A*STAR's Institute for Infocomm Research (I^2R), Singapore. It was launched in July 2019 to Singapore's public service sectors via the Government intranet.

As SG Translate is trained with localised data such as government communications materials, it is able to produce first-cut translations that are suited to Singapore's context in Singapore's four (4) official languages – English (Singapore), Chinese (Singapore), Bahasa Melayu (Singapore) and Tamil (Singapore). The engine's performance has indicated its capability to translate localised content, especially local terms related to the Singapore Government's policies and operations, as well as those related to local culture, such as the names of local delicacies.

SG Translate was originally developed to help public officers in Singapore manage the increasing demand for government communications materials to be made available in all four official languages. The localised translations generated by SG Translate serve as drafts and reduce the need for translators to start from scratch, thereby improving work productivity and efficiency. In addition, the time saved can be channeled into post-editing and vetting to ensure that the translations are properly nuanced and are able to accurately convey the information to citizens. Response from public officers to the initial roll-out was positive and encouraging. Many lauded the quality of the machine's first-cut translation, and found the translation generated by SG Translate to be more accurate and suitable for the local audience than those produced by other translation engines in the market. During the height of the COVID-19 pandemic, MCI's Translation Department (MCI-TD) officers used SG Translate to generate first-cut translations before refining the text further. This shortened the time taken to translate relevant materials into the other three official languages, and allowed Singaporeans to receive timely updates on evolving situations.

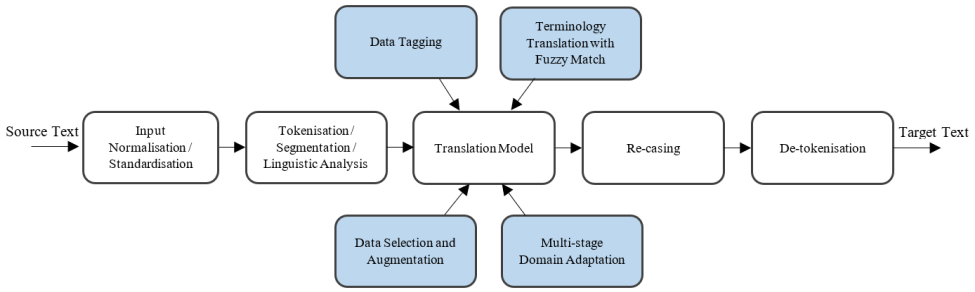
After the successful delivery of SG Translate, the idea of a collaborative web portal was mooted in late 2019 to extend SG Translate to the public, as well as to create more opportunities to work with people in their translation journeys. This fosters partnership and strengthens communications with all segments of society. This led to the establishment of SG Translate Together (SGTT), an online web portal that allows members of the public to use the SG Translate MT engine on the internet.

The purpose of this paper is to highlight the technical aspects behind building SG Translate and the community engagement aspect of SG Translate Together. For the technical perspective, the paper seeks to showcase key methodology of how the translation engine was developed to cater specifically to Singapore's linguistic use. Additionally, the paper will also cover how the SG Translate Together Web Portal harnesses the benefits of community engagement by involving members of the public to contribute training data to SG Translate. Through this process of citizenry engagement and partnership, the Singapore Government hopes to co-create a better MT engine that belongs to Singaporeans and for all to use.

2. SG Translate Neural Machine Translation Engine

In recent years, Neural Machine Translation (NMT) has made remarkable progress in the field of natural language processing (Wu et al., 2016; Gehring et al., 2017; Vaswani et al., 2017; Chen et al., 2018). However, when the translation model is limited by the amount of data, developing an engine with good translation performance in a specific domain is a challenge. Whether the language pair is low- or high- resource, domain-specific training

data are often rare, and it becomes a problem for data-hungry neural networks. This paper focuses on the development of SG Translate bi-directional translation engines for three language pairs (English-Chinese, English-Tamil, English-Malay) using NMT technology and the exploration of multi-stage domain adaptation and data augmentation methods to advance the engine's translation performance. Additionally, the placeholder-based fuzzy match mechanism for local terminology translation and data tagging strategy were employed to emphasise translation learning of Singapore-contextualised content. Experiments show that the present translation system generates more localised translations in Tamil, Malay and Chinese to English translations and vice versa as compared to commercially available solutions.



* The blue boxes are the methodologies applied to the translation model.

Figure 1. Overview of the translation system

Our translation system adopts standard sequence-to-sequence Transformer architecture (Vaswani et al., 2017). Figure 1 shows an overview of the translation system.

3. Methodology

SG Translate consists of three language pairs (English-Chinese, English-Malay, English-Tamil), of which English-Malay and English-Tamil are low-resource language pairs where both out-domain and domain-specific data are limited. This paper proposes strategies including data selection and augmentation, fuzzy terminology match, multi-stage domain adaptation and data tagging, which are proven to be effective for both low- and high-resource language pairs. The eventual product produces translation that suits the Singapore context, and is complemented by accurate translation of unique terminologies.

3.1. Data Selection and Augmentation

As deep learning requires large volumes of training data, the back translation (BT) method (Sennrich et al., 2015) was adopted to augment the training data. Back translation is proven to be effective under both low- and high-resource settings by exploiting monolingual data which is abundant and easily obtained. Firstly, an existing bilingual parallel corpus is used to construct a target-to-source NMT model, which is then used to translate target monolingual data into the source language. In doing so, a certain amount of pseudo bilingual parallel data is generated. This pseudo bilingual parallel data is used to augment the original bilingual dataset to train a new source-to-target model. For the selection of monolingual corpus to be used for back

translation, the selection strategy targeting difficult words (Fadaee et al., 2018) is adopted. Word frequency counting is adopted for all words in the original training corpus. Sentences containing low-frequency words and out-of-vocabulary (OOV) words are allocated higher priority for selection. Through this method, the diversity of the training set is increased and the ability of the source-to-target translation model to process low-frequency words and OOV words is enhanced.

3.2. Terminology Translation with Fuzzy Match

Data augmentation helps the engine to gain translation knowledge of common words. However, it is challenging to source for a large amount of training data belonging to the domains of our interest. Therefore, the engine could not provide accurate translation for domain-specific terminologies as their occurrences were low and the model could not pick up the necessary translation knowledge. We then propose to leverage a terminology dictionary with placeholders to address this problem.

A terminology dictionary containing a list of terms and their corresponding reference translations, is first created. Terms in the terminology dictionary would need to be equivalent, specific and unique. The engines utilise the terminology dictionary and a placeholder-based mechanism to translate terms such as person and entity names more accurately. Since the NMT model is trained to translate placeholder tokens into themselves, the placeholder tokens in the translation output are substituted with pre-specified translations in the terminology dictionary.

This placeholder replacement works well when the sentence contains a small number of terms that need to be replaced (e.g. one to three) and there is sufficient contextual information left other than the placeholder tokens in the replaced sentence. However, when the number of replaced terms in a sentence increases (e.g. more than three) or the sentence after placeholder replacement has little contextual information left other than the placeholder tokens, translation errors may occur. Therefore, the information of the replaced words is kept together with the corresponding placeholder tokens in the source sentence to enable the translation model to learn the context of the source terms. (Wang T., 2019).

As shown in Figure 2, in the source sentence of the training data, both the replaced words and the placeholder tokens are kept in the replaced sentence, and separators such as "<s>, <m>, < e>" are introduced to identify the boundaries of the replaced part; in the replaced target sentence, only the placeholder and the boundary identifier exist, which are consistent with those in the source sentence.

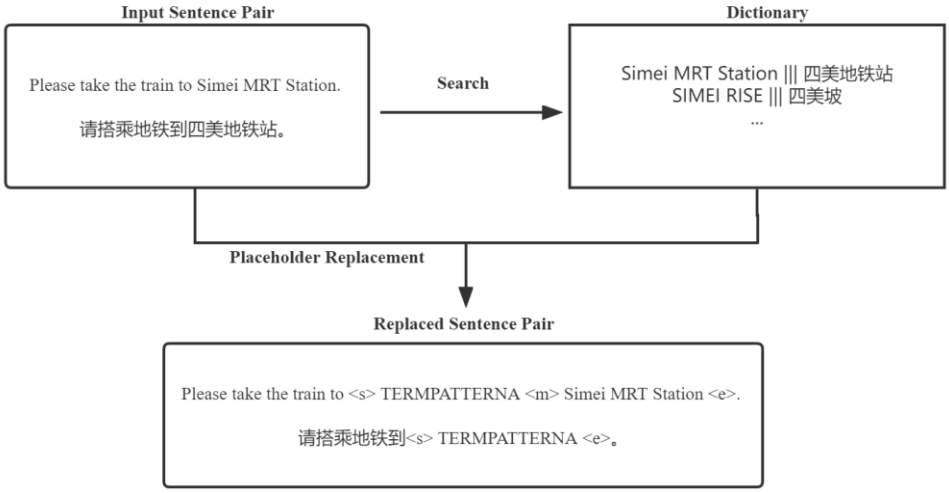


Figure 2. Example of retaining information of replaced terms in the source sentence

For terminology matching, if the match is only limited to an exact match between a term in the input sentence and the term in the terminology dictionary where only one form (usually root form) exists, variants of a term may be overlooked. For example, the term 'Community-in-Translation Events Grant' in the term dictionary will not be matched with 'Community in Translation Events Grant' (missing hyphens) or 'Community-in-Translation Event Grant (plural to singular form of 'event')' in an input sentence.

Therefore, fuzzy matching mechanism is deployed to resolve this issue in the term matching phase. Prior to term matching, the input sentence and the terms found in the terminology dictionary undergo a process known as de-punctuation. After term matching and replacement, the remaining punctuation in the sentence will be restored. At the same time, chained dictionaries and the stemming algorithm (Lovins J B., 1968) are introduced to rectify the issue of matching failures caused by tense differences or singular-plural form differences.

The principle of a chained dictionary is similar to that of a linked table, where a phrase is being split into words, with the preceding word in the phrase serving as the key to the following word, and the following word serving as the value of the preceding word. At each step of the chain dictionary query, if the word in the sentence or its stemmed form can match the key of the chain dictionary or the stemmed form of the key, the next query is performed. Otherwise, the query of the chain dictionary ends. The chained dictionary is also applicable to Tamil, Malay and Chinese.

Table 1. Comparison of model performance with and without fuzzy match

Language Pair	**BLEU**	
	With fuzzy match	**Without fuzzy match**
English to Tamil (EN2TA)	15.76	13.12
English to Chinese (EN2ZH)	16.55	13.7
English to Malay (EN2MS)	16.93	14.57
Tamil to English (TA2EN)	17.31	15.75

Chinese to English (ZH2EN)	19.77	18.82
Malay to English (MS2EN)	18.03	17.14

Table 1 shows that when fuzzy match is applied, the BLEU score for EN2TA, EN2ZH, EN2MS, TA2EN, ZH2EN and MS2EN improves by 2.64, 2.85, 2.36, 1.56, 0.95 and 0.89 respectively when translating 500 sentences containing input variants. The BLEU score improvement is remarkable in EN2TA, EN2ZH, EN2MS and TA2EN (more than 1 BLEU score). As only the de-punctuation operation was applied to the input sentences for ZH2EN and MS2EN translation engines, the performance of translation engines improved, but not significantly. Nonetheless, the results show that allowing the engine to recognise input variants via fuzzy match, improves the translation quality.

3.3. Multi-stage Domain Adaptation

To improve the translation performance of Singapore-contextualised content, adaptation (Chen et al., 2016) on domain-specific data related to Singapore content was carried out. Domain adaptation is performed in multiple stages. In the first stage, all domain bilingual data, including data not relevant to our domain and back-translation data, are used for building a base translation model which can acquire general translation knowledge. In the second stage, all high quality but non-domain specific training data are selected and used to further improve the translation quality of the model. In the final stage, high quality domain-specific data is used to further adapt the model finetuned in the second stage, thus emphasising translation learning of localised content.

3.4. Data Tagging

The training data sources for SGTT engine mainly include back-translation (BT) data, out-domain data and localised bilingual data. Among them, BT data and some out-domain data contain a certain degree of noise. Since the deep neural network is data-sensitive, when a translation system over-fits to certain features of noisy data, it will lead to the degradation of translation quality. Drawing on the approach mentioned by Marie B et al. (2020), we classify the data from different sources into two categories based on the quality of the data. A tag is added to the beginning of each sentence at the source side of each category of data to guide the model to gain data category information in the training process. We use "<BT>" for the data containing noise and "<PA>" for the data which is of good quality.

4. Illustrative Performance

This section illustrates SG Translate's capability to produce localised translations which are specific to Singapore's context.

4.1. Translation Related to Local Culture

SG Translate can translate sentences carrying local cultural context. In the example below, 'Hungry Ghost Festival' is a festival that is observed by many Chinese in the region and 'getai' is a live stage performance which usually takes place during the seventh month of the Chinese lunar calendar. The MT engine is able to recognise the cultural context and provide the translation suited for local audiences. Other MT engines may not be able to recognise this unique festival and this special genre of stage performance, which is seen only around this region.

Source (Chinese):	一提到**中元节**，人们一般上会想到**歌台**。
Target (English):	When it comes to **Hungry Ghost Festival**, people generally think of **getai**.

4.2. Translation Related to Everyday Life

'NRIC' is being used in Singaporeans' everyday life. It is a colloquial way of referring to one's identity card and stands for 'National Registration Identity Card'. The example below illustrates the positive outcome of domain adaptation where SG Translate is able to recognise 'NRIC' and provide an accurate translation of it in Tamil.

Source (English):	Bring your **NRIC** or valid passport, and poll card.
Target (Tamil):	*உங்கள் **அடையாள அட்டை** அல்லது செல்லுபடியாகும் கடவுச்சீட்டு மற்றும் வாக்கு அட்டை ஆகியவற்றைக் கொண்டு வாருங்கள்.*

4.3. Translation Related to the Government

The translation of government terms is standardised and has been included in terminology dictionaries to ensure that when the public translates local content related to government policies and schemes, the correct translation will be generated. As seen in the example below, the term 'Medishield Life' is a uniquely Singaporean term, as it is a healthcare insurance scheme administered by the Singapore Government. SG Translate is able to render the correct translation of 'Medishield Life' and its Malay equivalent, 'Medishield Hayat' as the term has been coined and added into the terminology dictionary of the MT Engine. Other MT engines may render it as 'Life Medishield' which is incorrect. Additionally, the Malay sentence is also a colloquial example of how Malay may be spoken in informal contexts in Singapore. SG Translate was able to render a satisfactory translation in English in spite of that, affirming the MT Engine's sensitivity to not only the local context but local linguistic patterns as well.

Source (Malay):	Awak tak tahu ke yang kita semua ada **MediShield Hayat?**
Target (English):	Don't you know we all have **MediShield Life?**

5. SG Translate Together Web Portal

Aligned with Singapore's Smart Nation initiatives, SG Translate was introduced to the public sphere via the SG Translate Together (SGTT) web portal (sgtranslatetogether.gov.sg). The portal aims to encourage more citizenry engagement to raise translation standards together. Besides performing translations via SG Translate, visitors can also access the one-stop repository of various translation events and translation-related resources on SGTT. Additionally, members of the public who are passionate about languages and translation can register for an account via Singapore's digital ID, Singpass, to take on translation tasks and contribute their post-edited translations to further train and improve the MT engine. The SGTT web portal was officially launched on 27 June 2022 and will be further enhanced with new features such as a community forum to promote interaction and collaboration between translation enthusiasts.

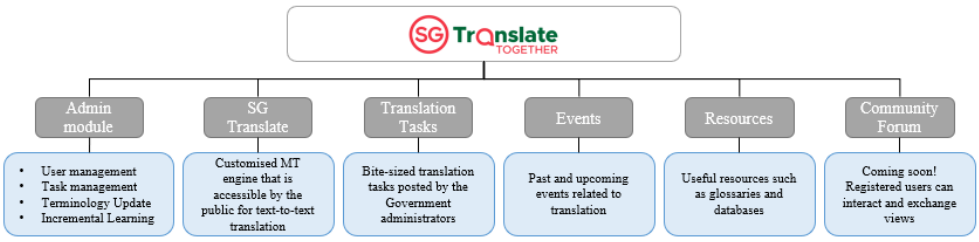


Figure 3. Overview of the SGTT web portal

5.1. Sustainable Approach

As continual training of the developed MT engine is paramount for improving its accuracy and ensuring its relevance in a rapidly-changing world, the incremental learning capability was developed to help the engine learn new knowledge and update the translation model when new data resources emerge. In addition, to overcome the problem of catastrophic forgetting caused by the introduction of new knowledge, the model ensemble technique (Garmash E et al., 2016) will be used in the final engine deployment phase.

As the Chinese, Malay and Tamil languages used in Singapore vary from those used in other regions, it is challenging to obtain new data to train the MT engine sustainably. Initially, MCI obtains quality bilingual data from Government agencies and partners from the private sector. With the establishment of the SGTT web portal, it has created an ecosystem that supports the sustainable training of SG Translate through the contribution of post-edited translations by registered users[1]. Many of them are volunteers called the Citizen Translators (CT) that MCI has recruited to work together to raise translation standards through a myriad of activities.

There are two types of registered users on the SGTT web portal: SGTT Translators and SGTT Proofreaders. Both types of users can take up translation tasks posted by the Singapore Government which includes a source text and the machine-generated translation, which is generated by SG Translate. They are required to post-edit the machine-generated translation and submit it for review. SGTT Proofreaders, who are registered users who are more experienced and well-versed with translations, will then review these translations by further editing them, and then providing ratings and feedback on the translation for the SGTT Translator. The reviewed translations are stored in the web portal and eventually extracted on a periodic basis to further train SG Translate via the incremental learning method designed by I^2R.

Through the abovementioned approach, the Singapore Government is able to sustainably obtain translation data through citizen engagement.

When registered users contribute their translations, their participation is recognised. Depending on their level of participation, they can receive e-certificates of recognition, e-vouchers or even being eligible to apply for training subsidies for translation-related courses. This mutually beneficial workflow allows the Singapore Government to obtain bilingual data by collaborating with citizens, consequently allowing them to hone their translation skills.

[1] Registered users are members of the public who have registered for a user account on SGTT via their digital ID, known as Singpass. Unlike non-registered users, this group of users are able to do more than just using the SG Translate MT engine to generate translations. They can contribute their own post-edited translation and/or give feedback to other translators on how to improve their translation. Registered users will also have access to the community forum which allows them to interact with one another and discuss translation matters.

5.2. Shaping Singapore's Translation Landscape and Nurturing the Next Generation of Translators

Translation is important in Singapore's multiracial and multilingual society. It not only bridges the communication gap between different communities, but also serves to strengthen mutual understanding. As we enter the digital era, there are more opportunities for the Singapore Government to harness technology to improve its work processes, as well as collaborate with the community. The SGTT web portal is one such opportunity.

By offering SG Translate to the public as a free-to-use tool, the Singapore Government hopes to lower the barriers in putting out content in the official languages to improve accessibility to information. Nonetheless, users are always reminded to check and edit the translation before disseminating the translated materials for public consumption. While technology can act as a catalyst, machines cannot replace human translators as communication is ultimately a connection between people, and translators are necessary to ensure that the content is best suited for the intended audience. Through the SGTT web portal, the Singapore Government hopes to encourage more people to adopt and develop translation technology to change the way translation is being done conventionally.

Prior to SGTT, there were numerous initiatives such as the Translation Talent Development Scheme (TTDS)[2] to support and nurture translation enthusiasts and practitioners. Complementing translation initiatives to date, the SGTT web portal is an addition to the suite of initiatives to uplift Singapore's translation standards. The web portal takes this a step further by providing interested individuals with a platform to learn from one another. As mentioned, SGTT Translators will receive feedback from SGTT Proofreaders on how to improve their translations. As users need not have prior experience to join SGTT as a Translator, anyone can join and hone their translation skills through the process. By gathering these passionate individuals - both inexperienced and experienced - "under one roof", it forms an active translation community in Singapore, fostering the spirit of sharing and learning, thereby nurturing the next generation of translators.

6. Summary

In this paper, we have presented the key methodology for the development of SG Translate – the Government customised MT engine. We also illustrated how SG Translate is able to translate content pertaining to local culture, everyday life and the Singapore Government. The results show that SG Translate has produced translations that are suited for Singapore's context. The paper also shared the conceptualisation and execution of SG Translate Together, and how it is used to obtain training data for SG Translate in a sustainable fashion.

Acknowledgement

This paper is co-tabled by the Translation Department from the Ministry of Communications and Information (Singapore), and the Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (Singapore). The SG Translate Customised Machine Translation Project is supported by the National Research Foundation, Singapore and Smart Nation and Digital Government Office, Singapore under its Translational R&D for Application

[2] The Translation Talent Development Scheme (TTDS) is a co-sponsorship grant set up by the National Translation Committee (NTC) to encourage Singaporean translation and interpretation (T&I) practitioners to further develop their capabilities and to attain mastery and deepening of their skills. The scheme also aims to nurture the next generation of translation talent in Singapore.

to Smart Nation (LOA No.: NRF2016IDM-TRANS001-062). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Smart Nation and Digital Government Office, Singapore.

Our authors would like to express our gratitude to the aforementioned agencies for supporting the development of the SG Translate MT engine, as well as all Government agencies and private entities which have contributed their monolingual and bilingual data to develop and further train SG Translate. In addition, MCI would like to thank Government Technology Agency (GovTech), Singapore for their support in building the SG Translate Together Web Portal.

References

Chen, M. X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., ... & Hughes, M. (2018). The best of both worlds: Combining recent advances in neural machine translation. *arXiv preprint arXiv:1804.09849*.

Chen, B., Kuhn, R., Foster, G., Cherry, C., & Huang, F. (2016). Bilingual methods for adaptive training data selection for machine translation. *In Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track (pp. 93-106).*

Fadaee, M., & Monz, C. (2018). Back-translation sampling by targeting difficult words in neural machine translation. *arXiv preprint arXiv:1808.09006.*

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.*

Garmash, E. & Monz, C. (2016). Ensemble Learning for Multi-Source Neural Machine Translation. *In Proceedings of the 26th International Conference on Computational Linguistics (COLING).*

Lovins, J. B. (1968). Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics, 11(1-2), 22-31*.

Marie, B., Rubino, R., & Fujita, A. (2020, July). Tagged back-translation revisited: Why does it really work?. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5990-5997).*

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30.*

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*.

Wang, T., Kuang, S., Xiong, D., & Branco, A. (2019). Merging external bilingual pairs into neural machine translation. *arXiv preprint arXiv:1912.00567.*