Singlish to English Translation with Precision: A Dataset and Language Detection-Driven Masked Modeling for Singlish to English Translation

Sujit Kumar, Gerome Kusuma, Stephanie Hilary Xinyi
Andy Hau Yan Ho, Andy Khong
Lee Kong Chian School of Medicine, Nanyang Technological University
Singapore
sujitkumar@iitg.ac.in, gerome.ang@gmail.com,
{stephaniehilary.ma, andyhyho, andykhong}@ntu.edu.sg

Abstract

Singlish, a creole rooted in English and influenced by Singapore's multilingual and multicultural environment, poses significant challenges for those proficient in standard English due to its unique and often complex lexical and syntactic structures. Despite significant advancements in language translation for both high- and low-resource languages, translating Singlish to English remains largely underexplored. This gap is primarily due to the lack of dedicated datasets for language detection and Singlish-to-English translation, as well as the absence of robust models capable of addressing the unique linguistic challenges posed by Singlish. In this work, we curate a word-level language detection dataset, a Singlish-to-English translation dataset, and propose a Language Detection-driven Masked Language Modelling approach for translating Singlish into English. We evaluate the performance of existing models and the proposed approach on two Singlish-to-English translation datasets, including our proposed SEAT dataset. The results demonstrate that the proposed LD-MLMTrans approach outperforms the baseline model and exhibits high proficiency in Singlish-to-English translation.

Keywords: Singlish, Language Identification, Masked Language Modelling, Creole Translation, Low-Resource NLP

1. Introduction

The increase in interconnectedness across diverse societal boundaries has driven the need for seamless and effective communication across linguistic and cultural boundaries. Advancements in Artificial Intelligence (AI) and, in particular, Natural Language Processing (NLP), have advanced both text-based and speech-based translation to facilitate the effective cross-linguistic communication and deeper cross-cultural engagement. In this aspect, the translation process involves conveying and adapting the unique linguistic characteristics of one language to another.

Singlish—or Colloquial Singapore English (CSE)—is an English-based creole language influenced by the diverse set of regional spoken languages spoken in Singapore. The predominant Chinese, Malay, and Indian population communicate through their mother-tongue languages including Mandarin, Hokkien, Teochew, Cantonese, Malay, and Tamil (Deterding, 2007; Leimgruber, 2011). Despite being a multi-racial society that values social integration and mixing facilitated by policies, cross-lingual influence results in Singlish with significant variations in lexicon, syntax, and semantics (Wee, 2004; Wang et al., 2017). Beyond the challenges associated with bilingual code-switching speech (Liu et al., 2021), the distinct lexical and syntactic features of Singlish make it even more challenging for native English speakers unfamiliar with Singlish (Harada, 2009; Liu et al., 2022).

Advances in LLMs have led to growing interest in Singlish-to-English translation. To this end, key linguistic and technical challenges remain and these include: (i) Lexical diversity: Singlish draws from multiple regional languages, leading to non-standard sentence structures that complicate semantic interpretation (Ningsih and Rahman, 2023); (ii) Implicit subjects and pronouns: Omitted subjects and pronouns—a common feature in Southeast Asian languages—that hinders speaker inference and discourse tracking (Ng and Chan, 2024); (iii) Tense ambiguity: Minimal use of morphological tense markerssimilar to the Chinese and Malay languagesleaves temporal intent implicit and hard to resolve (Ng and Chan, 2024); (iv) Use of discourse particles: Context-sensitive particles such as *lah*, *lor*, and *leh* convey pragmatic nuance absent in standard English, confounding direct translation (Foo and Ng, 2024); (v) Compressed syntactic structure: Omitting auxiliary components and mid-sentence punctuation makes segmentation and meaning extraction difficult (Ng and Chan, 2024); (vi) Creative morphological forms: Non-standard verb forms (for example, *stone-ed*, *sabo-ed*) reflect playful linguistic innovation that challenges model generalisation (Ng and Chan, 2024).

Notwithstanding the above, the scarcity of high-quality Singlish–English datasets hinders the development of an effective translation system (Ng and

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3506–3516
11–16 May 2026.

Chan, 2024). While recent efforts have created some Singlish-English datasets, these datasets are synthetically generated and limited in size (Ng and Chan, 2024; Liu et al., 2022), rendering the available data insufficient for developing robust models for language detection and Singlish to English translations. A recent study (Chow et al., 2024) curated an open-source Singlish dictionary to facilitate natural language understanding. However, the curated dictionary is limited in size and lacks lexical diversity. A joint paraphrasing task involving the translation and normalisation of Singlish text has been proposed at three linguistic levels (Liu et al., 2022): lexical level normalisation, syntactic level editing, and semantic level rewriting. Furthermore, few-shot prompting with GPT-2 for Singlish-to-English translation (Liu et al., 2022) and few-shot prompting with LLMs for phrase-level language detection and Singlish-to-English translation (Ng and Chan, 2024) have also been explored.

Motivated by existing challenges in Singlish-to-English translation, this study makes two key contributions. First, we extend the open-source Singlish dictionary (Chow et al., 2024) by incorporating frequently used non-English Singlish words from everyday Singaporean conversations. These words are subsequently used to curate a dataset for word-level language detection. We then propose two datasets for Singlish-to-English translation: (i) Singlish-English Aligned Translation (SEAT): a human-annotated Singlish to English translation dataset and (ii) Synthetic Singlish-English Translation (SSET): a synthetic Singlish to English translation dataset annotated using GPT-4. We subsequently propose a Language Detection-driven Masked Language Modelling LD-MLMTrans approach for Singlish-to-English translation. The proposed LD-MLMTrans approach first applies word-level language detection in Singlish text and replaces the non-English words in Singlish text with $\langle \texttt{MASK} \rangle$ tokens. The masked Singlish text is then passed to a language model that predicts appropriate replacements for the $\langle \texttt{MASK} \rangle$ tokens based on the surrounding English context and source Singlish text sequence, generating a complete translation into formal English.

We conducted experiments to evaluate the effectiveness of models trained on our proposed datasets for Singlish-to-English translation and study the performance of the proposed model. Results indicate that our proposed method outperforms existing baseline models and demonstrates high proficiency in detecting the language of words in Singlish text and translating Singlish to English.

Our key contributions can be summarised as follows: (i) We enrich the open-source Singlish dictionary (Chow et al., 2024) by incorporating additional Singlish words, and propose the **Augmented Singlish Dictionary**. Our proposed dictionary is one of the largest Singlish word resource for natural language understanding tasks involving Singaporean languages. Furthermore, we propose **Augmented Singlish Language Detection Corpus** dataset for word-level language detection in Singlish text by utilising *Augmented Singlish Dictionary*. (ii) This study proposes the Singlish-English Aligned Translation (SEAT) dataset to foster the development of Singlish-to-English translation models; (iii) We propose the LD-MLMTrans approach for Singlish-to-English translation.

2. Related Work

Recent surveys (Gemechu and Kanagachidambaresan, 2023; Gain et al., 2025; Zhu et al., 2024; Nam and Jang, 2024; Maruf et al., 2021; Datta et al., 2024) provide overviews and analyses of existing methods and challenges associated with machine translation. On the contrary, we review existing works associated with Singlish-to-English translation. In particular, a task-driven representation in learning has been explored to disentangle Singlish discourse particles such as *lah*, *meh*, and *hor* (Liu et al., 2022). These discourse particles were subsequently clustered based on their pragmatic functions which, in turn, facilitate downstream tasks such as Singlish-to-English machine translation.

Few-shot prompting with GPT-2 for Singlish to English translation has also been investigated (Foo and Ng, 2024) while a multi-step prompting method leveraging LLaMA-3.1 has been proposed for Singlish to English translation (Ng and Chan, 2024). However, such a prompting-based approach relies heavily on handcrafted prompts, lacks task-specific adaptation through fine-tuning, and fails to incorporate structured linguistic resources that are lacking in Singlish lexicon. This, in turn, limits its ability to accurately interpret idiomatic, culturally embedded, and contextually nuanced expressions present in Singlish.

Further to the above, phrase-level language detection assumes uniformity within text segments, often neglecting intra-phrase code-switching, which is a frequent characteristic of Singlish. This coarse granularity results in inaccurate labelling of mixed expressions and adversely affects downstream tasks (Liu et al., 2025). Datasets and models have also been proposed for language detection and machine translation across two hundred low-resource languages in the world in the spirit of achieving "No Language Left Behind" (Costa-Jussà et al., 2022). However, the dataset does not include Singlish Language.

3. Proposed Dataset

Singlish is a low-resource language with no publicly available datasets for word-level language detec-

Table 1: Comparison of English and non-English word distributions in the Baseline Singlish Language Detection Corpus (BSLDC) and the Augmented Singlish Language Detection Corpus (ASLDC).

Dictionary words	Singlish	English	Total
BSLDC	1,218	48,595	49,813
ASLDC	2,622	48,595	51,217

tion in Singlish text, and only synthetically generated datasets are available for training Singlish-to-English translation models. Motivated by this, we propose the *Augmented Singlish Language Detection Corpus* for word-level language detection by collecting English and non-English words commonly used in everyday Singaporean conversations, as will be described in Section 3.2. To facilitate the development of NLP models for Singlish, we also curate a Singlish-to-English translation dataset, as described in Section 3.3.

3.1. Augmented Singlish Dictionary for Language Detection

We curate an Augmented Singlish Dictionary by augmenting words into the open-sourced Singlish dictionary (Chow et al., 2024) by collecting non-English Singlish words *Cambridge English Dictionary*[1], *Dictionary of Singlish and Singapore English*[2], *Singlish Wikipedia*[3], Reddit posts, conversational datasets, and the *National Speech Corpus* (Koh et al., 2019). Subsequently, we manually annotate the English meaning of the newly collected words to form the *Augmented Singlish Dictionary* by combining the newly collected Singlish words and Singlish words in the open-sourced Singlish dictionary (Chow et al., 2024).

3.2. Word-Level Language Identification Dataset for Singlish Text

We curated two word-level language detection datasets for detecting English and non-English Singlish words in Singlish text: (i) the *Baseline Singlish Language Detection Corpus* (BSLDC) and (ii) the *Augmented Singlish Language Detection Corpus* (ASLDC). We first built a dictionary of English words from texts in news publications, media outlets, and the *National Speech Corpus* (Koh et al., 2019). Next, we combined the English words and Singlish words from the English Dictionary and open-source Singlish dictionary (Chow et al., 2024) to form the *Baseline Singlish Language Detection Corpus* for the word-level language detection task.

Similarly, we form *Augmented Singlish Language Detection Corpus* for word-level language detection in Singlish text by combining the English words with the Singlish words from the *Augmented Singlish Dictionary*. Both the *Baseline Singlish Language Detection Corpus* and the *Augmented Singlish Language Detection Corpus* are curated using the same collection of English words to foster the development of word-level language detection models in Singlish text. However, *Baseline Singlish Language Detection Corpus* includes Singlish words from the open-sourced Singlish dictionary (Chow et al., 2024). In contrast, *Augmented Singlish Language Detection Corpus* includes Singlish words from *Augmented Singlish Dictionary* described in Subsection 3.1.

3.3. Singlish-English Aligned Translation (SEAT) Dataset

This study proposes the Singlish-English Aligned Translation (SEAT) dataset to foster the training and development of robust models for translating Singlish to English. We consider the National University of Singapore Short Message Service Corpus (Chen and Min-Yen, 2015), which contains only Singlish text without any ground truth English translations. Our SEAT dataset curation process is organised as follows: (i) We first curated a subset of the Singlish text corpus $\mathcal{C}$ by randomly selecting 1,500 samples from the complete SMS Corpus; (ii) We subsequently obtained English translations for each Singlish text within the set $\mathcal{C}$ through manual translation by human annotators. To ensure quality translation and annotation, we selected two native Singlish speakers, a student and a research fellow, both fluent in Singlish and English. The instructions provided to annotators for translating Singlish to English are as follows: (i) *Preserve Intended Meaning, Not Literal Structure* : Translate each Singlish sentence into fluent Standard English by conveying the intended meaning, even if it requires rephrasing or reordering words. Avoid literal word for word translation that may distort the meaning or sound unnatural in English. (ii) *Handle Discourse Particles Appropriately*: Interpret Singlish discourse particles such as *lah*, *lor*, *meh*, *leh* based on their pragmatic function and reflect their tone or emphasis appropriately in English. For example, *Don't be late lah* may be translated as *Don't be late* or *Please don't be late*, depending on the conversational tone.

Furthermore, We consider the National University of Singapore Short Message Service Corpus (Chen and Min-Yen, 2015), to curate a large-scale Synthetic Singlish-English Translation (SSET) dataset by considering GPT-4o as an annotator. We prompted GPT-4o to generate English translation of corresponding input Singlish text. To prevent any overlap between the samples of the SEAT and SSET datasets, the 1,500 samples $\mathcal{C}$

[1] Dictionary of Singlish and Singapore English
[2] Dictionary of Singlish and Singapore English
[3] Singlish Wikipedia

Table 2: Definitions of the 1–5 Likert scale scores used for human evaluation of Adequacy, Fluency, and Consistency.

Score	Adequacy	Fluency	Consistency
1	Unrelated meaning; mistranslation or nonsense.	Incomprehensible or ungrammatical.	Completely inconsistent terminology or style.
2	Major meaning loss; only vague fragments retained.	Hard to read with many grammar mistakes.	Frequent inconsistencies that confuse meaning.
3	Partial meaning preserved; some key info missing.	Understandable but with noticeable grammar errors.	Some inconsistencies affect readability.
4	Meaning mostly preserved; minor omissions.	Minor grammar issues; overall natural.	Mostly consistent with minor variations.
5	Fully preserves meaning; no loss or addition.	Fully grammatical and natural English.	Terminology and style fully consistent.

Table 3: Inter-annotator agreement measured using Fleiss Kappa for Adequacy, Fluency, and Consistency.

Parameter	Fleiss Kappa (k)	Agreement Level
Adequacy	0.792	Substantial Agreement
Fluency	0.846	Almost Perfect Agreement
Consistency	0.812	Almost Perfect Agreement

used for the SEAT dataset curations were intentionally excluded during the GPT-4o-based annotation of the SSET dataset. Our prompt instruction to the GPT-4o annotator was as follows: *"You are an expert translator specialising in Singapore Colloquial English (Singlish). Translate the Singlish sentence into fluent, grammatically correct English while preserving meaning and cultural context."*

To assess the quality of annotations by GPT-4o, we conducted a human evaluation of a small subset of 250 Singlish and English pairs from the newly curated SSET datasets. The 250 samples were assigned to four annotators who rated the English translations generated by GPT-4o for each Singlish text on a 1–5 scale along the dimensions of adequacy, fluency, and consistency. Table 2 presents the descriptions and scoring criteria for each dimension. The inter-annotator agreement using Fleiss Kappa (κ), shown in Table 3, indicates substantial agreement for adequacy and almost perfect agreement for fluency and consistency, demonstrating high reliability of the human evaluations and ensuring that the curated dataset provides credible and trustworthy quality assessments for subsequent model evaluations.

4. The Proposed LD-MLMTrans Method

As discussed in Section 1, we propose a Language Detection-driven Masked Language Modelling LD-MLM) MLMTrans method for translating Singlish to English, as shown in Figure 1. Our proposed LD-MLMTrans method consist of two main stages: (i) Selective Masking for Contextual Translation and (ii) Contextual Masked Language Translation.

4.1. Selective Masking for Contextual Translation

Given a text sequence $\mathcal{S} = w_1, w_2, \ldots, w_n$ with n words, we identified each word w_i as either an English word or a non-English word commonly used in Singapore. To indefiy English word or a non-English word commonly used in Singlish text we fine-tuned RoBERTa (Liu et al., 2019) we adopted continuous training using Language Identification (LID) models, namely LID-176 by fastText (Joulin et al., 2017, 2016)[4] and lid218e from Meta AI's No Language Left Behind project (NLLB Team et al., 2022; Costa-Jussà et al., 2022)[5], for Singlish language detection on the BSLD and ASLDC datasets curated in Section 3.2. Ideally, any LID model may be employed to identify the language of each word w_i in the Singlish text sequence $\mathcal{S}$. However, given the superior performance of the RoBERTa model reported in Table 6 for word-level language detection on Singlish text, we adopted RoBERTa to identify the language of each word w_i in the text sequence $\mathcal{S}$. Subsequently, we masked the non-English words w_i in the text sequence $\mathcal{S}$ using the $\langle\texttt{MASK}\rangle$ token to obtain a new sequence $\mathcal{M}$, with all non-English words replaced with the $\langle\texttt{MASK}\rangle$ token.

4.2. Contextual Masked Language Translation

. Given a text sequence $\mathcal{M}$ with all non-English words masked, the objective of *Contextual Masked Language Translation* is to generate the corresponding English words to replace the $\langle\texttt{MASK}\rangle$ tokens based on the context of unmasked English words in $\mathcal{M}$. We consider two Contextual Masked Language Translation methodologies leveraging the LLMs: (i) masked language modelling and (ii) instruction tuning with LLMs. As shown in Figure 1 in the masked language modelling setup, the masked text sequence $\mathcal{M}$ along with a prompt instruction and Singlish text sequence $\mathcal{S}$ is fed as input to an LLM. The model is expected to infer and generate the appropriate English words to replace the $\langle\texttt{MASK}\rangle$ tokens by utilising the contextual information from the surrounding unmasked English words in $\mathcal{M}$ and the Singlish text sequence $\mathcal{S}$. Our proposed LD-MLMTrans leverages the training paradigm and characteristics of auto-regressive LLMs, which are trained to predict the next token or $\langle\texttt{MASK}\rangle$ token based on the given context and previously generated tokens (Merrill et al., 2024; Gong et al.; Olsson et al., 2022). Similarly, LD-MLMTrans generates replacements for the masked tokens $\langle\texttt{MASK}\rangle$ in the masked text sequence $\mathcal{M}$ by utilis-

[4] LID-176
[5] NLLB

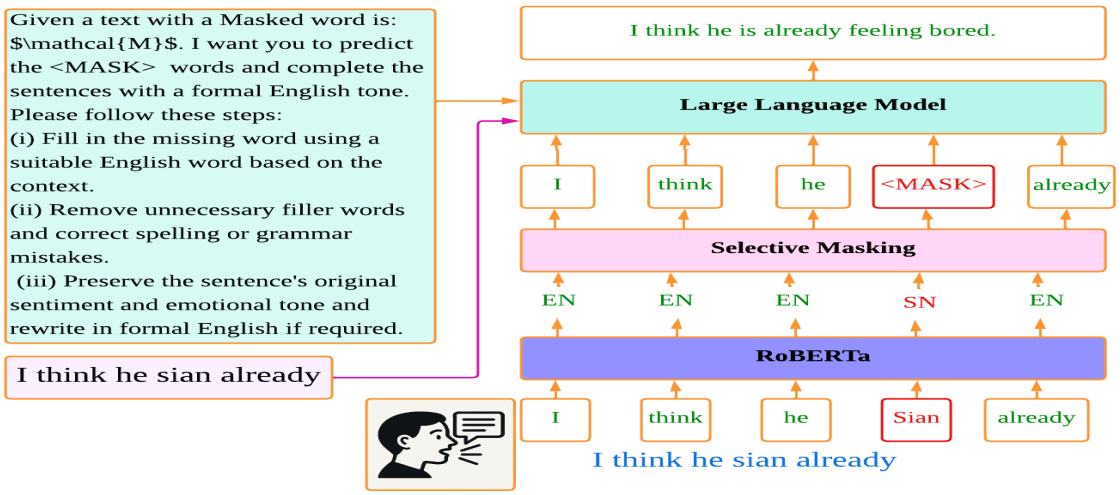

Figure 1: The proposed Language Detection-driven Masked Language Modeling Translation LD-MLMTrans.

ing the context from the corresponding Singlish text sequence $\mathcal{S}$ and the unmasked English words present in $\mathcal{M}$.

For instruction tuning, we fine-tune an LLM using the following steps: *(a)* The original Singlish text sequence $\mathcal{S}$ was transformed into a masked sequence $\mathcal{M}$ by replacing all non-English words with the $\langle$MASK$\rangle$ token, as detailed in Section 4.1. *(b)* Next, we finetune LLMs using $\mathcal{M}$ and a prompt instruction as input, with the corresponding English translation as the target output. Our prompt instruction for instruction tuning and masked language modelling with Llama is as follows: *Given a text with a Masked word $\mathcal{M}$, predict the $\langle$MASK$\rangle$ words and complete the sentences with a formal English tone. Please follow these steps: (i) Fill in the missing word using a suitable English word based on the context. (ii) Remove unnecessary filler words and correct spelling or grammatical errors. (iii) Preserve the original sentiment of the sentence and emotional tone and rewrite in formal English if required.* We present eight pairs of Singlish sentences and their corresponding English translations, accompanied by step-by-step reasoning, as examples to LLMs while instruction tuning.

Table 4: Characteristics of Singlish-to-English translation datasets. Here, *Singlish Words* and *English Words* indicate the average number of Singlish and English words per sample, respectively.

Dataset	Split	#Samples	Singlish Words	English Words
SEAT	Train	1,150	8.672	9.899
	Dev	150	8.738	9.150
	Test	200	8.835	9.593
SSET	–	51,530	10.646	11.178

5. Experiment Configurations

To study the performance of the proposed and baseline models, we consider a set of commonly used performance metrics, including BLEU score (Papineni et al., 2002), METEOR score (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), BERTScore (Zhang* et al., 2020), and MoverScore (Zhao et al., 2019). These metric captures different and distinct dimensions of model performance , including lexical overlap, semantic alignment, and the contextual similarity between the ground truth English text and the translated English text of Singlish text. We consider overall accuracy and class-wise F1 Scores as performance metrics for evaluating the word-level language detection model. This study consider RoBERTa (Liu et al., 2019)[6] for the language detection task. Similarly we also consider LLaMA-3.3-70B-Instruct (Grattafiori et al., 2024)[7], Mistral-Nemo-Instruct-2407 [8], Gemma-3 (Team et al., 2025)[9] and Phi-4 (Abdin et al., 2024)[10] as large language models in the proposed LD-MLMTrans, and instruction tuning. We adopted LoRA (Low-Rank Adaptation of Large Language Models) (Hu et al., 2022) for fine-tuning and instruction tuning. Table 5 presents the details of the hyperparameters used to produce the results presented in this paper.

5.1. Datasets

We conducted our experiments on our proposed SEAT and SSET datasets to evaluate the performance of the baseline and proposed models. Table 4 presents the characteristics of *SSET* and

[6] RoBERTa
[7] Llama-3
[8] Mistral
[9] Gemma-3
[10] Phi-4

Table 5: Details of Experimental Setups and Hyperparameters.

Hyperparameter	Value
Batch Size	16
Learning Rate	0.01
Maximum Number of Beams	4
Number of Epochs	100
Gradient Accumulation Steps	4
Weight Decay	0.01
Quantization	4-bit
LoRA Rank (r)	16
LoRA Alpha	32
Max New Tokens (Inference)	150
Max New Tokens (Inference, CoT)	250
LR Scheduler Type	Linear
Optimizer	AdamW 8-bit

Table 6: Performance comparison of RoBERTa and Language Identification (LID) models for word-level language detection on Singlish text. Here, **Sing.** and **Eng.** denote the F1 scores for detecting non-English and English words, respectively.

Type	Model	BSLDC			ASLDC		
		Acc	Sing.	Eng.	Acc	Sing.	Eng.
Pretrained	LID (fastText) (Joulin et al., 2017)	0.689	0.190	0.800	0.689	0.190	0.809
	LID (NLLB) (Costa-Jussà et al., 2022)	0.628	0.211	0.756	0.630	0.162	0.762
Fine-tuned	LID (fastText) (Joulin et al., 2017)	0.932	0.449	0.964	0.978	0.764	0.988
	LID (NLLB) (Costa-Jussà et al., 2022)	0.934	0.437	0.965	0.979	0.776	0.989
	RoBERTa	**0.935**	**0.527**	**0.966**	**0.981**	**0.862**	**0.989**

SEAT datasets. We used the training set of the SEAT dataset to train the models and their test set to evaluate their performance. Furthermore, the *SSET* dataset is used exclusively to train the instruction-tuned variant of our proposed method, **LD-MLMTrans(IT)**, on a larger and more diverse Singlish-to-English corpus. Notably, the *SSET* dataset is not used in the evaluations due to its synthetic curation process. Our proposed SEAT dataset is publicly available at Hugging Face[11], and the language detection dataset ASLDC dataset is available at Hugging Face [12].

5.2. Baseline Models

We consider **GPT-2** (Foo and Ng, 2024) and **Multi-step Prompting** (Ng and Chan, 2024) as baseline models to compare the performance of the proposed models in translating Singlish to English. Additionally, we consider chain-of-thought prompting with eight-shot examples using LLaMA-3 **Llama-3 (CoT)** and supervised fine-tuning of LLaMA-3 for Singlish to English translation **Llama-3 (SFT)** as the baseline model. Our eight-shot Prompting and prompt instruction for **Llama-3 (CoT)** is as follows: *Translate the following Singlish text into standard English. Singlish sentences may contain a mixture of English and regional language words. Please follow the step-by-step process below:(i) Identify the language of each word in the sentence. (ii) Translate non-English words into their English equivalents. (iii) Remove unnecessary filler words and correct spelling or grammatical errors. (iv) Reconstruct and translate the Singlish text into English using the information from the above steps to produce a complete sentence in standard English.* Whereas in case **Llama-3 (SFT)** we provide a prompt instruction and Singlish as input and corresponding English text as ground truth to Llama. Our prompt instruction for **Llama-3 (SFT)** is as follows: *Translate the following Singlish sentence to English:*.

[11] SEAT Dataset
[12] ASLDC Dataset

Table 7: presents performance comparisons of **LD-MLMTrans(IT)** when trained on the training set of the SEAT dataset, and the SSET dataset, and evaluated on the test set of the SEAT dataset.

Method	BLEU	METEOR	ROUGE-L	MoverScore	BERTScore
Trained on the Training set of SEAT					
LD-MLMTrans (Llama, *IT*)	**0.642**	**0.732**	**0.746**	**0.683**	**0.986**
LD-MLMTrans (Mistral, *IT*)	0.637	0.727	0.741	0.678	0.982
LD-MLMTrans (Gemma-3, *IT*)	0.637	0.727	0.741	0.678	0.982
LD-MLMTrans (Phi-4, *IT*)	0.635	0.725	0.739	0.676	0.981
Trained on SSET					
LD-MLMTrans (Llama, *IT*)	**0.741**	**0.787**	**0.814**	**0.689**	**0.989**
LD-MLMTrans (Mistral, *IT*)	0.738	0.784	0.811	0.686	0.986
LD-MLMTrans (Gemma-3, *IT*)	0.735	0.781	0.808	0.684	0.984
LD-MLMTrans (Phi-4, *IT*)	0.733	0.779	0.806	0.682	0.983

Table 8: Performance comparison of baseline and proposed methods for Singlish to English translation on the SEAT dataset. The models are trained on the training set of SEAT and evaluated on the test set of SEAT.

Method	BLEU	METEOR	ROUGE-L	MoverScore	BERTScore
GPT-2 (Foo and Ng, 2024)	0.226	0.346	0.379	0.370	0.805
Multi step Prompting (Ng and Chan, 2024)	0.284	0.428	0.463	0.482	0.847
Llama-3 (CoT)	0.351	0.567	0.671	0.510	0.853
Llama-3 (SFT)	0.402	0.621	0.726	0.561	0.896
LD-MLMTrans (Llama)	0.586	0.627	0.677	0.595	0.944
LD-MLMTrans (Mistral)	0.579	0.619	0.669	0.588	0.938
LD-MLMTrans (Gemma-3)	0.582	0.621	0.671	0.590	0.940
LD-MLMTrans (Phi-4)	0.581	0.620	0.670	0.589	0.939
LD-MLMTrans (Llama, *IT*)	**0.642**	**0.732**	**0.746**	**0.683**	**0.986**
LD-MLMTrans (Mistral, *IT*)	0.637	0.727	0.741	0.678	0.982
LD-MLMTrans (Gemma-3, *IT*)	0.637	0.727	0.741	0.678	0.982
LD-MLMTrans (Phi-4, *IT*)	0.635	0.725	0.739	0.676	0.981

6. Result and Discussion

Table 6 presents the performance comparison of RoBERTa, LID-176 by fastText (Joulin et al., 2017, 2016), and lid218e from the No Language Left Behind project by Meta AI (NLLB Team et al., 2022; Costa-Jussà et al., 2022) on the BSLDC and ASLDC datasets for word-level language detection in Singlish text. As language detection is an integral component of our proposed LD-MLMTrans approach for Singlish to English translation, we first study the performance of the RoBERTa model on the language detection task. With reference to Table 6, it is evident that the performance of

LID (fastText) and LID (NLLB) is relatively poor in detecting Singlish words (non-English words) in Singlish text. Although the No Language Left Behind (NLLB) LID is trained for language detection across more than two hundred languages, it still lacks the ability to detect non-English words in Singlish text sequences accurately. Motivated by this limitation of existing LID models (e.g. fastText and NLLB) this study curates two datasets, BSLDC and ASLDC (refer to Section 3.2), to further train and fine-tune models for word-level language detection in Singlish. Table 6 shows that continuous training of LID models, such as fastText and NLLB, and fine-tuning of RoBERTa models on our proposed datasets, BSLDC and ASLDC, significantly improve their performance in word-level language detection for Singlish text sequences. These observations demonstrate that our proposed datasets are effective for training language detection models on Singlish. Table 6 also shows that the fine-tuned RoBERTa model outperforms both fastText and NLLB LID models on the BSLDC and ASLDC datasets. Furthermore, the performance of the RoBERTa model and both LID are significantly higher when fine-tuned on the proposed ASLDC corpus compared to their performance when trained on the BSLDC corpus. From results presented in Table 6, we conclude that our proposed language detection dataset is more effective for language detection in Singlish texts, as it includes a larger, comprehensive set of non-English Singlish words.

Table 8 presents the performance comparison between the baseline and the proposed methods to translate Singlish into English. Our proposed method comprises two setups: LD-MLMTrans and LD-MLMTrans(IT). While both approaches are similar, LD-MLMTrans uses a pre-trained LLM model with prompt instructions and masked language modelling, whereas LD-MLMTrans(IT) adopts instruction tuning, as discussed in Section 4.2. Table 8 presents the performance of baseline and proposed models when trained on the training set of SEAT and evaluated on the test set of SEAT datasets. From Table 8, it is evident that both setups of our proposed method $\mathbf{LD-MLMTrans}$ and $\mathbf{LD-MLMTrans(IT)}$ outperform **GPT-2** (Foo and Ng, 2024), **Multi-step Prompting** (Ng and Chan, 2024), and **Llama-3 (CoT)**. The superior performance of $\mathbf{LD\text{-}MLMTrans}$ and $\mathbf{LD\text{-}MLMTrans(IT)}$ over **Llama-3 (CoT)** and **Llama-3 (SFT)** underscores the effectiveness of integrating language detection with masked language modelling. The results clearly highlight the incremental gains achieved through LID-driven masking, beyond what is attainable with supervised finetuning with prompt instructions alone using the same pretrained Llama-3 model and training data. Such observations from comparison between performance of $\mathbf{LD\text{-}MLMTrans}$ and $\mathbf{LD\text{-}MLMTrans(IT)}$ over **Llama-3 (CoT)** and **Llama-3 (SFT)** further validate that explicit token-level language cues and selective masking provide complementary supervision signals, enabling the model to better capture cross-lingual alignments and code-mixed nuances in Singlish-to-English translation.

Furthermore, Table 8 demonstrate that $\mathbf{LD\text{-}MLMTrans(IT)}$ outperforms $\mathbf{LD\text{-}MLMTrans}$. This performance gain arises because $\mathbf{LD\text{-}MLMTrans}$ suffers from performance degradation when Singlish sentences are extremely short or from the lack of English words. In such cases, the *Selective Masking for Contextual Translation* module (Section 4.1) masks all tokens, leaving the model with insufficient context to predict the masked tokens. In contrast, instruction tuning in $\mathbf{LD\text{-}MLMTrans(IT)}$ enables the model to explicitly learn Singlish-to-English translation rather than relying solely on contextual prediction of masked tokens. Consequently, $\mathbf{LD\text{-}MLMTrans(IT)}$ achieves superior performance. Table 7 presents the performance comparisons of the $\mathbf{LD\text{-}MLMTrans(IT)}$ model when trained on the training set of the SEAT dataset and the SSET dataset and evaluated on the test set of the SEAT dataset. From the table, it is evident that training the $\mathbf{LD\text{-}MLMTrans(IT)}$ model on the SSET dataset significantly improves its performance compared to when it is trained on the SEAT dataset. From this observation, we conclude that our proposed SSET dataset is effective and reliable for training robust models for Singlish-to-English translation.

Furthermore, the following conclusions can be made from Table 7 and Table 8: (i) *Substantial Gains in Semantic Metrics*: LD-MLMTrans achieves significant improvements in semantic evaluation metrics, such as BERTScore and MoverScore, suggesting that LD-MLMTrans preserves the semantic meaning of the original Singlish sentences during translation. (ii) *Effective Low-resource Adaptation*: The superior performance of LD-MLMTrans over baseline models indicates its effectiveness in handling low-resource and code-mixed language scenarios in the Singlish language, where standard models tend to underperform; (iii) *Impact of Language Detection and Masked Modeling*: The superior performance of LD-MLMTrans over baseline models validates the effectiveness of integrating language detection with masked language modelling, underscoring the value of structured linguistic cues in guiding LLMs for more accurate and fluent translations in code-mixed contexts.

Table 9: Performance comparison of $\mathbf{Llama\text{-}3\ (CoT)}$ and the proposed methods on the **HinglishBuzz** dataset.

Method	BLEU	METEOR	ROUGE-L	Mover	BERT
$\mathbf{Llama\text{-}3\ (CoT)}$	0.358	0.503	0.679	0.528	0.871
$\mathbf{Llama\text{-}3\ (SFT)}$	0.382	0.526	0.698	0.552	0.888
LD-MLMTrans	0.546	0.562	0.705	0.571	0.918
LD-MLMTrans(IT)	0.612	0.620	0.719	0.584	0.926

6.1. Error Analysis

We manually analyze SEAT test set translations to identify and characterize errors produced by the LD-MLMTrans and LD-MLMTrans(IT) models in Singlish-to-English translation. Comparing human-annotated translations and those generated by the proposed LD-MLMTrans and LD-MLMTrans(IT) models, we conclude the following: (i) *Variations in Lexical Choices*: Translation by both human and LD-MLMTrans and LD-MLMTrans(IT) model preserve meaning but differ subtly in word choice.

(ii) *Cultural Expressions and Localised Terms*: For cultural terms such as *kiasu* or *sial*, both human and model translations provide reasonable interpretations. (iii) *Fluency and Structural Differences*: The outputs from LD-MLMTrans and LD-MLMTrans(IT) are structurally fluent and grammatically sound. However, in a few cases, the model generates more elaborate sentence structures than human translations, which are often more concise and direct. This indicates that the model tends to produce more explicit or complete sentence forms. iv) *Faithfulness to Implicit Meaning:* Human annotations tend to preserve the implicit tone and brevity characteristic of Singlish, while LD-MLMTrans occasionally expands on the content (v) *Stylistic Differences in Intensity and Emotion*: The model and human translation of Singlish occasionally differ in how emotional or emphatic expressions are rendered.

Our error analysis and comparison between human and D-MLMTrans translation of Singlish into English reveal that LD-MLMTrans and LD-MLMTrans(IT) perform reasonably well in generating fluent and semantically appropriate translations. The observed differences primarily relate to stylistic variation, interpretation of localised terms, and degrees of expressiveness.

6.2. Evaluation of LD-MLMTrans Beyond Singlish

To further examine the effectiveness and robustness of our proposed models, LD-MLMTrans and LD-MLMTrans(IT), we evaluated their performance on a Hindi-English code-mixed dataset. This study curates a small Hindi-English code-mixed dataset, **HinglishBuzz**, by collecting five hundred and fifty sentences from conversations on social media platforms and translating them into Hindi with the help of human annotators. We collected unique Hindi words from the newspaper corpus curated in the study by (Kumar et al., 2024) and unique English words from the Global News Dataset[13] to construct the vocabulary used for word-level language detection. Once we obtained the vocabulary of English and Hindi words, we fine-tuned the RoBERTa model for word-level language detection. Following the approach outlined in Subsection 4.1, we then masked the English words in Hindi-English mixed sentences. In this case, we aimed to translate the sentences into Hindi using sentences with mixed Hindi and English words. Masking is explicitly applied to the English words within the Hindi-English mixed input. Subsequently, we adopted the Contextual Masked Language Translation outline in subsection 4.2 with Llama-3 to translate the sentences into Hindi. Considering the superior performance of Llama-3 in Table 8, we consider only Llama-3 models as LLMs in our proposed LD-MLMTrans and LD-MLMTrans (IT) over the **HinglishBuzz** dataset. Table 9 presents the performance of the $LD-MLMTrans$ and $LD-MLMTrans\ (IT)$ and the baseline Llama-3 (CoT) model on the HinglishBuzz dataset. Comparing the performance of the proposed methods on the SSET and SEAT datasets for Singlish to English translation (Table 7, 8) and on the **HinglishBuzz** dataset (Table 9), we can conclude that our methods demonstrate consistent performance across different code-mixed settings. This suggests the proposed approach is robust to new languages and unseen datasets in code-mixed translation tasks.

7. Conclusion and Future Work

This study enhances the existing open-source Singlish dictionary by introducing an Augmented Singlish Dictionary. Next, we propose the Augmented Singlish Language Detection Corpus and the Singlish-English Aligned Translation (SEAT) dataset to foster the development of word-level language detection for Singlish text and Singlish-to-English translation, respectively. Furthermore, we propose the LD-MLMTrans approach for Singlish-to-English translation. Our proposed method first performs word-level language detection on the Singlish text to detect and mask non-English words. Subsequently, the masked Singlish sentence is then passed to an LLM, which predicts the masked tokens based on the surrounding context and translates the entire sentence into formal English. We conduct our experiments on two datasets, including our proposed SEAT dataset. The results demonstrate that LD-MLMTrans outperforms baseline models from the literature for Singlish-to-English translation and exhibits high proficiency in Singlish-to-English translation. This study identifies spoken

[13] Global News Dataset

Singlish Automatic Speech Recognition with translations as the scope of future work, as such undertaking holds immense potential to address and resolve communication challenges in real time between speakers with varying language capacities and preferences, most prominently found in healthcare contexts between patients and care providers.

Limitations

Despite the overall effectiveness of LD-MLMTrans in generating fluent and semantically appropriate translations, our error analysis reveals two key limitations. First, the model tends to generalise culturally specific expressions such as *kiasu* or *sial*, leading to translations that, while valid, lack the nuanced interpretation often present in human annotations.

Second, LD-MLMTrans occasionally produces overly explicit or elaborated outputs, which can diminish the implicit tone and brevity that are characteristic of Singlish. Another limitation of our proposed method, Language Detection-driven Masked Language Modeling (LD-MLMTrans), is that the model using only a masked language modelling setup may produce suboptimal results when the input sentence consists entirely of Singlish words. However, the instruction-tuned variant of our proposed method, LD-MLMTrans(IT), performs well in such cases due to its fine-tuning with prompt-based instructions and task-specific learning during training. These limitations highlight the need for further refinement in handling localised content and preserving stylistic subtleties in future work.

Ethical Consideration

The Augmented Singlish Dictionary, Augmented Singlish Language Detection Corpus and Singlish-English Aligned Translation (SEAT) datasets are curated without collecting any Personally Identifiable Information and Data (PII). The Augmented Singlish Dictionary, Augmented Singlish Language Detection Corpus, is curated by collecting Singlish words from an open-source available Singlish dictionary. In contrast, Singlish-English Aligned Translation (SEAT) is curated by manually translating Singlish text in an open-source available Singlish Short Message Service Corpus (Chen and Min-Yen, 2015). The Singlish Short Message Service Corpus (Chen and Min-Yen, 2015) is an open-source Singlish text used in this study to propose the SEAT dataset. We manually translated the Singlish text in the Singlish Short Message Service Corpus (Chen and Min-Yen, 2015) to obtain English translations and to be used in training or evaluating models for Singlish to English translations. The dataset is released under the Creative Commons Attribution 4.0 International license, and the accompanying code is released under the MIT license.

Acknowledgments

This work was supported by CArdiovascular DiseasE National Collaborative Enterprise (CADENCE) National Clinical Translational Program (MOH-001277-01).

Bibliographical References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

T Chen and Kan Min-Yen. 2015. The national university of singapore sms corpus.

Siew Yeng Chow, Chang-Uk Shin, and Francis Bond. 2024. This word mean what: Constructing a singlish dictionary with chatgpt. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)@ LREC-COLING 2024*, pages 41–50.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Soumi Datta, Deena Joseph Wilson, and Sainik Kumar Mahata. 2024. A survey on recent advancements in neural machine translation. In *International Conference on Smart Systems and Wireless Communication*, pages 187–200. Springer.

David Deterding. 2007. *Singapore English*. Edinburgh University Press.

Linus Tze En Foo and Lynnette Hui Xian Ng. 2024. Disentangling singlish discourse particles with task-driven representation. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, pages 1–6.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2025. Bridging the linguistic divide: A survey on leveraging large language models for machine translation. *arXiv preprint arXiv:2504.01919*.

Ebisa Gemechu and GR Kanagachidambaresan. 2023. Text-text neural machine translation: A survey. *Optical Memory and Neural Networks*, 32(2):59–72.

Zixuan Gong, Xiaolin Hu, Huayi Tang, and Yong Liu. Towards auto-regressive next-token prediction: In-context learning emerges from generalization. In *The Thirteenth International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shinichi Harada. 2009. The roles of singapore standard english and singlish. *Information Research*, 40:69–81.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Jia Xin Koh, Aqilah Mislan, Kevin Khoo, Brian Ang, Wilson Ang, Charmaine Ng, and Ying-Ying Tan. 2019. Building the singapore english national speech corpus. *Malay*, 20(25.0):19–3.

Sujit Kumar, Anant Shankhdhar, Divyam Singal, Bhuvan Aggarwal, Ahaan Sameer Malhotra, and Sanasam Ranbir Singh. 2024. Fake news article detection datasets for hindi language. *Language Resources and Evaluation*, pages 1–36.

Jakob RE Leimgruber. 2011. Singapore english. *Language and Linguistics Compass*, 5(1):47–62.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hexin Liu, Leibny Paola García Perera, Xinyi Zhang, Justin Dauwels, Andy W. H. Khong, Sanjeev Khudanpur, and Suzy J. Styles. 2021. End-to-End Language Diarization for Bilingual Code-Switching Speech. In *Interspeech 2021*, pages 1489–1493.

Hexin Liu, Xiangyu Zhang, Haoyang Zhang, Leibny Garcia-Perera, Andy W. H. Khong, Eng Chng, and Shinji Watanabe. 2025. Aligning speech to languages to enhance code-switching speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, PP:1–14.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu, Shikang Ni, Aiti Aw, and Nancy Chen. 2022. Singlish message paraphrasing: A joint task of creole translation and text normalization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3924–3936.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

William Merrill, Zhaofeng Wu, Norihito Naka, Yoon Kim, and Tal Linzen. 2024. Can you learn semantics through next-word prediction? the case of entailment. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2752–2773.

Wongyung Nam and Beakcheol Jang. 2024. A survey on multimodal bidirectional machine learning translation of image and natural language processing. *Expert Systems with Applications*, 235:121168.

Lynnette Hui Xian Ng and Luo Qi Chan. 2024. What talking you?: Translating code-mixed messaging texts to english. *arXiv preprint arXiv:2411.05253*.

Nourma Silvia Ningsih and Fadhlur Rahman. 2023. Exploring the unique morphological and syntactic features of singlish (singapore english). *Journal of English in Academic and Professional Communication*, 9(2):72–80.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao,

Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *CoRR*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal dependencies parsing for colloquial singaporean english. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1732–1744.

Lionel Wee. 2004. Singapore english: morphology and syntax. *A handbook of varieties of English*, 2:1058–1072.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781.